Engineering Differential Equations: Theory and Applications

Bill Goodwine

March 19, 2008

Preface

This book is the result of course notes that were created for a sequence of new courses in the Department of Aerospace and Mechanical Engineering at the University of Notre Dame. The new sequence of courses was comprised of two courses, titled Differential Equations, Vibrations and Control I and II, which cover material typically covered in three engineering courses: differential equations, vibrations and controls (obviously).

The consolidation of the three courses into two was brought about to streamline the curriculum for undergraduate students in our department. It was felt, by me at least, that if the most direct engineering applications were presented in conjunction with the study of differential equations and the associated solution techniques, that students would be more motivated to study the material and hence learn and retain it better. Also, some consolidation could be accomplished because the inevitable review of the relevant differential equations subject matter in controls and vibrations would be obviated. Since the author was the primary advocate of this consolidation, he was naturally assigned to develop the courses. This book is the result.

With regard to the efficacy of this approach, it is clear, based upon graduation survey results, that the students are very satisfied with it. The author can state, based upon teaching experience, that the amount of controls material that can be covered and understood is the same as when controls was an independent course. With respect to the students' ultimate understanding and retention of the differential equations subjects as a purely mathematical matter the result is less clear. Informal investigations by the author indicate that that it is neither drastically better nor worse in comparison to the prior situation when differential equations was an independent course. So, at least to the extent that consolidation was accomplished the result is successful since the students seem no worse off than before and the material is covered in two courses instead of three; furthermore, it is successful to the extent that student satisfaction is greatly increased.

Prerequisites

The student is assumed to have a good background in calculus and perhaps an introduction to linear algebra. A dynamics course would be useful, but the basic mechanics from the typical undergraduate engineering physics sequence seems to suffice.

Course structure

The material is organized in a manner that is most logical from my perspective. However, curricular realities prevent me from teaching it in the sequential order of the chapters. In particular, it seems logical to consider first order systems, second order systems, systems of first order equations (nth order equations) and then infinite dimensional systems (partial differential equations). However, due to the fact that partial differential equations need to be covered in the first semester in our curriculum, that subject is covered before systems of first order equations when I teach the courses. What I typically cover is organized as follows.

First semester

- Chapter 1, introduction: classification of differential equations (sections 1.5), the definition of different types of solutions (section 1.6) and an introduction to numerical methods (section and 1.10)
- Chapter 2, first order ordinary differential equations: fast review.
- Chapter 3, second order, constant coefficient ordinary differential equations: covered in detail.
- Chapter 4, linear oscillations: covered in detail.
- Chapter 9, introduction to PID control (section 9.2).
- Chapter 13, numerical methods: covered in detail.
- Chapter 12, separation of variables for partial differential equations: covered in detail, with the numerical methods subjects covered in parallel.
- Chapter 14, nonlinear equations and linear approximations: time permitting.

Second semester

- Chapter 6, systems of first order ordinary differential equations: covered in detail.
- Chapter 7, multiple degree of freedom linear oscillations, covered in detail.
- Chapter 8, Laplace transform methods, covered in detail.
- Chapter 9, classical control theory, covered in detail.
- Chapter 16, Lagrange's equations, covered in detail.

What is not covered

The most conspicuous differential equations subject that is not covered in this text is the use of power series solutions, particularly as applied to second order linear ordinary differential equations with variable coefficients. Also, the controls material is limited to the most basic subjects in classical control, transfer functions, the root locus design method and frequency analysis ("Bode plots")

Bill Goodwine
University of Notre Dame
Notre Dame, Indiana
USA

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Preliminaries

1.1 The Engineering Utility of Differential Equations

Nearly all the fundamental principles that govern physical processes of engineering interest are described by differential equations. Hence, it is fair to say that the ability to analyze, solve and understand differential equations is fundamentally important for engineers. This book is intended to make differential equations more accessible to engineering students by presenting and developing some application areas in parallel with the presentation of the mathematics. This is done sometimes by way of simply using the application as a motivational problem and other times by fully developing the application material. The main two applications areas in this book are mechanical vibrations and basic feedback control theory. Those two areas are completely presented. Many other applications areas are also presented in the book, but are not presented in a necessarily comprehensive manner.

Additionally, there is an emphasis on analyzing the solutions to each problem, *e.g.,* instead of the "answer" being simply a mathematical expression of the form

$$x(t) = \frac{F_0\left(k - m\omega^2\right)}{\left(k - m\omega\right)^2 + \left(c\omega\right)^2}\cos\omega t + \frac{c\omega F_0}{\left(k - m\omega\right)^2}\sin\omega t \tag{1.1}$$

the question may be to determine the frequency, ω, at which $x(t)$ obtains the greatest magnitude, which, of course, requires not only determining the "answer" in Equation 1.1, but analyzing it as well.

1.2 Mathematical Approach of this Book

The main approach to categorize differential equation solution methods in this book is to distinguish differential equations using the following five criteria:

1. whether the equation is *ordinary* or *partial*;
2. the *order* of the equation;
3. whether the equation is *linear* or *nonlinear*;
4. if the equation is linear, whether the equation is *homogeneous* or *inhomogeneous*; and,
5. if the equation is linear, whether the equation has *constant* or *variable* coefficients.

Without elaborating upon any of these distinctions, and for the moment restricting our attention to only first and second order equations, it is apparent that already we are dealing with $2^5 = 32$ different possible categorizations. Fortunately, some solution techniques apply to more than one category of equations. Also, for the purposes of this book, some of the distinctions are only important when coupled with the others. Hence learning 32 different solution methods is not necessary. However, complicating matters is the fact that some categories have more than one solution method and which one is desirable depends, perhaps, upon which is simply the easiest to apply.

This book will outline a rather conventional set of solution methods with more emphasis than usual on the mundane, but crucial, ability of categorizing the equation and with much greater emphasis than usual on the immediate engineering applications of each category of differential equation.

1.3 Notation

Throughout this book, the following rules for notation are usually observed.

1. Dots above variables indicate differentiation with respect to time, *e.g.*,

$$\begin{aligned}\dot{x}(t) &= \frac{dx}{dt}(t)\\ \ddot{x}(t) &= \frac{d^2x}{dt^2}(t).\end{aligned}$$

2. A natural number in parentheses as a superscript to a function indicates the number of times the function is differentiated with respect to the independent variable, *e.g.,*

$$x^{(n)}(t) = \frac{d^nx}{dt^n}(t).$$

3. Generally, the uppercase Roman letters A and B represent matrices, *e.g.*,

$$A = \begin{bmatrix} 2 & 3 & \cdots & 8 \\ 3 & 7 & \cdots & 8 \\ \vdots & & \ddots & \vdots \\ 8 & 2 & \cdots & 9 \end{bmatrix}.$$

4. The letter I usually represents the identity matrix, *i.e.,*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

   The dimension of I is usually obvious from the context of its use. For example, if A is a 3×3 matrix and the expression $A + I$ is used, then I is 3×3.

   The symbol $I(s)$ may also be used to represent the Laplace transform of the variable $i(t)$ representing current.

   The letter I may also represent an interval of real numbers.

5. The letter i usually represents the *imaginary unit*, which satisfies $i^2 = -1$. The letter i may also be used to represent electric current.

6. The letter e represents the base of the natural logarithm, *i.e.,* $e \approx 2.71828$.

7. The letter m usually represents a mass.

8. The letter k usually represents a spring constant.

9. The letter b usually represents a viscous damping constant.

10. The letter R usually represents electrical resistance.

11. The letter C usually represents electrical capacitance.

12. The letter L usually represents electrical inductance.

13. The letter v usually represents a voltage or velocity.

14. The letter λ usually represents an eigenvalue.

15. The letter ξ usually represents a vector.

16. A symbol that is a vector with a hat usually represents an eigenvector of a matrix, *e.g.*, $\hat{\xi}$.

17. The symbol $\mathbb{R}$ represents the set of real numbers.

18. The symbol $\mathbb{C}$ represents the set of complex numbers.

1.4 Sets, Relations and Functions

Most engineering students have a pretty decent grip on the idea of a function, and functions are pretty important in this book because the solution to a differential equation is a function. Slightly more complicated subjects probably need a quick review, however. This section will first deal briefly with *sets* since functions are relationships between sets. Then a function is defined as well as implicit functions. Implicit functions arise in this book because they are the natural representation of a solution to certain differential equations considered in Section 2.3.4. Next follows the definition of multivariable and multivalued functions and a review of their calculus.

1.4.1 Sets

Without getting bogged down in the nuances of basic set theory, we will consider a *set* to be a collection of *elements.*[1] We assume that there is a way for us to determine whether or not an element is in a set[2] and whether or not two elements are equal. Many sets have common names. The two sets we will be most concerned with are the set of real numbers, typically denoted by $\mathbb{R}$ and the set of complex numbers, typically denoted by $\mathbb{C}$. We will often deal with particular subsets, the most common of which are *intervals of* $\mathbb{R}$, such as

$$[a, b] = \{x \in \mathbb{R} \,|a \leq x \leq b\},$$

i.e., real numbers that are either a, b or between a and b, or

$$(a, b] = \{x \in \mathbb{R} \,|a < x \leq b\},$$

i.e., real numbers that are between a and b or are b. An *open interval* is an interval of the form

$$(a, b) = \{x \in \mathbb{R} \,|a < x < b\},$$

where the term "open" connotes the fact that the interval does not include its boundary or endpoints.

Sometimes we will put more than one set together to make a new set. A common way in which this is done is called the *Cartesian product.*

Definition 1.4.1 Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$ be sets. The *Cartesian product* of $\mathcal{D}_1$, $\mathcal{D}_2$, $\ldots$, $\mathcal{D}_n$, is the set

$$\mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_n = \{(x_1, x_2, \ldots, x_n) \,|x_1 \in \mathcal{D}_1, x_2 \in \mathcal{D}_2, \cdots, x_n \in \mathcal{D}_n\}.$$

[1] More precisely, a collection of elements is a *class* and a set is a certain kind of class. A reader interested in the distinction is referred to [11].

[2] *Fuzzy logic* is the branch of logic and mathematics where set theory is generalized to include the notion of partial set membership. In this book an element is either in a set or it is not in the set. In contrast, in fuzzy logic an element may be partially in a set. A classic example of a fuzzy set is the set of "warm days." It is natural to think of some days as "kind of" warm, which is represented in fuzzy logic by a kind of warm day being partially in the set of warm days and partially not in it. There is a vast literature on fuzzy logic and the interested reader is referred to the original paper [22].

Elements of $\mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_n$ are called *n-tuples*. Elements of $\mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_n$ are *ordered* which means that

$$(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_n)$$

if and only if $x_1 = y_1$, $x_2 = y_2$, $\cdots$, and $x_n = y_n$. ⋄

Example 1.4.2 As sets,

$$\{1, 4, 6, 8\}$$

and

$$\{8, 4, 6, 1\}$$

are the same. As ordered sets they are not the same. ■

An example of the way the Cartesian product is used is when vectors are used to represent something.

Example 1.4.3 An example of a Cartesian product is the set of vectors in three dimensional Euclidean space. To specify a point in space, a set of three basis vectors is needed, and the point is then represented by its component along each of these three basis vectors. In this book we will write

$$\xi = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

to represent the point. The set to which this point belongs is

$$\mathbb{R} \times \mathbb{R} \times \mathbb{R} = \mathbb{R}^3.$$ ■

1.4.2 Relations and Functions

In this book, a *relation* between elements of sets may be defined by equation or set of equations. Elements of the sets satisfy the relation if they satisfy the equation. A special kind of relation is a function.

Definition 1.4.4 Given two sets, $\mathcal{D}$ and $\mathcal{R}$, if, for each element of $x \in \mathcal{D}$ there is an assignment of one and only one element of $y \in \mathcal{R}$ then we say that y is a *function* of x. The set $\mathcal{D}$ is called the *domain* and the set $\mathcal{R}$ the *range*.

The variable x denoting an element of the domain is called the *independent variable* and the variable y denoting the elements of the range is called the *dependent variable*. It is common to write $y = f(x)$ to indicate that y is a function of x. ⋄

Note that it may be necessary to specify which set is the domain and which is the range. Of course, we do not usually bother to do that and it is normally clear from the context which set is the domain and which is the range. We will often indirectly specify the domain and range by saying that a function is *from* the domain *to* the range.

Example 1.4.5 If $s = \alpha + i\beta$, the equation

$$r = \|s\| = \sqrt{\alpha^2 + \beta^2}$$

defines a function from the complex numbers to the real numbers (the complex numbers are the domain and the real numbers are the range) since there is one and only one real number for each complex number that satisfies the equation. The equation does not define a function from the real numbers to the complex numbers because for most real numbers, r, there are many complex numbers with $\|s\| = r$. ■

So far we have been considering functions between two sets. Of course, functions may exist between multiple sets, which is manifested in the case where the dependent variable depends upon more than one independent variable. In such a case, the dependent variable is a function of the independent variables if, for each possible combination of the independent variables, there corresponds only one value of the dependent variable. Solutions to partial differential equations are multi-variable functions.

Definition 1.4.6 Given $m + 1$ sets, $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$, and $\mathcal{R}$, element of $x_1 \in \mathcal{D}_1, x_2 \in \mathcal{D}_2, \ldots, x_m \in \mathcal{D}_m$ there corresponds one and only one element of $y \in \mathcal{R}$, then we say that y is a *function* of $x_1, x_2, \ldots, x_m$. The variables $x_1, x_2, \ldots, x_m$ are called the *independent variables* and the variable y denoting the elements of the range is called the *dependent variable*. Using the Cartesian product, the domain is given by

$$\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_m$$

and the function is a function from $\mathcal{D}$ to $\mathcal{R}$. It is common to write $y = f(x_1, x_2, \ldots, x_m)$ to indicate that y is a function of $x_1, x_2, \ldots, x_m$. ⋄

Example 1.4.7 The equation

$$r = \sqrt{x^2 + y^2}$$

defines a function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ since there is only one r for any specified values for x and y. ■

1.4.3 The derivative

The derivative is given by the usual limit definition.

Definition 1.4.8 Let $x(t)$ be a function with the single independent variable t. The *derivative of x with respect to t* is denoted by $\frac{dx}{dt}$ and is defined by

$$\frac{dx}{dt}(t) = \lim_{\Delta t \to 0} \frac{x(t + \Delta t) - x(t)}{\Delta t}.$$ ⋄

Of course the usual interpretation of the derivative is that it is the rate of change of the function with respect to the independent variable. If graphed, it is the slope of the curve of $x(t)$. If the function depends on more than one independent variable, then we must consider the partial derivative.

Definition 1.4.9 Let $x(t_1, \ldots, t_n)$ be a function with independent variables $t_1, \ldots, t_n$. The *partial derivative of x with respect to t_m* is denoted by $\frac{\partial x}{\partial t_m}$ and is defined by

$$\frac{\partial x}{\partial t_m}(t_1, \ldots, t_n) = \lim_{\Delta t \to 0} \frac{x\left(t, \ldots, t_m + \Delta t, \ldots, t_n\right) - x\left(t_1, \ldots, t_m, \ldots, t_n\right)}{\Delta t}. \qquad \diamond$$

This book will use practically all the usual notational means to represent derivatives. Which one is used will typically depend on the conventional notation used by various application areas. In particular, because it can be difficult to interpret an equation with many parentheses, we will often use a "subscript" notation to indicate the values at which a derivative function is evaluated instead of following the function name by parentheses, *i.e.,*

$$\left.\frac{df}{dx}\right|_{x=x_0} = \frac{df}{dx}\left(x_0\right).$$

1.4.4 Implicit Functions

So far things are simple: given an element of the domain, if we have a way to determine one and only one element of the range, then we have a function. In some cases, however, it naturally arises that for a function of more than one variable, we are interested not so much in what element of the range corresponds to elements of the domain, but rather in the relationship among the elements of the domain that correspond to *one* particular element in the range. A circle is an obvious example.

> **Example 1.4.10** Returning to Example 1.4.7, we may consider the set of points that satisfy
>
> $$x^2 + y^2 = 1. \tag{1.2}$$
>
> A plot of all points that satisfy this equation is illustrated in Figure 1.1. ■

In Example 1.4.7 we had a function of two variables, and in Example 1.4.10 we studied the set of points that satisfy $x^2 + y^2 = 1$. This second example defines a *relation*, which is more general than a function. Two points $x \in \mathbb{R}$ and $y \in \mathbb{R}$ satisfy the relation if they satisfy Equation 1.2. Mathematically a relation is defined to be a subset of the domain. For purposes of this book we will consider them to be the subset of the domain that satisfy some equation, such as $f(x, y) = 1$, or $f(x, y) \geq 2$.

It is logical in the second example to study the relationship between x and y beyond simply asking whether or not they satisfy the relation. By referring to Figure 1.1, it is clear that x and y are not related by a function since for any $x \in (-1, 1)$ there are *two* values for y (and *vice-versa*). In the next example, we show that is possible to make the relationship between x and y that satisfy $f(x, y) = 1$ into a function, at least for a limited domain and/or range.

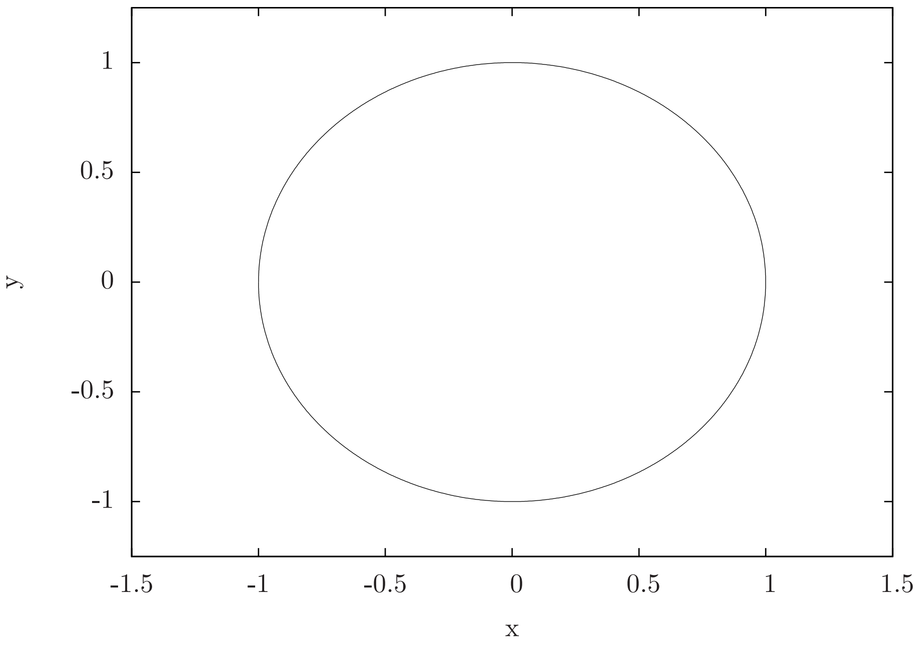


Figure 1.1. A plot of the subset of points in $\mathbb{R}^2$ that satisfy $x^2 + y^2 = 1$.

Example 1.4.11 Consider the set of points that satisfy

$$x^2 + y^2 = 1. \tag{1.3}$$

One way to make this relation into a function is to appropriately restrict the domain and range. It is clear from Figure 1.1 that, at most, the domain must be limited at least to the interval $\mathcal{D} = [-1, 1]$. With respect to the range, it must also be restricted so that only the top half or bottom half of the circle is included in the range.

So, in this example, Equation 1.3 defines a function $y = f(x)$ if we restrict the domain to be

$$\mathcal{D} = \{x \in \mathbb{R} \,|{-1} \leq x \leq 1\}$$

and specify either

$$y = \sqrt{1 - x^2}$$

or

$$y = -\sqrt{1 - x^2},$$

which corresponds to either the top or bottom half of the circle, respectively. ■

Do not infer from Example 1.4.7 that it will always be the case that an equation that defines an implicit function can be "solved" for one of the variables. In fact doing so will typically be difficult.

In Example 1.4.7 we were able to solve for y in terms of x for some region of the domain. Motivated by this we define an *implicit function* as follows.

Definition 1.4.12 The equation $f\left(x, y\right) = c$ where c is a constant, defines an implicit function if and only if there exists a function $g\left(x\right)$ such that

$$f\left(x, g\left(x\right)\right) = c.$$

⋄

The more natural way to write this is to write $g(x) = y(x)$, *i.e.,* the y variable is actually a function of x. The idea is, that we want to be able to ask how y should change if we want to vary x and simultaneously require $f(x, y) = c$. To determine when such a $y(x)$ will exist is not too hard. To do so, we differentiate $f\left(x, y\left(x\right)\right) = c$ with respect to x and solve for $\frac{dy}{dx}$ (how y changes with x) as follows. Differentiating and using the chain rule for the second component gives

$$\frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}.$$

Since $y(x)$ is defined so that $f(x, y(x)) = c$, then

$$\frac{df}{dx} = 0$$

and hence

$$\frac{dy}{dx} = -\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}}. \tag{1.4}$$

Intuitively, in order to determine how y should change as a function of x, we need that the denominator on the right hand side of Equation 1.4 be nonzero. In fact, this is exactly what is required and is the basis for the proof of the following theorem.

Theorem 1.4.13 *Let $f(x,y)$ be a continuously differentiable real-valued function defined on an open set and let (x_0, y_0) be a point such that $f(x_0, y_0) = c$ and such that*

$$\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} \neq 0.$$

Then there exists a function $y(x)$ on an open interval containing x_0 such that $y(x)$ is continuously differentiable and

$$f(x, y(x)) = c.$$

This theorem is called the *implicit function theorem* and is one of the most fundamental and useful tools in analysis. Unfortunately for our purposes it is not so useful since it only tells us when y may be considered a function of x, but it does not tell us what that function is.

1.5 Types of Differential Equations

This section provides the basic definitions necessary to categorize a given differential equation (or set of differential equations) according to the five criteria outlined above. The solution methods developed subsequently will only be applicable to certain types of differential equations; hence, it is critical from the beginning to be able to properly categorize them. Before that, however, we must first consider what exactly a "differential equation" is.

Definition 1.5.1 Let $x(t_1, t_2, \ldots, t_m)$ be a function of the m independent variables ~~t_i~~ $\underline{t_1, t_2, \ldots, t_m}$. A *differential equation* is ~~simply~~ an equation that ~~involves $t, x(t_1, \ldots, t_m)$ and~~ contains at least one derivative $\underline{\text{(of any order)}}$ of $x(t_1, \ldots, t_m)$.⋄

Example 1.5.2 The equation

$$\frac{1}{t}\ddot{x}(t) = 3$$

is a differential equation with dependent variable x and independent variable t. ■

Sometimes we will have to consider a set of differential equations, which will be called a system of differential equations.

Definition 1.5.3 Let each function in the set of functions $\{x_1, x_2, \ldots, x_n\}$ be a function of the m independent variables $t_1, t_2, \ldots, t_m$. A *system of differential equation* is a set of n equations that contains at least one derivative of each of the functions in the set $\{x_1, x_2, \ldots, x_n\}$. $\diamond$

Example 1.5.4 The set of equtions

$$\begin{aligned} x_1(t_1, t_2, t_3) + \frac{\partial^5 x_2}{\partial t_1^3 \partial t_2^2}(t) &= \sin 3t_3 \\ \csc(t_1)\frac{\partial x_2}{\partial t_2} &= \frac{\partial^2 x_1}{\partial t_3^2}(t) \end{aligned}$$

is a system of two differential equations with dependent variables x_1 and x_2 and independent variables t_1, t_2 and t_3. ■

In general, because they can be determined from fundamental scientific principles, the differential equation governing a system is known, but the solution is unknown. "Solving" a differential equation amounts to determining the function (dependent variable) of the independent variable which satisfies the differential equation.

1.5.1 Ordinary *vs.* partial differential equations

~~If~~ In a differential equation, if the dependent variable is a function of only one independent variable, then the differential equation is an *ordinary differential equation*. If the dependent variable depends on more than one independent variable then the differential equation is a *partial differential equation.* Generally it is trivial to distinguish between ordinary and partial differential equations since the derivatives are notationally different.

Example 1.5.5 The equation describing a mass-spring-damper system under the influence of a forcing function given by

$$\ddot{x}(t) + 3\dot{x}(t) + 5x(t) = \cos(t) \tag{1.5}$$

is an ordinary differential equation with independent variable t and dependent variable x. ■

Example 1.5.6 The equation that described the shape of a vibrating string

$$\frac{\partial^2 u}{\partial t^2}(x, t) = \frac{\partial^2 u}{\partial x^2}(x, t)$$

where $u(x, t)$ gives the displacement of the string, u, at position x at time t, is a partial differential equation. ■

> A system of differential equastions is ordinary if each of the dependent variables are a function of one and the same independent variable.

