Lecture notes for Math63CM last week of classes

Lenya Ryzhik

March 17, 2020

1 Kepler's laws

Lenya's comment: The section on Kepler's laws notes is based on the lecture notes by Brian White for 63CM that he graciously allowed me to use for this class. I have added some comments here and there, and made other changes, but am mostly following Brian's notes closely. All mistakes are mine.

In the 1600s, Kepler, after analyzing the the astronomical observations of Tycho Brahe, formulated the following laws about the motion of the planets around the sun:

1. The orbit of each planet is an ellipse with the sun at one focus.
2. The line segment from the sun to the planet sweeps out equal areas in equal times. (That is, it sweeps out area at a constant rate).
3. The square of the period is proportional to the cube of the major axis of the ellipse.

In 1687, Newton deduced Kepler's Laws from his general laws of motion and his law of gravitation attraction. Here, we give such a derivation, following ideas of Simon Kochen (as presented in *Differential Equations* by Simmons and Kranz).

We will treat the sun and the planet as point masses. Actually, Newton took some care to justify that: he showed that if the bodies in question are spherically symmetric about their centers, then they move exactly as the would if they were point masses. Of course we are also ignoring the forces that other planets, moons, etc exert on a given planet, since those are much smaller than the force exerted by the sun. All our considerations will be in $\mathbb{R}^3$, the three-dimensional space.

Let $y(t) \in \mathbb{R}^3$ be the position of the planet at time t and $z(t)$ be the position of the sun (remember that we assume that both the sun and the planet are point masses) and let m be the mass of the planet, and M be the mass of the sun. The Newton's second law says that the equation of motion for the planet and the sun are

$$my''(t) = F, \quad Mz''(t) = -F \tag{1.1}$$

where F is the gravitational force of the sun acting on the planet. It is given by

$$F = GMm\frac{z(t) - y(t)}{\|z(t) - y(t)\|^3},$$

where G is the gravitational constant. Note that the gravitational force acts along the vector $y(t) - z(t)$ that is in the direction connecting the planet and the sun and its strength is

$$\|F\| = GMm\frac{\|z(t) - y(t)\|}{\|z(t) - y(t)\|^3} = \frac{GMm}{\|z(t) - y(t)\|^2}.$$

Now, (1.1) takes the form

$$my''(t) = GMm\frac{z(t) - y(t)}{\|z(t) - y(t)\|^3}, \quad Mz''(t) = GMm\frac{y(t) - z(t)}{\|z(t) - y(t)\|^3}. \tag{1.2}$$

Dividing by m in the first equation in (1.2), and by M in the second, gives

$$y''(t) = GM\frac{z(t) - y(t)}{\|z(t) - y(t)\|^3}, \quad z''(t) = Gm\frac{y(t) - z(t)}{\|z(t) - y(t)\|^3}. \tag{1.3}$$

Exercise 1.1. Define the center of mass

$$X(t) = \frac{my(t) + Mz(t)}{m + M},$$

and the total momentum

$$p(t) = my'(t) + Mz'(t),$$

and show that $p(t) = p(0)$ for all $t \geq 0$. This is the law of conservation of the total momentum. Deduce then that $X(t) = X(0) + p(0)t$ – the center of mass moves with a constant speed. Hint: use (1.2).

Consider the position of the planet relative to the sun: set

$$x(t) = y(t) - z(t). \tag{1.4}$$

Subtracting the second equation in (1.3) from the first gives

$$x''(t) = -G(M + m)\frac{x(t)}{\|x(t)\|^3}. \tag{1.5}$$

We set

$$G_0 = G(M + m), \tag{1.6}$$

and write (1.5) as

$$x''(t) = -G_0\frac{x(t)}{\|x(t)\|^3}. \tag{1.7}$$

Remark 1.2. We will always assume that initially the sun and the planet are not at the same point, so that $x(0) \neq 0$. This will ensure that solution to (1.7) exists for some time interval $0 \leq t < T$, with some $T > 0$. We will later see that the existence time $T = +\infty$ except in some very special cases.

The conservation of the angular momentum and Kepler's second law

Here is the first result, which is the conservation of the angular momentum. Recall that $x \times y$ is the standard vector product if $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ are two vectors in $\mathbb{R}^3$, then $x \times y$ is the vector in $\mathbb{R}^3$ with the components

$$x \times y = (x_2y_3 - x_3y_2, x_3y_1 - x_1y_3, x_1y_2 - x_2y_1).$$

It is easy to check that $(x \times y) \cdot x = (x \times y) \cdot y = 0$, so that $x \times y$ is orthogonal both to x and to y.

Exercise 1.3. Let $f(t)$ and $g(t)$ be two differentiable vector-valued functions of $t \in \mathbb{R}$. Show that

$$\frac{d}{dt}(f(t) \times g(t)) = f'(t) \times g(t) + f(t) \times g'(t). \tag{1.8}$$

Theorem 1.4. *Let $x(t)$ be a solution to (1.7), then $x(t) \times x'(t) = x(0) \times x'(0)$ for all $t \geq 0$.*

Proof. Using the product differentition formula (1.8) gives

$$\frac{d}{dt}(x(t) \times x'(t)) = x'(t) \times x'(t) + x(t) \times x''(t) = x(t) \times x''(t).$$

We used above the fact that we have $x'(t) \times x'(t) = 0$ automatically from the definition of the vector product. In addition, we see from (1.7) that $x(t) \times x''(t) = 0$ because $x''(t)$ is a scalar multiple of $x(t)$. We deduce that

$$\frac{d}{dt}(x(t) \times x'(t)) = 0,$$

which means that $x(t) \times x'(t) = x(0) \times x'(0)$, as claimed. □

As a consequence, if $\Omega_0 \neq 0$, then both $x(t)$ and $x'(t)$ are orthogonal to $\Omega_0 = x(0) \times x'(0)$ for all $t \geq 0$. In particular, this means that $x(t)$ lies in the plane that passes through the origin and is orthogonal to Ω_0 – the planet motion is restricted to this plane for all $t \geq 0$. The case $\Omega_0 = 0$ is special, we will consider it separately below. Let us record this for the future.

Corollary 1.5. *If $\Omega_0 = x(0) \times x'(0) \neq 0$, then the trajectory $x(t)$ lies in the two dimensional plane that passes through $x(0)$ and the origin, and is orthogonal to the vector Ω_0.*

We may now understand Kepler's second law. Assume that $\Omega_0 \neq 0$ and let $A(t)$ be the area swept by the vector $x(t)$ in the plane

$$\Pi_0 = \{y \in \mathbb{R}^3 : \ y \cdot \Omega_0 = 0\} \tag{1.9}$$

in the time interval $[0, t]$.

Exercise 1.6. Show that

$$A'(t) = \frac{1}{2}|x(t) \times x'(t)|.$$

Hint: this is because the area swept out during a small time interval $[t, t + \Delta t]$ is approximately the area of the triangle with the sides $x(t)$ and $x'(t)\Delta t$.

Corollary 1.7. *[Kepler's second law] The area $A(t)$ swept out in the time interval $[0, t]$ is*

$$A(t) = \frac{t}{2}\|x(0) \times x'(0)\|.$$

Proof. This is an immediate consequence of Theorem 1.4 and Exercise 1.6. □

Remark 1.8. Note that the theorem and its corollary hold for any "central" force, i.e., any force that at each time is a scalar multiple of the position vector. (Newton already pointed out this out.)

