Nonlinear Ordinary Differential Equations

by Peter J. Olver
University of Minnesota

1. Introduction.

These notes are concerned with initial value problems for systems of ordinary differential equations. Here our emphasis will be on nonlinear phenomena and properties, particularly those with physical relevance. Finding a solution to a differential equation may not be so important if that solution never appears in the physical model represented by the system, or is only realized in exceptional circumstances. Thus, equilibrium solutions, which correspond to configurations in which the physical system does not move, only occur in everyday situations if they are stable. An unstable equilibrium will not appear in practice, since slight perturbations in the system or its physical surroundings will immediately dislodge the system far away from equilibrium.

Of course, very few nonlinear systems can be solved explicitly, and so one must typically rely on a numerical scheme to accurately approximate the solution. Basic methods for initial value problems, beginning with the simple Euler scheme, and working up to the extremely popular Runge–Kutta fourth order method, will be the subject of the final section of the chapter. However, numerical schemes do not always give accurate results, and we briefly discuss the class of stiff differential equations, which present a more serious challenge to numerical analysts.

Without some basic theoretical understanding of the nature of solutions, equilibrium points, and stability properties, one would not be able to understand when numerical solutions (even those provided by standard well-used packages) are to be trusted. Moreover, when testing a numerical scheme, it helps to have already assembled a repertoire of nonlinear problems in which one already knows one or more explicit analytic solutions. Further tests and theoretical results can be based on first integrals (also known as conservation laws) or, more generally, Lyapunov functions. Although we have only space to touch on these topics briefly, but, we hope, this will whet the reader's appetite for delving into this subject in more depth. The references [**2**, **9**, **13**, **15**, **17**] can be profitably consulted.

2. First Order Systems of Ordinary Differential Equations.

Let us begin by introducing the basic object of study in discrete dynamics: the initial value problem for a first order system of ordinary differential equations. Many physical applications lead to higher order systems of ordinary differential equations, but there is a simple reformulation that will convert them into equivalent first order systems. Thus, we do not lose any generality by restricting our attention to the first order case throughout. Moreover, numerical solution schemes for higher order initial value problems are entirely based on their reformulation as first order systems.

As always, when confronted with a new problem, it is essential to fully understand the simplest case first. Thus, we begin with a single scalar, first order ordinary differential equation

$$\frac{du}{dt} = F(t,u). \tag{2.1}$$

In many applications, the independent variable t represents time, and the unknown function $u(t)$ is some dynamical physical quantity. Throughout this chapter, all quantities are assumed to be real. (Results on complex ordinary differential equations can be found in [**14**].) Under appropriate conditions on the right hand side (to be formalized in the following section), the solution $u(t)$ is uniquely specified by its value at a single time,

$$u(t_0) = u_0. \tag{2.2}$$

The combination (2.1–2) is referred to as an *initial value problem*, and our goal is to devise both analytical and numerical solution strategies.

A differential equation is called *autonomous* if the right hand side does not explicitly depend upon the time variable:

$$\frac{du}{dt} = F(u). \tag{2.3}$$

All autonomous scalar equations can be solved by direct integration. We divide both sides by $F(u)$, whereby

$$\frac{1}{F(u)}\,\frac{du}{dt} = 1,$$

and then integrate with respect to t; the result is

$$\int \frac{1}{F(u)}\,\frac{du}{dt}\,dt = \int dt = t + k,$$

where k is the constant of integration. The left hand integral can be evaluated by the change of variables that replaces t by u, whereby $du = (du/dt)\,dt$, and so

$$\int \frac{1}{F(u)}\,\frac{du}{dt}\,dt = \int \frac{du}{F(u)} = G(u),$$

where $G(u)$ indicates a convenient anti-derivative[†] of the function $1/F(u)$. Thus, the solution can be written in implicit form

$$G(u) = t + k. \tag{2.4}$$

If we are able to solve the implicit equation (2.4), we may thereby obtain the explicit solution

$$u(t) = H(t + k) \tag{2.5}$$

[†] Technically, a second constant of integration should appear here, but this can be absorbed into the previous constant k, and so proves to be unnecessary.

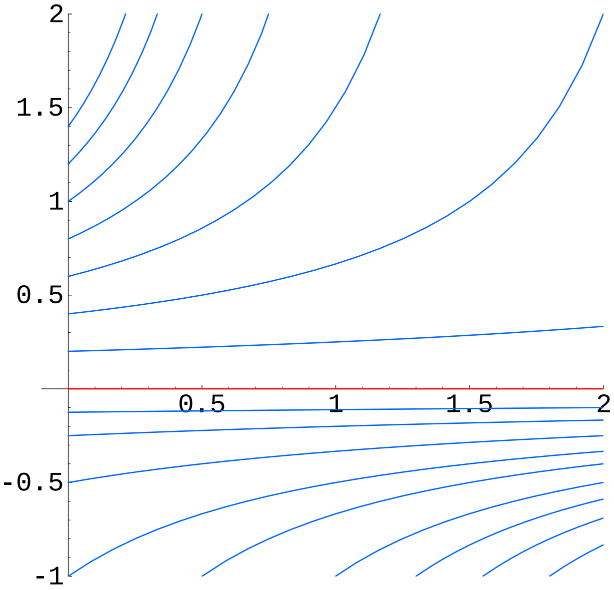


Figure 1. Solutions to $\dot{u} = u^2$.

in terms of the inverse function $H = G^{-1}$. Finally, to satisfy the initial condition (2.2), we set $t = t_0$ in the implicit solution formula (2.4), whereby $G(u_0) = t_0 + k$. Therefore, the solution to our initial value problem is

$$G(u) - G(u_0) = t - t_0, \qquad \text{or, explicitly,} \qquad u(t) = H\big(t - t_0 + G(u_0)\big). \tag{2.6}$$

Remark: A more direct version of this solution technique is to rewrite the differential equation (2.3) in the "separated form"

$$\frac{du}{F(u)} = dt,$$

in which all terms involving u, including its differential du, are collected on the left hand side of the equation, while all terms involving t and its differential are placed on the right, and then formally integrate both sides, leading to the same implicit solution formula:

$$G(u) = \int \frac{du}{F(u)} = \int dt = t + k. \tag{2.7}$$

Before completing our analysis of this solution method, let us run through a couple of elementary examples.

Example 2.1. Consider the autonomous initial value problem

$$\frac{du}{dt} = u^2, \qquad\qquad u(t_0) = u_0. \tag{2.8}$$

To solve the differential equation, we rewrite it in the separated form

$$\frac{du}{u^2} = dt, \qquad \text{and then integrate both sides:} \qquad -\frac{1}{u} = \int \frac{du}{u^2} = t + k.$$

Solving the resulting algebraic equation for u, we deduce the solution formula

$$u = -\frac{1}{t+k}\,. \tag{2.9}$$

To specify the integration constant k, we evaluate u at the initial time t_0; this implies

$$u_0 = -\frac{1}{t_0+k}\,, \qquad \text{so that} \qquad k = -\frac{1}{u_0} - t_0.$$

Therefore, the solution to the initial value problem is

$$u = \frac{u_0}{1 - u_0(t-t_0)}\,. \tag{2.10}$$

Figure 1 shows the graphs of some typical solutions.

As t approaches the critical value $t^\star = t_0 + 1/u_0$ from below, the solution "blows up", meaning $u(t) \to \infty$ as $t \to t^\star$. The blow-up time $t^\star$ depends upon the initial data — the larger $u_0 > 0$ is, the sooner the solution goes off to infinity. If the initial data is negative, $u_0 < 0$, the solution is well-defined for all $t > t_0$, but has a singularity in the past, at $t^\star = t_0 + 1/u_0 < t_0$. The only solution that exists for all positive and negative time is the constant solution $u(t) \equiv 0$, corresponding to the initial condition $u_0 = 0$.

In general, the constant *equilibrium solutions* to an autonomous ordinary differential equation, also known as its *fixed points*, play a distinguished role. If $u(t) \equiv u^\star$ is a constant solution, then $du/dt \equiv 0$, and hence the differential equation (2.3) implies that $F(u^\star) = 0$. Therefore, the equilibrium solutions coincide with the *roots* of the function $F(u)$. In point of fact, since we divided by $F(u)$, the derivation of our formula for the solution (2.7) assumed that we were *not* at an equilibrium point. In the preceding example, our final solution formula (2.10) happens to include the equilibrium solution $u(t) \equiv 0$, corresponding to $u_0 = 0$, but this is a lucky accident. Indeed, the equilibrium solution does *not* appear in the "general" solution formula (2.9). One must typically take extra care that equilibrium solutions do not elude us when utilizing this basic integration method.

Example 2.2. Although a population of people, animals, or bacteria consists of individuals, the aggregate behavior can often be effectively modeled by a dynamical system that involves continuously varying variables. As first proposed by the English economist Thomas Malthus in 1798, the population of a species grows, roughly, in proportion to its size. Thus, the number of individuals $N(t)$ at time t satisfies a first order differential equation of the form

$$\frac{dN}{dt} = \rho N, \tag{2.11}$$

where the proportionality factor $\rho = \beta - \delta$ measures the rate of growth, namely the difference between the birth rate $\beta \geq 0$ and the death rate $\delta \geq 0$. Thus, if births exceed deaths, $\rho > 0$, and the population increases, whereas if $\rho < 0$, more individuals are dying and the population shrinks.

In the very simplest model, the growth rate ρ is assumed to be independent of the population size, and (2.11) reduces to a simple linear ordinary differential equation whose solutions satisfy the Malthusian exponential growth law $N(t) = N_0 \, e^{\rho t}$, where $N_0 = N(0)$ is the initial population size. Thus, if $\rho > 0$, the population grows without limit, while if $\rho < 0$, the population dies out, so $N(t) \to 0$ as $t \to \infty$, at an exponentially fast rate. The Malthusian population model provides a reasonably accurate description of the behavior of an isolated population in an environment with unlimited resources.

In a more realistic scenario, the growth rate will depend upon the size of the population as well as external environmental factors. For example, in the presence of limited resources, relatively small populations will increase, whereas an excessively large population will have insufficient resources to survive, and so its growth rate will be negative. In other words, the growth rate $\rho(N) > 0$ when $N < N^\star$, while $\rho(N) < 0$ when $N > N^\star$, where the *carrying capacity* $N^\star > 0$ depends upon the resource availability. The simplest class of functions that satifies these two inequalities are of the form $\rho(N) = \mu(N^\star - N)$, where $\mu > 0$ is a positive constant. This leads us to the nonlinear population model

$$\frac{dN}{dt} = \mu \, N \, (N^\star - N). \tag{2.12}$$

In deriving this model, we assumed that the environment is not changing over time; a dynamical environment would require a more complicated non-autonomous differential equation.

Before analyzing the solutions to the nonlinear population model, let us make a preliminary change of variables, and set $u(t) = N(t)/N^\star$, so that u represents the size of the population in proportion to the *carrying capacity* $N^\star$. A straightforward computation shows that $u(t)$ satisfies the so-called *logistic differential equation*

$$\frac{du}{dt} = \lambda \, u \, (1-u), \qquad\qquad u(0) = u_0, \tag{2.13}$$

where $\lambda = N^\star \mu$, and, for simplicity, we assign the initial time to be $t_0 = 0$. The logistic differential equation can be viewed as the continuous counterpart of the logistic map studied in my Notes on Nonlinear Systems. However, unlike its discrete namesake, the logistic differential equation is quite sedate, and its solutions easily understood.

First, there are two equilibrium solutions: $u(t) \equiv 0$ and $u(t) \equiv 1$, obtained by setting the right hand side of the equation equal to zero. The first represents a nonexistent population with no individuals and hence no reproduction. The second equilibrium solution corresponds to a static population $N(t) \equiv N^\star$ that is at the ideal size for the environment, so deaths exactly balance births. In all other situations, the population size will vary over time.

To integrate the logistic differential equation, we proceed as above, first writing it in the separated form

$$\frac{du}{u(1-u)} = \lambda \, dt.$$

Integrating both sides, and using partial fractions,

$$\lambda \, t + k = \int \frac{du}{u(1-u)} = \int \left[\frac{1}{u} + \frac{1}{1-u} \right] du = \log \left| \frac{u}{1-u} \right| ,$$

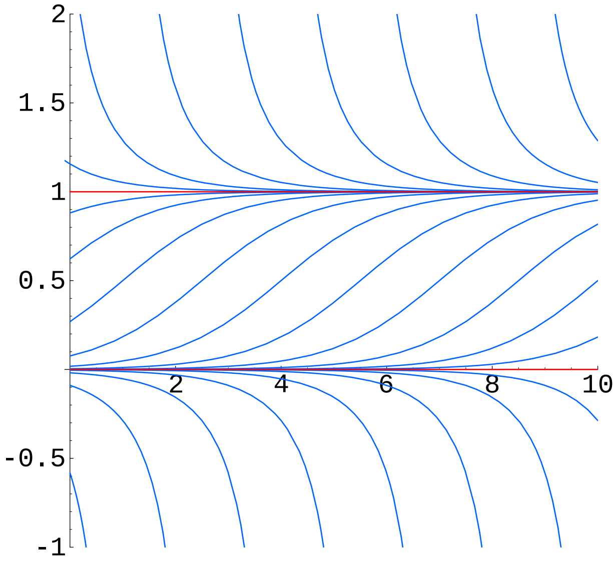


Figure 2. Solutions to $u' = u(1-u)$.

where k is a constant of integration. Therefore

$$\frac{u}{1-u} = c\,e^{\lambda t}, \qquad \text{where} \qquad c = \pm e^{k}.$$

Solving for u, we deduce the solution

$$u(t) = \frac{c\,e^{\lambda t}}{1 + c\,e^{\lambda t}}\,. \tag{2.14}$$

The constant of integration is fixed by the initial condition. Solving the algebraic equation

$$u_0 = u(0) = \frac{c}{1+c} \qquad \text{yields} \qquad c = \frac{u_0}{1-u_0}\,.$$

Substituting the result back into the solution formula (2.14) and simplifying, we find

$$u(t) = \frac{u_0\,e^{\lambda t}}{1 - u_0 + u_0\,e^{\lambda t}}\,. \tag{2.15}$$

The resulting solutions are illustrated in Figure 2. Interestingly, while the equilibrium solutions are not covered by the integration method, they reappear in the final solution formula, corresponding to initial data $u_0 = 0$ and $u_0 = 1$ respectively. However, this is a lucky accident, and cannot be anticipated in more complicated situations.

When using the logistic equation to model population dynamics, the initial data is assumed to be positive, $u_0 > 0$. As time $t \to \infty$, the solution (2.15) tends to the equilibrium value $u(t) \to 1$ — which corresponds to $N(t) \to N^\star$ approaching the carrying capacity in the original population model. For small initial values $u_0 \ll 1$ the solution initially grows at an exponential rate λ, corresponding to a population with unlimited resources. However, as the population increases, the gradual lack of resources tends to slow down the growth rate, and eventually the population saturates at the equilibrium value. On the other hand, if $u_0 > 1$, the population is too large to be sustained by the available resources, and so dies off until it reaches the same saturation value. If $u_0 = 0$, then the solution remains at equilibrium $u(t) \equiv 0$. Finally, when $u_0 < 0$, the solution only exists for a finite amount of time, with

$$u(t) \longrightarrow -\infty \qquad \text{as} \qquad t \longrightarrow t^\star = \frac{1}{\lambda} \log\left(1 - \frac{1}{u_0}\right).$$

Of course, this final case does appear in the physical world, since we cannot have a negative population!

The separation of variables method used to solve autonomous equations can be straightforwardly extended to a special class of non-autonomous equations. A *separable* ordinary differential equation has the form

$$\frac{du}{dt} = a(t)\, F(u), \tag{2.16}$$

in which the right hand side is the product of a function of t and a function of u. To solve the equation, we rewrite it in the separated form

$$\frac{du}{F(u)} = a(t)\, dt.$$

Integrating both sides leads to the solution in implicit form

$$G(u) = \int \frac{du}{F(u)} = \int a(t)\, dt = A(t) + k. \tag{2.17}$$

The integration constant k is then fixed by the initial condition. And, as before, one must properly account for any equilibrium solutions, when $F(u) = 0$.

Example 2.3. Let us solve the particular initial value problem

$$\frac{du}{dt} = (1 - 2t)\, u, \qquad u(0) = 1. \tag{2.18}$$

We begin by writing the differential equation in separated form

$$\frac{du}{u} = (1 - 2t)\, dt.$$

Integrating both sides leads to

$$\log u = \int \frac{du}{u} = \int (1 - 2t)\, dt = t - t^2 + k,$$

where k is the constant of integration. We can readily solve for

$$u(t) = c\, e^{t - t^2},$$

where $c = \pm e^k$. The latter formula constitutes the general solution to the differential equation, and happens to include the equilibrium solution $u(t) \equiv 0$ when $c = 0$. The given initial condition requires that $c = 1$, and hence $u(t) = e^{t - t^2}$ is the unique solution to the initial value problem. The solution is graphed in Figure 3.

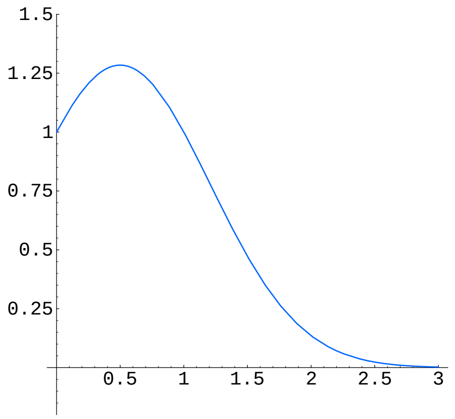


Figure 3. Solution to the Initial Value Problem $\dot{u} = (1 - 2t)\, u$, $u(0) = 1$.

First Order Systems

A *first order system of ordinary differential equations* has the general form

$$\frac{du_1}{dt} = F_1(t, u_1, \ldots, u_n), \qquad \cdots \qquad \frac{du_n}{dt} = F_n(t, u_1, \ldots, u_n). \tag{2.19}$$

The unknowns $u_1(t), \ldots, u_n(t)$ are scalar functions of the real variable t, which usually represents time. We shall write the system more compactly in vector form

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}), \tag{2.20}$$

where $\mathbf{u}(t) = (\, u_1(t), \ldots, u_n(t)\,)^T$, and $\mathbf{F}(t, \mathbf{u}) = (\, F_1(t, u_1, \ldots, u_n), \ldots, F_n(t, u_1, \ldots, u_n)\,)^T$ is a vector-valued function of $n + 1$ variables. By a *solution* to the differential equation, we mean a vector-valued function $\mathbf{u}(t)$ that is defined and continuously differentiable on an interval $a < t < b$, and, moreover, satisfies the differential equation on its interval of definition. Each solution $\mathbf{u}(t)$ serves to parametrize a curve $C \subset \mathbb{R}^n$, also known as a *trajectory* or *orbit* of the system.

In this chapter, we shall concentrate on initial value problems for such first order systems. The general initial conditions are

$$u_1(t_0) = a_1, \qquad u_2(t_0) = a_2, \qquad \cdots \qquad u_n(t_0) = a_n, \tag{2.21}$$

or, in vectorial form,

$$\mathbf{u}(t_0) = \mathbf{a} \tag{2.22}$$

Here t_0 is a prescribed initial time, while the vector $\mathbf{a} = (\, a_1, a_2, \ldots, a_n\,)^T$ fixes the initial position of the desired solution. In favorable situations, as described below, the initial conditions serve to uniquely specify a solution to the differential equations — at least for nearby times. The general issues of existence and uniqueness of solutions will be addressed in the following section.

A system of differential equations is called *autonomous* if the right hand side does not explicitly depend upon the time t, and so takes the form

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}). \tag{2.23}$$

One important class of autonomous first order systems are the steady state fluid flows. Here $\mathbf{F}(\mathbf{u}) = \mathbf{v}$ represents the fluid velocity vector field at the position $\mathbf{u}$. The solution $\mathbf{u}(t)$ to the initial value problem (2.23, 22) describes the motion of a fluid particle that starts at position $\mathbf{a}$ at time t_0. The differential equation tells us that the fluid velocity at each point on the particle's trajectory matches the prescribed vector field.

An *equilibrium solution* is constant: $\mathbf{u}(t) \equiv \mathbf{u}^\star$ for all t. Thus, its derivative must vanish, $d\mathbf{u}/dt \equiv \mathbf{0}$, and hence, every equilibrium solution arises as a solution to the system of algebraic equations

$$\mathbf{F}(\mathbf{u}^\star) = \mathbf{0} \tag{2.24}$$

prescribed by the vanishing of the right hand side of the system (2.23).

Example 2.4. A *predator-prey system* is a simplified ecological model of two species: the predators which feed on the prey. For example, the predators might be lions roaming the Serengeti and the prey zebra. We let $u(t)$ represent the number of prey, and $v(t)$ the number of predators at time t. Both species obey a population growth model of the form (2.11), and so the dynamical equations can be written as

$$\frac{du}{dt} = \rho u, \qquad\qquad \frac{dv}{dt} = \sigma v, \tag{2.25}$$

where the growth rates ρ, σ may depend upon the other species. The more prey, i.e., the larger u is, the faster the predators reproduce, while a lack of prey will cause them to die off. On the other hand, the more predators, the faster the prey are consumed and the slower their net rate of growth.

If we assume that the environment has unlimited resources for the prey, which, barring drought, is probably valid in the case of the zebras, then the simplest model that incorporates these assumptions is the *Lotka–Volterra system*

$$\frac{du}{dt} = \alpha u - \delta u v, \qquad\qquad \frac{dv}{dt} = -\beta v + \gamma u v, \tag{2.26}$$

corresponding to growth rates $\rho = \alpha - \delta v$, $\sigma = -\beta + \gamma u$. The parameters $\alpha, \beta, \gamma, \delta > 0$ are all positive, and their precise values will depend upon the species involved and how they interact, as indicated by field data, combined with, perhaps, educated guesses. In particular, α represents the unrestrained growth rate of the prey in the absence of predators, while $-\beta$ represents the rate that the predators die off in the absence of their prey. The nonlinear terms model the interaction of the two species: the rate of increase in the predators is proportional to the number of available prey, while the rate of decrease in the prey is proportional to the number of predators. The initial conditions $u(t_0) = u_0$, $v(t_0) = v_0$ represent the initial populations of the two species.

We will discuss the integration of the Lotka–Volterra system (2.26) in Section 4. Here, let us content ourselves with determining the possible equilibria. Setting the right hand sides of the system to zero leads to the nonlinear algebraic system

$$0 = \alpha u - \delta u v = u(\alpha - \delta v), \qquad 0 = -\beta v + \gamma u v = v(-\beta + \gamma u).$$

Thus, there are two distinct equilibria, namely

$$u_1^\star = v_1^\star = 0, \qquad u_2^\star = \beta/\gamma, \quad v_2^\star = \alpha/\delta.$$

The first is the uninteresting (or, rather catastrophic) situation where there are no animals — no predators and no prey. The second is a nontrivial solution in which both populations maintain a steady value, for which the birth rate of the prey is precisely sufficient to continuously feed the predators. Is this a feasible solution? Or, to state the question more mathematically, is this a stable equilibrium? We shall develop the tools to answer this question below.

Higher Order Systems

A wide variety of physical systems are modeled by nonlinear systems of differential equations depending upon second and, occasionally, even higher order derivatives of the unknowns. But there is an easy device that will reduce any higher order ordinary differential equation or system to an equivalent first order system. "Equivalent" means that each solution to the first order system uniquely corresponds to a solution to the higher order equation and vice versa. The upshot is that, for all practical purposes, one only needs to analyze first order systems. Moreover, the vast majority of numerical solution algorithms are designed for first order systems, and so to numerically integrate a higher order equation, one must place it into an equivalent first order form.

We have already encountered the main idea in our discussion of the phase plane approach to second order scalar equations

$$\frac{d^2u}{dt^2} = F\left(t, u, \frac{du}{dt}\right). \tag{2.27}$$

We introduce a new dependent variable $v = \dfrac{du}{dt}$. Since $\dfrac{dv}{dt} = \dfrac{d^2u}{dt^2}$, the functions u, v satisfy the equivalent first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = F(t, u, v). \tag{2.28}$$

Conversely, it is easy to check that if $\mathbf{u}(t) = \left(\, u(t), v(t)\,\right)^T$ is any solution to the first order system, then its first component $u(t)$ defines a solution to the scalar equation, which establishes their equivalence. The basic initial conditions $u(t_0) = u_0$, $v(t_0) = v_0$, for the first order system translate into a pair of initial conditions $u(t_0) = u_0$, $\dot u(t_0) = v_0$, specifying the value of the solution and its first order derivative for the second order equation.