Some nuances exist, but generally speaking if there are partial derivative signs in the equation it is a partial differential equation and if there are only ordinary derivative operators (d's) or "dots," *e.g.,* $\dot{x}$ or "primes," *e.g.,* y' then the equation is ordinary.

1.5.2 The order of a differential equation

The *order* of an ordinary differential equation is simply the order of the highest derivative in the equation.

Example 1.5.7 The equation

$$\sin t + x(t) + \ddot{x}(t) = 35\dot{x}(t)\cos(t)$$

is second order. ■

For a partial differential equation, the order is also the order of the highest derivative of the independent variable. We may also express the order with respect to each of the independent variables.

Example 1.5.8 The wave equation

$$\frac{\partial^2 u}{\partial x^2}(x,t) = \frac{\partial^2 u}{\partial t^2}(x,t)$$

is second order in both x and t.

The heat equation

$$\frac{\partial^2 u}{\partial x^2}(x,t) = \frac{\partial u}{\partial t}(x,t)$$

is second order in the independent variable x and first order in the independent variable t. ■

Example 1.5.9 The equation

$$\frac{\partial^3 u}{\partial^2 x \partial t}(x,t) = 5$$

is third order and it is second order with respect to x and first order with respect to t. ■

1.5.3 Linear *vs.* nonlinear differential equations

This is perhaps the most important distinction of all. With the exception of some first order equations and other very specific examples, nonlinear differential equations do not have any known solution techniques; in contrast, linear differential equations have some very nice properties. A differential equation is *linear* if all the terms in the equation are linear in the dependent variable and its derivatives; otherwise, it is *nonlinear*.

Considering first an nth order ordinary differential equation with independent variable t and dependent variable x, if the equation can be put in the form

$$f_n(t)\frac{d^n x}{dt^n}(t) + f_{n-1}(t)\frac{d^{n-1}x}{dt^{n-1}}(t) + \cdots + f_1(t)\frac{dx}{dt}(t) + f_0(t)x(t) = g(t) \quad (1.6)$$

it is linear.

Remark 1.5.10 The functions $f_i(t)$ and $g(t)$ do *not* have to be linear functions of t in order for the equation to be linear. Only linearity in the dependent variable matters. ⋄

Extending this to the partial differential equation case is straight-forward. The equation is linear if all the terms containing the dependent variable or any of its derivatives appears linearly in the equation; otherwise, it is nonlinear.

Considering an nth order partial differential equation with independent variables x and t and dependent variable u, if the equation can be put in the form

$$\sum_{i,j,i+j\leq n} f_{i,j}(x,t)\frac{\partial^{i+j}u}{\partial x^i \partial t^j}(x,t) = g(x,t)$$

it is linear.

Example 1.5.11 The following differential equations are linear or nonlinear as indicated:

$$
\begin{array}{rl}
\ddot{x}(t) + t^2 \sin(t)\, x(t) = 5t & \text{linear} \\
\ddot{x}(t) + t^2 \sin(t)\, x^2(t) = 5t & \text{nonlinear} \\
\ddot{x}(t) + t^2 \sin(t) \sin(x(t)) = 5t & \text{nonlinear} \\
\ddot{x}(t) + t^2 \sin(t)\, x(t) = 5tx(t) & \text{linear} \\
\ddot{x}(t) + 2t\dot{x}(t) = 5x(t) & \text{linear} \\
\ddot{x}(t) + 2\dot{x}(t) = 5x(t) & \text{linear} \\
\ddot{x}(t) + 2x(t)\, t = 5x(t) & \text{linear} \\
\ddot{x}(t) + 2x(t)\, \dot{x}(t) = 5x(t) & \text{nonlinear} \\
\ddot{x}(t) + 2x(t) = 5\sin(t) & \text{linear} \\
\ddot{x}(t) + 2x(t) = 5\sin(t)\, \dot{x}(t) & \text{linear} \\
\ddot{x}(t) + 2x(t) = 5\sin(t) \sin(\dot{x}(t)) & \text{nonlinear} \\
\dfrac{\partial^2 u}{\partial x^2}(x,t) = \dfrac{\partial u}{\partial t}(x,t) & \text{linear} \\
\dfrac{\partial^2 u}{\partial x^2}(x,t) = u\dfrac{\partial u}{\partial t}(x,t) & \text{nonlinear} \\
\dfrac{\partial^2 u}{\partial x^2}(x,t) = \dfrac{\partial u}{\partial t}(x,t)\, \dfrac{\partial u}{\partial t}(x,t) & \text{nonlinear} \\
\dfrac{\partial^2 u}{\partial x^2}(x,t) = \dfrac{\partial u}{\partial t}(x,t) + x & \text{linear} \\
\dfrac{\partial^2 u}{\partial x^2}(x,t) = \dfrac{\partial u}{\partial t}(x,t)\, x + x & \text{linear.}
\end{array}
$$

1.5.4 Homogeneous *vs.* inhomogeneous linear ordinary differential equations

If any of the terms of a linear ordinary differential equation are only a function of the independent variable(s) or are a constant, then the equation is inhomogeneous; otherwise, it is homogeneous. The “terms” of a differential equation are the elements of the equation that are on either side of the equality and that are combined by addition or subtraction. Note that determining whether or not an equation is homogeneous or inhomogeneous will require being precise about what variables are dependent and independent. An equation that is not already in a convenient or standard form may take more than cursory study to determine homogeneity or inhomogeneity.

Example 1.5.12 The following linear ordinary differential equations with dependent variable x and independent variable t are homogeneous or inhomogeneous as indicated:

$$\begin{aligned} \ddot{x}\left(t\right)+t^{2}\sin\left(t\right)x\left(t\right)&=5t &\qquad \text{inhomogeneous}\\ \ddot{x}\left(t\right)+t^{2}\sin\left(t\right)x\left(t\right)&=5tx\left(t\right) &\qquad \text{homogeneous}\end{aligned}$$

1.5.5 Constant *vs.* variable coefficient linear ordinary differential equations

The coefficients in question are the terms which multiply the dependent variable and its derivatives in a linear differential equation. If they are constants the equation is constant coefficient; otherwise, it is variable coefficient. Note that if the equation is inhomogeneous, then there may be terms that are functions of the independent variable, but if they are not coefficients of the dependent variable it will still be a constant coefficient differential equation. Especially in control theory and in dynamical systems, constant coefficient equations are often referred to as *time invariant*.

Example 1.5.13 The following linear ordinary differential equations with dependent variable x and independent variable t are either constant or variable coefficient as indicated:

$$\begin{aligned} \ddot{x}\left(t\right)+t^{2}\sin\left(t\right)x\left(t\right)&=5t &\qquad \text{variable coefficient}\\ \ddot{x}\left(t\right)+t^{2}\sin\left(t\right)x\left(t\right)&=5tx\left(t\right) &\qquad \text{variable coefficient}\\ \ddot{x}\left(t\right)+2\dot{x}\left(t\right)&=5x\left(t\right) &\qquad \text{constant coefficient}\\ \ddot{x}\left(t\right)+2\dot{x}\left(t\right)t&=5x\left(t\right) &\qquad \text{variable coefficient}\\ \ddot{x}\left(t\right)+2x\left(t\right)&=5\sin\left(t\right) &\qquad \text{constant coefficient}\end{aligned}$$

1.6 Solutions of Differential Equations

There are, in fact, several different types of solutions to differential equations. Distinguishing among them is important, not only for fundamentally understanding the subject, but also for avoiding frustration subsequently when "solving" problems so that the right type of solution is actually obtained.

Definition 1.6.1 An *explicit solution* (usually just called "a *solution*") of a differential equation is a function that satisfies the differential equation.

Example 1.6.2 The function, $x(t) = \sin t$ is a solution to

$$\ddot{x} + x = 0. \tag{1.7}$$

Actually, any function of the form $x(t) = c \sin t$, where c is a constant, is a solution to Equation 1.7. ■

Example 1.6.3 The function $x(t) = \cos t + t \sin t$ is an explicit solution to

$$\ddot{x} + x = 2\cos t. \qquad \blacksquare$$

Sometimes a solution is only defined on a particular interval of the independent variable. In such a case the interval must be stated as part of the solution.

Example 1.6.4 The function

$$x(t) = 2\ln t + c$$

where c is an arbitrary constant is a solution to

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

for $t > 0$. ■

Observe that that in Examples 1.6.2 and1.6.4 there are arbitrary constants in the solution. We will also be interested in the case where there are no arbitrary constants.

Definition 1.6.5 A *particular solution* of a differential equation is a function which satisfies the differential equation, but contains no arbitrary constants. ⋄

Example 1.6.6 The function

$$x(t) = 2\ln t + 6$$

is a particular solution to

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

for $t > 0$.

Also, $x(t) = \sin t$ is a particular solution to Equation 1.7 as are $x(t) = 2\sin t$, $x(t) = 3\sin t$, *etc.* ■

Example 1.6.7 The function $x(t) = t\sin t$ is a particular solution to

$$\ddot{x} + x = 2\cos t. \qquad \blacksquare$$

Sometimes we will determine solutions that are described by implicit functions, as is illustrated by the following example. These arise in this book for certain types of first order, nonlinear, ordinary differential equations.

Example 1.6.8 Consider the relation

$$f(x,t) = x^2 + t^2 = c. \tag{1.8}$$

Since

$$\frac{\partial f}{\partial x} = 2x \neq 0$$

as long as $x \neq 0$, $f(x,t) = c$ defines an implicit function $x(t)$ according to Theorem 1.4.13 away from $x = 0$.

For a specified constant c, the implicit function $x(t)$ defined by the relation in Equation 1.8 is a particular solution to

$$2x\dot{x} + 2t = 0. \tag{1.9}$$

■

If c is arbitrary then $x(t)$ is a solution to Equation 1.9.

This is verified by differentiating Equation 1.7 with respect to time and noting that x is a function of t. In particular,

$$\begin{aligned} \frac{df}{dt}(x(t),t) &= \left[\frac{\partial f}{\partial x}(x(t),t)\right]\frac{dx}{dt}(t) + \frac{\partial f}{\partial t}(x(t),t) \\ &= [2x(t)]\,\dot{x}(t) + 2t \\ &= 0. \end{aligned}$$

If some solutions contain arbitrary constants, then it is natural to ask when a solution has enough arbitrary constants to represent *every* solution to a differential equation.

Definition 1.6.9 The *general solution* of a differential equation is a function from which every particular solution may be obtained by an appropriate choice of values for arbitrary constants. ⋄

It should be apparent that it will typically be very difficult to know whether or not a given solution is a general solution, even if it contains many arbitrary constants. In certain cases where we can make some definite theoretical statements regarding the uniqueness of solutions, it may be possible to assert that a given solution is a general solution.

Example 1.6.10 At this point we have not developed to prove it, but the function $x(t) = c_1 \sin t + c_2 \cos t$ happens to be the general solution to

$$\ddot{x} + x = 0.$$

Every particular solution may be obtained by the appropriate choice for c_1 and c_2. ■

Example 1.6.11 The function $x(t) = c_1 \sin t + c_2 \cos t + t \sin t$ is the general solution to

$$\ddot{x} + x = 2\cos t.$$

■

For most engineering problems it is natural to expect that if a differential equation describes some dynamical process, then there will be only one solution, *i.e.,* a unique solution. We will defer the issue of existence and uniqueness of solutions to later. For present purposes we will address the question of exactly how the arbitrary constants are determined. The data that is used to determine the arbitrary constants in a general solution to determine a specific particular solution are called either the *initial conditions* or *boundary conditions.*

Example 1.6.12 Consider the differential equation and solution from Example 1.6.11 again, and assume that we desire the solution that satisfies

$$\begin{aligned} x(0) &= 0 \\ \dot{x}(0) &= 1. \end{aligned}$$

Substituting this data into the general solution gives the two equations

$$\begin{aligned} x(0) &= c_1 \sin 0 + c_2 \cos 0 + 0 \sin 0 \\ &= c_2. \end{aligned}$$

Since it was specified that $x(0) = 0$,then $c_2 = 0$. Similarly,

$$\begin{aligned} \dot{x}(0) &= c_1 \cos 0 - c_2 \sin 0 + \sin 0 + 0 \cos 0 \\ &= c_1. \end{aligned}$$

Since $\dot{x}(0) = 1$, then $c_1 = 1$. Hence, the function

$$x(t) = \sin t + t \sin t$$

is the particular solution that satisfies, in addition to the differential equation, the two additional criteria that $x(0) = 0$ and $\dot{x}(0) = 1$. ■

Reviewing the examples presented so far regarding general solutions, it appears that the general solution to an nth order differential equation will have n arbitrary constants. This is generally, although not always the case. For all the general solutions we will consider in this book is, in fact, the case.

The term *homogeneous solution* means the general solution to an ordinary, homogeneous differential equation. If the equation is inhomogeneous, the homogeneous solution is the general solution obtained by setting the inhomogeneous term to zero. If the equation is inhomogeneous, then a subscript h will be used to designate the fact that the solution is the homogeneous solution as opposed to the particular, general or explicit solution. For the types of differential equations will consider in this book where we deal with homogeneous solutions, an nth order differential equation will have n different[3] homogeneous solutions. An alternative common name for homogeneous solutions is *complementary functions.*

Example 1.6.13 The function $x_h(t) = c_1 \sin t + c_2 \cos t$ is the homogeneous solution to all of the following differential equations:

$$\begin{aligned} \ddot{x} + x &= \sin t \\ \ddot{x} + x &= \cos t \\ \ddot{x} + x &= t \\ \ddot{x} + x &= e^t \\ \ddot{x} + x &= \frac{\sin t + 35 \cos t}{e^t + 6} \\ \ddot{x} + x &= 5. \end{aligned}$$

[3] What exactly constitutes “different” solutions is a bit subtle. In fact, we will require *linearly independent* solutions, which is detailed subsequently in Section 3.2.2.

As will be seen in Chapter 3, the homogeneous solution is usually added to a particular solution to determine a general solution to an ordinary, inhomogeneous, linear equation.

1.7 Existence and Uniqueness of Solutions

Given a differential equation, the issue of whether or not it actually has a solution, and if it does, whether or not that solution is unique, is clearly of great importance.

1.8 Stability

1.9 A Few Fundamental Principles from Science

Differential equations arise in engineering because the fundamental laws governing many physical processes are known relationships between various quantities and their derivatives. Hence, the fundamental law is known, and often quite simple such as Newton's second law, $F = ma$; however, the ultimate consequences of this law may be quite complicated. This section reviews a few fundamental laws of science, some of which are the foundation which gives rise to differential equations that have engineering importance.

1.9.1 Units

In order for numeric descriptions of quantities to be meaningful, a system of units must be employed. As is conventional this book will use the following as the base units for the seven base quantities, and all other units will be derived from these. These base units are

1. the *meter,* m, which is a unit for the base quantity of *length*;

2. the *second,* s, which is a unit for the base quantity of *time*;

3. the *kilogram,* kg, which is a unit for the base quantity of *mass*;

4. the *ampere,* A, which is a unit for the base quantity of *electric current*;

5. the *kelvin,* K, which is a unit for the base quantity of *thermodynamic temperature*;

6. the *mole,* mol, which is a unit for the base quantity of the *amount of substance*; and,

7. the *candela,* cd, which is a unit for the base quantity of the *luminous intensity* of light.

Derived quantity	Name	Symbol	Base Units	Other Units
area	square meter		m^2	
volume	cubic meter		m^3	
plane angle	radian	rad	$\frac{\mathrm{m}}{\mathrm{m}}$	
speed, velocity	meters per second		$\frac{\mathrm{m}}{\mathrm{s}}$	
angular velocity	radians per second		$\frac{1}{\mathrm{s}}$	$\frac{\mathrm{rad}}{\mathrm{s}}$
acceleration	meters per second squared		$\frac{\mathrm{m}}{\mathrm{s}^2}$	
mass density	kilograms per cubic meter		$\frac{\mathrm{kg}}{\mathrm{m}^3}$	
frequency	Hertz	Hz	$\frac{1}{\mathrm{s}}$	
force	Newton	N	$\frac{\mathrm{kg}\cdot\mathrm{m}}{\mathrm{s}^2}$	
moment	Newton meter	$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2}$	$\mathrm{N}\cdot\mathrm{m}$	
energy, work	Joule	J	$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2}$	$\mathrm{N}\cdot\mathrm{m}$
power	Watt	W	$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^3}$	$\frac{\mathrm{J}}{\mathrm{s}}$
electric charge	Coulomb	C	$\mathrm{A}\cdot\mathrm{s}$	
electric potential	Volt	V	$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^3\cdot\mathrm{A}}$	$\frac{\mathrm{W}}{\mathrm{A}}$
electric capacitance	Farad	F	$\frac{\mathrm{A}^2\cdot\mathrm{s}^4}{\mathrm{kg}\cdot\mathrm{m}^2}$	$\frac{\mathrm{C}}{\mathrm{V}}$
electric resistance	Ohm	Ω	$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^3\cdot\mathrm{A}^2}$	$\frac{\mathrm{V}}{\mathrm{A}}$
electric inductance	Henry	H	$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2\cdot\mathrm{A}^2}$	
heat capacity	Joules per Kelvin		$\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2\cdot\mathrm{K}}$	$\frac{\mathrm{J}}{\mathrm{K}}$
thermal conductivity	Watt per meter Kelvin		$\frac{\mathrm{kg}\cdot\mathrm{m}}{\mathrm{s}^3\cdot\mathrm{K}}$	$\frac{\mathrm{W}}{\mathrm{m}\cdot\mathrm{K}}$

Table 1.1. Some derived units based upon the seven base units in the SI system adapted from [20].

magnitude	name	symbol	magnitude	name	symbol
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a
10^{2}	hecto	h	10^{-21}	zepto	z
10^{1}	deka	da	10^{-24}	yocto	y

Table 1.2. The standard prefixes corresponding to different orders of magnitude [20].

See [20] for further information. Units derived from these base units that are used in this book are presented in Table 1.1. For completeness, the usual prefixes used for different orders of magnitude are presented in Table 1.2.

The calculus operations of differentiation and integration change the units of a function in an intuitive manner. By the definition of the derivative,

$$\frac{df}{dt}(t) = \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t}$$

the units of a derivative of a function will be the units of that function divided by the units of the independent variable with respect to which it is being differentiated.

Example 1.9.1 If $x(t)$ has units of meters, then $\dot{x}(t)$ will have units of meters divided by seconds. ■

Example 1.9.2 If $u(x,t)$ has units of Kelvin, then $\underline{\frac{\partial^2}{\partial x^2 u}(x,t)}$ $\underline{\frac{\partial^2}{\partial x^2}u(x,t)}$ will have units of Kelvin divided by meters squared. ■

Conversely, the units of the integral of a function will be the units of that function multiplied by the units of the independent variable with respect to which it is being integrated.

Example 1.9.3 If $f(t)$ has units of meters and t has units of seconds, then $\int f(t)dt$ will have units of meters times seconds. ■

1.9.2 Mechanical Systems

In this section we will consider some basic ways to determine the equations of motion for mechanical systems. This text is not intended to be a mechanics book; however, it is important to consider the manner in which differential equations arise. Keep in mind that the point of this section is only the means to determine the right equations. The rest of the book is about how to solve them.

This section is intended as a summary of basic results from dynamics. A complete exposition requires a much more comprehensive treatment. An interested reader is referred to, for example, [15] for an introductory treatment, [10] for an intermediate treatment or to [8, 2, 4, 1] for a more advanced treatment.

In *The Principia*, [17, 18], Isaac Newton states the following three laws of motion.

Law 1.9.4 Every body preserves in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon.[4]

[4] As originally published, it states "Lex I: Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus a viribus impressis cogitur statum illum mutare."

The modern expression of this law is *conservation of momentum.*

Law 1.9.5 The alteration of motion is ever proportional to the motive forces impressed; and is made in the direction of the right line in which that force is impressed.[5]

This gives rise to the familiar “force equals mass times acceleration” rule.

Law 1.9.6 To every action there is always opposed an equal reaction: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.[6]

In other words, forces occur in equal and opposite pairs. If you push on a body, the force you exert is exactly the same as the force that the body exerts on you. This law plays a critical role in the development of rigid body mechanics.

Application of Newton’s Laws to translational motion of a particle

Newton’s first law speaks of a “body.” We need to be a bit more precise and distinguish between two types of bodies. In particular, we will consider particles and rigid bodies. A *particle* is a object that generally has a finite mass, but has no appreciable physical extent compared to its range of motion. In such a case it is valid to assume that the mass is concentrated at a point. A *rigid body* is a collection of particles where the distance between any two particles remains fixed. In this text, unless otherwise indicated, all vectors describing physical systems will be with respect to an *inertial coordinate system,* which is a coordinate system that is not rotating and has an origin that is not accelerating.[7]

First, we will define some fundamental quantities for particles, collections of particles and rigid bodies, and then express various forms of Newton’s laws for each.

Definition 1.9.7 The *linear momentum,* $\mathbf{p}$, of a particle of mass m with velocity $\mathbf{v}$ measured relative to an inertial coordinate system is given by

$$\mathbf{p} = m\mathbf{v}. \tag{1.10}$$

◇

[5] “Lex II: Mutationem motus proportionalem esse vi motrici impressae, et fieri secundum lineam rectam qua vis illa imprimitur.”

[6] “Lex III: Actioni contrariam semper et qualem esse reactionem: sive corporum duorum actiones in se mutuo semper esse quales et in partes contrarias dirigi.”

[7] Exactly how to determine whether or not a coordinate system is inertial is not an easy thing. Generally, however, on the earth if it is not accelerating with respect to the surface of the earth, then it is approximately inertial unless the acceleration is extremely small or the extent of motion is large. Sometimes an inertial frame is defined to be one in which Newton’s laws hold; however, this is not of much use if our purpose is to apply Newton’s laws! Appealing to Einstein’s general theory of relativity gives a complete answer, but is beyond the scope of this text.

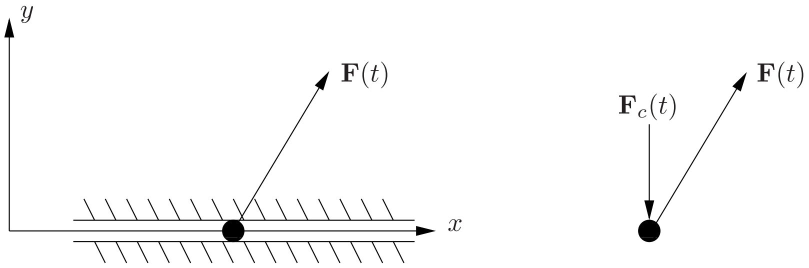


Figure 1.2. System for Example 1.9.8.

Considering a particle of mass m, where $\mathbf{x}$ denotes its position, Newton's second law states that if $\mathbf{F}$ represents the vector sum of the forces acting on the particle,

$$\frac{d\,(m\mathbf{v})}{dt}(t) = \mathbf{F}, \tag{1.11}$$

and in the case where m is constant,

$$m\frac{d^2\mathbf{x}}{dt^2}(t) = \mathbf{F}. \tag{1.12}$$

In the case where there are no forces, the first law follows. and is given by

$$\frac{d\,(\mathbf{p})}{dt}(t) = \mathbf{0} \qquad \Longrightarrow \qquad \mathbf{p}(t) = \text{const.}$$

Equations 1.11 and 1.12 are the primary equations used to determine equations describing the translational motion of a particle. To use it for a problem where the motion of the particle is constrained, the constraint forces must either be determined or be orthogonal to the directions of motion under consideration.

Most of the use of Newton's laws in this book will be concerned with the special case of *rectilinear motion,* which is motion along a straight line. This case is nice because the equations of motion will reduce to a scalar differential equation and the application of Newton's law will simply be to write $F = ma$ in the relevant direction. The student is cautioned to be cognizant of how restrictive this case actually is and to exercise care in applying Newton's law in the appropriate form, in particular Equation 1.11, when the motion is not necessarily rectilinear.

Example 1.9.8 Consider a particle of mass m constrained to move along the x-axis and subjected to an applied force $\mathbf{F}(t)$ as is illustrated in Figure 1.2. The force $\mathbf{F}(t)$ may have both a magnitude and orientation that changes with time. Assume that the constraint is frictionless.

