1 Modeling

Humans can't be trusted to develop realistic models. They can't be trusted with developing toy models either, since they soon forget that these are tinker toys, and this way endanger all of us.
—Doron Zeilberger

Applied math has three major components: First-principle-based modeling, data-driven methods, and algorithms.
—Weinan E

1.1 The Art and Science of Mathematical Modeling

Budding mathematicians in their infancy often suppose that mathematics is the science of analyzing the equations that are *known* to model the universe. As these mathematicians mature in their mathematical knowledge and abilities they come to realize that the equations that govern all motion and activity in the universe are actually not known (this was a very disappointing moment in one of the authors' education), and if they were, there would be no hope of finding 'solutions' to them anyway. Reality, it turns out, is a bit messier than a nice set of equations, and much messier than the homework you have been assigned previously or will be assigned in this course. What are we mathematicians to do then, if we do not always have an equation to start with? Model, my young apprentice, we must model.

How do we model? There are two fundamental approaches to mathematical modeling, one of which we will address in this text, and the other is addressed in Volume 3. The basic choice comes down to deciding to focus the modeling on a data driven approach (an option that has only been relevant recently), versus developing a model based on physical laws (such as gravity) that can be isolated and observed in the natural world. Although the curriculum provided here appears to imply that these two approaches are separate and distinct from each other, the best solutions are typically achieved by merging both concepts together; for example, deriving a model from first principles and then 'tuning' parts of that model to fit existing data in order to obtain the most accurate predictions. Even at the most basic level, the laws of physics didn't just appear in Newton's mind, but came as a result of observations he made (data) of the natural world.

In an ideal universe the models would be perfect and known, and we would have little need for the current discussion. But we do not have the information necessary to completely model everything, so we must iteratively refine our models of whatever physical phenomenon we are interested in. Before we start happily iterating, though, we need to decide what we are iterating on, that is, how do we model something to begin with?

In this book we are concerned primarily with models that deal with evolution in time of some given quantity or entity. We focus on dynamic modeling, meaning models that are meant to capture dynamic changes or evolution of different objects. It is important to remember, as we go through the mathematics underlying these models, that they are just models. While we can definitively prove various things about a dynamic model, certain hypotheses must hold for the conclusions to apply. This also provides us with a way of testing the validity of our model. If the mathematics behind a dynamic model for a physical system yields un-physical conclusions, then the model needs to be revised or thrown out entirely.

This mentality is at the heart of all mathematical modeling, and hopefully is illustrated in the examples that follow in this chapter. In other fields the idea is often to try to include every detail in the model, resulting in a complex system for which there is little to no hope of analyzing the mathematics underlying the system, and hence little hope of rigorously saying anything about the model's predictive power.

In mathematics the approach is most often the opposite: we sacrifice detail for the sake of simplicity and practicality. What is the use of including every detail if we don't know whether such details affect the outcomes we are interested in? Instead, mathematical modeling starts with the simplest model possible, potentially omitting relevant details so that the model can be completely understood and the different components can be analyzed completely. Once this simple model is understood, we usually see that the model is too simplistic and omits some pertinent detail of the problem in question. We must then return to the model and try to add in the pertinent details, but only one or two at a time, so as to understand the affect of these details on the model's outcomes. Then we analyze the more advanced model to determine its shortcomings and repeat the process.

1.1.1 An example: the law of gravity

It took years of empirical analysis (an earlier version of what we might call data analysis today) to obtain the inverse square law for gravity. Modeling gravity without the notion of a derivative was quite the feat. While most people credit Newton with the famous inverse square law for gravity, there is written evidence that Hooke[1], a contemporary of Newton, was the first to propose the inverse square law. Newton essentially invented calculus to mathematically derive Kepler's three laws for planetary motion from the inverse square law of gravity. Kepler's three laws were originally empirical laws based upon massive amounts of very accurate naked-eye observational data collected by Tycho Brahe (or more likely his underlings), some of which Kepler confessed to obtaining from Brahe by less-than-ethical means. Establishing Kepler's laws of planetary motion as a consequence of mathematical derivation, rather than as matching observational data, marked a significant paradigm shift in modeling.

Remark 1.1.1. As increasingly large data sets become available, the trend in modeling is moving away from the Newtonian approach of deriving models from first principles. Instead many modelers strive to fit the data to arbitrarily defined functions, chosen without regard to the source of the data (this is machine learning). The most powerful modelers use both—incorporating what they know about first principles into machine learning models fit to data.

Apart from the historical drama of the origins of the inverse square law and Kepler's three laws, the process of modeling is often a matter of guess and check. Because of the prevailing theory of the Earth-centric solar system at the time, Kepler at first disregarded the notion of an ellipse for the shape of the orbit of the planet Mars, the only batch of data that Brahe gave Kepler access to at the time. Kepler tried all kinds of curves to no avail. Only later, as the Sun-centric version of the Solar System gained more popularity, did Kepler return to the ellipse and got it to match the data.

Dynamical modeling is modeling with differential equations. We want to track quantities that depend on time, and often it is most natural to describe these in terms of how they change in time (that is, their time derivative) or how their derivative changes in time (the second time derivative) and so forth. This results in a system of differential equations.

The law of gravity is an example of this. For N bodies in space at positions $\mathbf{q}_i \in \mathbb{R}^3$ with masses m_i, $i = 1, 2, \ldots, N$, the positions $\mathbf{q}_i$ change over time (they are functions of time), and the masses are constant. The velocity of the ith body is $\dot{\mathbf{q}}_i = \frac{d}{dt}\mathbf{q}_i$, and the acceleration is $\ddot{\mathbf{q}}_i = \frac{d^2}{dt^2}\mathbf{q}_i$. The three principles used in creating this model are the following:

(i) Newton's Second Law: The force F on body i is equal to mass m_i times acceleration $\ddot{\mathbf{q}}_i$; that is $F = m_i a_i = m_i \ddot{\mathbf{q}}_i$.

(ii) The total force on a body is the sum of the forces exerted on it by all the other bodies.

[1] This is the person after whom Hooke's law for springs is named.

(iii) The force exerted on body i by body j in is directly proportional to the masses m_i and mass m_j and inversely proportional to the square $\|\mathbf{q}_j - \mathbf{q}_i\|^2$ of the distance between them. This force is exerted in the direction of the unit vector $\frac{\mathbf{q}_j - \mathbf{q}_i}{\|\mathbf{q}_j - \mathbf{q}_i\|}$.

Putting these together gives

$$m_i \ddot{\mathbf{q}}_i = \sum_{j \neq i} \left(\frac{G m_i m_j}{\|\mathbf{q}_j - \mathbf{q}_i\|^2} \right) \frac{\mathbf{q}_j - \mathbf{q}_i}{\|\mathbf{q}_j - \mathbf{q}_i\|} = \sum_{j \neq i} \frac{G m_i m_j (\mathbf{q}_j - \mathbf{q}_i)}{\|\mathbf{q}_j - \mathbf{q}_i\|^3}, \tag{1.1}$$

where G is a constant (called the *universal gravitational constant*). It has been shown experimentally that $G \approx 6.6732 \times 10^{-11} \, \frac{\mathrm{m}^3}{\mathrm{s}^2 \mathrm{kg}}$. Alternatively, (1.1) can be written as

$$m_i \ddot{\mathbf{q}}_i = -\frac{\partial U}{\partial \mathbf{q}_i}, \tag{1.2}$$

where

$$U = \sum_{1 \leq j < k \leq N} \frac{G m_j m_k}{\|\mathbf{q}_j - \mathbf{q}_k\|}$$

and $\frac{\partial U}{\partial \mathbf{q}_i} = D_{\mathbf{q}_i}^{\mathsf{T}} U$ denotes the gradient of U with respect to the three coordinates (q_{i1}, q_{i2}, q_{i3}) of $\mathbf{q}_i$. This system of nonlinear differential equations has been researched since Newton first formulated it. We'd like is to be able solve this system, by which we mean, given initial values $\mathbf{q}_i(0)$ and velocities $\dot{\mathbf{q}}_i(0)$ write down equations for each $\mathbf{q}_i(t)$ as a function of t. Sadly, this can only be done for the case of $N = 2$, and that only implicitly. Fortunately, we don't necessarily need a complete solution in order to understand some of the most important aspects of a solution.

1.1.2 Example: Simple Harmonic Motion

Imagine a frictionless block of mass m (call the block *Bob*) attached to one end of a spring with the other end fixed to the wall. Assume that Bob and the spring move in only one direction, that is, we are only concerned with Bob's horizontal position $y(t)$, and need not worry about movement in other directions. Newton's second law guarantees that the mass times Bob's acceleration must be equal to the forces acting on Bob. In other words

$$m \, \ddot{y}(t) = \text{sum of forces}, \tag{1.3}$$

where $\ddot{y}(t) = \frac{d^2 y}{dt^2}$. We often omit the t and just write $m\ddot{y} = F$ where F is the sum of all the forces.

In our setup there is no friction and the only force acting on Bob is due to the spring. If $y = 0$ is the spring's resting position, then Hooke's Law for springs [2] states that the force F_s due to the spring is proportional to Bob's distance from the resting position:

$$F_s = -ky$$

[2] This is the same Hooke who accused Newton of stealing the discovery of the inverse-square law of gravity.

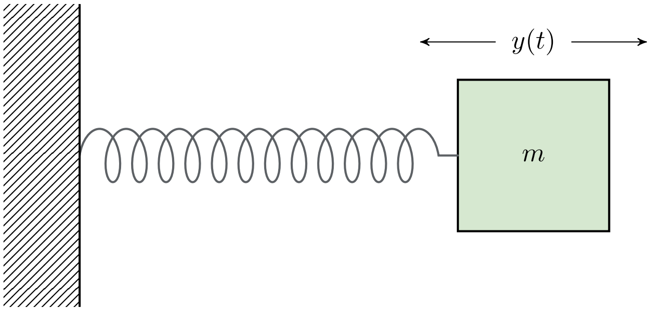

Figure 1.1: An illustration of a simple harmonic oscillator. The box represents a mass named Bob, who is at the sole will of the spring, whether compressing or expanding. The rest position is at $y = 0$.

for some positive k. Equating the two parts gives a second-order linear equation for Bob's position

$$m\ddot{y} = -ky \qquad \text{or} \qquad m\ddot{y} + ky = 0.$$

Dividing by m and introducing the parameter

$$\omega^2 = \frac{k}{m},$$

we can rewrite this equation as

$$\ddot{y} + \omega^2 y = 0, \tag{1.4}$$

It is straightforward to verify that any y of the form

$$y = a\cos(\omega t) + b\sin(\omega t), \tag{1.5}$$

for arbitrary $a, b \in \mathbb{R}$ is a solution of (1.4). In Chapter 4 we show that all solutions of (1.4) have this form.

If we know Bob's initial position $y(0) = y_0$, then that puts constraints on the possible values of a by

$$y_0 = y(0) = a\cos(0) + b\sin(0) = a$$

Similarly, if we know Bob's initial velocity $\dot{y}(0) = v_0$, then that constrains b by

$$v_0 = \dot{y}(0) = -a\omega\sin(0) + b\omega\cos(0) = b\omega.$$

Together, these initial conditions and (1.4) uniquely determine Bob's position at any subsequent moment in time.

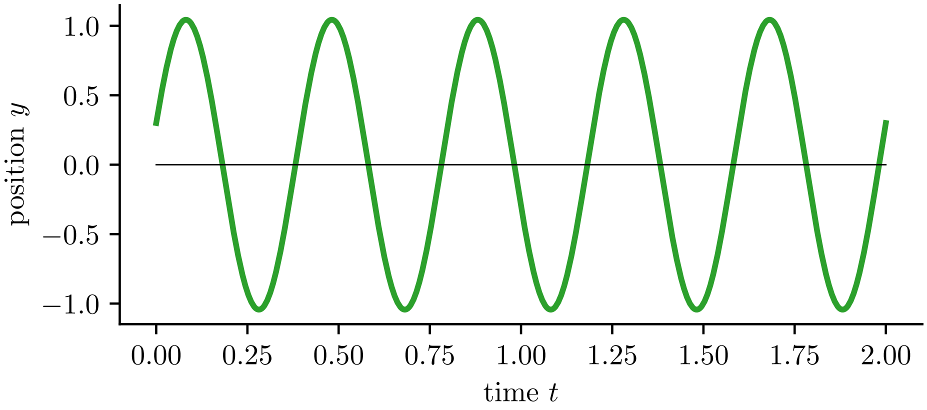

Figure 1.2: The motion of a simple spring over time. Note that the initial position in this case is at $y(0) = 0.3$. A different value of the initial state would significantly alter the overall motion of the spring.

Remark 1.1.2. The fact that Bob's motion depends on both his initial position and initial velocity should not be surprising. For anyone who has played with springs, experience suggests that if Bob starts with no initial velocity, but starts away from resting position at y_0 (say we have pulled Bob to point y_0 and then release him), then he will initially be drawn toward 0 by the spring, will overshoot 0, then slow down and eventually start moving back toward 0, and repeat the cycle, oscillating back and forth. The exact form of that oscillation will depend on how far away from 0 Bob started. This is completely compatible with the solution (1.5) with $a = y_0$ and $b = 0$, giving $y(t) = y_0 \cos(\omega t)$. Similarly, if Bob starts in the resting position with nonzero velocity (perhaps Bob is struck by another object), then Bob will travel some distance in that same direction before being pulled back by the spring, oscillating again, but differently from the first solution. And the exact form of this oscillation will depend on Bob's initial velocity. This is compatible with the solution (1.5) with $a = 0$ and $b = \frac{v_0}{\omega}$, giving $y(t) = \frac{v_0}{\omega} \sin(\omega t)$. Finally, if Bob starts in the resting state and has initial velocity 0, then $a = b = 0$ and $y(t) = 0$ for all t, agreeing with the intuition that Bob should never move in this case.

This example shows that a differential equation alone is usually inadequate to fully describe the physical phenomenon we are trying to model, and we must include some type of auxiliary conditions, like an initial position and velocity, in order to completely specify a solution. The number of auxiliary conditions and the type will be problem dependent; we discuss this in much more detail in the next chapter.

We cannot always completely describe a unique solution to a given differential equation, even knowing all the appropriate auxiliary conditions. But even without a complete description, we can often understand the qualitative properties of the solutions, which may be more important. These include questions like "Does the solution oscillate back and forth forever, like Bob on the spring?" or "Does it eventually approach a fixed position?" These questions lead us to the study of dynamical systems and questions of stability and long-term behavior. In many cases it doesn't matter how the system begins, if we wait long enough we know where it will end up.

1.1.3 Damped Harmonic Motion

The solution (1.5) oscillates forever, but this is not very realistic—our physical experience with springs suggests that the amplitude of the oscillation gradually reduces and eventually the motion comes to a stop, that is, the motion is *damped.*[3] The problem with the simple model is that it neglects the force of friction.

We expect friction to be resistant to any motion; that is, friction will be an additional force that acts in the direction opposite to Bob's velocity $\dot{y}(t)$, and it seems reasonable to assume (and experiments approximately confirm) that the magnitude of the force of friction is proportional to the magnitude of the velocity. That is, we assume $F_f = -c\dot{y}(t)$, for some $c > 0$. The precise value of c will depend on Bob's mass and how slippery the surface is. Thus the total force should be the sum of the force of friction and the spring force $F = F_s + F_f$, which gives a new, improved model

$$m\ddot{y}(t) = -c\dot{y}(t) - ky(t). \tag{1.6}$$

Using the same substitution $\omega^2 = \frac{k}{m}$, and setting $\gamma = \frac{c}{2m}$, we can write (1.6) as

$$\ddot{y}(t) + 2\gamma\dot{y}(t) + \omega^2 y(t) = 0, \tag{1.7}$$

The solutions of (1.7) break into three cases, based on whether $\gamma^2 < \omega^2$, $\gamma^2 = \omega^2$, or $\gamma^2 > \omega^2$. These are called the *underdamped*, *critically damped*, and *overdamped* cases, respectively.

The Underdamped Case

A straightforward but tedious computation shows that for any $d \geq 0$ and $\phi \in \mathbb{R}$ the function

$$y = de^{-\gamma t}\cos\left(\left(\sqrt{\omega^2 - \gamma^2}\right)t - \phi\right). \tag{1.8}$$

is a solution of (1.7) in the underdamped case ($\gamma^2 < \omega^2$). We show later that all solutions of (1.7) are of this form if $\gamma^2 < \omega^2$.

The initial conditions $y(0) = y_0$ and $\dot{y}(0) = v_0$ uniquely determine the values d and ϕ. But every solution has limit 0 as $t \to \infty$, as shown in Exercise 1.3.

[3] The verb *damp* should not be confused with the verb *dampen*. To *dampen* means to make slightly wet. To *damp* means to restrict the amplitude or intensity. A damper on the piano can be useful to make the sound quieter by reducing the amplitude of the string vibrations, but a dampener on a piano would probably make the pianist upset. It is interesting to note that a small fire can be damped by dampening.

Remark 1.1.3. We had supposed that it was unrealistic for the spring to oscillate for eternity with no change in its amplitude. The solution (1.8) doesn't completely fix that problem, because technically the amplitude of its oscillations is always nonzero, but it is exponentially decaying, so for all practical purposes it is effectively zero after some time In this sense we are satisfied that the model (1.6) for including friction is valid: it still allows for oscillations (subject to certain conditions on the damping coefficient γ relative to the oscillation frequency ω), but these oscillations rapidly decay until they are practically zero.

Overdamped Case

In the overdamped case ($\gamma^2 > \omega^2$), a straightforward check shows that for any $a, b \in R$ the function

$$y = a\exp(-(\gamma - \sqrt{\gamma^2 - \omega^2})t) + b\exp(-(\gamma + \sqrt{\gamma^2 - \omega^2})t)$$

is a solution of (1.7). Again, additional conditions are needed determine the choice of a and b.

In this case the damping force plays the dominant role and Bob moves quickly toward the resting position 0 with no oscillation.

Critically Damped Case

In the critically damped case ($\gamma^2 = \omega^2$), a straightforward check shows that for any $a, b \in R$ the function

$$y = (a + bt)\exp(-\gamma t)$$

is a solution of (1.7). In this case the damping force and the spring force are perfectly balanced to give at most one oscillation before converging toward the resting position 0.

1.1.4 Forced and Damped Harmonic Motion

Now suppose that are we interested in Bob on a spring with friction and an additional force $F(t)$ acting on Bob. For example, we might be pulling or pushing on Bob at different times and would like to know how that affects things. The resulting differential equation is developed in Exercise 1.5

1.1.5 Oscillators Abound

Simple harmonic motion shows up everywhere. We show later in the book that the motion of a pendulum can be reduced under certain assumptions to a simple harmonic oscillator. It is also common to use oscillators as models in bridges and buildings. In fact oscillators can be used to model far more complicated motion, and generalizations of the simple model considered here arise in quantum mechanics and thermodynamics. In other contexts oscillators can also arise in models for the oscillatory nature of biological and physical processes, and in financial markets.

1.2 Modeling Population Dynamics

In this section we discuss some models that deal with populations, such as bacteria, people, or rabbits. Of course, we can make up models for anything we like, but to get a good model that makes sense and agrees with what happens in reality we need to put in a lot of work, and get help from domain experts.

Although biological populations change discretely (they change by ± 1 at instances in time), they can be modeled by continuous differential equations. Rather than counting actual numbers of biological entities, we track percentages or some other ratio. This makes the discrete changes in time seem more continuous and lends itself to better numerical simulations. In this section we explore a few of the many types of population models employed by researchers.

1.2.1 Compartmental Models

Consider a disease spreading rapidly through a homogeneous population. We make the following assumptions, some more realistic than others:

(i) This is a mild disease such as the common cold with little to no fatality rate.

(ii) The population can be divided into three groups:

  (a) Susceptible $S(t)$

  (b) Infected $I(t)$

  (c) Recovered $R(t)$.

(iii) The population is constant: there are no births, deaths, or migration.

(iv) The population values are normalized so that

$$S(t) + I(t) + R(t) = 1 \text{ for all } t \geq 0,$$

In other words, we are looking at the percentage or proportion of the population in each of these categories.

(v) The progression of the disease has no spatial dependence; that is, everyone is in contact with everyone else in the population, regardless of whether they live near or interact with each other often. This may seem to be a rather bold assumption, but for a sufficiently large population in a small enough land mass (think Manhattan or Singapore), it is not unreasonable.

Now to model the progress of the disease:

(i) Recovered individuals cannot get sick again. They remain recovered, have full immunity, and can never be susceptible again. Thus $R(t)$ should be increasing in time and $S(t)$ decreasing.

(ii) The rate at which $S(t)$ decreases should be proportional to the current number of susceptible individuals and how they interact with the infected $I(t)$.

(iii) The rate of infected individuals recovering at any moment should be proportional to the total number of infected in the population.

Putting these all together yields the *SIR model* for disease spread:

$$\dot{S}(t) = -bS(t)I(t) \tag{1.9a}$$

$$\dot{I}(t) = bS(t)I(t) - kI(t) \tag{1.9b}$$

$$\dot{R}(t) = kI(t). \tag{1.9c}$$

We do not yet have the adequate machinery in place to analyze this system extensively, but we can check this model in one way immediately. We claimed that this model was built on the premise that the total population would remain fixed. This should mean that the time derivative of the total population is zero, which is verified in Exercise 1.6. Once we understand how to approach these problems better, you will be able to prove that if all of $S(t)$, $I(t)$, and $R(t)$ are positive initially, then they will remain positive for all time.

Remark 1.2.1. Returning to our previous discussion, can we say anything about the long-term behavior of this model? We know that the total population is constant, but do we know anything else?

Remark 1.2.2. The assumptions for this model are not entirely realistic, and one can quickly see how things might go wrong. For instance, one assumption is that there are no deaths and no births in this model population. In certain populations, and with certain diseases this may be a fair assumption in the short term, but it is unrealistic over a long time period. We should consider how to add such effects into the model, and how such effects would influence the evolution of each sub-population. One can also see how to generalize this to far more complicated situations where perhaps there are different stages to the disease (see the exercises). In addition, one might consider a disease from which some individuals may temporarily recover, but their immunity isn't permanent, and they can become susceptible again after some period of time, or perhaps a subset of the recovered individuals become susceptible again, but others retain full immunity.

1.2.2 Exponential Growth Models

The prototypical example of population growth, ignoring diseases and natural predators, is usually rabbits. Until the 18th century, there were no rabbits on the Australian continent. The furry long-eared rodents were introduced sometime in the 18th century, likely as adorable house pets who then escaped into the great outback. With almost no natural predators and a vegetation that they clearly enjoyed, the rabbit population in Australia increased exponentially (see (1.10) below) to the point that by the 20th century they were considered a national pest, and devious methods of curbing the population were devised.[4]

This drastic example of population growth would seem nonsensical if it were not real, and illustrates what can happen in ideal conditions for any population. Most populations have natural predators, or a finite carrying capacity that the natural environment can sustain. Australia has never reached its carrying capacity for rabbits, as human intervention has occurred first.

[4] Google rabbits in Australia...it is an interesting read.

We can model the growth of the Australian rabbit population by noting that the rate at which the population is increasing, is proportional to the current size of the population (the number of births and deaths at each moment in time is proportional to the population), with a proportionality constant inversely related to the rabbit's rate of maturation (a very short time period by the standards of almost any other mammalian species). Letting $x(t)$ be the rabbit population, gives the simple differential equation:

$$\dot{x}(t) = rx(t), \quad x(0) \geq 0, \tag{1.10}$$

where r is the growth rate. The solution to (1.10) is $x(t) = e^{rt}x(0)$, where $x(0)$ is the initial population.

Remark 1.2.3. Equation (1.10) applies to population growth only if $r > 0$. If $r < 0$, then the population is exponentially decaying, that is the death rate is greater than the birth rate. Examples of decaying dynamics include, but are not limited to, radioactive decay, the rate of a chemical reaction, amplitude of overdamped oscillators, and the depreciation of an automobile's value over time.

1.2.3 Logistic Growth Models

In most cases (1.10) is not a good model for populations because populations typically do not have unlimited resources. Populations that are relatively small will tend to grow almost exponentially, but as they get large, relative to their resources, they grow more slowly. If there is only enough food for 10,000 rabbits, the population of rabbits is unlikely to grow beyond 10,000. To model this requires that we adjust the exponential growth model (1.10) so that growth $\dot{x}$ starts out close to rx when x is small but shrinks toward 0 as x approaches the *carrying capacity* (maximum number of rabbits the environment can support). One way to do this is the *logistic growth equation*:

$$\dot{x} = rx\left(1 - \frac{x}{k}\right), \tag{1.11}$$

where k is the carrying capacity of the environment. This differs from the exponential model (1.10) only by the factor $\left(1 - \frac{x(t)}{k}\right)$ on the right. This factor is close to 1 when x is small and close to 0 when x is near k. Thus when x is small, then (1.11) becomes $\dot{x} \approx rx$, which means that the population is growing approximately exponentially. However, as x gets near k, the derivative $\dot{x}$ approaches zero and the population growth slows.

A straightforward check shows that functions of the form

$$x(t) = \frac{k}{1 + Ce^{-rt}}, \tag{1.12}$$

where $C = \frac{k - x(0)}{x(0)}$ is constant, are solutions of (1.11). All solutions to (1.11) are of this form, as can be shown by rewriting (1.11) as

$$\frac{k}{x(t)(k - x(t))}\frac{dx}{dt} = r,$$

and integrating both sides with respect to t (a method called *separation of variables*).