Similarly, given a third order equation

$$\frac{d^3u}{dt^3} = F\left(t, u, \frac{du}{dt}, \frac{d^2u}{dt^2}\right),$$

we set

$$v = \frac{du}{dt}\,, \qquad\qquad w = \frac{dv}{dt} = \frac{d^2u}{dt^2}\,.$$

The variables u, v, w satisfy the equivalent first order system

$$\frac{du}{dt} = v, \qquad\qquad \frac{dv}{dt} = w, \qquad\qquad \frac{dw}{dt} = F(t,u,v,w).$$

The general technique should now be clear.

Example 2.5. The forced *van der Pol equation*

$$\frac{d^2u}{dt^2} + (u^2-1)\,\frac{du}{dt} + u = f(t) \tag{2.29}$$

arises in the modeling of an electrical circuit with a triode whose resistance changes with the current. It also arises in certain chemical reactions and wind-induced motions of structures. To convert the van der Pol equation into an equivalent first order system, we set $v = du/dt$, whence

$$\frac{du}{dt} = v, \qquad\qquad \frac{dv}{dt} = f(t) - (u^2-1)\,v - u, \tag{2.30}$$

is the equivalent phase plane system.

Example 2.6. The Newtonian equations for a mass m moving in a potential force field are a second order system of the form

$$m\,\frac{d^2\mathbf{u}}{dt^2} = -\nabla F(\mathbf{u})$$

in which $\mathbf{u}(t) = \big(\,u(t), v(t), w(t)\,\big)^T$ represents the position of the mass, while $F(\mathbf{u}) = F(u,v,w)$ is the potential function. In components,

$$m\,\frac{d^2u}{dt^2} = -\,\frac{\partial F}{\partial u}\,, \qquad m\,\frac{d^2v}{dt^2} = -\,\frac{\partial F}{\partial v}\,, \qquad m\,\frac{d^2w}{dt^2} = -\,\frac{\partial F}{\partial w}\,. \tag{2.31}$$

For example, a planet moving in the sun's gravitational field satisfies the Newtonian system for the gravitational potential

$$F(\mathbf{u}) = -\,\frac{\alpha}{\|\,\mathbf{u}\,\|} = -\,\frac{\alpha}{\sqrt{u^2+v^2+w^2}}\,, \tag{2.32}$$

where α depends on the masses and the universal gravitational constant. (This simplified model ignores any additional interplanetary forces.) Thus, the mass' motion in such a gravitational force field follows the solution to the second order Newtonian system

$$m\,\frac{d^2\mathbf{u}}{dt^2} = -\,\nabla F(\mathbf{u}) = -\,\frac{\alpha\,\mathbf{u}}{\|\,\mathbf{u}\,\|^3} = \frac{\alpha}{(u^2+v^2+w^2)^{3/2}}\begin{pmatrix} u \\ v \\ w \end{pmatrix}.$$

The same system of ordinary differential equations describes the motion of a charged particle in a Coulomb electric force field, where the sign of α is positive for attracting opposite charges, and negative for repelling like charges.

To convert the second order Newton equations into a first order system, we set $\mathbf{v} = \dot{\mathbf{u}}$ to be the mass' velocity vector, with components

$$p = \frac{du}{dt}\,, \qquad q = \frac{dv}{dt}\,, \qquad r = \frac{dw}{dt}\,,$$

and so

$$\begin{aligned} \frac{du}{dt} &= p, & \frac{dv}{dt} &= q, & \frac{dw}{dt} &= r, \\ \frac{dp}{dt} &= -\,\frac{1}{m}\,\frac{\partial F}{\partial u}\,(u,v,w), & \frac{dq}{dt} &= -\,\frac{1}{m}\,\frac{\partial F}{\partial v}\,(u,v,w), & \frac{dr}{dt} &= -\,\frac{1}{m}\,\frac{\partial F}{\partial w}\,(u,v,w). \end{aligned} \tag{2.33}$$

One of Newton's greatest acheivements was to solve this system in the case of the central gravitational potential (2.32), and thereby confirm the validity of Kepler's laws of planetary motion.

Finally, we note that there is a simple device that will convert any non-autonomous system into an equivalent autonomous system involving one additional variable. Namely, one introduces an extra coordinate $u_0 = t$ to represent the time, which satisfies the elementary differential equation $du_0/dt = 1$ with initial condition $u_0(t_0) = t_0$. Thus, the original system (2.19) can be written in the autonomous form

$$\frac{du_0}{dt} = 1, \qquad \frac{du_1}{dt} = F_1(u_0,u_1,\ldots,u_n), \qquad \cdots \qquad \frac{du_n}{dt} = F_n(u_0,u_1,\ldots,u_n). \tag{2.34}$$

For example, the autonomous form of the forced van der Pol system (2.30) is

$$\frac{du_0}{dt} = 1, \qquad \frac{du_1}{dt} = u_2, \qquad \frac{du_2}{dt} = f(u_0) - (u_1^2 - 1)u_2 - u_1, \tag{2.35}$$

in which u_0 represents the time variable.

3. Existence, Uniqueness, and Continuous Dependence.

It goes without saying that there is no general analytical method that will solve all differential equations. Indeed, even relatively simple first order, scalar, non-autonomous ordinary differential equations cannot be solved in closed form. For example, the solution to the particular *Riccati equation*

$$\frac{du}{dt} = u^2 + t \tag{3.1}$$

cannot be written in terms of elementary functions, although it can be solved in terms of Airy functions, [**25**]. The *Abel equation*

$$\frac{du}{dt} = u^3 + t \tag{3.2}$$

fares even worse, since its general solution cannot be written in terms of even standard special functions — although power series solutions can be tediously ground out term by term. Understanding when a given differential equation can be solved in terms of elementary functions or known special functions is an active area of contemporary research, [**3**]. In this vein, we cannot resist mentioning that the most important class of exact solution techniques for differential equations are those based on symmetry. An introduction can be found in the author's graduate level monograph [**26**]; see also [**5**, **16**].

Existence

Before worrying about how to solve a differential equation, either analytically, qualitatively, or numerically, it behooves us to try to resolve the core mathematical issues of existence and uniqueness. First, does a solution exist? If, not, it makes no sense trying to find one. Second, is the solution uniquely determined? Otherwise, the differential equation probably has scant relevance for physical applications since we cannot use it as a predictive tool. Since differential equations inevitably have lots of solutions, the only way in which we can deduce uniqueness is by imposing suitable initial (or boundary) conditions.

Unlike partial differential equations, which must be treated on a case-by-case basis, there are complete general answers to both the existence and uniqueness questions for initial value problems for systems of ordinary differential equations. (Boundary value problems are more subtle.) While obviously important, we will not take the time to present the proofs of these fundamental results, which can be found in most advanced textbooks on the subject, including [**2**, **13**, **15**, **17**].

Let us begin by stating the Fundamental Existence Theorem for initial value problems associated with first order systems of ordinary differential equations.

Theorem 3.1. *Let $\mathbf{F}(t,\mathbf{u})$ be a continuous function. Then the initial value problem*[†]

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t,\mathbf{u}), \qquad\qquad \mathbf{u}(t_0) = \mathbf{a}, \tag{3.3}$$

admits a solution $\mathbf{u} = \mathbf{f}(t)$ that is, at least, defined for nearby times, i.e., when $|\,t - t_0\,| < \delta$ for some $\delta > 0$.

Theorem 3.1 guarantees that the solution to the initial value problem exists — at least for times sufficiently close to the initial instant t_0. This may be the most that can be said, although in many cases the maximal interval $\alpha < t < \beta$ of existence of the solution might be much larger — possibly infinite, $-\infty < t < \infty$, resulting in a *global solution*. The interval of existence of a solution typically depends upon both the equation and the particular initial data. For instance, even though its right hand side is defined everywhere, the solutions to the scalar initial value problem (2.8) only exist up until time $1/u_0$, and so, the larger the initial data, the shorter the time of existence. In this example, the only global solution is the equilibrium solution $u(t) \equiv 0$. It is worth noting that this short-term

[†] If $\mathbf{F}(t,\mathbf{u})$ is only defined on a subdomain $\Omega \subset \mathbb{R}^{n+1}$, then we must assume that the point $(t_0, \mathbf{a}) \in \Omega$ specifying the initial conditions belongs to its domain of definition.

existence phenomenon does not appear in the linear regime, where, barring singularities in the equation itself, solutions to a linear ordinary differential equation are guaranteed to exist for all time.

In practice, one always extends a solutions to its maximal interval of existence. The Existence Theorem 3.1 implies that there are only two possible ways in which a solution cannot be extended beyond a time $t^\star$: Either

(i) the solution becomes unbounded: $\| \mathbf{u}(t) \| \to \infty$ as $t \to t^\star$, or

(ii) if the right hand side $F(t, \mathbf{u})$ is only defined on a subset $\Omega \subset \mathbb{R}^{n+1}$, then the solution $\mathbf{u}(t)$ reaches the boundary $\partial\Omega$ as $t \to t^\star$.

If neither occurs in finite time, then the solution is necessarily global. In other words, a solution to an ordinary differential equation cannot suddenly vanish into thin air.

Remark: The existence theorem can be readily adapted to any higher order system of ordinary differential equations through the method of converting it into an equivalent first order system by introducing additional variables. The appropriate initial conditions guaranteeing existence are induced from those of the corresponding first order system, as in the second order example (2.27) discussed above.

Uniqueness and Smoothness

As important as existence is the question of uniqueness. Does the initial value problem have more than one solution? If so, then we cannot use the differential equation to predict the future behavior of the system from its current state. While continuity of the right hand side of the differential equation will guarantee that a solution exists, it is not quite sufficient to ensure uniqueness of the solution to the initial value problem. The difficulty can be appreciated by looking at an elementary example.

Example 3.2. Consider the nonlinear initial value problem

$$\frac{du}{dt} = \frac{5}{3}\, u^{2/5}, \qquad\qquad u(0) = 0. \tag{3.4}$$

Since the right hand side is a continuous function, Theorem 3.1 assures us of the existence of a solution — at least for t close to 0. This autonomous scalar equation can be easily solved by the usual method:

$$\int \frac{3}{5}\,\frac{du}{u^{2/5}} = u^{3/5} = t + c, \qquad\text{and so}\qquad u = (t+c)^{5/3}.$$

Substituting into the initial condition implies that $c = 0$, and hence $u(t) = t^{5/3}$ is a solution to the initial value problem.

On the other hand, since the right hand side of the differential equation vanishes at $u = 0$, the constant function $u(t) \equiv 0$ is an equilibrium solution to the differential equation. (Here is an example where the integration method fails to recover the equilibrium solution.) Moreover, the equilibrium solution has the same initial value $u(0) = 0$. Therefore, we have constructed two different solutions to the initial value problem (3.4). Uniqueness is *not*

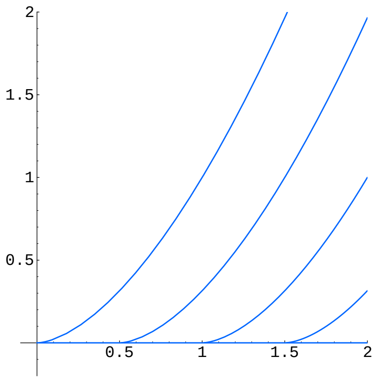


Figure 4. Solutions to the Differential Equation $\dot{u} = \frac{5}{3}\,u^{2/5}$.

valid! Worse yet, there are, in fact, an *infinite* number of solutions to the initial value problem. For *any* $a > 0$, the function

$$u(t) = \begin{cases} 0, & 0 \le t \le a, \\ (t-a)^{5/3}, & t \ge a, \end{cases} \tag{3.5}$$

is differentiable everywhere, even at $t = a$. (Why?) Moreover, it satisfies both the differential equation and the initial condition, and hence defines a solution to the initial value problem. Several of these solutions are plotted in Figure 4.

Thus, to ensure uniqueness of solutions, we need to impose a more stringent condition, beyond mere continuity. The proof of the following basic uniqueness theorem can be found in the above references.

Theorem 3.3. *If* $\mathbf{F}(t,\mathbf{u}) \in \mathrm{C}^1$ *is continuously differentiable, then there exists one and only one solution*[†] *to the initial value problem* (3.3)*.*

Thus, the difficulty with the differential equation (3.4) is that the function $F(u) = \frac{5}{3}\,u^{2/5}$, although continuous everywhere, is not differentiable at $u = 0$, and hence the Uniqueness Theorem 3.3 does not apply. On the other hand, $F(u)$ is continuously differentiable away from $u = 0$, and so any nonzero initial condition $u(t_0) = u_0 \ne 0$ will produce a unique solution — for as long as it remains away from the problematic value $u = 0$.

Blanket Hypothesis: From now on, all differential equations must satisfy the uniqueness criterion that their right hand side is continuously differentiable.

While continuous differentiability is sufficient to guarantee uniqueness of solutions, the smoother the right hand side of the system, the smoother the solutions. Specifically:

Theorem 3.4. *If* $\mathbf{F} \in \mathrm{C}^n$ *for* $n \ge 1$*, then any solution to the system* $\dot{\mathbf{u}} = \mathbf{F}(t,\mathbf{u})$ *is of class* $\mathbf{u} \in \mathrm{C}^{n+1}$*. If* $\mathbf{F}(t,\mathbf{u})$ *is an analytic function, then all solutions* $\mathbf{u}(t)$ *are analytic.*

[†] As noted earlier, we extend all solutions to their maximal interval of existence.

The basic outline of the proof of the first result is clear: Continuity of $\mathbf{u}(t)$ (which is a basic prerequisite of any solution) implies continuity of $\mathbf{F}(t, \mathbf{u}(t))$, which means $\dot{\mathbf{u}}$ is continuous and hence $\mathbf{u} \in \mathrm{C}^1$. This in turn implies $\mathbf{F}(t, \mathbf{u}(t)) = \dot{\mathbf{u}}$ is a continuously differentiable of t, and so $\mathbf{u} \in \mathrm{C}^2$. And so on, up to order n. The proof of analyticity follows from a detailed analysis of the power series solutions, [**14**]. Indeed, the analytic result underlies the method of power series solutions of ordinary differential equations, [**2, 13**].

Uniqueness has a number of particularly important consequences for the solutions to autonomous systems, i.e., those whose right hand side does not explicitly depend upon t. Throughout the remainder of this section, we will deal with an autonomous system of ordinary differential equations

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}), \qquad \text{where} \qquad \mathbf{F} \in \mathrm{C}^1, \tag{3.6}$$

whose right hand side is defined and continuously differentiable for all $\mathbf{u}$ in a domain $\Omega \subset \mathbb{R}^n$. As a consequence, each solution $\mathbf{u}(t)$ is, on its interval of existence, uniquely determined by its initial data. Autonomy of the differential equation is an essential hypothesis for the validity of the following properties.

The first result tells us that the solution trajectories of an autonomous system do not vary over time.

Proposition 3.5. *If $\mathbf{u}(t)$ is the solution to the autonomous system* (3.6) *with initial condition $\mathbf{u}(t_0) = \mathbf{u}_0$, then the solution to the initial value problem $\widetilde{\mathbf{u}}(t_1) = \mathbf{u}_0$ is $\widetilde{\mathbf{u}}(t) = \mathbf{u}(t - t_1 + t_0)$.*

Proof: Let $\widetilde{\mathbf{u}}(t) = \mathbf{u}(t - t_1 + t_0)$, where $\mathbf{u}(t)$ is the original solution. In view of the chain rule and the fact that t_1 and t_0 are fixed,

$$\frac{d}{dt}\,\widetilde{\mathbf{u}}(t) = \frac{d\mathbf{u}}{dt}\,(t - t_1 + t_0) = \mathbf{F}(\mathbf{u}(t - t_1 + t_0)) = \mathbf{F}(\widetilde{\mathbf{u}}(t)),$$

and hence $\widetilde{\mathbf{u}}(t)$ is also a solution to the system (3.6). Moreover,

$$\widetilde{\mathbf{u}}(t_1) = \mathbf{u}(t_0) = \mathbf{u}_0$$

has the indicated initial conditions, and hence, by uniqueness, must be the one and only solution to the latter initial value problem. *Q.E.D.*

Note that the two solutions $\mathbf{u}(t)$ and $\widetilde{\mathbf{u}}(t)$ parametrize the *same* curve in $\mathbb{R}^n$, differing only by an overall “phase shift”, $t_1 - t_0$, in their parametrizations. Thus, all solutions passing through the point $\mathbf{u}_0$ follow the same trajectory, irrespective of the time they arrive there. Indeed, not only is the trajectory the same, but the solutions have identical speeds at each point along the trajectory curve. For instance, if the right hand side of (3.6) represents the velocity vector field of steady state fluid flow, Proposition 3.5 implies that the stream lines — the paths followed by the individual fluid particles — do not change in time, even though the fluid itself is in motion. This, indeed, is the meaning of the term “steady state” in fluid mechanics.

One particularly important consequence of uniqueness is that a solution $\mathbf{u}(t)$ to an autonomous system is either stuck at an equilibrium for all time, or is always in motion. In other words, either $\dot{\mathbf{u}} \equiv \mathbf{0}$, in the case of equilibrium, or, otherwise, $\dot{\mathbf{u}} \neq \mathbf{0}$ wherever defined.

Proposition 3.6. *Let $\mathbf{u}^\star$ be an equilibrium for the autonomous system* (3.6), *so $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$. If $\mathbf{u}(t)$ is any solution such that $\mathbf{u}(t^\star) = \mathbf{u}^\star$ at some time $t^\star$, then $\mathbf{u}(t) \equiv \mathbf{u}^\star$ is the equilibrium solution.*

Proof: We regard $\mathbf{u}(t^\star) = \mathbf{u}^\star$ as initial data for the given solution $\mathbf{u}(t)$ at the initial time $t^\star$. Since $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$, the constant function $\mathbf{u}^\star(t) \equiv \mathbf{u}^\star$ is a solution of the differential equation that satisfies the same initial conditions. Therefore, by uniqueness, it coincides with the solution in question. *Q.E.D.*

In other words, it is mathematically impossible for a solution to reach an equilibrium position in a finite amount of time — although it may well approach equilibrium in an asymptotic fashion as $t \to \infty$; see Proposition 3.9 below for details. Physically, this observation has the interesting and physically counterintuitive consequence that a mathematical system never actually attains an equilibrium position! Even at very large times, there is always some very slight residual motion. In practice, though, once the solution gets sufficiently close to equilibrium, we are unable to detect the motion, and the physical system has, in all but name, reached its stationary equilibrium configuration. And, of course, the inherent motion of the atoms and molecules not included in such a simplified model would hide any infinitesimal residual effects of the mathematical solution. Without uniqueness, the result is false. For example, the function $u(t) = (t - t^\star)^{5/3}$ is a solution to the scalar ordinary differential equation (3.4) that reaches the equilibrium point $u^\star = 0$ in a finite time $t = t^\star$.

Continuous Dependence

In a real-world applications, initial conditions are almost never known exactly. Rather, experimental and physical errors will only allow us to say that their values are approximately equal to those in our mathematical model. Thus, to retain physical relevance, we need to be sure that small errors in our initial measurements do not induce a large change in the solution. A similar argument can be made for any physical parameters, e.g., masses, charges, spring stiffnesses, frictional coefficients, etc., that appear in the differential equation itself. A slight change in the parameters should not have a dramatic effect on the solution.

Mathematically, what we are after is a criterion of *continuous dependence* of solutions upon both initial data and parameters. Fortunately, the desired result holds without any additional assumptions, beyond requiring that the parameters appear continuously in the differential equation. We state both results in a single theorem.

Theorem 3.7. *Consider an initial value problem problem*

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}, \boldsymbol{\mu}), \qquad \mathbf{u}(t_0) = \mathbf{a}(\boldsymbol{\mu}), \tag{3.7}$$

in which the differential equation and/or the initial conditions depend continuously on one or more parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$*. Then the unique*[†] *solution* $\mathbf{u}(t, \boldsymbol{\mu})$ *depends continuously upon the parameters.*

Example 3.8. Let us look at a perturbed version

$$\frac{du}{dt} = \alpha\, u^2, \qquad\qquad u(0) = u_0 + \varepsilon,$$

of the initial value problem that we considered in Example 2.1. We regard ε as a small perturbation of our original initial data u_0, and α as a variable parameter in the equation. The solution is

$$u(t, \varepsilon) = \frac{u_0 + \varepsilon}{1 - \alpha\,(u_0 + \varepsilon)\, t}\,. \tag{3.8}$$

Note that, where defined, this is a continuous function of both parameters α, ε. Thus, a small change in the initial data, or in the equation, produces a small change in the solution — at least for times near the initial time.

Continuous dependence *does not* preclude nearby solutions from eventually becoming far apart. Indeed, the blow-up time $t^\star = 1/\big[\,\alpha\,(u_0 + \varepsilon)\,\big]$ for the solution (3.8) depends upon both the initial data and the parameter in the equation. Thus, as we approach the singularity, solutions that started out very close to each other will get arbitrarily far apart; see Figure 1 for an illustration.

An even simpler example is the linear model of exponential growth $\dot{u} = \alpha u$ when $\alpha > 0$. A very tiny change in the initial conditions has a negligible short term effect upon the solution, but over longer time intervals, the differences between the two solutions will be dramatic. Thus, the "sensitive dependence" of solutions on initial conditions already appears in very simple linear equations. For similar reasons, continuous dependence does *not* prevent solutions from exhibiting chaotic behavior. Further development of these ideas can be found in [**1**, **8**] and elsewhere.

As an application, let us show that if a solution to an autonomous system converges to a single limit point, then that point is necessarily an equilibrium solution. Keep in mind that, owing to uniqueness of solutions, the limiting equilibrium cannot be mathematically achieved in finite time, but only as a limit as time goes to infinity.

Proposition 3.9. *Let* $\mathbf{u}(t)$ *be a solution to the autonomous system* $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$*, with* $\mathbf{F} \in \mathrm{C}^1$*, such that* $\displaystyle\lim_{t\to\infty} \mathbf{u}(t) = \mathbf{u}^\star$*. Then* $\mathbf{u}^\star$ *is an equilibrium solution, and so* $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$*.*

Proof: Let $\mathbf{v}(t, \mathbf{a})$ denote the solution to the initial value problem $\dot{\mathbf{v}} = \mathbf{F}(\mathbf{v}),\ \mathbf{v}(0) = \mathbf{a}$. (We use a different letter to avoid confusion with the given solution $\mathbf{u}(t)$.) Theorem 3.7 implies that $\mathbf{v}(t, \mathbf{a})$ is a continuous function of the initial position $\mathbf{a}$. and hence $\mathbf{v}(t, \mathbf{u}(s))$ is a continuous function of $s \in \mathbb{R}$. Since $\displaystyle\lim_{s\to\infty} \mathbf{u}(s) = \mathbf{u}^\star$, we have

$$\lim_{s\to\infty} \mathbf{v}(t, \mathbf{u}(s)) = \mathbf{v}(t, \mathbf{u}^\star).$$

[†] We continue to impose our blanket uniqueness hypothesis.

On the other hand, since the system is autonomous, Proposition 3.5 implies that $\mathbf{v}(t,\mathbf{u}(s)) = \mathbf{u}(t+s)$, and hence

$$\lim_{s \to \infty} \mathbf{v}(t,\mathbf{u}(s)) = \lim_{s \to \infty} \mathbf{u}(t+s) = \mathbf{u}^\star.$$

Equating the preceding two limit equations, we conclude that $\mathbf{v}(t,\mathbf{u}^\star) = \mathbf{u}^\star$ for all t, and hence the solution with initial value $\mathbf{v}(0) = \mathbf{u}^\star$ is an equilibrium solution. *Q.E.D.*

The same conclusion holds if we run time backwards: if $\lim_{t \to -\infty} \mathbf{u}(t) = \mathbf{u}_\star$, then $\mathbf{u}_\star$ is also an equilibrium point. When they exist, solutions that start and end at equilibrium points play a particularly role in the dynamics, and are known as *heteroclinic*, or, if the start and end equilibria are the same, *homoclinic orbits*. Of course, limiting equilibrium points are but one of the possible long term behaviors of solutions to nonlinear ordinary differential equations, which can also become unbounded in finite or infinite time, or approach periodic orbits, known as *limit cycles*, or become completely chaotic, depending upon the nature of the system and the initial conditions. Resolving the long term behavior os solutions is one of the many challenges awaiting the detailed analysis of any nonlinear ordinary differential equation.