Let

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix},$$

denote the position of the particle and

$$\mathbf{F}(t) = \left[\begin{array}{c} F_x(t) \\ F_y(t) \end{array} \right],$$

denote the two components of the force. A *free body diagram*[8] of the particle is illustrated on the right of Figure 1.2. There are two forces acting on the particle; the applied force $\mathbf{F}(t)$ and some unknown constraint force, $\mathbf{F}_c(t)$. Since the constraint is frictionless, $\mathbf{F}_c(t)$ must be purely in the y-direction with no component in the x-direction, so we may write

$$\mathbf{F}_c(t) = \left[\begin{array}{c} 0 \\ F_c(t) \end{array} \right].$$

For this system, Equation 1.12 is of the form

$$m\frac{d^2\mathbf{x}}{dt^2} = \mathbf{F}(t) + \mathbf{F}_c(t)$$

Writing the vectors in terms of their components gives

$$m \left[\begin{array}{c} \frac{d^2x}{dt^2} \\ \frac{d^2y}{dt^2} \end{array} \right] = \left[\begin{array}{c} F_x(t) \\ F_y(t) \end{array} \right] + \left[\begin{array}{c} 0 \\ F_c(t) \end{array} \right] = \left[\begin{array}{c} F_x(t) \\ F_y(t) + F_c(t) \end{array} \right]$$

which is equivalent to the two scalar equations

$$\begin{aligned} m\ddot{x} &= F_x(t) && (1.13) \\ m\ddot{y} &= F_y(t) + F_c(t). && (1.14) \end{aligned}$$

Since the motion is constrained to be only in the x-direction, Equation 1.14 reduces to

$$F_y(t) + F_c(t) = 0.$$

Observe that if the point of interest in the problem were only the motion in the x-direction, we could have easily determined Equation 1.13 by only considering the forces in the x-direction. In such a case there is no need to even determine Equation 1.14. This nice form of the equations occurred not only because the direction of motion and constraint force were orthogonal, but because they were in *constant* directions, which is a consequence of the motion being rectilinear. ■

The following example illustrates that things become more complicated when the motion is not rectilinear.

Example 1.9.9 Consider a particle constrained to move along a curve described by $y = f(x)$ as illustrated in Figure 1.3 subjected to a known external force, $\mathbf{F}(t)$. This may be thought of as a bead moving along a frictionless wire with the prescribed shape.

[8] A free body diagram is an illustration of the particle isolated from the environment wherein all the forces acting on the body are illustrated.

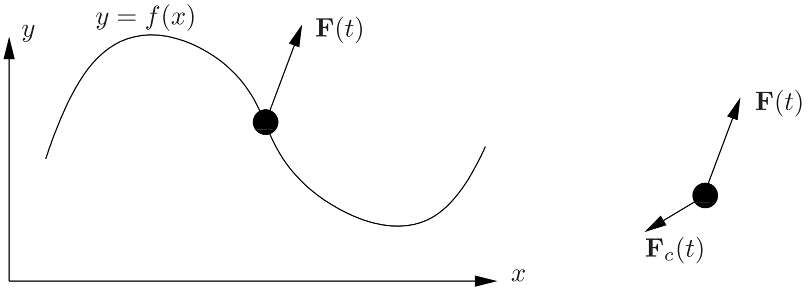


Figure 1.3. System for Example 1.9.9.

The particle *must* obey Newton's second law; hence,

$$m\frac{d^2\mathbf{x}}{dt^2} = \mathbf{F}(t) + \mathbf{F}_c(t), \tag{1.15}$$

where $\mathbf{F}_c(t)$ is the constraint force between the bead and wire.

This seems like two equations for the two components of $\mathbf{x}(t)$. However, those two components are not actually the unknowns. Since the particle must stay on the wire, if we know $x(t)$, then we know $y(t)$ since $y = f(x)$. So, only *one* of the two components is really not known. The other unknown is ~~actually~~ the magnitude of $\mathbf{F}_c(t)$. Since the wire is frictionless, $\mathbf{F}_c(t)$ must be orthogonal to the wire at the location of the bead. So we know its direction at any location $\mathbf{x}(t)$, but not its magnitude. Since the slope of $y = f(x)$ is given by the derivative at x, then the vector

$$\mathbf{t}(x) = \begin{bmatrix} 1 \\ \left.\frac{df}{dx}\right|_x \end{bmatrix}$$

will be in the direction tangent to the curve. Computing the normal vector, $n(x)$ such that $n(x) \cdot t(x) = 0$, we have that the normal vector at the point x is

$$\mathbf{n}(x) = \begin{bmatrix} -\left.\frac{df}{dx}\right|_x \\ 1 \end{bmatrix}.$$

Hence the constraint force is in the direction of this normal vector, but not necessarily of the same magnitude. It is a vector that has the same direction as the normal, but not the same magnitude *i.e.,*

$$\mathbf{F}_c(t) = F_c \begin{bmatrix} -\left.\frac{df}{dx}\right|_{x(t)} \\ 1 \end{bmatrix} \tag{1.16}$$

If we write

$$\mathbf{x}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

since $y(t) = f(x(t))$, then the chain rule for differentiation gives

$$\dot{y}(t) = \left.\frac{df}{dx}\right|_{x(t)} \dot{x}(t)$$

and then the chain rule and product rule for differentiation gives

$$\ddot{y}(t) = \left.\frac{d^2 f}{dx^2}\right|_{x(t)} \dot{x}^2(t) + \left.\frac{df}{dx}\right|_{x(t)} \ddot{x}(t). \tag{1.17}$$

Now we can substitute Equation 1.17 into the second component of Equation 1.15 and substitute Equation 1.16 into Equation 1.15, solve one of them for $F_c(t)$ and substitute into the other to result in a single differential equation in dependent variable x and independent variable t. Substituting Equation 1.16 into Equation 1.15 and writing it in components gives

$$\begin{aligned} m\ddot{x}(t) &= F_x(t) - F_c(t) \left.\frac{df}{dx}\right|_{x(t)} \qquad (1.18)\\ m\ddot{y}(t) &= F_y(t) + F_c(t). \end{aligned}$$

Hence

$$F_c(t) = m\ddot{y}(t) - F_y(t)$$

and substituting from Equation 1.17 gives

$$F_c(t) = m\left(\left.\frac{d^2 f}{dx^2}\right|_{x(t)} \dot{x}^2(t) + \left.\frac{df}{dx}\right|_{x(t)} \ddot{x}(t) \right) - F_y(t).$$

Finally, substituting for $F_c(t)$ in Equation 1.18 gives

$$m\ddot{x}(t) = F_x(t) - \left(m\left(\left.\frac{d^2 f}{dx^2}\right|_{x(t)} \dot{x}^2(t) + \left.\frac{df}{dx}\right|_{x(t)} \ddot{x}(t) \right) - F_y(t) \right) \left.\frac{df}{dx}\right|_{x(t)}. \tag{1.19}$$

■

The point of Example 1.9.9 was to demonstrate that the case of constrained non-rectilinear motion is rather involved. There are other ways to approach the problem (see Example 16.0.2), which in some cases may be more efficient, but there is basically no way to just to be able to write the equations as *one* equation that is of the form $m\ddot{x} = F$, where F is simply the applied forces projected onto the direction of motion. This is because, in general, the motion of the particle, the applied force and the constraint force have components in both coordinate directions. Of course, one may attempt to realign the coordinate axes with the curve $y = f(x)$, but how to do this depends on the location of the particle, and if the particle is moving, the realigned coordinates will have to move with it, which, except in the simplest case of pure translation, will make the realigned coordinates non-inertial. Such an approach is actually probably the most intuitive, but there is the complication of properly taking care of the fact that the coordinate system is non-inertial. An interested reader is referred to [10] for a comprehensive treatment of this issue.

Fortunately, for most problems in this book rectilinear motion is what is considered and hence, it is usually relatively easy to write $F = ma$ in the correct direction.

Application of Newton's laws to rotational motion of a particle

We now consider a formulation that is amenable to rotational motion, by basically taking the cross product of a vector from some point with each side of Equations 1.10, 1.11 and 1.12.

Definition 1.9.10 The *angular momentum*, $\mathbf{h}_O(t)$, of a particle of mass m with velocity $\mathbf{v}(t)$ about a point O

$$\begin{aligned} \mathbf{h}_O(t) &= \mathbf{r}(t) \times \mathbf{p}(t) \\ &= m\left(\mathbf{r}(t) \times \mathbf{v}(t)\right) \qquad (1.20) \end{aligned}$$

where $\mathbf{r}$ is measured from O to the position of the particle and × is the normal cross product in $\mathbb{R}^3$ and the second equation holds if the mass of the particle is constant. ⋄

Computing the derivative of angular momentum with respect to time

$$\begin{aligned} \frac{d\mathbf{h}_O}{dt}(t) &= \frac{d\left(\mathbf{r}(t) \times \mathbf{p}(t)\right)}{dt} \\ &= \frac{d\mathbf{r}}{dt}(t) \times \mathbf{p}(t) + \mathbf{r}(t) \times \frac{d\mathbf{p}}{dt}(t) \\ &= \mathbf{v}(t) \times \mathbf{p}(t) + \mathbf{r}(t) \times \frac{d\mathbf{p}}{dt}(t) \\ &= \mathbf{v}(t) \times \left(m\mathbf{v}(t)\right) + \mathbf{r}(t) \times \frac{d\mathbf{p}}{dt}(t) \\ &= \mathbf{r}(t) \times \frac{d\mathbf{p}}{dt}(t), \end{aligned}$$

and by Equation 1.11, we have

$$\frac{d\mathbf{h}_O}{dt}(t) = \mathbf{r}(t) \times \mathbf{F}(t). \qquad (1.21)$$

Equation 1.21 is the usual, "the rate of change of angular momentum about a point is equal to the sum of the moments about that point." If all the forces acting on the particle are parallel to $\mathbf{r}$, then angular momentum about the point O is conserved.

Example 1.9.11 Consider a particle with mass m constrained to move along a frictionless circular hoop with radius r, as illustrated in Figure 1.4.

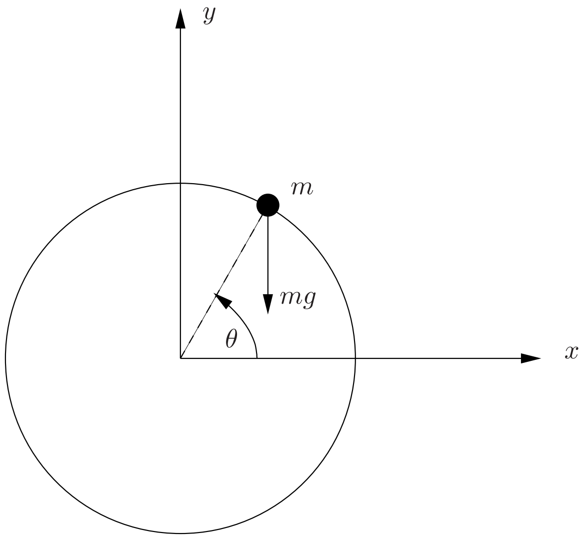


Figure 1.4. Hoop for Example 1.9.11.

Determine a differential equation that describes the motion of the particle using θ as the dependent variable and t as the independent variable.

Since the particle is constrained to move along the hoop the magnitude of the velocity will be

$$\|\mathbf{v}\| = r\dot{\theta}.$$

The angular momentum about the center of the hoop is

$$\mathbf{h} = \mathbf{r} \times \mathbf{v}.$$

We *could* determine the components of $\mathbf{v}$ and $\mathbf{r}$ as a function of θ.[9] However, it is easier to observe that because of the geometry of the hoop, $\mathbf{v}$ will always be orthogonal to $\mathbf{r}$ and $\mathbf{h}$ about the center of the hoop will always be orthogonal to the plane of the hoop. Hence, we can let h be the scalar that represents the magnitude of the angular momentum of the particle along

[9] Heck, let's work it out anyway. If we take the x-axis to the right, the y-axis up and the z-axis out of the page in Figure 1.4, then

$$\mathbf{v} = \begin{bmatrix} -r\dot{\theta}\sin\theta \\ r\dot{\theta}\cos\theta \\ 0 \end{bmatrix}$$

and

$$\mathbf{r} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \\ 0 \end{bmatrix}.$$

the axis orthogonal to and out of the plane of the hoop and write

$$h = r^2\dot{\theta}.$$

To determine the motion of the particle, we will use Equation 1.21, so we need to compute $\mathbf{r} \times \mathbf{F}$. This cross product will also be along the axis orthogonal to the plane of the hoop, so we may write

$$\|\mathbf{r} \times \mathbf{F}\| = rmg\sin\theta.$$

Finally, since r is constant,

$$\frac{dh}{dt}(t) = r^2\ddot{\theta}(t)$$

and by Equation 1.21

$$r^2\ddot{\theta} = rmg\sin\theta$$

or

$$\ddot{\theta} = \frac{mg}{r}\sin\theta,$$

which is a second order, nonlinear, ordinary differential equation. ■

Application of Newton's Laws to a System of Particles

In order to extend Newton's laws to rigid bodies, we must consider the application of them to systems of particles, which is simply a collection of particles. In the next section will consider the special case where the system of particles makes a rigid body which requires the additional constrain that the distance between any two particles remains fixed. Also, generally, there will be an infinite number of particles in a typical rigid body.

Application of Newton's Laws to a Rigid Body

Now, for a rigid body is simply a system of particles where the particles are constrained by internal forces to remain a fixed distance from each other. Fortunately, a non-obvious consequence of Newton's third law is that we may express the equation of motion for the translational motion of the rigid body in a

Hence

$$\begin{aligned}\mathbf{r} \times \mathbf{v} &= \begin{bmatrix} r\cos\theta \\ r\sin\theta \\ 0 \end{bmatrix} \times \begin{bmatrix} -r\dot{\theta}\sin\theta \\ r\dot{\theta}\cos\theta \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \\ r^2\dot{\theta}\cos^2\theta + r^2\dot{\theta}\sin^2\theta \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \\ r^2\dot{\theta}. \end{bmatrix}\end{aligned}$$

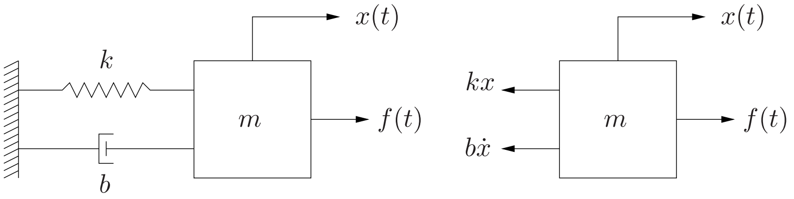


Figure 1.5. Mechanical system with a mass, spring and damper and its free body diagram for Example 1.9.12.

simple form with respect to its center of mass. In particular, if a rigid body is subjected to external forces with sum $\mathbf{F}$, then

$$m\frac{d^2\mathbf{x}_{com}}{dt^2} = \mathbf{F}. \tag{1.22}$$

where m is the total mass of the rigid body, *i.e.,*

$$m = \int dm$$

and $\mathbf{x}_{com}$ is the center of mass of the body, defined by

$$\mathbf{x}_{com} = \frac{\int \mathbf{x} dm}{m},$$

where the integrals are over the extent of the rigid body. So, we have the convenient result that we may simply think of $\mathbf{F} = m\mathbf{a}$ as correct for a rigid body as long as $\mathbf{a}$ represents the acceleration of the center of mass of the body and $\mathbf{F}$ represents the sum of the external forces only.

Example 1.9.12 Determine the equation of motion for the mass-spring-damper system illustrated in Figure 1.5. Assume that $x = 0$ when the spring is unstretched.

A free body diagram of the mass is illustrated on the right in Figure 1.5. So, since the acceleration of the mass is equal to $\ddot{x}$, we have

$$f(t) - kx - b\dot{x} = m\ddot{x}$$

which, is usually expressed in the form

$$m\ddot{x} + b\dot{x} + kx = f(t). \qquad \blacksquare$$

Definition 1.9.13 The *scalar moment of inertia*, denoted by J, of a mass particle about a specified point is

$$J = mr^2,$$

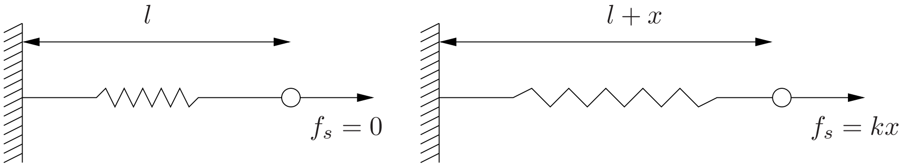


Figure 1.6. The relationship between force and extension of an ideal spring.

where m is the mass of the particle and r is the distance from the point to the particle, and for a collection of N particles,

$$J = \sum_{i=1}^{N} m_i r_i^2.$$

Extending this to a planar rigid body

$$J = \int_A r^2 dm,$$

where A is the planar area of the body. ◇

1.9.3 Mechanical Components

In this book we will be primarily concerned with interconnected rigid body systems. It is presumed that the student has at least a basic introduction to dynamics and is familiar with applying Newton's laws to point masses and rigid bodies constrained to motion in the plane. The two main components we need to properly model are linear *springs* and *viscous dampers.*

Springs

An ideal linear spring is a mechanical device which requires a force to extend it that is proportional to the amount of extension. Mathematically,

$$f_s = kx,$$

where f_s is the force required to extend the length of the spring, x is the amount by which the length of the spring has been extended and k is the *spring constant*, which is a characteristic of the spring. The force, f_s and extension x must be defined in a manner so that they have the same sign when a positive force and extension are in the same direction. Negative extension is compression, and the equation still holds. The relationship is illustrated in Figure 1.6 where the unstretched length of the spring is l.

Throughout this book we will make an important assumption regarding the reference point from which spring displacements are measured.

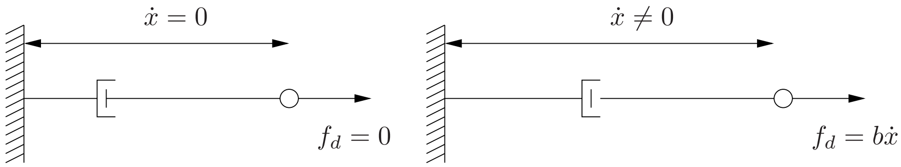


Figure 1.7. The relationship between force and the rate of extension of an ideal viscous damper.

Assumption 1.9.14 *Unless stated otherwise, any variable that represents the extension or compression of a spring is assumed to have a value of zero when the spring is at an equilibrium. If there is no gravity, then the variable will be zero when the spring is unstretched. If there is gravity acting on a mass that is supported by the spring, then the variable will be zero when the spring is stretched by an amount that results in a force equal to the weight of the mass.*

Another important assumption is that, unless otherwise specified, the mass of the spring itself may be neglected.

Viscous dampers

A viscous damper[10] is a mechanical device that requires a force to extend it that is proportional to the *rate* at which it is being extended. A common example of such a device is an automobile shock absorber. Mathematically,

$$f_d = b\dot{x},$$

where f_d is the force required to extend the damper, $\dot{x}$ is the rate at which the damper is being extended and b is the *damper constant*, which is a characteristic of the damper. The force, f_d and the rate of extension $\dot{x}$ must be defined in a manner so that they have the same sign when a positive force and rate of extension are in the same direction. Negative extension is compression, and the equation still holds. The common schematic representation of a viscous damper as well as the relationship between force and rate of displacement is illustrated in Figure 1.7. Note that for an ideal damper, the force is independent of the length of the damper. Unless otherwise specified, throughout this book we will assume that the mass of the damper itself is negligible, and hence may be omitted from any model.

Cantilever beams

Procedure to model mechanical systems

Now we will apply Newton's second law and the definition of these mechanical components to determine the equation of motion for a system.

[10] Another common term used to refer to these devices is *viscous dashpot.*

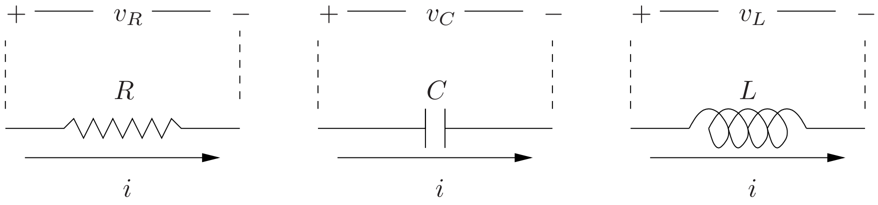


Figure 1.8. Representation of an ideal resistor, capacitor and inductor.

1.9.4 Kirchhoff's Laws

Law 1.9.15 *Kirchhoff's voltage law* states that the sum of the voltage drops around any closed loop in a circuit is zero.

This is basically conservation of energy.

Law 1.9.16 *Kirchhoff's current law* states that the sum of the currents into any point in a circuit is zero.

This is basically conservation of charge.

1.9.5 Electronic Components

There are many types of electronic components, and properly modeling some of them are necessary in this book. In particular, we will consider resistors, capacitors, inductors, voltage sources, current sources, direct current motors ("d.c. motors") and operational amplifiers ("op-amps").

Resistors

In an ideal *resistor,* the voltage drop across the resistor is proportional to the current passing through the resistor. The constant of proportionality is called the *resistance* and the equation describing this property is

$$v_R = iR, \tag{1.23}$$

where v_R is the voltage across the resistor, i is the current passing through it and R is the resistance of the resistor. The typical schematic representation of a resistor is illustrated in Figure 1.8.

Capacitors

In an ideal *capacitor,* the time rate of change of the voltage across the capacitor is proportional to the current through it. The constant of proportionality is

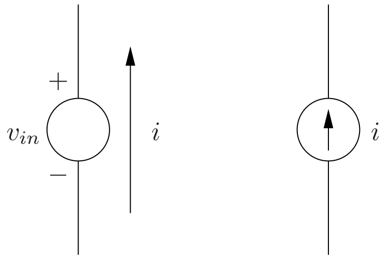


Figure 1.9. An ideal voltage source (left) and current source (right).

called the capacitance, is represented by the symbol C, and has units of Farads. The equation describing it is given by

$$i = C\frac{dv_C}{dt}.$$

This should make sense since charge will not flow through the capacitor, the effect of current flow will be the accumulation of charge on the plates of the capacitor, which results in a change in voltage across the plates. The schematic representation of a capacitor is illustrated in Figure 1.8.

Inductors

In an ideal *inductor,* the voltage drop across the inductor is proportional to the time rate of change of the current through it. The constant of proportionality is called the inductance, is represented by the symbol L and has units of Henrys. The equation governing it is

$$v_L = L\frac{di}{dt}.$$

The schematic representation of an inductor is illustrated in Figure 1.8.

Voltage source

An ideal *voltage source* supplies a specified voltage that is independent of the current that the circuit draws. A schematic illustration of an ideal voltage source is illustrated in Figure 1.9. The voltage, $v(t)$ is specified; whereas, the current through the voltage source, i is determined by the circuit to which it is attached. Of course, a real voltage source cannot maintain a specified voltage if it would require a very high current, for example in a short circuit.

Current source

An ideal *current source* supplies a specified current that is independent of the terminal voltage across the source. Its schematics representation is illustrated

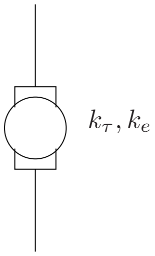


Figure 1.10. Schematic representation of a direct current motor.

in Figure 1.9.

Direct current motors

The schematic representation for a direct current motor ("d.c. motor") is illustrated in Figure 1.10. The two idealized properties of a d.c. motor we will need in this book relate the output torque of the motor to the current flowing through it and the voltage drop across the motor to the angular velocity of the shaft of the motor. Mathematically

$$\begin{aligned} \tau &= k_\tau i \\ v_m &= k_e \dot{\theta} \end{aligned}$$

where τ is the torque produced by the motor, v_m is the voltage drop across the motor, $\dot{\theta}$ is the angular velocity of the shaft of the motor and k_τ and k_e are the *torque* and *back e.m.f.* proportionality constants of the motor.

Operational amplifier

An operational amplifier ("op-amp") scales an input voltage difference by an amount called the *gain*[11] The mathematical description is

$$v_{out} = kv_{in}$$

where v_{in} is the potential difference across the two input pins and k is the *open loop gain*. An ideal op-amp has infinite input impedance, which means that no current flows across the input pins. Ideal op-amps are frequently assumed to have infinite open loop gain. In this text, we will assume the open loop gain is large, but not necessarily infinite. A schematic representation of an op-amp is illustrated in Figure 1.11.

[11] Sometimes the gain is specifically called the *open loop gain* to distinguish it from the closed loop gain. The closed loop gain of an op-amp will be discussed in Chapter 9.

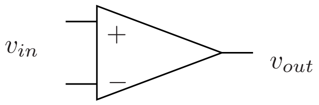


Figure 1.11. Schematic representation of an ideal operational amplifier.

1.9.6 Fourier's Law

Gives rise to the heat equation:

$$\frac{\partial^2 u}{\partial x^2} = \alpha \frac{\partial u}{\partial t}$$

1.9.7 Lagrange's Equations

This is complicated enough to warrant its own chapter.