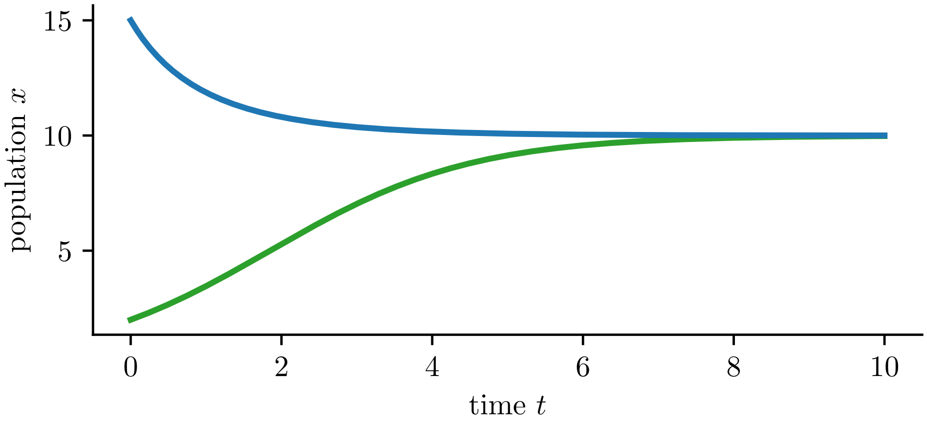

Figure 1.3: The population of a given species in an environment with carrying capacity 10 with two different initial conditions. The blue line starts at $x(0) = 15$ and the green line starts at $x(0) = 2$. Note that both solutions rapidly tend to the carrying capacity.

Of course the actual trajectory of the population depends on the the constant C, whose value depends on the initial population $x(0)$, but as $t \to \infty$ the exponential term in (1.12) goes to 0 and $x(t) \to k$, regardless of the initial value x_0 (see Figure 1.3). This agrees with our expectation that the population of rabbits won't grow beyond what the environment can support, and if it happens to start at a level greater than the population can support, it will shrink down toward the carrying capacity.

Remark 1.2.4. There are many natural variations on the logistic model, the first of which we consider below, where there are two species in the environment, one a predator of the other. Other considerations that we may want to consider include the effect of seasonal changes in the environment. For instance, rabbits' reproductive cycle is much faster than a calendar year and the carrying capacity of the environment for rabbits is lower in winter than in summer. Or perhaps the carrying capacity of the environment is tied to multi-year weather patterns such as El Niño, for instance, some semi-arid regions in Nevada and New Mexico are known to blossom on roughly seven year cycles. Each of these variants needs a new model.

1.2.4 Modeling predator-prey dynamics

One of the most fundamental models in modern quantitative population dynamics and epidemiology is the *Lotka–Volterra predator-prey model.* To motivate the model, consider the population dynamics of two species, one a predator and the other the prey or primary food source for the predator. In such a system, it has been observed that the predator and prey populations are cyclical, with the prey population increasing when the predator population is low, the predator population decreasing when the prey begin to die off, and the opposite effect when either population is increased. Such a fundamental system was first investigated in the context of a dynamical system in the 1920s by Alfred J. Lotka and Vito Volterra separately.[5]

As an initial version of the model, consider the evolution of a species we will refer to as the *prey* whose population $r(t)$ grows exponentially in the absence of a predator

$$\dot{r} = \rho r, \tag{1.13}$$

where $\rho > 0$. Of course this is not realistic because the prey probably don't live in an environment of infinite resources. We revisit this assumption later, but for now assume that constraints on the environment do not influence the dynamics of $r(t)$. We also consider the evolution of the population $w(t)$ of the predator species. Assume that the predator species can only eat this specific prey species, so when the prey is absent the predator population dies off exponentially as

$$\dot{w}(t) = -\mu w(t), \tag{1.14}$$

where $\mu > 0$.

Because the predator eats the prey, the prey population is affected in part by the number of predators, and the predator population is affected by the amount of food (prey) available, so a model needs to account for both of these effects. A simple, reasonable assumption is that the population loss of the prey from predation should be proportional to the number of predators times the number of prey, and similarly, the population growth of the predators should be proportional to the number of predators times the number of prey

This suggests the following model:

$$\dot{r}(t) = \rho r(t) - ar(t)w(t), \tag{1.15a}$$
$$\dot{w}(t) = -\mu w(t) + \varepsilon a r(t)w(t), \tag{1.15b}$$

[5] You may appropriately question why we devote so much space to such a simple two-dimensional model. A potential reason is that this must have been how Charlie Weasley ensured that the dragon population in Romania didn't get out of control (muggles in Romania owe Dumbledore for insisting that Charlie take the course 'Nonlinear dynamics in the wizarding world'). A more realistic answer is likely that one of the authors spent an inordinate amount of time on this and similar models as a student, and is now thrilled to pass on such knowledge and wisdom to a captive audience.

where $a > 0$ represents a constant that relates how many prey each predator consumes in a given interaction (this depends on the metabolism of the predator, and the size of the prey), and $\varepsilon > 0$ is a measure of how much the prey eaten contribute to growth in the predator population. A value of ε near 1 indicates that the predator doesn't need to eat very often (think of large cats such as lions or tigers where the prey is gazelle or some other large herbivore) and ε near 0 indicates a predator that must consume a large amount of prey to be satiated or satisfied (consider a frog eating flies).

A clever change of variables

$$t = \frac{\tau}{\mu}, \qquad r = \beta x, \qquad w = \gamma y \tag{1.16a}$$

$$b = \frac{\rho}{\mu} \qquad \gamma = \frac{\mu}{a}, \qquad \beta = \frac{\mu}{\varepsilon a}. \tag{1.16b}$$

turns (1.15) into

$$\frac{dx}{d\tau} = bx - xy, \tag{1.17a}$$

$$\frac{dy}{d\tau} = -y + xy, \tag{1.17b}$$

which has only one parameter b. This parameter b can be interpreted as the ratio of the growth rate of the prey to the death rate of the predator in isolation. This change of variables may seem to come out of nowhere, but it is motivated by a general strategy called *nondimensionalization* where we choose scalings in such a way to make the entire system unitless. For more on this method see Subsection 1.4.2.

We call the system (1.17) the *first-order Lotka–Volterra system.* It is our first approximation to predator-prey dynamics. As shown in Subsection 1.2.5, this is not the best model of predator-prey interactions, but it is a good first start. Plots of one solution to the model are given in Figures 1.4 and 1.8. The plot in Figure 1.4 plots each pair of values (prey population, predator population) as a single point in the plane. The space of all such pairs is called *phase space*[6]

[6] In the study of dynamical systems, the *phase space* is the space of all possible states. In this example that's $\mathbb{R}^2$, corresponding to pairs (prey, predator). Beware that the use of the word *phase* in *phase space* has nothing to do with the *phase* ϕ of a periodic function $f(t + \phi)$ (meaning a horizontal offset). Unfortunately we need both meanings of the word in this book.

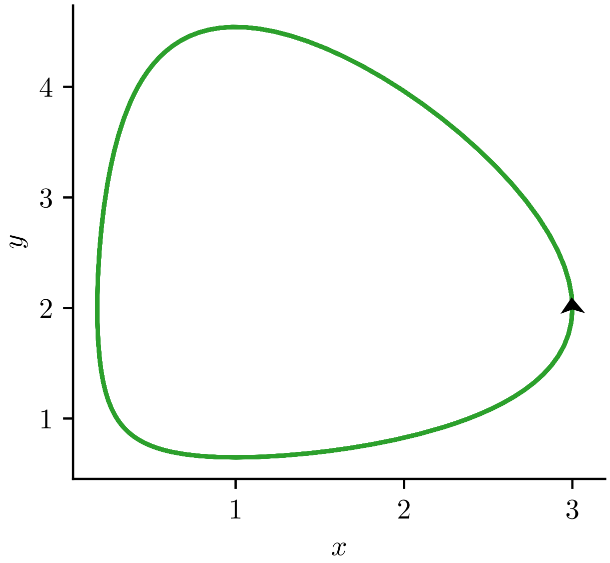

Figure 1.4: Plotting all pairs $(x(t), y(t))$ in the plane, for all values of $t > 0$ gives the *phase* plot of a solution of a two-dimensional model. This figure shows the phase plot for the predator-prey species in the Lotka–Volterra first-order system (1.17) for $b = 2$ and initial populations of $x(0) = 3$ and $y(0) = 2$ (black arrowhead). Note how the trajectory is a closed curve, indicating that over time the same values of x and y recur, in a cyclical pattern. When the predator population (y-axis) increases that leads to a decrease in the prey, leading to a decrease in predator, and so forth

1.2.5 Structural stability

It is important to recognize that you often don't get exactly what you want in your first round of modeling. Modeling is a process that takes many iterations to get 'correct.' One can show (and we do later on) that the system (1.17) has solutions that are oscillatory in nature (as illustrated in Figures 1.4 and 1.8. A precursory look at these figures makes it seem like we have solved the problem. We see the correct type of oscillatory behavior, almost exactly how we would expect it to occur. If the predator population gets too strong then the prey is eaten too quickly, the prey population declines and the predators have a corresponding lag in population. This cycle is repeated, and the two competing species have an oscillatory (in time) population structure. This explains the dynamics we set out to capture. This model looks great, and with a graphical representation of the situation this nice, you may wonder why you didn't hear about Lotka and Volterra in grade school!

But there are a few problems with this model. We don't give the details here, but one issue is that the dynamics are changed significantly if the initial conditions are changed even imperceptibly. In other words, if we started the populations of the respective species with slightly altered values, then we would get a completely different cyclical feature than that illustrated in Figure 1.4.

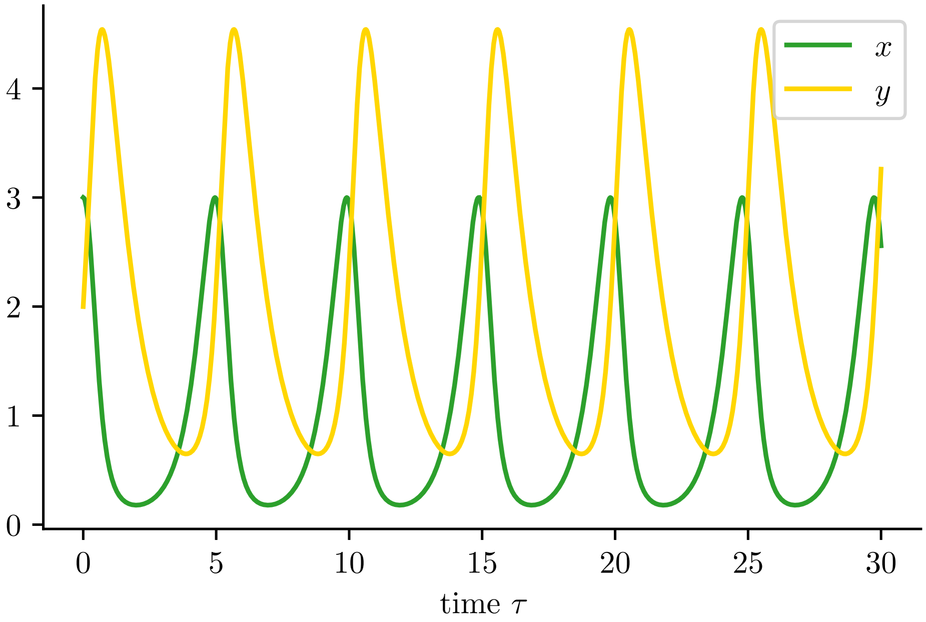

Figure 1.5: The evolution in time of the y (yellow) and x (green) values over time for the for the first-order Lotka–Volterra model. This is the same solution obtained for Figure 1.4, but where the two populations are each plotted in time, instead of as pairs of points. Starting at the initial values of $x_0 = 3$ and $y_0 = 2$, the value of x initially decreases, while the value of y increases, corresponding to $\dot{x} < 0$ and $\dot{y} > 0$. This tells us that in the phase plot in Figure 1.4 the model moves counterclockwise away from the initial point.

This is disappointing for a variety of reasons, the first being that we cannot realistically suppose that we know the initial conditions exactly. We may have a good guess, and in some controlled laboratory environments perhaps we know the initial populations exactly, but if we thought we would explain all population dynamics with this model then we would be disappointed.

Here's another problem. For a given choice of b, the initial conditions determine a unique solution $(x(t), y(t))$; but the initial conditions do not determine the parameter b, and b does not determine the initial conditions. In fact we can obtain an infinite number of these cyclical solutions, dependent on the initial conditions, without changing the parameter b. This is disappointing, because it means that in order to use this model we must somehow know b—the ratio of the growth rate of the prey to the death rate of the predator in isolation.

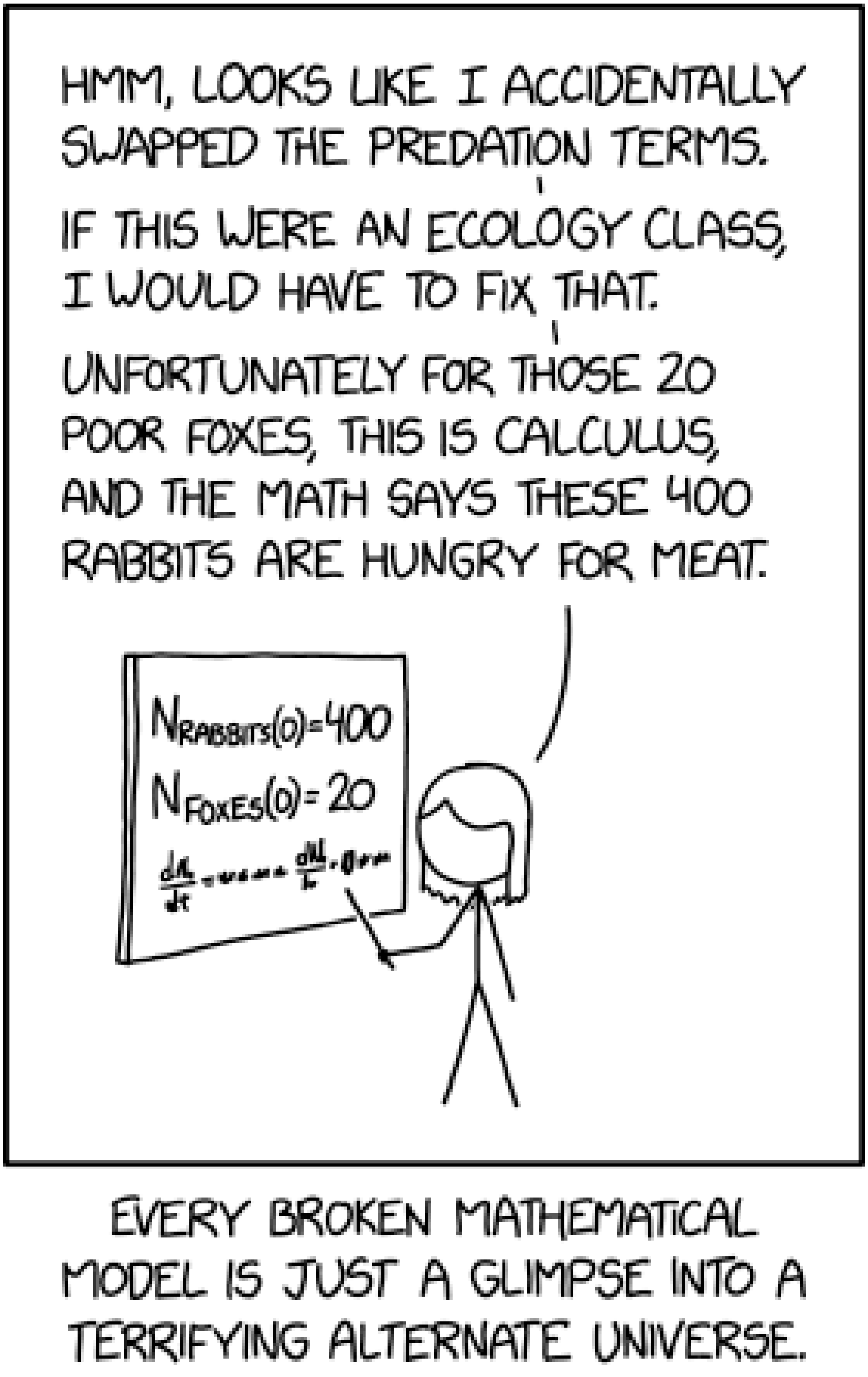
HMM, LOOKS LIKE I ACCIDENTALLY SWAPPED THE PREDATION TERMS.
IF THIS WERE AN ECOLOGY CLASS, I WOULD HAVE TO FIX THAT.
UNFORTUNATELY FOR THOSE 20 POOR FOXES, THIS IS CALCULUS, AND THE MATH SAYS THESE 400 RABBITS ARE HUNGRY FOR MEAT.
N RABBITS(0)=400
N FOXES(0)=20
EVERY BROKEN MATHEMATICAL MODEL IS JUST A GLIMPSE INTO A TERRIFYING ALTERNATE UNIVERSE.

Figure 1.6: Source: XKCD, Randall Munroe `https://xkcd.com/2945/`.

A more significant problem is that this model is based on some simplifying assumptions with the hope that is would be ‘stable’ to ignoring the details we left out. That is to say, we assumed that the model would qualitatively (approximately) yield the same results even if we used assumptions that were not quite as simple. This is, unfortunately, not true for (1.17). Consider for example if there was a limit on the resources for the prey in the absence of a predator (say water), implying a ‘carrying capacity’ for the environment that the prey live in. Rather than a simple linear growth rate in the population r, this can be modeled by an evolution equation that looks more like

$$\dot{r} = r\left(\rho - \tilde{g}r - aw\right), \tag{1.18}$$

with a corresponding nondimensionalized version of the same. It is relatively easy to check that solutions to such a system are most definitely not the same qualitatively: they no longer have equal period oscillatory behavior, but rather they decay in amplitude and eventually converge to a fixed equilibrium state.

A solution to the nondimensional version of this system (given below for convenience) is shown in Figure 1.7 for the exact same initial conditions as that shown in Figure 1.4. The nondimensional version of this system is given by:

$$\begin{aligned} \dot{x} &= bx - gx^2 - xy, && (1.19)\\ \dot{y} &= -y + xy, && (1.20) \end{aligned}$$

where the nondimensionalization and the value of g with respect to the physical parameters of the system is carried out in the exercises. We interpret g as a nondimensional measure of the carrying capacity of the environment, that is, the population density of the prey that can be supported when no predators are present.

As already mentioned, a solution to this second-order Lotka–Volterra system with the same initial conditions used in Figure 1.4 is shown in Figure 1.7. In this case, there is an equilibrium solution at which both populations can co-exist, but now there is no cyclical nature to the solution; that is, the key part of the solution that we were most excited about has disappeared entirely.

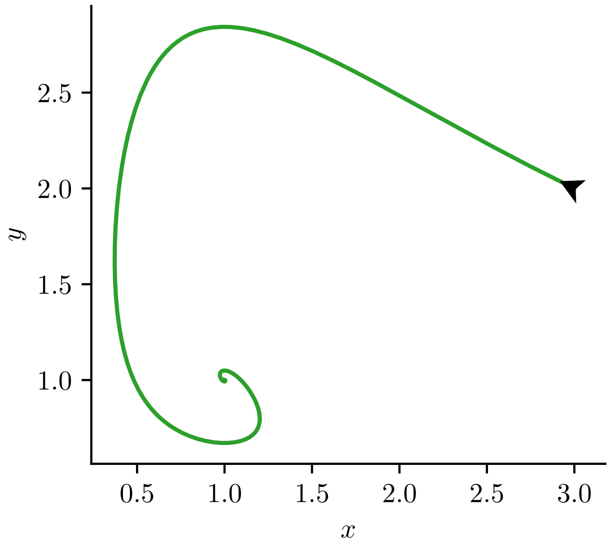

Figure 1.7: The phase space evolution of the predator-prey species in the Lotka–Volterra second-order system for $b = 2$, $g = 1$, and an initial nondimensional populations of $x(0) = 3$ and $y(0) = 2$ (black arrowhead). Note that the population dynamics move toward a fixed point at $x = 1, \ y = 1$.

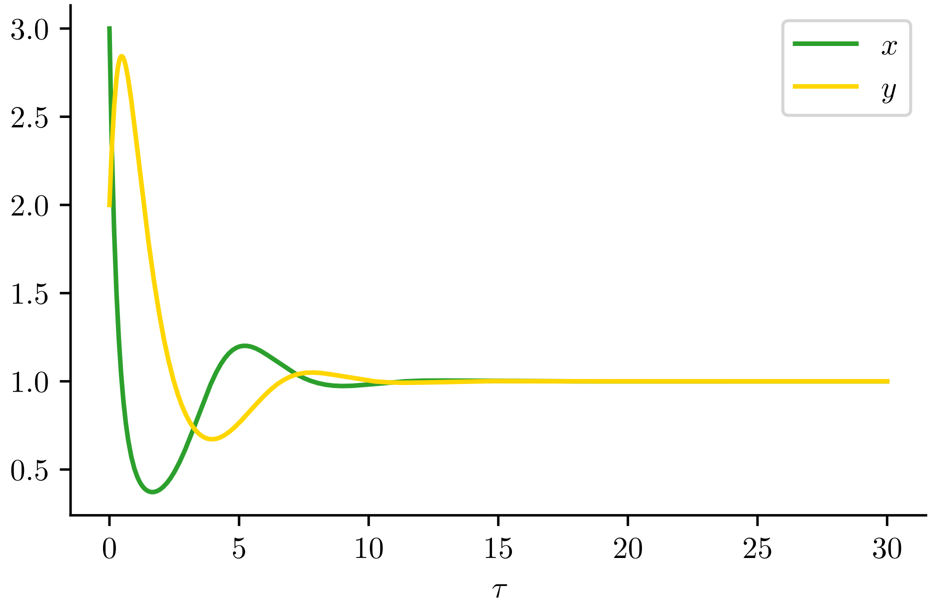

Figure 1.8: The evolution in time of the y (yellow) and x (green) values over time for the for the second-order Lotka–Volterra model with $b = 2$, $g = 1$, and an initial nondimensional populations of $x(0) = 3$ and $y(0) = 2$. This is the same solution obtained for Figure 1.7, but where the two populations are each plotted in time, instead of as pairs of points. Both the x and the y values converge to 1 as $t \to \infty$, corresponding to the the path in the phase plot Figure 1.7 moving toward the limiting point $(1, 1)$.

1.3 Dimensional Analysis and Nondimensionalization

When making mathematical models, it is very useful to think carefully about the dimensions of the physical quantities being modeled.

1.3.1 Dimensions

Physical quantities are usually measured in some sort of units, for example distance traveled could be measured in yards, inches, meters, kilometers. But regardless of the units used, distance is a length, and we say that the *dimension* of distance is length, which we denote by L. Similarly, area could be measured in square inches, or square kilometers, or acres, or hectares, but again, regardless of the units used, area has dimensions of L^2.

Most physical quantities can be expressed in terms of seven basic *dimensions*: time (T), length (L), mass (M), electric current (I), absolute temperature (Θ), amount of substance (N) and luminous intensity (J). For example, velocity is a distance (length) per time, so the dimension of velocity is LT^{-1}. Acceleration is a velocity per time so the dimensions of acceleration are LT^{-2}. The dimension of a quantity x is denoted by square brackets $[x]$ around the quantity. For example, if v is a quantity of velocity, then $[v] = \mathsf{LT}^{-1}$. Table 1.1, describes several physical quantities and their dimensions.

In order to compare any two dimensional quantities they must have the same dimension. It does not make sense to say that a 5 meters is greater than (or less than, or the same as) 3 seconds, because the first quantity has dimension L, which is not the same as the dimension T of the second quantity. Similarly, addition and subtraction of quantities only makes sense if they have the same dimension.

Definition 1.3.1. *Two quantities are* commensurable *if they are of the same kind*[7] *and have the same dimensions. Otherwise they are* incommensurable.

It does not make sense to compare, add, or subtract, incommensurable quantities. But, as mentioned above, they can be multiplied or divided. For example length L can be multiplied or divided by time T.[8] When quantities are multiplied or divided or exponentiated, then their dimensions are also multiplied, divided, or exponentiated, respectively. Force is mass times acceleration, so its dimensions are MLT^{-2}. Differentiation is a limit of ratios, so an expression of the form $\frac{df}{dq}$ has dimensions $[f][q]^{-1}$. Velocity along a line is the derivative of position with respect to time, so its dimensions are LT^{-1}. Power is the derivative of energy ($\mathsf{ML}^2\mathsf{T}^{-2}$) with respect to time, so the dimensions of power are $\mathsf{ML}^2\mathsf{T}^{-3}$.

A mathematical modeler can get a lot of mileage just from the requirement that in order to compare or add two dimensional quantities, they must be of the same dimension.

[7] Being the same *kind* is a dodgy concept, not really formally defined here; but it essentially means that it makes sense to compare the things. Torque and energy both have dimensions of $\mathsf{ML}^2\mathsf{T}^{-2}$, but they don't measure the same thing—torque is moment of rotational force, while energy is ability to do work—so it doesn't make sense to compare or add energy to torque.