4. Stability.

Once a solution to a system of ordinary differential equations has settled down, its limiting value is an equilibrium solution; this is the content of Proposition 3.9. However, not all equilibria appear in this fashion. The only steady state solutions that one directly observes in a physical system are the stable equilibria. Unstable equilibria are hard to sustain, and will disappear when subjected to even the tiniest perturbation, e.g., a breath of air, or outside traffic jarring the experimental apparatus. Thus, finding the equilibrium solutions to a system of ordinary differential equations is only half the battle; one must then understand their stability properties in order to characterize those that can be realized in normal physical circumstances.

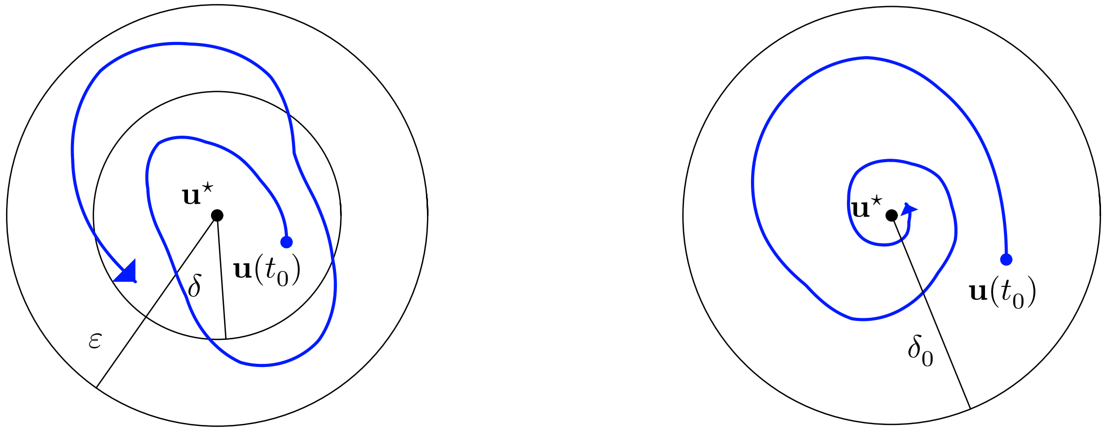
We will focus our attention on autonomous systems

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$$

whose right hand sides are at least continuously differentiable, so as to ensure the uniqueness of solutions to the initial value problem. If *every* solution that starts out near a given equilibrium solution tends to it, the equilibrium is called *asymptotically stable*. If the solutions that start out nearby stay nearby, then the equilibrium is *stable*. More formally:

Definition 4.1. An equilibrium solution $\mathbf{u}^\star$ to an autonomous system of first order ordinary differential equations is called

- *stable* if for every (small) $\varepsilon > 0$, there exists a $\delta > 0$ such that every solution $\mathbf{u}(t)$ having initial conditions within distance $\delta > \| \mathbf{u}(t_0) - \mathbf{u}^\star \|$ of the equilibrium remains within distance $\varepsilon > \| \mathbf{u}(t) - \mathbf{u}^\star \|$ for all $t \geq t_0$.
- *asymptotically stable* if it is stable and, in addition, there exists $\delta_0 > 0$ such that whenever $\delta_0 > \| \mathbf{u}(t_0) - \mathbf{u}^\star \|$, then $\mathbf{u}(t) \to \mathbf{u}^\star$ as $t \to \infty$.



Stability

Asymptotic Stability

Figure 5. Stability of Equilibria.

Thus, although solutions nearby a stable equilibrium may drift slightly farther away, they must remain relatively close. In the case of asymptotic stability, they will eventually return to equilibrium. This is illustrated in Figure 5

Example 4.2. As we saw, the logistic differential equation

$$\frac{du}{dt} = \lambda\, u\,(1-u)$$

has two equilibrium solutions, corresponding to the two roots of the quadratic equation $\lambda\, u(1-u) = 0$. The solution graphs in Figure 1 illustrate the behavior of the solutions. Observe that the first equilibrium solution $u_1^\star = 0$ is unstable, since all nearby solutions go away from it at an exponentially fast rate. On the other hand, the other equilibrium solution $u_2^\star = 1$ is asymptotically stable, since any solution with initial condition $0 < u_0$ tends to it, again at an exponentially fast rate.

Example 4.3. Consider an autonomous (meaning constant coefficient) homogeneous linear planar system

$$\frac{du}{dt} = a\,u + b\,v, \qquad\qquad \frac{dv}{dt} = c\,u + d\,v,$$

with coefficient matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The origin $u^\star = v^\star = 0$ is an evident equilibrium, solution, and, moreover, is the only equilibrium provided A is nonsingular. According to the results in [**27**; Section 9.3], the stability of the origin depends upon the eigenvalues of A: It is (globally) asymptotically stable if and only if both eigenvalues are real and negative, and is stable, but not asymptotically stable if and only if both eigenvalues are purely imaginary, or if 0 is a double eigenvalue and so $A = \mathrm{O}$. In all other cases, the origin is an unstable equilibrium. Later, we will see how this simple linear analysis has a direct bearing on the stability question for nonlinear planar systems.

Before looking at any further examples, we need to develop some basic mathematical tools for investigating the stability of equilibria. We begin at the beginning. The stability analysis for first order scalar ordinary differential equations

$$\frac{du}{dt} = F(u) \tag{4.1}$$

is particularly easy. The first observation is that all non-equilibrium solutions $u(t)$ are *strictly monotone* functions, meaning they are either always increasing or always decreasing. Indeed, when $F(u) > 0$, then (4.1) implies that the derivative $\dot{u} > 0$, and hence $u(t)$ is increasing at such a point. Vice versa, solutions are decreasing at any point where $F(u) < 0$. Since $F(u(t))$ depends continuously on t, any non-monotone solution would have to pass through an equilibrium value where $F(u^\star) = 0$, in violation of Proposition 3.6. This proves the claim.

As a consequence of monotonicity, there are only three possible behaviors for a non-equilibrium solution:

(a) it becomes unbounded at some finite time: $|\,u(t)\,| \to \infty$ as $t \to t^\star$; or

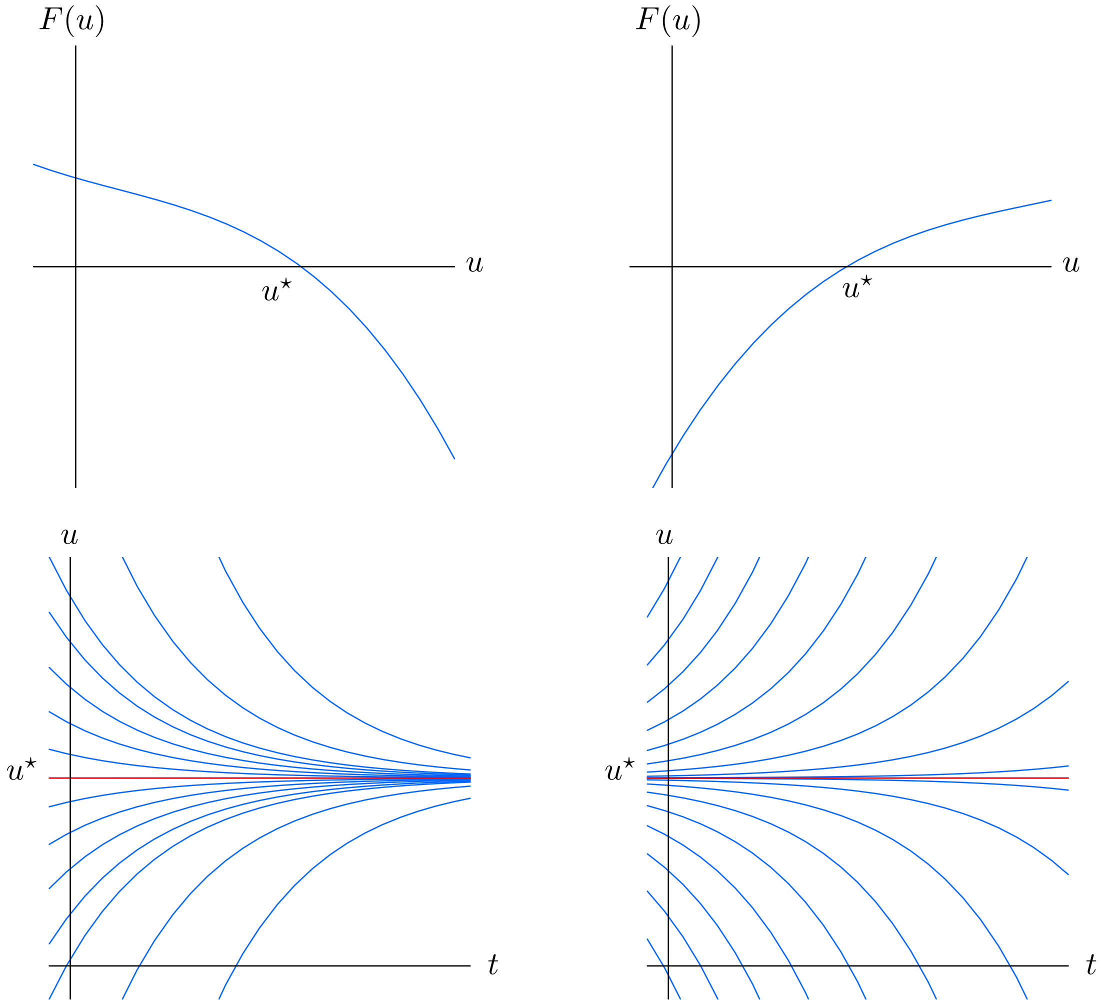
(b) it exists for all $t \geq t_0$, but becomes unbounded as $t \to \infty$; or

(c) it exists for all $t \geq t_0$ and has a limiting value, $u(t) \to u^\star$ as $t \to \infty$, which, by Proposition 3.9 must be an equilibrium point.

Let us look more carefully at the last eventuality. Suppose $u^\star$ is an equilibrium point, so $F(u^\star) = 0$. Suppose that $F(u) > 0$ for all u lying slightly below $u^\star$, i.e., on an interval of the form $u^\star - \delta < u < u^\star$. Any solution $u(t)$ that starts out on this interval, $u^\star - \delta < u(t_0) < u^\star$ must be increasing. Moreover, $u(t) < u^\star$ for all t since, according to Proposition 3.6, the solution cannot pass through the equilibrium point. Therefore, $u(t)$ is a solution of type (c). It must have limiting value $u^\star$, since by assumption, this is the only equilibrium solution it can increase to. Therefore, in this situation, the equilibrium point $u^\star$ is *asymptotically stable from below*: solutions that start out slightly below return to it in the limit. On the other hand, if $F(u) < 0$ for all u slightly below $u^\star$, then any solution that starts out in this regime will be monotonically decreasing, and so will move downwards, away from the equilibrium point, which is thus *unstable from below*.

By the same reasoning, if $F(u) < 0$ for u slightly above $u^\star$, then solutions starting out there will be monotonically decreasing, bounded from below by $u^\star$, and hence have no choice but to tend to $u^\star$ in the limit. Under this condition, the equilibrium point is *asymptotically stable from above*. The reverse inequality, $F(u) > 0$, corresponds to solutions that increase away from $u^\star$, which is hence *unstable from above*. Combining the two stable cases produces the basic asymptotic stability criterion for scalar ordinary differential equations.

Theorem 4.4. *A equilibrium point $u^\star$ of an autonomous scalar differential equation is asymptotically stable if and only if $F(u) > 0$ for $u^\star - \delta < u < u^\star$ and $F(u) < 0$ for $u^\star < u < u^\star + \delta$, for some $\delta > 0$.*



Stable Equilibrium

Unstable Equilibrium

Figure 6. Equilibria of Scalar Ordinary Differential Equations.

In other words, if $F(u)$ switches sign from positive to negative as u increases through the equilibrium point, then the equilibrium is asymptotically stable. If the inequalities are reversed, and $F(u)$ goes from negative to positive, then the equilibrium point is unstable. The two cases are illustrated in Figure 6. An equilibrium point where $F(u)$ is of one sign on both sides, e.g., the point $u^\star = 0$ for $F(u) = u^2$, is stable from one side, and unstable from the other.

Example 4.5. Consider the differential equation

$$\frac{du}{dt} = u - u^3. \tag{4.2}$$

Solving the algebraic equation $F(u) = u - u^3 = 0$, we find that the equation has three equilibria: $u_1^\star = -1$, $u_2^\star = 0$, $u_3^\star = +1$, As u increases, the graph of the function $F(u) = u - u^3$ switches from positive to negative at the first equilibrium point $u_1^\star = -1$, which proves its stability. Similarly, the graph goes back from negative to positive at $u_2^\star = 0$,

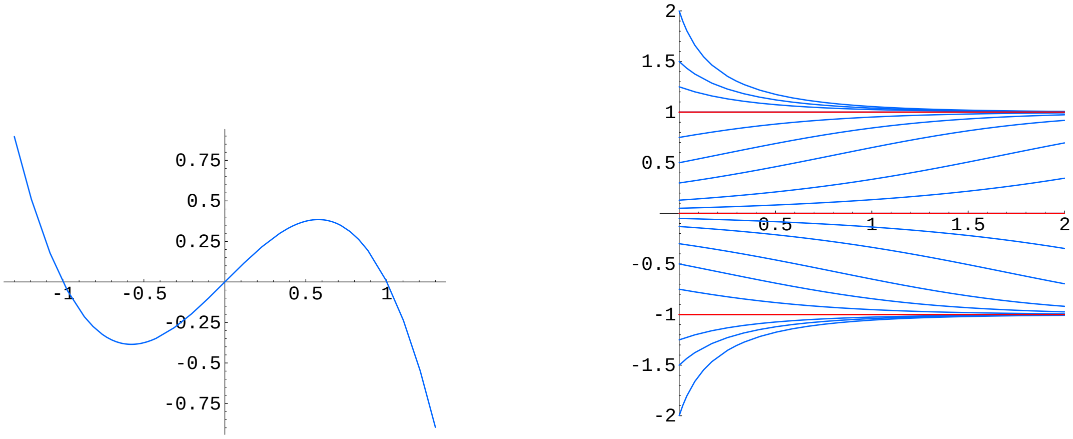


Figure 7. Stability of $\dot{u} = u - u^3$.

establishing the instability of the second equilibrium. The final equilibrium $u_3^\star = +1$ is stable because $F(u)$ again changes from positive to negative there.

With this information in hand, we are able to completely characterize the behavior of all solutions to the system. Any solution with negative initial condition, $u_0 < 0$, will end up, asymptotically, at the first equilibrium, $u(t) \to -1$ as $t \to \infty$. Indeed, if $u_0 < -1$, then $u(t)$ is monotonically increasing to -1, while if $-1 < u_0 < 0$, the solution is decreasing towards -1. On the other hand, if $u_0 > 0$, the corresponding solution ends up at the other stable equilibrium, $u(t) \to +1$; those with $0 < u_0 < 1$ are monotonically increasing, while those with $u_0 > 1$ are decreasing. The only solution that does not end up at either -1 or $+1$ as $t \to \infty$ is the unstable equilibrium solution $u(t) \equiv 0$. Any perturbation of it, no matter how tiny, will force the solutions to choose one of the stable equilibria. Representative solutions are plotted in Figure 7. Note that all the curves, with the sole exception of the horizontal axis, converge to one of the stable solutions ±1, and diverge from the unstable solution 0 as $t \to \infty$.

Thus, the sign of the function $F(u)$ nearby an equilibrium determines its stability. In most instances, this can be checked by looking at the derivative of the function at the equilibrium. If $F'(u^\star) < 0$, then we are in the stable situation, where $F(u)$ goes from positive to negative with increasing u, whereas if $F'(u^\star) > 0$, then the equilibrium $u^\star$ unstable on both sides.

Theorem 4.6. *Let $u^\star$ be a equilibrium point for a scalar ordinary differential equation $\dot{u} = F(u)$. If $F'(u^\star) < 0$, then $u^\star$ is asymptotically stable. If $F'(u^\star) > 0$, then $u^\star$ is unstable.*

For instance, in the preceding example,

$$F'(u) = 1 - 3\,u^2,$$

and its value at the equilibria are

$$F'(-1) = -2 < 0, \qquad F'(0) = 1 > 0, \qquad F'(1) = -2 < 0.$$

The signs reconfirm our conclusion that ±1 are stable equilibria, while 0 is unstable.

In the borderline case when $F'(u^\star) = 0$, the derivative test is inconclusive, and further analysis is needed to resolve the status of the equilibrium point. For example, the equations $\dot{u} = u^3$ and $\dot{u} = -u^3$ both satisfy $F'(0) = 0$ at the equilibrium point $u^\star = 0$. But, according to the criterion of Theorem 4.4, the former has an unstable equilibrium, while the latter's is stable. Thus, Theorem 4.6 is not as powerful as the direct algebraic test in Theorem 4.4. But it does have the advantage of being a bit easier to use. More significantly, unlike the algebraic test, it can be directly generalized to systems of ordinary differential equations.

Linearization and Stability

In higher dimensional situations, we can no longer rely on simple monotonicity properties, and a more sophisticated approach to stability issues is required. The key idea is already contained in the second characterization of stable equilibria in Theorem 4.6. The derivative $F'(u^\star)$ determines the slope of the tangent line, which is a linear approximation to the function $F(u)$ near the equilibrium point. In a similar fashion, a vector-valued function $\mathbf{F}(\mathbf{u})$ is replaced by its linear approximation near an equilibrium point. The basic stability criteria for the resulting linearized differential equation were established in [**27**; Section 9.2]. and, in most situations, the linearized stability or instability carries over to the nonlinear regime.

Let us first revisit the scalar case

$$\frac{du}{dt} = F(u) \tag{4.3}$$

from this point of view. *Linearization* of a scalar function at a point means to replace it by its tangent line approximation

$$F(u) \approx F(u^\star) + F'(u^\star)(u - u^\star) \tag{4.4}$$

If $u^\star$ is an equilibrium point, then $F(u^\star) = 0$, and so the first term disappears. Therefore, we anticipate that, near the equilibrium point, the solutions to the nonlinear ordinary differential equation (4.3) will be well approximated by its linearization

$$\frac{du}{dt} = F'(u^\star)(u - u^\star).$$

Let us rewrite the linearized equation in terms of the deviation $v(t) = u(t) - u^\star$ of the solution from equilibrium. Since $u^\star$ is fixed, $dv/dt = du/dt$, and so the linearized equation takes the elementary form

$$\frac{dv}{dt} = a\,v, \qquad \text{where} \qquad a = F'(u^\star) \tag{4.5}$$

is the value of the derivative at the equilibrium point. Note that the original equilibrium point $u^\star$ corresponds to the zero equilibrium point $v^\star = 0$ of the linearized equation (4.5). We already know that the linear differential equation (4.5) has an asymptotically stable equilibrium at $v^\star = 0$ if and only if $a = F'(u^\star) < 0$, while for $a = F'(u^\star) > 0$ the origin is unstable. In this manner, the linearized stability criterion reproduces that established in Theorem 4.6.

The same linearization technique can be applied to analyze the stability of an equilibrium solution $\mathbf{u}^\star$ to a first order autonomous system

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u}). \tag{4.6}$$

We approximate the function $\mathbf{F}(\mathbf{u})$ near an equilibrium point, where $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$, by its first order Taylor polynomial:

$$\mathbf{F}(\mathbf{u}) \approx \mathbf{F}(\mathbf{u}^\star) + \mathbf{F}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) \ = \ \mathbf{F}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star). \tag{4.7}$$

Here, $\mathbf{F}'(\mathbf{u}^\star)$ denotes its $n \times n$ Jacobian matrix at the equilibrium point. Thus, for nearby solutions, we expect that the deviation from equilibrium, $\mathbf{v}(t) = \mathbf{u}(t) - \mathbf{u}^\star$, will be governed by the linearized system

$$\frac{d\mathbf{v}}{dt} = A\,\mathbf{v}, \qquad \text{where} \qquad A = \mathbf{F}'(\mathbf{u}^\star). \tag{4.8}$$

Now, we already know the complete stability criteria for linear systems, [**27**; Section 9.2]. The zero equilibrium solution to (4.8) is asymptotically stable if and only if all the eigenvalues of the coefficient matrix $A = \mathbf{F}'(\mathbf{u}^\star)$ have negative real part. In contrast, if one or more of the eigenvalues has positive real part, then the zero solution is unstable. Indeed, it can be proved, [**13, 15**], that these linearized stability criteria are also valid in the nonlinear case.

Theorem 4.7. *Let $\mathbf{u}^\star$ be an equilibrium point for the first order ordinary differential equation $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If all of the eigenvalues of the Jacobian matrix $\mathbf{F}'(\mathbf{u}^\star)$ have negative real part,* $\operatorname{Re}\lambda < 0$, *then $\mathbf{u}^\star$ is asymptotically stable. If, on the other hand, $\mathbf{F}'(\mathbf{u}^\star)$ has one or more eigenvalues with positive real part,* $\operatorname{Re}\lambda > 0$, *then $\mathbf{u}^\star$ is an unstable equilibrium.*

Intuitively, the additional nonlinear terms in the full system should only slightly perturb the eigenvalues, and hence, at least for those with nonzero real part, not alter their effect on the stability of solutions. The borderline case occurs when one or more of the eigenvalues of $\mathbf{F}'(\mathbf{u}^\star)$ is either 0 or purely imaginary, i.e., $\operatorname{Re}\lambda = 0$, while all other eigenvalues have negative real part. In such situations, the linearized stability test is inconclusive, and we need more detailed information (which may not be easy to come by) to resolve the status of the equilibrium.

Example 4.8. The second order ordinary differential equation

$$m\,\frac{d^2\theta}{dt^2} + \mu\,\frac{d\theta}{dt} + \kappa \sin\theta = 0 \tag{4.9}$$

describes the damped oscillations of a rigid pendulum that rotates on a pivot subject to a uniform gravitational force in the vertical direction. The unknown function $\theta(t)$ measures the angle of the pendulum from the vertical, as illustrated in Figure 8. The constant $m > 0$ is the mass of the pendulum bob, $\mu > 0$ is the coefficient of friction, assumed here to be strictly positive, and $\kappa > 0$ represents the gravitational force.

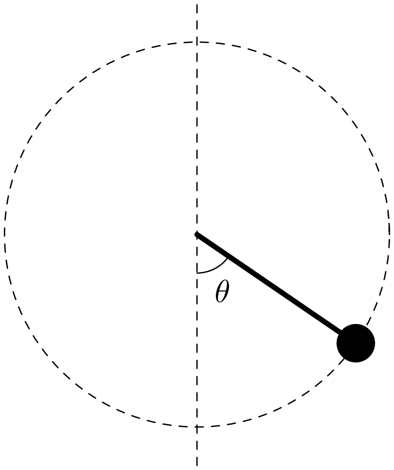


Figure 8. The Pendulum.

In order to study the equilibrium solutions and their stability, we must first convert the equation into a first order system. Setting $u(t) = \theta(t)$, $v(t) = \dfrac{d\theta}{dt}$, we find

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\,\alpha\,\sin u - \beta\,v, \qquad \text{where} \qquad \alpha = \frac{\kappa}{m}\,, \qquad \beta = \frac{\mu}{m}\,, \tag{4.10}$$

are both positive constants. The equilibria occur where the right hand sides of the first order system (4.10) simultaneously vanish, that is,

$$v = 0, \qquad -\,\alpha\,\sin u - \beta\,v = 0, \qquad \text{and hence} \qquad u = 0,\ \pm\pi,\ \pm 2\pi,\ \ldots\ .$$

Thus, the system has infinitely many equilibrium points:

$$\mathbf{u}_k^\star = (k\pi, 0) \qquad \text{where} \qquad k = 0, \pm 1, \pm 2, \ldots \quad \text{is any integer.} \tag{4.11}$$

The equilibrium point $\mathbf{u}_0^\star = (0,0)$ corresponds to $u = \theta = 0$, $v = \dot\theta = 0$, which means that the pendulum is at rest at the bottom of its arc. Our physical intuition leads us to expect this to describe a stable configuration, as the frictional effects will eventually damp out small nearby motions. The next equilibrium $\mathbf{u}_1^\star = (\pi, 0)$ corresponds to $u = \theta = \pi$, $v = \dot\theta = 0$, which means that the pendulum is sitting motionless at the top of its arc. This is a theoretically possible equilibrium configuration, but highly unlikely to be observed in practice, and is thus expected to be unstable. Now, since $u = \theta$ is an angular variable, equilibria whose u values differ by an integer multiple of 2π define the same physical configuration, and hence should have identical stability properties. Therefore, all the remaining equilibria $\mathbf{u}_k^\star$ physically correspond to one or the other of these two possibilities: when $k = 2j$ is even, the pendulum is at the bottom, while when $k = 2j+1$ is odd, the pendulum is at the top.