1.10 Introduction to Numerical Methods

Because it is a sad, but true, fact that *most* differential equations cannot be solved using methods in this book (and any other book, for that matter) methods that use computers to determine *approximate* solutions are extremely important. Even if we can solve a differential equation, it may be the case that the solution is given as an implicit function, which is of somewhat limited use. This section considers *Euler's method* for solving initial value problems for ordinary differential equations, which is the most basic, and perhaps most common, method to use a computer to determine an approximate solution to a differential equation. Chapter 13 considers more advanced topics on numerical methods including more sophisticated methods for initial value problems, as well as techniques for boundary value problems and partial differential equations.

As should be clear subsequently there are two major shortcomings to resorting to numerical techniques. First, only explicit solutions may be obtained, *i.e.,* general solutions that can be used for any initial conditions cannot be determined using numerical methods (with exceptions). Therefore, if the initial conditions to a problem change, the entire method must be used again. It is not simply a matter of computing different coefficients within a solution. Secondly, the "answer" is only an approximate answer and will be in the form of tabulated data. If a more accurate solution is required, then more computer resources must be allocated to the problem and if an expression of the solution in terms of close formed functions is required, the method is not appropriate. Even with these two caveats, however, numerical methods are extremely useful and commonplace in engineering.

1.10.1 Euler's Method

Consider an ordinary, first order differential equation of the form

$$\dot{x} = f(x(t), t), \tag{1.24}$$

and assume that either we do not know how to solve it, or we are too lazy to solve it using analytical techniques. In order to derive an algorithm to determine an approximate solution, recall the definition of the derivative from calculus

$$\dot{x}(t) = \frac{dx(t)}{dt} = \lim_{\Delta t \to 0} \left(\frac{x\left(t + \Delta t\right) - x\left(t\right)}{\Delta t} \right). \tag{1.25}$$

Another way to interpret this equation is that, if the limit exists and Δt is small, then

$$\dot{x}(t) \approx \frac{x\left(t + \Delta t\right) - x\left(t\right)}{\Delta t}.$$

Keep in mind that the typical scenario is that the differential equation is known, *i.e.,* $f(x, t)$ in equation 1.24 is known. The solution, $x(t)$ is unknown. This is in contrast to the usual use of equation 1.25 where $x(t)$ is known and the derivative is unknown.

Now, assume that an initial condition is known as well, so that

$$x(t_0) = x_0 \tag{1.26}$$

has been specified. So, what is known is $f(x, t)$ in equation 1.24, the initial condition in equation 1.26 and also the definition of the derivative in equation 1.25. Now, at $t = t_0$, the approximate derivative is given by

$$\dot{x}(t_0) \approx \frac{x(t_0 + \Delta t) - x(t_0)}{\Delta t}.$$

For a specified Δt, everything in the preceding equation is known except $x(t + \Delta t)$, so it can be solved for $x(t + \Delta t)$ as

$$x(t + \Delta t) \approx \dot{x}(t_0)\Delta t + x(t_0)$$

or, from equation 1.24

$$x(t_0 + \Delta t) \approx f(x(t_0), t_0)\Delta t + x(t_0). \tag{1.27}$$

In words, if $x(t_0)$ is known and the differential equation, $\dot{x} = f(x, t)$, is known, then an approximation for $x(t + \Delta t)$ is given by equation 1.27. Also, given normal convergence properties, it will be the case that as Δt gets smaller, the approximation will be more accurate. The final piece of the puzzle is to note that once $x(t+\Delta t)$ is computed, $x(t+2\Delta t)$ can be computed from equation 1.27 by substituting the value for $x(t + \Delta t)$ for $x(t_0)$ and $t_0 + \Delta t$ for $t_0$in the right hand side of equation 1.27, *i.e.,*

$$x(t + 2\Delta t) \approx f(x(t_0 + \Delta t), t_0 + \Delta t)\Delta t + x(t_0 + \Delta t),$$

and by recursion, then

$$x\left(t + n\Delta t\right) \approx f\left(x\left(t_0 + \left(n-1\right)\Delta t\right), t_0 + \left(n-1\right)\Delta t\right)\Delta t + x\left(t_0 + \left(n-1\right)\Delta t\right). \tag{1.28}$$

Equation 1.28 is "the answer." The algorithm to implement it for a given Δt is called Euler's method, and is as follows.

1. Let $x(t_0) = x_0$;

2. let $n = 0$;

3. let $n = n + 1$;

4. let $x(t+n\Delta t) = f\left(x\left(t_0 + \left(n-1\right)\Delta t\right), t_0 + \left(n-1\right)\Delta t\right) + x\left(t_0 + \left(n-1\right)\Delta t\right)$;

5. if $n\Delta t$ is less than the time to which the approximate solution is needed, return to step 3.

At this point, things may be a bit abstract, so an example may be helpful.

Example 1.10.1 Determine an approximate numerical solution to

$$\begin{aligned} \dot{x} &= \sin 2t \\ x(0) &= 3. \end{aligned} \tag{1.29}$$

This equation is the type that will be considered in detail in section 2.3.4. Note that since the left hand side is only a function of x and the right hand side is only a function of t, both sides may be directly integrated. Since we can find the exact solution, it will be useful to compare with the approximate solution. The exact solution (which can be verified by differentiating it and substituting into the equation 1.29) is

$$x(t) = \frac{7}{2} - \frac{1}{2}\cos 2t.$$

In this example, $t_0 = 0$, $x_0 = 3$ and $f(x,t) = -3x * e^{\sin 2t}$. Picking $\Delta t = 0.5$ (a discussion on how to choose Δt appears subsequently), the first 20 steps of the algorithm are as follows. The last column is the exact solution, which is included for comparison. A plot of the approximate numerical solution and the exact solution are illustrated in Figure 1.12.

t	n	$x(t)$	$f(x(t),t)$	$x(t+\Delta t)$	$\frac{7}{2}-\frac{1}{2}\cos 2(t+\Delta t)$
0.000000	0	3.000000	0.000000	3.000000	3.229849
0.500000	1	3.000000	0.841471	3.420735	3.708073
1.000000	2	3.420735	0.909297	3.875384	3.994996
1.500000	3	3.875384	0.141120	3.945944	3.826822
2.000000	4	3.945944	−0.756802	3.567543	3.358169
2.500000	5	3.567543	−0.958924	3.088081	3.019915
3.000000	6	3.088081	−0.279415	2.948373	3.123049
3.500000	7	2.948373	0.656987	3.276866	3.572750
4.000000	8	3.276866	0.989358	3.771545	3.955565
4.500000	9	3.771545	0.412118	3.977605	3.919536
5.000000	10	3.977605	−0.544021	3.705594	3.497787
5.500000	11	3.705594	−0.999990	3.205599	3.078073
6.000000	12	3.205599	−0.536573	2.937312	3.046277
6.500000	13	2.937312	0.420167	3.147396	3.431631
7.000000	14	3.147396	0.990607	3.642699	3.879844
7.500000	15	3.642699	0.650288	3.967844	3.978830
8.000000	16	3.967844	−0.287903	3.823892	3.637582
8.500000	17	3.823892	−0.961397	3.343193	3.169842
9.000000	18	3.343193	−0.750987	2.967700	3.005648
9.500000	19	2.967700	0.149877	3.042638	3.295959

Note that the numerical solution is approximate in two ways. First, in between the times $n\Delta t$, the solution can only be interpolated. In Figure 1.12 the interpolation is linear; regardless, even if more sophisticated interpolation methods are used, the solution will only be approximately correct between the times $n\Delta t$. Second, even for the exact times $n\Delta t$, the solution still does not exactly match the exact solution. This is due to the fact that each computation for $x(t_0+n\Delta t)$ is only an approximation. Thus, except for the single point $t=t_0$, the solution at the times $t+n\Delta t$ is only approximately correct.

Decreasing the time step to $\Delta t = 0.1$ gives the result illustrated in Figure 1.13. Note that decreasing the step size by a factor of 5 greatly improves the accuracy of the approximate solution. A code listing using C is included in Appendix D.1.1. A code listing using FORTRAN is included in Appendix D.2.1. ■

1.10.2 Determining an Appropriate Step Size

A more detailed and theoretically rigorous analysis of the types of errors introduced by numerical methods will be considered in Chapter 13. At this point a heuristic approach will be used, which is simply to continue to reduce the step size by a certain factor (say by a factor of 2, or perhaps even by an order of magnitude) until the answer seems to have converged to a fixed solution. This is best illustrated by means of an example.

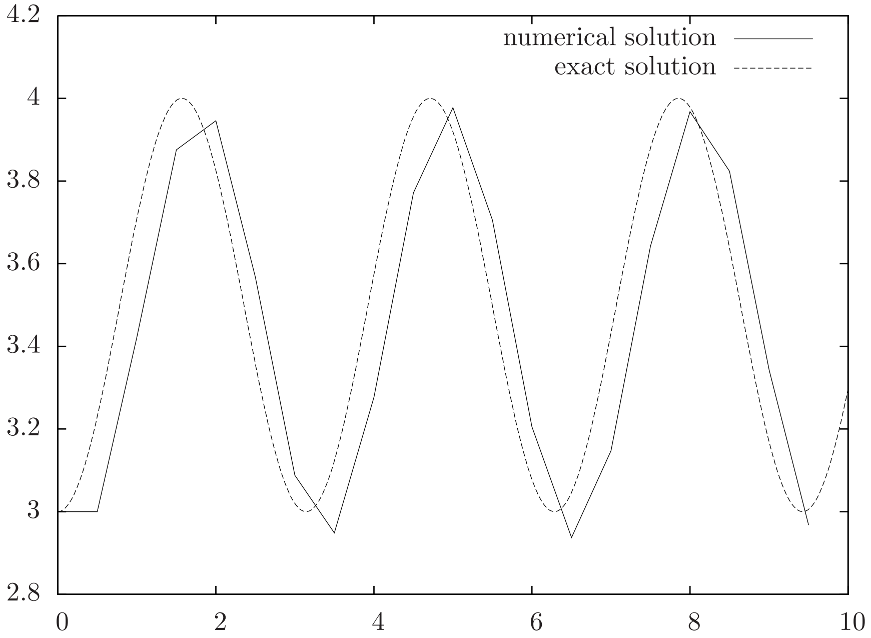


Figure 1.12. Approximate and exact solutions for example 1.10.1 with $\Delta t = 0.5$.

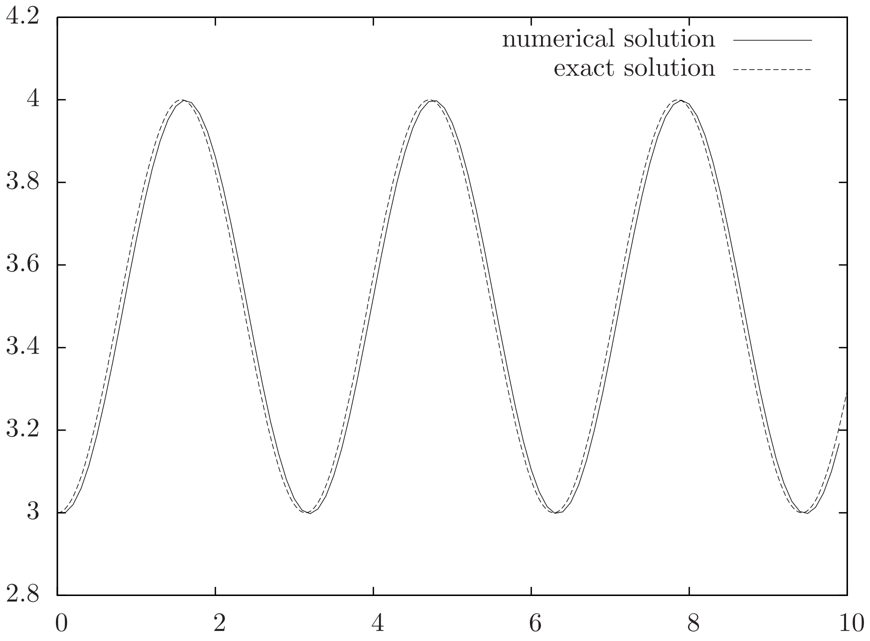


Figure 1.13. Approximate and exact solutions for example 1.10.1 with $\Delta t = 0.1$.

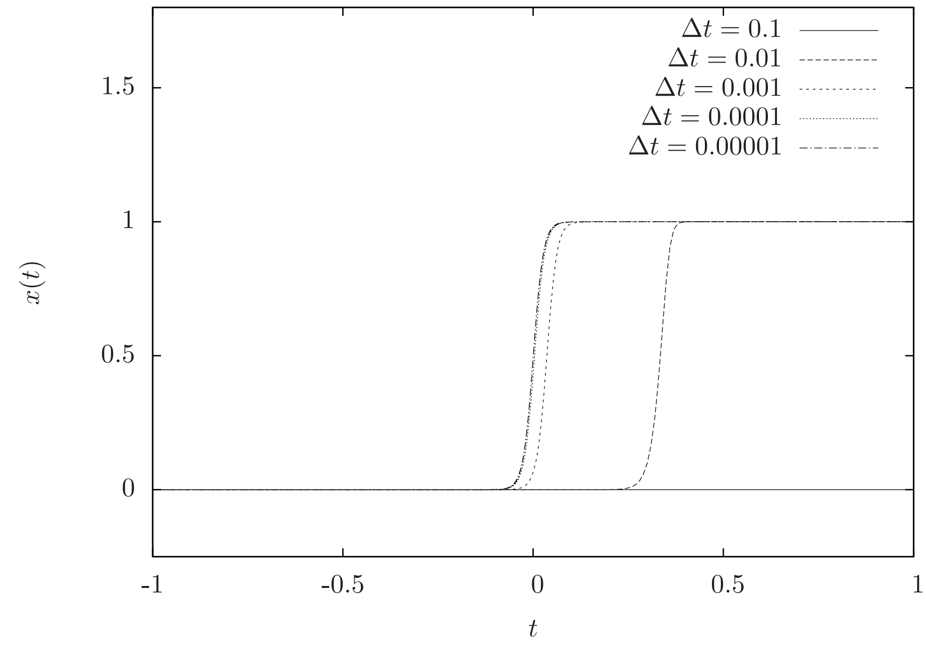


Figure 1.14. Approximate solutions for equation 1.30 using various Δt values.

Example 1.10.2 Find an approximate solution to

$$\begin{aligned} \dot{x} &= 75x(1-x) & (1.30) \\ x(-1) &= \frac{1}{1+e^{75}} & (1.31) \end{aligned}$$

using Euler's method on the time interval $-1 \leq t \leq 1$. The solution to this problem is simply implementing Euler's method using

$$\begin{aligned} t_0 &= -1 \\ x_0 &= \frac{1}{1+e^{75}} \\ f(x(t),t) &= 75x(1-x). \end{aligned}$$

Figure 1.14 illustrates the solution for a variety of values for Δt. Note that Δt must be quite small before the solution converges. A code listing using C is included in Appendix D.1.1. A code listing using FORTRAN is included in Appendix D.2.1. ■

1.10.3 Numerical Methods for Higher Order Differential Equations

The development so far is limited to ordinary, first order differential equations. This section will extend the approach to higher order ordinary differential equations by using a straight-forward reformulation of the problem to convert it into a system of first order equations. First an example will be presented which gives the main idea. It will be followed by a more general theorem which states the main result of this section.

Example 1.10.3 Find an approximate numerical solution to

$$\begin{aligned} \ddot{x} + \sin(t)\dot{x} + \cos(t)x &= e^{-5t} \qquad (1.32)\\ x(0) &= 2 \\ \dot{x}(0) &= 5. \end{aligned}$$

The main idea is the following. Consider the following change of variables

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= \dot{x}(t). \end{aligned}$$

Then the following equations are equivalent

$$\ddot{x}+\sin t\dot{x}+\cos tx = e^{-5t} \quad \Longleftrightarrow \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ e^{-5t} - \sin(t)x_2 - \cos(t)x_1 \end{bmatrix}. \qquad (1.33)$$

This is because the second line of the right hand equation is determined by simply solving equation 1.32 for $\ddot{x}$ and recognizing that $\ddot{x} = \dot{x}_2$ since $x_2 = \dot{x}$. The initial value problems are equivalent as well if

$$\begin{aligned} x_1(0) &= 2 \\ x_2(0) &= 5. \end{aligned}$$

Observe that in general terms, the right hand formulation of equation 1.33 is simply of the form

$$\begin{aligned} \dot{x}_1(t) &= f_1(x_1(t), x_2(t), t) \\ \dot{x}_2(t) &= f_1(x_2(t), x_2(t), t). \end{aligned}$$

Hence, for this case Euler's method, expressed in equation 1.28 has the simple reformulation of

$$\begin{aligned} x_1(t+n\Delta t) &\approx f_1(x_1(t+(n-1)\Delta t), x_2(t+(n-1)\Delta t), t+(n-1)\Delta t)\Delta t \\ &+ x_1(t+(n-1)\Delta t) \\ x_2(t+n\Delta t) &\approx f_2(x_1(t+(n-1)\Delta t), x_2(t+(n-1)\Delta t), t+(n-1)\Delta t)\Delta t \\ &+ x_2(t+(n-1)\Delta t), \end{aligned}$$

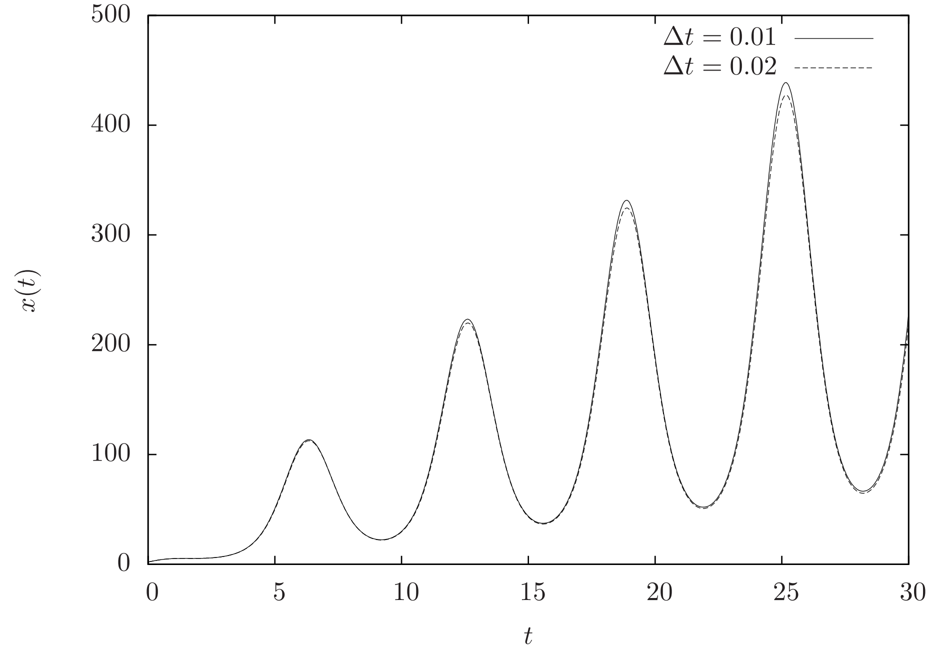


Figure 1.15. Solution for equation 1.33.

or using the particular equations of this example

$$
\begin{array}{rcl}
x_1(t+x\Delta t) & \approx & x_2(t+(n-1)\Delta t)\Delta t + x_1(t+(n-1)\Delta t) \\
x_2(t+x\Delta t) & \approx & \Big[e^{-5(t+(n-1)\Delta t)} - \sin\left(t+(n-1)\,\Delta t\right) x_2(t+(n-1)\Delta t) \\
 & - & \cos\left(t+(n-1)\,\Delta t\right) x_1(t+(n-1)\Delta t)\Big]\,\Delta t \\
 & + & x_2(t+(n-1)\Delta t).
\end{array}
$$

Since this is notationally a bit cumbersome, it may be easier to refer to the example code in the appendix. A code listing using C is included in Appendix D.1.1. A code listing using FORTRAN is included in Appendix D.2.1. A plot of the solution for $\Delta t = 0.02$ and $\Delta t = 0.01$ is illustrated in Figure 1.15. ■

1.10.4 Using Matlab to Solve Differential Equations

The `ode` series[12] of functions in Matlab provide the basic functionality for solving initial value problems for ordinary differential equations. Perhaps the most

[12] The functions include, `ode43()`, `ode23()`, `ode113()`, `ode15s()`, `ode23s()`, `ode23t()`, `ode23tb()`, `ode15i()` which provide functionality using a variety of solution methods applicable to a variety of differential equations.

common of these is `ode45()`, the usage of which will be outlined here. This function used the fourth order Runge-Kutta method, the details of which are included in Chapter 13. The basic usage is

```
>> [T,Y] = ODE45(ODEFUN,TSPAN,Y0,OPTIONS)
```

where

```
T
```

is the time vector,

```
Y
```

is the solution vector (or matrix),

```
ODEFUN
```

is a function that provides the derivative information (the right hand side of the equation),

```
Y0
```

in the initial condition, and

```
OPTIONS
```

is a list of optional parameters sent to the solver. The following example illustrates its basic use.

Example 1.10.4 Use Matlab to determine an approximate numerical solution to the set of equations from example 1.10.3.

The file "secondorder.m" contains the following.

```
function xdot = secondorder(t,x)
  xdot = zeros(2,1);
  xdot(1) = x(2);
  xdot(2) = exp(-5.0*t) - sin(t)*x(2) - cos(t)*x(1);
```

and in the command window

```
>> [t,y] = ode45(@secondorder,[0 30],[2 5]);
>> plot(t,y(:,1));
>> xlabel('t');
>> ylabel('x(t)');
```

will produce a plot similar to that illustrated in Figure 1.16. ■

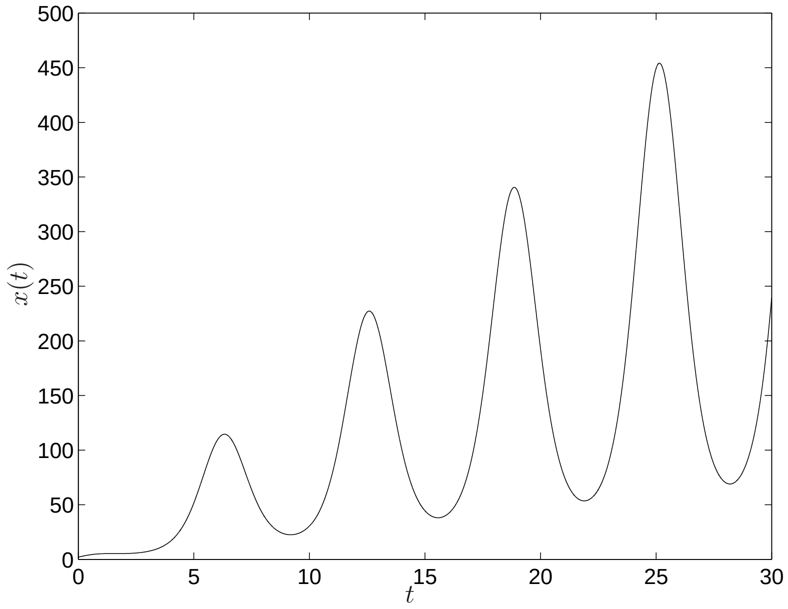


Figure 1.16. Solution output from Matlab for example 1.10.4.

1.10.5 Using Octave to Solve Differential Equations

Octave[13] is free software designed primarily for Linux and Unix operating systems, but also available on Windows and Mac OS X and has many features similar to Matlab. The main function for computing approximate solutions for ordinary differential equations in Octave is `lsode()`. This function does not use Euler integration as the default method; however, the usage is straightforward and the use of numerical computational environments is sufficiently commonplace that a description of their use should appear with the introductory material. The basic usage is

```
y = lsode("f",x0,t)
```

where

```
"f"
```

is a function (which must be defined in a file with the same name),

```
"x_0"
```

is the initial condition and

[13] For more information visit `http://www.octave.org`.

```
"t"
```

is a time vector. The following example illustrates its use.

Example 1.10.5 Use Octave to determine a solution to the set of equations from example 1.10.3.

The file “secondorder.m” contains the following.

```
function xdot = secondorder(x,t)
  xdot = zeros(2,1);
  xdot(1) = x(2);
  xdot(2) = exp(-5.0*t) - sin(t)*x(2) - cos(t)*x(1);
endfunction
```

Within the octave command line interface, the steps

```
octave:1> t = linspace(0,30,10000);
octave:2> y = lsode("secondorder",[2;5],t);
octave:3> plot(t,y(:,1),’;’)
octave:4> xlabel(’t’)
octave:5> ylabel(’x(t)’)
```

produces a solution vector called `y`. The solution is illustrated in Figure 1.17.