[8] The fancy mathematical way of talking about this is to say that dimensions *form an Abelian group under multiplication.*

Example 1.3.2. Consider the following example from the wonderful book *Street-fighting Mathematics*: A ball initially at rest falls from a height h and hits the ground at speed v. Find v assuming a gravitational acceleration g and neglecting air resistance.

The traditional way (or at least the way most people learn in calculus) to solve this problem is to solve the initial value problem $\ddot{x}(t) = -g$ for $x(t)$ with $x(0) = h$ and $\dot{x}(0) = 0$, finding the value t_1 for which $x(t_1) = 0$, and then evaluating $v = \dot{x}(t_1)$.

But dimensional analysis can be used here. The answer to the question is a velocity, so it has dimension $\mathsf{L}\mathsf{T}^{-1}$. The other quantities we have to work with are h with dimension L, and g, with dimensions $\mathsf{L}\mathsf{T}^{-2}$. None of these quantities can be compared or added or subtracted; they are all incommensurable. But we can multiply or divide or exponentiate them to try to get the dimensions to match. Specifically, $[gh] = \mathsf{L}^2\mathsf{T}^{-2}$, which is the square of $[v] = \mathsf{L}\mathsf{T}^{-1}$. Taking the square root makes the dimensions match:

$$[\sqrt{gh}] = \mathsf{L}\mathsf{T}^{-1} = [v].$$

This implies that

$$v = c\sqrt{gh}, \tag{1.21}$$

where c is a dimensionless constant, that is the dimensions of c are $[c] = \mathsf{L}^0\mathsf{T}^0 = 1$.

You might think that (1.21) is not so useful, because the dimensionless constant could be anything. But in fact, many qualitative properties can be identified from (1.21). Among other things, it tells us how the impact velocity changes when either the initial height or gravity is changed; for example, quadrupling the initial height h doubles the impact velocity. It also gives a useful check to any solution we find using other techniques.

Definition 1.3.3. *When working with a collection of fundamental (independent) dimensions* $\mathsf{L}_1, \ldots, \mathsf{L}_m$*, we say that a quantity* c *is* dimensionless *if its dimensions are* $[c] = \mathsf{L}_1^0 \cdots \mathsf{L}_m^0$.

Of course, even if the dimensions of two quantities are the same, in order to compute with them, we expect the units to be the same. Although it makes sense to compare or add 5 m to 3 in, it's much easier to work with when all the units match. In this case we change inches to meters by the relation $1\,\mathrm{in} = 0.0254\,\mathrm{m}$, so

$$5\,\mathrm{m} + 3\,\mathrm{in} = 5\,\mathrm{m} + 3\cancel{\mathrm{in}}\,0.0254\frac{\mathrm{m}}{\cancel{\mathrm{in}}} = 5.0756\,\mathrm{m}.$$

In general, changing units involves multiplying each dimension by an appropriate constant.

Example 1.3.4. The previous example showed that, neglecting air resistance, the impact velocity v_i of an object dropped from height h is $c\sqrt{gh}$, where g is acceleration due to gravity and h is the initial height. It is straightforward to show (using calculus, for example) that the constant in this law is $\frac{1}{2}$, so the physical law is

$$v_i = \frac{1}{2}\sqrt{gh}. \tag{1.22}$$

If time is measured in seconds (s) and length in meters (m), then $h = \eta\,\mathrm{m}$ for some (dimensionless) η, acceleration $g = \gamma\frac{\mathrm{m}}{\mathrm{s}^2}$ for some (dimensionless) γ, and $v_i = \nu\frac{\mathrm{m}}{\mathrm{s}}$ for some dimensionless ν. The law (1.22) written in these units is

$$\nu\frac{\mathrm{m}}{\mathrm{s}} = \frac{1}{2}\sqrt{\eta\gamma\frac{\mathrm{m}^2}{s^2}} = \frac{1}{2}\sqrt{\eta\gamma}\frac{\mathrm{m}}{\mathrm{s}}. \tag{1.23}$$

Notice that the units on both sides of (1.23) are the same—as expected when the dimensions are comparable. Dividing out the units, this implies that

$$\nu = \frac{1}{2}\sqrt{\eta\gamma}. \tag{1.24}$$

Changing the seconds and meters to minutes and inches, respectively, in (1.23) gives

$$\eta\,\mathrm{m} = \eta\cancel{\mathrm{m}}\frac{39.3701\,\mathrm{in}}{\cancel{\mathrm{m}}} = 39.3701\eta\,\mathrm{in},$$

and

$$\gamma\,\mathrm{m/s}^2 = \gamma\frac{\cancel{\mathrm{m}}}{\cancel{\mathrm{s}}^2}\frac{39.3701\,\mathrm{in}}{\cancel{\mathrm{m}}}\left(\frac{60\cancel{\mathrm{s}}}{\mathrm{min}}\right)^2 = 39.3701\times 60^2\gamma\frac{\mathrm{in}}{\mathrm{min}^2},$$

and

$$\nu\,\mathrm{m/s} = \nu\frac{\cancel{\mathrm{m}}}{\cancel{\mathrm{s}}}\frac{39.3701\,\mathrm{in}}{\cancel{\mathrm{m}}}\frac{60\cancel{\mathrm{s}}}{\mathrm{min}} = 39.3701\times 60\nu\frac{\mathrm{in}}{\mathrm{min}}.$$

Note that the law (1.22) still holds in these units; that is,

$$v_i = 39.3701\times 60\nu\frac{\mathrm{in}}{\mathrm{min}} \overset{\star}{=} \frac{1}{2}\sqrt{(39.3701^2)(60^2)\eta\gamma\frac{\mathrm{in}^2}{\mathrm{min}^2}} = \frac{1}{2}\sqrt{gh},$$

where the starred equality holds by (1.24).

Any equality $q_1 = q_2$ of commensurable quantities can be rewritten as $q_1 - q_2 = 0$. But if the quantities are not commensurable, neither of these equalities makes sense. More generally, it does not make physical sense to talk about a function of dimensional quantities unless the function has the property of being *dimensionally homogeneous* or *unit free*, as defined below.

Definition 1.3.5. *For any dimension* L*, replacing* L *by* $\lambda\mathsf{L}$ *for some nonzero, dimensionless constant* λ *is called a* change of units *of* L*. We sometimes write this as* $\widetilde{\mathsf{L}} = \lambda\mathsf{L}$*. For any dimensional quantity* q *expressed in terms of* m *fundamental (independent) dimensions* $\mathsf{L}_1, \dots, \mathsf{L}_m$ *(so* $[q] = \mathsf{L}_1^{a_1} \cdots \mathsf{L}_n^{a_n}$*), a change of units via* $\lambda_1, \dots, \lambda_m$ *causes a rescaling of* q *by* $\lambda_1^{a_1} \cdots \lambda_m^{a_m}$*, which we denote as* $\widetilde{q}$ *and call the* change of units *of* q *by* $\lambda_1, \dots, \lambda_m$*:*

$$\widetilde{q} = q(\lambda_1^{a_1}\mathsf{L}_1, \dots, \lambda_m^{a_m}\mathsf{L}_m) = \lambda_1^{a_1} \cdots \lambda_m^{a_m} q(\mathsf{L}_1, \dots, \mathsf{L}_m) = \lambda_1^{a_1} \cdots \lambda_m^{a_m} q.$$

A function $f(q_1, \dots, q_n)$ *of dimensional quantities* $q_1, \dots, q_n$ *expressed in terms of* n *fundamental (independent) dimensions* $\mathsf{L}_1, \dots, \mathsf{L}_m$*, is called* dimensionally homogeneous *(or* unit free*) if any change of units by* $\lambda_1, \dots, \lambda_m$ *of the dimensions results satisfies*

$$f(\widetilde{q_1}, \dots, \widetilde{q_n}) = \lambda_1^{b_1} \cdots \lambda^{b_m} f(q_1, \dots, q_n)$$

for some choice of exponents $b_1, \dots, b_m$*. In this case the dimension of* f *is* $[f] = \mathsf{L}_1^{b_1} \cdots \mathsf{L}_m^{b_m}$*.*

If an expression $f(q_1, \dots, q_n)$ is not dimensionally homogeneous, then it does not have a well-defined dimension and changing units gives fundamentally different results.

In Example 1.3.4, the law $v_i = \frac{1}{2}\sqrt{gh}$ can be rewritten as $v_i - \frac{1}{2}\sqrt{gh} = 0$, where the left side $f(v_i, g, h) = v_i - \frac{1}{2}\sqrt{gh}$ is a dimensionally homogeneous function f of the dimensional quantities v_i, g and h. It is homogeneous because changing units by $\mathsf{L} \mapsto \lambda_1\mathsf{L}$ and $\mathsf{T} \mapsto \lambda_2\mathsf{T}$ gives $f(\widetilde{v}_i, \widetilde{g}, \widetilde{h}) = \lambda_1\lambda_2^{-1} f(v_i, g, h)$, corresponding to the fact that every term in f, and f itself, has dimension LT^{-1}.

Remark 1.3.6. Since $0 \cdot \mathsf{L}_1^{a_1} \cdots \mathsf{L}_M^{a_m} = 0$ for any exponents $a_1, \dots, a_m$, the number zero has the same dimension as any dimensional quantity or any dimensionally homogeneous function. Thus the equality $f(q_1, \dots, q_n) = 0$ makes sense whenever f is dimensionally homogeneous.

It does not make sense to talk about exp of a dimensional quantity because every term in the expansion $\exp(x) = 1 + x + \frac{x^2}{2} + \cdots$ has a different dimension. If $[x] = \mathsf{L}$, then the $[x^2] = \mathsf{L}^2$, and each term in the sum involves a different power of L. Thus the sum does not make sense. In general, transcendental functions only make sense for dimensionless quantities.

But integration and differentiation do make sense for dimensional quantities. Integration is a limiting process applied to a sum and product: $\int_a^b f(x)\,dx = \lim_{n\to\infty} f(x_i)\Delta x$, where $\Delta x = \frac{b-a}{n}$. The numbers b and a have the same dimension as x and n is a dimensionless count (the number of terms in a Riemann sum), so Δx has the same dimensions as x. As long as $f(x)$ has a meaningful consistent dimension and x has a meaningful consistent dimension, then it is permissible to multiply $f(x_i)\,\Delta x$ and add the terms. The limit preserves dimensions, so $[dx] = [x]$ and $\left[\int_a^b f(x)\,dx\right] = [f(x)\,dx] = [f(x)][dx]$.

Similarly, derivatives are a limit of a difference quotient $\frac{d}{dx}f(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$, where h must have the same dimension as x (otherwise $x + h$ wouldn't make sense), so $\left[\frac{d}{dx}f(x)\right] = \frac{[f(x)]}{[x]}$, provided $[f(x)]$ makes sense.

Quantity	Dimensions	Quantity	Dimensions
velocity	LT^{-1}	frequency	T^{-1}
acceleration	LT^{-2}	volume	L^3
force	MLT^{-2}	pressure	$\mathsf{ML}^{-1}\mathsf{T}^{-2}$
momentum	MLT^{-1}	angular momentum	$\mathsf{ML}^2\mathsf{T}^{-1}$
energy, heat, work	$\mathsf{ML}^2\mathsf{T}^{-2}$	torque	$\mathsf{ML}^2\mathsf{T}^{-2}$
power	$\mathsf{ML}^2\mathsf{T}^{-3}$	density	ML^{-3}

Table 1.1: Common physical quantities expressed in terms of three of the fundamental dimensions: length (L), time (T), and mass (M).

1.3.2 Dimensional Analysis for Problems Without Dimensions

Even if a problem doesn't come with a natural choice of dimensions, if we thoughtfully give dimensions (in a consistent way) to some of the quantities in the problem, then dimensional analysis can often be useful.

Example 1.3.7. Consider the problem of computing the integral $\int_0^\infty e^{-\alpha t^3}\,dt$, for a constant α. There are no dimensions given here, but we might choose to give t the dimension T. Note that exponents must be dimensionless because in an expressions like $a^m = \underbrace{a \cdot a \cdots a}_{m}$, the exponent m is the number of times that a is multiplied by itself. Therefore, the exponent αt^3 should be dimensionless, and thus $[\alpha] = \mathsf{T}^{-3}$. The number e is dimensionless, so the integrand $e^{-\alpha t^3}$ is also dimensionless. The term dt in the integral represents the limit of quantities Δt of the same dimension as t, so $[dt] = \mathsf{T}$. The integral itself is just the limit of a sum, so it doesn't change the dimension. Thus the solution to this problem must have dimension T.

The variable t is integrated away, so the solution cannot involve t and must be constructed from α alone. That means the solution must be of the form α^m for some m. But $[\alpha^m] = \mathsf{T}^{-3m}$ must be T, so $m = -\frac{1}{3}$. This implies that

$$\int_0^\infty e^{-\alpha t^3}\,dt = \frac{c}{\alpha^{1/3}},$$

for some dimensionless constant c. To find the constant, we could compute the integral for one value of α. For example, taking $\alpha = 1$, we can integrate numerically to get $c = 0.8929795$.

1.3.3 Nondimensionalization

Many differential equations can be simplified by a change of variables to make all the quantities nondimensional. This is best illustrated with an example.

Example 1.3.8. Consider the ODE

$$\dot{x} = a + bx + cx^2, \tag{1.25}$$

where a, b, and c are (dimensional) constants. Assume that $[x] = \mathsf{L}$, and $[t] = \mathsf{T}$. To remove the dimensions from x and t we make a substitution $\xi = \frac{x}{d}$ and $\tau = \frac{t}{k}$ for yet-to-be-determined constant values of d and k, satisfying $[d] = [x] = \mathsf{L}$ and $[k] = [t] = \mathsf{T}$. Substituting $x(t) = d\xi(t)$ and $t = k\tau$ into the original equation (and using the chain rule) gives

$$\frac{dx}{dt} = \frac{d}{k}\frac{d\xi}{d\tau} = a + (bd)\xi + (cd^2)\xi^2,$$

which we can rewrite as

$$\frac{d\xi}{d\tau} = \frac{k}{d}a + (bk)\xi + (ckd)\xi^2.$$

Choosing $k = \frac{1}{b}$ and $d = ak = \frac{a}{b}$ turns this into

$$\frac{d\xi}{d\tau} = 1 + \xi + \gamma\xi^2, \tag{1.26}$$

where $\gamma = \frac{a^2c}{b^3}$. The original problem in now not only dimensionless, but it is also simpler than the original, since it depends on only one parameter instead of three. This means that to solve an equation of the form (1.26) for any choice of a, b and c (with $a \neq 0 \neq b$), we need only solve the simpler equation (**??**).

1.4 Buckingham π Theorem

The previous example of nondimensionalization might seem a little ad hoc, but the Buckingham π theorem tells us that this can be done in great generality and it gives tools for how to find the right substitutions. It is a fundamental tool in dimensional analysis and says, essentially, that any dimensionally homogeneous law

$$f(q_1, \ldots, q_n) = 0$$

relating n dimensional quantities $q_1, \ldots, q_n$, is equivalent to a law

$$g(\pi_1, \ldots, \pi_s) = 0$$

relating $s < n$ dimensionless quantities $\pi_1, \ldots, \pi_s$. This equivalence can lead to new not-otherwise-obvious relations among the quantities involved. Moreover, the dimensionless relation is usually simpler because the number s of dimensionless quantities involved is strictly less than the number of dimensional quantities, and the dimensionless law has no units.

1.4.1 Buckingham π

Theorem 1.4.1 (Buckingham π). [9] *Given n dimensional quantities $q_1, q_2, \ldots, q_n$ satisfying*

$$[q_j] = \mathsf{L}_1^{a_{1j}} \mathsf{L}_2^{a_{2j}} \cdots \mathsf{L}_m^{a_{mj}}, \tag{1.27}$$

where $\mathsf{L}_1, \mathsf{L}_2, \ldots, \mathsf{L}_m$ are fundamental dimensions, assume the $m \times n$ matrix $A = [a_{ij}]$ of exponents has nullity s (the dimension of its kernel). If we have a dimensionally homogeneous physical law

$$f(q_1, q_2, \ldots, q_n) = 0,$$

then there exists dimensionless quantities $\pi_1, \pi_2, \ldots, \pi_s$ formed from $q_1, q_2, \ldots, q_n$ and an equivalent dimensionless physical law

$$g(\pi_1, \pi_2, \ldots, \pi_s) = 0. \tag{1.28}$$

We give a proof here only for the special case that $f(q_1, \ldots, q_n)$ can be written as a sum of monomials of the form $q_1^{x_1} \cdots q_n^{x_n}$ for various choices of $\mathbf{x} = (x_1, \ldots, x_n)$. The general proof is given at the end of this section.

Proof. Because f is dimensionally homogeneous, there exists $\mathbf{b} = (b_1, \ldots, b_m)$ such that changing units of the dimensions by $\lambda_1, \ldots, \lambda_m$ scales f by $\lambda_1^{b_1} \cdots \lambda_m^{b_m}$. For cleaner notation, define the function

$$\mathbf{q}^{\mathbf{x}} = q_1^{x_1} \cdots q_n^{x_n}$$

for any $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Note that $\mathbf{q}^{\mathbf{x}+\mathbf{y}} = \mathbf{q}^{\mathbf{x}}\mathbf{q}^{y}$ for any vectors $\mathbf{x}$ and $\mathbf{y}$. The assumption about f being a sum of monomials can be expressed using $\mathbf{q}^{\mathbf{x}}$ as saying that $f(q_1, \ldots, q_n)$ is a sum of terms of the form $\mathbf{q}^{\mathbf{x}}$ for various choices of $\mathbf{x}$. Combining this assumption with (1.27) shows that $A\mathbf{x} = \mathbf{b}$ for each $\mathbf{x}$ that appears as a term in f.

Let $\mathbf{x}_0$ be one such solution, so that $A\mathbf{x}_0 = \mathbf{b}$. Any $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{b}$ must be of the form $\mathbf{x} = \mathbf{v} + \mathbf{x}_0$ for some $\mathbf{v} \in \mathscr{N}(A)$, where $\mathscr{N}(A)$ is the kernel (nullspace) of A. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$ be a basis for $\mathscr{N}(A)$ with $\mathbf{v}_j = (v_{1j}, \ldots, v_{nj})$ for each $j \in \{1, \ldots, s\}$ and define

$$\pi_k = \mathbf{q}^{\mathbf{v}_k} = q_1^{v_{1k}} \cdots q_n^{v_{nk}} \quad \text{for each } k \in \{1, \ldots, s\}.$$

For any $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{b}$ we have $\mathbf{x} = \mathbf{x}_0 + \sum_{k=1}^{s} c_k \mathbf{v}_k$, for some choice of $\mathbf{c} = (c_1, \ldots, c_s)$, which gives

$$\mathbf{q}^{\mathbf{x}} = \mathbf{q}^{\mathbf{x}_0} \prod_{k=1}^{s} \mathbf{q}^{c_k \mathbf{v}_k} = \mathbf{q}^{\mathbf{x}_0} \prod_{k=1}^{s} \pi_k^{c_k}.$$

This implies that every summand in f is of the form $\mathbf{q}^{\mathbf{x}_0} \prod_{k=1}^{s} \pi_k^{c_k}$ for some choice of $\mathbf{c} = (c_1, \ldots, c_s)$. Thus every term of f is divisible by $\mathbf{q}^{\mathbf{x}_0}$, and

$$0 = f(q_1, \ldots, q_n) = \mathbf{q}^{\mathbf{x}_0} g(\pi_1, \ldots, \pi_s),$$

[9] Of course the Buckingham π theorem is not actually due to Buckingham, but rather Joseph Bertrand, inspired by ideas from Rayleigh. This is an example of *Stigler's Law of Eponomy*, which says that no scientific discovery is named after its discoverer, not even Stigler's Law of Eponomy.

for some function $g(\pi_1, \ldots, \pi_s)$ depending only on the dimensionless quantities $\pi_1, \ldots, \pi_s$. Dividing by $\mathbf{q}^{\mathbf{x}_0}$ gives (1.28). which is the desired dimensionless physical law. □

Remark 1.4.2. If there is only one dimensionless quantity π expressed in terms of the q_i, then Buckingham π says that there are terms of the form π^c that sum to 0. That means that π itself must be constant. The quantity π need not be zero, because one of the terms in $g(\pi)$ could, itself, be a (dimensionless) nonzero constant.

With s dimensionless quantities $g(\pi_1, \ldots, \pi_s) = 0$, the implicit function theorem guarantees that $\pi_s = h(\pi_1, \ldots, \pi_{s-1})$ for some function h. Of course the choice of which of the π_i to put last and label as π_s is arbitrary.

Example 1.4.3. Consider a projectile of mass m launched vertically into the air with initial velocity v. What is the maximal height h that the projectile will reach before coming down? The dimensional quantities are mass m, velocity v, height h, and the gravitational constant g, with dimensions $[m] = \mathsf{M}$, $[v] = \mathsf{LT}^{-1}$, $[h] = \mathsf{L}$, and $[g] = \mathsf{LT}^{-2}$. The matrix A has rows corresponding to M, L, and T, respectively, and columns corresponding to m, v, h, and g, respectively.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & -1 & 0 & -2 \end{bmatrix}$$

This has rank 3 and nullity 1, so the Buckingham π theorem says there is one dimensionless quantity π and a dimensionless physical law $g(\pi) = 0$, so as pointed out in Remark 1.4.2, the quantity π must be constant. A little linear algebra shows that the null space $\mathscr{N}(A)$ is spanned by the vector $\mathbf{v} = (0, -2, 1, 1)$, which gives $\pi = \mathbf{q}^{\mathbf{v}} = \frac{gh}{v^2} = c$ for some constant c. This implies

$$h = \frac{cv^2}{g}. \tag{1.29}$$

It is interesting to note that this guarantees the solution is independent of mass.

We cannot determine the constant c with dimensional analysis, but the following physics argument shows that $c = \frac{1}{2}$. The kinetic energy at launch $\frac{1}{2}mv^2$ is entirely converted into potential energy mgh when the projectile reaches its apex, therefore $\frac{1}{2}mv^2 = mgh$, which shows, after substituting (1.29), that $c = \frac{1}{2}$.

Example 1.4.4. I ride my bike to work most days and notice that when I ride faster, air resistance exerts a force on me. I'd like to model the effect of wind resistance with a differential equation. The force F of wind resistance seems to depend on my velocity v and how upright I am sitting—so I guess it depends on the area a that my body presents to the oncoming air. Air resistance is stronger at lower altitudes, and I'm told it is zero in a vacuum (I've never tested that), so it seems reasonable to think that the force is also a function of air density ρ. So we expect to be able to make a model like $F = h(v, a, \rho)$ for some function h. The relevant dimensions are

$$[F] = \mathsf{MLT}^{-2} \qquad [v] = \mathsf{LT}^{-1}$$
$$[a] = \mathsf{L}^{2} \qquad [\rho] = \mathsf{ML}^{-3}$$

So we may assume the fundamental dimensions are M, L, and T. Assume there is a physical law of the form $F - h(v, a, \rho) = 0$. The exponent matrix A is

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 2 & -3 \\ -2 & -1 & 0 & 0 \end{bmatrix},$$

which has rank 3 and nullity 1. It is straightforward to check that the kernel $\mathscr{N}(A)$ is spanned by $\mathbf{v} = (1, -2, -1, -1)$, so $\pi = Fv^{-2}a^{-1}\rho^{-1}$ is a dimensionless constant. This implies that $F = cv^2a^{-1}\rho^{-1}$ for some dimensionless constant. If $y(t)$ is my position at time t and the density ρ and the area a are constant in time, then this can be modeled with a differential equation

$$m\ddot{y} = c\rho a(\dot{y})^2,$$

where m is my mass (plus the mass of the bicycle). We know that mass term m is necessary not only because of Newton's second law, but also because the the dimensions of force always include M and must match the dimensions of the right side, which includes M (from the density term ρ).

Example 1.4.5. The famous fluid dynamicist G.I. Taylor combined the images published in September 17, 1945 issue of *Time Magazine* (see Figure 1.9 for two of the 25 images) with some basic empirical knowledge from small explosives to estimate the amount of energy in the world's first atomic explosion, code named Trinity. He was able to reveal top secret information from publicly available data using Buckingham's theorem.

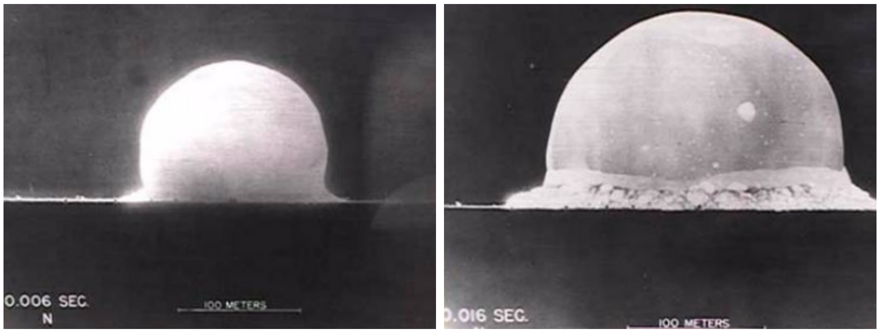

Figure 1.9: Two of the 25 images taken from the Trinity explosion. Using the scale on the image and time stamp, Taylor was able to determine several points of radius vs. time graph and used that to help fit the curve to determine the energy of the explosion.