Conservation of energy and boundedness of orbits

We now explain the connection between the total energy of the planet and the sun to the boundedness of its trajectory. In order to define the total physical energy, let us recall that $y(t)$ is the position of the planet, $z(t)$ is the position of the sun, and the physical kinetic energy of the sun and the planet are, respectively:

$$E_{kin,s} = \frac{M\|z'\|^2}{2}, \quad E_{kin,s} = \frac{m\|y'\|^2}{2}, \tag{1.10}$$

so that the total kinetic energy is

$$\tilde{E}_{kin} = \frac{M\|z'\|^2}{2} + \frac{m\|y'\|^2}{2}. \tag{1.11}$$

It is convenient to re-write the energy in terms of $x(t) = y(t) - z(t)$ and the center of mass

$$X(t) = \frac{my(t) + Mz(t)}{m + M},$$

so that

$$y(t) = X(t) + \frac{M}{m+M}x(t), \quad z(t) = X(t) - \frac{m}{m+M}x(t), \tag{1.12}$$

and

$$\begin{aligned} \tilde{E}_{kin} &= \frac{M\|z'\|^2}{2} + \frac{m\|y'\|^2}{2} = \frac{M}{2}\left\|X'(t) - \frac{m}{m+M}x'(t)\right\|^2 + \frac{m}{2}\left\|X'(t) + \frac{M}{m+M}x'(t)\right\|^2 \\ &= \frac{M+m}{2}\|X'(t)\|^2 + \frac{(Mm^2 + mM^2)\|x'(t)\|^2}{2(M+m)^2} = \frac{M+m}{2}\|X'(t)\|^2 + \frac{Mm}{2(M+m)}\|x'(t)\|^2. \end{aligned} \tag{1.13}$$

The physical potential energy of the sun and the planet is

$$\tilde{E}_{pot} = -\frac{GMm}{\|y(t) - z(t)\|} = -\frac{GMm}{\|x(t))\|}. \tag{1.14}$$

Thus, the total physical energy of the system is

$$\begin{aligned}\tilde{E}_{tot} &= \frac{M+m}{2}\|X'(t)\|^2 + \frac{Mm}{2(M+m)}\|x'(t)\|^2 - \frac{Mm}{\|x(t))\|} \\ &= \frac{M+m}{2}\|X'(t)\|^2 + \frac{Mm}{M+m}\Big(\frac{\|x'(t)\|^2}{2} - \frac{G(M+m)}{\|x(t))\|}\Big) \\ &= E_{c,kin}(t) + \frac{Mm}{M+m}\Big(\frac{\|x'(t)\|^2}{2} - \frac{G_0}{\|x(t))\|}\Big).\end{aligned} \tag{1.15}$$

We used the definition (1.6) of G_0 in the last step above and also defined the kinetic energy of the center of mass as

$$E_{c,kin}(t) = \frac{M+m}{2}\|X'(t)\|^2.$$

Recall that the center of mass $X(t)$ moves with a constant speed, as shown in Exercise 1.1, so that $E_{c,kin}(t)$ is constant in time. Thus, to simplify slightly the notation, we define the mathematical total energy of the system as

$$E(t) = \frac{\|x'(t)\|^2}{2} - \frac{G_0}{\|x(t))\|}. \tag{1.16}$$

It is related to the total physical energy $\tilde{E}_{tot}(t)$ by

$$\tilde{E}_{tot}(t) = E_{c,kin}(t) + \frac{Mm}{M+m}E(t). \tag{1.17}$$

We also define the mathematical kinetic energy as

$$E_{kin}(t) = \frac{1}{2}\|x'(t)\|^2, \tag{1.18}$$

and the mathematical potential energy as

$$E_{pot}(t) = -\frac{G_0}{\|x(t)\|},$$

Let $x(t)$ be the solution to (1.7)

$$x''(t) = -G_0\frac{x(t)}{\|x(t)\|^3}. \tag{1.19}$$

Let us show that the total energy is preserved.

Theorem 1.9. *We have $E(t) = E(0)$ for all $t \geq 0$.*

Proof. Let us first compute the evolution of the kinetic energy:

$$\frac{dE_{kin}}{dt} = \frac{1}{2}\frac{d}{dt}(\|x'(t)\|^2) = (x'(t)\cdot x''(t)) = -\frac{G_0}{\|x(t)\|^3}(x(t)\cdot x'(t)). \tag{1.20}$$

We used (1.19) in the last step. For the potential energy we have

$$\frac{dE_{pot}}{dt} = -G_0\frac{d}{dt}\Big(\frac{1}{\|x(t)\|}\Big) = \frac{G_0}{\|x(t)\|^2}\frac{d\|x(t)\|}{dt} = \frac{G_0}{2\|x(t)\|^3}\frac{d\|x(t)\|^2}{dt}. \tag{1.21}$$

For the last step, we used the identity

$$\frac{dg}{dt} = \frac{1}{2g}\frac{d(g^2)}{dt},$$

with $g(t) = \|x(t)\|$. We continue (1.21):

$$\frac{dE_{pot}}{dt} = \frac{G_0}{2\|x(t)\|^3}\frac{d\|x(t)\|^2}{dt} = \frac{G_0}{\|x(t)\|^3}(x(t)\cdot x'(t)). \tag{1.22}$$

We deduce from (1.20) and (1.22) that the total energy is conserved:

$$\frac{dE}{dt} = \frac{dE_{kin}}{dt} + \frac{dE_{pot}}{dt} = 0, \tag{1.23}$$

hence $E(t) = E(0)$, and the proof is complete. □

The conservation of energy has an interesting implication for the trajectories. It follows from Theorem 1.9 that

$$\frac{\|x'(t)\|^2}{2} - \frac{G_0}{\|x(t)\|} = E(0), \tag{1.24}$$

with the initial energy

$$E(0) = \frac{\|x'(0)\|^2}{2} - \frac{G_0}{\|x(0)\|}. \tag{1.25}$$

As a consequence, if $E(0) < 0$, we have

$$\frac{G_0}{\|x(t)\|} = \frac{\|x'(t)\|^2}{2} + |E(0)| \geq |E(0)|, \tag{1.26}$$

and thus

$$\|x(t)\| \leq \frac{G_0}{|E(0)|}. \tag{1.27}$$

In particular, if $E(0) < 0$ the the trajectory $x(t)$ is bounded. We have proved the following theorem.

Theorem 1.10. *Assume that*

$$\frac{\|x'(0)\|^2}{2} < \frac{G_0}{\|x(0)\|}, \tag{1.28}$$

so that $E(0)$ given by (1.25) is negative, then

$$\|x(t)\| \leq R_0, \quad \textit{for all } t \geq 0, \tag{1.29}$$

with

$$R_0 = \frac{G_0}{|E(0)|}. \tag{1.30}$$

In other words, given the initial position $x(0)$ of the planet, if the initial velocity is not too large, in the sense that (1.28) holds, then the planet always stays within the distance R_0 from the sun. That is, the trajectory is a curve in the plane Π_0 defined in (1.9) that stays in the disk of radius R_0 around the origin.

Note that R_0 defined in (1.30) tends to $+\infty$ as $|E(0)| \to 0$. This indicates that if $E(0) \geq 0$, that is, the initial velocity is sufficiently large, so that

$$\frac{\|x'(0)\|^2}{2} \geq \frac{G_0}{\|x(0)\|}, \tag{1.31}$$

then we may expect that

$$\|x(t)\| \to +\infty \text{ as } t \to +\infty, \tag{1.32}$$

and the planet moves further and further away from the sun as $t \to +\infty$. However, at the moment we do not know that: R_0 is just an upper bound for the trajectory distance to the origin, hence all we know is that if (1.31) holds then (1.32) would not contradict conservation of energy, not that (1.32) is actually true.