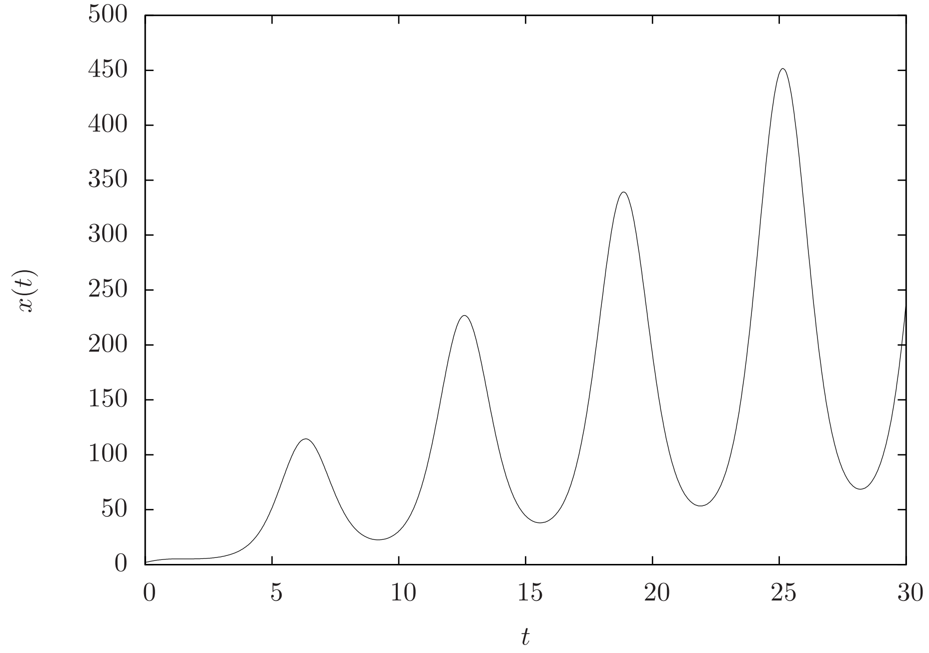


Figure 1.17. Solution output from Octave for example 1.10.5. ■

1.11 Exercises

Problem 1.1 Classify each of the following differential equations according to whether it is

- ordinary or partial; and
- linear or nonlinear.

If it is linear, indicate whether it is

- constant or variable coefficient; and,
- homogeneous or inhomogeneous.

Also determine

- its order; and,
- the dependent and independent variables.

1.

$$\begin{aligned} 5\ddot{x}+6\dot{x}+\sin(t)\,x &= \cos\left(t^2\right) \\ x(0) &= 1 \\ \dot{x}(0) &= \pi \end{aligned}$$

2.

$$\begin{aligned} \cos(t)\,\dot{x}+e^t x &= x^2 \\ x(\xi) &= e \end{aligned}$$

3.

$$\begin{aligned} \cos(t)\,\dot{x}+e^t x &= x \\ x(\xi) &= e \end{aligned}$$

4.

$$\begin{aligned} \cos(t)\,\dot{x}+e^t x &= 2 \\ x(\xi) &= e \end{aligned}$$

5.

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

6.

$$\begin{aligned} 2\ddot{x} + 19\dot{x} + 24x &= 0 \\ x(0) &= 1 \\ \dot{x}(0) = 0 \end{aligned}$$

7.

$$\begin{aligned} 2\frac{\partial^2 \zeta}{\partial \gamma^2} + 19\frac{\partial \zeta}{\partial \alpha} + 24\zeta &= \gamma^2 + \alpha^2 \\ \zeta(0) &= 1 \\ \dot{\zeta}(0) &= 0 \end{aligned}$$

8.

$$\begin{aligned} 6\ddot{x} + 23\dot{x} + t^3 x^2 &= 0 \\ x(0) &= 1 \\ \dot{x}(0) &= 0 \end{aligned}$$

9.

$$\begin{aligned} 6\ddot{x} + 23\dot{x} + x^3 &= \sin\left(t^2\right) \\ x(0) &= 1 \\ \dot{x}(0) &= 0 \end{aligned}$$

10.

$$\begin{aligned} 2\frac{d^2 \xi}{d\eta^2} + 19\frac{d\xi}{d\eta} + 25\xi &= 0 \\ \xi(0) &= 1 \\ \dot{\xi}(0) &= 0 \end{aligned}$$

11.

$$\begin{aligned} \pi\ddot{x} + e\dot{x} + x &= \sin(t) \\ x(0) &= 1 \\ \dot{x}(0) &= 0 \end{aligned}$$

12.

$$\begin{aligned} 2\frac{d^2 \zeta}{d\gamma^2} + 19\frac{d\zeta}{d\gamma} + \gamma 24\zeta &= 0 \\ \zeta(0) &= 1 \\ \dot{\zeta}(0) &= 0 \end{aligned}$$

13.

$$\begin{aligned} 2\frac{d^2\zeta}{d\gamma^2} + 19\frac{d\zeta}{d\gamma} + 24\zeta &= \gamma \\ \zeta(0) &= 1 \\ \dot{\zeta}(0) &= 0 \end{aligned}$$

Problem 1.2 Write a computer program to determine an approximation numerical solution to

$$\begin{aligned} \dot{x} + x &= e^{3t} \\ x(0) &= 1 \end{aligned}$$

using Euler's method. Determine an appropriate step size by decreasing the step size until the solution seems to converge. Compare your answer with a solution determined using Matlab. Submit your computer code as well as a plot of the approximate solution.

Problem 1.3 Write a computer program to determine an approximate numerical solution to

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

using Euler's method. Be sure to continue to decrease the step size until the solution seems to converge. Compare your answer with a solution determined using Matlab or Octave. Submit your computer code as well as a plot of the approximate solution.

Problem 1.4 Write a computer program to determine an approximate numerical solution to

$$\begin{aligned} \ddot{x} + t\dot{x} + 2x &= 0 \\ x(0) &= 3 \\ \dot{x}(0) &= -2 \end{aligned}$$

using Euler's method. Be sure to continue to decrease the step size until the solution seems to converge. Compare your answer with a solution determined using Matlab or Octave. Submit your computer code as well as a plot of the approximate solution.

Chapter 2

Ordinary First Order Equations

2.1 Introduction

First order ordinary differential equations are nice because they have some rather special properties. All linear first order ordinary differential equations can easily be solved. This is in contrast to higher order ordinary differential equations that become much more complicated when, for example, they contain variable coefficients. Furthermore, even some methods exist to solve nonlinear first order ordinary differential equations. Such methods do not exist, in general, for higher order equations.

This chapter considers methods to solve first order ordinary differential equations of the form

$$\dot{x} = f\left(x(t), t\right). \tag{2.1}$$

If the differential equation that must be solved is not of the form of Equation 2.1 care must be taken that solutions are neither gained nor lost.

Example 2.1.1 Any function $x(t)$ that satisfies

$$5\dot{x}(t) - \cos t = 0$$

also satisfies

$$\dot{x} = \frac{1}{5}\cos t.$$

■

Example 2.1.2 Clearly not all functions $x(t)$ that satisfy

$$\dot{x}^2(t) = \cos^2(t)$$

also satisfy

$$\dot{x}(t) = \cos t.$$

Both $x(t) = \sin(t)$ and $x(t) = -\sin(t)$ satisfy the former, but only $x(t) = -\sin(t)$ satisfies the latter. ■

Most of the examples and exercises in this chapter start with an equation of the form of Equation 2.1. The reader is cautioned to exercise care if a lot of manipulation is required to convert the equation to be solved into the form of Equation 2.1, particularly when making use of inverse functions, that the equation that is solved actually is equivalent to the problem at hand.

2.2 Motivational Examples

The first example of a first order differential equation comes from heat transfer.

Example 2.2.1 Consider the problem of determining the temperature of an object paced in an oven (or conversely, a refrigerator). If the inside of the oven is at temperature T_a, and is constant, and the initial temperature of the body is $T(0)$, we want to determine $T(t)$.

While a complete exposition of heat transfer requires an entire course and hence is obviously outside the scope of this book, a couple relevant concepts can be considered here. First, temperature can be considered as a measure of the amount of thermal energy which a body contains. Second, heat transfer, then, is a measure of how much energy is transferred between systems in a given amount of time. Let q denote the rate of heat transfer. The units for q will be energy per unit time, or $\frac{J}{s}$ or watts W.

Considering an energy balance on the body, we have that the rate of change of the internal energy of the body must be equal to the rate of energy transfer into (or out of) the body from the surrounding air. A basic result from heat transfer is that the heat transfer from a surrounding fluid to a body is given by

$$q(t) = hA\left(T_a - T(t)\right), \tag{2.2}$$

where A is the surface area of the body and h is the *convection heat transfer coefficient* which will have units of $\frac{W}{m^2}K$. Equation 2.2 should make perfect sense. The rate at which energy is transferred from the body to the fluid, or *vice-versa* is proportional to the difference in their temperatures and the amount of area over which it may occur.

Since temperature is a measure of the amount of thermal energy contained in the body, the rate of change of temperature should be proportional to the rate at which energy is transferred into the body. This is true, and in particular,

$$q(t) = \rho V c \frac{dT}{dt}(t), \tag{2.3}$$

where ρ is the density of the fluid, V is the volume and c is the *specific heat* of the material, which has units of $\frac{J}{kg}K$.

Since conservation of energy requires that the rate of heat transfer into the body must equal the rate of change of its internal energy, Equation 2.2

and 2.3 must be equal, so we have

$$hA\left(T_a - T(t)\right) = \rho V c \dot{T}(t).$$

If we let $\theta(t) = T(t) - T_a$, then

$$-hA\theta(t) = \rho V c \dot{\theta}(t)$$

or

$$\dot{\theta}(t) + \frac{hA}{\rho V c}\theta(t) = 0.$$

Usually, this equation is written in the form

$$\dot{\theta}(t) + \frac{1}{RC}\theta(t) = 0, \tag{2.4}$$

where R is the resistance to convective heat transfer and C is called the lumped thermal capacitance.[1] Equation 2.4 is a linear, first order, ordinary, constant coefficient, homogeneous differential equation. ■

The next section outlines how to solve various forms of first order equations. As it turns out, there are multiple ways to solve equation 2.4, and in particular, the two different methods from section 2.3.2 may be used to solve this problem.

The next examples come from the field of bioengineering. First we need to consider some basic reaction rate concepts.

The *Michaelis-Menton equation* describes many physiological processes; among other things, biological process catalyzed by enzymes and protein facilitated diffusion of substances into or out of cells. The form of the equation is

$$v_o = \frac{v_{max}[s]}{k_m + [s]} \tag{2.5}$$

where v_o is the reaction rate or uptake rate, $[s]$ is the concentration of some substrate and v_{max} and k_m are constants which depend upon the particular process under consideration. A plot of v_o *vs.* $[s]$ for various values of v_{max} and k_m are illustrated in Figures 2.1 and 2.2.

Example 2.2.2 The rate of uptake of blood plasma glucose into skeletal muscle, the brain, liver and other organs for oxidation (use for energy) is regulated by hormones such as insulin and are facilitated in the different organs by the GLUT family of proteins. Thus if we let g represent plasma

[1] A careful reader, or one with a background in heat transfer, will recognize that fact that when we use $T(t)$ to represent the temperature of the body, it is implicitly assuming that the temperature distribution in the body is uniform. This is intuitively appropriate in some cases, and is rigorously justified when the *Boit number,* , which is defined as the dimensionless quantity $Bi = \frac{hk}{L}$, where k is the *thermal conductivity* of the body and L is a characteristic length of the body. When $Bi \ll 1$, then the approach taken in this example problem, which is called the *lumped capacitance method,* is a justified approximation. See [12] for a complete exposition.

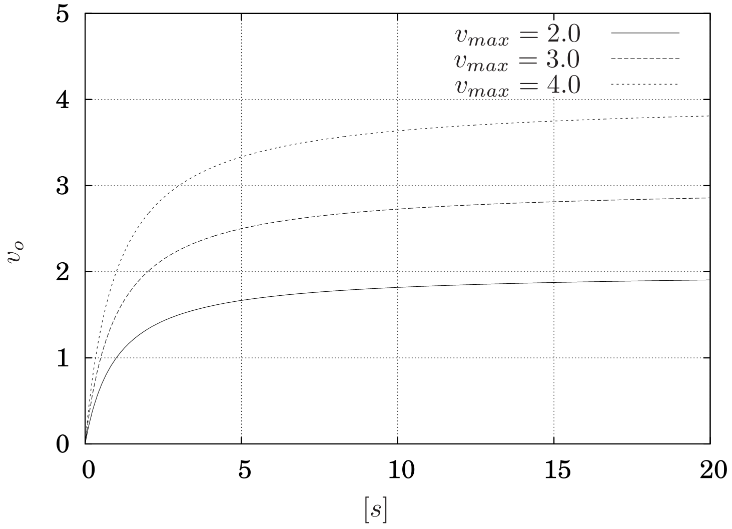


Figure 2.1. Reaction rate for various v_{max} and $k_m = 1.0$.

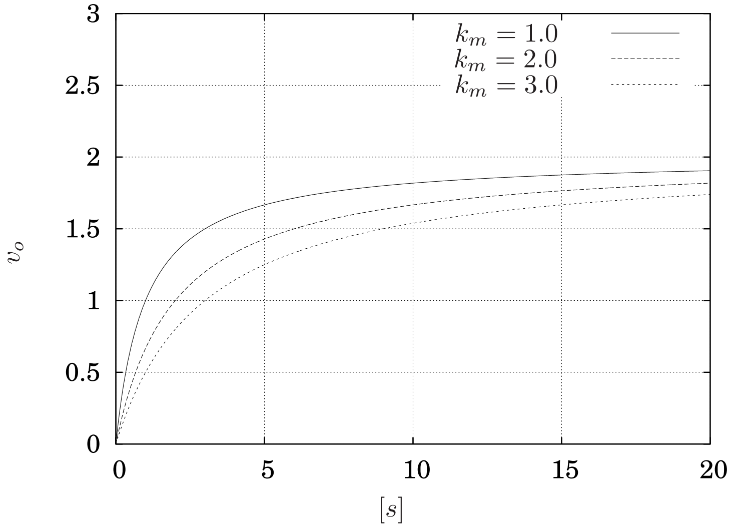


Figure 2.2. Reaction rate for various k_m and $v_{max} = 2.0$.

glucose levels, the change in plasma glucose concentrations due to uptake by, say in, skeletal muscle, is given by

$$\dot{g} = -\frac{v_{max}g}{k_m + g}$$

or

$$k_m\dot{g} + g\dot{g} + v_{max}g = 0. \tag{2.6}$$

This is an ordinary, first order, nonlinear differential equation. ■

Example 2.2.3 The rates of metabolism of many drugs are described by equation 2.5 as well. In some cases, the constant k_m is either very large or very small compared to the blood concentration of the drug so that some simplifications are possible.

For example, alcohol is such that if x represents blood alcohol concentrations, x is always much larger than k_m. In this case the denominator of equation 2.5 can be approximated by $k_m + x \approx x$, and then the equation describing the blood alcohol concentration as a function of time is

$$\dot{x} = -v_{max}. \tag{2.7}$$

This is an ordinary, first order, constant coefficient, inhomogeneous, linear differential equation. ■

Example 2.2.4 For other drugs, cocaine is an illicit example but there are many pharmaceutical examples, metabolism is such that the constant k_m is very large compared to the drug concentration levels. In that case, the denominator of equation 2.5 can be approximated simply by k_m, *i.e.,* $k_m + x \approx k_m$ and the blood drug concentration as a function of time is given by

$$k_m\dot{x} = -v_{max}x \tag{2.8}$$

which is an ordinary, first order, constant coefficient, homogeneous, linear differential equation. ■

2.3 Solution Methods

Section 2.2 presents several first order differential equations. This section considers various methods to solve them.

Since a first order ordinary differential equation can contain, at most, three terms, its simple structure lends itself to certain solution methods which will not generally apply to higher order equations. In particular, the methods outlined in section 2.3.4 make use of this fact.

2.3.1 Ordinary First Order Linear Homogeneous Constant Coefficient Differential Equations

Because it is the case that the coefficients of the dependent variable terms in engineering differential equations are often parameters which describe the physical properties of a system, and it is also often the case that such parameters are constant (mass, thermal capacitance, *etc.*), it is often the case that differential equations in engineering have constant coefficients. This section presents a method to solve ordinary, first order, constant coefficient, linear differential equations.

Before addressing ordinary, first order, constant coefficient, linear homogeneous differential equations, consider the following fact regarding ordinary, constant coefficient, linear, homogeneous differential equations of *any order*.

If you remember anything from differential equations, remember the following: **ordinary, linear, constant coefficient, homogeneous differential equations of any order have exponential solutions**. To emphasize the fact, let us make it a theorem.

Theorem 2.3.1 *Ordinary, linear, constant coefficient, homogeneous differential equations with dependent variable x and independent variable t have solutions of the form $x = ce^{\lambda t}$ where c is a non-zero constant.*

Proof Consider an nth order, ordinary, linear, constant coefficient, homogeneous differential equation of the form

$$\alpha_n \frac{d^n x}{dt^n} + \alpha_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \cdots + \alpha_1 \frac{dx}{dt} + \alpha_0 x = 0. \tag{2.9}$$

To verify the form of the solution, simply substitute $x = ce^{\lambda t}$ into Equation 2.9:

$$\alpha_n \lambda^n ce^{\lambda t} + \alpha_{n-1} \lambda^{n-1} ce^{\lambda t} + \cdots + \alpha_1 \lambda ce^{\lambda t} + \alpha_0 e^{\lambda t} = 0.$$

Since $ce^{\lambda t}$ is never zero, it is legitimate to divide each side of the equation by it which gives

$$\alpha_n \lambda^n + \alpha_{n-1} \lambda^{n-1} + \cdots + \alpha_1 \lambda + \alpha_0 = 0 \tag{2.10}$$

which is an nth order polynomial in λ. Since, by the fundamental theorem of algebra, Equation 2.10 has n solutions, there may be, in fact, up to n different solutions of the form $x = e^{\lambda t}$. □

Remark 2.3.2 The fact bears repeating: **ordinary, linear, constant coefficient, homogeneous differential equations of any order have exponential solutions**. ◇

Armed with this knowledge, we now consider solutions to ordinary, first order, constant coefficient, linear differential equations. This will provide the general solution to Equation 2.8 (since it is already homogeneous) as well as the homogeneous solution to equations 2.4 and 2.7. It will do nothing for us for Equation 2.6 since it is not linear.

Any ordinary, first order, homogeneous, linear, constant coefficient differential equation can be written as

$$\dot{x} + \alpha x = 0.$$

Note that there is no restriction that α may not be zero. Assuming a solution of the form

$$x(t) = ce^{\lambda t}$$

and substituting gives

$$\lambda = -\alpha$$

or

$$x(t) = ce^{-\alpha t}. \tag{2.11}$$

Example 2.3.3 Returning to example 2.2.4, we have a general solution of the form

$$x(t) = ce^{-\frac{v_{max}}{k_m}t}. \tag{2.12}$$

To determine c, we would have to know the initial blood concentration of the drug. Assuming $x(0) = x_0$, substituting $t = 0$ into Equation 2.12 gives $c = x_0$, so the solution to

$$\begin{aligned} k_m\dot{x} &= -v_{max}x \\ x(0) &= x_0 \end{aligned}$$

is

$$x(t) = x_0 e^{-\frac{v_{max}}{k_m}t}.$$

■

Remark 2.3.4 It is practically worth memorizing that the solution to

$$\begin{aligned} \dot{x} + \alpha x &= 0 \\ x(0) &= x_0 \end{aligned}$$

is

$$x(t) = x_0 e^{-\alpha t}. \tag{2.13}$$

⋄

2.3.2 Ordinary First Order Linear Inhomogeneous Constant Coefficient Differential Equations

Now we consider the same case as in the previous section but where the equation is inhomogeneous. Two solution methods will be presented. The first is easier, but only works when the inhomogeneous term is in a particular class of functions, and the second is computationally a bit harder, but will always work. Both approaches require that a homogeneous solution be known, so the first order of business is to determine the homogeneous solution in the form of Equation 2.11 (not in the form of Equation 2.13) as outlined in the previous section.

Undetermined coefficients

The idea behind undetermined coefficients is relatively simple, as is illustrated by the following example. The approach has two components. First a homogeneous and particular solution are determined separately and then combined for the solution (this will be mathematically justified after the example). Second, a particular form of the particular solution is assumed, which is then substituted into the differential equation which will give rise to equations for some undetermined coefficients in the particular solution.

Example 2.3.5 Solve

$$\begin{aligned} \dot{x} + 3x &= \sin 2t \qquad (2.14) \\ x(0) &= 1. \end{aligned}$$

This is an ordinary, first order, linear, constant coefficient, inhomogeneous differential equation. From Equation 2.11, the homogeneous solution is

$$x_h(t) = ce^{-3t}$$

where c is an arbitrary real number.

To determine the particular solution, consider the following logic. We seek a function, $x(t)$ such that if we take its derivative and add it to three times itself we will obtain the function $\sin 2t$. A moment's reflection will result in the conclusion that the only sorts of functions that can be combined with their derivative to obtain a sine function are sines and cosines that are a function of the same argument. So, it is logical to assume that the particular solution is of the form

$$x_p(t) = c_1 \cos 2t + c_2 \sin 2t$$

where c_1 and c_2 are coefficients that are yet to be determined, *i.e.,* the undetermined coefficients. The manner in which to compute the undetermined coefficients is obvious: substitute x_p into the differential equation to see if equations for c_1 and c_2 can be derived. So, since $\dot{x}_p(t) = -2c_1 \sin 2t + 2c_2 \cos 2t$, and substituting gives

$$\begin{aligned} \dot{x} + 3x &= (-2c_1 \sin 2t + 2c_2 \cos 2t) + 3\left(c_1 \cos 2t + c_2 \sin 2t\right) \\ &= (2c_2 + 3c_1) \cos 2t + (-2c_1 + 3c_2) \sin 2t \\ &= \sin 2t, \end{aligned}$$

where the last $\sin 2t$ term is the inhomogeneous term from Equation 2.14. Since the second and third lines of the above equation must be true for all time, then

$$\begin{aligned} 3c_1 + 2c_2 &= 0 \\ -2c_1 + 3c_2 &= 1, \end{aligned}$$

which gives $c_1 = -\frac{2}{13}$ and $c_2 = \frac{3}{13}$, so the particular solution is

$$x_p(t) = -\frac{2}{13}\cos 2t + \frac{3}{13}\sin 2t.$$

The final task is to ensure that the initial condition is satisfied, *i.e.*, $x(0) = 1$. Note the following two facts:

1. The particular solution satisfies Equation 2.14 but does not satisfy the initial condition.
2. The homogeneous solution does not satisfy the differential equation in Equation 2.14, but does have a coefficient that has not yet been specified that perhaps may be used to in some way to satisfy the initial condition.

Now observe that since x_h is a homogeneous solution, by definition when it is substituted into Equation 2.14 the result will be zero. So, *since the equation is linear*, it may be added to the particular solution and the sum will still satisfy the differential equation. In particular, using $x = x_h + x_p$ and substituting gives

$$\begin{aligned} \dot{x} + 3x &= (\dot{x}_h + \dot{x}_p) + 3(x_h + x_p) \\ &= (\dot{x}_h + 3x_h) + (\dot{x}_p + 3x_p) \\ &= 0 + (\dot{x}_p + 3x_p) \\ &= \sin 2t. \end{aligned}$$

So, since $x = x_h + x_p$ satisfies Equation 2.14 and also contains a coefficient that has not yet been specified (the c in x_h) evaluating $x(0)$ and setting it equal to the initial condition will give an equation for c. So,

$$\begin{aligned} x(0) &= x_h(0) + x_p(0) \\ &= c + -\frac{2}{13}. \end{aligned}$$

Since the initial condition was $x(0) = 1$, clearly $c = \frac{15}{13}$ and the solution to the differential equation is

$$x(t) = \frac{15}{13}e^{-3t} + -\frac{3}{13}\cos 2t + \frac{3}{13}\sin 2t.$$ ■

At first glance, the main idea behind the undetermined coefficients approach may seem to be simply educated guesswork. However, the method is actually guaranteed to work if the right conditions are met. Insight into the method is obtained by observing that certain functions have only a finite number of linearly independent derivatives.