Assume the dimensional quantities for a nuclear explosion are time t, energy E, air density ρ, and the radius r of the shock wave. The dimensions are $[t] = \mathsf{T}$, $[E] = \mathsf{ML}^2\mathsf{T}^{-2}$, $[\rho] = \mathsf{ML}^{-3}$, and $[r] = \mathsf{L}$, depending on only three fundamental dimensions M, L and T. The matrix A is

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 2 & -3 & 1 \\ 1 & -2 & 0 & 0 \end{bmatrix},$$

which has rank 3 and nullity 1. The kernel $\mathscr{N}(A)$ is spanned by $\mathbf{v} = (-2, -1, 1, 5)$. This gives

$$\pi = \frac{\rho r^5}{Et^2}.$$

As before, we have $\pi = c$ for some constant c, and thus

$$E = \frac{\rho r^5}{ct^2}$$

From small explosives tests, it has been estimated that $c = (1.033)^5$. Taking average air density in the region of New Mexico to be $\rho = 1.21\,\mathrm{kg/m}^3$, and putting in the measured radius at the time of each photograph (marked on the photographs), gives an estimated value of E from each photograph. Taylor fit the values from the 25 published images to get the estimate $E = 7.19 \times 10^{13}\,J$, which equates to 17.5 kilotons of TNT. The actual amount was reported to be between 15–20 kilotons, so this was a pretty good estimate. Supposedly Taylor was admonished by the U.S. government for revealing classified details of the atomic bomb even though he derived it all from unclassified, publicly available information.

1.4.2 Nondimensionalization with Buckingham π

Many models can be simplified by recasting them in nondimensional form. This process is often called *nondimensionalization.* We illustrate how to use the idea to simplify the Lotka–Volterra model introduced in the previous section.

In the Lotka–Volterra model (1.15) we are concerned with one independent variable (time), two dependent variables (prey r and predator w), and four parameters (ρ, a, μ, and ε). The dimensions of r and w are $[r] = \mathsf{R}$ (number of rabbits) and $[w] = \mathsf{W}$ (number of wolves). The rates ρ and μ have dimension

$$[\rho] = \mathsf{T}^{-1} = [\mu]$$

because the equation (1.13) defining ρ equates $\dot{r}$ with ρr, so the dimensions must match: $[\dot{r}] = \mathsf{RT}^{-1} = [\rho r] = [\rho]\mathsf{R}$, and similarly for $[\mu]$ from (1.14). Matching dimensions in (1.15a) and (1.15b) shows that

$$[a] = \mathsf{W}^{-1}\mathsf{T}^{-1} \qquad \text{and} \qquad [\varepsilon] = \mathsf{R}^{-1}\mathsf{W}.$$

So there are three fundamental dimensions R, W, and T. The exponent matrix A for the nine quantities t, r, w, $\dot{r}$, $\dot{w}$, ρ, a, μ, ε is

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 \end{bmatrix},$$

which has rank 3 and nullity 6. Buckingham π guarantees six nondimensional quantities, which reduces the total number of quantities in our model by three. There are (infinitely) many ways to construct the nondimensional quantities, because there are infinitely many different choices of basis for $\mathscr{N}(A)$.[10] We choose the basis

$$\begin{aligned}
\mathbf{v}_1 &= (1,0,0,0,0,0,0,1,0) && \text{corresponding to } \pi_1 = t\mu \\
\mathbf{v}_2 &= (0,1,0,0,0,0,1,-1,1) && \text{corresponding to } \pi_2 = ra\mu^{-1}\varepsilon \\
\mathbf{v}_3 &= (0,0,1,0,0,0,1,-1,0) && \text{corresponding to } \pi_3 = wa\mu^{-1} \\
\mathbf{v}_4 &= (0,0,0,0,0,1,0,-1,0) && \text{corresponding to } \pi_6 = \rho\mu^{-1} \\
\mathbf{v}_5 &= (0,0,0,1,0,0,1,-2,1) && \text{corresponding to } \pi_4 = \dot{r}a\mu^{-2}\varepsilon \\
\mathbf{v}_6 &= (0,0,0,0,1,0,1,-2,0) && \text{corresponding to } \pi_5 = \dot{w}a\mu^{-2}
\end{aligned}$$

It is straightforward to check that this set is linearly independent and lies in the kernel $\mathscr{N}(A)$. We rename four of these new nondimensional parameters as

$$\begin{array}{lll}
\tau = \pi_1 = t\mu & x = \pi_2 = ra\mu^{-1}\varepsilon & \dot{y} = \\
y = \pi_3 = wa\mu^{-1} & b = \pi_4 = \rho\mu^{-1}, &
\end{array}$$

[10] When doing nondimensionalization, there are always free choices, but typically one choice is preferred. Sometimes this is because it makes the parameters fundamentally easier to deal with than any other choice, but usually the reason is just that it was the choice learned by everyone who previously studied the particular model. It usually pays to use the standard choice so that you don't have to try to get octogenarian scientists to take you seriously when you use a different form of the equations than they have ever seen.

which matches the "clever change of variables" (1.16) that we used in the previous section. We still haven't dealt with the parameters π_5 and π_6. These are clearly related to the derivatives $\frac{d}{d\tau}x$ and $\frac{d}{d\tau}y$ of x and y, respectively. But we want to use the derivatives in the final model rather than π_4 and π_5. Because x and y and τ are dimensionless, the derivatives $\frac{d}{d\tau}x$ and $\frac{d}{d\tau}y$ are, too, so they should be expressible in terms of the various terms π_k. The chain rule gives

$$\begin{aligned}\frac{dx}{d\tau} &= \frac{dx}{dt}\frac{dt}{d\tau} = \dot{r}a\mu^{-1}\varepsilon\mu^{-1} = \pi_4\\ \frac{dy}{d\tau} &= \frac{dy}{dt}\frac{dt}{d\tau} = \dot{r}a\mu^{-1}\varepsilon\mu^{-1} = \pi_4\end{aligned}$$

To express the original Lotka–Voltera model (1.15) in terms of these new, nondimensional variables, first solve for the variables $t = \tau\mu^{-1}$, $r = xa^{-1}\mu\varepsilon^{-1}$, $w = ya^{-1}\mu$ and $\rho = b\mu$, and then substitute these expressions into (1.15). Simplifying yields the final nondimensionalized form (1.17).

1.4.3 Proof of Buckingham π in the General Case

Here we give the proof of the Buckingham π Theorem (Theorem 1.4.1) in full generality, that is, not assuming that f can be written as a sum of monomials of the form $\mathbf{q}^{\mathbf{x}}$.

Proof. If A has nullity s and rank r with $r+s=n$, then we may choose r columns of A that span that column space of A. By reindexing the quantities $q_1,\ldots,q_n$, we may assume that the first r columns $\mathbf{a}_1,\ldots,\mathbf{a}_r$ of A span the column space of A.

Changing the units of the fundamental dimensions by $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m)$ (that is, $\mathsf{L}_i \mapsto \lambda_i\mathsf{L}_i$) induces a rescaling of the quantities $q_1,\ldots,q_n$ by $\boldsymbol{\lambda}^{\mathsf{T}}A$; that is, the jth entry of the row vector $\boldsymbol{\lambda}^{\mathsf{T}}A$ is the rescaling factor of q_j induced by the change of units. The first r columns of A are linearly independent, which implies that for any values $q_1,\ldots,q_r$ there is at least one solution $\boldsymbol{\lambda}$ to the system $\boldsymbol{\lambda}^{\mathsf{T}}A = (q_1^{-1},\ldots,q_r^{-1})^{\mathsf{T}}$. That means we may change units to make the quantities $q_1,\ldots,q_r$ all equal to 1 for an appropriate choice of units.

The fact that the first r columns span the column space of A implies that the remaining $s=n-r$ columns can be written as a linear combination of the first r columns:

$$\mathbf{a}_{r+k} = \sum_{j=1}^{r} c_{jk}\mathbf{a}_j.$$

Let $\mathbf{e}_\ell = (0,\ldots,0,1,0,\ldots,0)$ be the standard basis vector in $\mathbb{R}^n$ with 1 appearing in only the ℓth position. Let $\mathbf{c}_k \in \mathbb{R}^n$ be the vector $\mathbf{c}_k = (c_{1k},\ldots,c_{rk},0,0,0,\ldots,0)$, and for each $k \in \{1,\ldots,s\}$ let $\mathbf{v}_k = \mathbf{e}_{r+k} - \mathbf{c}_k$ The set $\mathbf{v}_1,\ldots,\mathbf{v}_s$ spans the kernel $\mathscr{N}(A)$, and we define $\pi_k = \mathbf{q}^{\mathbf{v}_k}$. The quantities π_k are dimensionless because $\mathbf{v}_k \in \mathscr{N}(A)$. Observe that

$$\pi_k\mathbf{q}^{\mathbf{c}_k} = \mathbf{q}^{\mathbf{e}_{r+k}-\mathbf{c}_k}\mathbf{q}^{\mathbf{c}_k} = \mathbf{q}^{\mathbf{e}_{r+k}} = q_{r+k}.$$

Thus any q_{r+k} can be written in terms of $q_1, \ldots, q_r$ and π_k alone. Substituting these into the original law $f(q_1, \ldots, q_n) = 0$, we have

$$f(q_1, \ldots, q_r, \pi_1 \mathbf{q}^{\mathbf{c}_1}, \pi_2 \mathbf{q}^{\mathbf{c}_2}, \ldots, \pi_s \mathbf{q}^{\mathbf{c}_s}) = 0,$$

which depends only on the $\pi_1, \ldots, \pi_s$ and $q_1, \ldots, q_r$. But, as argued earlier, there is a change of units to make $q_1, \ldots, q_r$ all equal to 1. Setting

$$g(\pi_1, \ldots, \pi_s) = f(1, 1, \ldots, 1, \pi_1, \ldots, \pi_s)$$

gives the desired dimensionless law

$$g(\pi_1, \ldots, \pi_s) = 0. \quad \Box$$

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

1.1. Verify that (1.5) is a solution of (1.4).

1.2. Prove that the initial conditions $y(0) = y_0$ and $\dot{y}(0) = v_0$ uniquely determine the solution (1.8). Hint: the value of ϕ may not be unique, but the solution (1.8) is. Explain.

1.3. Prove that every solution of the damped harmonic oscillator discussed in Section 1.1.3 (underdamped, overdamped, and critically damped) has limit 0 as $t \to \infty$, regardless of the initial conditions.

1.4. Using Newton's second law $F = ma$, model (give a differential equation for) the situation where Bob is subject not only to damping (from friction) but also another force $F(t)$. This is called the *forced-and-damped harmonic oscillator*.

1.5. The solution of the forced-and-damped harmonic oscillator depends on the exact form of the force $F(t)$, and, as you might guess, this additional force can drastically affect the dynamical evolution of the spring.

(i) Given a constant force $F(t) = c$ in one direction, predict, from your physical experience, what Bob's final limiting position will be.

(ii) Consider a periodic force $f(t) = A_0 \cos(\omega_0 t)$ for some amplitude A_0 and frequency ω_0 (think of giving Bob a push at regular time intervals). If ω_0 has just the right value ($\omega_0 = \sqrt{\omega^2 - \gamma^2}$), then something new happens, giving a solution that we haven't seen in any of the other harmonic oscillators. What would you expect to happen to Bob in this case as $t \to \infty$? Explain.

1.6. Show that the total population of the SIR model in (1.9) does not change in time.

1.7. Develop (with brief, reasonable explanation) a variant of the SIR model in which there are two stages to the disease. Stage-one infected individuals $I_1(t)$ are the only ones that can infect the susceptible population $S(t)$, while stage-two infected individuals $I_2(t)$, while sick, can not infect others. Make certain that the total population stays constant.

1.8. Develop yet another variant of the SIR model, again with a reasonable explanation, that includes the possibility that a small percentage of the recovered population $R(t)$ can become susceptible again. Make certain that the total population stays constant in this model.

1.9. Modify the logistic equation to consider the generic situation where the carrying capacity of the environment is time-dependent. Justify your choices. What type of time dependence makes sense, and for what reasons?

1.10. In your own words, explain why the differences between the plots in Figures 1.7 and 1.4 indicate a fundamental problem with the simplifying assumptions that went into the Lotka–Volterra first-order model (1.17).

1.11. Assume that a pendulum of length ℓ with a bob on the end of mass m moves only through a very small angle. Explain why its period P cannot depend only on ℓ and m. Show that if P is known to depend only on ℓ and m and the acceleration g due to gravity, then $P = c\sqrt{\frac{\ell}{g}}$. This tells us that changing the mass of the bob does not change the period P, so adding a larger bob to your grandfather clock won't change how it keeps time, but changing the length of the pendulum will, and so will changing the altitude of the clock (since g depends on the altitude).

1.12. The speed v of a wave in deep water is determined by its wavelength λ and the acceleration g due to gravity. What does dimensional analysis imply regarding the relationship between v, λ, and g?

1.13. A small sphere of radius r and density ρ is falling at constant velocity v under the influence of gravity g in a liquid of density ρ_ℓ and viscosity μ (given in mass per length per time). It is observed experimentally that

$$v - \frac{2}{9}r^2\rho g\mu^{-1}\left(1 - \frac{\rho_\ell}{\rho}\right) = 0$$

Prove that the left side of this law is dimensionally homogeneous.

1.14. Use dimensional analysis and the fact that $\int \frac{dx}{1+x^2} = \arctan(x) + C$ (but don't use u-substitution), to give a general formula for the integral $\int \frac{dx}{a^2+x^2}$. Do the same for $\int \sqrt{1-3x^2}\,dx$. Hint: you may use the fact that $\int \sqrt{1-x^2}\,dx = \frac{1}{2}(\arcsin(x) + x\sqrt{1-x^2}) + C$, but don't use u-substitution.

1.15. Water is poured into a large inverted cone (with an opening angle of $\frac{\pi}{2}$) at a rate of $\frac{dV}{dt} = 10m^3s^{-1}$. Use dimensional analysis to estimate the rate at which the depth is increasing when the water depth is $h = 5m$. Then use calculus to find the exact rate.

1.16. Find the dimensions of the constants a, b, c, d, k in Example 1.3.8 and show that the constant α is dimensionless.

1.17. Assume that $f(t)$ is dimensionless and well defined, and make an appropriate change of variables to nondimensionalize the ODE $a\dot{x} + bx = cf(t)$. Your final answer should have no free parameters despite the fact that the original equation has three.

1.18. The speed v of a wave in deep water is determined by its wavelength λ and the acceleration g due to gravity. What does dimensional analysis imply about the relationship between v, λ, and g?

1.19. A perfect gas in equilibrium has specific energy E (Energy per mass), temperature T and Boltzman constant k (specific energy per degree). Derive a functional relationship of the form $E = f(k, T)$.

1.20. The length L of an organism depends upon how long t it has been alive, its density ρ, its resource assimilation rate a (mass per area per time), and its resource use rate b (mass per volume per time). Show that there is a physical law involving two dimensionless quantities. Give two independent dimensionless quantities π_1 and π_2 expressed in terms of L, t, ρ, a, and b.

1.21. Drag D on a sphere of radius r moving with velocity v in a fluid with density ρ and the fluid viscosity μ depends on all of these variables $D = D(r, v, \rho, \mu)$, which gives a relationship $F(D, r, v, \rho, \mu) = 0$.

(i) Prove that there is a relationship that depends on only two dimensionless variables.

(ii) The quantity $\pi_1 = \frac{\rho v r}{\mu}$ is called the Reynolds number. Prove that π_1 is dimensionless.

(iii) Let $\pi_2 = D\rho^\alpha v^\beta r^\gamma$. Find the values of α, β, and γ that make π_2 dimensionless. This quantity is called the *dimensionless drag force*.

(iv) Show that there is a relation of the form $\pi_2 = G(\pi_2)$.

(v) Experiments show that when the Reynolds number π_1 is large then $G \approx 1$. What does that tell us about D as a function of r, v, ρ, μ?

(vi) Experiments show that when the Reynolds number π_1 is small, then $G \propto \frac{1}{\pi_1}$. What does this tell us about the functional form of $D(r, v, \rho, \mu)$ for low Reynolds number?

Without the Buckingham π theorem we would need four sets of experiments to determine the drag as a function of the four variables, but now we need only one set of experiments using the dimensionless parameters to obtain all the information we need.

1.22. Use a different choice of variables for the nondimensionalization of the first-order Lotka–Volterra system that makes the birth/growth rate of the prey 1; that is, rather than having the parameter b appear in the nondimensional form, you will have a different dimensionless parameter appear in a different location. What does the final system look like?

1.23. Carry out the necessary dimensional analysis to convert (1.18) to (1.19) while guaranteeing that the predator population follows (1.20) in nondimensional form.

Notes

Problems Exercise 1.11–20 are based on exercises from David Logan's book [**?**].

2 Existence and Uniqueness of Solutions

Uniqueness is something my mother pounded into me.
—Sandra Bullock

The primary goal of the mathematical modeler and the mathematical scientist is to produce a function or algorithm that describes the input–output relationship of a given physical, biological, social, or otherwise interesting phenomenon. This process can be a difficult one, especially in situations where there is little data or where the future inputs are not similar enough to past inputs to be able to rely heavily on pattern recognition as the primary means of prediction.[11]

Although it may be difficult or impossible to explicitly write down a function that describes the behavior of a given phenomenon over time, it may be possible to express how the phenomenon changes in space and time as a function of its current state and its corresponding derivatives. For example it is not generally possible to explicitly write down an exact function for the trajectories that a system of planets will take, but it is possible to describe the forces acting on those planets throughout space and time and express them as a system of differential equations that the planets approximately[12] satisfy. We can then use computers to numerically approximate solutions to these differential equations and chart approximate paths for where the planets will be in the future. In other words, by modeling phenomena with differential equations and then approximating the solutions of the differential equations numerically, we can often make good predictions and decisions even though we cannot "solve" the differential equations in closed form.

[11] But it is astounding how far pattern recognition can take you—see Volume 3 for more on this.

[12] Remember that all models are wrong, some are useful. Even in physics there are relativistic effects, quantum effects, and approximate physical constants that always reduce a model from an absolute truth to an approximation of reality.

Of course, in order for any of this to make sense, we need to know that a solution exists. Otherwise, an algorithm might still give a result, but that result would be meaningless, since it is not an approximation of the actual solution. We also need to know under what circumstances a solution is unique. If the model does not have a unique solution, the true state of the system (the actual physical reality we are trying to model) could correspond to one solution, while the trajectory produced by our algorithms might approximate very different solutions, and hence bear no similarity to reality.

In this text we are concerned with modeling, analyzing, computing, and, when possible, solving differential equations. We begin with *ordinary differential equations* (ODEs), which have a single, scalar, independent variable, usually representing time. We are particularly interested in the dynamics of various phenomena, meaning the temporal (time) evolution of the system. Examples include the position and momentum of a particle or object in physics, the concentration of a fluid in chemistry, or the population of a species in a natural habitat.

In this chapter, we provide somewhat general conditions under which we can guarantee the existence and uniqueness of solutions of ODEs. We also prove that solutions depend continuously or smoothly on the initial conditions of the model. When we have existence, uniqueness, and continuous dependence on initial data, we say that the problem is *well posed*.

2.1 Ordinary Differential Equations and the Contraction Mapping Principle

A differential equation, or system of differential equations, is said to be *ordinary* when it has exactly one independent, scalar, real-valued variable. We focus on ordinary[13] differential equations (ODEs) in the first part of this text. We are primarily interested in the case where the independent variable represents time and the model describes the temporal evolution of a given phenomenon. Such mathematical models (or mathematical *systems*) are called *dynamical systems*.

2.1.1 Basic Notions of Ordinary Differential Equations

Definition 2.1.1. *By* ordinary differential equation (ODE) *we mean a relation of the form*

$$x^{(k)} = f(t, x, \dot{x}, \ddot{x}, \dddot{x}, \ldots, x^{(k-1)}), \tag{2.1}$$

where the function $x \in C^k(U;\mathbb{R})$ has k continuous derivatives on an open set $U \subset \mathbb{R}$, the function $f \in C(V;\mathbb{R})$ is continuous on some open set $V \subset \mathbb{R}^{k+1}$. We often assume that the coordinate on U is time, denoted by t, and we write $\dot{x} = x^{(1)} = \frac{d}{dt}x(t)$, and $\ddot{x} = x^{(2)} = \frac{d^2}{dt^2}x(t)$, and so forth. The function f is considered known, and the function $x(t)$ is called the solution *of the differential equation* (2.1)*. The highest derivative k in the equation is the* order *of the differential equation.*

[13] In contrast, partial differential equations (PDEs) have two or more independent variables. We discuss PDEs in the next part of the text.

Example 2.1.2. The function $x = t\tan t$ is a solution of the first-order differential equation

$$\dot{x} = \frac{x + t^2 + x^2}{t}$$

because

$$\begin{aligned}\dot{x} &= \tan t + t\sec^2 t = \tan t + t(1+\tan^2 t)\\ &= \frac{t\tan t + t^2 + t^2\tan^2 t}{t} = \frac{x + t^2 + x^2}{t}\end{aligned}$$

More generally, we want to consider systems of differential equations.

Definition 2.1.3. *A collection of equations of the form*

$$\begin{aligned} x_1^{(k)} &= f_1(t, \mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}}, \dots, \mathbf{x}^{(k-1)})\\ \vdots \quad & \quad \vdots \\ x_n^{(k)} &= f_n(t, \mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}}, \dots, \mathbf{x}^{(k-1)})\end{aligned}$$

or, equivalently, as

$$\mathbf{x}^{(k)} = \mathbf{f}(t, \mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}}, \dots, \mathbf{x}^{(k-1)}) \tag{2.2}$$

with $\mathbf{x} = (x_1, \dots, x_n) \in C^k(U; \mathbb{R}^n)$ *for* $U \subset \mathbb{R}$ *and* $\mathbf{f} = (f_1, \dots, f_n) \in C(V; \mathbb{R}^n)$ *for* $V \subset \mathbb{R}^{1+kn}$*, is called a* system of ODEs of order k*.*

Example 2.1.4. Suppose that $A \in M_n$ is a matrix and $\mathbf{x} \in C^1(U; \mathbb{R}^n)$ is a function of t. The equation

$$\dot{\mathbf{x}} = A\mathbf{x}. \tag{2.3}$$

defines a first-order system of ODEs, that can also be written as

$$\begin{aligned}\dot{x}_1 &= a_{11}x_1 + \cdots + a_{1n}x_n\\ \vdots \quad & \quad \vdots\\ \dot{x}_n &= a_{n1}x_1 + \cdots + a_{nn}x_n.\end{aligned}$$

Remark 2.1.5. There are many examples of differential equations that provably cannot be "solved" in closed form in terms of elementary functions. However, we can often approximate solutions numerically and therefore still make good predictions and decisions, and that is really the goal of every mathematical modeler.

2.1.2 ODEs as First-Order Systems

It's often useful to convert higher-order ODEs to first-order systems of ODEs.

Example 2.1.6 (Second-Order Linear Oscillator). The *second-order linear oscillator* is the differential equation

$$\ddot{x} + p(t)\dot{x} + q(t)x = f(t). \tag{2.4}$$

This represents the most general model for Bob on the spring, with time-varying damping and forcing (compare Section 1.1.4 and Exercise 1.5). The special case (1.7) of Section 1.1.3 where Bob has mass m, there is no forcing $f(t) = 0$, the spring constant is k, and force of friction is $F_f = -c\dot{y}(t)$ for some constant c, corresponds to taking the special case of $p(t) = \frac{c}{m}$ and $q(t) = \frac{k}{m}$.

This model can be recast as a first-order system of equations by setting $\mathbf{y} = (y_1, y_2)$ with

$$y_1 = x \qquad \text{and} \qquad y_2 = \dot{x},$$

which gives

$$\dot{\mathbf{y}} = \begin{bmatrix} 0 & 1 \\ -q(t) & -p(t) \end{bmatrix} \mathbf{y} + \begin{bmatrix} 0 \\ f(t) \end{bmatrix}. \tag{2.5}$$

It turns out that the procedure of Example 2.1.6 works in general. Any k-th order system (2.2) of ODEs can be rewritten as a first-order system by defining the function

$$\mathbf{y} = \begin{bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \\ \vdots \\ \mathbf{x}^{(k-2)} \\ \mathbf{x}^{(k-1)} \end{bmatrix} \quad \text{so that} \quad \dot{\mathbf{y}} = \begin{bmatrix} \dot{\mathbf{x}} \\ \ddot{\mathbf{x}} \\ \vdots \\ \mathbf{x}^{(k-1)} \\ \mathbf{f}(t, \mathbf{y}) \end{bmatrix}$$

If $\mathbf{x}$ has dimension n, then $\mathbf{y}$ has dimension nk.

Example 2.1.7. The fourth-order equation

$$\frac{d^4 z}{dx^4} + \lambda \frac{d^2 z}{dx^2} = 0, \;\; 0 < x < L,$$

models the buckling of a uniform elastic column of length L by an axial load P where λ is proportional to P.