The case of zero angular momentum

Before we analyze the trajectories in general, let us consider the simplest case of zero anguklar momentum:

$$\Omega_0 = x(0) \times x'(0) = 0. \tag{1.33}$$

As we always assume that $x(0) \neq 0$, this means that the initial velocity $x'(0)$ is either parallel to $x(0)$ or zero:

$$x'(0) = v_0 e, \tag{1.34}$$

with $e = x(0)/\|x(0)\|$ being the unit vector in the direction of $x(0)$. Conservation of the angular momentum implies that then

$$x(t) \times x'(t) = 0, \tag{1.35}$$

so that, as long as $x(t)$ does not hit the origin, the velocity $x'(t)$ is parallel to the position $x(t)$. We claim that then, as long as the solution $x(t)$ exists, it has the form

$$x(t) = r(t)e, \tag{1.36}$$

with a scalar-valued function $r(t) \geq 0$. Indeed, let the scalar-valued functions $r(t)$ and $v(t)$ be the solutions to

$$\begin{aligned} r'(t) &= v(t), \\ v'(t) &= -\frac{G_0}{r^2(t)}, \end{aligned} \tag{1.37}$$

with the initial condition $r(0) = \|x(0)\|$ and $v(0) = v_0$, as in (1.34). Then it is straightforward to check that $\tilde{x}(t) = r(t)e$ is the solution to (1.7):

$$\tilde{x}''(t) = -\frac{G_0 \tilde{x}(t)}{\|\tilde{x}(t)\|^3}, \tag{1.38}$$

with the initial condition

$$\begin{aligned} \tilde{x}(0) &= r(0)e = \|x(0)\| \frac{x(0)}{\|x(0\|} = x(0), \\ \tilde{x}'(0) &= r'(0)e = v_0 e = x'(0). \end{aligned} \tag{1.39}$$

It follows from the uniqueness of the solutions to (1.7) with a prescribed $x(0)$ and $x'(0)$ that $x(t) = \tilde{x}(t)$, and thus $x(t)$ has the form (1.36).

Let us now discuss how solutions to (1.37) may behave. It is clear that the function $v(t)$ is decreasing as $v'(t) < 0$, and that solution exists as long as $r(t)$ does not hit zero. The next exercise says that as soon as we know that the planet velocity points toward the sun at some time $t_0 > 0$ then the planet will reach the sun in a finite time.

Exercise 1.11. Show that if there exists $t_0 > 0$ so that $v(t_0) \leq 0$ then there exists $T > 0$ such that $r(t) \to 0$ as $t \to T^-$.

The next exercise shows that if the initial velocity of the planet points away from the sun and is sufficiently large then the planet always moves away from the sun.

Exercise 1.12. Show that for each $r > 0$ there exists $\bar{v}(r) > 0$ so that if $r(0) = r$ and $v(0) > \bar{v}(r)$ then we have $v(t) > v(0)/2$ for all $t \geq 0$. Hint: consider the first time t_0 such that $v(t_0) = v(0)/2$ and get a contradiction if $v(0)$ is sufficiently large. You may need to estimate $r(t)$ from below for all $0 \leq t \leq t_0$, and the integral

$$\int_0^t \frac{G_0}{r^2(s)} ds$$

from above.

As a consequence of Exercise 1.12 we deduce that for $v(0) > \bar{v}(r)$. the function $r(t)$ is increasing in t, and

$$r(t) \geq r(0) + \frac{v(0)}{2} t,$$

so that $r(t) \to +\infty$ as $t \to +\infty$. We also have a comparison principle.

Exercise 1.13. Show that if $(r_1(t), v_1(t))$ and $(r_2(t), v_2(t))$ are two solutions to (1.37) such that $r_1(0) \le r_2(0)$ and $v_1(0) \le v_2(0)$, then $r_1(t) \le r_2(t)$ for all $t \ge 0$.

Corollary 1.14. *For each $r_0 > 0$ here exists $\bar{v}_0(r)$ such that if $r(0) = r_0$ and $v(0) > \bar{v}_0(r)$ then the solution to (1.37) satisfies*

$$\lim_{t \to +\infty} r(t) = +\infty, \tag{1.40}$$

while if $v(0) < \bar{v}_0(r)$, then there exists $T > 0$ so that

$$\lim_{t \to T^-} r(t) = 0. \tag{1.41}$$

Thus, in the case of zero angular momentum we have a clear dichotomy: either the planet trajectory escapes, in the sense that (1.40) holds, or it collapses onto the sun, in the sense that (1.41) holds at a finite time $T > 0$.

Conic sections

We will see below that when the angular momentum does not vanish, the trajectory of the planet is either an ellipse, a parabola or a hyperbola in the plane Π_0 that passes through the origin and is orthogonal to Ω_0:

$$\Pi_0 = \{y \in \mathbb{R}^3 : \ y \cdot \Omega_0 = 0\}. \tag{1.42}$$

Let us recall the following characterization of these curves. Let $\alpha \in \mathbb{R}$ and $k > 0$ be fixed and consider the set of points in the plane that is described in the polar coordinates (r, θ) by

$$\alpha^2 = r(1 + k\cos\theta). \tag{1.43}$$

Proposition 1.15. *If $k = 0$, the curve given by (1.43) is a circle centered at the origin, if $0 < k < 1$, it is an ellipse with one of the two foci at the origin, if $k = 1$, it is a parabola with focus at the origin, and if $k > 1$, it is a hyperbola with one of the two foci at the origin.*

Proof. If $k = 0$, then (1.43) is simply $r = \alpha^2$, which is a circle of radius α centered at the origin. Let us assume that $0 < k < 1$. The curve (1.43) intersects the x axis at the points $(-r_1, 0)$ and $(r_2, 0)$ with

$$r_1 = \frac{\alpha^2}{1-k}, \quad r_2 = \frac{\alpha^2}{1+k}. \tag{1.44}$$

Hence, if the curve is, indeed, an ellipse, the second focus should be located at the point $\bar{x} = (-\bar{r}, 0)$, with

$$\bar{r} = r_1 - r_2 = \frac{2\alpha^2 k}{1-k^2}. \tag{1.45}$$

Let $x = (r\cos\theta, r\sin\theta)$ be a point on the curve. Its distance to $\bar{x}$ is

$$\|x - \bar{x}\| = \Big((r\cos\theta + \bar{r})^2 + r^2\sin^2\theta\Big)^{1/2} = \Big(r^2 + \bar{r}^2 + 2r\bar{r}\cos\theta\Big)^{1/2} = \Big(r^2 + \bar{r}^2 + \frac{2\bar{r}(\alpha^2 - r)}{k}\Big)^{1/2}. \tag{1.46}$$

We used (1.43) in the last step. Continuing (1.46) gives

$$\|x - \bar{x}\| = \Big(r^2 + \bar{r}^2 + \frac{2\bar{r}(\alpha^2 - r)}{k}\Big)^{1/2} = \Big(r^2 - \frac{2\bar{r}r}{k} + \bar{r}^2 + \frac{2\bar{r}\alpha^2}{k}\Big)^{1/2}. \tag{1.47}$$

Note that, miraculously, we have

$$\bar{r}^2 + \frac{2\bar{r}\alpha^2}{k} = \bar{r}\Big(\bar{r} + \frac{2\alpha^2}{k}\Big) = 2\alpha^2\bar{r}\Big(\frac{k}{1-k^2} + \frac{1}{k}\Big) = \frac{2\alpha^2\bar{r}}{k(1-k^2)} = \frac{\bar{r}^2}{k^2}. \tag{1.48}$$

Using this in (1.47) gives

$$\|x - \bar{x}\| = \Big(r^2 - \frac{2\bar{r}r}{k} + \bar{r}^2 + \frac{2\bar{r}\alpha^2}{k}\Big)^{1/2} = \Big(r^2 - \frac{2\bar{r}r}{k} + \frac{\bar{r}^2}{k^2}\Big)^{1/2} = \Big|\frac{\bar{r}}{k} - r\Big| = \Big|\frac{2\alpha^2}{1-k^2} - r\Big|. \tag{1.49}$$

However, (1.43) implies that

$$r = \frac{\alpha^2}{1 + k\cos\theta} \leq \frac{\alpha^2}{1-k} < \frac{2\alpha^2}{(1-k)(1+k)} = \frac{\bar{r}}{k}. \tag{1.50}$$

We used the fact that $0 \leq k < 1$ in the second inequality above. This gives

$$\|x - \bar{x}\| = \Big|\frac{2\alpha^2}{1-k^2} - r\Big| = \frac{2\alpha^2}{1-k^2} - r. \tag{1.51}$$

It follows that

$$\|x - \bar{x}\| + \|x\| = \frac{2\alpha^2}{1-k^2}. \tag{1.52}$$

Hence, (1.43), indeed, is an ellipse with the foci at the points $(-\bar{r}, 0)$ and $(0, 0)$. □

Exercise 1.16. Check that if $k = 1$ then (1.43) describes a parabola, and if $k > 1$ then it is one branch of a hyperbola with one focus at the origin.