Let us now confirm our intuition by applying the linearization stability criterion of Theorem 4.7. The right hand side of the system (4.10), namely

$$\mathbf{F}(u,v) = \begin{pmatrix} v \\ -\,\alpha\,\sin u - \beta\,v \end{pmatrix}, \qquad \text{has Jacobian matrix} \qquad \mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ -\,\alpha\,\cos u & -\beta \end{pmatrix}.$$

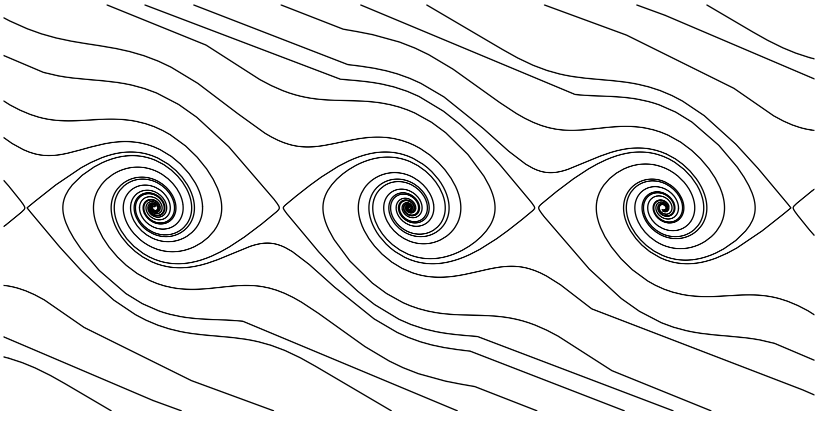


Figure 9. The Underdamped Pendulum.

At the bottom equilibrium $\mathbf{u}_0^\star = (0,0)$, the Jacobian matrix

$$\mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ -\alpha & -\beta \end{pmatrix} \qquad \text{has eigenvalues} \qquad \lambda = \frac{-\beta \,\pm\, \sqrt{\beta^2 - 4\,\alpha}}{2}\,.$$

Under our assumption that $\alpha, \beta > 0$, both eigenvalues have negative real part, and hence the origin is a stable equilibrium. If $\beta^2 < 4\,\alpha$ — the *underdamped* case — the eigenvalues are complex, and hence, in the terminology of [**27**; Section 9.3], the origin is a *stable focus*. In the phase plane, the solutions spiral in to the focus, which corresponds to a pendulum with damped oscillations of decreasing magnitude. On the other hand, if $\beta^2 > 4\,\alpha$, then the system is *overdamped*. Both eigenvalues are negative, and the origin is a *stable node*. In this case, the solutions decay exponentially fast to $\mathbf{0}$. Physically, this would be like a pendulum moving in a vat of molasses. The exact same analysis applies at all even equilibria $\mathbf{u}_{2j}^\star = (2\,j\,\pi, 0)$ — which really represent the same bottom equilibrium point.

On the other hand, at the top equilibrium $\mathbf{u}_1^\star = (\pi, 0)$, the Jacobian matrix

$$\mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ \alpha & -\beta \end{pmatrix} \qquad \text{has eigenvalues} \qquad \lambda = \frac{-\beta \,\pm\, \sqrt{\beta^2 + 4\,\alpha}}{2}\,.$$

In this case, one of the eigenvalues is real and positive while the other is negative. The linearized system has an unstable saddle point, and hence the nonlinear system is also unstable at this equilibrium point. Any tiny perturbation of an upright pendulum will dislodge it, causing it to swing down, and eventually settle into a damped oscillatory motion converging on one of the stable bottom equilibria.

The complete phase portrait of an underdamped pendulum appears in Figure 9. Note that, as advertised, almost all solutions end up spiraling into the stable equilibria. Solutions with a large initial velocity will spin several times around the center, but eventually the cumulative effect of frictional forces wins out and the pendulum ends up in a damped oscillatory mode. Each of the the unstable equilibria has the same saddle form as its linearizations, with two very special solutions, corresponding to the stable eigenline of the linearization, in which the pendulum spins around a few times, and, in the $t \to \infty$ limit, ends up standing upright at the unstable equilibrium position. However, like unstable

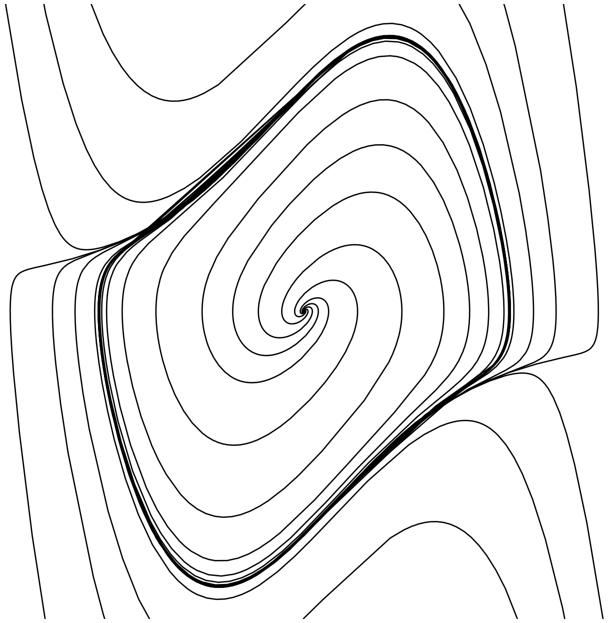


Figure 10. Phase Portrait of the van der Pol System.

equilibria, such solutions are practically impossible to achieve in a physical environment as any tiny perturbation will cause the pendulum to sightly deviate and then end up eventually decaying into the usual damped oscillatory motion at the bottom.

A deeper analysis demonstrates the local *structural stability* of any nonlinear equilibrium whose linearization is structurally stable, and hence has no eigenvalues on the imaginary axis: $\operatorname{Re} \lambda \neq 0$. Structural stability means that, not only are the stability properties of the equilibrium dictated by the linearized approximation, but, nearby the equilibrium point, all solutions to the nonlinear system are slight perturbations of solutions to the corresponding linearized system, and so, close to the equilibrium point, the two phase portraits have the same qualitative features. Thus, stable foci of the linearization remain stable foci of the nonlinear system; unstable saddle points remain saddle points, although the eigenlines become slightly curved as they depart from the equilibrium. Thus, the structural stability of linear systems, as discussed at the end of [**27**; Section 9.3] also carries over to the nonlinear regime near an equilibrium. A more in depth discussion of these issues can be found, for instance, in [**13**, **15**].

,

Example 4.9. Consider the unforced *van der Pol system*

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -(u^2 - 1)\, v - u. \tag{4.12}$$

that we derived in Example 2.5. The only equilibrium point is at the origin $u = v = 0$. Computing the Jacobian matrix of the right hand side,

$$\mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ 2\,u\,v - 1 & 1 \end{pmatrix}, \qquad \text{and hence} \qquad \mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

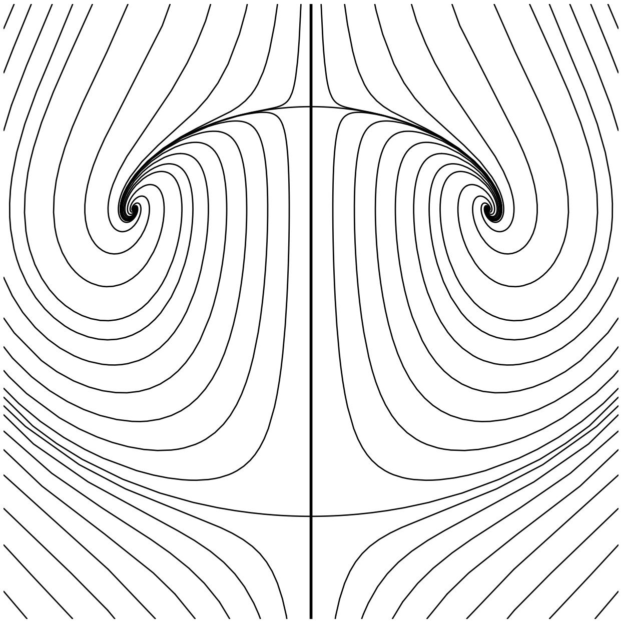


Figure 11. Phase Portrait for $\dot{u} = u\,(v-1),\ \dot{v} = 4 - u^2 - v^2.$.

The eigenvalues of $\mathbf{F}'(0,0)$ are $\frac{1}{2} \pm \mathrm{i}\,\frac{\sqrt{3}}{2}$, and correspond to an unstable focus of the linearized system near the equilibrium point. Therefore, the origin is an unstable equilibrium for nonlinear van der Pol system, and all non-equilibrium solutions starting out near $\mathbf{0}$ eventually spiral away.

On the other hand, it can be shown that solutions that are sufficiently far away from the origin spiral in towards the center. So what happens to the solutions? As illustrated in the phase plane portrait sketched in Figure 10, all non-equilibrium solutions spiral towards a stable periodic orbit, known as a *limit cycle* for the system. Any non-zero initial data will eventually end up closely following the limit cycle orbit as it periodically circles around the origin. A rigorous proof of the existence of a limit cycle relies on the more sophisticated *Poincaré–Bendixson Theory* for planar autonomous systems, discussed in detail in [**13**].

Example 4.10. The nonlinear system

$$\frac{du}{dt} = u\,(v-1), \qquad\qquad \frac{dv}{dt} = 4 - u^2 - v^2,$$

has four equilibria: $(0, \pm 2)$ and $(\pm\sqrt{3}\,, 1)$. Its Jacobian matrix is

$$\mathbf{F}'(u,v) = \begin{pmatrix} v-1 & u \\ -2\,u & -2\,v \end{pmatrix}.$$

A table of the eigenvalues at the equilibrium points and their stability follows. These results are reconfirmed by the phase portrait drawn in Figure 11.

Equilibrium Point	Jacobian matrix	Eigenvalues	Stability
$(0,2)$	$\begin{pmatrix} 1 & 0 \\ 0 & -4 \end{pmatrix}$	$1,\ -4$	unstable saddle
$(0,-2)$	$\begin{pmatrix} -3 & 0 \\ 0 & 6 \end{pmatrix}$	$-3, 6$	unstable saddle
$(\sqrt{3}\,,1)$	$\begin{pmatrix} 0 & -\sqrt{3} \\ 2\sqrt{3} & -2 \end{pmatrix}$	$-1 \pm \mathrm{i}\sqrt{5}$	stable focus
$(-\sqrt{3}\,,1)$	$\begin{pmatrix} 0 & -\sqrt{3} \\ 2\sqrt{3} & -2 \end{pmatrix}$	$-1 \pm \mathrm{i}\sqrt{5}$	stable focus

Conservative Systems

When modeling a physical system that includes some form of damping — due to friction, viscosity, or dissipation — linearization will usually suffice to resolve the stability or instability of equilibria. However, when dealing with conservative systems, when damping is absent and energy is preserved, the linearization test is often inconclusive, and one must rely on more sophisticated stability criteria. In such situations, one can often exploit conservation of energy, appealing to our general philosophy that minimizers of an energy function should be stable (but not necessarily asymptotically stable) equilibria.

By saying that energy is *conserved*, we mean that it remains constant as the solution evolves. Conserved quantities are also known as *first integrals* for the system of ordinary differential equations. Additional well-known examples include the laws of conservation of mass, and conservation of linear and angular momentum. Let us mathematically formulate the general definition.

Definition 4.11. A *first integral* of an autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ is a real-valued function $I(\mathbf{u})$ which is constant on solutions.

In other words, for each solution $\mathbf{u}(t)$ to the differential equation,

$$I(\mathbf{u}(t)) = c \qquad \text{for all} \qquad t, \tag{4.13}$$

where c is a fixed constant, which will depend upon which solution is being monitored. The value of c is fixed by the initial data since, in particular, $c = I(\mathbf{u}(t_0)) = I(\mathbf{u}_0)$. Or, to rephrase this condition in another, equivalent manner, every solution to the dynamical system is constrained to move along a single *level set* $\{ I(\mathbf{u}) = c \}$ of the first integral, namely the level set that contains the initial data $\mathbf{u}_0$.

Note first that any constant function, $I(\mathbf{u}) \equiv c_0$, is trivially a first integral, but this provides no useful information whatsoever about the solutions, and so is uninteresting. We will call any autonomous system that possesses a nontrivial first integral $I(\mathbf{u})$ a *conservative system*.

How do we find first integrals? In applications, one often appeals to the underlying physical principles such as conservation of energy, momentum, or mass. Mathematically, the most convenient way to check whether a function is constant is to verify that its derivative is identically zero. Thus, differentiating (4.13) with respect to t and invoking the chain rule leads to the basic condition

$$0 = \frac{d}{dt}\, I(\mathbf{u}(t)) = \nabla I(\mathbf{u}(t)) \cdot \frac{d\mathbf{u}}{dt} = \nabla I(\mathbf{u}(t)) \cdot \mathbf{F}(\mathbf{u}(t)). \tag{4.14}$$

The final expression can be identified as the directional derivative of $I(\mathbf{u})$ with respect to the vector field $\mathbf{v} = \mathbf{F}(\mathbf{u})$ that specifies the differential equation. Writing out (4.14) in detail, we find that a first integral $I(u_1, \ldots, u_n)$ must satisfy a first order linear partial differential equation:

$$F_1(u_1, \ldots, u_n)\, \frac{\partial I}{\partial u_1} + \;\cdots\; + F_n(u_1, \ldots, u_n)\, \frac{\partial I}{\partial u_n} = 0. \tag{4.15}$$

As such, it looks harder to solve than the original ordinary differential equation! Often, one falls back on either physical intuition, intelligent guesswork, or, as a last resort, a lucky guess. A deeper fact, due to the pioneering twentieth century mathematician Emmy Noether, cf. [**24**, **26**], is that first integrals and conservation laws are the result of underlying symmetry properties of the differential equation. Like many nonlinear methods, it remains the subject of contemporary research.

Let us specialize to planar autonomous systems

$$\frac{du}{dt} = F(u,v), \qquad\qquad \frac{dv}{dt} = G(u,v). \tag{4.16}$$

According to (4.15), any first integral $I(u,v)$ must satisfy the linear partial differential equation

$$F(u,v)\, \frac{\partial I}{\partial u} \;+\; G(u,v)\, \frac{\partial I}{\partial v} = 0. \tag{4.17}$$

This nonlinear first order partial differential equation can be solved by the method of characteristics. Consider the auxiliary first order scalar ordinary differential equation[†]

$$\frac{dv}{du} = \frac{G(u,v)}{F(u,v)} \tag{4.18}$$

for $v = h(u)$. Note that (4.18) can be formally obtained by dividing the second equation in the original system (4.16) by the first, and then canceling the time differentials dt. Suppose we can write the general solution to the scalar equation (4.18) in the implicit form

$$I(u,v) = c, \tag{4.19}$$

[†] We assume that $F(u,v) \not\equiv 0$. Otherwise, $I(u) = u$ is itself a first integral, and the system reduces to a scalar equation for v.

where c is a constant of integration. We claim that the function $I(u,v)$ is a first integral of the original system (4.16). Indeed, differentiating (4.19) with respect to u, and using the chain rule, we find

$$0 = \frac{d}{du}\, I(u,v) = \frac{\partial I}{\partial u} + \frac{dv}{du}\,\frac{\partial I}{\partial v} = \frac{\partial I}{\partial u} + \frac{G(u,v)}{F(u,v)}\,\frac{\partial I}{\partial v}\,.$$

Clearing the denominator, we conclude that $I(u,v)$ solves the partial differential equation (4.17), which justifies our claim.

Example 4.12. As an elementary example, consider the linear system

$$\frac{du}{dt} = -\,v, \qquad\qquad \frac{dv}{dt} = u. \tag{4.20}$$

To construct a first integral, we form the auxiliary equation (4.18), which is

$$\frac{dv}{du} = -\,\frac{u}{v}\,.$$

This first order ordinary differential equation can be solved by separating variables:

$$v\,dv = -\,u\,du, \qquad\quad \text{and hence} \qquad\quad \tfrac{1}{2}\,u^2 + \tfrac{1}{2}\,v^2 = c,$$

where c is the constant of integration. Therefore, by the preceding result,

$$I(u,v) = \tfrac{1}{2}\,u^2 + \tfrac{1}{2}\,v^2$$

is a first integral. The level sets of $I(u,v)$ are the concentric circles centered at the origin, and we recover the fact that the solutions of (4.20) go around the circles. The origin is a stable equilibrium — a center.

This simple example hints at the importance of first integrals in stability theory. The following key result confirms our general philosophy that energy minimizers, or, more generally, minimizers of first integrals, are stable equilibria.

Theorem 4.13. *Let $I(\mathbf{u})$ be a first integral for the autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If $\mathbf{u}^\star$ is a strict local extremum — minimum or mximum — of I, then $\mathbf{u}^\star$ is a stable equilibrium point for the system.*

Remark: At first sight, the fact that strict maxima are also stable equilibria appears to contradict our intuition. However, energy functions typically do not have local maxima. Indeed, physical energy is the sum of kinetic and potential contributions. While potential energy can admit maxima, e.g., the pendulum at the top of its arc, these are only unstable saddle points for the full energy function, since the kinetic energy can always be increased by moving a bit faster.

Proof: We first prove that $\mathbf{u}^\star$ is an equilibrium. Indeed, the solution $\mathbf{u}(t)$ with initial condition $\mathbf{u}(t_0) = \mathbf{u}^\star$ must maintain the value of $I(\mathbf{u}(t)) = I(\mathbf{u}^\star)$. But, by definition of a strict local minimum, $I(\mathbf{u}) > I(\mathbf{u}^\star)$ for all $\mathbf{u}$ near $\mathbf{u}^\star$, and hence, by continuity, the solution has no choice but to remain at the point $\mathbf{u}^\star$.

To prove stability, set

$$M(r) = \max \{\, I(\mathbf{u}) \mid \| \mathbf{u} - \mathbf{u}^\star \| \le r \,\}, \qquad m(r) = \min \{\, I(\mathbf{u}) \mid \| \mathbf{u} - \mathbf{u}^\star \| = r \,\}.$$

Thus, $M(r)$ is the maximum value of the first integral over a ball[†] of radius r centered at the minimum, while $m(r)$ is the minimum over its boundary sphere of radius r. Since I is continuous, so are m and M. Since $\mathbf{u}^\star$ is a strict local minimum, $M(r) \ge m(r) > m(0) = M(0) = I(\mathbf{u}^\star)$ for any $0 < r < \varepsilon$ sufficiently small.

For each $\varepsilon > 0$, we can choose a $\delta > 0$ such that $M(\delta) < m(\varepsilon)$. Then, whenever $\| \mathbf{u}(t_0) - \mathbf{u}^\star \| \le \delta$, then $I(\mathbf{u}(t)) = I(\mathbf{u}(t_0)) \le M(\delta) < m(\varepsilon)$. Since $m(\varepsilon)$ is the minimum possible value for $I(\mathbf{u})$ when $\| \mathbf{u}(t) - \mathbf{u}^\star \| = \varepsilon$, the solution $\mathbf{u}(t)$ cannot cross the sphere of radius ε at any t, and so $\| \mathbf{u}(t) - \mathbf{u}^\star \| < \varepsilon$ for all $t \ge t_0$. Hence, we have fulfilled the stability criteria of Definition 4.1. *Q.E.D.*

Example 4.14. Consider the specific predator-prey system

$$\frac{du}{dt} = 2u - uv, \qquad \frac{dv}{dt} = -9v + 3uv, \tag{4.21}$$

modeling populations of, say, lions and zebra, and a special case of (2.26). According to Example 2.4, there are two possible equilibria:

$$u_1^\star = v_1^\star = 0, \qquad u_2^\star = 3, \quad v_2^\star = 2.$$

Let us first try to determine their stability by linearization. The Jacobian matrix for the system is

$$\mathbf{F}'(u,v) = \begin{pmatrix} 2-v & -u \\ 3v & 3u-9 \end{pmatrix}.$$

At the first, trivial equilibrium,

$$\mathbf{F}'(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -9 \end{pmatrix}, \qquad \text{with eigenvalues } 2 \text{ and } -9.$$

Since there is one positive and one negative eigenvalue, the origin is an unstable saddle point. On the other hand, at the nonzero equilibrium, the Jacobian matrix

$$\mathbf{F}'(3,2) = \begin{pmatrix} 0 & -3 \\ 6 & 0 \end{pmatrix}, \qquad \text{has purely imaginary eigenvalues} \qquad \pm 3\sqrt{2}\,\mathrm{i}\,.$$

So the linearized system has a stable center. However, as purely imaginary eigenvalues is a borderline situation, Theorem 4.7 cannot be applied. Thus, the linearization stability test is *inconclusive*.

It turns out that the predator-prey model is a conservative system. To find a first integral, we need to solve the auxiliary equation (4.18), which is

$$\frac{dv}{du} = \frac{-9v + 3uv}{2u - uv} = \frac{-9/u + 3}{2/v - 1}\,.$$

[†] We write as if the norm is the Euclidean norm, but any other norm will work equally well for this proof.

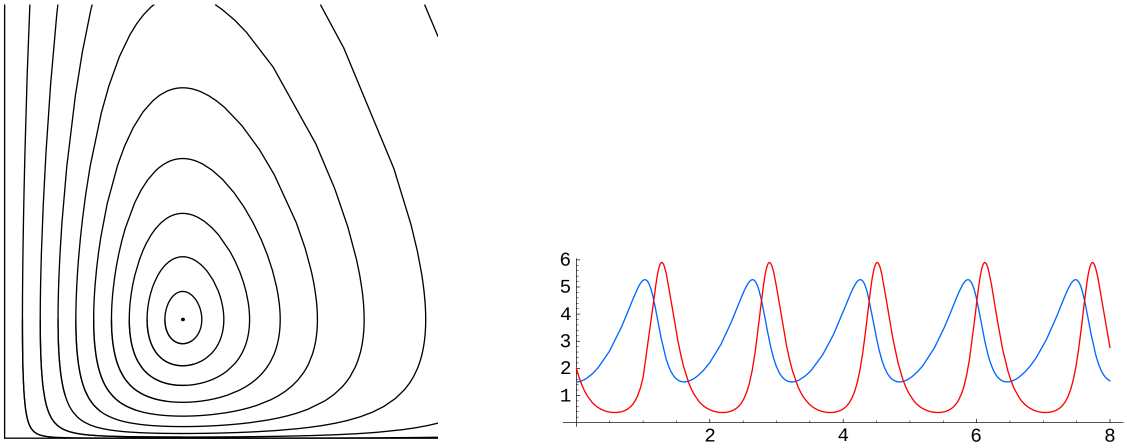


Figure 12. Phase Portrait and Solution of the Predator-Prey System.

Fortunately, this is a separable first order ordinary differential equation. Integrating,

$$2\,\log v - v = \int \left(\frac{2}{v} - 1 \right) dv = \int \left(-\frac{9}{u} + 3 \right) du = -9 \log u + 3\,u + c,$$

where c is the constant of integration. Writing the solution in the form (4.19), we conclude that

$$I(u,v) = 9\,\log u - 3\,u + 2\,\log v - v = c,$$

is a first integral of the system. The solutions to (4.21) must stay on the level sets of $I(u,v)$. Note that

$$\nabla I(u,v) = \begin{pmatrix} 9/u - 3 \\ 2/v - 1 \end{pmatrix}, \qquad \text{and hence} \qquad \nabla I(3,2) = \mathbf{0},$$

which shows that the second equilibrium is a critical point. (On the other hand, $I(u,v)$ is not defined at the unstable zero equilibrium.) Moreover, the Hessian matrix at the critical point,

$$\nabla^2 I(3,2) = \begin{pmatrix} -3 & 0 \\ 0 & -1 \end{pmatrix},$$

is negative definite, and hence $\mathbf{u}_2^\star = (\,3,2\,)^T$ is a strict local maximum of the first integral $I(u,v)$. Thus, Theorem 4.13 proves that the equilibrium point is a stable center.