Setting

$$y_1 = z, \quad y_2 = \frac{dz}{dx}, \quad y_3 = \frac{d^2 z}{dx^2}, \quad y_4 = \frac{d^3 z}{dx^3},$$

gives

$$\begin{aligned} \frac{dy_1}{dx} &= \frac{dz}{dx} = y_2, & \frac{dy_2}{dx} &= \frac{d^2 z}{dx^2} = y_3, \\ \frac{dy_3}{dx} &= \frac{d^3 z}{dx^3} = y_4, & \frac{dy_4}{dx} &= \frac{d^4 z}{dx^4} = -\lambda \frac{d^2 z}{dx^2} = -\lambda y_3. \end{aligned}$$

Setting $\mathbf{y} = (y_1, y_2, y_3, y_4)$, the fourth-order equation becomes the system

$$\dot{\mathbf{y}} = A\mathbf{y}, \qquad \text{where} \qquad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\lambda & 0 \end{bmatrix}.$$

This reduction of higher-order systems to first-order systems means that we can focus primarily on solutions and properties of first-order systems of the following form:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t). \tag{2.6}$$

Here $\mathbf{x}$ is a continuously differentiable function on $U \subset \mathbb{R}$ taking values in $X = \mathbb{R}^n$ and $\mathbf{f} : X \times U \to X$ is sufficiently smooth (something we will return to later on).

Remark 2.1.8. In most of what follows, it's no harder to replace $\mathbb{R}^n$ with an arbitrary Banach space[14] $(X, \|\cdot\|)$ in (2.6). For a brief review of Banach spaces, see Section]2.1.5. For full details, see Volume 1, Chapter 5.

Definition 2.1.9. *The system* (2.6) *is called*

(*i*) Autonomous *if* $\mathbf{f}$ *doesn't explicitly depend on time, that is,* $\mathbf{f}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x})$*. In this case* $\mathbf{f}$ *is called a* vector field*. A system that is not autonomous is said to be* nonautonomous*.*

(*ii*) Linear *if* $\mathbf{f}(\mathbf{x}, t) = A(t)\mathbf{x} + \mathbf{b}(t)$*, where* $A(t) \in C(U; \mathscr{B}(X; X))$ *is a continuous function from* U *to the space of bounded linear operators*[15] *on* X *and* $\mathbf{b}(t) \in C(U; X)$ *is a continuous function from* U *to* X*. A system that is not linear is called* nonlinear*. If* $\mathbf{b}(t) \equiv \mathbf{0}$ *is the zero function, then we say that the system is* homogeneous*.*

Note that one can remove time dependence from an ODE to make it autonomous by using essentially the same idea we used to reduce the system to first order.

Example 2.1.10. The forced second-order linear oscillator

$$\ddot{x} + p(t)\dot{x} + q(t)x = f(t) \tag{2.7}$$

can be cast as a system of three first-order (nonlinear) ODEs by setting

$$y_1 = x, \quad y_2 = \dot{x}, \qquad \text{and} \qquad y_3 = t.$$

[14] This is often a way to make the connection between ODEs and PDEs, but we will not focus on that particular approach very much in this text.

[15] See Volume 1 Section 5.7.1. for a review of bounded linear operators. In the case where $X = \mathbb{R}^n$, then $A(t)$ can be written in a given basis as a matrix with entries that are single-valued functions on U.

This gives

$$\begin{aligned}
\dot{y}_1 &= y_2,\\
\dot{y}_2 &= \ddot{x} = -p(t)\dot{x} - q(t)x + f(t) = -p(y_3)y_2 - q(y_3)y_2 + f(y_3),\\
\dot{y}_3 &= 1
\end{aligned}$$

2.1.3 Initial Value Problems

Usually there are infinitely many solutions to a differential equation, and additional information is needed to specify a unique solution. For many of the examples of Chapter 1 the additional information given was an initial value of the solution and its derivatives. Taken together, an ODE and the necessary initial conditions to make the solution unique are called an *initial value problem (IVP)*. Of course, the "initial" point in time t_0 need not be $t = 0$. Often t_0 is considered an initial starting point and we are concerned with the behavior for $t > t_0$, but we may also be interested in the case $t < t_0$, which describes how the system got to its condition at t_0.

Because all ODEs can be reformulated as first-order systems, we give the formal definition of an initial value problem only for first-order systems.

Definition 2.1.11. *An* initial value problem (IVP) *is a first-order system of ODEs of the form 2.6 and an initial value* $\mathbf{x}(t_0) = \mathbf{x}_0$ *for some* $t_0 \in U$*.*

Example 2.1.12. Given $c \in \mathbb{R}$, consider the IVP

$$\dot{x}(t) = c, \qquad \text{and} \qquad x(0) = x_0. \tag{2.8}$$

With a little work, a calculus student should be able to show that $x(t) = x_0 + ct$ is the unique solution.

Example 2.1.13. Given $c \in \mathbb{R}$, consider the IVP

$$\dot{x}(t) = cx(t), \qquad \text{and} \qquad x(0) = x_0. \tag{2.9}$$

With a little more work than the previous example, one can show that $x(t) = x_0e^{ct}$ is the unique solution.

2.1.4 Finite-Time Blow Up

One of the most fundamental questions we can ask of a differential equation is if its solutions are well behaved[16] This leads us to consider one of the most classical examples in all of differential equations.

[16] The notion of *well-behaved* is most certainly not well-defined yet.

Example 2.1.14. Consider the IVP

$$\dot{x} = x^2 \qquad \text{and} \qquad x(t_0) = x_0. \tag{2.10}$$

Separating variables and integrating gives

$$\int_{x_0}^{x(t)} \frac{dx}{x^2} = \int_{t_0}^{t} dt,$$

which implies

$$x(t) = \frac{x_0}{1 - x_0(t - t_0)}.$$

There are two cases:

(i) If $x_0 > 0$ then, as depicted in Figure 2.1, the solution tends to infinity as $t \to t_* = t_0 + \frac{1}{x_0}$, at which point $x(t_*)$ and its derivative are undefined and the IVP is no longer valid. This is an example of what is often called *finite-time blowup.* This IVP and solution are valid only on the interval $(-\infty, t_*)$.

(ii) If $x_0 < 0$ then the solution is well behaved as $t \to \infty$, but going backward in time results in finite-time blowup at $t = t_* < t_0$. This IVP and solution are valid only on the interval (t_*, ∞).

Example 2.1.14 shows that an IVP with a solution on a given interval may not have solutions for all time. In the next section we discuss when solutions exist and give conditions under which the solution is guaranteed to extend to a larger interval. Proving that an IVP has a valid solution for all $t \in (-\infty, \infty)$ is generally difficult problem, even for some of the most simple IVPs.

We are also concerned that our solutions to an IVP are unique, that is, if $x(t)$ and $y(t)$ are both solutions to the same IVP, then we expect $x(t) = y(t)$ for all t (or at least for all t where the solutions make sense). This is true for all of the models discussed in Chapter 1. When the solutions of the IVP are not unique, the model is much less useful for predictive purposes.

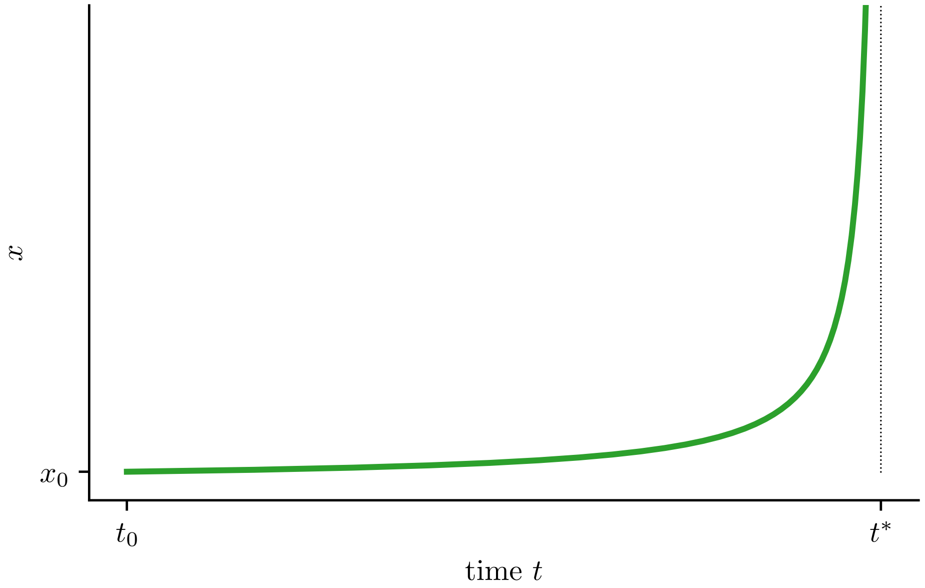

Figure 2.1: The solution of the IVP (2.10) with $x_0 > 0$ goes to ∞ as $t \to t_* = t_0 + \frac{1}{x_0}$. This is an example of finite-time blowup. This solution is not valid for $t > t_*$.

2.1.5 Review of Banach, Lipschitz, and Contraction Mapping

The next section gives conditions that guarantee the existence and uniqueness of solutions to an IVP, but the proof requires some concepts and results from real analysis (Volume 1), which we review here briefly.

Some Banach Spaces

Assume that $X, \|\cdot\|$ is a Banach space, that is, a normed vector space in which every Cauchy sequence converges. The following normed vector spaces are also Banach spaces:

- $X \times \mathbb{R}$ with norm $\|(\boldsymbol{\xi}, t)\| = \|\boldsymbol{\xi}\|_\infty + |t|$. The proof of this is Exercise 2.4.
- The space $L^\infty(\overline{I}; X)$ of bounded linear functions from a compact interval $\overline{I} \subset \mathbb{R}$ to X with norm
$$\|\mathbf{h}\|_\infty = \sup_{t \in \overline{I}} \|\mathbf{h}(t)\|.$$
See Volume 1, Theorem 5.7.5 for a proof.
- The subspace $C(\overline{I}; X) \subset L^\infty(\overline{I}; X)$ with the same norm $\|\cdot\|_\infty$. See Proposition 2.1.15 below.

Proposition 2.1.15. *The subset $C(\overline{I}; X)$ of continuous functions is a vector subspace of $L^\infty(\overline{I}; X)$ and is also a Banach space with the same norm $\|\cdot\|_\infty$.*

Proof. The fact that $C(\overline{I};X)$ is a vector subspace of $L^\infty(\overline{I};X)$ is a routine check. What is less obvious is the fact that any Cauchy sequence in $C(\overline{I};X)$ converges to a function in $C(\overline{I};X)$. To see this, consider a Cauchy sequence $(g_i)_{i\in\mathbb{N}}$ in $C(\overline{I};X)$. Because $L^\infty(\overline{I};X)$ is complete, the sequence $(g_i)_{i\in\mathbb{N}}$ converges to some $g \in L^\infty(\overline{I};X)$.

To see that g is in $C(\overline{I};X)$, is suffices to show it is a continuous function from $\overline{I}$ to X. To that end consider any $t \in \overline{I}$ and a given $\varepsilon > 0$. Since $(g_i)_{i\in\mathbb{N}}$ converges to g, there exists $N > 0$ such that $\|g_i - g\|_\infty < \frac{\varepsilon}{4}$ if $i > N$. Choose one $i > N$. Since g_i is continuous, there exists $\delta > 0$ so that

$$\|g_i(s) - g_i(t)\| < \frac{\varepsilon}{2} \qquad \text{whenever } |s-t| < \delta.$$

This implies that

$$\begin{aligned}\|g(s) - g(t)\| &\leq \|g(s) - g_i(s)\| + \|g_i(s) - g_i(t)\| + \|g(t) - g_i(t)\| \\ &\leq \|g - g_i\|_\infty + \|g_i(s) - g_i(t)\| + \|g - g_i\|_\infty \\ &< \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \varepsilon,\end{aligned}$$

whenever $|s-t| < \delta$. Hence, g is continuous at every $t \in \overline{I}$, as required. □

Corollary 2.1.16. *The set $C(\overline{I};X)$ is a closed subset of $L^\infty(\overline{I};X)$.*

Proof. It suffices to show that every convergent sequence in $C(\overline{I};X)$ converges to a point in $C(\overline{I};X)$. Every convergent sequence in $C(\overline{I};X)$ is Cauchy, and Proposition 2.1.15 guarantees that it converges to a point in $C(\overline{I};X)$. □

Uniformly Lipschitz

The Picard-Lindelöf theorem requires an open set $E \subset X \times \mathbb{R}$ on which the function $\mathbf{f} : E \to X$ is continuous in the second variable t and *uniformly Lipschitz* in the first variable. That means there is a constant $L \geq 0$ such that

$$\|\mathbf{f}(\boldsymbol{\xi}_1, t) - \mathbf{f}(\boldsymbol{\xi}_2, t)\| \leq L\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\| \;\text{ for all }\; (\boldsymbol{\xi}_1, t), (\boldsymbol{\xi}_2, t) \in E. \tag{2.11}$$

We call L the *Lipschitz constant.* Note that the second coordinate t is the same for both points $(\boldsymbol{\xi}_i, t)$ in (2.11), and there is just one Lipschitz constant L that must work no matter what t, $\boldsymbol{\xi}_1$, and $\boldsymbol{\xi}_2$ are.

Example 2.1.17. The function $g : \mathbb{R} \times [-2,2] \to \mathbb{R}$ given by $g(x,t) = \sin(tx)$ is uniformly Lipschitz continuous in the first coordinate with Lipschitz constant 2. To see this, note that by Taylor's Theorem,

$$\sin(t(x+h)) = \sin(tx) + th\cos(\xi),$$

for some $\xi \in [x, x+h]$. Thus

$$|\sin(t(x+h)) - \sin(xt)| = |th\cos(\xi)| \leq |t||h| \leq 2|h|.$$

Note that the Lipschitz constant of 2 holds for all values of $t \in [-2, 2]$, so g is uniformly Lipschitz in the first variable.

If the domain of g is extended to $\mathbb{R} \times \mathbb{R}$ rather than $\mathbb{R} \times [-2, 2]$, then g is not uniformly Lipschitz in the first variable. To see this, consider any $L > 0$. For all $y \in \left(0, \frac{\pi}{3}\right)$, a straightforward check shows that $\sin(y) > \frac{y}{2}$. This implies that for $x = 0$, for $t > 2L$, and for $h \in (0, \frac{\pi}{3t})$ we have

$$|g(t(0+h)) - g(0)| = |\sin(th)| > \frac{|t|}{2}|h| > L|h|.$$

Therefore, no L satisfies the requirements to be is a uniform Lipschitz constant for the function g.

Proposition 2.1.18. *For an open set $E \subset X \times \mathbb{R}$, a sufficient condition for $\mathbf{f} : E \to X$ to be uniformly Lipschitz in the first variable is if $\overline{E}$ is compact and $\mathbf{f}$ is C^1 on an open set containing $\overline{E}$.*

Proof. Using the norm $\|(\mathbf{x}, t)\| = \|\mathbf{x}\| + |t|$ on the product $X \times \mathbb{R}$, it follows from the Mean Value Theorem that

$$\begin{aligned}\|\mathbf{f}(\mathbf{x}_1, t) - \mathbf{f}(\mathbf{x}_2, t)\| &\leq \sup_{(\mathbf{x},s)\in\overline{E}} \|D\mathbf{f}(\mathbf{x}, s)\| \, \|(\mathbf{x}_1, t) - (\mathbf{x}_2, t)\| \\ &= \sup_{(\mathbf{x},s)\in\overline{E}} \|D\mathbf{f}(\mathbf{x}, s)\| \, \|(\mathbf{x}_1 - \mathbf{x}_2, 0)\| \\ &= \sup_{(\mathbf{x},s)\in\overline{E}} \|D\mathbf{f}(\mathbf{x}, s)\| \, \big(\|\mathbf{x}_1 - \mathbf{x}_2\| + |0|\big) \\ &= \sup_{(\mathbf{x},s)\in\overline{E}} \|D\mathbf{f}(\mathbf{x}, s)\| \, \|\mathbf{x}_1 - \mathbf{x}_2\|.\end{aligned}$$

Setting

$$L = \sup_{(\mathbf{x},s)\in\overline{E}} \|D\mathbf{f}(\mathbf{x}, s)\|$$

gives the uniform Lipschitz property in the first variable:

$$\|\mathbf{f}(\mathbf{x}_1, t) - \mathbf{f}(\mathbf{x}_2, t)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\| \quad \text{for all} \quad (\mathbf{x}_1, t), (\mathbf{x}_2, t) \in E. \qquad \square$$

Uniform Contraction Mapping

Recall (Volume 1, Definition 7.1.3) that a *contraction mapping* in a normed space is a function $\phi : D \to D$ for which there exists $0 \leq k < 1$ such that

$$\|\phi(\mathbf{x}) - \phi(\mathbf{y})\| \leq k\|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in D. \tag{2.12}$$

This is important because of the *Contraction Mapping Principle* (Volume 1, Theorem 7.1.8) which states that when D is a closed subset of a Banach space, then every contraction mapping $\phi : D \to D$ has a unique fixed point $\mathbf{x} \in D$, that is, $\phi(\mathbf{x}) = \mathbf{x}$.

We use this in the next section to construct a solution to initial value problems by constructing a particular contraction mapping $\phi : C(\overline{I}; X) \to C(\overline{I}; X)$ on the Banach space $C(\overline{I}; X)$ of continuous functions from a closed interval $\overline{I}$ to another Banach space X. The fixed point $\mathbf{x}$ will be a function that satisfies $\phi(\mathbf{x}) = \mathbf{x}$, and thus if ϕ is chosen appropriately, this shows $\mathbf{x}$ is the desired solution.

Remark 2.1.19. If f is a Lipschitz continuous map from a space to itself and has Lipschitz constant $0 \leq L < 1$, then f is a contraction mapping.

We need a stronger form of the Contraction Mapping Principle, namely the *Uniform* Contraction Mapping Principle (Volume 1, Theorem 7.2.4), which applies to a family of contraction mappings, one for each $b \in B$, meaning a function $\Phi : D \times B \to D$, where for each $b \in B$ the map $\Phi(\cdot, b) : D \to D$ is a contraction mapping. The Uniform Contraction Mapping Principle says, first, that if there is one k that works in (2.12) for all $b \in B$, then every $b \in B$ has a corresponding unique fixed point in D, so that defines a function $g : B \to D$; and second, if Φ is C^ℓ (that is, continuously differentiable $\ell > 0$ times), then the map $g : B \to D$ is also C^ℓ.

2.2 Existence and Uniqueness

Let $(X, \|\cdot\|)$ be a Banach space. Consider the IVP

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) \tag{2.13a}$$

$$\mathbf{x}(t_0) = \mathbf{x}_0. \tag{2.13b}$$

In this section, we prove the existence and uniqueness of solutions to this IVP. First we prove that there is a unique function $\mathbf{x}(t)$ that satisfies (2.13) in a neighborhood of t_0. This is considered a "local" existence result because the theorem only gives a neighborhood for existence, not existence on a fixed interval. Then we prove existence and uniqueness over larger intervals by gluing local solutions together.

2.2.1 Local Existence and Uniqueness Overview

Local existence and uniqueness of solutions of the IVP (2.13) is a result of the *Picard–Lindelöf* Theorem (see Theorem 2.2.2). This theorem is also often called the *Cauchy–Lipschitz* Theorem.

The first step in the existence and uniqueness proof is the following crucial observation: integrating both sides of (2.13a) from t_0 to t gives

$$\int_{t_0}^{t} \dot{\mathbf{x}}(t)\, dt = \int_{t_0}^{t} f(\mathbf{x}(s), s)\, ds,$$

or equivalently

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} f(\mathbf{x}(s), s)\, ds. \tag{2.14}$$

Note that $\mathbf{x}(t)$ appears on both sides. This suggests that we should be looking for some kind of fixed-point theorem (hence the review of contraction mappings in the previous section).

Remark 2.2.1. The integral in (2.14) makes sense even when X is an infinite-dimensional Banach space. For details on the definition of the integral in that case see Volume 1, Section 5.10.2.

The idea for both existence and uniqueness is to form an operator $\phi : C(\overline{I}; X) \to C(\overline{I}; X)$ defined by

$$\phi[\mathbf{h}](t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\mathbf{h}(s), s)\, ds$$

for a suitably chosen compact interval $\overline{I}$ containing t_0 that makes ϕ a contraction mapping. Thus ϕ has a unique fixed point $\mathbf{x} \in C(\overline{I}; X)$. But $\phi[\mathbf{x}] = \mathbf{x}$ means that (2.14) holds. Substituting $t = t_0$ into (2.14) gives $\mathbf{x}(t_0) = \mathbf{x}_0$, and the Fundamental Theorem of Calculus implies that (2.13a) holds for all t in the interior of $\overline{I}$ (the derivative of $\mathbf{x}$ might not be defined at the endpoints). Thus the fixed point $\mathbf{x}$ is a solution of the IVP (2.13) on the interior I. The solution is unique, because any solution of the IVP must satisfy (2.14), and hence must be a fixed point of the contraction mapping ϕ.

The function ϕ depends also on $\mathbf{x}_0 \in X$, so it really is a function $\phi : C(\overline{I}; X) \times X \to C(\overline{I}; X)$, which turns out to be a *uniform* contraction mapping. It's not hard to show that ϕ is a continuous function on $C(\overline{I}; X) \times X$, so the Uniform Contraction Mapping Principle (Volume 1, Theorem 7.2.4) guarantees that the solution $\mathbf{x}$ of the IVP(2.13) depends continuously on the choice of $\mathbf{x}_0$. That is, a small change in $\mathbf{x}_0$ makes a small change in the function $\mathbf{x} \in C(\overline{I}; X)$.

2.2.2 Picard-Lindelöf/Cauchy-Lipschitz

Now that the main idea of the proof has been sketched, we are now ready to carefully state and prove the main theorem of this section.

Theorem 2.2.2 (Picard–Lindelöf/Cauchy–Lipschitz). *For a Banach space $(X, \|\cdot\|)$ and an open subset E of $X \times \mathbb{R}$, suppose $\mathbf{f} : E \to X$ is continuous on E and uniformly Lipschitz in the first variable, with Lipschitz constant $L > 0$. If $(\mathbf{x}_0, t_0) \in E$, then there exists an open interval $I \subset \mathbb{R}$ containing t_0 such that the IVP (2.13) has a unique solution $\mathbf{x}(t)$ for $t \in I$, and $(\mathbf{x}(t), t) \in E$ for all $t \in I$. Moreover, the function $\mathbf{x}$ depends continuously on $\mathbf{x}_0$.*

Proof. The first step is to select the open interval $I = (t_0 - \delta, t_0 + \delta)$, by choosing an appropriate $\delta > 0$. To do this, first note that since E is an open subset of $X \times \mathbb{R}$, there exists $r > 0$ so that

$$\overline{B((\mathbf{x}_0, t_0), r)} \subset E.$$

Let $U = B((\mathbf{x}_0, t_0), r) \subset E \subset X \times \mathbb{R}$. Also because $\mathbf{f}$ is continuous on E, it is bounded on the compact set $\overline{B((\mathbf{x}_0, t_0), r)}$ by

$$M = \sup\{\|\mathbf{f}(\mathbf{x}, t)\| : (\mathbf{x}, t) \in \overline{B(\mathbf{x}_0, t_0), r)}\} < \infty. \tag{2.15}$$

Choose δ to be

$$\delta = \min\left\{r, \frac{r}{M}, \frac{1}{2L}\right\}. \tag{2.16}$$

If $L = 0$, then $\frac{1}{2L} = \infty$ does not contribute to the choice of δ. Note that $\overline{I} = [t_0 - \delta, t_0 + \delta]$ is compact.

In what follows we understand the initial condition $\mathbf{x}_0$ to be the constant function in $C(\overline{I}; X)$ given by $\mathbf{x}(t) = \mathbf{x}_0$ for all $t \in \overline{I}$. Consider the closed set $F = \overline{B(\mathbf{x}_0, r)}$ containing the constant function $\mathbf{x}_0$ in the Banach space $C(\overline{I}; X)$.[17]

We show that ϕ is a contraction mapping on the closed set F, but the first requirement for that to hold is that ϕ must map F to F. This is the case because any $\mathbf{h} \in F$ satisfies

$$\begin{aligned}
\|\phi[\mathbf{h}] - \mathbf{x}_0\|_\infty &= \sup_{t \in \overline{I}} \left\| \int_{t_0}^{t} \mathbf{f}(\mathbf{h}(s), s)\, ds \right\| \\
&\leq \sup_{t \in \overline{I}} \left| \int_{t_0}^{t} \|\mathbf{f}(\mathbf{h}(s), s)\|\, ds \right| \\
&\leq M \sup_{t \in \overline{I}} \left| \int_{t_0}^{t} ds \right| = M\delta \leq r.
\end{aligned}$$

Here the second line follows from the triangle inequality, and the third line follows from the definition of M in (2.15) and of δ in (2.16).