Elliptic orbits and Kepler's first law

Let us go back to Kepler's laws, with the assumption that the angular momentum $\Omega_0 = x(0) \times x'(0)$ is non-zero. So far, we have shown that the trajectory $x(t)$ stays in the plane Π_0 passing through the origin that is orthogonal to the vector Ω_0. We will now use Proposition 1.15 to show that if the trajectory is bounded then it has to be an ellipse. In other words, if the planet stays within a finite distance from the sun for all $t \geq 0$, then its trajectory is an ellipse. We will see that this is a condition on the initial position $x(0)$ and velocity $x'(0)$. If this condition does not hold, then the trajectory may be a hyperbola or a parabola. In order to set-up the polar coordinates system to which we will apply Proposition 1.15, let $r(t) = \|x(t)\|$ be the distance from the sun to the planet, and

$$\omega(t) = \frac{x(t)}{|x(t)|} \in \mathbb{R}^3,$$

be the direction of the vector connecting the sun and the planet, so that $x(t) = r(t)\omega(t)$. Note that $\|\omega(t)\| = 1$ for all $t \geq 0$. To define the angle variable, consider the vector

$$Y = \Big(\frac{\|x'(0)\|^2}{G_0} - \frac{1}{\|x(0)\|}\Big)x(0). \tag{1.53}$$

Note that Y lies in the plane Π_0 because $x(0)$ does. If $Y \neq 0$, then, we let $\theta(t)$ be the angle formed by $\omega(t)$ with respect to the vector Y, so that

$$\cos\theta(t) = \frac{(\omega(t) \cdot Y)}{k},$$

where $k = \|Y\|$. If it happens that $Y = 0$, then we choose $\theta(t)$ to be the angle formed by $\omega(t)$ and any fixed unit vector in Π_0. Finally, set

$$\alpha = \frac{\|\Omega_0\|}{\sqrt{G_0}}, \tag{1.54}$$

where $\Omega_0 = x(0) \times x'(0)$. Our goal is to prove the following.

Theorem 1.17. *Let $x(t)$ be a solution to (1.7), then, for all $t \geq 0$ we have*

$$\alpha^2 = r(t)(1 + k\cos\theta(t)). \tag{1.55}$$

Proof. First, note that

$$x'(t) = \frac{d}{dt}(r(t)\omega(t)) = r'(t)\omega(t) + r(t)\omega'(t).$$

Since we also have $\omega(t) \times \omega(t) = 0$, we can write, using conservation of the angular momentum in Theorem 1.4

$$\begin{aligned} \Omega_0 = x(0) \times x'(0) &= x(t) \times x'(t) = r(t)\omega(t) \times (r'(t)\omega(t) + r(t)\omega'(t)) \\ &= r(t)r'(t)\omega(t) \times \omega(t) + r^2(t)\omega(t) \times \omega'(t) = r^2(t)\omega(t) \times \omega'(t). \end{aligned} \tag{1.56}$$

Make sure you understand which quantities are scalars and which ones are vectors in the above computation! As $\|\omega(t)\|^2 = 1$ for all $t \geq 0$, we see that

$$\frac{d}{dt}\|\omega(t)\|^2 = (\omega(t) \cdot \omega'(t)) = 0, \tag{1.57}$$

and thus $\omega(t)$ is orthogonal to $\omega'(t)$. As, in addition, we have $\|\omega(t)\| = 1$, we conclude that

$$\|\Omega_0\| = r^2(t)\|\omega(t) \times \omega'(t)\| = r^2(t)\|\omega(t)\|\|\omega'(t)\| = r^2(t)\|\omega'(t)\|.$$

It follows, in particular, that if $\Omega_0 = 0$, then $\|\omega'(t)\| = 0$, so that $\omega(t)$ is constant. In this case, the planet moves in a straight line through the origin, as we have discussed in detail above.

Thus, from now on we will assume that $\Omega_0 \neq 0$. Let us recall the equation of motion (1.7) and write it as

$$x''(t) = -G_0 \frac{x(t)}{\|x(t)\|^3} = -G_0 \frac{r(t)\omega(t)}{r(t)^3} = -G_0 \frac{\omega(t)}{r(t)^2}. \tag{1.58}$$

We have then

$$\begin{aligned} x''(t) \times \Omega_0 = x''(t) \times (x(t) \times x'(t)) &= \left(-G_0 \frac{\omega(t)}{r^2(t)}\right) \times (r^2(t)\omega \times \omega'(t)) \quad \text{(by (1.56))} \\ &= -G_0 \omega(t) \times (\omega(t) \times \omega'(t)). \end{aligned} \tag{1.59}$$

Exercise 1.18. Show that if $\mathbf{u}$ is a unit vector in $\mathbb{R}^3$, and $\mathbf{v}$ is orthogonal to $\mathbf{u}$, then $\mathbf{u} \times (\mathbf{u} \times \mathbf{v}) = -\mathbf{v}$. Hint: first, show that we may assume that $\mathbf{u} = \mathbf{i}$ and $\mathbf{v} = \|\mathbf{v}\|\,\mathbf{j}$, where $\mathbf{i}, \mathbf{j}, \mathbf{k}$ form the standard basis for $\mathbb{R}^3$. In that case, observe that

$$\mathbf{u} \times (\mathbf{u} \times \mathbf{v}) = \mathbf{i} \times (\mathbf{i} \times (\|\mathbf{v}\|\,\mathbf{j})) = \mathbf{i} \times (\|\mathbf{v}\|\,\mathbf{k}) = -\|\mathbf{v}\|\,\mathbf{j} = -\mathbf{v}.$$

As $\omega(t)$ is a unit vector that is orthogonal to $\omega'(t)$, we deduce from (1.59) and Exercise 1.18 that

$$x''(t) \times \Omega_0 = G_0 \omega'(t),$$

so that

$$\frac{d}{dt}(x'(t) \times \Omega_0) = G_0 \omega'(t).$$

Integrating in t gives

$$x'(t) \times \Omega_0 - x'(0) \times \Omega_0 = G_0(\omega(t) - \omega(0)),$$

or

$$x'(t) \times \Omega_0 = G_0\omega(t) + x'(0) \times \Omega_0 - G_0\omega(0) = G_0\omega(t) + G_0 Y, \tag{1.60}$$

with

$$\begin{aligned} Y &= \frac{1}{G_0}\Big(x'(0) \times \Omega_0 - G_0\omega(0)\Big) = \frac{1}{G_0}\Big(x'(0) \times [x(0) \times x'[0] - \frac{G_0}{\|x(0)\|}x(0)\Big) \\ &= \Big(\frac{\|x'(0)\|^2}{G_0} - \frac{1}{\|x(0)\|}\Big)x(0), \end{aligned} \tag{1.61}$$

independent of t, as in (1.53). We used Exercise 1.18 once again above. Note that the vector Y also lies in the plane orthogonal to Ω_0 that passes through the origin, as does $x(t)$ for all $t \geq 0$. Let's take the inner product of (1.60) with $x(t)$:

$$x(t) \cdot (x'(t) \times \Omega_0) = x(t) \cdot (G_0\omega(t) + G_0 Y) = r(t)\omega(t) \cdot (G_0\omega(t) + Y) = r(t)G_0(1 + \omega(t) \cdot Y). \tag{1.62}$$