The first integral serves to completely characterize the qualitative behavior of the system. In the physically relevant region, i.e., the upper right quadrant $Q = \{\,u > 0,\ \ v > 0\,\}$ where both populations are positive, all of the level sets of the first integral are closed curves encircling the equilibrium point $\mathbf{u}_2^\star = (\,3,2\,)^T$. The solutions move in a counterclockwise direction around the closed level curves, and hence all non-equilibrium solutions in the positive quadrant are periodic. The phase portrait is illustrated in Figure 12, along with a typical periodic solution. Thus, in such an idealized ecological model, for any initial conditions starting with some zebra and lions, i.e., where $u(t_0), v(t_0) > 0$, the populations will maintain a balance over the long term, each varying periodically between maximum and minimum values. Observe also that the maximum and minimum values of the two populations are not achieved simultaneously. Starting with a small number of predators, the number of prey will initially increase. The predators then have more food available, and so also start to increase in numbers. At a certain critical point, the predators are sufficiently numerous as to kill prey faster than they can reproduce. At this point, the prey population has reached its maximum, and begins to decline. But it takes a while for the predator population to feel the effect, and so it continues to increase. Eventually the increasingly rapid decline in the number of prey begins to affect the predators. After the predators reach their maximum number, both populations are in decline. Eventually, enough predators have died off so as to relieve the pressure on the prey, whose population bottoms out, and then slowly begins to rebound. Later, the number of predators also reaches a minimum, at which point the entire growth and decay cycle starts over again.

In contrast to a linear system, the period of the population cycle is not fixed, but depends upon how far away from the stable equilibrium the solution orbit lies. Near equilibrium, the solutions are close to those of the linearized system which, in view of its eigenvalues $\pm 3\,\mathrm{i}\,\sqrt{2}$, are periodic of frequency $3\sqrt{2}$ and period $\sqrt{2}\,\pi/3$. However, solutions that are far away from equilibrium have much longer periods, and so greater imbalances between predator and prey populations leads to longer periods, and more radically varying numbers. Understanding the mechanisms behind these population cycles is of increasing important in the ecological management of natural resources.

Example 4.15. In our next example, we look at the undamped oscillations of a pendulum. When we set the friction coefficient $\mu = 0$, the nonlinear second order ordinary differential equation (4.9) reduces to

$$m\,\frac{d^2\theta}{dt^2} + \kappa \sin\theta = 0. \tag{4.22}$$

As before, we convert this into a first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\,\alpha \sin u, \tag{4.23}$$

where

$$u(t) = \theta(t), \qquad v(t) = \frac{d\theta}{dt}\,, \qquad \alpha = \frac{\kappa}{m}\,.$$

The equilibria,

$$\mathbf{u}_k^\star = (k\,\pi, 0) \qquad \text{for} \qquad k = 0, \pm 1, \pm 2, \ldots\,,$$

are the same as in the damped case, i.e., the pendulum is either at the top (k even) or the bottom (k odd) of the circle.

Let us see what the linearization stability test tells us. In this case, the Jacobian matrix of (4.23) is

$$\mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ -\,\alpha\,\cos u & 0 \end{pmatrix}.$$

At the top equilibria

$$\mathbf{F}'(\mathbf{u}_{2j+1}^\star) = \mathbf{F}'\bigl((2j+1)\pi, 0\bigr) = \begin{pmatrix} 0 & 1 \\ \alpha & 0 \end{pmatrix} \qquad \text{has real eigenvalues} \qquad \pm\sqrt{\alpha}\,,$$

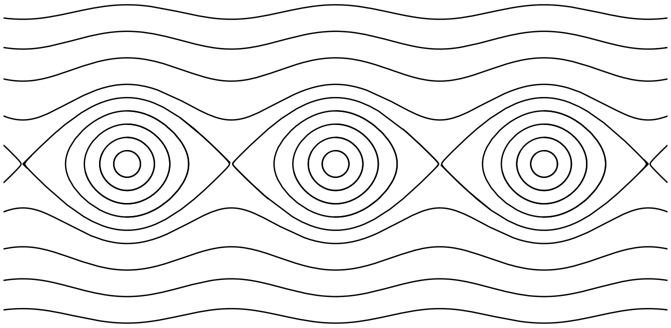


Figure 13. The Undamped Pendulum.

and hence these equilibria are unstable saddle points, just as in the damped version. On the other hand, at the bottom equilibria

$$\mathbf{F}'(\mathbf{u}_{2j}^\star) = \mathbf{F}'(2j\pi, 0) = \begin{pmatrix} 0 & 1 \\ -\alpha & 0 \end{pmatrix}, \quad \text{has purely imaginary eigenvalues} \quad \pm\,\mathrm{i}\,\sqrt{\alpha}\,.$$

Without the benefit of damping, the linearization test is inconclusive, and the stability of the bottom equilibria remains in doubt.

Since we are dealing with a conservative system, the total energy of the pendulum, namely

$$E(u,v) = \tfrac{1}{2}\,m v^2 + \kappa\,(1-\cos u) = \frac{m}{2}\left(\frac{d\theta}{dt}\right)^2 + \kappa\,(1-\cos\theta) \tag{4.24}$$

should provide us with a first integral. Note that E is a sum of two terms, which represent, respectively, the kinetic energy due to the pendulum's motion, and the potential energy[†] due to the height of the pendulum bob. To verify that $E(u,v)$ is indeed a first integral, we compute

$$\frac{dE}{dt} = \frac{\partial E}{\partial u}\,\frac{du}{dt} + \frac{\partial E}{\partial v}\,\frac{dv}{dt} = (\kappa \sin u)\,v + (m v)(-\,\alpha \sin u) = 0, \qquad \text{since} \qquad \alpha = \frac{\kappa}{m}\,.$$

Therefore, E is indeed constant on solutions, reconfirming the physical basis of the model.

The phase plane solutions to the pendulum equation will move along the level sets of the energy function $E(u,v)$, which are plotted in Figure 13. Its critical points are the equilibria, where

$$\nabla E(\mathbf{u}) = \begin{pmatrix} \kappa \sin u \\ m v \end{pmatrix} = \mathbf{0}, \qquad \text{and hence} \qquad \mathbf{u} = \mathbf{u}_k^\star = (\,k\pi, 0\,), \quad k = 0, \pm 1, \pm 2, \ldots\,.$$

To characterize the critical points, we appeal to the second derivative test, and so evaluate the Hessian

$$\nabla^2 E(u,v) = \begin{pmatrix} \kappa \cos u & 0 \\ 0 & m \end{pmatrix}.$$

[†] In a physical system, the potential energy is only defined up to an additive constant. Here we have fixed the zero energy level to be at the bottom of the pendulum's arc.

At the bottom equilibria,

$$\nabla^2 E(\mathbf{u}^\star_{2j}) = \nabla^2 E(2j\pi, 0) = \begin{pmatrix} \kappa & 0 \\ 0 & m \end{pmatrix}$$

is positive definite, since κ and m are positive constants. Therefore, the bottom equilibria are strict local minima of the energy, and so Theorem 4.13 guarantees their stability.

Each stable equilibrium is surrounded by a family of closed oval-shaped level curves, and hence forms a center. Each oval corresponds to a periodic solution[‡] of the system, in which the pendulum oscillates back and forth symmetrically around the bottom of its arc. Near the equilibrium, the period is close to that of the linearized system, namely $2\pi/\sqrt{\alpha}$ as prescribed by the eigenvalues of the Jacobian matrix. This fact underlies the use of pendulum-based clocks for keeping time, first recognized by Galileo. A grandfather clock is accurate because the amplitude of its pendulum's oscillations is kept relatively small. However, moving further away from the equilibrium point in the phase plane, we find that the periodic solutions with very large amplitude oscillations, in which the pendulum becomes nearly vertical, have much longer periods, and so would lead to inaccurate time-keeping.

The large amplitude limit of the periodic solutions is of particular interest. The pair of trajectories connecting two distinct unstable equilibria are known as the *homoclinic orbits*. Physically, a homoclinic orbit corresponds to a pendulum that starts out just shy of vertical, goes through exactly one full rotation, and eventually (as $t \to \infty$) ends up vertical again. The homoclinic orbits play an essential role in the analysis of the chaotic behavior of a periodically forced pendulum, [**1**, **8**, **23**].

Finally, the level sets lying above and below the "cat's-eyes" formed by the homoclinic and periodic orbits are known as the *running orbits*. Since $u = \theta$ is a 2π periodic angular variable, a running orbit solution $(\,u(t), v(t)\,)^T = (\theta(t), \dot\theta(t))^T$, in fact, also corresponds to a periodic physical motion, in which the pendulum spins around and around its pivot. The larger the total energy $E(u, v)$, the farther away from the u–axis the running orbit lies, and the faster the pendulum spins.

In summary, the qualitative behavior of a solution to the pendulum equation is almost entirely characterized by its energy:

• $E = 0$,	stable equilibria,
• $0 < E < 2\kappa$,	periodic oscillating orbits,
• $E = 2\kappa$,	unstable equilibria and homoclinic orbits,
• $E > 2\kappa$,	running orbits.

Example 4.16. The system governing the dynamical rotations of a rigid solid body around its center of mass are known as the *Euler equations* of rigid body mechanics, in honor of the prolific eighteenth century Swiss mathematician Leonhard Euler, cf. [**11**].

[‡] More precisely, a family of periodic solutions indexed by their initial condition on the oval, and differing only by a phase shift: $\mathbf{u}(t - \delta)$.

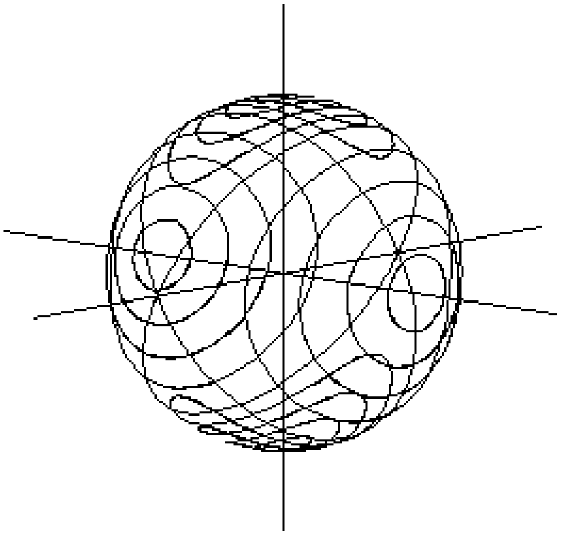


Figure 14. The Rigid Body Phase Portrait.

The eigenvectors of the positive definite inertia tensor of the body prescribe its three mutually orthogonal *principal axes*. The corresponding eigenvalues $I_1, I_2, I_3 > 0$ are called the *principal moments of inertia*. Let $u_1(t), u_2(t), u_3(t)$ denote the angular momenta of the body around its three principal axes. In the absence of external forces, the dynamical system governing the body's rotations around its center of mass takes the symmetric form

$$\frac{du_1}{dt} = \frac{I_2 - I_3}{I_2\, I_3}\, u_2 u_3, \qquad \frac{du_2}{dt} = \frac{I_3 - I_1}{I_1\, I_3}\, u_1 u_3, \qquad \frac{du_3}{dt} = \frac{I_1 - I_2}{I_1\, I_2}\, u_1 u_2. \tag{4.25}$$

Th Euler equations model, for example, the dynamics of a satellite spinning in its orbit above the earth. The solution will prescribe the angular motions of the satellite around its center of mass, but not the overall motion of the center of mass as the satellite orbits the earth.

Let us assume that the moments of inertia are all different, which we place in increasing order $0 < I_1 < I_2 < I_3$. The equilibria of the Euler system (4.25) are where the right hand sides simultaneously vanish, which requires that either $u_2 = u_3 = 0$, or $u_1 = u_3 = 0$, or $u_1 = u_2 = 0$. In other words, every point on the three coordinate axes is an equilibrium configuration. Since the variables represent angular momenta, these equilibria correspond to the body spinning around one of its principal axes at a fixed angular velocity.

Let us analyze the stability of these equilibrium configurations. The linearization test fails completely — as it must whenever dealing with a non-isolated equilibrium. But the Euler equations turn out to admit two independent first integrals:

$$E(\mathbf{u}) = \frac{1}{2}\left(\frac{u_1^2}{I_1} + \frac{u_2^2}{I_2} + \frac{u_3^2}{I_3}\right), \qquad A(\mathbf{u}) = \frac{1}{2}\left(u_1^2 + u_2^2 + u_3^2\right). \tag{4.26}$$

The first is the total kinetic energy, while the second is the total angular momentum. The proof that $dE/dt = 0 = dA/dt$ for any solution $\mathbf{u}(t)$ to (4.25) is left as an exercise for the reader.

Since both E and A are constant, the solutions to the system are constrained to move along a common level set $C = \{\, E = e,\ A = a \,\}$. Thus, the solution trajectories are the curves obtained by intersecting the sphere $S_a = \{\, A(\mathbf{u}) = a \,\}$ of radius $\sqrt{2a}$ with the ellipsoid $L_e = \{\, E(\mathbf{u}) = e \,\}$. In Figure 14, we have graphed the solution trajectories on a fixed sphere. (To see the figure, make sure the left hand periodic orbits are perceived on the back side of the sphere.) The six equilibria on the sphere are at its intersections with the coordinate axes. Those on the x and z axes are surrounded by closed periodic orbits, and hence are stable equilibria; indeed, they are, respectively, local minima and maxima of the energy when restricted to the sphere. On the other hand, the two equilibria on the y axis have the form of unstable saddle points, and are connected by four distinct homoclinic orbits. We conclude that a body that spins around either of its principal axes with the smallest or the largest moment of inertia is stable, whereas one that spins around the axis corresponding to the intermediate moment of inertia is unstable. This mathematical deduction can be demonstrated physically by flipping a solid rectangular object, e.g., this book, up into the air. It is easy to arrange it to spin around its long axis or its short axis in a stable manner, but it will balk at attempts to make it rotate around its middle axis!

Lyapunov's Method

Systems that incorporate damping, viscosity and/or frictional effects do not typically possess non-constant first integrals. From a physical standpoint, the damping will cause the total energy of the system to be a decreasing function of time. Asymptotically, the system returns to a (stable) equilibrium, and the extra energy has been dissipated away. However, this physical principle captures important mathematical implications for the behavior of solutions. It leads to a useful alternative method for establishing stability of equilibria, even in cases when the linearization stability test is inconclusive. The nineteenth century Russian mathematician Alexander Lyapunov was the first to pinpoint the importance of such functions in dynamics.

Definition 4.17. A *Lyapunov function* for the first order autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ is a continuous real-valued function $L(\mathbf{u})$ that is non-increasing on all solutions $\mathbf{u}(t)$, meaning that

$$L(\mathbf{u}(t)) \ \leq\ L(\mathbf{u}(t_0)) \qquad \text{for all} \qquad t > t_0. \tag{4.27}$$

A *strict Lyapunov function* satisfies the strict inequality

$$L(\mathbf{u}(t)) \ <\ L(\mathbf{u}(t_0)) \qquad \text{for all} \qquad t > t_0, \tag{4.28}$$

whenever $\mathbf{u}(t)$ is a *non-equilibrium* solution to the system. (Clearly, the Lyapunov function must be constant on an equilibrium solution.)

We can characterize continuously differentiable Lyapunov functions by using the elementary calculus results that a scalar function is non-increasing if and only if its derivative is non-negative, and is strictly decreasing if its derivative is strictly less than 0. We can compute the derivative of $L(\mathbf{u}(t))$ by applying the same chain rule computation used to establish (4.14). As a result, we establish the basic criteria for Lyapunov functions.

Proposition 4.18. *A continuously differentiable function $L(\mathbf{u})$ is a Lyapunov function for the system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ if and only if it satisfies the* Lyapunov inequality

$$\frac{d}{dt}\, L(\mathbf{u}(t)) = \nabla L(\mathbf{u}) \cdot \mathbf{F}(\mathbf{u}) \ \leq\ 0 \qquad \text{for all solutions} \qquad \mathbf{u}(t). \tag{4.29}$$

Furthermore, $L(\mathbf{u})$ is a strict Lyapunov function if and only if

$$\frac{d}{dt}\, L(\mathbf{u}(t)) = \nabla L(\mathbf{u}) \cdot \mathbf{F}(\mathbf{u}) \ <\ 0 \qquad \text{whenever} \qquad \mathbf{F}(\mathbf{u}) \neq \mathbf{0}. \tag{4.30}$$

The main result on stability and instability of equilibria of a system that possesses a Lyapunov function follows.

Theorem 4.19. *Let $L(\mathbf{u})$ be a Lyapunov function for the autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If $\mathbf{u}^\star$ is a strict local minimum of L, then $\mathbf{u}^\star$ is a stable equilibrium point. If $L(\mathbf{u})$ is a strict Lyapunov function, then $\mathbf{u}^\star$ is an asymptotically stable equilibrium. On the other hand, any critical point of a strict Lyapunov function $L(\mathbf{u})$ which is not a local minimum is an unstable equilibrium point.*

In particular, local maxima of strict Lyapunov functions are *not* stable equilibria. In outline, the proof relies on the fact that the Lyapunov function is non-increasing on solutions, and hence any solution that starts out sufficiently near a minimum value cannot go far away, demonstrating stability. Figuratively, if you start near the bottom of a valley and never walk uphill you stay near the bottom. In the strict case, the Lyapunov function must be strictly decreasing on the non-equilibrium solutions, which thus must go to the minimum value of L as $t \to \infty$. Again, starting near the bottom of a valley and always walking downhill takes you eventually to the bottom. On the other hand, if a non-equilibrium solution starts near an equilibrium point that is not a local minimum, then the fact that the Lyapunov function is steadily decreasing implies that the solution must move further and further away from the equilibrium, which suffices to demonstrate instability.

Unlike first integrals, which can, at least in principle, be systematically constructed by solving a first order partial differential equation, finding Lyapunov functions is much more of an art form, usually requiring some combination of physical intuition and inspired guesswork.

Example 4.20. Return to the planar system

$$\frac{du}{dt} = v, \qquad\qquad \frac{dv}{dt} = -\,\alpha \sin u - \beta\, v,$$

describing the damped oscillations of a pendulum, as in (4.10). Physically, we expect that the damping will cause a continual decrease in the total energy in the system, which, by (4.24), is

$$E(u,v) = \tfrac{1}{2}\, m v^2 + \kappa\,(1 - \cos u).$$

Let us prove that E is, indeed, a Lyapunov function. We compute its time derivative, when $u(t), v(t)$ is a solution to the damped system. Recalling that $\alpha = \kappa/m$, $\beta = \mu/m$, we find

$$\frac{dE}{dt} = \frac{\partial E}{\partial u}\,\frac{du}{dt} + \frac{\partial E}{\partial v}\,\frac{dv}{dt} = (\kappa \sin u)\, v + (m\, v)\,(-\,\alpha \sin u - \beta\, v) = -\,\mu\, v^2 \leq 0,$$

since we are assuming that the frictional coefficient $\mu > 0$. Therefore, the energy satisfies the Lyapunov stability criterion (4.29). Consequently, Theorem 4.19 re-establishes the stability of the energy minima $u_{2k}^\star = 2\,k\,\pi$, $v_{2k}^\star = 0$, where the damped pendulum is at the bottom of the arc. In fact, since $dE/dt < 0$ except when $v = 0$, with a little more work, the Lyapunov criterion can be used to establish their asymptotic stability.

Hamiltonian and Poisson Systems

One particularly important class of conservative systems were first introduced in the work of William Rowan Hamilton and Siméon–Dennis Poisson in the early nineteenth century. Their research was concerned with methods for solving the conservative systems arising in classical mechanics. Remarkably, Hamiltonian systems turned out to be the mathematical vehicle that unlocked the modern world of subatomic quantum mechanics.

Definition 4.21. A *Poisson system* is a first order system of ordinary differential equations of the form

$$\dot{\mathbf{u}} = J\,\nabla H(\mathbf{u}), \qquad \text{where} \qquad J^T = -\,J \tag{4.31}$$

is a constant[†] skew-symmetric matrix. The real-valued function $H(\mathbf{u})$ is known as the *Hamiltonian function* for the system.

The Hamiltonian function is automatically a first integral for the Poisson system (4.31). Indeed, to verify (4.14), we compute

$$\frac{dH}{dt} = \nabla H \cdot \frac{d\mathbf{u}}{dt} = \nabla H \cdot \mathbf{F} = \nabla H^T\, \mathbf{F} = \nabla H^T\, J\, \nabla H = 0.$$

The final equality follows from the fact that J is skew-symmetric, and hence $\mathbf{v}^T\, J\,\mathbf{v} = 0$ for any vector $\mathbf{v}$. In applications, the Hamiltonian function often represents the total energy of the system, and so a Poisson system necessarily conserves energy. Thus, friction and damping are not typically included in the Poisson framework.

Conservation of the Hamiltonian function means that the solutions of the Poisson system (4.31) move along its level sets $\{\, H(\mathbf{u}) = c \,\}$. In particular, if a level set is a single point $\mathbf{u}_0$, then the corresponding solution cannot move and hence is an equilibrium solution: $\mathbf{u}(t) \equiv \mathbf{u}_0$. If $\mathbf{u}_0$ is an isolated maximum or minimum of H, then it is necessarily stable, since the nearby level sets remain close to $\mathbf{u}_0$ and hence nearby solutions can never

[†] Nonconstant Poisson matrices have additional, more subtle requirements, [**26**; Chapter 6].

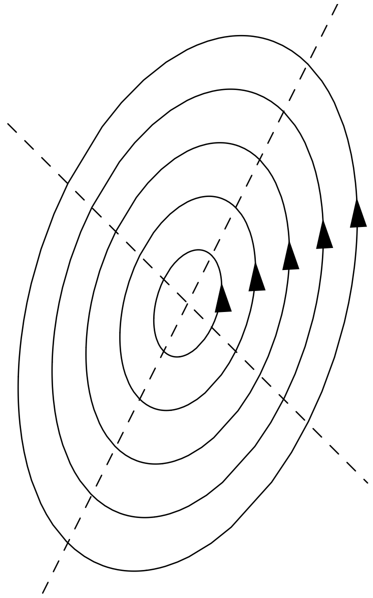


Figure 15. A Stable Planar Hamiltonian System.

get far away from the equilibrium solution. In Figure 15 we plot the elliptical level sets of a typical stable quadratic Hamiltonian in the plane — the solutions move periodically around the ellipses, all with the same period.

The simplest example is when $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is a 2×2 matrix. In this case, we set $\mathbf{u}(t) = (\, p(t), q(t)\,)^T$, and the Poisson system (4.31) corresponding to a Hamiltonian function $H(p,q)$ takes the form

$$\frac{dp}{dt} = -\,\frac{\partial H}{\partial q}\,, \qquad\qquad \frac{\partial q}{\partial t} = \frac{\partial H}{\partial p}\,. \tag{4.32}$$

For example, consider the undamped pendulum equation (4.23). Let us introduce the position variable $q = \theta$ representing the angular coordinate of the pendulum. Let $p = m\dot{\theta}$ be its angular momentum. The pendulum energy function (4.24) is rewritten in terms of the position and momentum variables:

$$H(p,q) = \frac{p^2}{2\,m} + \kappa\,(1 - \cos q).$$

The resulting Hamiltonian system is (4.32)

$$\frac{dp}{dt} = \kappa \sin q, \qquad\qquad \frac{\partial q}{\partial t} = \frac{p}{m}. \tag{4.33}$$

If we convert back to $u = q$ and velocity $v = \dot{\theta} = p/m$ we immediately recover the phase plane form (4.23) of the pendulum equation. Thus, we recover the constancy of the energy first integral directly from the general Hamiltonian framework.