The map ϕ is a contraction mapping on F with constant $\frac{1}{2}$ because for any $\mathbf{g}, \mathbf{h} \in F$ we have

$$\begin{aligned}
\|\phi[\mathbf{h}] - \phi[\mathbf{g}]\|_\infty &= \sup_{t \in \overline{I}} \left\| \int_{t_0}^{t} \big[\mathbf{f}(\mathbf{h}(s), s) - \mathbf{f}(\mathbf{g}(s), s)\big]\, ds \right\| \\
&\leq \sup_{t \in \overline{I}} \left| \int_{t_0}^{t} \|\mathbf{f}(\mathbf{h}(s), s) - \mathbf{f}(\mathbf{g}(s), s)\|\, ds \right| \\
&\leq \sup_{t \in \overline{I}} \left| \int_{t_0}^{t} L\|\mathbf{h}(s) - \mathbf{g}(s)\|\, ds \right| \\
&\leq \sup_{t \in \overline{I}} \left| \int_{t_0}^{t} L\|\mathbf{h} - \mathbf{g}\|_\infty\, ds \right| \\
&= \sup_{t \in \overline{I}} L\|\mathbf{h} - \mathbf{g}\|_\infty \left| \int_{t_0}^{t} ds \right| \\
&\leq L\|\mathbf{h} - \mathbf{g}\|_\infty \delta \leq \frac{1}{2}\|\mathbf{h} - \mathbf{g}\|_\infty.
\end{aligned} \tag{2.17}$$

The Contraction Mapping Principle guarantees there exists a unique fixed point $\mathbf{x} \in F$ satisfying $\phi[\mathbf{x}] = \mathbf{x}$, which is equivalent to (2.14) for all $t \in \overline{I}$. By the Fundamental Theorem of Calculus, the function $\mathbf{x}$ is differentiable on the interior I of $\overline{I}$ and satisfies the differential equation (2.13a). Moreover, $t_0 \in I$, and setting $t = t_0$ in (2.14) shows that x satisfies (2.13b), so there is unique function $\mathbf{x}(t)$ satisfying the IVP (2.13) on the open interval $I = (t_0 - \delta, t_0 + \delta)$.

[17] The r used here is the same r that was chosen to make a neighborhood $B((\mathbf{x}_0, t_0), r)$ of the *point* $(\mathbf{x}_0, t_0)$ in $X \times \mathbb{R}$, but here it is used to construct a neighborhood $B(\mathbf{x}_0, r)$ of the constant *function* $\mathbf{x}_0$ in the function space $C(\overline{I}; X)$.

Finally, the mapping ϕ depends on $\mathbf{x}_0$, so it defines a map $\Phi : C(\overline{I}; X) \times X \to C(\overline{I}; X)$. The proof that ϕ is a contraction mapping shows that the contraction inequality $\|\phi[\mathbf{g}] - \phi[\mathbf{h}]\| \leq \frac{1}{2}\|\mathbf{g} - \mathbf{h}\|$ is independent of $\mathbf{x}_0$, so Φ is a uniform contraction mapping. Moreover, giving $C(\overline{I}; X) \times X$ the metric $\|(\mathbf{h}, \boldsymbol{\xi})\| = \|\mathbf{h}\|_\infty + \|\boldsymbol{\xi}\|$ makes $C(\overline{I}; X) \times X$ into a Banach space, and Φ is Lipschitz continuous because for all $(\mathbf{h}, \boldsymbol{\xi}), (\mathbf{g}, \boldsymbol{\chi}) \in C(\overline{I}; X) \times X$ we have

$$\begin{aligned}\|\Phi(\mathbf{h}, \boldsymbol{\xi}) - \Phi(\mathbf{g}, \boldsymbol{\chi})\| &\leq \|\boldsymbol{\xi} - \boldsymbol{\chi}\| + \sup_{t \in \overline{I}} \| \int_{t_0}^{t} \mathbf{f}(\mathbf{h}(s), s) - \mathbf{f}(\mathbf{g}(s), s)\, ds\| \\ &\leq \|\boldsymbol{\xi} - \boldsymbol{\chi}\| + \frac{1}{2}\|\mathbf{h} - \mathbf{g}\| \\ &\leq \|(\mathbf{h}, \boldsymbol{\xi}) - (\mathbf{g}, \boldsymbol{\chi})\|.\end{aligned}$$

Therefore, by the Uniform Contraction Mapping Principle, the unique fixed point $\mathbf{x}$ is a continuous function of $\mathbf{x}_0$. □

Remark 2.2.3. The proof of the theorem above shows that the unique solution $\mathbf{x}$ of the IVP is always differentiable on I. But since $\mathbf{f}(\mathbf{x}, t)$ is continuous, then $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), t)$ is a continuous function of t and so $\mathbf{x}$ is C^1. If we assume $\mathbf{f}(\mathbf{x}, t)$ is C^1, then differentiating $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t)$ with respect to t on I gives

$$\ddot{\mathbf{x}}(t) = D_{\mathbf{x}}\mathbf{f}(\mathbf{x}(t), t)\dot{\mathbf{x}}(t) + D_t\mathbf{f}(\mathbf{x}(t), t).$$

Since the right-hand side is continuous, so is $\ddot{\mathbf{x}}$, and thus $\mathbf{x}$ is C^2. A similar argument shows that whenever $\mathbf{f}(\mathbf{x}, t)$ is C^k, then $\mathbf{x}$ is C^{k+1}.

Unexample 2.2.4. Both $x(t) \equiv 0$ and $x(t) = t^4$ are solutions of the IVP

$$\begin{aligned}\dot{x}(t) &= 4|x(t)|^{3/4} \\ x(0) &= 0.\end{aligned} \tag{2.18}$$

Thus this IVP does not have a unique solution. This does not violate the Picard-Lindelöf theorem because $f(x) = |x|^{3/4}$ is not Lipschitz continuous on any open set containing $x = 0$.

Example 2.2.5. Consider

$$\begin{aligned}\dot{x}(t) &= t^2 x(t) \\ x(0) &= 1\end{aligned}$$

where $t \in (-3, 3)$ and $x \in (-2, 2)$.

The function $f(x,t)$ is continuous and bounded by $M = 18$ on $[-2,2] \times [-3,3]$, and $|t^2x_1 - t^2x_2| \leq 9|x_1 - x_2|$, so $f(x,t)$ is Lipschitz continuous with Lipschitz constant $K = 9$ on $[-2,2]$ for all $t \in [-3,3]$. We also note that any ball with radius $r = 1$ in $\mathbb{R}^2$ centered about the initial conditions $(x_0, t_0) = (1,0)$ will remain in the indicated domain. Thus, by Theorem 2.2.2, there is a unique solution for $t \in [-\delta, \delta]$ where $\delta = \min\left(r, \frac{r}{M}, \frac{1}{2L}\right) = \min(1, 1/18, 1/18) = \frac{1}{18}$.

It is straightforward to check that $x(t) = e^{t^3/3}$ is a solution to the IVP, so it must be the unique solution. Note that $x(t) = e^{t^3/3}$ is actually defined and is a solution for all $t \in \mathbb{R}$, even thought the theorem only guarantees a solution locally.

Remark 2.2.6. The last part of the theorem, that the solution $\mathbf{x}$ depends continuously on $\mathbf{x}_0$, means that a small change in $\mathbf{x}_0$ produces only a small change in the function $\mathbf{x}(t)$. With a little more work, one can show that when $\mathbf{f}$ is C^k, then the map Φ is also C^k, and so the Universal Contraction Mapping Principle guarantees that the resulting function $\mathbf{x}_0 \mapsto \mathbf{x}(t)$ mapping initial values to solutions of the IVP is a C^k function of the initial value $\mathbf{x}_0$.

2.2.3 Continuation to Larger Intervals

The existence and uniqueness result of in Theorem 2.2.2 is a local result that is only guaranteed on a small open set. But, to be useful we need to be able to understand the long-term behavior of ODEs, so that we can make predictions. It is not enough to know what happens at time $t_0 + \delta$ for some small δ, we are interested in what happens as $t \to \infty$.

In this section we prove the Unique Extension Theorem, which guarantees that any solution of the IVP(2.13) on any interval containing t_0 must agree with any other solution on that interval. That doesn't mean that the unique solution can be extended or "continued" beyond the original interval $(t_0 - \delta, t_0 + \delta)$, but if it can be extended, that extension is unique. This means it makes sense to talk about a largest interval containing t_0 on which the IVP has a solution, and that solution is unique.

Theorem 2.2.7 (Unique Extension). *If the unique solution of the IVP 2.13 on the interval I given by Theorem 2.2.2 extends to a larger open interval, it extends uniquely.*

Proof. Suppose that there are two solutions $\mathbf{x} = \phi(t)$ and $\mathbf{x} = \psi(t)$ of the IVP (2.13) on an open interval (r,s) containing the interval I given by the Picard-Lindelöf Theorem (one or both of the endpoints r and s could be infinite). By uniqueness of the solution on I, the two functions must agree on I, that is, $\phi(t) = \psi(t)$ for all $t \in I$.

Consider the union J of all open intervals J_α containing I on which $\phi(t) = \psi(t)$ for all $t \in J_\alpha$. The set J is an open interval because any union of open sets is open, and these all contain I, so their union is connected, hence it is an interval (a,b) (one or both of the endpoints could be infinite).

If $(r,s) \not\subset J$, then either $r < a$ or $b < s$. In either case, there is a point $t_1 \in (r,s)$ on the boundary of J. Since ϕ and ψ are both differentiable, they are continuous, and they agree on all of J, so they must also agree at t_1 (the set of points where two continuous functions agree is always closed).

Consider a new IVP

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, t), \\ \mathbf{x}(t_1) &= \phi(t_1), \end{aligned} . \tag{2.19}$$

By the uniqueness from the Picard-Lindelöf theorem, there is an open interval U containing t_1 where $\phi(t) = \psi(t)$, for all $t \in U$. This shows that $J \subset J \cup U$ is an open interval containing I and t_1 on which $\phi = \psi$, but $t_1 \notin J$, contradicting the definition of J. □

Definition 2.2.8. *We say the IVP* $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$, $\mathbf{x}(t_0) = \mathbf{x}_0$, *with* $\mathbf{x}$ *taking values in an open set* $U \subset \mathbb{R}^n$, *is* well posed *if*

(a) $\mathbf{x}(t)$ *exists for arbitrary* $\mathbf{x}_0 \in U$

(b) $\mathbf{x}(t)$ *is determined uniquely by* $\mathbf{x}_0$, *and*

(c) $\mathbf{x}(t)$ *depends continuously on* $\mathbf{x}_0$.

The Picard-Lindelöf Theorem combined with the Unique Extension theorem guarantee that when $\mathbf{f}$ is continuous and uniformly Lipschitz in the first variable, then the corresponding IVP is well posed.

2.2.4 Finite-Time Blowup

The Unique Extension Theorem shows there is a maximal open interval of existence for each IVP. Finding this maximal open interval of existence is not necessarily easy for nonlinear systems, and, as we have seen with the finite-time blowup, it may depend on the initial conditions.

Example 2.1.14 shows an IVP with finite-time blow up, where a solution has a vertical asymptote and yet the function $f(x,t) = x^2$ is definitely well behaved, that is, Lipschitz on any compact domain. So, how does this solution blow up and cease to exist in finite time without contradicting the Picard-Lindelf̈ Theorem?

The answer lies in the fact that at any point along the trajectory there is a neighborhood for which the solution continues to exist for a little longer, but the size of that neighborhood (the size of δ) gets smaller and smaller as we get closer to the point of blow up. This is why it is so important to understand that the solutions guaranteed to exist are local solutions and that Theorem 2.2.2 in no way suggests that a solution will exist for all time.

But the following result gives some information on how a solution must behave topologically if the maximal open interval of existence for the unique solution of an IVP is not $(-\infty, \infty)$.

Theorem 2.2.9 (Finite Time Blowup). *Suppose U is an open subset of $\mathbb{R}^n$ and I is an interval of $\mathbb{R}$. If $\mathbf{f} : U \times I \to \mathbb{R}^n$ is a C^1 function and if $(\alpha, \beta) \subset I$ is the maximal open interval of existence of the solution of the IVP (2.13) for $t_0 \in (\alpha, \beta)$ and $\mathbf{x}_0 \in U$, with $\beta < \infty$, then for each compact subset $K \subset U$ there exists $t \in (t_0, \beta)$ such that $\mathbf{x}(t) \notin K$. Similarly, if $\alpha > -\infty$, then for every compact $K \subset U$, there is a $t \in (\alpha, t_0)$ such that $\mathbf{x}(t) \notin K$.*

2.2.5 Picard Iteration

The nice thing about using the contraction mapping principle is that it gives an algorithm, called *Picard iteration* (sometimes also called the *Method of Successive Approximation*), to approximate the unique solution. Recall that the proof of the contraction mapping principle finds the unique fixed point by starting at any point $\mathbf{q}_0$ and repeatedly applying the contraction mapping ϕ to get a Cauchy sequence $\mathbf{q}_0$, $\mathbf{q}_1 = \phi(q_0)$, $\mathbf{q}_2 = \phi(q_1)$, $\ldots$. This Cauchy sequence must converge because it lies in a Banach space; and the point it converges to is the unique fixed point. In the case of an IVP, a natural place to start the iteration is with the constant function $\mathbf{q}_0(t) = \mathbf{x}_0$. This gives the sequence

$$\begin{aligned}
\mathbf{q}_0(t) &= \mathbf{x}_0 \\
\mathbf{q}_1(t) &= \mathbf{x}_0 + \int_{t_0}^{t} f(\mathbf{q}_0, s)ds \\
\mathbf{q}_2(t) &= \mathbf{x}_0 + \int_{t_0}^{t} f(\mathbf{q}_1(s), s)ds \\
&\vdots \\
\mathbf{q}_{n+1}(t) &= \mathbf{x}_0 + \int_{t_0}^{t} f(\mathbf{q}_n(s), s)ds,
\end{aligned}$$

which must converge to the correct answer on a small neighborhood of t_0.

Example 2.2.10. Consider the ODE

$$\begin{aligned}
\dot{x}(t) &= 2t(1 + x(t)), \\
x(0) &= 0.
\end{aligned} \tag{2.20}$$

Following Picard iteration, let

$$q_1(t) = \int_0^t 2s\, ds = t^2$$

and

$$q_2(t) = \int_0^t 2s(1 + s^2)\, ds = t^2 + \frac{1}{2}t^4.$$

Inductively, if we assume that

$$q_n(t) = \sum_{k=1}^{n} \frac{t^{2k}}{k!},$$

then we can see that

$$q_{n+1}(t) = \int_0^t 2s(1 + x_n(s))\, ds = \int_0^t \sum_{k=0}^{n-1} \frac{s^{2k}}{k!}\, ds = \sum_{k=0}^{n} \frac{s^{2(k+1)}}{(k+1)!} = \sum_{k=1}^{n+1} \frac{t^{2k}}{k!}.$$

Taking the limit, we get that $q(t) = e^{t^2} - 1$, which is the solution to (2.20).

2.2.6 Gronwall's Inequality

Gronwall's inequality gives an alternative way to prove the continuous dependence of $\mathbf{x}$ on $\mathbf{x}_0$, and it gives more detailed information about how much $\mathbf{x}$ can change as a function of $\mathbf{x}_0$.

Lemma 2.2.11 (Gronwall's Inequality). *For continuous functions $\phi(t) \geq 0$ and $u(t) \geq 0$ with domain $[t_0, T]$, if*

$$u(t) \leq u(t_0) + \int_{t_0}^{t} \phi(s)u(s)\, ds \quad \text{for all } t \in [t_0, T],$$

then

$$u(t) \leq u(t_0) \exp\left\{\int_{t_0}^{t} \phi(s)\, ds\right\} \quad \text{for all } t \in [t_0, T].$$

The proof of this inequality is somewhat involved and is left to the following section.

As a consequence of Gronwall's Inequality we get an alternative proof of the fact that solutions of IVPs depend continuously on the choice of initial value.

Proposition 2.2.12. *Suppose $\mathbf{f}(\mathbf{x}, t)$, for $(\mathbf{x}, t) \in D = U \times I \subset \mathbb{R}^n \times \mathbb{R}$, is uniformly Lipschitz in the first variable with Lipschitz constant L. The unique solutions of*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t),\ \mathbf{x}(t_0) = \mathbf{x}_0,\ (\mathbf{x}_0, t) \in D,$$
$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t),\ \mathbf{y}(t_0) = \mathbf{y}_0,\ (\mathbf{y}_0, t) \in D,$$

satisfy the inequality

$$\|\mathbf{x}(t) - \mathbf{y}(t)\| \leq \|\mathbf{x}_0 - \mathbf{y}_0\| \exp\{L(t - t_0)\}, \qquad \forall t \in I.$$

Proof. The unique solutions satisfy the integral equations

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(s), s)\, ds \quad \text{and} \quad \mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(\mathbf{y}(s), s)\, ds.$$

Thus for $t \geq t_0$ and $t \in I$ (similar argument for $t \leq t_0$ and $t \in I$), we have

$$\begin{aligned}\|\mathbf{x}(t)-\mathbf{y}(t)\| &= \left\|\mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(s),s)\,ds - \mathbf{y}_0 - \int_{t_0}^{t} \mathbf{f}(\mathbf{y}(s),s)\,ds\right\| \\ &\leq \|\mathbf{x}_0-\mathbf{y}_0\| + \left\|\int_{t_0}^{t} \mathbf{f}(\mathbf{x}(s),s)\,ds - \int_{t_0}^{t} \mathbf{f}(\mathbf{y}(s),s)\,ds\right\| \\ &\leq \|\mathbf{x}_0-\mathbf{y}_0\| + \int_{t_0}^{t} \|\mathbf{f}(\mathbf{x}(s),s)-\mathbf{f}(\mathbf{y}(s),s)\|\,ds \\ &\leq \|\mathbf{x}_0-\mathbf{y}_0\| + \int_{t_0}^{t} L\|\mathbf{x}(s)-\mathbf{y}(s)\|.\end{aligned}$$

With

$$u(t) = \|\mathbf{x}(t)-\mathbf{y}(t)\| \quad \text{and } \phi(t) = L,$$

Gronwall's inequality gives

$$\|\mathbf{x}(t)-\mathbf{y}(t)\| \leq \|\mathbf{x}_0-\mathbf{y}_0\| \exp\{L(t-t_0)\}.$$

This implies continuous dependence of the solutions on the initial conditions. □

2.2.7 *Proof of Gronwall's Inequality

Proof. Suppose that $u(t_0) > 0$ (we deal with the case of $u(t_0) = 0$ later). Set

$$x(t) = u(t_0) + \int_{t_0}^{t} \phi(s)u(s)\,ds.$$

Then $u(t) \leq x(t)$ for all $t \in [t_0, T]$ and

$$\dot{x}(t) = \phi(t)u(t) \quad \text{and} \quad x(t_0) = u(t_0).$$

It follows that

$$\frac{d}{dt}\ln x(t) = \frac{\dot{x}(t)}{x(t)} = \frac{\phi(t)u(t)}{x(t)} \leq \phi(t) \quad \text{for} \quad \text{all } t \in [t_0, T].$$

Integrating this inequality over the interval $[t_0, t]$ for $t \in [t_0, T]$ gives

$$\ln x(t) - \ln x(t_0) \leq \int_{t_0}^{t} \phi(s)\,ds.$$

Rearranging this gives

$$\ln x(t) \leq \ln x(t_0) + \int_{t_0}^{t} \phi(s)\,ds.$$

Exponentiating both sides gives

$$x(t) \leq \exp\left\{\ln x(t_0) + \int_{t_0}^{t} \phi(s)\,ds\right\} = x(t_0)\exp\left\{\int_{t_0}^{t} \phi(s)\,ds\right\}.$$

Since $u(t) \leq x(t)$ and $x(t_0) = u(t_0)$ we obtain

$$u(t) \leq x(t) = x(t_0) \exp\left\{\int_{t_0}^{t} \phi(s)\,ds\right\} = u(t_0) \exp\left\{\int_{t_0}^{t} \phi(s)\,ds\right\}.$$

Now supposing $u(t_0) = 0$ we show for each $\varepsilon > 0$ that the function $u(t) + \varepsilon$ satisfies

$$u(t) + \varepsilon \leq u(t_0) + \varepsilon + \int_{t_0}^{t} \phi(s)[u(s) + \varepsilon]\,ds,\ t \in [t_0, T].$$

By hypothesis the function $u(t)$ satisfies

$$u(t) \leq u(t_0) + \int_{t_0}^{t} \phi(s)u(s)\,ds,\ t \in [t_0, T].$$

Adding ε to both sides gives

$$u(t) + \varepsilon \leq u(t_0) + \varepsilon + \int_{t_0}^{t} \phi(s)u(s)\,ds,\ t \in [t_0, T].$$

Since $\phi(s)u(s) \leq \phi(s)[u(s) + \varepsilon]$ for all $s \in [t_0, T]$ it follows that

$$u(t) + \varepsilon \leq u(t_0) + \varepsilon + \int_{t_0}^{t} \phi(s)[u(s) + \varepsilon]\,ds,\ t \in [t_0, T].$$

With $u(t_0) + \varepsilon > 0$ we apply the case of when $u(t_0) > 0$ from above to obtain

$$u(t) + \varepsilon \leq (u(t_0) + \varepsilon) \exp\left\{\int_{t_0}^{t} \phi(s)\,ds\right\},\ t \in [t_0, T].$$

Taking the limit in this as $\varepsilon \to 0^+$ gives the result. □

Remark 2.2.13. Note that using Gronwall's inequality does not give a very good bound on the difference between solutions, particularly if L is very large. It does, however, guarantee that as long as $\mathbf{f}(\mathbf{x}, t)$ is uniformly Lipschitz in $\mathbf{x}$, then solutions must depend at least continuously on the initial data.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

2.1. Write $\dddot{y}(t) - 3\ddot{y}(t) + 5\cos(t)\sin(y(t)) = 0$ as a first-order system.

2.2. Write the ODE in the previous exercise as an autonomous first-order system.

2.3. Solve the IVP

$$\begin{aligned}\dot{x}(t) &= x^4(t)\\ x(t_0) &= x_0\end{aligned}$$

and determine on what interval the solution is valid, that is, where does blowup occur? Hint: Consider using separation of variables, as in Example 2.1.14.

2.4. Prove that if $X, \|\cdot\|$ is a Banach space, then $X \times \mathbb{R}$ with norm $\|(\boldsymbol{\xi}, t)\| = \|\boldsymbol{\xi}\|_\infty + |t|$ is also a Banach space.

2.5. Prove that $f(x) = x^a$ is not Lipschitz continuous on $[0, 1]$ for any $0 < a < 1$.

2.6. Prove that the function $f(x) = \frac{1}{2}\left(x + \frac{3}{x}\right)$ is a contraction mapping with constant $\frac{1}{2}$ on the interval $[\frac{3}{2}, \infty)$, and its unique fixed point is $x = \sqrt{3}$. Hint: To show that $f : [\frac{3}{2}, \infty) \to [\frac{3}{2}, \infty)$ you need only show that $f\left(\frac{3}{2} + a\right) \geq \frac{3}{2}$ when $a \geq 0$. To show the contraction mapping property, it may be helpful first to show that $|f(x) - f(y)|$ can be rewritten as $\frac{1}{2}|x - y|\left|1 - \frac{3}{xy}\right|$.

2.7. Find a solution to the IVP

$$\begin{aligned}\dot{y}(t) &= ty(t)\\ y(0) &= 2\end{aligned}$$

and explain why it is the only solution in any neighborhood of $t = 0$.

2.8. For each of the following ODEs with given initial values, reformulate it as an IVP in the form (2.13) and decide whether it meets the conditions of Theorem 2.2.2. Prove your answers are correct. Hint: Proposition 2.1.18 may be useful.

(i) The SIR model (1.9) for known constants b and k, with initial conditions $S(0) = S_0$, $I(0) = I_0$, and $R(0) = R_0$.

(ii) The logistic equation (1.11) with fixed real constants $k > 0$ and $r \in \mathbb{R}$ and initial value $x(0) = x_0$.

(iii) The Lotka–Volterra Predator–prey model (1.17) with fixed constant b and initial values $x(0) = x_0$ and $y(0) = y_0$.

(iv) The forced linear oscillator (2.7) with initial conditions $x(0) = x_0$ and $\dot{x}(0) = v_0$. What are the weakest conditions that can be placed on the functions $p(t)$, $q(t)$, and $f(t)$ to ensure that the conditions of Theorem 2.2.2 are satisified?

2.9. Consider the IVP $\dot{x} = f(x)$ with $x(0) = x_0$.

(i) Suppose that $f(x) > 0$ in the interval $[x_0, x_1]$. Show that the solution of the IVP reaches x_1 at time

$$T = \int_{x_0}^{x_1} \frac{1}{f(x)} dx. \tag{2.21}$$

Hint: Since $dx/dt > 0$ consider the inverse function $t = t(x)$ (the inverse function theorem applies here) and compute its derivative $\frac{dt}{dx}$.

(ii) Suppose that $f(x) > 0$ on $[x_0, \infty)$. Show that the solution of the IVP diverges to ∞ at a positive finite time T if and only if

$$T = \int_{x_0}^{\infty} \frac{dx}{f(x)} < \infty. \tag{2.22}$$

2.10. Consider the initial-value problem

$$\begin{aligned} \dot{x}(t) &= x(t) \\ x(t_0) &= x_0, \end{aligned}$$

(i) Compute the first four Picard iterations starting with $q_0(t) = x_0$. Do these iterates appear to be converging to the actual solution? Justify your answer.

(ii) Compute the first four Picard iterations starting with $q_0(t) = \cos(t)$ Do these solutions appear to converge to the actual solution? Justify your answer.