Example 2.3.6 Returning to Example 2.3.5, we computed that if

$$x(t) = c_1 \cos 2t + c_2 \sin 2t,$$

If the inhomogeneous term, $g(t)$ is	then assume for $x_p(t)$
$\hat{c}\cos\omega t$	$c_1\cos\omega t + c_2\sin\omega t$
$\hat{c}\sin\omega t$	$c_1\cos\omega t + c_2\sin\omega t$
$\hat{c}e^{\lambda t}$	$ce^{\lambda t}$
$\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0$	$c_n t^n + \cdots + c_1 t + c_0$
sum of above terms	sum of corresponding terms
product of above terms	product of corresponding terms

Table 2.1. Forms to assume for x_p depending on the inhomogeneous term $g(t)$.

then

$$\dot{x}(t) + 3x(t) = (3c_1 + 2c_2)\cos 2t + (-2c_1 + 3c_2)\sin 2t.$$

The critical observation is that we started with a function of the form

$$x(t) = c_1\cos 2t + c_2\sin 2t,$$

and after substituting it into the differential equation obtained a function of the form

$$x(t) = k_1\cos 2t + k_2\sin 2t.$$

Specifically, a linear combination of the function $x(t)$ and its derivative is exactly the same form as the original function, albeit with different coefficients. ■

As the following theorem shows that if the inhomogeneous term, $g(t)$ is such that it only has a finite number of linearly independent[2] derivatives, then, assuming a solution that is a linear combination of $g(t)$ and its derivatives will always lead to a set of equations that will give a solution for the undetermined coefficients. First we need to define what it means for functions to be linearly independent.

Definition 2.3.7 A set of functions, $f_1(t), \ldots, f_n(t)$ is *linearly dependent* on an interval $\mathcal{I} = (t_0, t_1)$ if there exists a set of constants, $c_1, \ldots, c_n$ which are not all zero such that

$$c_1 f_1(t) + c_2 f_2(t) + \cdots + c_n f_n(t) = 0, \quad \forall t \in \mathcal{I}. \qquad \diamond$$

If the functions are not linearly dependent, then they are *linearly independent.*

[2] A necessary consition for functions to be linearly independent is provided subsequently in Section 3.2.2.

Theorem 2.3.8 *An nth order, linear, ordinary, constant coefficient, inhomogeneous differential of the form*

$$\alpha_n \frac{d^n x}{dt^n}(t) + \alpha_{n-1} \frac{d^{n-1} x}{dt^{n-1}}(t) + \cdots + \alpha_1 \frac{dx}{dt}(t) + \alpha_0 x(t) = g(t) \tag{2.15}$$

where $g(t)$ has only a finite number of linearly independent derivatives has a particular solution

$$x_p(t) = c_0 g(t) + c_1 \frac{dg}{dt}(t) + c_2 \frac{d^2 g}{dt^2}(t) + \cdots + c_m \frac{d^m g}{dt^m}(t)$$

as long as none of the functions $x_p(t)$ is not a homogeneous solution to Equation 2.15 for any combination of the coefficients c_i where not all of them are zero and where m is the number of linearly independent derivatives of $g(t)$.

Proof Consider the vector space

$$V = \left\{ c_0 g\,(t) + c_1 \frac{dg}{dt}\,(t) + \cdots + c_m \frac{d^m g}{dt^m}\,(t) \quad | \quad c_i \in \mathbb{R}, i \in \{1, \ldots, m\} \right\}.$$

The functions $g(t), \frac{dg}{dt}(t), \ldots, \frac{d^m g}{dt^m}(t)$ are the basis elements for V. By assumption the operator $\frac{d}{dt}$ is a linear operator on V. Conseqently

$$\mathrm{D} = \alpha_0 + \alpha_1 \frac{d}{dt} + \cdots + \alpha_m \frac{d^m}{dt^m}$$

is also a linear operator on V. The null space of D is the empty set since by assumption no element of V is a homogeneous solution to Equation 2.15. This implies that the set of functions $\mathrm{D}\,g(t), \mathrm{D}\,\frac{dg}{dt}(t), \ldots, \mathrm{D}\,\frac{d^m g}{dt^m}(t)$ also is a basis for V. Hence,

$$\begin{aligned} \mathrm{D}\,x_p(t) &= c_1\,\mathrm{D}\,g(t) + \cdots + c_m\,\mathrm{D}\,\frac{d^m g}{dt^m}(t) \\ &= g(t) \end{aligned}$$

will be satisfied by a unique set of coefficients, c_i. □

Example 2.3.9 Determine the particular solution to the ordinary, first order, linear, constant coefficient, inhomogeneous differential equation

$$3\dot{x} + 6x = 9e^t.$$

Assume $x_p(t) = c_1 e^t$. Then $\dot{x}_p(t) = c_1 e^t$ and substituting gives

$$3c_1 e^t + 6c_1 e^t = 9e^t \quad \Longrightarrow \quad c_1 = 1.$$

Hence

$$x_p(t) = e^t.$$

■

Complication: when the inhomogeneous term contains a homogeneous solution

It may be the case that the form of the particular solution that table 2.1 indicates also happens to be a homogeneous solution to the differential equation. The reason this is problematic is that it will be impossible to determine the undetermined coefficients since the assumed form of the solution must be equal to zero rather than the inhomogeneous term. An example illustrates this conundrum and a subsequent example illustrates the solution.

Example 2.3.10 Use the method of undetermined coefficients to determine the general solution to

$$\dot{x} + 3x = e^{-3t} + \sin 2t.$$

Referring to table 2.1, it is logical to assume

$$x_p(t) = c_1 e^{-3t} + c_2 \sin 2t + c_3 \cos 2t.$$

Differentiating and substituting gives

$$\begin{aligned} &\left(-3c_1 e^{-3t} + 2c_2 \cos 2t - 2c_3 \sin 2t\right) \\ &\quad + \quad 3\left(c_1 e^{-3t} + c_2 \sin 2t + c_3 \cos 2t\right) = e^{-3t} + \sin 2t. \end{aligned}$$

Collecting coefficients of $e^{-3t}, \sin 2t$ and $\cos 2t$ respectively gives the following set of equations

$$\begin{aligned} -3c_1 + 3c_1 &= 1 \\ -2c_3 + 3c_2 &= 1 \\ 2c_2 + 3c_3 &= 0. \end{aligned}$$

Note that the first equation is $0 = 1$, *i.e.,* there does not exist any c_1 that will satisfy the equations.

This problem is due to the fact that e^{-3t} is, in addition to being a component of the inhomogeneous term, a homogeneous solution to the differential equation. When it is substituted into the differential equation it *must* evaluate to zero, by definition. ■

If appropriately using the method of undetermined coefficients for first order equations, this complication can only happen with an equation of the form

$$\dot{x} + \alpha x = e^{-\alpha t}. \tag{2.16}$$

Table 2.1 would indicate to choose $x_p(t) = ce^{-\alpha t}$; however, this is also the homogeneous solution. Using a technique that actually foreshadows the method of variation of parameters presented subsequently, the approach will be to assume a particular solution of the form

$$x_p(t) = \mu(t) e^{-\alpha t},$$

substitute into Equation 2.16 and use the result to (hopefully) determine $\mu(t)$. Differentiating $x_p(t)$ and substituting gives

$$\left(\dot{\mu}(t)e^{-\alpha t} - \alpha\mu(t)e^{-\alpha t}\right) + \alpha\mu(t)e^{-\alpha t} = e^{-\alpha t},$$

which simplified to

$$\dot{\mu}(t) = 1$$

or

$$\mu(t) = t + c.$$

Hence,

$$x_p(t) = (t + c)\, e^{-\alpha t}.$$

Note that since the term $ce^{-\alpha t}$ is actually a homogeneous solution, it is not necessary to add it to the particular solution at this stage in the process of determining the solution since it will be added to it subsequently anyway. So the simplest form for the particular solution is

$$x_p(t) = te^{-\alpha t}.$$

Hence, when the assumed form of the particular solution is also the homogeneous solution to the differential equation, the solution is to multiply the assumed form by the independent variable.

Example 2.3.11 Continuing from example 2.3.10, instead of assuming

$$x_p(t) = c_1 e^{-3t} + c_2 \sin 2t + c_3 \cos 2t$$

assume

$$x_p(t) = c_1 te^{-3t} + c_2 \sin 2t + c_3 \cos 2t.$$

Differentiating and substituting gives

$$\begin{aligned} &\left(c_1 e^{-3t} - 3c_1 te^{-3t} + 2c_2 \cos 2t - 2c_3 \sin 2t\right) \\ &+ \quad 3\left(c_1 te^{-3t} + c_2 \sin 2t + c_3 \cos 2t\right) = e^{-3t} + \sin 2t. \end{aligned}$$

Collecting terms now gives

$$\begin{aligned} c_1 &= 1 \\ -2c_3 + 3c_2 &= 1 \\ 2c_2 + 3c_3 &= 0 \end{aligned}$$

which has the solution

$$\begin{aligned} c_1 &= 1 \\ c_2 &= \frac{3}{13} \\ c_3 &= -\frac{2}{13}, \end{aligned}$$

and hence

$$x_p(t) = te^{-3t} + \frac{2}{13}\sin 2t - \frac{2}{13}\cos 2t,$$

and the general solution is

$$x(t) = x_h(t) + x_p(t) = ce^{-3t} + te^{-3t} + \frac{2}{13}\sin 2t - \frac{2}{13}\cos 2t. \qquad \blacksquare$$

So, the recommended procedure in using the method of undetermined coefficients is to solve for the homogeneous solution first. In that way it will be possible to identify *a priori* if the assumed form of $x_p(t)$ will contain any homogeneous solutions. Alternatively, this scenario can be recognized if the situation arises when it is algebraically impossible to determine the coefficients that will make the particular solution satisfy the differential equation. In this latter case, however, one must be careful to be sure that the form for $x_p(t)$ is otherwise correct.

Variation of parameters

This method will always work for linear first order ordinary differential equations. As long as one is willing to evaluate the integrals required, it will yield the solution.

The idea behind the variation of parameters method is, that if a homogeneous solution for a differential equation is known, denoted by x_h, then assume a solution of the form $x(t) = \mu(t)x_h(t)$. Substituting the assumed form of the solution into the differential equation will yield and equation for μ that, if it can be solved, will give the solution. Unlike the method for undetermined coefficients, this method will work for variable coefficients as well, but this section will limit the coverage to the constant coefficient case. Also unlike the case for undetermined coefficients, no special form of the inhomogeneous term is necessary.

Consider the ordinary, first order, linear, constant coefficient, inhomogeneous differential equation

$$\begin{aligned} \dot{x} + \alpha x &= g(t) \\ x(t_0) &= x_0. \end{aligned}$$

From before, $x_h(t) = ce^{-\alpha t}$. Assume $x(t) = c\mu(t)e^{-\alpha t}$. Substituting into the differential equation gives

$$\begin{aligned} c\dot{\mu}e^{-\alpha t} - c\mu\alpha e^{-\alpha t} + \alpha c\mu e^{-\alpha t} &= c\dot{\mu}e^{-\alpha t} \\ &= g(t). \end{aligned}$$

Hence

$$\dot{\mu} = \frac{1}{c}e^{\alpha t}g(t)$$

which can be directly integrated. So

$$\mu(t) - \mu(t_0) = \int_{t_0}^{t} \frac{1}{c} e^{\alpha s} g(s) ds$$

or

$$\mu(t) = \int_{t_0}^{t} \frac{1}{c} e^{\alpha s} g(s) ds + \mu(t_0)$$

where $\mu(t_0)$ is arbitrary. So

$$\begin{aligned} x(t) &= \mu(t) c e^{-\alpha t} \\ &= c e^{-\alpha t} \int_{t_0}^{t} \frac{1}{c} e^{\alpha s} g(s) ds + \mu(t_0) c e^{-\alpha t} \\ &= e^{-\alpha t} \int_{t_0}^{t} e^{\alpha s} g(s) ds + c_1 e^{-\alpha t} \end{aligned}$$

where $c_1 = \mu(t_0)c$. Evaluating $x(t_0)$ gives

$$\begin{aligned} x(t_0) &= e^{-\alpha t} \int_{t_0}^{t} e^{\alpha s} g(s) ds + c_1 e^{-\alpha t_0} \\ &= c_1 e^{-\alpha t_0} \\ &= x_0. \end{aligned}$$

Thus $c_1 = x_0 e^{\alpha t_0}$ and

$$x(t) = e^{-\alpha t} \int_{t_0}^{t} e^{\alpha s} g(s) ds + x_0 e^{\alpha t_0} e^{-\alpha t}. \tag{2.17}$$

Remark 2.3.12 If the initial condition were not specified and a general solution were desired, the integral in the above method would become an indefinite integral and a constant of integration would be necessary. It is left as an exercise for the student to prove that the general solution to the ordinary, first order, linear, constant coefficient, inhomogeneous differential equation

$$\dot{x} + \alpha x = g(t) \tag{2.18}$$

is

$$x(t) = e^{-\alpha t} \int e^{\alpha t} g(t) dt + c e^{-\alpha t}. \tag{2.19}$$

◇

2.3.3 Variation of Parameters for Ordinary First Order Linear Inhomogeneous Variable Coefficient Differentail Equations

The same procedure as above may be used in the case of ordinary, first order, linear, variable coefficient, inhomogeneous differential equations. Consider

$$\begin{aligned} \dot{x} + h(t)x &= g(t) \qquad (2.20) \\ x(t_0) &= x_0. \qquad (2.21) \end{aligned}$$

The procedure will be the same as before: find a homogeneous solution, $x_h(t)$, assume the solution of the form $x(t) = \mu(t)x_h(t)$, substitute to determine an equation for $\mu(t)$, and if possible, solve for $\mu(t)$. The first task is to determine the homogeneous solution, which is not simply $x_h(t) = ce^{\lambda t}$ in the case of a variable coefficient.

First consider the corresponding homogeneous equation

$$\frac{dx_h(t)}{dt} + h(t)x_h(t) = 0.$$

Rearranging gives

$$\frac{1}{x_h(t)}\frac{dx_h(t)}{dt} = -h(t).$$

Integrating each side with respect to t gives

$$\begin{aligned}\int \frac{1}{x_h(t)}\frac{dx_h(t)}{dt}dt &= \int \frac{d}{dt}\left(\ln\left(x_h(t)\right)\right)dt \\ &= \ln\left(x_h(t)\right) + c \\ &= -\int h(t)dt.\end{aligned}$$

Hence

$$x_h(t) = ke^{-\int h(t)dt}, \tag{2.22}$$

where $k = -e^{-c}$.

Remark 2.3.13 This procedure to find the homogeneous solution is a special case of the method for separable equations outlined in section 2.3.4. ⋄

Now armed with the homogeneous solution, assume a solution of the form

$$x(t) = \mu(t)x_h(t) = \mu(t)ke^{\int h(t)dt}.$$

Substituting gives

$$\begin{aligned}(\dot{\mu}x_h + \mu\dot{x}_h) + h(t)\left(\mu x_h\right) &= \left(\dot{\mu}ke^{-\int h(t)dt} - \mu h(t)ke^{-\int h(t)dt}\right) + h(t)\left(\mu ke^{-\int h(t)dt}\right) \\ &= \dot{\mu}ke^{-\int h(t)dt} \\ &= g(t).\end{aligned}$$

Hence

$$\dot{\mu} = \frac{1}{k}g(t)e^{\int h(t)dt} \quad \Longrightarrow \quad \mu(t) = \int\left(\frac{1}{k}g(t)e^{\int h(t)dt}\right)dt + c.$$

and

$$\begin{aligned}x(t) &= \left(\int\left(\frac{1}{k}g(t)e^{\int h(t)dt}\right)dt + c\right)\left(ke^{-\int h(t)dt}\right) \\ &= \left(\int\left(g(t)e^{\int h(t)dt}\right)dt + c\right)\left(e^{-\int h(t)dt}\right).\end{aligned} \tag{2.23}$$

Remark 2.3.14 Occasionally it will be convenient to combine arbitrary constants but not change the name of the variable, as was done in Equation 2.23. The constant k was distributed across both terms in the left term of the equation, so the constant term c is now actually ck; however, since both c and k are arbitrary, it is most convenient just to keep the variable name as c. ◇

At this point it is worth observing that Equation 2.23 is the solution to Equation 2.20. The only possible complication is that sometimes the integrals may not have a closed-form solution, or may simply be difficult to evaluate.

Example 2.3.15 Determine the general solution to

$$\dot{x} + \frac{3}{t}x = \sin t.$$

■

Since this equation is of the form of Equation 2.20, so the general solution is given by Equation 2.23 where $h(t) = \frac{3}{t}$ and $g(t) = \sin t$. Substituting into the solution gives

$$\begin{aligned}
x(t) &= \left(\int \left(g(t) e^{\int h(t)dt} \right) dt + c \right) \left(e^{-\int h(t)dt} \right) \\
&= \left(\int \left((\sin t)\, e^{\int \frac{3}{t}dt} \right) dt + c \right) \left(e^{-\int \frac{3}{t}dt} \right) \\
&= \left(\int \left((\sin t)\, e^{3\ln t} \right) dt + c \right) \left(e^{-3\ln t} \right) \\
&= \left(\int \left((\sin t)\, e^{3\ln t} \right) dt + c \right) \left(e^{-3\ln t} \right)
\end{aligned}$$

2.3.4 Ordinary First Order Nonlinear Differential Equations

Unfortunately, it is generally the case that nonlinear differential equations are difficult, at best, and generally do not even have closed-form solutions. In the case of first order equations, however, there is one case in which a solution may be obtained, and that case is the so-called exact equation. Before presenting the theory and method of exact equations, the next section presents a simplified, special case of exact equations, namely, separable equations.

Separable equations

A notationally simplistic, yet nonetheless useful, description of the idea behind separable equations is, that if it is possible to put all the terms that are a function of the dependent variable on one side of the equation and all the terms that are a function of the independent variable on the other side of the equation the equation is separable. In such a case, both sides may be directly integrated.

Example 2.3.16 Find the general solution to

$$(x+1)\left(t^2+5t+3\right)=x\dot{x}.$$

This may be rearranged as

$$t^2+5t+3=\frac{x}{x+1}\frac{dx}{dt}$$

and each side may be integrated with respect to t

$$\int t^2+5t+3dt=\int\frac{x(t)}{x(t)+1}\frac{dx(t)}{dt}dt.$$

Recall from calculus the substitution rule for integration, namely,

$$\int_{t_0}^{t} f\left(x(s)\right)\frac{dx(s)}{ds}ds=\int_{x(t_0)}^{x(t)} f(x)dx.$$

Using this fact,

$$\begin{aligned}\int t^2+5t+3dt &= \int\frac{x(t)}{x(t)+1}\frac{dx(t)}{dt}dt\\ &= \int\frac{x}{x+1}dx,\end{aligned}$$

so

$$\frac{t^3}{3}+\frac{5t^2}{2}+3t=x(t)-\ln\left(x(t)+1\right)+c.$$

Note that one problem is that the solution, $x(t)$ may be, as is the case in this example, only determined as an implicit function of the dependent variable. ■

The preceding example was rather precise and in practice the approach is a bit more formal. In words, the simplest way to approach the problem is to notationally treat $\dot{x}$ as $\frac{dx}{dt}$ and try to manipulate the equation so that all the x terms are on one side of the equation along with the dx term and all the t terms are on the other side with the dt term. While this casual use of notational convenience works correctly in this case, it is important to recognize that what is actually going on is an integration by substitution on the x side of the equation. Another example will illustrate this point and complete the treatment of separable equations. It will also illustrate the slight variation in the approach when the problem is an initial value problem rather than finding a general solution. The only difference being that data is now available to make the integrals definite integrals.

Example 2.3.17 Determine the solution to

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

Note this can perhaps be more easily solved by directly using Equation 2.23 with $g(t) = 0$; however, just for the fun of it this example will solve it by separation of variables.

A bit of manipulation gives

$$\frac{dx}{dt} + \sin(t)\, x = 0 \qquad \Longleftrightarrow \qquad \frac{dx}{x} = -\sin(t)\, dt,$$

so

$$\int_{x(t_0)}^{t} \frac{1}{x} dx = -\int_{t_0}^{t} \sin(s)\, ds$$

or

$$\int_{2}^{x} (t)\frac{1}{x} dx = \int_{1}^{t} \sin(t)\, dt \qquad \Longleftrightarrow \qquad \ln x - \ln 2 = \cos t - \cos 1,$$

which gives, upon taking the exponential of each side

$$x(t) = 2e^{\cos t - \cos 1}.$$

■

Exact equations

While actually *using* it is another matter, the *idea* behind exact equations is actually quite simple. Consider a function, $\psi(x(t), t)$ (as usual, t is the independent variable and x is the dependent variable) and consider the level sets of ψ; namely, $\psi(x(t), t) = c$. Differentiating ψ with respect to time gives

$$\frac{d\psi(x(t), t)}{dt} = \frac{\partial \psi}{\partial x}\frac{dx}{dt} + \frac{\partial \psi}{\partial t} = 0.$$

Note that this is of the form

$$f(x, t)\dot{x} + g(x, t) = 0, \tag{2.24}$$

where f and g are functions of both the independent variable, t, and the dependent variable x. If a differential equation just so happens to be of the form of Equation 2.24 such that there exists a $\psi(x(t), t)$ such that $\frac{\partial \psi}{\partial x} = f(x, t)$ and $\frac{\partial \psi}{\partial t} = g(x, t)$, then solving Equation 2.24 is simply a matter of determining ψ and setting $\psi(x, t) = c$ for the general solution. The correct value of c will be determined from the initial condition in the case of the initial value problem.

Since the order of differentiating the partial derivatives does not matter, *i.e.,*

$$\frac{\partial^2 \psi}{\partial x \partial t} = \frac{\partial^2 \psi}{\partial t \partial x}$$

and since

$$\frac{\partial \psi}{\partial x} = f(x, t) \quad \text{and} \quad \frac{\partial \psi}{\partial t} = g(x, t)$$

the following are equivalent

$$\frac{\partial^2 \psi}{\partial x \partial t} = \frac{\partial^2 \psi}{\partial t \partial x} \qquad \Longleftrightarrow \qquad \frac{\partial f}{\partial t} = \frac{\partial g}{\partial x}.$$

In other words, this proves the following theorem.

Theorem 2.3.18 *For the ordinary, first order differential equation*

$$f(x,t)\dot{x} + g(x,t) = 0, \tag{2.25}$$

if

$$\frac{\partial f}{\partial t} = \frac{\partial g}{\partial x}$$

then there exists a function $\psi(x(t),t)$ *such that*

$$\frac{\partial \psi}{\partial x} = f(x,t) \quad \textit{and} \quad \frac{\partial \psi}{\partial t} = g(x,t).$$

The general solution to Equation 2.25 is given implicitly by

$$\psi(x(t),t) = c.$$

So far, so good, but while the theory is nice and tidy, there are still two practical problems. First, the solution is only given implicitly by ψ. Second, we still need to determine a way to find ψ. The first problem is inherent in the method and is unavoidable. The second problem will be addressed subsequently. First, an example.

Example 2.3.19 Consider

$$2x\dot{x} = -2t - 1.$$

In this case $f(x,t) = 2x$ and $g(x,t) = 2t+1$.

$$\begin{aligned}\frac{\partial f}{\partial t} &= 0\\ \frac{\partial g}{\partial x} &= 0,\end{aligned}$$

the equation is exact. Note that

$$\psi(x,t) = x^2 + t^2 + t$$

is such that

$$\dot{\psi} = 0 \qquad \Longleftrightarrow \qquad 2x\dot{x} + 2t + 1 = 0,$$

so

$$x^2 + t^2 + t = c$$

gives the solution $x(t)$ implicitly. ■

Determining $\psi(x,t)$ is actually rather straight-forward. Since

$$\frac{\partial \psi}{\partial x} = f(x,t)$$

then

$$\psi(x,t) = \int f(x,t)dx + h(t),$$

and

$$\begin{aligned} g(x,t) &= \frac{\partial \psi}{\partial t} \\ &= \frac{\partial}{\partial t}\left(\int f(x,t)dx + h(t)\right) \\ &= \frac{\partial}{\partial t}\left(\int f(x,t)dx\right) + \dot{h}(t). \end{aligned}$$

Thus,

$$h(t) = \int \left(g(x,t) - \frac{\partial}{\partial t}\left(\int f(x,t)dx\right)\right) dt$$

and the general solution is given by

$$\psi(x,t) = \int f(x,t)dx + \int \left(g(x,t) - \frac{\partial}{\partial t}\left(\int f(x,t)dx\right)\right) dt = c.$$

2.4 Stability

2.5 Summary

Ordinary first order differential equations are solved using the following methods.

- If the equation is *linear, constant coefficient and homogeneous*, then assuming a solution of the form $x(t) = ce^{\lambda t}$ is probably the easiest method.

- If the equation is *linear, variable coefficient and homogeneous*, then using equation 2.22 is probably the easiest method.

- If the equation is *linear, constant coefficient and inhomogeneous* with an inhomogeneous term of the form given in Table 2.1, then the method of undetermined coefficients outlined in section 2.3.2 is probably the easiest.

- If the equation is *linear, constant coefficient and inhomogeneous* the method of variation of parameters with a solution given by equation 2.19 will work. If the inhomogeneous term is not given in Table 2.1 then this is probably the easiest method.

- If the equation is *linear, variable coefficient and inhomogeneous* the method of variation of parameters with a solution given by equation 2.23 will work.