A Poisson system is called *Hamiltonian* if J is nonsingular: $\det J \neq 0$. Since J is a skew-symmetric matrix, this can only happen if its size, which is also the dimension of the underlying space is even: $n = 2k$. The most important case, generalizing the two-dimensional version (4.32), is when

$$J = \begin{pmatrix} \mathrm{O} & -\mathrm{I} \\ \mathrm{I} & \mathrm{O} \end{pmatrix}$$

where O denotes the $k \times k$ zero matrix and I denotes the $k \times k$ identity matrix. In this case, the variables are split into two vector components:

$$\mathbf{u} = (\,\mathbf{p}, \mathbf{q}\,)^T = (\,p_1, \ldots, p_k, q_1, \ldots, q_k\,)^T .$$

We find

$$\nabla H(\mathbf{p}, \mathbf{q}) = \left(\frac{\partial H}{\partial p_1}, \;\ldots\; \frac{\partial H}{\partial p_k}, \frac{\partial H}{\partial q_1}, \;\ldots\; \frac{\partial H}{\partial q_k} \right)^T$$

and so (4.31) takes on the explicit form

$$\frac{dp_1}{dt} = -\,\frac{\partial H}{\partial q_1}, \quad \ldots \quad \frac{dp_k}{dt} = -\,\frac{\partial H}{\partial q_k}, \qquad \frac{dq_1}{dt} = \frac{\partial H}{\partial p_1}, \quad \ldots \quad \frac{dq_k}{dt} = \frac{\partial H}{\partial p_k}\,. \tag{4.34}$$

In physical applications, the $\mathbf{p}$ variables typically represent momenta, while the $\mathbf{q}$ variables represent positions of the objects in the system; this was already noted in the pendulum example.

Conservative mechanical systems can always be represented in Hamiltonian form, [**28**]. The prototypical example is to take the Hamiltonian function of the particular form

$$H(\mathbf{p}, \mathbf{q}) = \frac{p_1^2}{2\,m_1} + \;\cdots\; + \frac{p_k^2}{2\,m_k} + V(q_1, \ldots, q_k), \tag{4.35}$$

where each $m_i > 0$ is a positive constant. The canonical system (4.34) has the form

$$\frac{dp_1}{dt} = -\,\frac{\partial V}{\partial q_1}\,, \quad \ldots \quad \frac{dp_k}{dt} = -\,\frac{\partial V}{\partial q_k}\,, \qquad \frac{dq_1}{dt} = \frac{p_1}{m_1}, \quad \ldots \quad \frac{dq_k}{dt} = \frac{p_k}{m_k}\,,$$

which we can rewrite in vectorial form

$$M\,\frac{d\mathbf{q}}{dt} = \mathbf{p}, \qquad\qquad \frac{d\mathbf{p}}{dt} = -\,\nabla V(\mathbf{q}), \tag{4.36}$$

where $M = \operatorname{diag}(m_1, \ldots, m_n)$. Eliminating the $\mathbf{p}$ variables, we find that (4.36) reduces to a second order system of ordinary differential equations in the Newtonian form

$$M\,\frac{d^2\mathbf{q}}{dt^2} = -\,\nabla V(\mathbf{q}).$$

The vector $\mathbf{q}$ represents the position vector for the mechanical system, while $V(\mathbf{q})$ is the potential energy. The constants m_i are masses, and each $p_i = m_i\,\dot{q}_i$ represents the *momentum* variable associated with the position variable q_i. The Hamiltonian function is the total energy, the first term being the kinetic energy since the term

$$\frac{p_i^2}{2\,m_i} = \frac{m_i}{2}\left(\frac{dq_i}{dt}\right)^2$$

is precisely one half mass times velocity squared.

Example 4.22. Consider a planetary system consisting of n bodies in three-dimensional space. We represent the planets as point masses, where the i^{th} planet is concentrated at position $\mathbf{q}_i(t) = (\,x_i(t), y_i(t), z_i(t)\,)^T$. The planet's linear momentum vector is $\mathbf{p}_i(t) = m_i \dot{\mathbf{q}}_i(t) = (\,\dot{x}_i(t), \dot{y}_i(t), \dot{z}_i(t)\,)^T$, and its kinetic energy is

$$\frac{\|\,\mathbf{p}_i\,\|^2}{2\,m_i} = \frac{m_i}{2}\,\|\,\dot{\mathbf{q}}_i\,\|^2 = \frac{m_i}{2}\,(\,\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2\,).$$

The Newtonian gravitational potential between two point masses is proportional to the inverse square of their distance; more specifically

$$V(\mathbf{q}_i, \mathbf{q}_j) = \frac{G\,m_i\,m_j}{\|\,\mathbf{q}_i - \mathbf{q}_j\,\|^2} \tag{4.37}$$

represents the gravitational potential between planets i and j, where m_i, m_j are their masses, and G the universal gravitational constant. The Hamiltonian of the system is the total kinetic plus potential energy, which is obtained by summing the individual contributions from (pairs of) planets:

$$H(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{\|\,\mathbf{p}_i\,\|^2}{2\,m_i} + \sum_{1 \le i < j \le n} \frac{G\,m_i\,m_j}{\|\,\mathbf{q}_i - \mathbf{q}_j\,\|^2}.$$

The resulting Hamiltonian system is known as the *n–body problem*, and it has been the subject of intense research since the time of Newton. For a two body system, e.g., a planet and a sun, the motion is well understood; the smaller body moves around either an ellipse with the sum at a focus, or a parabola, or a hyperbola — the latter two cases apply to comets. Even today, there are fundamental unanswered questions about the nature of solutions to this basic physical system for 3 or more bodies. Recent developments have been the discovery of systems of planets in which one or more goes off to infinity in finite time, [**20**], and systems of planets that move around a figure 8 and even more complicated curves, [**6**, **22**].

Remark: For systems in Hamiltonian form, Theorem 4.13 tells us that the strict local minima (and maxima) of the Hamiltonian function are necessarily stable equilibrium points! This reconfirms our general observation. Saddle points are typically unstable.

The Hamiltonian form of a classical mechanical system plays a profound role in the construction of its quantum mechanics version. There is a canonical quantization procedure that replaces the nonlinear Hamiltonian system of ordinary differential equations with a linear, quantum mechanics partial differential equation known as the Schrödinger equation. For example, the quantum mechanical system describing an electron orbiting a proton in a hydrogen atom is the quantized version of the classical 2–body problem describing the motion of a single planet around the sun. For details, we refer the reader to any basic text in quantum mechanics, e.g., [**10**, **19**, **21**].

5. Numerical Methods.

Since we have no hope of solving the vast majority of differential equations in explicit, analytic form, the design of suitable numerical algorithms for accurately approximating solutions is essential. The ubiquity of differential equations throughout mathematics and its applications has driven the tremendous research effort devoted to numerical solution schemes, some dating back to the beginnings of the calculus. Nowadays, one has the luxury of choosing from a wide range of excellent software packages that provide reliable and accurate results for a broad range of systems, at least for solutions over moderately long time periods. However, all of these packages, and the underlying methods, have their limitations, and it is essential that one be able to to recognize when the software is working as advertised, and when it produces spurious results! Here is where the theory, particularly the classification of equilibria and their stability properties, as well as first integrals and Lyapunov functions, can play an essential role. Explicit solutions, when known, can also be used as test cases for tracking the reliability and accuracy of a chosen numerical scheme.

In this section, we survey the most basic numerical methods for solving initial value problems. For brevity, we shall only consider so-called single step schemes, culminating in the very popular and versatile fourth order Runge–Kutta Method. This should only serve as a extremely basic introduction to the subject, and many other important and useful methods can be found in more specialized texts, [**12**, **18**]. It goes without saying that some equations are more difficult to accurately approximate than others, and a variety of more specialized techniques are employed when confronted with a recalcitrant system. But all of the more advanced developments build on the basic schemes and ideas laid out in this section.

Euler's Method

The key issues confronting the numerical analyst of ordinary differential equations already appear in the simplest first order ordinary differential equation. Our goal is to calculate a decent approximation to the (unique) solution to the initial value problem

$$\frac{du}{dt} = F(t,u), \qquad\qquad u(t_0) = u_0. \tag{5.1}$$

To keep matters simple, we will focus our attention on the scalar case; however, all formulas and results written in a manner that can be readily adapted to first order systems — just replace the scalar functions $u(t)$ and $F(t,u)$ by vector-valued functions $\mathbf{u}$ and $\mathbf{F}(t,\mathbf{u})$ throughout. (The time t, of course, remains a scalar.) Higher order ordinary differential equations are inevitably handled by first converting them into an equivalent first order system, as discussed in Section 2, and then applying the numerical scheme.

The very simplest numerical solution method is named after Leonhard Euler — although Newton and his contemporaries were well aware of such a simple technique. Euler's Method is rarely used in practice because much more efficient and accurate techniques can be implemented with minimal additional work. Nevertheless, the method lies at the core of the entire subject, and must be thoroughly understood before progressing on to the more sophisticated algorithms that arise in real-world computations.

Starting at the initial time t_0, we introduce successive *mesh points* (or sample times)

$$t_0 < t_1 < t_2 < t_3 < \ \cdots \ ,$$

continuing on until we reach a desired final time $t_n = t^\star$. The mesh points should be fairly closely spaced. To keep the analysis as simple as possible, we will always use a uniform *step size*, and so

$$h = t_{k+1} - t_k > 0, \tag{5.2}$$

does not depend on k and is assumed to be relatively small. This assumption serves to simplify the analysis, and does not significantly affect the underlying ideas. For a uniform step size, the k^{th} mesh point is at $t_k = t_0 + k\,h$. More sophisticated *adaptive* methods, in which the step size is adjusted in order to maintain accuracy of the numerical solution, can be found in more specialized texts, e.g., [**12, 18**]. Our numerical algorithm will recursively compute approximations $u_k \approx u(t_k)$, for $k = 0, 1, 2, 3, \ldots$, to the sampled values of the solution $u(t)$ at the chosen mesh points. Our goal is to make the *error* $E_k = u_k - u(t_k)$ in the approximation at each time t_k as small as possible. If required, the values of the solution $u(t)$ between mesh points may be computed by a subsequent interpolation procedure, e.g., cubic splines.

As you learned in first year calculus, the simplest approximation to a (continuously differentiable) function $u(t)$ is provided by its tangent line or first order Taylor polynomial. Thus, near the mesh point t_k

$$u(t) \;\approx\; u(t_k) + (t - t_k)\,\frac{du}{dt}(t_k) \;=\; u(t_k) + (t - t_k)\,F(t_k, u(t_k)),$$

in which we replace the derivative du/dt of the solution by the right hand side of the governing differential equation (5.1). In particular, the approximate value of the solution at the subsequent mesh point is

$$u(t_{k+1}) \;\approx\; u(t_k) + (t_{k+1} - t_k)\,F(t_k, u(t_k)). \tag{5.3}$$

This simple idea forms the basis of Euler's Method.

Since in practice we only know the approximation u_k to the value of $u(t_k)$ at the current mesh point, we are forced to replace $u(t_k)$ by its approximation u_k in the preceding formula. We thereby convert (5.3) into the iterative scheme

$$u_{k+1} = u_k + (t_{k+1} - t_k)\,F(t_k, u_k). \tag{5.4}$$

In particular, when based on a uniform step size (5.2), *Euler's Method* takes the simple form

$$u_{k+1} = u_k + h\,F(t_k, u_k). \tag{5.5}$$

As sketched in Figure 16, the method starts off approximating the solution reasonably well, but gradually loses accuracy as the errors accumulate.

Euler's Method is the simplest example of a *one-step* numerical scheme for integrating an ordinary differential equation. This refers to the fact that the succeeding approximation, $u_{k+1} \approx u(t_{k+1})$, depends only upon the current value, $u_k \approx u(t_k)$, which is one mesh point or "step" behind.

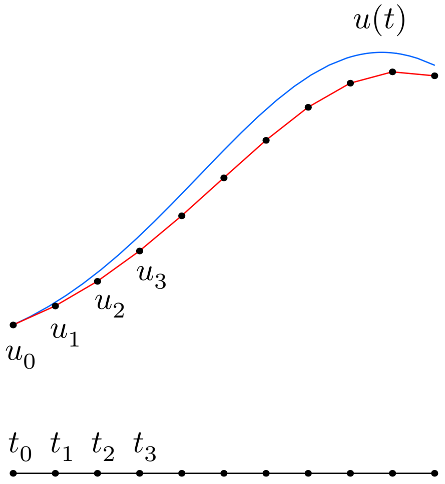


Figure 16. Euler's Method.

To begin to understand how Euler's Method works in practice, let us test it on a problem we know how to solve, since this will allow us to precisely monitor the resulting errors in our numerical approximation to the solution.

Example 5.1. The simplest "nontrivial" initial value problem is

$$\frac{du}{dt} = u, \qquad\qquad u(0) = 1,$$

whose solution is, of course, the exponential function $u(t) = e^t$. Since $F(t,u) = u$, Euler's Method (5.5) with a fixed step size $h > 0$ takes the form

$$u_{k+1} = u_k + h\,u_k = (1+h)\,u_k.$$

This is a linear iterative equation, and hence easy to solve:

$$u_k = (1+h)^k u_0 = (1+h)^k$$

is our proposed approximation to the solution $u(t_k) = e^{t_k}$ at the mesh point $t_k = k\,h$. Therefore, the Euler scheme to solve the differential equation, we are effectively approximating the exponential by a power function:

$$e^{t_k} = e^{k\,h} \;\approx\; (1+h)^k$$

When we use simply t to indicate the mesh time $t_k = k\,h$, we recover, in the limit, a well-known calculus formula:

$$e^t = \lim_{h\to 0}\,(1+h)^{t/h} = \lim_{k\to\infty}\left(1+\frac{t}{k}\right)^k. \tag{5.6}$$

A reader familiar with the computation of compound interest will recognize this particular approximation. As the time interval of compounding, h, gets smaller and smaller, the amount in the savings account approaches an exponential.

How good is the resulting approximation? The *error*

$$E(t_k) = E_k = u_k - e^{t_k}$$

measures the difference between the true solution and its numerical approximation at time $t = t_k = k\,h$. Let us tabulate the error at the particular times $t = 1, 2$ and 3 for various values of the step size h. The actual solution values are

$$e^1 = e = 2.718281828\ldots, \qquad e^2 = 7.389056096\ldots, \qquad e^3 = 20.085536912\ldots.$$

In this case, the numerical approximation always underestimates the true solution.

h	$E(1)$	$E(2)$	$E(3)$
.1	$-.125$	$-.662$	-2.636
.01	$-.0134$	$-.0730$	$-\ \ .297$
.001	$-.00135$	$-.00738$	$-\ \ .0301$
.0001	$-.000136$	$-.000739$	$-\ \ .00301$
.00001	$-.0000136$	$-.0000739$	$-\ \ .000301$

Some key observations:

- For a fixed step size h, the further we go from the initial point $t_0 = 0$, the larger the magnitude of the error.
- On the other hand, the smaller the step size, the smaller the error at a fixed value of t. The trade-off is that more steps, and hence more computational effort[†] is required to produce the numerical approximation. For instance, we need $k = 10$ steps of size $h = .1$, but $k = 1000$ steps of size $h = .001$ to compute an approximation to $u(t)$ at time $t = 1$.
- The error is more or less in proportion to the step size. Decreasing the step size by a factor of $\frac{1}{10}$ decreases the error by a similar amount, but simultaneously increases the amount of computation by a factor of 10.

The final observation indicates that the Euler Method is of *first order*, which means that the error depends *linearly* on the step size h. More specifically, at a fixed time t, the error is bounded by

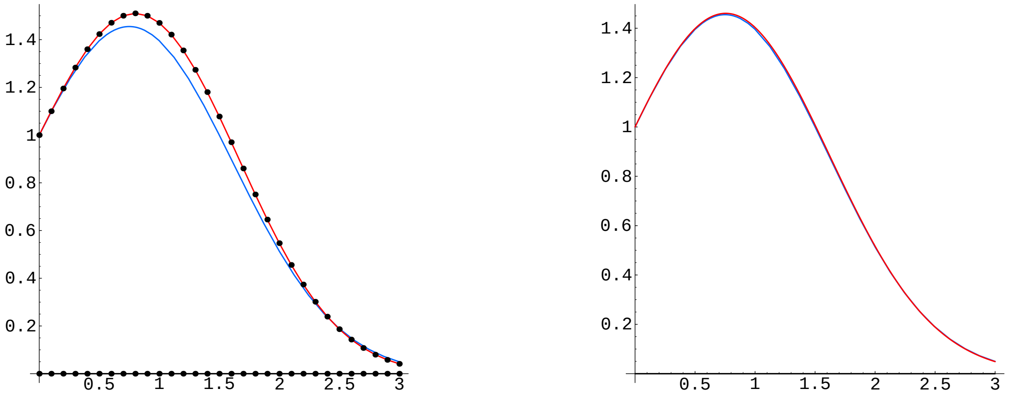
$$|\,E(t)\,| = |\,u_k - u(t)\,| \le C(t)\,h, \qquad \text{when} \qquad t = t_k = k\,h, \tag{5.7}$$

for some positive $C(t) > 0$ that depends upon the time, and the initial condition, but not on the step size.

Example 5.2. The solution to the initial value problem

$$\frac{du}{dt} = \left(1 - \tfrac{4}{3}\,t\right) u, \qquad u(0) = 1, \tag{5.8}$$

[†] In this case, there happens to be an explicit formula for the numerical solution which can be used to bypass the iterations. However, in almost any other situation, one cannot compute the approximation u_k without having first determined the intermediate values $u_0, \ldots, u_{k-1}$.



$h = .1$ $\qquad$ $h = .01$

Figure 17. Euler's Method for $\dot{u} = \left(1 - \frac{4}{3}\,t\right)u$.

was found in Example 2.3 by the method of separation of variables:

$$u(t) = \exp\left(t - \tfrac{2}{3}\,t^2\right). \tag{5.9}$$

Euler's Method leads to the iterative numerical scheme

$$u_{k+1} = u_k + h\left(1 - \tfrac{4}{3}\,t_k\right)u_k, \qquad u_0 = 1.$$

In Figure 17 we compare the graphs of the actual and numerical solutions for step sizes $h = .1$ and $.01$. In the former plot, we explicitly show the mesh points, but not in the latter, since they are too dense; moreover, the graphs of the numerical and true solutions are almost indistinguishable at this resolution.

The following table lists the numerical errors $E(t_k) = u_k - u(t_k)$ between the computed and actual solution values

$$u(1) = 1.395612425\ldots, \qquad u(2) = .513417119\ldots, \qquad u(3) = .049787068\ldots,$$

for several different step sizes:

h	$E(1)$	$E(2)$	$E(3)$
.1000	.07461761	.03357536	−.00845267
.0100	.00749258	.00324416	−.00075619
.0010	.00074947	.00032338	−.00007477
.0001	.00007495	.00003233	−.00000747

As in the previous example, each decrease in step size by a factor of 10 leads to one additional decimal digit of accuracy in the computed solution.

Taylor Methods

In general, the order of a numerical solution method governs both the accuracy of its approximations and the speed at which they converge to the true solution as the step size is decreased. Although the Euler Method is simple and easy to implement, it is only a first order scheme, and therefore of limited utility in serious computations. So, the goal is to devise simple numerical methods that enjoy a much higher order of accuracy.

Our derivation of the Euler Method was based on a first order Taylor approximation to the solution. So, an evident way to design a higher order method is to employ a higher order Taylor approximation. The Taylor series expansion for the solution $u(t)$ at the succeeding mesh point $t_{k+1} = t_k + h$ has the form

$$u(t_{k+1}) = u(t_k + h) = u(t_k) + h\,\frac{du}{dt}(t_k) + \frac{h^2}{2}\,\frac{d^2u}{dt^2}(t_k) + \frac{h^3}{6}\,\frac{d^3u}{dt^3}(t_k) + \ \cdots\ . \tag{5.10}$$

As we just saw, we can evaluate the first derivative term through use of the underlying differential equation:

$$\frac{du}{dt} = F(t,u). \tag{5.11}$$

The second derivative term can be found by differentiating the equation with respect to t. Invoking the chain rule[†],

$$\begin{aligned}\frac{d^2u}{dt^2} = \frac{d}{dt}\,\frac{du}{dt} = \frac{d}{dt}\,F(t,u(t)) &= \frac{\partial F}{\partial t}(t,u) + \frac{\partial F}{\partial u}(t,u)\,\frac{du}{dt} \\ &= \frac{\partial F}{\partial t}(t,u) + \frac{\partial F}{\partial u}(t,u)\ F(t,u) \equiv F^{(2)}(t,u).\end{aligned} \tag{5.12}$$

This operation is known as the *total derivative*, indicating that that we must treat the second variable u as a function of t when differentiating. Substituting (5.11–12) into (5.10) and truncating at order h^2 leads to the *Second Order Taylor Method*

$$\begin{aligned}u_{k+1} &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k) \\ &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,\left(\frac{\partial F}{\partial t}(t_k,u_k) + \frac{\partial F}{\partial u}(t_k,u_k)\ F(t_k,u_k)\right),\end{aligned} \tag{5.13}$$

in which, as before, we replace the solution value $u(t_k)$ by its computed approximation u_k. The resulting method is of second order, meaning that the error function satisfies the quadratic error estimate

$$|\,E(t)\,| = |\,u_k - u(t)\,| \le C(t)\,h^2 \qquad \text{when} \qquad t = t_k = k\,h. \tag{5.14}$$

[†] We assume throughout that F has as many continuous derivatives as needed.

Example 5.3. Let us explicitly formulate the second order Taylor Method for the initial value problem (5.8). Here

$$\begin{aligned}\frac{du}{dt} &= F(t,u) = \left(1 - \tfrac{4}{3}\,t\right)u,\\ \frac{d^2u}{dt^2} &= \frac{d}{dt}\,F(t,u) = -\tfrac{4}{3}\,u + \left(1 - \tfrac{4}{3}\,t\right)\frac{du}{dt} = -\tfrac{4}{3}\,u + \left(1 - \tfrac{4}{3}\,t\right)^2 u,\end{aligned}$$

and so (5.13) becomes

$$u_{k+1} = u_k + h\left(1 - \tfrac{4}{3}\,t_k\right)u_k + \tfrac{1}{2}\,h^2\left[-\tfrac{4}{3}\,u_k + \left(1 - \tfrac{4}{3}\,t_k\right)^2 u_k\right], \qquad u_0 = 1.$$

The following table lists the errors between the values computed by the second order Taylor scheme and the actual solution values, as given in Example 5.2.

h	$E(1)$	$E(2)$	$E(3)$
.100	.00276995	−.00133328	.00027753
.010	.00002680	−.00001216	.00000252
.001	.00000027	−.00000012	.00000002

Observe that, in accordance with the quadratic error estimate (5.14), a decrease in the step size by a factor of $\frac{1}{10}$ leads in an increase in accuracy of the solution by a factor $\frac{1}{100}$, i.e., an increase in 2 significant decimal places in the numerical approximation of the solution.

Higher order Taylor methods are obtained by including further terms in the expansion (5.10). For example, to derive a third order Taylor method, we include the third order term $(h^3/6)d^3u/dt^3$ in the Taylor expansion, where we evaluate the third derivative by differentiating (5.12), and so

$$\begin{aligned}\frac{d^3u}{dt^3} &= \frac{d}{dt}\,\frac{d^2u}{dt^2} = \frac{d}{dt}\,F^{(2)}(t,u) = \frac{\partial F^{(2)}}{\partial t} + \frac{\partial F^{(2)}}{\partial u}\frac{du}{dt} = \frac{\partial F^{(2)}}{\partial t} + F\,\frac{\partial F^{(2)}}{\partial u}\\ &= \frac{\partial^2 F}{\partial t^2} + 2\,F\,\frac{\partial^2 F}{\partial t\,\partial u} + F^2\,\frac{\partial^2 F}{\partial u^2} + \frac{\partial F}{\partial t}\,\frac{\partial F}{\partial u} + F\left(\frac{\partial F}{\partial u}\right)^2 \equiv F^{(3)}(t,u),\end{aligned} \tag{5.15}$$

where we continue to make use of the fact that $du/dt = F(t,u)$ is provided by the right hand side of the differential equation. The resulting third order Taylor method is

$$u_{k+1} = u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k) + \frac{h^3}{6}\,F^{(3)}(t_k,u_k), \tag{5.16}$$

where the last two summand are given by (5.12), (5.15), respectively. The higher order expressions are even worse, and a good symbolic manipulation system is almost essential for accurate computation.

Whereas higher order Taylor methods are easy to motivate, they are rarely used in practice. There are two principal difficulties:

- Owing to their dependence upon the partial derivatives of $F(t,u)$, the right hand side of the differential equation needs to be rather smooth.
- Even worse, the explicit formulae become exceedingly complicated, even for relatively simple functions $F(t,u)$. Efficient evaluation of the multiplicity of terms in the Taylor approximation and avoidance of round off errors become significant concerns.

As a result, mathematicians soon abandoned the Taylor series approach, and began to look elsewhere for high order, efficient integration methods.