2.11. For each of the equalities or inequalities in (2.17), explain why it is valid.

2.12.* Assume that $\phi(t)$ is continuous and that $u(t)$ is differentiable and satisfies the inequality $\dot{u}(t) \leq \phi(t)u(t)$. Define $v(t) = \exp\left(\int_0^t \phi(t)\, dt\right)$. Perform the following steps:

(i) Show that $v'(t) = \phi(t)v(t)$ and that $v(t) \geq 0$ for all $t \geq 0$.

(ii) Show that

$$\frac{d}{dt}\left(\frac{u(t)}{v(t)}\right) \leq 0.$$

(iii) Using the fact that

$$\frac{u(t)}{v(t)} \leq \frac{u(0)}{v(0)}$$

to conclude that $u(t) \leq u(0)v(t)$ for all $t \geq 0$.

This is an extension of Gronwall's inequality (Lemma 2.2.11). Notice that we did not require positivity among the functions.

Notes

3 Numerical Approximation to Solutions of ODEs

Computers are composed of nothing more than logic gates stretched out to the horizon in a vast numerical irrigation system.
—Stan Augarten

When you first encountered differential equations it was likely in a classically motivated course where you were given several different types of ODEs and taught how to compute exact solutions to them. Every week you were introduced to a different type of ODE with a different type of solution, and a different approach to finding that solution. This probably all seemed wonderfully convenient, as if the natural world were made to fit nicely into a semester's worth of specific examples and categories. But reality is much messier. Aside from linear systems of ODEs and a few other special cases, most randomly selected ODEs do not have a closed-form solution. Indeed we claim (without proof or verifiable evidence) that the set of ODEs with a closed-form solution is a set of measure zero among the set of all ODEs. We must find other ways to understand the solutions of most differential equations.

In the chapters following this one, we focus on identifying the long-term behavior of a class of solutions for a class of initial conditions, without explicitly computing a solution for any specific set of initial conditions. This is a powerful technique that allows us to determine the behavior of a dynamically evolving system even if we don't know the exact starting location.

But before looking at the long-term behavior of dynamical systems, it is useful to have some numerical methods for approximating solutions to differential equations. There are entire courses devoted to this topic, and a plethora of textbooks, as well as many academic careers, so this chapter should not be viewed as comprehensive. We give a brief introduction to numerical methods for ODEs, biased by the authors' personal views and opinions. We will work through a few key examples to illustrate some of the wide variety of approaches available.

This chapter focuses primarily on approximating the solution to the scalar-valued IVP

$$\dot{u} = f(u), \quad u(0) = u_0, \tag{3.1}$$

where $f(u)$ is at least C^2 on its maximal open interval of existence I. Approximation of vector-valued ODEs is similar, with the standard complications that come from high-order derivatives in multiple dimensions. In all the derivations that follow, similar steps can be taken to approximate solutions of ODEs when multiple independent variables are involved.

3.1 The Simplest Methods

Suppose we know the value of $u(t)$ at a fixed time t and are interested in determining the value of the solution $u(t + \Delta t)$ at time $t + \Delta t$ (we assume that Δt is small). Assuming that $u(t) \in C^2$, the Taylor series expansion of $u(t + \Delta t)$ about t gives

$$u(t + \Delta t) \approx u(t) + (\Delta t)\dot{u}(t). \tag{3.2}$$

The numerical methods in this section are all simple variations built on this basic approximation.

3.1.1 Forward Euler

Substituting the relation $\dot{u} = f(u)$ into (3.2) gives

$$u(t + \Delta t) \approx u(t) + (\Delta t)f(u(t)). \tag{3.3}$$

This means that if we know the value of $u(t)$, and if we can evaluate $f(u(t))$, then we can estimate $u(t + \Delta t)$. This is the simplest numerical approximation for an initial value problem and is called the *forward first-order Euler*[18] *finite difference method*. This is simple to implement and easy to analyze, as we show later.

Begin the numerical algorithm by selecting a compact interval $[t_0, T]$ that is (hopefully) inside the maximal open interval of existence, and a small positive number $\Delta t > 0$ to be the time step. Let n be the greatest integer such that $t_0 + n\Delta t \leq T$. To keep things simple, we assume from now on that $T = t_0 + n\Delta t$. Partition the interval $[t_0, T]$ into subintervals $[t_{i-1}, t_i]$ where

$$t_i = t_0 + i\Delta t, \qquad \text{for } i = 0, 1, \ldots, n. \tag{3.4}$$

The algorithm produces a sequence $v_0, v_1, \ldots, v_n$, which we hope are close approximations to the true (but unknown) values $u_0, u_1 = u(t_1), \ldots, u_n = u(t_n)$. For the first step of the algorithm use the initial condition $v_0 = u(t_0) = u_0$ of the IVP, since u_0 is known. Now iteratively estimate each subsequent value v_{i+1} from the previous estimate v_i by

$$v_{i+1} = v_i + (\Delta t)f(v_i), \qquad \text{for } i = 1, 2, \ldots, n.$$

Approximate points between (t_i, v_i) and (t_{i+1}, v_{i+1}) linearly, which corresponds graphically to drawing straight line segments to connect (t_i, v_i) to (t_{i+1}, v_{i+1}).

Example 3.1.1. Consider the very simple IVP

$$\dot{x} = ax + b, \qquad x(0) = 1, \tag{3.5}$$

[18] *Euler* is pronounced Oy-ler. If you say You-ler, no one will know who you are talking about.

whose explicit solution is

$$x(t) = \left(1 + \frac{b}{a}\right) e^{at} - \frac{b}{a}.$$

To approximate this solution using Euler's method, first choose a time-step Δt, and then, beginning at $v_0 = x(0) = 1$, compute

$$\begin{aligned} v_1 &= v_0 + \Delta t(av_0 + b) = 1 + \Delta t(a + b) \\ v_2 &= v_1 + \Delta t(av_1 + b) = 1 + 2\Delta t(a + b) + (\Delta t)^2(a^2 + ab) \\ \vdots &= \vdots \end{aligned}$$

Algorithm 3.1 gives Python code to implement this. Figure 3.1 shows the results of this algorithm for the IVP when $a = -1$ and $b = 2$ on the interval $[0, 3]$. The exact solution is

$$x(t) = 2 - e^{-t},$$

which asymptotically approaches 2, that is

$$\lim_{t\to\infty} x(t) = 2.$$

For large Δt (such as $\Delta t = 1$ in the left panel of the figure) the approximated solution is not very good. But the smaller the time step the more accurate the solution is. The approximated solution for $\Delta t = 0.01$ is essentially indistinguishable from the exact solution. Since we only needed the approximate solution up to time 3, the time cost of this computation for $\Delta t = 0.01$ is small, and if we needed it to be even more accurate, it still would not be costly to take Δt very small (like 10^{-5}).

In this case we can easily compute the error of each approximation because we know the exact solution. The right panel of Figure 3.1 shows the logarithm of the error plotted against time for various choices of Δt for this problem. But normally the reason you need a numerical method is precisely because you don't know the exact solution, so estimating errors is usually more difficult than this.

Remark 3.1.2. Euler's method works unusually well in Example 3.1.1 because the true solution $x(t)$ converges rapidly toward its final state $\lim_{t\to\infty} x(t) = 2$. In this special case, any choice of Δt gives a numerical approximation that also converges to that value. But that isn't the case with all IVPs, even simple ones, as the next example illustrates.

```python
def euler_1st_linear_ode(a, b, delta_t, T, ini_val):
    """Approximate the IVP  x' = ax+b with x(0) = ini_val
    using the 1st-order forward Euler method with time-step
    delta_t from time 0 to the final time T.

    Returns
    =======
        t_vals :  array of all times t_i
        v :       array of approximations of x(t_i)
    """

    # Find the points t_i where x will be estimated.
    t_vals = np.linspace(0, T, int(np.floor(T / delta_t)) + 1)

    v = np.zeros(np.shape(t_vals))  # Initialize the solution
    v[0] = ini_val                  # Set the initial value
    for i in range(1, len(t_vals)): # Compute subsequent values
        v[i] = v[i - 1] + delta_t * (a * v[i - 1] + b) # Euler

    return t_vals, v
```

Algorithm 3.1: Python implementation of the forward Euler method for the problem (3.5), as discussed in Example 3.1.1.

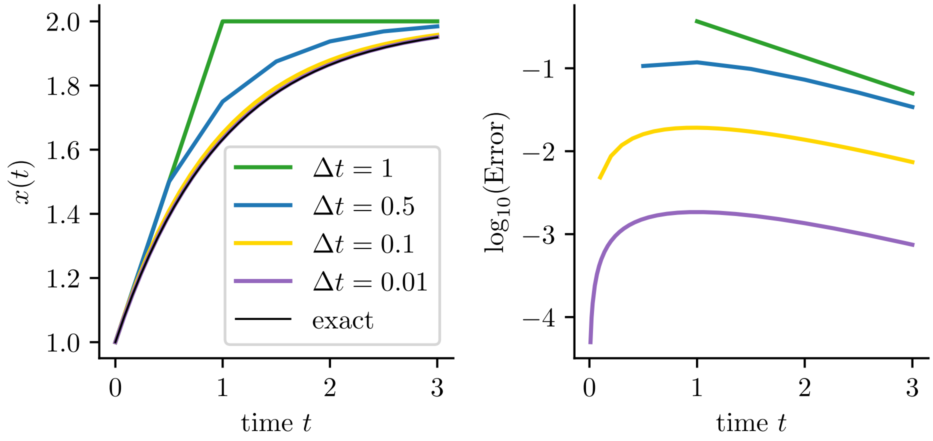

Figure 3.1: Estimated solutions (left panel) and logarithm of errors (right panel) of $\dot{x} = -x+2$, $x(0) = 1$ approximated using Euler's forward method with different time steps, as discussed in Example 3.1.1. In the left panel, the numerically approximated solution (violet) when $\Delta t = 0.01$ can only be seen under the exact solution (black) because it is plotted with a broader line width, but these are essentially the same. In this special case, all the different times steps seem to converge to a pretty good answer fairly quickly because the true solution converges rapidly to the limiting value of 2. But in cases where the true solution diverges rapidly toward infinity, as in Figure 3.2, then the size of ΔT has a significant long-term effect on the error of the approximation. Each of these methods starts with the correct initial condition, so the initial error (at $t = 0$) is zero (with logarithm $-\infty$), so that's not plotted here. The next step is at $t = \Delta T$, which is where each of these error plots starts in the right panel. Even though the errors are initially not good for some of the larger time steps, they all get better over time for this particular IVP. Unfortunately that steady improvement is not always the case for other IVPs.

Example 3.1.3. Consider the same IVP as in Example 3.1.1, but now with $a = 1$ and $b = 2$:

$$\dot{x} = x + 2, \quad x(0) = 1.$$

The exact solution is

$$x(t) = 3e^t - 2,$$

which has exponential growth, rather than exponential decay. That means as $t \to \infty$, the solution is unbounded.

Using the same time steps as in Example 3.1.1 we now get approximated solutions and errors plotted in Figure 3.2. Although the error is initially not terrible, it grows with each step, no matter how small Δt is. This is, of course, a problem if we are interested in the dynamics of the IVP after a long time has passed.

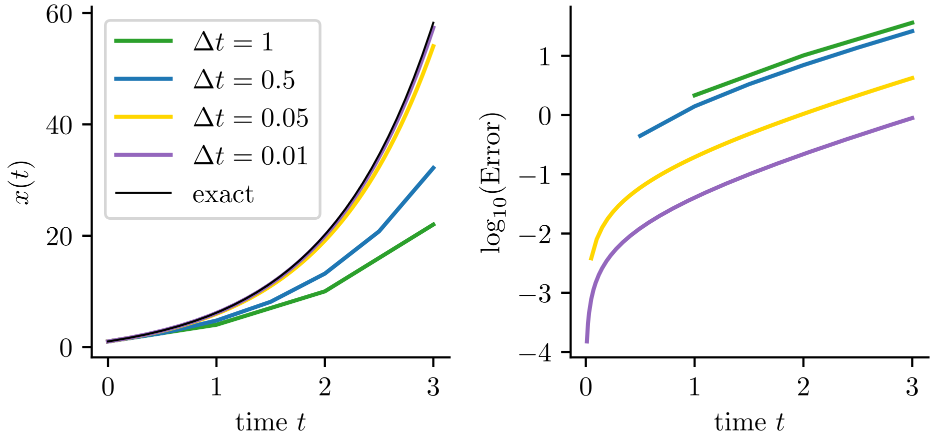

Figure 3.2: Plots of the approximated solutions (left panel) and log of errors (right panel) of $\dot{x} = x + 2$, $x(0) = 1$ found via Euler's forward method using different values of the time step Δt. As with Figure 3.1, each of these methods starts with the correct initial condition, so the initial error is zero (with logarithm $-\infty$), so that's not plotted here. The next step is at $t = \Delta T$, which is where each of these error plots starts. Unlike the case plotted in Figure 3.1, these errors all get worse over time, regardless of the choice of time step ΔT.

Remark 3.1.4. This last example illustrates an issue that is often neglected in numerical analysis and in mathematical modeling in general, but is extremely important. Even if we have well-posed IVP modeling a given phenomenon, with guarantees of existence and uniqueness of the solutions, our job is not finished.

Although solutions may be guaranteed to exist, if they are exponentially growing in time, and have exponential dependence on initial conditions and parameters, then small changes in initial values result in large changes in the final prediction. That is a big problem if the initial conditions and parameters are not known exactly or are represented imprecisely (say as a floating-point number), as is the case for almost all physical quantities and measurements. Solutions that are growing exponentially (in any sense) diverge from their numerical approximation as time progresses, and we can only rely on the approximate solution for a small window of time.

Be careful with your models, and with what you claim the model can predict reliably. In such a setting, we can't really say anything about the exact answer, given a set of parameters or initial conditions. But, as we show later in the book, we can often accurately understand some average behavior of the system. This is why climate scientists make no claims about specific weather patterns on 16 January 2116 but they can make general statements about the state of the climate (mean temperature, precipitation rate, etc.) in the year 2116 under certain conditions. In the climate/weather setting different solutions are exponentially dependent on parameters and initial conditions, but we also recognize that the wind speed in the upper atmosphere can't grow exponentially to infinity, so somehow things are contained.

Consider now what what happens if we apply a numerical method to solve a problem that we know doesn't have a unique solution. Without checking for convergence of your numerics and/or analyzing the nature of the solution beforehand, it is generally impossible to know what the numerical method will do, and at times it gives an answer that is utterly incorrect, as the next example shows.

Example 3.1.5. Recall that the solution for

$$\dot{x} = x^2, \quad x(0) = 1,$$

is given by

$$x(t) = \frac{1}{1-t},$$

which is **not** defined for $t \geq 1$.

Ignoring this fact, if we just march forward and apply forward Euler's method to this equation we end up with a 'solution' like those depicted in Figure 3.3. The approximated solution for all choices of Δt continues merrily beyond $t = 1$, despite the fact that we know the exact solution blows up to infinity as $t \to 1$. In most cases you do not know the exact solution, but the approximate solution does not readily reveal that the true solution blows up; that is to say, Euler's method (and most other numerical methods) lies to you like GPT[a] would, pretending it can compute the solution even where none exists at all.

[a] GPT is a deep learning architecture for large language models popular in 2018–2023 that were deceptively billed as 'artificial intelligence' because of their ability to mimic human-generated text. They were also infamous for their inability to admit, or even recognize, when their answers were spectacularly wrong.

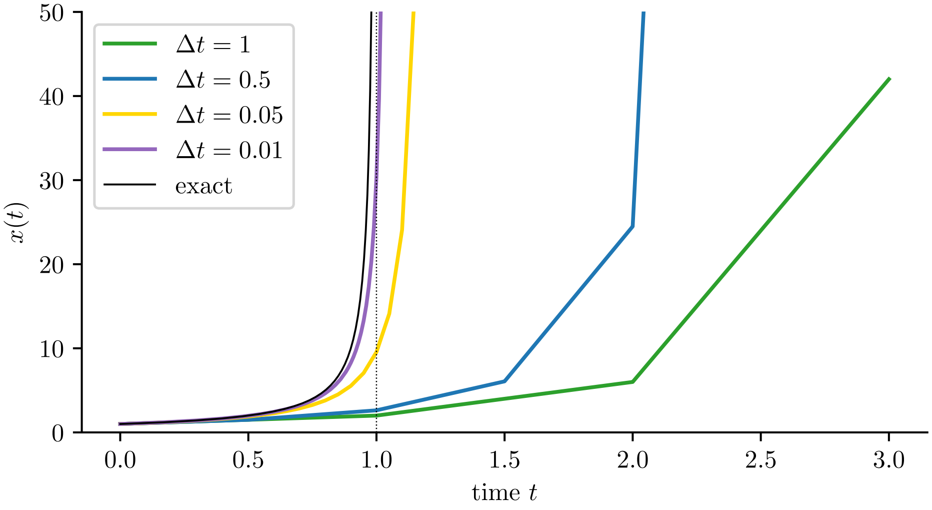

Figure 3.3: Numerical solutions of $\dot{x} = x^2, x(0) = 1$ using forward Euler at various time steps. All of the numerical solutions blithely cross $t = 1$ with a finite value of $x(1)$, while the true solution blows up (approaches ∞) as $t \to 1$ from the left.

3.1.2 Backward Euler

Consider a different approach to approximating the solutions of (3.1). In the formula (3.4) for forward Euler, we really had no special mathematical reason for choosing to evaluate $f(u(t))$ rather than $f(u(t + \Delta t))$. Using $f(u(t + \Delta t))$ instead gives the *backward Euler method*:

$$u(t + \Delta t) \approx u(t) + (\Delta t) f(u(t + \Delta t)), \tag{3.6}$$

or, as an iterative algorithm

$$v_{i+1} = v_i + (\Delta t) f(v_{i+1}). \tag{3.7}$$

If the problem is linear with $f(v) = av + b$, then (3.6) becomes

$$u(t + \Delta t) = u(t) + (\Delta t)[au(t + \Delta t) + b],$$

which can be solved for $u(t + \Delta t)$ provided that $1 - (\Delta t)a \neq 0$, that is,

$$\begin{aligned} u(t + \Delta t) - (\Delta t)au(t + \Delta t) &= u(t) + (\Delta t)b, \\ \Rightarrow u(t + \Delta t)[1 - (\Delta t)a] &= u(t) + (\Delta t)b, \\ u(t + \Delta t) &= \frac{u(t) + (\Delta t)b}{1 - (\Delta t)a}. \end{aligned}$$

In practice, we want to not only avoid $1 = (\Delta t)a$, but we also want to make sure that we don't even get close to that. We would prefer for Δt to small enough that $(\Delta t)a \ll 1$.

But there is a problem with backward Euler when the function $f(u)$ is nonlinear, because we are trying to find v_{i+1}, and (3.7) puts it on both sides of the equation, generally making it hard to solve for. To deal with this, we can use Newton's method or one of its many variants to estimate v_{i+1}. Newton's method and its many variations work magnificently well here for nonlinear problems, because Newton's method has a good starting point, namely v_i.[19]

There are real circumstances where using (3.7) for a nonlinear $f(u)$, is preferable to forward Euler, particularly for large systems of differential equations. This is true despite the need for Newton (or quasiNewton) methods, with all the increased temporal complexity that implies for each time step. We describe those benefits in a later chapter...something to look forward to!

The two methods discussed so far are classical examples of two different classes of numerical time-stepping methods.

(i) Forward Euler is an *explicit* time stepping method because if we know u_i, then u_{i+1} is explicitly described as $u_{i+1} = u_i + \Delta t f(u_i)$, and it can be directly evaluated at each time step.

(ii) Backward Euler is an *implicit* method because each u_{i+1} is only defined implicitly by the relation (3.7). That is, even if we know u_i, the next step u_{i+1} is only defined implicitly as the value that solves the equation $u_{i+1} = u_i + \Delta t f(u_{i+1})$.

3.1.3 Trapezoidal Method

Another implicit method can be derived by averaging the forward and backward Euler methods, which gives the *trapezoidal* method:

$$u(t+\Delta t) \approx u(t) + \frac{\Delta t}{2}\left[f(u(t)) + f(u(t+\Delta t))\right]. \tag{3.8}$$

It may seem strange that using (3.8) would be preferred to (3.6) since they are both implicit methods and (3.8) seems more complicated. The answer lies in the fact that (3.8) gives more information at each time step, so the error is smaller for the same size time step Δt. The trapezoid method is still implicit though, because the next step v_{i+1} is not explicitly written as a function of v_i, but rather must be found by solving an equation:

$$v_{i+1} = v_i + \frac{\Delta t}{2}\left[f(v_i) + f(v_{i+1})\right] \tag{3.9}$$

3.2 Error and Multistep Methods

All the numerical methods in the previous section are examples of *finite difference* methods for solving the IVP (3.1). That means they first uniformly discretize time into a finite number of values $t_0, t_1, \ldots, t_n$ with each $t_i = t_0 + i\Delta t$, and then they

[19] Recall that Newton's method converges quadratically (that is, very quickly) if has a good initial guess, sufficiently close to the correct answer. But Newton can converge slowly, or not at all, if its initial starting point is not close enough to the final answer. If Δt is very small, and $u(t)$ is continuous, then $u(t_i)$ should be close to $u(t_{i+1})$, usually making $u(t_i)$ a good initial starting point.

(i) Approximate $\dot{u}(t)$ with some function $\dot{U}(t, \Delta t)$, expressed as a finite linear combination of values of the form $u(t)$, $u(t + \Delta t)$, $u(t - \Delta t)$, and so forth.

(ii) Approximate $f(u(t))$ with some function $F(u(t))$. For many difference methods (but not all) $F(u) = f(u)$.

Using these approximations and the relation

$$\dot{U}(t, \Delta t) \approx F(u(t)) \tag{3.10}$$

one then solves for $u(t + \Delta t)$ (corresponding to v_{i+1}) in terms of $u(t)$, $u(t - \Delta t)$, and so forth, to get an approximate expression for v_{i+1} in terms of v_i, v_{i-1} and so forth.

Example 3.2.1. The forward Euler method is a finite difference method with $\dot{U}(t, \Delta t) = \frac{u(t+\Delta t)-u(t)}{\Delta t}$, and $F(u(t)) = f(u(t))$. Equation 3.10 is essentially equivalent to (3.2).

Example 3.2.2. The backward Euler method is a finite difference method with $\dot{U}(t, \Delta t) = \frac{u(t)-u(t-\Delta t)}{\Delta t}$, and $F(u(t)) = f(u(t))$. Equation (3.10) is essentially equivalent to (3.7) after making a substitution of $t \mapsto t + \Delta t$ (so t in (3.10) corresponds to $t + \Delta t$ in (3.6).

Example 3.2.3. The trapezoidal method is a finite difference method with $\dot{U}(t, \Delta t) = \frac{u(t+\Delta t)-u(t)}{\Delta t}$, and $F(u(t)) = \frac{1}{2}[f(u(t)) + f(u(t + \Delta t))]$. Equation (3.10) is essentially equivalent to (3.8) after making a substitution of $t \mapsto t + \Delta t$ (so t in (3.10) corresponds to $t + \Delta t$ in (3.6).

3.2.1 Types of Error

The errors of the approximations tell us how good the approximations are. Here are three different ways of thinking about the error.

Definition 3.2.4. *If the method begins with the true initial value $v_0 = u_0$ and if v_i is the resulting estimate for $u_i = u(t_i)$, then the* global error *on the interval $[t_0, T]$ (where $T = t_0 + n\Delta t$) is*

$$\max_{1 \le i \le n} |u_i - v_i|. \tag{3.11}$$

The local truncation error (LTE) *$\tau(t)$ of a finite difference approximation at time t measures how bad the approximation $\dot{U} \approx F(u(t))$ is after one step (locally), namely,*

$$\tau(t) = |\dot{U}(t) - F(u(t))|. \tag{3.12}$$

Finally, the one-step error $\mathscr{L}(t)$ *is the error of the method after one step: let* $v(t+\Delta t)$ *be the estimate produced by one step (also called one* update*) of the difference method, using an exact value of* $u(t)$*. The* one-step error[20] *is*

$$\mathscr{L}(t) = |u(t+\Delta t) - v(t+\Delta t)|. \tag{3.13}$$

Finally, we say that the method has order p *if* p *is the largest value for which the LTE is* $O((\Delta t)^p)$*.*

The LTE $\tau(t)$ measures how bad the approximation $\dot{U} \approx F(u(t))$ is after one step. If $F = f$ and $\dot{U}(t,\Delta) = \dot{u}(t)$, then the ODE $\dot{u}(t) = f(u(t))$ guarantees that $\tau(t) = 0$, but this is rarely the case. The one-step error $\mathscr{L}(t)$ measures how far apart u and the approximation v can get in one step. And the global error measures how far apart the solution values u_i and the approximations v_i can get over the whole interval $[t_0, T]$. The global error is what we usually want to know when we are talking about the error of a numerical method, but it is usually not possible to compute unless we already know the exact solution for u (in which case we don't really need a numerical method).

Both the one-step error and the LTE are usually computable, using Taylor's theorem, even if we don't know the true values u_i. To use Taylor's theorem we need f to be sufficiently differentiable, so we'll assume for the rest of this section that $f \in C^r$ for r large enough that Taylor's theorem applies to whatever order we happen to need for our analysis.