- If the equation is *nonlinear*, then first check if it is separable. If it is not, then check if it is exact.

2.6 Exercises

Problem 2.1 For each of the first order differential equations listed in Problem 1.1, determine which, if any, of the following solution methods apply based upon what has been covered in this book so far.

1. Assuming exponential solutions
2. Undetermined coefficients.
3. Variation of parameters.
4. Using the fact that the equation is separable.
5. Using the fact that the equation is exact.
6. Determining an approximate numerical solution.

It may be the case that no method, one method or more than one method may apply.

Problem 2.2 In dead organic matter, the C^{14} isotope decays at a rate proportional to the amount of it that is present. Furthermore, it takes approximately 5600 years for half of the original amount present to decay.

1. If $x(0)$ denotes the amount present when the organism is alive, determine a differential equation that describes the amount of the C^{14} isotope present if $x(t)$ represents the amount present after time t elapses after the organism dies.
2. In contrast, to C^{14}, the C^{12} isotope does not decay and the ratio of C^{12} to C^{14} is constant while an organism is alive. Hence, one should be able to compare the ratio of the two isotopes in a dead specimen to that of a live specimen. Determine how many years have elapsed if the ratio of the the amount of C^{14} to C^{12} is 30% of the original value.

Problem 2.3 You are in desperate need to determine (as in make up), by hand, 100 different exact first order differential equations in less than one hour. What would be a good way to do that? Determine 10 different exact first order ordinary differential equations using your method.

Problem 2.4 Use two different methods to determine the general solution to

$$\dot{x} + x = \sin 5t.$$

Also, find the solution if $x(0) = 0$.

Problem 2.5 Determine the general solution to

$$\dot{x} + \frac{x}{t} = \cos 5t.$$

Problem 2.6 Use two different methods to determine the general solution to

$$\dot{x} + 5x = e^{-5t}.$$

Also, find the solution if $x(0) = 1$ and plot the solution *versus* time for a length of time that is appropriate to demonstrate the qualitative nature of the solution.

Problem 2.7 Determine the general solution to

$$(2x + 1)\,\dot{x} = 3t^2.$$

If necessary, you may express the solution as an implicit function.

Problem 2.8 Use two different methods to determine the general solution to

$$3t^2\dot{x} + 6tx + 5 = 0.$$

Problem 2.9 Determine the solution to

$$\begin{aligned} t\dot{x} + 2x &= t^2 - t + 1 \\ x(1) &= \frac{1}{2} \\ t &> 0. \end{aligned}$$

Problem 2.10 Prove that all separable first order ordinary differential equations are exact. In other words, show that separable first order differential equations are a special case of exact first order ordinary differential equations.

Problem 2.11 Prove that Equation 2.19 is the solution to Equation 2.18.

Problem 2.12 Consider the first order, linear, variable coefficient, homogeneous ordinary differential equation

$$\dot{x} + tx = 0.$$

Does assuming a solution of the form

$$x(t) = e^{\lambda t}$$

work? Why or why not?

Problem 2.13 Consider the first order, nonlinear, ordinary differential equation

$$\dot{x} + x^2 = 0.$$

Does assuming a solution of the form

$$x(t) = e^{\lambda t}$$

work? Why or why not?

Chapter 3

Ordinary Second Order Linear Constant Coefficient Equations

3.1 Introduction

Second order equations arise quite frequently in engineering. In mechanical and aerospace engineering, in particular, they arise often in the context of the study of *vibrations.* First let us consider a few prototypical example problems example to help motivate the importance of second order ordinary differential equations as well as to illustrate their apparent importance.

Example 3.1.1 Consider the very simple mechanical system illustrated on the left in Figure 3.1. The scenario modeled by the problem is that a mass is attached to a moving base by a spring (on the left) and viscous damper (on the right). The base is moving with a specified motion $z(t)$. The question is what is the resulting motion of the mass?

A free body diagram of the mass is illustrated on the right in Figure 3.1, where f_s and f_d are the forces that the spring and damper exert on the mass, respectively. Assume that y and z are measured from a configuration where the spring is unstretched, *i.e.,* at $y = 0$ and $z = 0$ the spring is unstretched. In that case, $f_s = k\,(z - y)$. A viscous damper is such that the force required to compress or extend it is proportional to the rate at which it is being compressed or extended, respectively. Or mathematically, $f_d = b\,(\dot{z} - \dot{y})$.

Newton's law gives

$$m\ddot{y} = f_s + f_d = k\,(z - y) + b\,(\dot{z} - \dot{y})$$

or rearranging

$$m\ddot{y} + b\dot{y} + ky = kz + b\dot{z}.$$

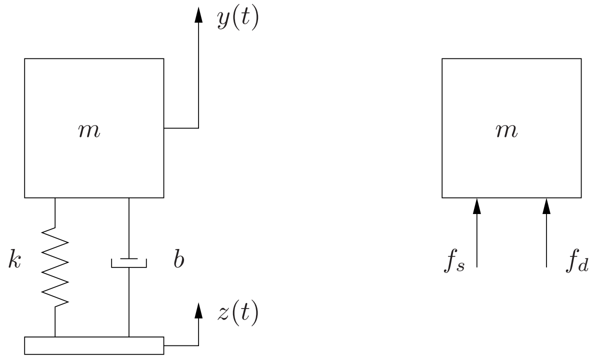


Figure 3.1. Mechanical system for example 3.1.1.

This is an ordinary, second order, linear, constant coefficient, inhomogeneous differential equation. Remember that $z(t)$ is assumed to be known. For example, if $z(t) = Z \sin \omega t$, then

$$m\ddot{y} + b\dot{y} + ky = kZ \sin \omega t + bZ \cos \omega t$$

which is more obviously in such a form.

For such a system, important and interesting questions may be the following:

1. Given $y(0)$ and $\dot{y}(0)$ what is the resulting motion of the mass, $y(t)$?

2. What is the magnitude of the resulting motion of the mass as a function of the magnitude of the base motion, Z?

3. What is the magnitude of the resulting motion of the mass as a function of the frequency of the motion of the base, ω?

4. Given either or both Z and ω, what are good choices for k and b so that the magnitude of either or both the motion or acceleration of the mass is minimized? (This is basically designing a suspension or vibration absorber). ■

3.2 Theory of Linear, Homogeneous Equations

In the study of mechanical vibrations, the starting point is usually the case of *free, undamped vibrations*. This case is illustrated by the following example and will be the starting point for the study of homogeneous equations.

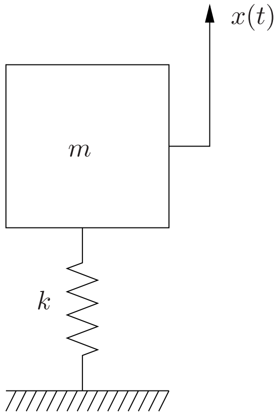


Figure 3.2. Mass spring system for example 3.2.1.

Example 3.2.1 Consider the mass–spring system illustrated in Figure 3.2. Assume for present purposes that there is no gravitational force acting on the spring and that $x = 0$ when the spring is unstretched. The only force on the mass is due to the spring and the equation of motion will be

$$m\ddot{x} + kx = 0. \tag{3.1}$$

Recalling theorem 2.3.1, this ordinary, second order, homogeneous, linear, constant coefficient differential equation must have solutions of the form

$$x(t) = ce^{\lambda t}.$$

Substituting this into the differential equation gives

$$m\lambda^2 ce^{\lambda t} + kce^{\lambda t} = 0.$$

Since $e^{\lambda t}$ is never zero and assuming that $c \neq 0$, we have the characteristic equation

$$m\lambda^2 + k = 0$$

or

$$\lambda = \pm\sqrt{-\frac{k}{m}}.$$

Since real spring constants and masses have only positive values,

$$\lambda = \pm i\sqrt{\frac{k}{m}}.$$

Let $\omega_n = \sqrt{\frac{k}{m}}$ denote the *natural frequency* of the system. Using this notation, there are two possible solutions

$$x_1(t) = e^{i\omega_n t} \quad \text{and} \quad x_2(t) = e^{-i\omega_n t}.$$

The fact that these two functions are indeed solutions to the differential equation may be verified by direct substitution. We will return to this example subsequently after a quick review of complex variables and then take a detour into the notion of a *fundamental set of solutions* for second order equations. ■

At this point it would behoove the reader to read Appendix A which gives a several page review of complex variable theory. The one fact that will be necessary and repeated here is the definition of Euler's formula, which relates the exponential of a complex number to trigonometric functions. Namely,

$$e^{(\alpha+i\beta)t} = e^{\alpha t}\left(\cos\beta t + i\sin\beta t\right).$$

Example 3.2.2 Returning to example 3.2.1 and using Euler's formula the two solutions may be rewritten as

$$\begin{aligned} x_1(t) &= \cos\omega_n t + i\sin\omega_n t \\ x_2(t) &= \cos\omega_n t - i\sin\omega_n t. \end{aligned}$$

Now consider the question: when will it be possible to combine the two solutions to satisfy any specified initial conditions? First, note that because the equation is linear and homogeneous, the solution

$$x(t) = c_1 x_1(t) + c_2 x_2(t)$$

also satisfies the differential equation. This may be verified by direct substitution using either form of the solutions. Using the sine and cosine form, then

$$\begin{aligned} \dot{x}_1(t) &= -\omega_n \sin\omega_n t + i\omega_n\cos\omega_n t \\ \dot{x}_2(t) &= -\omega_n \sin\omega_n t - i\omega_n\cos\omega_n t \\ \ddot{x}_1(t) &= -\omega_n^2 \cos\omega_n t - i\omega_n^2\sin\omega_n t \\ \ddot{x}_2(t) &= -\omega_n^2 \cos\omega_n t + i\omega_n^2\sin\omega_n t \end{aligned}$$

and substituting into equation 3.1 and using the fact that $\omega_n^2 = \frac{k}{m}$ gives

$$\begin{aligned} m\ddot{x} + kx &= m\left(c_1\ddot{x}_1 + c_2\ddot{x}_2\right) + k\left(c_1 x_1 + c_2 x_2\right) \\ &= mc_1\left(-\omega_n^2\cos\omega_n t - i\omega_n^2\sin\omega_n t\right) \\ &\quad + mc_2\left(-\omega_n^2\cos\omega_n t + i\omega_n^2\sin\omega_n t\right) \\ &\quad + kc_1\left(\cos\omega_n t + i\sin\omega_n t\right) \\ &\quad + kc_2\left(\cos\omega_n t - i\sin\omega_n t\right) \\ &= -c_1 k\cos\omega_n t - ic_1 k\sin\omega_n t \\ &\quad - c_2 k\cos\omega_n t + ic_2 k\sin\omega_n t \\ &\quad + kc_1\cos\omega_n t + ikc_1\sin\omega_n t \\ &\quad + kc_2\cos\omega_n t - ikc_2\sin\omega_n t \\ &= 0. \end{aligned}$$

The fact that a linear combination of the two solutions of equation 3.1 also satisfies the equation is a particular example of the *principle of superposition.* ■

3.2.1 The principle of superposition

The principle of superposition states that any linear combination of solutions to an ordinary, linear, homogeneous differential equation is also a solution. An example is the final computation in example 3.2.2. The following theorem proves the principle for the second order case. Proving it for nth order equations is left as an exercise.

Theorem 3.2.3 *Let the functions $x_1(t)$ and $x_2(t)$ each satisfy the ordinary, second order, linear, homogeneous differential equation*

$$f_2(t)\ddot{x}(t) + f_1(t)\dot{x}(t) + f_0(t)x = 0. \tag{3.2}$$

Then any linear combination of $x_1(t)$ and $x_2(t)$, i.e.,

$$x(t) = c_1 x_1(t) + c_2 x_2(t),$$

also satisfies equation 3.2.

PROOF The proof is simply by direct substitution.

$$\begin{aligned} f_2(t)\ddot{x}(t) + f_1(t)\dot{x}(t) + f_0(t)x(t) &= f_2(t)\left(c_1\ddot{x}_1(t) + c_2\ddot{x}_2(t)\right) \\ &\quad + f_1(t)\left(c_1\dot{x}_1(t) + c_2\dot{x}_2(t)\right) \\ &\quad + f_0(t)\left(c_1 x_1(t) + c_2 x_2(t)\right) \\ &= c_1\left(f_2(t)\ddot{x}_1(t) + f_1(t)\dot{x}_1(t) + f_0(t)x_1(t)\right) \\ &\quad + c_2\left(f_2(t)\ddot{x}_2(t) + f_1(t)\dot{x}_2(t) + f_0(t)x_2(t)\right) \\ &= 0 + 0 \\ &= 0. \end{aligned}$$

Note that the principle of superposition does not require that the equation have constant coefficients; however, that will be the predominant case in which we will use it.

Example 3.2.4 Returning to example 3.2.2, at this point it has been shown that the two functions

$$\begin{aligned} x_1(t) &= \cos\omega_n t + i\sin\omega_n t \\ x_2(t) &= \cos\omega_n t - i\sin\omega_n t \end{aligned}$$

are solutions to equation 3.1 and that any linear combination

$$x(t) = c_1 x_1(t) + c_2 x_2(t)$$

is also a solution.

Because this text is also concerned with the solutions and analyses of vibrations problems, an alternative, simpler form of the combination of the two solutions would be nice. To achieve this, it is possible to simply rearrange the linear combination as follows:

$$\begin{aligned} x(t) &= c_1 x_1(t) + c_2 x_2(t) \\ &= c_1 (\cos \omega_n t + i \sin \omega_n t) + c_2 (\cos \omega_n t - i \sin \omega_n t) \\ &= (c_1 + c_2) \cos \omega_n t + i (c_1 - c_2) \sin \omega_n t \\ &= \hat{c}_1 \cos \omega_n t + \hat{c}_2 \sin \omega_n t, \end{aligned}$$

where

$$\begin{aligned} \hat{c}_1 &= c_1 + c_2 \\ \hat{c}_2 &= i (c_1 - c_2). \end{aligned}$$

Note also, by direct substitution, it may be verified that the two functions

$$\begin{aligned} \hat{x}_1(t) &= \sin \omega_n t \\ \hat{x}_2(t) &= \cos \omega_n t \end{aligned}$$

are also solutions to equation 3.1. This is an otherwise unremarkable fact, but they are generally a more convenient representation of two homogeneous solutions than complex exponentials. ■

3.2.2 Linear independence

Now consider the question of determining when it will be the case that any initial conditions can be satisfied by appropriately determining the two unspecified coefficients in the various forms of the solutions above.

Example 3.2.5 Adding initial conditions to the problem statement corresponding to the system illustrated in Figure 3.2 gives

$$\begin{aligned} m\ddot{x} + kx &= 0 \\ x(0) &= x_0 \\ \dot{x}(0) &= \dot{x}_0. \end{aligned}$$

The examples above showed that

$$\begin{aligned} x(t) &= c_1 \hat{x}_1(t) + c_2 \hat{x}_2(t) \\ &= c_1 \cos \omega_n t + c_2 \sin \omega_n t \end{aligned}$$

is a solution to the differential equation. Now to determine the values for c_1 and c_2 that satisfy the initial conditions, simply evaluate $x(0)$ and $\dot{x}(0)$ and set them equal to x_0 and $\dot{x}_0$ respectively

$$\begin{aligned} x(0) &= c_1 \\ \dot{x}(0) &= c_2 \omega_n \end{aligned}$$

so

$$\begin{aligned} c_1 &= x_0 \\ c_2 &= \frac{\dot{x}_0}{\omega_n}. \end{aligned}$$

The main point of this example is the fact that regardless of the values given for x_0 and $\dot{x}_0$, there are values for c_1 and c_2 that satisfy the differential equation as well as the initial conditions and the solution is

$$x(t) = x_0 \cos \omega_n t + \frac{\dot{x}_0}{\omega_n} \sin \omega_n t.$$

■

Now consider the more general question: given two solutions, $x_1(t)$ and $x_2(t)$ of an ordinary, second order, linear, homogeneous differential equation, when will it be the case that the two coefficients in the general solution

$$x(t) = c_1 x_1(t) + c_2 x_2(t)$$

may be used to satisfy *any* given initial conditions?

Stated a bit more mathematically, consider the initial value problem

$$\begin{aligned} \ddot{x} + p(t)\dot{x} + q(t)x &= 0 \\ x(t_0) &= x_0 \\ \dot{x}(t_0) &= \dot{x}_0 \end{aligned}$$

and assume that $x_1(t)$ and $x_2(t)$ are solutions. From the principle of superposition in theorem 3.2.3, since the equation is ordinary, linear and homogeneous, then

$$x(t) = c_1 x_1(t) + c_2 x_2(t)$$

also satisfies the differential equation.

Now solving $x(t_0) = x_0$ and $\dot{x}(t_0) = \dot{x}_0$ for c_1 and c_2 gives

$$\begin{aligned} x(t_0) &= c_1 x_1(t_0) + c_2 x_2(t_0) = x_0 \\ \dot{x}(t_0) &= c_1 \dot{x}_1(t_0) + c_2 \dot{x}_2(t_0) = \dot{x}_0 \end{aligned}$$

which yields

$$c_1 = \frac{\dot{x}_0 x_2(t_0) - x_0 \dot{x}_2(t_0)}{x_1(t_0)\dot{x}_2(t_0) - x_2(t_0)\dot{x}_1(t_0)} \tag{3.3}$$

$$c_2 = \frac{\dot{x}_0 x_1(t_0) - x_0 \dot{x}_1(t_0)}{x_1(t_0)\dot{x}_2(t_0) - x_2(t_0)\dot{x}_1(t_0)}, \tag{3.4}$$

so the only time there will be a problem with solving for the coefficients is then the denominator is equal to zero (note that both denominators are equal). Observe furthermore that the denominators are only a function of the two solutions, $x_1(t)$ and $x_2(t)$ and *not* the initial conditions. This leads to the following definition.

Definition 3.2.6 Given n functions, $x_1(t), x_2(t), \ldots, x_n(t)$ define the *Wronskian*, W as the following determinant

$$W\left(x_1(t), x_2(t), \ldots, x_n(t)\right) = \begin{vmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \\ \frac{dx_1(t)}{dt} & \frac{dx_2(t)}{dt} & \cdots & \frac{dx_n(t)}{dt} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^n x_1(t)}{dt^n} & \frac{d^n x_2(t)}{dt^n} & \cdots & \frac{d^n x_n(t)}{dt^n} \end{vmatrix}.$$

$\diamond$

In particular, relevant to the case of second order differential equations, the Wronskian for two functions, $x_1(t)$ and $x_2(t)$ is

$$W\left(x_1(t), x_2(t)\right) = \begin{vmatrix} x_1(t) & x_2(t) \\ \dot{x}_1(t) & \dot{x}_2(t) \end{vmatrix} = x_1(t)\dot{x}_2(t) - x_2(t)\dot{x}_1(t).$$

Because it is commonly used, the following definition introduces the notion of linear independence for a set of functions.

Definition 3.2.7 A set of functions, $x_1(t), x_2(t), \ldots, x_n(t)$ is called *linearly independent on an interval $\mathcal{I}$* if there exist constants $c_1, \ldots, c_n$ that are not all zero such that

$$c_1 x_1(t) + c_2 x_2(t) + \cdots + c_n x_n(t) = 0$$

for all $t \in \mathcal{I}$. $\diamond$

A necessary condition for functions to be independent is given by the Wronskian

~~If~~

Theorem 3.2.8 *If*

$$W\left(x_1(t), x_2(t), \ldots, x_n(t)\right) \neq 0$$

for any $t \in \mathcal{I}$, then the set of functions, $x_1(t), x_2(t), \ldots, x_n(t)$ is linearly independent on $\mathcal{I}$.

Proof If the functions are linearly dependent, then there exists $c_1, \ldots, c_n$, not all zero such that

$$c_1 x_1(t) + c_2 x_2(t) + \cdots + c_n x_n(t) = 0.$$

Differentiating this equation with repsect to t gives

$$c_1 \dot{x}_1(t) + c_2 \dot{x}_2(t) + \cdots + c_n \dot{x}_n(t) = 0,$$

and differentiating it $n-1$ more times gives the system of equations

$$\begin{aligned} c_1 x_1(t) + c_2 x_2(t) + \cdots + c_n x_n(t) &= 0 \\ c_1 \dot{x}_1(t) + c_2 \dot{x}_2(t) + \cdots + c_n \dot{x}_n(t) &= 0 \\ c_1 \ddot{x}_1(t) + c_2 \ddot{x}_2(t) + \cdots + c_n \ddot{x}_n(t) &= 0 \\ &\vdots \\ c_1 x_1^{(n-1)}(t) + c_2 \ddot{x}_2^{(n-1)}(t) + \cdots + c_n x_n^{(n-1)}(t) &= 0. \end{aligned}$$

In order for there to be a nonzero solution to this system of equations, the Wronskian must be zero for all $t \in \mathcal{I}$, *i.e.*,,

$$W\left(x_1(t), x_2(t), \ldots, x_n(t)\right) = 0$$

Hence, if the Wronskian is nonzero for any $t \in \mathcal{I}$, the set of functions must be linearly independent. □

In particular, for the case of two functions, $x_1(t)$ and $x_2(t)$ are linearly independent if

$$W\left(x_1(t), x_2(t)\right) = x_1(t)\dot{x}_2(t) - x_2(t)\dot{x}_1(t) \neq 0.$$

Finally, putting all results of the previous few pages together gives the following theorem.

Theorem 3.2.9 *If $x_1(t)$ and $x_2(t)$ satisfy*

$$\ddot{x} + p(t)\dot{x} + q(t)x = 0$$

and if

$$W\left(x_1(t), x_2(t)\right) = x_1(t)\dot{x}_2(t) - x_2(t)\dot{x}_1(t) \neq 0,$$

then

$$x(t) = c_1x_1(t) + c_2x_2(t)$$

satisfies the initial value problem

$$\begin{aligned} \ddot{x} + p(t)\dot{x} + q(t)x &= 0 \\ x(t_0) &= x_0 \\ \dot{x}(t_0) &= \dot{x}_0 \end{aligned}$$

where c_1 and c_2 are given by equations 3.3 and 3.4.

So, if the goal is to solve an ordinary, linear, homogeneous, second order initial value problem, then it will not suffice to find *any* two homogeneous solutions to combine, but two *linearly independent* solutions. Fortunately, as will be illustrated subsequently, the methods developed in the next few sections will all generate linearly independent solutions, so careful attention to this detail will not be necessary.

3.3 Constant Coefficient, Homogeneous Equations

Recalling theorem 2.3.1 it is clear that ordinary, linear, constant coefficient, homogeneous, second order differential equations have solutions of the form $x(t) = e^{\lambda t}$. However, in contrast to the case of first order equations of this type, there will generally be *two* solutions for λ which complicates matters somewhat, as is illustrated by the following example.

Example 3.3.1 Determine the general solution to

$$\ddot{x} + 5\dot{x} + 6x = 0. \tag{3.5}$$

Assuming

$$x(t) = e^{\lambda t}$$

and substituting gives

$$\begin{aligned} \ddot{x} + 5\dot{x} + 6x &= \lambda^2 e^{\lambda t} + 5\lambda e^{\lambda t} + e^{\lambda t} \\ &= 0 \end{aligned}$$

and since $e^{\lambda t}$ is never equal to zero

$$\lambda^2 + 5\lambda + 6 = 0.$$

Using the quadratic formula (or simply factoring, as is possible in this case) gives

$$\lambda = -2 \quad \text{or} \quad \lambda = -3$$

so

$$\begin{aligned} x_1(t) &= e^{-2t} \\ x_2(t) &= e^{-3t} \end{aligned}$$

both satisfy equation 3.5 and

$$x(t) = c_1 e^{-2t} + c_2 e^{-3t}$$

is a general solution as long as the Wronskian, $W(x_1, x_2)$ is nonzero. Checking the Wronskian gives

$$W(x_1, x_2) = \begin{vmatrix} e^{-2t} & e^{-3t} \\ -2e^{-2t} & -3e^{-3t} \end{vmatrix} = -3e^{-5t} + 2e^{-5t} = -1e^{-5t} \neq 0. \quad \blacksquare$$

Now, it should be clear that assuming exponential solutions for second order equations of this type will result in a quadratic characteristic equation which, in general, will have two roots. Unfortunately, a quadratic equation may have either distinct roots, a complex conjugate pair of roots or a repeated root, and each case results in a solutions of a different type. Hence, each case must be considered separately.