Using the identity $A \cdot (B \times C) = \det(A, B, C) = \det(C, A, B) = C \cdot (A \times B)$, we have from (1.62)

$$\Omega_0 \cdot (x(t) \times x'(t)) = r(t)G_0(1 + \omega(t) \cdot Y). \tag{1.63}$$

But $x(t) \times x'(t) = \Omega_0$, so (1.63) is simply

$$\|\Omega_0\|^2 = r(t)G_0(1 + \omega(t) \cdot Y). \tag{1.64}$$

Let $k = \|Y\|$ and let $\theta(t)$ be the angle between $\omega(t)$ and Y, so that

$$\omega(t) \cdot Y = \|\omega(t)\|\|Y\| \cos\theta(t) = k\cos\theta(t).$$

Then (1.64) takes the form

$$\|\Omega_0\|^2 = r(t)G_0(1 + k\cos\theta(t)), \tag{1.65}$$

finishing the proof of Theorem 1.17. if we take into account the definition (1.54) of α. □

Kepler's first law is a consequence of Theorem 1.17 and Proposition 1.15, which shows that (1.65) describes a circle, an ellipse, a parabola or a hyperbola. Therefore, any bounded trajectory is an ellipse (of which a circle is a special case), which proves Kepler's first law. We also now have a sharp criterion for when the planet will forever stay in a bounded region around the sun: for that, we need $k = \|Y\| < 1$. Recall that Y is given explicitly in terms of $x(0)$ and $x'(0)$ by (1.53), so that the condition $\|Y\| < 1$ is equivalent to

$$\Big|\frac{\|x'(0)\|^2}{G_0} - \frac{1}{\|x(0)\|}\Big| < \frac{1}{\|x(0)\|}. \tag{1.66}$$

This should be compared to the sufficient condition (1.25) for this to be true in Theorem 1.10, that can be written as

$$\frac{\|x'(0)\|^2}{2G_0} < \frac{1}{\|x(0)\|}. \tag{1.67}$$

We see that actually these two conditions are equivalent, so that (1.25) is not only a sufficient but also a necessary condition for the planet trajectory to stay a bounded distance from the sun. Of course, the analysis in the present section tells us much more than just that: it says that every bounded trajectory is an ellipse.

Kepler's third law: the periods

We now look at Kepler's third law that says that the square of the period T of the planet trajectory (which we by now know is an ellipse) is proportional to the cube of the major axis of the ellipse (the longest chord of the ellipse). More precisely, we will show that

$$T = \frac{2\pi a^{3/2}}{\sqrt{G_0}}, \tag{1.68}$$

where a is the major semi-axis (half of the longest chord of the ellipse). Note that the constant G_0 does not depend on the trajectory: this relation holds for all trajectories. The period T and the major semi-axis a do, of course, depend on the individual trajectory.

To show that (1.68) holds, let us first compute the area of an ellipse given by an equation of the form

$$\alpha^2 = r(1 + k\cos\theta), \tag{1.69}$$

as in Proposition 1.15. To do this, we need to find its two semi-axes: the area of an ellipse is

$$A_0 = \pi ab, \tag{1.70}$$

where a and b are, respectively, its major semi-axis, and minor semi-axis (half the length of the longest chord that is perpendicular to the chord of length $2a$). As we have seen in the proof of Proposition 1.15, the major axis is along the x-axis and connects the points $(-r_1, 0)$ and $(r_2, 0)$, with r_1 and r_2 as in (1.44):

$$r_1 = \frac{\alpha^2}{1-k}, \quad r_2 = \frac{\alpha^2}{1+k}, \tag{1.71}$$

so that

$$a = \frac{r_1 + r_2}{2} = \frac{\alpha^2}{2(1-k)} + \frac{\alpha^2}{2(1+k)} = \frac{\alpha^2}{1-k^2}. \tag{1.72}$$

To find the minor semi-axis, note that, as we have shown in the proof of Proposition 1.15, the two foci are at the points $\bar{x} = (-\bar{r}, 0)$ and $(0, 0)$, with $\bar{r}$ given by (1.45):

$$\bar{r} = r_1 - r_2 = \frac{2\alpha^2 k}{1 - k^2}. \tag{1.73}$$

Moreover, every point on the ellipse satisfies (1.52):

$$\|x - \bar{x}\| + \|x\| = \frac{2\alpha^2}{1 - k^2}. \tag{1.74}$$

Note that the minor axis connects the points $x_r = (-r_m, -b)$ and $x_b = (-r_m, b)$ that both lie on the ellipse, where $r_m = \bar{r}/2$. Since $\|x_b - \bar{x}\| = \|x_b\|$ from symmetry, using $x = x_b$ in (1.74) gives

$$\|x_b - \bar{x}\| + \|x_b\| = 2(r_m^2 + b^2)^{1/2} = \frac{2\alpha^2}{1 - k^2}, \tag{1.75}$$

hence

$$b^2 = \Big(\frac{\alpha^2}{1 - k^2}\Big)^2 - r_m^2 = \Big(\frac{\alpha^2}{1 - k^2}\Big)^2 - \Big(\frac{\alpha^2 k}{1 - k^2}\Big)^2 = \frac{\alpha^4(1 - k^2)}{(1 - k^2)^2} = \frac{\alpha^4}{1 - k^2}, \tag{1.76}$$

so that

$$b = \frac{\alpha^2}{(1 - k^2)^{1/2}}. \tag{1.77}$$

We conclude that the area of the ellipse is

$$A_0 = \pi a b = \frac{\pi \alpha^4}{(1 - k^2)^{3/2}}. \tag{1.78}$$

In order to compute the period of the planet motion around the ellipse we recall Corollary 1.7, Kepler's second law: the area $A(t)$ swept out in the time interval $[0, t]$ is

$$A(t) = \frac{t}{2}\|x(0) \times x'(0)\| = \frac{t}{2}\|\Omega_0\|.$$

The period T of the planet motion satisfies $A_0 = A(T)$, which, after using (1.78), becomes

$$\frac{\pi \alpha^4}{(1 - k^2)^{3/2}} = \frac{T}{2}\|\Omega_0\|,$$

which can be re-written with the help of (1.54) as

$$\frac{\pi \alpha^4}{(1 - k^2)^{3/2}} = \frac{T \alpha \sqrt{G_0}}{2},$$

or

$$T = \frac{2\pi \alpha^3}{\sqrt{G_0}(1 - k^2)^{3/2}}.$$

Finally, (1.72) gives

$$T = \frac{2\pi a^{3/2}(1 - k^2)^{3/2}}{\sqrt{G_0}(1 - k^2)^{3/2}} = \frac{2\pi a^{3/2}}{\sqrt{G_0}}, \tag{1.79}$$

which is Kepler's third law (1.68).

2 Population models

The Lotka-Volterra predator-prey model

The Lotka-Volterra model was originally introduced by Alfred Lotka in 1910 to study chemical reactions. Later, he realized that the model can also be used to model interactions of predatory and prey populations. Independently, Vito Volterra was asked by his future son-in-law, who was a marine biologist, to explain why the fraction of predatory fish caught by fishermen increased dramatically during World War I. This led to his work in this area.

The model describes the prey population, of the size $x(t)$ and the predator population, of the size $y(t)$. If there is no interaction between the two populations, then each develops according to the following ODEs:

$$\frac{dx(t)}{dt} = ax(t), \quad \frac{dy(t)}{dt} = -by(t). \tag{2.1}$$

Here, $a > 0$ is the balance between the birth rate of the prey and their natural death rate in the absence of predators. The assumption that $a > 0$ means that the prey would survive if no predators are present. The second parameter $b > 0$ measures the balance between the birth rate of the predators and their natural death rate in the absence of prey. The assumption that $b > 0$ means that in the absence of prey the population of predators would die out.