In some cases we can estimate the global error from the one-step error, but in most cases the LTE gives a better bound on the global error than you get from the one-step error. To estimate the global error from the one-step error when the IVP is well behaved, we would expect the global error after $n = \frac{T}{\Delta t}$ steps to be on the order of

$$n\mathscr{L} = \frac{T}{\Delta t}\mathscr{L} = O\left(\frac{\mathscr{L}}{\Delta t}\right), \tag{3.14}$$

because at each step we should be introducing approximately $\mathscr{L}$ more error into the system. But the one-step error is measuring the error you'd get in approximating u_{i+1} if you had started at the true value of u_i, while the numerical method does not start at u_i to compute v_{i+1}—it starts instead at the approximation v_i.

It can be shown that when f is sufficiently well behaved, that the global error E is bounded by

$$E \leq C \max_{1 \leq i \leq n} \tau(t_i), \tag{3.15}$$

where C is independent of Δt. Thus the error of the method is no larger than $O(\tau(t))$ for some $t \in [t_0, T]$. This is why we say that the method has order p if p is the largest integer such that $\tau(t) = O((\Delta t)^p)$ for all $t \in [t_0, T]$.

[20] Some authors call this the *local truncation error*, which makes it important to check what people really mean when they talk about local error—they could mean either the one-step error of (3.13) or they could match our definition (3.12) of LTE.

Example 3.2.5 (Error of Forward Euler). The LTE for forward Euler is

$$\begin{aligned}\tau(t) &= |\dot{U}(t) - F(u(t))| \\ &= \left|\frac{u(t+\Delta t) - u(t)}{\Delta t} - f(u(t))\right| \\ &= \left|\frac{u(t+\Delta t) - u(t)}{\Delta t} - \dot{u}(t)\right|, \qquad (3.16)\end{aligned}$$

where the last line follows from the ODE $\dot{u} = f(u)$. Taylor's theorem gives

$$u(t+\Delta t) = u(t) + (\Delta t)\dot{u}(t) + \frac{(\Delta t)^2}{2}\ddot{u}(t) + O\left((\Delta t)^3\right). \qquad (3.17)$$

Substituting (3.17) into (3.16) gives

$$\begin{aligned}\tau(t) &= \left|\frac{u(t) + \Delta t\,\dot{u}(t) + \frac{(\Delta t)^2}{2}\ddot{u}(t) + O\left((\Delta t)^3\right) - u(t)}{\Delta t} - \dot{u}(t)\right| \qquad (3.18) \\ &= \left|\frac{\Delta t}{2}\ddot{u}(t) + O\left((\Delta t)^2\right)\right| \qquad (3.19) \\ &= O(\Delta t). \qquad (3.20)\end{aligned}$$

The one-step error of forward Euler is

$$\begin{aligned}\mathscr{L}(t) &= |u(t+\Delta t) - (u(t) + \Delta t f(u(t))) \\ &= \frac{(\Delta t)^2}{2}\ddot{u}(t) + O\left((\Delta t)^3\right) \\ &= O((\Delta t)^2)\end{aligned}$$

The global error of forward Euler is not immediately computable, but the LTE is $O(\Delta t)$, so (3.15) shows that the order of forward Euler is 1. This also agrees with (3.14), since $\frac{\mathscr{L}}{\Delta t} = \frac{O((\Delta t)^2}{\Delta t} = O(\Delta t)$. Both of these arguments agree with the numerical results seen in Figures 3.1 and 3.2. Those figures show an order of magnitude (factor of ten) increase in Δt (say from 0.05 (yellow) to 0.5 (blue)) results in approximately one order of magnitude increase (corresponding to a gap of 1 on the $\log_{10}$ scale) in the global error, which is the maximum distance between the yellow and the blue curves).

Backward Euler is also a first-order method, although the derivation takes a bit more effort (Yay for exercises!).

Example 3.2.6. The LTE of the trapezoid method is

$$
\begin{aligned}
\tau(t) &= \left| \frac{u(t+\Delta t) - u(t)}{\Delta t} - \frac{1}{2}\left[f(u(t)) + f(u(t+\Delta t)) \right] \right| \\
&= \left| \frac{u(t) + (\Delta t)\dot{u}(t) + \frac{(\Delta t)^2}{2}\ddot{u}(t) + \frac{(\Delta t)^3}{6}\dddot{u}(t) + O\left((\Delta t)^4\right) - u(t)}{\Delta t} \right. \\
&\qquad \left. - \frac{1}{2}\left[\dot{u}(t) + \dot{u}(t+\Delta t) \right] \right| \\
&= \left| \dot{u}(t) + \frac{\Delta t}{2}\ddot{u}(t) + \frac{(\Delta t)^2}{6}\dddot{u}(t) + O\left((\Delta t)^3\right) - \dot{u}(t) - \frac{\Delta t}{2}\ddot{u}(t) \right. \\
&\qquad \left. - \frac{(\Delta t)^2}{4}\dddot{u}(t) - O\left((\Delta t)^3\right) \right| \\
&\leq \frac{(\Delta t)^2}{12} |\dddot{u}(t)| + O\left((\Delta t)^3\right) = O((\Delta t)^2).
\end{aligned}
$$

Hence the trapezoid method is a second-order method. The effects of using a second-order rather than a first-order method should be apparent from the exercises. The difference in error can be significant, and if an implicit method is workable the trapezoid method doesn't cost much more than the backward Euler method.

3.2.2 Multistep Methods

Another method we can derive from the forward and backward Euler methods comes from averaging them in a different way, approximating the derivative by what is called a *centered difference*

$$
\dot{u}(t) \approx \frac{u(t+\Delta t) - u(t-\Delta t)}{2\Delta t}. \tag{3.21}
$$

This leads to the *midpoint* or *Leapfrog* method:

$$
u(t+\Delta t) \approx u(t-\Delta t) + 2(\Delta t) f(u(t)) \tag{3.22}
$$

or

$$
v_{i+1} = v_{i-1} + 2(\Delta t) f(v_i). \tag{3.23}
$$

The midpoint method is also second order, but it is fundamentally different from the trapezoid method because the midpoint method uses two time steps (both v_i and v_{i-1}) to construct the next state v_{i+1}. This is what is referred to as a *multistep* method, and is one of the simplest in an entire class of multistep methods. We can also create an implicit midpoint method if desired, but as we noticed earlier, this doesn't affect the order of the method.

There are many other multistep methods. In fact, you can come up with as many multistep methods as you like at whatever order of accuracy you prefer, both explicit and implicit. We will only highlight a few examples here. The most common multistep methods are linear multistep methods (LMMs), which are essentially a weighted interpolation between current and past information (and future if the method is implicit). We consider two classes of LMMs, both of which are referred to as *Adams* methods. These have the general form:

$$u(t+r(\Delta t)) = u(t+(r-1)(\Delta t)) + (\Delta t)\sum_{j=0}^{r} b_j f(u(t+j(\Delta t))), \tag{3.24}$$

or

$$v_{i+r} = v_{i+r-1} + (\Delta t)\sum_{j=0}^{r} b_j f(v_{i+j}), \tag{3.25}$$

where the term on the left hand side of (3.25) is the next step, and the first term on the right is the current step with all the terms in the summation being the information at the past time steps. The coefficients b_j are chosen to maximize the order of accuracy of the method, and possibly to simplify the algebra in the algorithm. Choosing $b_r = 0$ makes the method explicit, and then the r coefficients $b_0, b_1, \ldots, b_{r-1}$ can be chosen so the method has order r. Following this procedure yields what is referred to as the r-step *Adams–Bashforth method*.

Example 3.2.7. Consider the two-step Adams–Bashforth method (the one-step method is just forward Euler), corresponding to (3.24) with $r = 2$.

$$u(t+2(\Delta t)) = u(t+\Delta t) + (\Delta t)\left[b_0 f(u(t)) + b_1 f(u(t+\Delta t))\right]. \tag{3.26}$$

As before, using Taylor's formula and the fact that $f(u(t)) = \dot{u}(t)$, reorganize (3.26) to try to make the LTE be $O\left((\Delta t)^2\right)$):

$$\begin{aligned}
\tau(t) &= \left|\frac{u(t+2(\Delta t)) - u(t+\Delta t)}{\Delta t} - b_0 f(u(t)) - b_1 f(u(t+\Delta t))\right| && (3.27)\\
&= \left|\dot{u}(t) + \frac{3(\Delta t)}{2}\ddot{u}(t) + \frac{7(\Delta t)^2}{6}\dddot{u}(t) - b_0\dot{u}(t) - b_1\dot{u}(t+\Delta t) + O\left((\Delta t)^3\right)\right| && (3.28)\\
&= \left|\left[1 - b_0 - b_1\right]\dot{u}(t) + (\Delta t)\left[\frac{3}{2} - b_1\right]\ddot{u}(t) + O\left((\Delta t)^2\right)\right|. && (3.29)
\end{aligned}$$

To ensure that the truncation error is $O\left((\Delta t)^2\right)$, choose b_0 and b_1 to satisfy the linear system

$$1 - b_0 - b_1 = 0, \qquad b_1 = \frac{3}{2}, \tag{3.30}$$

which has the unique solution

$$b_0 = -\frac{1}{2}, \qquad b_1 = \frac{3}{2}, \tag{3.31}$$

and hence the two-step Adams–Bashforth method is given by:

$$v_{i+2} = v_{i+1} + (\Delta t)\left[-\frac{1}{2}f(v_i) + \frac{3}{2}f(v_{i+1})\right]. \tag{3.32}$$

Remark 3.2.8. Returning to (3.25), now consider what happens if we don't set $b_r = 0$, so the method is implicit. This gives one more degree of freedom, so that for an r-step method (with appropriate choice of coefficients b_j) the method has order $r+1$. These methods are referred to as r*-step Adams–Moulton* methods and are also frequently used in practice. The one-step Adams–Moulton method is simply the trapezoid method we introduced earlier, and higher-order Adams–Moulton methods can be derived in a straightforward fashion.

3.2.3 Benefits and downsides to multistep methods

High order multistep methods are a relatively straightforward way to generate a highly accurate time stepping scheme, but they also have some significant drawbacks. Here are a few things to keep in mind.

(i) There is a fine balance in high-order multistep methods between the spatial complexity, the temporal complexity, and the accuracy of the method. Because of increased accuracy, a higher-order method can use larger time steps, which can reduce the temporal complexity of the algorithm. Each time step requires evaluation of the function $f(u(t))$ at several previous time steps, but each of these evaluations has already occurred (at previous time steps) and can be stored in memory, so each forward time step only requires a single evaluation of the function $f(u(t))$. Especially in higher dimensions, evaluation of $f(u(t))$ may be expensive, so avoiding unnecessary computations is helpful. But this require more storage—an r-step method requires storing the previous $r-1$ evaluations of $f(u(t))$. In low dimensions, this is not a significant cost, but for more complicated problems in high dimensions, the spatial complexity may be significant.

(ii) A single-step method can just start with the initial value u_0, but multistep methods need more starting values than are usually given by the problem. If we are using a 3-step Adams–Bashforth method for an IVP that only specifies $u(0) = u_0$, then how do we come up with the other two values $u(-\Delta t)$ and $u(-2(\Delta t))$? A simple alternative is to use forward or backward Euler for the first time step then switch to a 2-step method for the second, and so forth. This approach sacrifices some accuracy in the first few time steps, and we have no guarantee that we can make it up later on. It turns out that this is 'sort of' ok. Because we are only using a lower-order method for the first step, the error doesn't cascade to the entire solution so long as the first time step is only one order less than the multistep method used through the rest of the simulation. This means that we can legitimately use forward Euler for the first step, and 2-step Adams–Bashforth for the rest and maintain a second order solution. Roughly speaking, this is because $\mathscr{L}(t)$ is $O\left((\Delta t)^3\right)$ for the 2-step Adams–Bashforth method so that over $T/(\Delta t)$ time steps the global error is $O\left((\Delta t)^2\right)$, and the initial forward Euler step is also $O\left((\Delta t)^2\right)$ and only happens once, so the global error is still approximately $O\left((\Delta t)^2\right)$.

(iii) Multistep methods rely on high-order derivatives canceling each other out when estimating the truncation error. This is only true if the true solution has continuous derivatives of high-enough order. As discussed later in this book, we are sometimes interested in solutions of differential equations that do not have a large number of continuous derivatives.

3.3 Higher-Order Single-Step Methods

There are clear reasons (as described above) for not using multistep methods, but higher order accuracy is often essential to the accurate approximation of a differential equation. Some alternative approaches are discussed here, beginning with the extremely popular Runge–Kutta methods.

3.3.1 Runge–Kutta or Multistage methods

Rather than use data from the past time steps, multistage methods approximate $f(u(t))$ at intermediate values such as $f(u(t) + (\Delta t)/2)$. These intermediate approximations are then used to approximate $u(t+\Delta t)$. The most popular multistage methods are known as *Runge–Kutta*[21] methods. There are numerous ways to derive these methods, but we won't go into a lot of detail for the derivation here. A quick Google search will identify Runge–Kutta methods of varying orders, and the different variations that one can make on them. We will only discuss a few of the more popular methods here, and give a brief explanation of how they work.

The setting for Runge–Kutta methods is to suppose that we will approximate the solution at the next time step as

$$u(t + \Delta t) = u(t) + (\Delta t)\phi(u(t); \Delta t), \tag{3.33}$$

[21] Pronounced *ROONG-uh COOT-uh*.

where (bear with us on the notation here, it is kind of hideous at first)

$$\phi(u(t); \Delta t) = \sum_{i=0}^{q-1} w_i k_i, \tag{3.34}$$

for some positive integer q is an approximation of $f(u(t))$. The w_i are coefficients that can be selected to achieve the desired order of accuracy (and the simplest algebraic expression), and the k_i are given by one of two expressions

$$k_i = \begin{cases} f\left(u(t) + \Delta t \sum_{j=0}^{i-1} \beta_{ij} k_j\right) & \text{(explicit)} \\ f\left(u(t) + \Delta t \sum_{j=0}^{q-1} \beta_{ij} k_j\right) & \text{(implicit)} \end{cases}, \tag{3.35}$$

where the β_{ij} for $i = 0, \ldots, q-1$ and $j = 0, \ldots, i-1$ are chosen to simplify the scheme or obtain the highest order of accuracy.

The simplest Runge–Kutta method is the single-stage forward Euler method, and isn't very interesting to us. The two-stage, explicit Runge–Kutta method is the next most straightforward to understand.

Example 3.3.1. The two-stage explicit Runge–Kutta method has

$$\phi(u(t); \Delta t) = w_0 k_0 + w_1 k_1, \tag{3.36}$$

where

$$k_0 = f\left(u(t)\right), \tag{3.37}$$

$$k_1 = f\left(u(t) + \Delta t \beta_{10} k_0\right) = f\left(u(t) + \Delta t \beta_{10} f(u(t))\right). \tag{3.38}$$

We want to choose w_0, w_1, and β_{10} so that the method is second order. Before doing so note the following useful identities (the second one comes from applying the chain rule to the first).

$$f(u(t)) = \dot{u}(t), \tag{3.39}$$

$$\dot{u}(t) \frac{\partial f(u)}{\partial u} = \dot{u}(t) f_u(u(t)) = \ddot{u}(t). \tag{3.40}$$

The LTE of the method is

$$\begin{aligned}
\tau(t) &= \frac{u(t+\Delta t)-u(t)}{\Delta t} - \phi(u(t);\Delta t) \qquad &(3.41)\\
&= \frac{u(t+\Delta t)-u(t)}{\Delta t} - w_0 f(u(t)) - w_1 f(u(t)+\Delta t\beta_{10} f(u(t))) \qquad &(3.42)\\
&= \dot{u}(t) + \frac{\Delta t}{2}\ddot{u}(t) - w_0\dot{u}(t) - w_1 f\left(u(t)+\beta_{10}\Delta t\dot{u}(t)\right) + O\left((\Delta t)^2\right) \qquad &(3.43)\\
&= \dot{u}(t) + \frac{\Delta t}{2}\ddot{u}(t) - w_0\dot{u}(t) - w_1 f(u(t)) - w_1\beta_{10}\Delta t\dot{u}(t) f_u(u(t)) \qquad &(3.44)\\
&+ O\left((\Delta t)^2\right) \qquad &(3.45)\\
&= \dot{u}(t)\left[1-w_0-w_1\right] + \Delta t\ddot{u}(t)\left[\frac{1}{2}-w_1\beta_{10}\right] + O\left((\Delta t)^2\right). \qquad &(3.46)
\end{aligned}$$

Thus, this method is second order so long as $w_0 + w_1 = 1$ and $w_1\beta_{10} = \frac{1}{2}$. This gives us two equations with three unknowns so there is a choice to be made among the infinite number of second-order explicit Runge–Kutta methods, depending on the solution of these two equations. To make things simple, let $w_0 = w_1 = \frac{1}{2}$ and $\beta_{10} = 1$. This gives the following two-stage, second-order, explicit Runge–Kutta scheme:

$$\begin{aligned}
u(t+\Delta t) &= u(t) + \frac{\Delta t}{2}\left[k_0 + k_1\right], \quad \text{where} \qquad &(3.47)\\
k_0 &= f(u(t)), \qquad &(3.48)\\
k_1 &= f(u(t)+\Delta t k_0). \qquad &(3.49)
\end{aligned}$$

As described above, there are infinitely many Runge–Kutta methods, corresponding to different choices of the coefficients w_i and β_{ij}. The degree of flexibility afforded by choosing the β_{ij} and w_i not only allows us to select the order of the scheme, but also can greatly reduce the complexity of the resulting coefficients.

Example 3.3.2 (Runge–Kutta fourth order). One of the most famous Runge–Kutta methods is the standard fourth-order 4-stage method which is given by:

$$\begin{aligned}
k_0 &= f(u(t)), \qquad &(3.50)\\
k_1 &= f\left(u(t) + \frac{\Delta t}{2}k_0\right), \qquad &(3.51)\\
k_2 &= f\left(u(t) + \frac{\Delta t}{2}k_1\right), \qquad &(3.52)\\
k_3 &= f\left(u(t) + (\Delta t)k_2\right), \qquad &(3.53)\\
u(t+\Delta t) &= u(t) + \frac{\Delta t}{6}\left(k_0 + 2k_1 + 2k_2 + k_3\right). \qquad &(3.54)
\end{aligned}$$

This method has been popular for decades because the coefficients are easy to remember, and the method is not that difficult to implement. Note that this method assumes that $f(u) \in C^5$.

Many other Runge–Kutta methods have been developed that satisfy certain desirable properties including different types of stability and desired order of accuracy. Implicit Runge–Kutta methods are possible, although rarely used, because each evaluation of the function f is costly in the implicit setting.

Some of the advantages and disadvantages of Runge–Kutta methods include:

(i) RK methods do not require as much storage as they rely only on data at the previous time step to compute the next.

(ii) RK methods can be used to initialize an approximation with only a single value given for the initial data.

(iii) On the negative side, RK methods require several evaluations of the function f for each step. This can be a problem if the function f is computationally expensive to evaluate.

(iv) Another issue with RK methods that only arises in particular circumstances, but is worth noting, is the assumption that interpolation to an intermediate time step is reasonable. In certain circumstances when the system is nonautonomous and $f(u,t)$ depends on t in a discrete way, interpolating between these discrete time steps is not reasonable, and can be misleading to the true solution.

3.3.2 Taylor-Series Methods

Perhaps the most obvious way of defining a time stepping method that satisfies a certain order of accuracy is to simply expand at the desired time step according to Taylor's Theorem, and define the appropriate terms according to derivatives of $\dot{u}(t) = f(u(t))$, for example $\ddot{u}(t) = \dot{u}(t) f_u(u(t)) = f(u(t)) f_u(u(t))$. Similar evaluations can be made for higher derivatives of u as well. Armed with this information consider the Taylor series expansion of the solution:

$$u(t+\Delta t) = u(t) + (\Delta t)\dot{u}(t) + \frac{(\Delta t)^2}{2}\ddot{u}(t) + O\left((\Delta t)^3\right) \tag{3.55}$$

$$= u(t) + (\Delta t) f(u(t)) + \frac{(\Delta t)^2}{2} f(u(t)) f_u(u(t)) + O\left((\Delta t)^3\right). \tag{3.56}$$

Using this we can identify a second order time stepping scheme, provided $f_u(u(t))$ can be evaluated.

Example 3.3.3. To construct a second-order Taylor-series finite difference approximation to the ODE

$$\dot{u} = u^2 \cos u, \tag{3.57}$$

first compute

$$f_u = 2u\cos u - u^2 \sin u. \tag{3.58}$$

Inserting (3.58) into the derivation above gives

$$\begin{aligned} u(t+\Delta t) &\approx u(t) + (\Delta t)u(t)^2 \cos u(t) \\ &\quad + \frac{(\Delta t)^2}{2} u(t)^3 (2\cos^2 u(t) - u(t)\sin u(t)\cos u(t)), \end{aligned}$$

and the update scheme

$$v_{i+1} = v_i + (\Delta t)v_i^2 \cos(v_i) + \frac{(\Delta t)^2}{2} v_i^3 (2\cos^2(v_i) - v_i \sin(v_i)\cos(v_i)).$$

The main drawback of the Taylor-series method is that each choice of $f(u)$ requires the derivation of a different update scheme. This makes these methods relatively impractical and unpopular. Nevertheless they can be useful in certain settings, particularly where Runge–Kutta and multistep methods are not desirable for some other reason.

Remark 3.3.4. All of the derivation we have shown here has focused on autonomous differential equations. This isn't necessary, and equivalent methods are available for the nonautonomous setting. We have focused on autonomous systems only to keep the presentation as simple as possible.

There are many other time stepping approaches, but the brief overview presented here should be sufficient to yield a rough picture of the potential methods. It is important to remember that different numerical methods are appropriate depending on the mathematical model and the type of observations sought from that model. The modeler must weigh the pros and cons of each potential algorithm and make an informed choice on which is most suited to their problem. This isn't an easy task, and one must keep in mind all of the benefits and costs of each approach, including issues of stability which are discussed later in this text.

Remark 3.3.5. Scientists often become enamored with a certain approach and try to mold or adapt that approach to every problem they encounter. While this may seem unproductive, it actually works out quite well as long as there are enough scientists to go around with for all the different obsessions. This means that every potential avenue is chased down by someone who is obsessed with their favorite approach, and eventually, with enough single-minded approaches, one of them will work. In today's world this process succeeds again and again, just beware as young scientists in training that you have prejudices about methods and approaches, even if you don't recognize them yet.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

3.1. Generalize Algorithm 3.1 to work for any ODE of the form $\dot{x} = f(x)$ where $f(x)$ is a function you provide. Turn in your code.

3.2. Using the algorithm you created for the previous problem, solve the ODE $\dot{x} = \sin(x)$, $x(0) = 1$ for $\Delta t = 0.01,\ 0.1, 1.0$, and create a plot similar to Figure 3.1. (Hint: don't worry about the exact solution in this case, just compare different values of Δt.)

3.3. Compute a big-Oh temporal complexity estimate for Algorithm 3.1 for each time step inside the main loop. If $N \approx T/\Delta t$ is the number of time steps to be taken, what is the approximate temporal complexity of Algorithm 3.1?

3.4. Adjust Algorithm 3.1 for the backward Euler method given in (3.7) for the same linear system described in Example 3.1.1.

3.5. Generate plots showing the approximated solution for $a = 1$ and $a = -1$ with $b = 2$ for forward and backward Euler with $\Delta t = 0.1$ and $\Delta t = 0.01$ to the IVP described in Example 3.1.1. Plot the solution on the interval $[0, 3]$.

3.6. Do the same thing as in the previous two exercises, but for the trapezoid method given in (3.9). What differences do you notice in the error of the approximation? Create a single figure that compares all three methods.

3.7. Show that the truncation error for backward Euler is $O(\Delta t)$.

3.8. Show that the midpoint method is second order.

3.9. Derive the 3-step Adams–Bashforth method, and show that it is really third order.

3.10. Derive the 2-step Adams–Moulton method, and show that it is third order.

3.11. Code up an approximate solution to Example 3.1.1 with $a = 1$ and $b = 2$, $\Delta t = 0.01$ and $x(0) = 1$, using forward Euler for the first step, and 2-step Adams–Bashforth for the rest. Run the simulation up to the final time $T = 3.0$ and generate plots of the solution and errors from the true solution.

3.12. Code up an approximation to the IVP $\dot{x} = x(1-x)$ with $x(0) = 0.1$ using the two-stage explicit Runge–Kutta method described in the text. Run the algorithm until final time $T = 4.0$ for time steps of $\Delta t = 0.1,\ 0.01,\ 0.001$, and create a plot that shows how the solution is converging.

3.13. Code up the four-stage, fourth order Runge–Kutta scheme described in the text for the same IVP in the previous problem. Compute the same plots for the same time steps. Can you tell that this scheme is fourth order? How?

3.14. Use a second-order Taylor series method to approximate the solution to $\dot{x}(t) = \sin(x(t))$ with $x(0) = 1$ for time steps $\Delta t = 0.1,\ 0.01$ and $\Delta t = 0.001$. Verify the LTE of this method analytically.

3.15.* Derive the linear system of equations for the coefficients of a three-stage, third-order Runge–Kutta scheme.

Notes