Error Analysis

Before pressing on, we need to engage in a more serious discussion of the error in a numerical scheme. A general *one-step* numerical method can be written in the form

$$u_{k+1} = G(h, t_k, u_k), \tag{5.17}$$

where G is a prescribed function of the current approximate solution value $u_k \approx u(t_k)$, the time t_k, and the step size $h = t_{k+1} - t_k$, which, for illustrative purposes, we assume to be fixed. (We leave the discussion of *multi-step methods*, in which G could also depend upon the earlier values $u_{k-1}, u_{k-2}, \ldots$, to more advanced texts, e.g., [**12**, **18**].)

In any numerical integration scheme there are, in general, three sources of error.

- The first is the *local error* committed in the current step of the algorithm. Even if we have managed to compute a completely accurate value of the solution $u_k = u(t_k)$ at time t_k, the numerical approximation scheme (5.17) is almost certainly not exact, and will therefore introduce an error into the next computed value $u_{k+1} \approx u(t_{k+1})$. Round-off errors, resulting from the finite precision of the computer arithmetic, will also contribute to the local error.
- The second is due to the error that is already present in the current approximation $u_k \approx u(t_k)$. The local errors tend to accumulate as we continue to run the iteration, and the net result is the *global error*, which is what we actually observe when comparing the numerical approximation with the exact solution.
- Finally, if the initial condition $u_0 \approx u(t_0)$ is not computed accurately, this *initial error* will also make a contribution. For example, if $u(t_0) = \pi$, then we introduce some initial error by using a decimal approximation, say $\pi \approx 3.14159$.

The third error source will, for simplicity, be ignored in our discussion, i.e., we will assume $u_0 = u(t_0)$ is exact. Further, for simplicity we will assume that round-off errors do not play any significant role — although one must always keep them in mind when analyzing the computation. Since the global error is entirely due to the accumulation of successive local errors, we must first understand the local error in detail.

To measure the local error in going from t_k to t_{k+1}, we compare the exact solution value $u(t_{k+1})$ with its numerical approximation (5.17) under the assumption that the current computed value is correct: $u_k = u(t_k)$. Of course, in practice this is never the case, and so the local error is an artificial quantity. Be that as it may, in most circumstances the local error is (*a*) easy to estimate, and, (*b*) provides a reliable guide to the global accuracy of the numerical scheme. To estimate the local error, we assume that the step size h is small and approximate the solution $u(t)$ by its Taylor expansion[†]

$$\begin{aligned} u(t_{k+1}) &= u(t_k) + h\,\frac{du}{dt}(t_k) + \frac{h^2}{2}\,\frac{d^2u}{dt^2}(t_k) + \;\cdots \\ &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k) + \frac{h^3}{6}\,F^{(3)}(t_k,u_k) + \;\cdots\;. \end{aligned} \tag{5.18}$$

In the second expression, we have employed (5.12, 15) and their higher order analogs to evaluate the derivative terms, and then invoked our local accuracy assumption to replace $u(t_k)$ by u_k. On the other hand, a direct Taylor expansion, in h, of the numerical scheme produces

$$\begin{aligned} u_{k+1} &= G(h,t_k,u_k) \\ &= G(0,t_k,u_k) + h\,\frac{\partial G}{\partial h}(0,t_k,u_k) + \frac{h^2}{2}\,\frac{\partial^2 G}{\partial h^2}(0,t_k,u_k) + \frac{h^3}{6}\,\frac{\partial^3 G}{\partial h^3}(0,t_k,u_k) + \;\cdots\;. \end{aligned} \tag{5.19}$$

The local error is obtained by comparing these two Taylor expansions.

Definition 5.4. A numerical integration method is of *order n* if the Taylor expansions (5.18, 19) of the exact and numerical solutions agree up to order h^n.

For example, the Euler Method

$$u_{k+1} = G(h,t_k,u_k) = u_k + h\,F(t_k,u_k),$$

is already in the form of a Taylor expansion — that has no terms involving $h^2, h^3, \ldots$. Comparing with the exact expansion (5.18), we see that the constant and order h terms are the same, but the order h^2 terms differ (unless $F^{(2)} \equiv 0$). Thus, according to the definition, the Euler Method is a first order method. Similarly, the Taylor Method (5.13) is a second order method, because it was explicitly designed to match the constant, h and h^2 terms in the Taylor expansion of the solution. The general Taylor Method of order n sets $G(h,t_k,u_k)$ to be exactly the order n Taylor polynomial, differing from the full Taylor expansion at order h^{n+1}.

Under fairly general hypotheses, it can be proved that, if the numerical scheme has order n as measured by the local error, then the *global error* is bounded by a multiple of h^n. In other words, assuming no round-off or initial error, the computed value u_k and the solution at time t_k can be bounded by

$$|\,u_k - u(t_k)\,| \le M\,h^n, \tag{5.20}$$

where the constant $M > 0$ may depend on the time t_k and the particular solution $u(t)$. The error bound (5.20) serves to justify our numerical observations. For a method of order n, decreasing the step size by a factor of $\frac{1}{10}$ will decrease the overall error by a factor of about 10^{-n}, and so, roughly speaking, we anticipate gaining an additional n digits of accuracy —

[†] In our analysis, we assume that the differential equation, and hence the solution, has sufficient smoothness to justify the relevant Taylor approximation.

at least up until the point that round-off errors begin to play a role. Readers interested in a more complete error analysis of numerical integration schemes should consult a specialized text, e.g., [**12**, **18**].

The bottom line is the higher its order, the more accurate the numerical scheme, and hence the larger the step size that can be used to produce the solution to a desired accuracy. However, this must be balanced with the fact that higher order methods inevitably require more computational effort at each step. If the total amount of computation has also decreased, then the high order method is to be preferred over a simpler, lower order method. Our goal now is to find another route to the design of higher order methods that avoids the complications inherent in a direct Taylor expansion.

An Equivalent Integral Equation

The secret to the design of higher order numerical algorithms is to replace the differential equation by an equivalent integral equation. By way of motivation, recall that, in general, differentiation is a badly behaved process; a reasonable function can have an unreasonable derivative. On the other hand, integration ameliorates; even quite nasty functions have relatively well-behaved integrals. For the same reason, accurate numerical integration is relatively painless, whereas numerical differentiation should be avoided unless necessary. While we have not dealt directly with integral equations in this text, the subject has been extensively developed by mathematicians, [**7**], and has many important physical applications.

Conversion of an initial value problem (5.1) to an integral equation is straightforward. We integrate both sides of the differential equation from the initial point t_0 to a variable time t. The Fundamental Theorem of Calculus is used to explicitly evaluate the left hand integral:

$$u(t) - u(t_0) = \int_{t_0}^{t} \dot{u}(s)\,ds = \int_{t_0}^{t} F(s,u(s))\,ds.$$

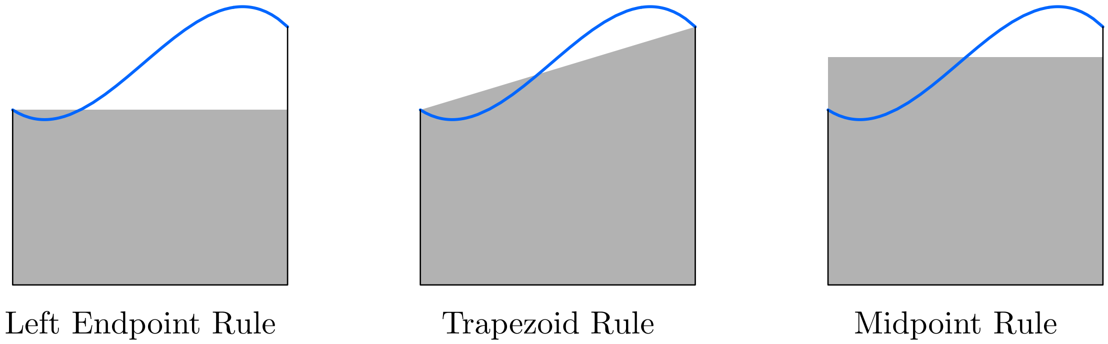
Rearranging terms, we arrive at the key result.

Lemma 5.5. *The solution $u(t)$ to the the integral equation*

$$u(t) = u(t_0) + \int_{t_0}^{t} F(s,u(s))\,ds \tag{5.21}$$

coincides with the solution to the initial value problem $\dfrac{du}{dt} = F(t,u)$, $u(t_0) = u_0$.

Proof: Our derivation already showed that the solution to the initial value problem satisfies the integral equation (5.21). Conversely, suppose that $u(t)$ solves the integral equation. Since $u(t_0) = u_0$ is constant, the Fundamental Theorem of Calculus tells us that the derivative of the right hand side of (5.21) is equal to the integrand, so $\dfrac{du}{dt} = F(t,u(t))$. Moreover, at $t = t_0$, the upper and lower limits of the integral coincide, and so it vanishes, whence $u(t) = u(t_0) = u_0$ has the correct initial conditions. *Q.E.D.*



Left Endpoint Rule Trapezoid Rule Midpoint Rule

Figure 18. Numerical Integration Methods.

Observe that, unlike the differential equation, the integral equation (5.21) requires no additional initial condition — it is automatically built into the equation. The proofs of the fundamental existence and uniqueness Theorems 3.1 and 3.3 for ordinary differential equations are, in fact, based on the integral equation reformulation of the initial value problem; see [**13**, **15**] for details.

The integral equation reformulation is equally valid for systems of first order ordinary differential equations. As noted above, $\mathbf{u}(t)$ and $\mathbf{F}(t,\mathbf{u}(t))$ become vector-valued functions. Integrating a vector-valued function is accomplished by integrating its individual components. Complete details are left to the exercises.

Implicit and Predictor–Corrector Methods

From this point onwards, we shall abandon the original initial value problem, and turn our attention to numerically solving the equivalent integral equation (5.21). Let us rewrite the integral equation, starting at the mesh point t_k instead of t_0, and integrating until time $t = t_{k+1}$. The result is the basic integral formula

$$u(t_{k+1}) = u(t_k) + \int_{t_k}^{t_{k+1}} F(s,u(s))\,ds \tag{5.22}$$

that (implicitly) computes the value of the solution at the subsequent mesh point. Comparing this formula with the Euler Method

$$u_{k+1} = u_k + h\,F(t_k,u_k), \qquad \text{where} \qquad h = t_{k+1} - t_k,$$

and assuming for the moment that $u_k = u(t_k)$ is exact, we discover that we are merely approximating the integral by

$$\int_{t_k}^{t_{k+1}} F(s,u(s))\,ds \;\approx\; h\,F(t_k,u(t_k)). \tag{5.23}$$

This is the Left Endpoint Rule for numerical integration — that approximates the area under the curve $g(t) = F(t,u(t))$ between $t_k \leq t \leq t_{k+1}$ by the area of a rectangle whose height $g(t_k) = F(t_k,u(t_k)) \approx F(t_k,u_k)$ is prescribed by the left-hand endpoint of the graph. As indicated in Figure 18, this is a reasonable, but not especially accurate method of numerical integration.

In first year calculus, you no doubt encountered much better methods of approximating the integral of a function. One of these is the *Trapezoid Rule*, which approximates the integral of the function $g(t)$ by the area of a trapezoid obtained by connecting the two points $(t_k, g(t_k))$ and $(t_{k+1}, g(t_{k+1}))$ on the graph of g by a straight line, as in the second Figure 18. Let us therefore try replacing (5.23) by the more accurate trapezoidal approximation

$$\int_{t_k}^{t_{k+1}} F(s,u(s))\,ds \;\approx\; \tfrac{1}{2}\,h\left[\,F(t_k,u(t_k)) + F(t_{k+1},u(t_{k+1}))\,\right]. \tag{5.24}$$

Substituting this approximation into the integral formula (5.22), and replacing the solution values $u(t_k), u(t_{k+1})$ by their numerical approximations, leads to the (hopefully) more accurate numerical scheme

$$u_{k+1} = u_k + \tfrac{1}{2}\,h\left[\,F(t_k,u_k) + F(t_{k+1},u_{k+1})\,\right], \tag{5.25}$$

known as the *Trapezoid Method*. It is an *implicit scheme*, since the updated value u_{k+1} appears on both sides of the equation, and hence is only defined implicitly.

Example 5.6. Consider the differential equation $\dot{u} = \left(1 - \frac{4}{3}\,t\right)u$ studied in Examples 5.2 and 5.3. The Trapezoid Method with a fixed step size h takes the form

$$u_{k+1} = u_k + \tfrac{1}{2}\,h\left[\,\left(1 - \tfrac{4}{3}\,t_k\right)u_k + \left(1 - \tfrac{4}{3}\,t_{k+1}\right)u_{k+1}\,\right].$$

In this case, we can explicit solve for the updated solution value, leading to the recursive formula

$$u_{k+1} = \frac{1 + \frac{1}{2}\,h\left(1 - \frac{4}{3}\,t_k\right)}{1 - \frac{1}{2}\,h\left(1 - \frac{4}{3}\,t_{k+1}\right)}\;u_k = \frac{1 + \frac{1}{2}\,h - \frac{2}{3}\,h\,t_k}{1 - \frac{1}{2}\,h + \frac{2}{3}\,h\,(t_k + h)}\;u_k. \tag{5.26}$$

Implementing this scheme for three different step sizes gives the following errors between the computed and true solutions at times $t = 1, 2, 3$.

h	$E(1)$	$E(2)$	$E(3)$
.100	−.00133315	.00060372	−.00012486
.010	−.00001335	.00000602	−.00000124
.001	−.00000013	.00000006	−.00000001

The numerical data indicates that the Trapezoid Method is of second order. For each reduction in step size by $\frac{1}{10}$, the accuracy in the solution increases by, roughly, a factor of $\frac{1}{100} = \frac{1}{10^2}$; that is, the numerical solution acquires two additional accurate decimal digits.

The main difficulty with the Trapezoid Method (and any other implicit scheme) is immediately apparent. The updated approximate value for the solution u_{k+1} appears on both sides of the equation (5.25). Only for very simple functions $F(t,u)$ can one expect to solve (5.25) explicitly for u_{k+1} in terms of the known quantities t_k, u_k and $t_{k+1} = t_k + h$. The alternative is to employ a numerical equation solver, such as the bisection algorithm or Newton's Method, to compute u_{k+1}. In the case of Newton's Method, one would use the current approximation u_k as a first guess for the new approximation u_{k+1}. The resulting scheme requires some work to program, but can be effective in certain situations.

An alternative, less involved strategy is based on the following far-reaching idea. We already know a half-way decent approximation to the solution value u_{k+1} — namely that provided by the more primitive Euler scheme

$$\widetilde{u}_{k+1} = u_k + h\,F(t_k, u_k). \tag{5.27}$$

Let's use this estimated value in place of u_{k+1} on the right hand side of the implicit equation (5.25). The result

$$\begin{aligned} u_{k+1} &= u_k + \tfrac{1}{2}\,h\left[\,F(t_k,u_k) + F(t_k+h,\widetilde{u}_{k+1})\,\right] \\ &= u_k + \tfrac{1}{2}\,h\left[\,F(t_k,u_k) + F\bigl(\,t_k+h, u_k + h\,F(t_k,u_k)\,\bigr)\,\right]. \end{aligned} \tag{5.28}$$

is known as the *Improved Euler Method*. It is a completely explicit scheme since there is no need to solve any equation to find the updated value u_{k+1}.

Example 5.7. For our favorite equation $\dot{u} = \bigl(1 - \frac{4}{3}\,t\bigr)\,u$, the Improved Euler Method begins with the Euler approximation

$$\widetilde{u}_{k+1} = u_k + h\,\bigl(1 - \tfrac{4}{3}\,t_k\bigr)\,u_k,$$

and then replaces it by the improved value

$$\begin{aligned} u_{k+1} &= u_k + \tfrac{1}{2}\,h\left[\,\bigl(1 - \tfrac{4}{3}\,t_k\bigr)\,u_k + \bigl(\,1 - \tfrac{4}{3}\,t_{k+1}\,\bigr)\,\widetilde{u}_{k+1}\,\right] \\ &= u_k + \tfrac{1}{2}\,h\left[\,\bigl(1 - \tfrac{4}{3}\,t_k\bigr)\,u_k + \bigl(\,1 - \tfrac{4}{3}\,(t_k+h)\,\bigr)\bigl(\,u_k + h\,\bigl(1 - \tfrac{4}{3}\,t_k\bigr)\,u_k\,\bigr)\,\right] \\ &= \left[\,\bigl(\,1 - \tfrac{2}{3}\,h^2\,\bigr)\left[\,1 + h\,\bigl(1 - \tfrac{4}{3}\,t_k\bigr)\,\right] + \tfrac{1}{2}\,h^2\,\bigl(\,1 - \tfrac{4}{3}\,t_k\,\bigr)^2\,\right]\,u_k. \end{aligned}$$

Implementing this scheme leads to the following errors in the numerical solution at the indicated times. The Improved Euler Method performs comparably to the fully implicit scheme (5.26), and significantly better than the original Euler Method.

h	$E(1)$	$E(2)$	$E(3)$
.100	−.00070230	.00097842	.00147748
.010	−.00000459	.00001068	.00001264
.001	−.00000004	.00000011	.00000012

The Improved Euler Method is the simplest of a large family of so-called *predictor–corrector algorithms*. In general, one begins a relatively crude method — in this case the Euler Method — to *predict* a first approximation $\widetilde{u}_{k+1}$ to the desired solution value u_{k+1}. One then employs a more sophisticated, typically implicit, method to *correct* the original prediction, by replacing the required update u_{k+1} on the right hand side of the implicit scheme by the less accurate prediction $\widetilde{u}_{k+1}$. The resulting explicit, corrected value u_{k+1} will, provided the method has been designed with due care, be an improved approximation to the true solution.

The numerical data in Example 5.7 indicates that the Improved Euler Method is of second order since each reduction in step size by $\frac{1}{10}$ improves the solution accuracy by, roughly, a factor of $\frac{1}{100}$. To verify this prediction, we expand the right hand side of (5.28) in a Taylor series in h, and then compare, term by term, with the solution expansion (5.18). First[†],

$$F\big(t_k + h, u_k + h\,F(t_k, u_k)\big) = F + h\big(F_t + F\,F_u\big) + \tfrac{1}{2}h^2\big(F_{tt} + 2F\,F_{tu} + F^2\,F_{uu}\big) + \ \cdots\ ,$$

where all the terms involving F and its partial derivatives on the right hand side are evaluated at t_k, u_k. Substituting into (5.28), we find

$$u_{k+1} = u_k + h\,F + \tfrac{1}{2}h^2\big(F_t + F\,F_u\big) + \tfrac{1}{4}h^3\big(F_{tt} + 2F\,F_{tu} + F^2\,F_{uu}\big) + \ \cdots\ . \qquad (5.29)$$

The two Taylor expansions (5.18) and (5.29) agree in their order $1, h$ and h^2 terms, but differ at order h^3. This confirms our experimental observation that the Improved Euler Method is of second order.

We can design a range of numerical solution schemes by implementing alternative numerical approximations to the basic integral equation (5.22). For example, the Midpoint Rule approximates the integral of the function $g(t)$ by the area of the rectangle whose height is the value of the function at the midpoint:

$$\int_{t_k}^{t_{k+1}} g(s)\,ds \;\approx\; h\,g\left(t_k + \tfrac{1}{2}\,h\right), \qquad \text{where} \qquad h = t_{k+1} - t_k. \qquad (5.30)$$

See Figure 18 for an illustration. The Midpoint Rule is known to have the same order of accuracy as the Trapezoid Rule, [**4**]. Substituting into (5.22) leads to the approximation

$$u_{k+1} = u_k + \int_{t_k}^{t_{k+1}} F(s, u(s))\,ds \;\approx\; u_k + h\,F\left(t_k + \tfrac{1}{2}\,h, u\left(t_k + \tfrac{1}{2}\,h\right)\right).$$

Of course, we don't know the value of the solution $u\left(t_k + \frac{1}{2}\,h\right)$ at the midpoint, but can predict it through a straightforward adaptation of the basic Euler approximation:

$$u\left(t_k + \tfrac{1}{2}\,h\right) \;\approx\; u_k + \tfrac{1}{2}\,h\,F(t_k, u_k).$$

The result is the *Midpoint Method*

$$u_{k+1} = u_k + h\,F\left(t_k + \tfrac{1}{2}\,h, u_k + \tfrac{1}{2}\,h\,F(t_k, u_k)\right). \qquad (5.31)$$

A comparison of the terms in the Taylor expansions of (5.18) and (5.31) reveals that the Midpoint Method is also of second order.

[†] We use subscripts to indicate partial derivatives to save space.

Runge–Kutta Methods

The Improved Euler and Midpoint Methods are the most elementary incarnations of a general class of numerical schemes for ordinary differential equations that were first systematically studied by the German mathematicians Carle Runge and Martin Kutta in the late nineteenth century. Runge–Kutta Methods are by far the most popular and powerful general-purpose numerical methods for integrating ordinary differential equations. While not appropriate in all possible situations, Runge–Kutta schemes are surprisingly robust, performing efficiently and accurately in a wide variety of problems. Barring significant complications, they are the method of choice in most basic applications. They comprise the engine that powers most computer software for solving general initial value problems for systems of ordinary differential equations.

The most general *Runge–Kutta Method* takes the form

$$u_{k+1} = u_k + h \sum_{i=1}^{m} c_i\, F(t_{i,k}, u_{i,k}), \tag{5.32}$$

where m counts the number of *terms* in the method. Each $t_{i,k}$ denotes a point in the k^{th} mesh interval, so $t_k \leq t_{i,k} \leq t_{k+1}$. The second argument $u_{i,k} \approx u(t_{i,k})$ can be viewed as an approximation to the solution at the point $t_{i,k}$, and so is computed by a similar, but simpler formula of the same type. There is a lot of flexibility in the design of the method, through choosing the coefficients c_i, the times $t_{i,k}$, as well as the scheme (and all parameters therein) used to compute each of the intermediate approximations $u_{i,k}$. As always, the *order* of the method is fixed by the power of h to which the Taylor expansions of the numerical method (5.32) and the actual solution (5.18) agree. Clearly, the more terms we include in the Runge–Kutta formula (5.32), the more free parameters available to match terms in the solution's Taylor series, and so the higher the potential order of the method. Thus, the goal is to arrange the parameters so that the method has a high order of accuracy, while, simultaneously, avoiding unduly complicated, and hence computationally costly, formulae.

Both the Improved Euler and Midpoint Methods are instances of a family of two term Runge–Kutta Methods

$$\begin{aligned} u_{k+1} &= u_k + h \left[\, a F(t_k, u_k) + b F\big(\, t_{k,2}, u_{k,2}\,\big)\,\right] \\ &= u_k + h \left[\, a F(t_k, u_k) + b F\big(\, t_k + \lambda h, u_k + \lambda h F(t_k, u_k)\,\big)\,\right], \end{aligned} \tag{5.33}$$

based on the current mesh point, so $t_{k,1} = t_k$, and one intermediate point $t_{k,2} = t_k + \lambda h$ with $0 \leq \lambda \leq 1$. The basic Euler Method is used to approximate the solution value

$$u_{k,2} = u_k + \lambda h F(t_k, u_k)$$

at $t_{k,2}$. The Improved Euler Method sets $a = b = \frac{1}{2}$ and $\lambda = 1$, while the Midpoint Method corresponds to $a = 0$, $b = 1$, $\lambda = \frac{1}{2}$. The range of possible values for a, b and λ is found by matching the Taylor expansion

$$\begin{aligned} u_{k+1} &= u_k + h \left[\, a F(t_k, u_k) + b F\big(\, t_k + \lambda h, u_k + \lambda h F(t_k, u_k)\,\big)\,\right] \\ &= u_k + h\,(a+b)\, F(t_k, u_k) + h^2 b \lambda \left[\, \frac{\partial F}{\partial t}\,(t_k, u_k) + F(t_k, u_k)\, \frac{\partial F}{\partial u}\,(t_k, u_k) \right] + \ \cdots\ . \end{aligned}$$

(in powers of h) of the right hand side of (5.33) with the Taylor expansion (5.18) of the solution, namely

$$u(t_{k+1}) = u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,\left[\,F_t(t_k,u_k) + F(t_k,u_k)\,F_u(t_k,u_k)\,\right] + \ \cdots\ ,$$

to as high an order as possible. First, the constant terms, u_k, are the same. For the order h and order h^2 terms to agree, we must have, respectively,

$$a + b = 1, \qquad\qquad b\,\lambda = \tfrac{1}{2}.$$

Therefore, setting

$$a = 1 - \mu, \qquad b = \mu, \qquad \text{and} \qquad \lambda = \frac{1}{2\,\mu}\,, \qquad \text{where } \mu \text{ is arbitrary}^{\dagger},$$

leads to the following family of two term, second order Runge–Kutta Methods:

$$u_{k+1} = u_k + h\,\left[\,(1-\mu)\,F(t_k,u_k) + \mu\,F\left(t_k + \frac{h}{2\,\mu}, u_k + \frac{h}{2\,\mu}\,F(t_k,u_k)\right)\right]. \tag{5.34}$$

The case $\mu = \frac{1}{2}$ corresponds to the Improved Euler Method (5.28), while $\mu = 1$ yields the Midpoint Method (5.31). Unfortunately, none of these methods are able to match all of the third order terms in the Taylor expansion for the solution, and so we are left with a one-parameter family of two step Runge–Kutta Methods, all of second order, that include the Improved Euler and Midpoint schemes as particular instances. The methods with $\frac{1}{2} \leq \mu \leq 1$ all perform more or less comparably, and there is no special reason to prefer one over the other.