Exercise 2.1. Show that if $x(0) > 0$ and $y(0) > 0$, and $a > 0$, $b > 0$, then $x(t) \to +\infty$ and $y(t) \to 0$ as $t \to +\infty$.

Next, we modify the system to account for the predator-prey encounters. If $N(t)$ is the number of interactions per unit time, then (2.2) should be modified to

$$\frac{dx(t)}{dt} = ax(t) - \tilde{\beta}N(t), \quad \frac{dy(t)}{dt} = -by(t) + \tilde{\gamma}N(t). \tag{2.2}$$

Here, $\tilde{\beta} > 0$ is the parameter that measures the rate of depletion of the population of prey and $\tilde{\gamma} > 0$ is the parameter that measures the rate of increase of the population of predators due to the presence of predator-prey encounters. The next modeling assumption is that

$$N(t) = \alpha x(t)y(t), \tag{2.3}$$

so that the chance of an encounter is proportional both to the population of the prey and of the predators. Using this in (2.2) gives the Lotka-Volterra system, also known as the predator-prey model:

$$\frac{dx(t)}{dt} = ax(t) - \beta x(t)y(t), \quad \frac{dy(t)}{dt} = -by(t) + \gamma x(t)y(t), \tag{2.4}$$

with $\beta = \tilde{\beta}\alpha$, $\gamma = \tilde{\gamma}\alpha$.

The first important observation is that if initially the populations are non-negative, then they stay non-negative: discussing negative populations would make no sense.

Exercise 2.2. Let $(x(t), y(t))$ be a solution to (2.4) for $0 \le t \le T$, with some $T > 0$. Show that if $x(0) \ge 0$ and $y(0) \ge 0$ then $x(t) \ge 0$ and $y(t) \ge 0$ for all $0 \le t \le T$. Hint: show that the set $\{(x, y) \in \mathbb{R}^2 : \ x \ge 0, y \ge 0\}$ is an invariant region for (2.4).

The second observation is that solutions to (2.4) exist for all times $t \ge 0$ even though the nonlinear term is quadratic in x and y.

Exercise 2.3. Let $(x(t), y(t))$ be a solution to (2.4) with $x(0) \ge 0$ and $y(0) \ge 0$. Show that its maximal existence time (for $t \ge 0$) is $T = +\infty$. Hint: use positivity of $y(t)$ to get an upper bound on $x(t)$, then use that bound on $x(t)$ to bound $y(t)$ from above. Note that this is not true without the assumption that $x(0) \ge 0$ and $y(0) \ge 0$. Take $a = \beta = \gamma = 1$, $b = -1$, and $x(0) = 1$, $y(0) = -1$. Show that then $x(t) = -y(t)$ for all $t \ge 0$ in the maximal existence time interval, and that the maximal time interval is finite.

From now on, we assume that $x(0) \geq 0$ and $y(t) \geq 0$, so that solutions exist for all $t \geq 0$. The basic question is the long time behavior of this system: whether in the long run both populations survive, only one does, or both die out. To see this at an informal level, let us re-write (2.4) as

$$\frac{dx(t)}{dt} = x(t)(a - \beta y(t)), \quad \frac{dy(t)}{dt} = y(t)(-b + \gamma x(t)). \tag{2.5}$$

We see that the prey population $x(t)$ is increasing if and only if $y(t) < \bar{y} = a/\beta$ – there are not too many predators around, and the predator population $y(t)$ is increasing if and only if $x(t) > \bar{x} = b/\gamma$ – there sufficiently many prey around. For example, if we start with $x(t) < \bar{x}$ and $y(t) < \bar{y}$ then the population of prey would increase but the population of predators would decrease, until $x(t)$ crosses $\bar{x}$ after which time the population of predators would start increasing.

To make the analysis precise, note that (2.5) has exactly two equilibria:

$$E_1 = (0, 0), \quad E_2 = (\bar{x}, \bar{y}) = \Big(\frac{b}{\gamma}, \frac{a}{\beta}\Big). \tag{2.6}$$

Let us compute the linearizations at these two points: the system (2.5) has the form

$$\frac{d}{dt}\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} F_1(x(t), y(t)) \\ F_2(x(t), y(t)) \end{pmatrix}, \tag{2.7}$$

with

$$F(x, y) = \begin{pmatrix} F_1(x, y) \\ F_2(x, y) \end{pmatrix} = \begin{pmatrix} ax - \beta xy \\ -by + \gamma xy \end{pmatrix}, \tag{2.8}$$

so that the derivative matrix is

$$DF(x, y) = \begin{pmatrix} \dfrac{\partial F_1(x,y)}{\partial x} & \dfrac{\partial F_1(x,y)}{\partial y} \\ \dfrac{\partial F_2(x,y)}{\partial x} & \dfrac{\partial F_2(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} a - \beta y & -\beta x \\ \gamma y & -b + \gamma x \end{pmatrix}. \tag{2.9}$$

Therefore, the derivative matrix at the point $E_1 = (0, 0)$ is

$$DF(E_1) = \begin{pmatrix} a & 0 \\ 0 & -b \end{pmatrix}. \tag{2.10}$$

This matrix has a positive eigenvalue $\lambda_1 = a$ that corresponds to the eigenvector $e_1 = (1, 0)$ and a negative eigenvalue $\lambda_2 = -b$ that corresponds to the eigenvector $e_2 = (0, 1)$. Hence, the equilibrium $(0, 0)$ is unstable – which is good news!

Exercise 2.4. Show that the stable manifold of E_1 is the line $x = 0$, and the unstable manifold is the line $y = 0$. Be careful not to confuse the stable manifold for the nonlinear problem with the eigenspace for the linearized problem. They do coincide in this particular case but they are different objects in general.

The derivative matrix at E_2 is

$$DF(E_2) = \begin{pmatrix} 0 & -\beta b/\gamma \\ \gamma a/\beta & 0 \end{pmatrix}, \tag{2.11}$$

and its eigenvalues are $\mu_{1,2} = \pm i\sqrt{ab}$. As the eigenvalues are purely imaginary, we can not immediately say how the solutions to the nonlinear system behave near E_2. Luckily, we have a conserved quantity (also known as an integral of motion): consider the function

$$L(x, y) = \gamma x - b \log x + \beta y - a \log y, \tag{2.12}$$

so that

$$\nabla L(x, y) = \Big(\frac{\partial L(x,y)}{\partial x}, \frac{\partial L(x,y)}{\partial y}\Big) = \Big(\gamma - \frac{b}{x}, \beta - \frac{a}{y}\Big). \tag{2.13}$$

Then we have

$$\begin{aligned}\nabla L(x,y)\cdot F(x,y) &= \Big(\gamma-\frac{b}{x},\beta-\frac{a}{y}\Big)\cdot(ax-\beta xy,-by+\gamma xy)\\ &= \frac{1}{xy}(\gamma xy-by,\beta xy-ax)\cdot(ax-\beta xy,-by+\gamma xy) \qquad (2.14)\\ &= \frac{1}{xy}\Big[(\gamma xy-by)(ax-\beta xy)+(\beta xy-ax)(-by+\gamma xy)\Big]=0.\end{aligned}$$

It follows that $L(x,y)$ is conserved along the trajectories:

$$L(x(t),y(t))=L(x(0),y(0)), \qquad (2.15)$$

so that the trajectory stays on a level set of $L(x,y)$. Also, we see from (2.13) that the only critical point of the function $L(x,y)$ in the positive quadrant $Q=\{x>0,y>0\}$ is the point E_2. As $L(x,y)\to+\infty$ as (x,y) approaches the boundary of Q or $\|(x,y)\|\to+\infty$, we deduce that E_2 is the unique minimum of the Lyapunov function $L(x,y)$ inside Q. It is a strict local minimum since $L(x,y)$ is strictly convex on any compact set $K\subset Q$. In addition, the above considerations show that the level sets of $L(x,y)$ inside Q are closed curves around E_2. We have proved the following.

Theorem 2.5. *Let $x(t),y(t)$ be a solution to (2.5) such that $x(0)>0$ and $y(0)>0$. If $(x(0),y(0))=(\bar{x},\bar{y})$, then $(x(t),y(t))=(\bar{x},\bar{y})$ for all $t\geq 0$, and if $(x(0),y(0))\neq(\bar{x},\bar{y})$ then $(x(t),y(t))$ is periodic in t.*

Exercise 2.6. Work out what happens if $x(0)=0$ or $y(0)=0$.

We see that if initially both predators and prey are present, then the populations oscillate in time, and neither population dies out, nor do they grow unboundedly in time.

Let us go back to the World War I fishing situation, where the fraction of the predator fish caught by the fishermen increased compared to the pre-war years. Volterra's observation was that one should look at the stable equilibrium E_2

$$E_2=(\bar{x},\bar{y})=\Big(\frac{b}{\gamma},\frac{a}{\beta}\Big). \qquad (2.16)$$

The direct effect of the war was the decrease in the rate of fishing. Let us see how this affects the parameters of the Lotka-Volterra system. First, γ and β are unaffected – they come from predator-prey encounters and are not affected by fishing one way or other. On the other hand, the parameter a is the rate of growth of the prey population if predators are absent, hence a becomes larger when fishing decreases. The parameter b is the rate of depletion of predators if the prey are absent. This rate decreases when fishing decreases, thus b gets smaller. Thus, during the war γ and β were unchanged, a became larger, and b got smaller. Looking at expression (2.16) we see that the fraction of predators at the equilibrium state is larger during the war, precisely in agreement with the observations.

A SIR-type epidemic model

A basic epidemic model typically includes three types of population: susceptible – those who can potentially get the disease, infected – those who are currently infected, and recovered – those who have had the disease but no longer do. Such models are known as SIR models. Let us denote by $S(t)$ the size of the susceptible population, by $I(t)$ the number of infected, and by $R(t)$ the number of recovered. We will assume that once you recover you can not get the disease again, and that anyone infected, no matter how recently, can transmit the disease.

The susceptible population decreases by means of being infected, and the rate of decrease is proportional to the number of infected individuals. This means that $S(t)$ satisfies

$$\frac{dS(t)}{dt}=-\beta S(t)I(t). \qquad (2.17)$$

Here, $\beta>0$ is the parameter measuring how high the transmission rate is. The infected population increases via infecting someone susceptible, and decreases via recover:

$$\frac{dI(t)}{dt}=\beta S(t)I(t)-\gamma I(t). \qquad (2.18)$$

Here, $\gamma > 0$ is the recovery rate from the disease. Finally, the recovered population grows simply by the recovery of the infected population:

$$\frac{dR(t)}{dt} = \gamma I(t). \tag{2.19}$$

Exercise 2.7. Show that if $S(0) \geq 0$, $I(0) \geq 0$ and $R(0) \geq 0$ then $S(t) \geq 0$, $I(t) \geq 0$ and $R(t) \geq 0$ for all $t \geq 0$ in the maximal interval of existence for the system (2.17)-(2.19). Show also that if $S(0) > 0$ and $I(0) > 0$, then $S(t) > 0$ and $I(t) > 0$ for all $t > 0$.

The total population is

$$N(t) = S(t) + I(t) + R(t). \tag{2.20}$$

Note that it does not change in time:

$$\frac{dN(t)}{dt} = \frac{dS(t)}{dt} + \frac{dI(t)}{dt} + \frac{dR(t)}{dt} = -\beta S(t)I(t) + \beta S(t)I(t) - \gamma I(t) + \gamma I(t) = 0, \tag{2.21}$$

so that $N(t) = N(0)$.

Exercise 2.8. *Use (2.21) to show that if $S(0) \geq 0$, $I(0) \geq 0$ and $R(0) \geq 0$ then solutions to (2.17)-(2.19) exist for all $t \geq 0$.*

From now on, we will assume that $S(0) > 0$ and $I(0) > 0$ – there are some susceptible and some infected people at $t = 0$. Note that the susceptible and infected populations form a closed system:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\beta S(t)I(t) \\ \frac{dI(t)}{dt} &= (\beta S(t) - \gamma)I(t), \end{aligned} \tag{2.22}$$

so we do not need to track $R(t)$.

Note that, as both $S(t)$ and $I(t)$ are positive, it follows from the first equation in (2.22) that the function $S(t)$ is strictly decreasing. This makes perfect sense; the number of susceptible people can not increase, as there is no source of them. It follows that the limit

$$S_\infty = \lim_{t \to +\infty} S(t) \tag{2.23}$$

exists and is non-negative: $S_\infty \geq 0$. The interpretation of S_∞ is very intuitive: it is the size of the population that will never get sick. Our goal is to understand if $S_\infty = 0$ or $S_\infty > 0$, and get a handle on how large S_∞ can be.

A very helpful observation is that the system (2.22) has a conserved quantity, similar in spirit to the function (2.12) in the predator-prey model

$$L(S, I) = S - \frac{\gamma}{\beta} \log S + I. \tag{2.24}$$

Indeed, we have

$$\frac{dL(S(t), I(t))}{dt} = \Big(1 - \frac{\gamma}{\beta S(t)}\Big)\frac{dS(t)}{dt} + \frac{dI(t)}{dt} = -\beta\Big(1 - \frac{\gamma}{\beta S(t)}\Big)S(t)I(t) + \beta S(t)I(t) - \gamma I(t) = 0, \tag{2.25}$$

so that

$$L(S(t), I(t)) = L(S(0), I(0)), \tag{2.26}$$

which is

$$S(t) - \frac{\gamma}{\beta} \log S(t) + I(t) = S(0) - \frac{\gamma}{\beta} \log S(0) + I(0). \tag{2.27}$$

As $I(t) \geq 0$ and $S(t) \geq 0$ for all $t \geq 0$, we deduce that

$$-\frac{\gamma}{\beta} \log S(t) \leq S(0) - \frac{\gamma}{\beta} \log S(0) + I(0), \tag{2.28}$$

hence

$$S(t) \geq \exp\Big[\log S(0) - \frac{\beta}{\gamma}\Big(S(0)+I(0)\Big)\Big] = S(0)\exp\Big[-\frac{\beta}{\gamma}\Big(S(0)+I(0)\Big)\Big]. \tag{2.29}$$

This gives a lower bound on the size of the susceptible population remaining after time t, and also a lower bound on S_∞:

$$S_\infty \geq S(0)\exp\Big[-\frac{\beta}{\gamma}\Big(S(0)+I(0)\Big)\Big]. \tag{2.30}$$

This is a lower bound on how many people will not be infected as $t \to +\infty$, after the epidemic "has done its run". This bound is not optimal but it already says that not everyone will be eventually infected: $S_\infty > 0$.

Let us now get a bit more information on the behavior of the solutions to (2.22), and a better handle on S_∞. Note that the size $I(t)$ of the infected population increases as long as $S(t) > \gamma/\beta$ – this follows from the second equation in (2.22). The next proposition shows that this can not continue forever.