To construct a third order Runge–Kutta Method, we need to take at least $m \geq 3$ terms in (5.32). A rather intricate computation (best done with the aid of computer algebra) will produce a range of valid schemes; the results can be found in [**12**, **18**]. The algebraic manipulations are rather tedious, and we leave a complete discussion of the available options to a more advanced treatment. In practical applications, a particularly simple fourth order, four term formula has become the most used. The method, often abbreviated as RK4, takes the form

$$u_{k+1} = u_k + \frac{h}{6}\,\left[\,F(t_k,u_k) + 2\,F(t_{2,k},u_{2,k}) + 2\,F(t_{3,k},u_{3,k}) + F(t_{4,k},u_{4,k})\,\right], \tag{5.35}$$

where the times and function values are successively computed according to the following procedure:

$$\begin{aligned} t_{2,k} &= t_k + \tfrac{1}{2}\,h, & u_{2,k} &= u_k + \tfrac{1}{2}\,h\,F(t_k,u_k),\\ t_{3,k} &= t_k + \tfrac{1}{2}\,h, & u_{3,k} &= u_k + \tfrac{1}{2}\,h\,F(t_{2,k},u_{2,k}),\\ t_{4,k} &= t_k + h, & u_{4,k} &= u_k + h\,F(t_{3,k},u_{3,k}). \end{aligned} \tag{5.36}$$

The four term RK4 scheme (5.35–36) is, in fact, a fourth order method. This is confirmed by demonstrating that the Taylor expansion of the right hand side of (5.35) in powers of

[†] Although we should restrict $\mu \geq \frac{1}{2}$ in order that $0 \leq \lambda \leq 1$.

h matches all of the terms in the Taylor series for the solution (5.18) up to and including those of order h^4, and hence the local error is of order h^5. This is not a computation for the faint-hearted — bring lots of paper and erasers, or, better yet, a good computer algebra package! The RK4 scheme is but one instance of a large family of fourth order, four term Runge–Kutta Methods, and by far the most popular owing to its relative simplicity.

Example 5.8. Application of the RK4 Method (5.35–36) to our favorite initial value problem (5.8) leads to the following errors at the indicated times:

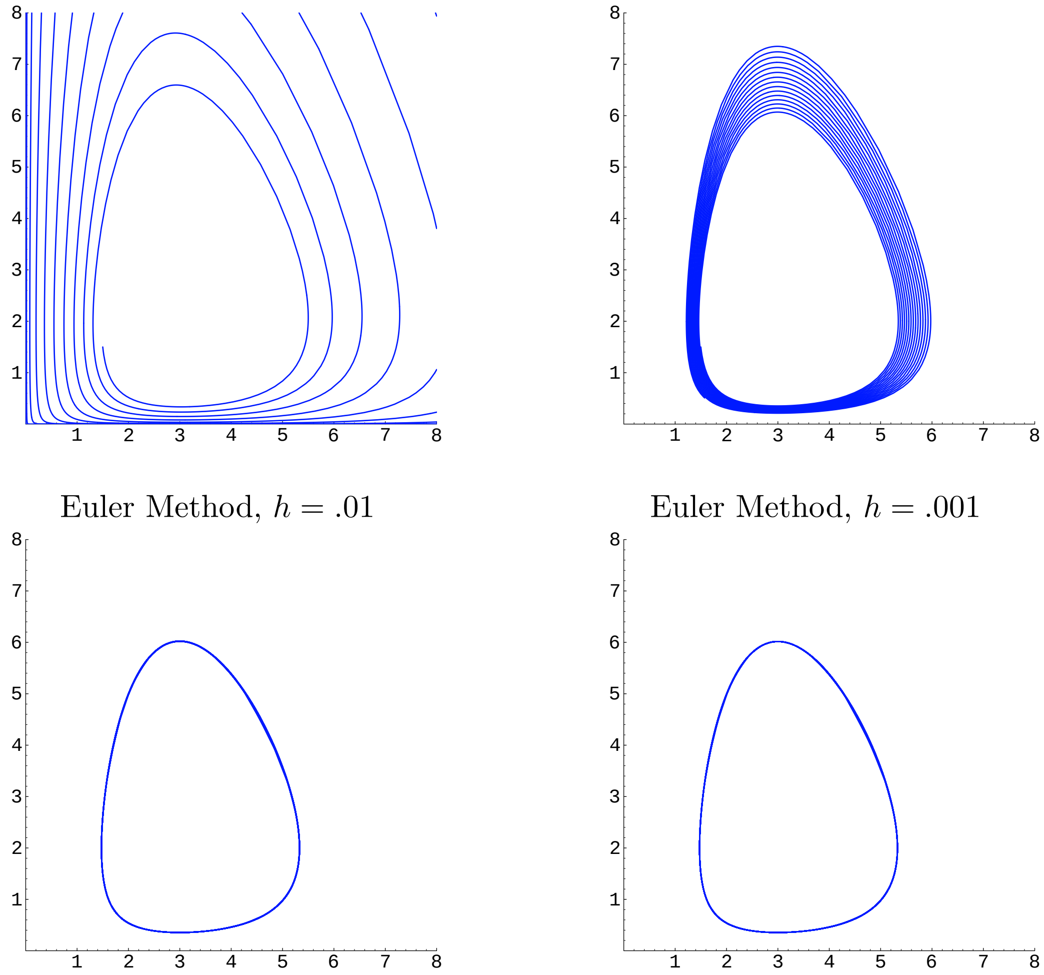
h	$E(1)$	$E(2)$	$E(3)$
.100	-1.944×10^{-7}	1.086×10^{-6}	4.592×10^{-6}
.010	-1.508×10^{-11}	1.093×10^{-10}	3.851×10^{-10}
.001	-1.332×10^{-15}	-4.741×10^{-14}	1.932×10^{-14}

The accuracy is phenomenally good — much better than any of our earlier numerical schemes. Each decrease in the step size by a factor of $\frac{1}{10}$ results in about 4 additional decimal digits of accuracy in the computed solution, in complete accordance with its status as a fourth order method.

Actually, it is not entirely fair to compare the accuracy of the methods using the same step size. Each iteration of the RK4 Method requires four evaluations of the function $F(t,u)$, and hence takes the same computational effort as four Euler iterations, or, equivalently, two Improved Euler iterations. Thus, the more revealing comparison would be between RK4 at step size h, Euler at step size $\frac{1}{4}h$, and Improved Euler at step size $\frac{1}{2}h$, as these involve roughly the same amount of computational effort. The resulting errors $E(1)$ at time $t = 1$ are listed in the following table.

Thus, even taking computational effort into account, the Runge–Kutta Method continues to outperform its rivals. At a step size of .1, it is almost as accurate as the Improved Euler Method with step size .0005, and hence 200 times as much computation, while the Euler Method would require a step size of approximately $.24 \times 10^{-6}$, and would be $4,000,000$ times as slow as Runge–Kutta! With a step size of .001, RK4 computes a solution value that is near the limits imposed by machine accuracy (in single precision arithmetic). The superb performance level and accuracy of the RK4 Method immediately explains its popularity for a broad range of applications.

h	Euler	Improved Euler	Runge–Kutta 4
.1	1.872×10^{-2}	-1.424×10^{-4}	-1.944×10^{-7}
.01	1.874×10^{-3}	-1.112×10^{-6}	-1.508×10^{-11}
.001	1.870×10^{-4}	-1.080×10^{-8}	-1.332×10^{-15}



Euler Method, $h = .01$

Euler Method, $h = .001$

Improved Euler Method, $h = .01$

RK4 Method, $h = .01$

Figure 19. Numerical Solutions of Predator–Prey Model.

Example 5.9. As noted earlier, by writing the function values as vectors $\mathbf{u}_k \approx \mathbf{u}(t_k)$, one can immediately use all of the preceding methods to integrate initial value problems for first order systems of ordinary differential equations $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. Consider, by way of example, the Lotka–Volterra system

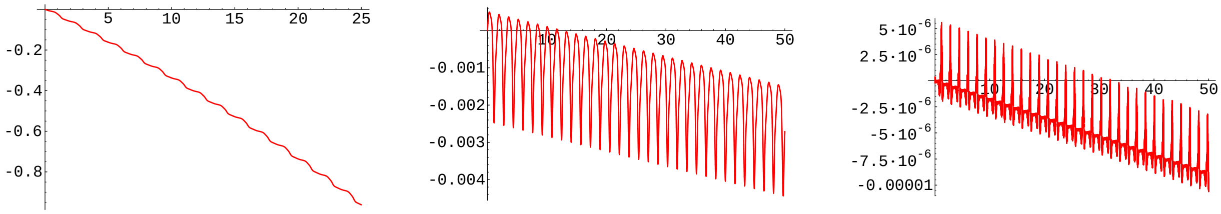
$$\frac{du}{dt} = 2u - uv, \qquad \frac{dv}{dt} = -9v + 3uv, \tag{5.37}$$

analyzed in Example 4.14. To find a numerical solution, we write $\mathbf{u} = (u, v)^T$ for the solution vector, while $\mathbf{F}(\mathbf{u}) = (2u - uv, -9v + 3uv)^T$ is the right hand side of the system. The Euler Method with step size h is given by

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + h\,\mathbf{F}(\mathbf{u}^{(k)}),$$

or, explicitly, as a first order nonlinear iterative system

$$u^{(k+1)} = u^{(k)} + h\,(2u^{(k)} - u^{(k)}\,v^{(k)}), \qquad v^{(k+1)} = v^{(k)} + h\,(-9v^{(k)} + 3u^{(k)}\,v^{(k)}).$$



Euler Method
$h = .001$

Improved Euler Method
$h = .01$

RK4 Method
$h = .01$

Figure 20. Numerical Evaluation of Lotka–Volterra First Integral.

The Improved Euler and Runge–Kutta schemes are implemented in a similar fashion. Phase portraits of the three numerical algorithms starting with initial conditions $u^{(0)} = v^{(0)} = 1.5$, and up to time $t = 25$ in the case of the Euler Method, and $t = 50$ for the other two, appear in Figure 19. Recall that the solution is supposed to travel periodically around a closed curve, which is the level set

$$I(u,v) = 9 \log u - 3\,u + 2 \log v - v = I(1.5, 1.5) = -1.53988$$

of the first integral. The Euler Method spirals away from the exact periodic solution, whereas the Improved Euler and RK4 Methods perform rather well. Since we do not have an analytic formula for the solution, we are not able to measure the error exactly. However, the known first integral is supposed to remain constant on the solution trajectories, and so one means of monitoring the accuracy of the solution is to track the variation in the numerical values of $I(u^{(k)}, v^{(k)})$. These are graphed in Figure 20; the Improved Euler keeps the value within .0005, while in the RK4 solution, the first integral only experiences change in its the fifth decimal place over the indicated time period. Of course, the longer one continues to integrate, the more error will gradually creep into the numerical solution. Still, for most practical purposes, the RK4 solution is indistinguishable from the exact solution.

In practical implementations, it is important to monitor the accuracy of the numerical solution, so to gauge when to abandon an insufficiently precise computation. Since accuracy is dependent upon the step size h, one may try adjusting h so as stay within a preassigned error. *Adaptive methods*, allow one to change the step size during the course of the computation, in response to some estimation of the overall error. Insufficiently accurate numerical solutions would necessitate a suitable reduction in step size (or increase in the order of the scheme). On the other hand, if the solution is more accurate than the application requires, one could increase the step size so as to reduce the total amount of computational effort.

How might one decide when a method is giving inaccurate results, since one presumably does not know the true solution and so has nothing to directly test the numerical approximation against? One useful idea is to integrate the differential equation using two different numerical schemes, usually of different orders of accuracy, and then compare the results. If the two solution values are reasonably close, then one is usually safe in assuming that the methods are both giving accurate results, while in the event that they differ beyond some preassigned tolerance, then one needs to re-evaluate the step size. The required adjustment to the step size relies on a more detailed analysis of the error terms. Several well-studied methods are employed in practical situations; the most popular is the Runge–Kutta–Fehlberg Method, which combines a fourth and a fifth order Runge–Kutta scheme for error control. Details can be found in more advanced treatments of the subject, e.g., [**12**, **18**].

Stiff Differential Equations

While the fourth order Runge–Kutta Method with a sufficiently small step size will successfully integrate a broad range of differential equations — at least over not unduly long time intervals — it does occasionally experience unexpected difficulties. While we have not developed sufficiently sophisticated analytical tools to conduct a thorough analysis, it will be instructive to look at why a breakdown might occur in a simpler context.

Example 5.10. The elementary linear initial value problem

$$\frac{du}{dt} = -\,250\,u, \qquad\qquad u(0) = 1, \tag{5.38}$$

is an instructive and sobering example. The explicit solution is easily found; it is a very rapidly decreasing exponential: $u(t) = e^{-250\,t}$.

$$u(t) = e^{-250\,t} \qquad\text{with}\qquad u(1) \;\approx\; 2.69 \times 10^{-109}.$$

The following table gives the result of approximating the solution value $u(1) \approx 2.69\times 10^{-109}$ at time $t = 1$ using three of our numerical integration schemes for various step sizes:

h	Euler	Improved Euler	RK4
.1	6.34×10^{13}	3.99×10^{24}	2.81×10^{41}
.01	4.07×10^{17}	1.22×10^{21}	1.53×10^{-19}
.001	1.15×10^{-125}	6.17×10^{-108}	2.69×10^{-109}

The results are not misprints! When the step size is .1, the computed solution values are perplexingly large, and appear to represent an exponentially growing solution — the complete opposite of the rapidly decaying true solution. Reducing the step size beyond a critical threshold suddenly transforms the numerical solution to an exponentially decaying function. Only the fourth order RK4 Method with step size $h = .001$ — and hence a total of $1,000$ steps — does a reasonable job at approximating the correct value of the solution at $t = 1$.

You may well ask, what on earth is going on? The solution couldn't be simpler — why is it so difficult to compute it? To understand the basic issue, let us analyze how the Euler Method handles such simple differential equations. Consider the initial value problem

$$\frac{du}{dt} = \lambda\,u, \qquad\qquad u(0) = 1, \tag{5.39}$$

which has an exponential solution

$$u(t) = e^{\lambda t}. \tag{5.40}$$

As in Example 5.1, the Euler Method with step size h relies on the iterative scheme

$$u_{k+1} = (1 + \lambda h)\, u_k, \qquad\qquad u_0 = 1,$$

with solution

$$u_k = (1 + \lambda h)^k. \tag{5.41}$$

If $\lambda > 0$, the exact solution (5.40) is exponentially growing. Since $1+\lambda h > 1$, the numerical iterates are also growing, albeit at a somewhat slower rate. In this case, there is no inherent surprise with the numerical approximation procedure — in the short run it gives fairly accurate results, but eventually trails behind the exponentially growing solution.

On the other hand, if $\lambda < 0$, then the exact solution is exponentially decaying and positive. But now, if $\lambda h < -2$, then $1 + \lambda h < -1$, and the iterates (5.41) grow exponentially fast in magnitude, with alternating signs. In this case, the numerical solution is nowhere close to the true solution; this explains the previously observed pathological behavior. If $-1 < 1 + \lambda h < 0$, the numerical solutions decay in magnitude, but continue to alternate between positive and negative values. Thus, to correctly model the qualitative features of the solution and obtain a numerically respectable approximation, we need to choose the step size h so as to ensure that $0 < 1 + \lambda h$, and hence $h < -1/\lambda$ when $\lambda < 0$. For the value $\lambda = -250$ in the example, then, we must choose $h < \frac{1}{250} = .004$ in order that the Euler Method give a reasonable numerical answer. A similar, but more complicated analysis applies to any of the Runge–Kutta schemes.

Thus, the numerical methods for ordinary differential equations exhibit a form of conditional stability. Paradoxically, the larger negative λ is — and hence the faster the solution tends to a trivial zero equilibrium — the *more* difficult and expensive the numerical integration.

The system (5.38) is the simplest example of what is known as a *stiff differential equation*. In general, an equation or system is stiff if it has one or more very rapidly decaying solutions. In the case of autonomous (constant coefficient) linear systems $\dot{\mathbf{u}} = A\,\mathbf{u}$, stiffness occurs whenever the coefficient matrix A has an eigenvalue with a large negative real part: $\operatorname{Re}\lambda \ll 0$, resulting in a very rapidly decaying eigensolution. It only takes one such eigensolution to render the equation stiff, and ruin the numerical computation of even the well behaved solutions! Curiously, the component of the actual solution corresponding to such large negative eigenvalues is almost irrelevant, as it becomes almost instantaneously tiny. However, the presence of such an eigenvalue continues to render the numerical solution to the system very difficult, even to the point of exhausting any available computing resources. Stiff equations require more sophisticated numerical procedures to integrate, and we refer the reader to [**12**, **18**] for details.

Most of the other methods derived above also suffer from instability due to stiffness of the ordinary differential equation for sufficiently large negative λ. Interestingly, stability for solving the trivial test scalar ordinary differential equation (5.39) suffices to characterize acceptable step sizes h, depending on the size of λ, which, in the case of systems, becomes the eigenvalue. The analysis is not so difficult, owing to the innate simplicity of the test ordinary differential equation (5.39). A significant exception, which also illustrates the test for behavior under rapidly decaying solutions, is the Trapezoid Method (5.25). Let us analyze the behavior of the resulting numerical solution to (5.39). Substituting $f(t,u) = \lambda\, u$ into the Trapezoid iterative equation (5.25), we find

$$u_{k+1} = u_k + \tfrac{1}{2}\, h \left[\, \lambda\, u_k + \lambda\, u_{k+1} \,\right],$$

which we solve for

$$u_{k+1} = \frac{1 + \frac{1}{2}\, h\lambda}{1 - \frac{1}{2}\, h\lambda}\, u_k \equiv \mu\, u_k.$$

Thus, the behavior of the solution is entirely determined by the size of the coefficient μ. If $\lambda > 0$, then $\mu > 1$ and the numerical solution is exponentially growing, as long as the denominator is positive, which requires $h < 2/\lambda$ to be sufficiently small. In other words, rapidly growing exponential solutions require reasonably small step sizes to accurately compute, which is not surprising. On the other hand, if $\lambda < 0$, then $|\,\mu\,| < 1$, *no matter how large negative λ gets*! (But we should also restrict $h < -2/\lambda$ to be sufficiently small, as otherwise $\mu < 0$ and the numerical solution would have oscillating signs, even though it is decaying, and hence vanishing small. If this were part of a larger system, such minor oscillations would not worry us because they would be unnoticeable in the long run.) Thus, the Trapezoid Method is *not* affected by very large negative exponents, and hence not subject to the effects of stiffness.

The Trapezoid Method is the simplest example of an A stable method. More precisely, a numerical solution method is called *A stable* if the zero solution is asymptotically stable for the iterative equation resulting from the numerical solution to the ordinary differential equation $\dot{u} = \lambda\, u$ for all $\lambda < 0$. The big advantage of A stable methods is that they are not affected by stiffness. Unfortunately, A stable methods are few and far between. In fact, they are all implicit one-step methods! *No explicit Runge–Kutta Method is A stable*; see [**18**] for a proof of this disappointing result. Moreover, multistep methods also suffer from the lack of A stability and so are all prone to the effects of stiffness. Still, when confronted with a seriously stiff equation, one should discard the sophisticated methods and revert to a low order, but A stable scheme like the Trapezoid Method.

References

[**1**] Alligood, K.T., Sauer, T.D., and Yorke, J.A., *Chaos. An Introduction to Dynamical Systems*, Springer-Verlag, New York, 1997.

[**2**] Birkhoff, G., and Rota, G.–C., *Ordinary Differential Equations*, Blaisdell Publ. Co., Waltham, Mass., 1962.

[**3**] Bronstein, M., *Symbolic integration I*: *Transcendental Functions*, Springer–Verlag, New York, 1997.

[**4**] Burden, R.L., and Faires, J.D., *Numerical Analysis*, Seventh Edition, Brooks/Cole, Pacific Grove, CA, 2001.

[**5**] Cantwell, B.J., *Introduction to Symmetry Analysis*, Cambridge University Press, Cambridge, 2003.

[**6**] Chenciner, A., and Montgomery, R., A remarkable periodic solution of the three-body problem in the case of equal masses, *Ann. Math.* **152** (2000), 881–901.

[**7**] Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, vol. I, Interscience Publ., New York, 1953.

[**8**] Devaney, R.L., *An Introduction to Chaotic Dynamical Systems*, Addison–Wesley, Redwood City, Calif., 1989.

[**9**] Diacu, F., *An Introduction to Differential Equations*, W.H. Freeman and Co., New York, 2000.

[**10**] Dirac, P.A.M., *The Principles of Quantum Mechanics*, 3rd ed., Clarendon Press, Oxford, 1947.

[**11**] Goldstein, H., *Classical Mechanics*, Second Edition, Addison–Wesley, Reading, Mass., 1980.

[**12**] Hairer, E., Nørsett, S.P., and Wanner, G., *Solving Ordinary Differential Equations*, 2nd ed., Springer–Verlag, New York, 1993–1996.

[**13**] Hale, J.K., *Ordinary Differential Equations*, Second Edition, R.E. Krieger Pub. Co., Huntington, N.Y., 1980.

[**14**] Hille, E., *Ordinary Differential Equations in the Complex Domain*, John Wiley & Sons, New York, 1976.

[**15**] Hirsch, M.W., and Smale, S., *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[**16**] Hydon, P.E., *Symmetry Methods for Differential Equations*, Cambridge Texts in Appl. Math., Cambridge University Press, Cambridge, 2000.

[**17**] Ince, E.L., *Ordinary Differential Equations*, Dover Publ., New York, 1956.

[**18**] Iserles, A., *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge, 1996.

[**19**] Landau, L.D., and Lifshitz, E.M., *Quantum Mechanics* (*Non-relativistic Theory*), Course of Theoretical Physics, vol. 3, Pergamon Press, New York, 1977.

[**20**] Mather, J.N., and McGehee, R., *Solutions of the collinear four body problem which become unbounded in finite time*, Dynamical Systems, Theory and Applications; J. Moser, ed., Springer, Berlin, 1975, pp. 573–597..

[**21**] Messiah, A., *Quantum Mechanics*, John Wiley & Sons, New York, 1976.

[**22**] Montaldi, J., and Steckles, K., Classification of symmetry groups for planar n-body choreographies, preprint, 2013, `arXiv:1305.0470v2`.

[**23**] Moon, F.C., *Chaotic Vibrations*, John Wiley & Sons, New York, 1987.

[**24**] Noether, E., Invariante Variationsprobleme, *Nachr. König. Gesell. Wissen. Göttingen, Math.–Phys. Kl.* (1918), 235–257. (See Kosmann-Schwarzbach, Y., *The Noether Theorems. Invariance and Conservation Laws in the Twentieth Century*, Springer, New York, 2011, for an English translation.)

[**25**] Olver, F.W.J., Lozier, D.W., Boisvert, R.F., and Clark, C.W., eds., *NIST Handbook of Mathematical Functions*, Cambridge University Press, Cambridge, 2010.

[**26**] Olver, P.J., *Applications of Lie Groups to Differential Equations*, 2nd ed., Graduate Texts in Mathematics, vol. 107, Springer–Verlag, New York, 1993.

[**27**] Olver, P.J., and Shakiban, C., *Applied Linear Algebra*, Second Edition, Undergraduate Texts in Mathematics, Springer, New York, 2018.

[**28**] Whittaker, E.T., *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Cambridge University Press, Cambridge, 1937.