Proposition 2.9. *There exists $t_0 \geq 0$ so that $S(t_0) \leq \gamma/\beta$.*

Proof. If $S(0) \leq \gamma/\beta$ then we are done as we can take $t_0 = 0$. Assume that $S(0) > \gamma/\beta$ and that $S(t) \geq \gamma/\beta$ for $0 \leq t \leq \tau$ for some $\tau > 0$. We will now provide an upper bound for such time τ. It follows from the above assumption and the second equation in (2.22) that

$$\frac{dI(t)}{dt} = (\beta S(t) - \gamma)I(t) \geq 0 \text{ for all } 0 \leq t \leq \tau, \tag{2.31}$$

thus, in particular, $I(t) \geq I(0)$ for all $0 \leq t \leq \tau$. Using this in the first equation in (2.22) gives

$$\frac{dS(t)}{dt} = -\beta S(t)I(t) \leq -\beta S(t)I(0), \quad \text{for all } 0 \leq t \leq \tau, \tag{2.32}$$

so that

$$\frac{dS(t)}{dt} + \beta S(t)I(0) \leq 0, \quad \text{for all } 0 \leq t \leq \tau. \tag{2.33}$$

Multiplying both sides by $\exp(I(0)t)$ gives

$$0 \geq \Big(\frac{dS(t)}{dt} + \beta S(t)I(0)\Big)e^{I(0)t} = \frac{d}{dt}\Big(e^{\beta I(0)t}S(t)\Big) \quad \text{for all } 0 \leq t \leq \tau. \tag{2.34}$$

Integrating in time from $t = 0$ to τ gives

$$0 \geq e^{\beta I(0)\tau}S(\tau) - S(0), \tag{2.35}$$

so that

$$S(\tau) \leq S(0)e^{-\beta I(0)\tau}. \tag{2.36}$$

As, by assumption, we have $S(\tau) \geq \gamma/\beta$, we deduce that

$$\frac{\gamma}{\beta} \leq S(0)e^{-\beta I(0)\tau}, \tag{2.37}$$

and thus

$$\tau \leq \tau_0 = \frac{1}{\beta I(0)}\log\Big(\frac{S(0)\beta}{\gamma}\Big). \tag{2.38}$$

It follows that there exists $t_0 \leq \tau_0$ so that $S(t_0) \leq \gamma/\beta$. As the function $S(t)$ is decreasing, we conclude that

$$S(t) \leq \frac{\gamma}{\beta} \text{ for all } t \geq \tau_0 = \frac{1}{\beta I(0)}\log\Big(\frac{S(0)\beta}{\gamma}\Big), \tag{2.39}$$

finishing the proof. □

The proof of Proposition 2.9 shows that $I(t)$ is decreasing for all $t \geq \tau_0$, and also that $S(t)$ drops below the value $S_c = \gamma/\beta$ at a finite time $t_0 \leq \tau_0$, and stays below this value for all $t \geq t_0$. An immediate consequence is an upper bound on S_∞:

$$S_\infty \leq S_c = \frac{\gamma}{\beta}. \tag{2.40}$$

Very remarkably, the number S_c, known as the epidemiological threshold, does not depend on the initial size $S(0)$ of the susceptible population – it only depends on the parameters γ and β. The particular form of this dependence $S_c = \gamma/\beta$ is quite intuitive: it is inversely proportional to the infection rate β, and is directly proportional to the inverse γ of the recovery time. In other words, the quicker people recover, the larger the epidemiological threshold, and the larger the infection rate, the smaller the epidemiological threshold. Thus, the epidemiological threshold plays two roles: first, the size of the infected population starts decreasing when the size of the susceptible population drops below S_c, and, second, it gives an upper bound on the size S_∞ of the susceptible population that will never get sick.

Now that we know that the size of the susceptible population has to drop below S_c, let us see if we can improve on the lower bound on S_∞ given in (2.30). For this, we need to show that that eventually the epidemic will die out.

Proposition 2.10. *We have*

$$\lim_{t \to +\infty} I(t) = 0. \tag{2.41}$$

Proof. Recall that Exercise 2.7 shows that $I(t) > 0$ and $S(t) > 0$ for all $t \geq 0$, so that the first equation in (2.22) shows that the function $S(t)$ is strictly decreasing. Hence, Proposition 2.9 shows that there exists $t_1 > 0$ so that $S(t) \leq S(t_1) < \gamma/\beta$ for all $t \geq t_1$ – note that here the last inequality is strict. Then, the second equation in (2.22) shows that

$$\frac{dI(t)}{dt} = (\beta S(t) - \gamma) I(t) \leq (\beta S(t_1) - \gamma) I(t), \quad \text{for } t \geq t_1. \tag{2.42}$$

Let us set $k = \gamma - \beta S(t_1) > 0$ and write (2.42) as

$$\frac{dI(t)}{dt} \leq -kI(t), \quad \text{for } t \geq t_1. \tag{2.43}$$

Multiplying (2.43) by $\exp(kt)$, we obtain

$$\frac{d}{dt}\Big(e^{kt} I(t)\Big) = \Big(\frac{dI(t)}{dt} + kI(t)\Big) e^{kt} \leq 0, \quad \text{for } t \geq t_1. \tag{2.44}$$

Integrating in t from t_1 to $t > t_1$ gives

$$e^{kt} I(t) - e^{kt_1} I(t_1) \leq 0, \quad \text{for } t \geq t_1, \tag{2.45}$$

so that

$$I(t) \leq I(t_1) e^{-k(t-t_1)}, \quad \text{for } t \geq t_1, \tag{2.46}$$

This implies that

$$\lim_{t \to +\infty} I(t) = 0, \tag{2.47}$$

and we are done. □

Finally, we can characterize the value of S_∞. Note that the function $g(x) = x - (\gamma/\beta) \log x$ has a minimum at $x = S_c = \gamma/\beta$, is decreasing for $0 < x < S_c$ and increasing for $x > S_c$, and also

$$\lim_{x \to 0^+} g(x) = \lim_{x \to +\infty} g(x) = +\infty. \tag{2.48}$$

Hence, an equation of the form

$$g(x) = y$$

with $y > \min_{x>0}(g(x))$ has exactly two solutions: $x_1(y) \in (0, S_c)$ and $x_2(y) > S_c$.

Proposition 2.11. *The value of S_∞ is the smaller root of the equation*

$$g(x) = S(0) - \frac{\gamma}{\beta}\log S(0) + I(0). \tag{2.49}$$

Proof. We go back to (2.27):

$$S(t) - \frac{\gamma}{\beta}\log S(t) + I(t) = S(0) - \frac{\gamma}{\beta}\log S(0) + I(0). \tag{2.50}$$

Passing to the limit $t \to +\infty$ with the help of Proposition 2.10 gives

$$S_\infty - \frac{\gamma}{\beta}\log S_\infty = S(0) - \frac{\gamma}{\beta}\log S(0) + I(0). \tag{2.51}$$

As we have already shown that $S_c < \gamma/\beta$, the conclusion of Proposition 2.11 follows. □

Interestingly, the graph of the function $g(x)$ shows that the value of S_∞ is larger if the right side of (2.49) is smaller. This is because S_∞ is the smaller root of (2.49), and

$$\lim_{x\to 0^+} g(x) = +\infty. \tag{2.52}$$

In particular, for a fixed initial size of the susceptible population $S(0)$, the size S_∞ decreases as $I(0)$ increases, which is quite natural.

The SIR model we have discussed in this section is very basic and does not take into account many important features. The most obvious are the spatial effects – we did not consider any spatial variations in the population size, and instead consider the population as a whole. The spatial dependence and connectivity between various geographic regions is, clearly, an important aspect. It can be modeled with partial differential equations, with SIR-type models for functions $S(t,x)$, $I(t,x)$ and $R(t,x)$ that now depend both on time and space. Such models lie outside the scope of this class. Another actively studied extension is taking into account the randomness and fluctuations in the interactions between the infected and susceptible populations. These are still a subject of active research with many open questions mathematically. Nevertheless, remarkably, even the very simple ODE model we have considered here gives qualitatively non-trivial predictions.