As a recurring example throughout the investigation of the three possible cases, consider the system mass-spring-damper illustrated in Figure 3.3 which will have the equation of motion

$$m\ddot{x} + b\dot{x} + kx = 0. \tag{3.6}$$

Assuming $x(t) = e^{\lambda t}$ will result in the characteristic equation

$$m\lambda^2 + b\lambda + k = 0$$

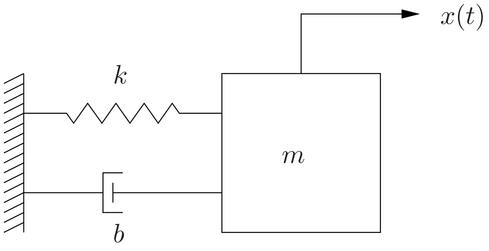


Figure 3.3. Mechanical system described by equation 3.6.

which has roots

$$\lambda_1 = \frac{-b+\sqrt{b^2-4mk}}{2m} \quad \text{and} \quad \lambda_2 = \frac{-b-\sqrt{b^2-4mk}}{2m}.$$

The roots, λ_1 and λ_2 will either be

- real and distinct (when $b^2 - 4mk > 0$);
- a complex conjugate pair (when $b^2 - 4mk < 0$); or,
- repeated (when $b^2 - 4mk = 0$).

As is already clear, the solutions to equation 3.6 will involve the parameters m, b and k. However, there exists a standard, canonical form for such equations. This form is valuable to know both because it is a standard formulation for second order problems and also because it simplifies notation the problem.

Canonical form for second order systems

Consider

$$m\ddot{x} + b\dot{x} + kx = 0$$

and the following definitions

$$\begin{aligned} \zeta &= \frac{b}{2\sqrt{mk}} \\ \omega_d &= \omega_n\sqrt{1-\zeta^2} \\ &= \sqrt{\frac{k}{m}}\sqrt{1-\frac{b^2}{4mk}} \\ &= \sqrt{\frac{k}{m}\frac{4mk-b^2}{4mk}} \\ &= \sqrt{\frac{4mk-b^2}{2m}}. \end{aligned}$$

The first term, ζ is called *the damping ratio* and the second term, ω_d is called *the damped natural frequency.* Observing that

$$\frac{b}{2m} = \zeta\omega_n$$

equation 3.10 can be rewritten as

$$\begin{aligned} m\ddot{x} + b\dot{x} + kx &= m\left(\ddot{x} + \frac{b}{m}\dot{x} + \frac{k}{m}x\right) \\ &= m\left(\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x\right). \end{aligned}$$

Since the equation is homogeneous and $m \neq 0$, the two differential equations are equivalent, *i.e.,*

$$m\ddot{x} + b\dot{x} + kx = 0 \qquad \Longleftrightarrow \qquad \ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 0.$$

In this case, the characteristic equation is

$$\lambda^2 + 2\zeta\omega_n\lambda + \omega_n^2 = 0$$

which gives

$$\begin{aligned} \lambda &= \frac{-2\zeta\omega_n \pm \sqrt{4\zeta^2\omega_n^2 - 4\omega_n^2}}{2} \\ &= -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}. \end{aligned}$$

The three cases corresponding to distinct, real roots, complex conjugate roots and repeated roots correspond to the cases where $\zeta > 1$, $\zeta < 1$ and $\zeta = 1$ respectively. The "simplification" is in that the roots only contain two parameters, ω_n and ζ instead of the three parameters, m, b and k.

3.3.1 Distinct, real roots

In the case that the quadratic equation has distinct real roots, as was illustrated in example 3.3.1, the two solutions

$$\begin{aligned} x_1(t) &= e^{\lambda_1 t} \\ x_2(t) &= e^{\lambda_2 t} \end{aligned}$$

where λ_1 and λ_2 are the roots of the characteristic equation, will both satisfy the differential equation and by the principle of superposition the linear combination

$$x(t) = c_1 x_1(t) + c_2 x_2(t)$$

will also satisfy it. Furthermore, in this case where $\lambda_1 \neq \lambda_2$ the Wronskian is always nonzero. This fact is illustrated by the direct computation,

$$\begin{aligned} W(x_1, x_2) &= \begin{vmatrix} e^{\lambda_1 t} & e^{\lambda_2 t} \\ \lambda_1 e^{\lambda_1 t} & \lambda_2 e^{\lambda_2 t} \end{vmatrix} \qquad (3.7) \\ &= \lambda_2 e^{(\lambda_1+\lambda_2)t} - \lambda_1 e^{(\lambda_1+\lambda_2)t} \\ &= (\lambda_2 - \lambda_1)\, e^{(\lambda_1+\lambda_2)t} \\ &\neq 0. \end{aligned}$$

To emphasize its importance, the above results are restated in the form of a theorem.

Theorem 3.3.2 *For an ordinary, second order, linear, constant coefficient, homogeneous differential equation, if the roots of the corresponding characteristic equation are real and distinct, denoted by λ_1 and λ_2, then the functions*

$$\begin{aligned} x_1(t) &= e^{\lambda_1 t} \\ x_2(t) &= e^{\lambda_2 t} \end{aligned}$$

both are solutions to the differential equation. Furthermore $x_1(t)$ and $x_2(t)$ are linearly independent and therefore

$$x(t) = c_1 e^{\lambda t} + c_2 t e^{\lambda t}$$

is a general solution of the differential equation.

In the case of the mass-spring-damper system illustrated in figure 3.3 and described by equation 3.6 the general solution will be

$$\begin{aligned} x(t) &= c_1 e^{\frac{-b+\sqrt{b^2-4mk}}{2m}t} + c_2 e^{\frac{-b-\sqrt{b^2-4mk}}{2m}t} \\ &= c_1 e^{\left(-\zeta\omega_n+\omega_n\sqrt{\zeta^2-1}\right)t} + c_2 e^{\left(-\zeta\omega_n-\omega_n\sqrt{\zeta^2-1}\right)t}. \end{aligned}$$

Note that since $b^2 - 4mk > 0$ then this solution will correspond to the sum of two decaying exponentials.

3.3.2 Complex roots

In the case in the mass-spring-damper problem where $b^2 - 4mk < 0$ or $\zeta < 1$, the sign of the term inside the radical will be negative and hence the two roots will be a complex conjugate pair given by

$$\begin{aligned} \lambda_1 &= \frac{-b + i\sqrt{4mk - b^2}}{2m} = -\zeta\omega_n + i\omega_n\sqrt{1-\zeta^2} \\ \lambda_1 &= \frac{-b - i\sqrt{4mk - b^2}}{2m} = -\zeta\omega_n - i\omega_n\sqrt{1-\zeta^2}. \end{aligned}$$

Using Euler's formula, the two solutions

$$\begin{aligned} \hat{x}_1(t) &= \hat{c}_1 e^{\lambda_1 t} \\ \hat{x}_2(t) &= \hat{c}_2 e^{\lambda_2 t} \end{aligned}$$

where, using the canonical formulation

$$\begin{aligned} \hat{x}_1(t) &= \hat{c}_1 e^{-\zeta\omega_n t}\left(\cos\left(\omega_n\sqrt{1-\zeta^2}t\right) + i\sin\left(\omega_n\sqrt{1-\zeta^2}t\right)\right) \\ &= \hat{c}_1 e^{-\zeta\omega_n t}\left(\cos\omega_d t + i\sin\omega_d t\right) \\ \hat{x}_2(t) &= \hat{c}_2 e^{-\zeta\omega_n t}\left(\cos\left(\omega_n\sqrt{1-\zeta^2}t\right) - i\sin\left(\omega_n\sqrt{1-\zeta^2}t\right)\right) \\ &= \hat{c}_2 e^{-\zeta\omega_n t}\left(\cos\omega_d t - i\sin\omega_d t\right), \end{aligned}$$

and, following the procedure outlined in example 3.2.4 and defining

$$\begin{aligned} c_1 &= \hat{c}_1 + \hat{c}_2 \\ c_2 &= i\left(\hat{c}_1 - \hat{c}_2\right) \end{aligned} \tag{3.8}$$

then the two solutions

$$\begin{aligned} x_1(t) &= e^{-\zeta\omega_n t}\cos\omega_d t \\ x_2(t) &= e^{-\zeta\omega_n t}\sin\omega_d t \end{aligned} \tag{3.9}$$

may be added in a linear combination

$$x(t) = c_1\left[e^{-\zeta\omega_n t}\cos\omega_d t\right] + c_2\left[e^{-\zeta\omega_n t}\sin\omega_d t\right] \tag{3.10}$$

to form a solution.

There is nothing wrong with repeating the Wronskian computation for this case; however, it is worth noting that the computation in equation 3.7 is valid for the case where the λ's are complex as well. Also, since the combination of solutions expressed by the constants in equation 3.8 is full rank, the sine and cosine combination of the solutions will be linearly independent as well. However, just to complete the picture, the detailed Wronskian computation is as follows

$$\begin{aligned} W(x_1, x_2) &= \begin{vmatrix} e^{-\zeta\omega_n t}\cos\omega_d t & e^{-\zeta\omega_n t}\sin\omega_d t \\ -e^{-\zeta\omega_n t}\left(\omega_d\cos\omega_d t + \zeta\omega_n\sin\omega_d t\right) & e^{-\zeta\omega_n t}\left(\omega_d\cos\omega_d t - \zeta\omega_n\sin\omega_d t\right) \end{vmatrix} \\ &= \omega_d e^{-2\zeta\omega_n t} \\ &\neq 0. \end{aligned}$$

So, the above proves the following theorem.

Theorem 3.3.3 *For an ordinary, second order, linear, constant coefficient, homogeneous differential equation, if the roots of the corresponding characteristic equation are a complex conjugate pair, denoted by λ_1 and λ_2, then the functions*

$$\begin{aligned} x_1(t) &= e^{\lambda_1 t} \\ x_2(t) &= e^{\lambda_2 t} \end{aligned}$$

both are solutions to the differential equation. Furthermore $x_1(t)$ and $x_2(t)$ are linearly independent and therefore

$$x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

is a general solution of the differential equation.

Using the sine and cosine formulation gives the following corollary to theorem 3.3.3.

Corollary 3.3.4 *Equivalently, if the two roots are denoted by*

$$\begin{aligned}\lambda_1 &= -\zeta\omega_n + i\omega_n\sqrt{1-\zeta^2}\\ \lambda_2 &= -\zeta\omega_n - i\omega_n\sqrt{1-\zeta^2}\end{aligned}$$

then the functions

$$\begin{aligned}x_1(t) &= e^{-\zeta\omega_n t}\cos\omega_d t\\ x_2(t) &= e^{-\zeta\omega_n t}\sin\omega_d t\end{aligned}$$

both are solutions to the differential equation. Furthermore $x_1(t)$ and $x_2(t)$ are linearly independent and therefore

$$x(t) = c_1 e^{-\zeta\omega_n t}\cos\omega_d t + c_2 e^{-\zeta\omega_n t}\sin\omega_d t$$

is a general solution of the differential equation.

A numerical example may be helpful at this point.

Example 3.3.5 Determine a general solution to

$$\ddot{x} + 2\dot{x} + 5x = 0.$$

Just for fun, let us solve this two ways.

1. Assuming a solution of the form $x(t) = e^{\lambda t}$ gives the characteristic equation

   $$\lambda^2 + 2\lambda + 5 = 0 \qquad \Longleftrightarrow \qquad \lambda = -1 \pm 2i.$$

   Immediately we can write either

   $$x(t) = c_1 e^{(-1-2i)t} + c_2 e^{(-1+2i)t}$$

   or

   $$x(t) = c_1 e^{-t}\cos 2t + c_2 e^{-t}\sin 2t.$$

2. Alternatively, using the definition of ω_n, ζ and ω_d,

   $$\begin{aligned}\omega_n &= \sqrt{\frac{k}{m}} = \sqrt{5}\\ \zeta &= \frac{b}{2\sqrt{mk}} = \frac{1}{\sqrt{5}} = \frac{\sqrt{5}}{5}\\ \omega_d &= \omega_n\sqrt{1-\zeta^2} = \sqrt{5}\sqrt{1-\frac{1}{5}} = 2.\end{aligned}$$

   and substituting into equation 3.10 gives

   $$x(t) = c_1 e^{-t}\cos 2t + c_2 e^{-t}\sin 2t.$$

■

3.3.3 Repeated roots

Now consider the case when $b^2 = 4mk$ or, equivalently, $\zeta = 1$. In this case, $\lambda = -\zeta\omega_n$, and at this point there is only one solution

$$x_1(t) = e^{-\zeta\omega_n t}.$$

To find another solution, assume a solution of the form

$$x_2(t) = \mu(t)x_1(t),$$

substitute to see if it determines $\mu(t)$. Computing

$$\begin{aligned}\dot{x}_2 &= \mu\dot{x}_1 + \dot{\mu}x_1 \\ \ddot{x}_2 &= 2\dot{\mu}\dot{x}_1 + \mu\ddot{x}_1 + \ddot{\mu}x_1.\end{aligned}$$

and substituting into

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 0$$

gives

$$\begin{aligned}(2\dot{\mu}\dot{x}_1 + \mu\ddot{x}_1 + \ddot{\mu}x_1) + 2\zeta\omega_n\left(\mu\dot{x}_1 + \dot{\mu}x_1\right) + \omega_n^2\mu x_1 &= \\ \mu\left(\ddot{x}_1 + 2\zeta\omega_n\dot{x}_1 + \omega_n^2 x_1\right) + (2\dot{\mu}\dot{x}_1 + \ddot{\mu}x_1 + 2\zeta\omega_n\dot{\mu}x_1) &= \\ \ddot{\mu}x_1 + 2\dot{\mu}\left(\dot{x}_1 + \omega_n x_1\right) &= \\ \ddot{\mu}x_1 &= \\ \ddot{\mu} &= 0 \quad \Longrightarrow \quad \mu(t) = t + c.\end{aligned}$$

Note that in the second line the term in the left pair of parentheses is zero because x_1 is a solution to the homogeneous equation. In the third line the term in parentheses is zero due to the definition of x_1. Finally, since c is arbitrary, it may be zero and hence, finally,

$$\begin{aligned}x_2(t) &= tx_1(t) \\ &= te^{\lambda t} \\ &= te^{-\omega_n t}.\end{aligned}$$

So, the two solutions

$$\begin{aligned}x_1(t) &= e^{\lambda t} = e^{\omega_n t} \\ x_2(t) &= te^{\lambda t} = te^{\omega_n t}\end{aligned}$$

both satisfy the differential equation.

A direct computation with the Wronskian shows they are linearly independent, which is left as an exercise.

So, the above proves the following theorem.

Theorem 3.3.6 *For an ordinary, second order, linear, constant coefficient, homogeneous differential equation, if the roots of the corresponding characteristic equation, are equal, i.e., the roots are repeated, and are denoted by λ, then the following two functions*

$$\begin{aligned} x_1(t) &= e^{\lambda t} \\ x_2(t) &= te^{\lambda t} \end{aligned}$$

both are solutions to the differential equation. Furthermore $x_1(t)$ and $x_2(t)$ are linearly independent and therefore

$$x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

is a general solution of the differential equation.

An example follows.

Example 3.3.7 Find a general solution to

$$\ddot{x} + 4\dot{x} + 4 = 0.$$

The corresponding characteristic equation is

$$\lambda^2 + 4\lambda + 4 = 0.$$

Hence, $\lambda = -2$ is the repeated solution. Therefore

$$x(t) = c_1 e^{-2t} + c_2 te^{-2t}$$

is the general solution. ■

3.4 Inhomogeneous Equations

The two methods for solving inhomogeneous, second order, ordinary, linear differential equations go by the same name, and are essentially equivalent in approach, to the methods outlined in Chapter 2 for inhomogeneous first order equations; namely, the method of undetermined coefficients and the method of variation of parameters.

Before presenting the two approaches, note that any equation of this type may be converted to the canonical form. In particular,

$$m\ddot{x} + b\dot{x} + kx = f(t) \quad \Longleftrightarrow \quad \ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = \frac{f(t)}{m}. \tag{3.11}$$

Hence, it will suffice to study solutions to equations of the form of the one on the right in equation 3.11.

3.4.1 The Method of Undetermined coefficients for Contant Coefficient Equations

The method of undetermined coefficients is essentially the same as was presented for first order equations in section 2.3.2. Thus this section will limit the presentation to a few examples.

Example 3.4.1 Find the general solution to

$$m\ddot{x} + kx = F\cos\omega t.$$

From examples 3.2.1 through 3.2.4, the homogeneous solution is

$$x_h(t) = c_1\cos\omega_n t + c_2\sin\omega_n t,$$

where, as usual, $\omega_n = \sqrt{\frac{k}{m}}$. While not necessary to simply find the solution, this example will work with the normal form

$$\ddot{x} + \omega_n x = \frac{F}{m}\cos\omega t.$$

Referring to Table 2.1, as long as $\omega \neq \omega_n$, then a correct assumption for the form of the particular solution is

$$x_p(t) = A\cos\omega t + B\sin\omega t.$$

Skipping the gory details, differentiating $x_p(t)$ and substituting into the differential equation gives

$$\begin{aligned} A &= \frac{F}{m\left(\omega_n^2 - \omega^2\right)} \\ B &= 0, \end{aligned}$$

so the entire solution is

$$\begin{aligned} x(t) &= c_1\cos\omega_n t + c_2\sin\omega_n t + \frac{F}{m\left(\omega_n^2 - \omega^2\right)}\cos\omega t \\ &= c_1\cos\omega_n t + c_2\sin\omega_n t + \frac{F}{k\left(1 - \left(\frac{\omega}{\omega_n}\right)^2\right)}\cos\omega t \end{aligned}$$

To normalize the solution, note that the quantity $\frac{F}{k}$ is simply the amount that the spring would displace under the action of a force of magnitude F. Hence, define the *static deflection* as

$$\delta = \frac{F}{k}.$$

Using this the solution becomes

$$x(t) = \delta\frac{1}{1 - \left(\frac{\omega}{\omega_n}\right)^2}\cos\omega t.$$

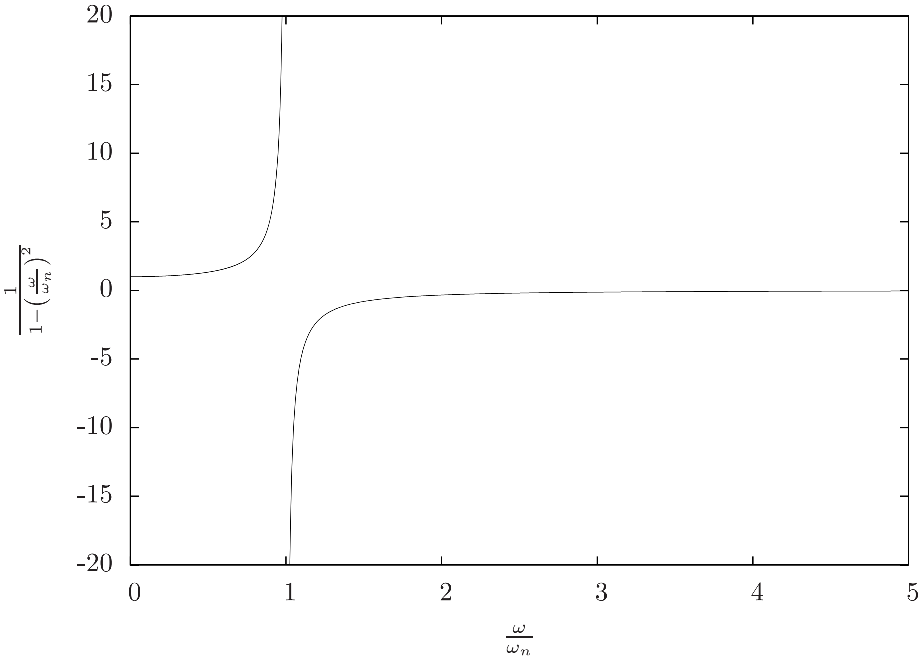


Figure 3.4. Magnification factor versus frequency ratio.

The term

$$M = \frac{1}{1-\left(\frac{\omega}{\omega_n}\right)^2}$$

is called the *magnification factor* because it multiplies the static deflection and hence determines the dependency of the magnitude of the response as a function of the forcing frequency, ω. A plot of the magnification factor versus frequency ratio, $\frac{\omega}{\omega_n}$ is illustrated in Figure 3.4. ■

The case where $\omega = \omega_n$ is referred to as *resonance*, and will be further explored in the exercises. At this point is suffices to note that it corresponds to the case where the initially assumed form of the particular solution is the same as a homogeneous solution; hence, the assumed form of $x_p(t)$ must be multiplied by t, which will result in a solution with a magnitude that increases linearly with time. This will be illustrated with a numerical example.

Example 3.4.2 Solve

$$\begin{aligned} \ddot{x} + x &= \cos t \\ x(0) &= 0 \\ \dot{x}(0) &= 0, \end{aligned}$$

which corresponds, for example, to $m = 1$, $k = 1$ and hence, $\omega_n = 1$. Assuming homogeneous solutions of the form $x_h(t) = e^{\lambda t}$ gives $\lambda = \pm i$, so

$$x_h(t) = c_1 \cos t + c_2 \sin t$$

is a homogeneous solution. Note that due to the inhomogeneous term, $\cos t$, one may be inclined to assume $x_p(t) = A\cos t + B\sin t$; however, since $\sin t$ and $\cos t$ are homogeneous solutions, then the appropriate particular solution is

$$x_p(t) = t\left(A\cos t + B\sin t\right).$$

Differentiating twice, substituting, equating coefficients of $\sin t$ and $\cos t$ and solving for A and B gives $A = 0$ and $B = \frac{1}{2}$; hence

$$x(t) = c_1 \cos t + c_2 \sin t + \frac{t\sin t}{2}.$$

Evaluating the initial conditions gives $c_1 = c_2 = 0$. Hence

$$x(t) = \frac{t\sin t}{2}$$

is the solution of the initial value problem. A plot of this solution is illustrated in Figure 3.5 and is an illustration of the phenomenon of *resonance.* Note that the solution grows unbounded, *i.e.,*

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

■

Example 3.4.3 Find the general solution to

$$\ddot{x} + \dot{x} + 4x = t\sin 2t.$$

Assuming

$$x_h(t) = e^{\lambda t}$$

gives the characteristic equation

$$\lambda^2 + \lambda + 4 = 0$$

so

$$\lambda = \frac{-1 \pm \sqrt{1-16}}{2} = -\frac{1}{2} \pm \frac{\sqrt{15}}{2} i.$$

Hence,

$$x_h(t) = e^{-\frac{1}{2}t}\left(c_1 \sin\frac{\sqrt{15}}{2}t + c_2 \cos\frac{\sqrt{15}}{2}t\right).$$

Since the inhomogeneous term is of the product of a polynomial in t and $\sin 2t$, we must assume a solution that contains the product of the all the corresponding linearly independent derivatives. Hence, assume

$$x_p(t) = At\sin 2t + Bt\cos 2t + C\sin 2t + D\cos 2t.$$

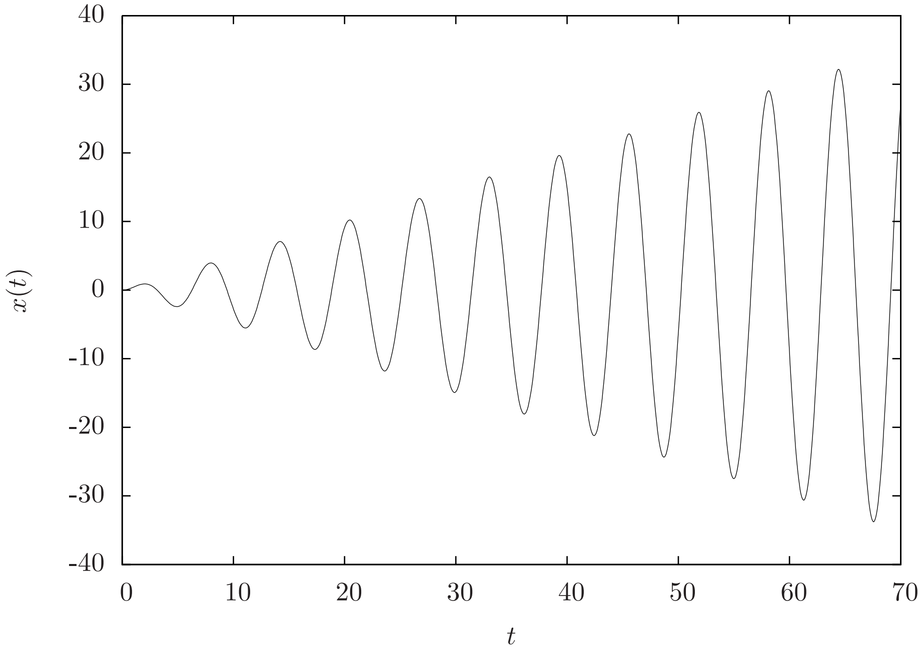


Figure 3.5. Resonance response of solution to example 3.4.2.

Differentiating gives

$$\begin{aligned}\dot{x}_p(t) &= A\sin 2t + 2At\cos 2t + B\cos 2t - 2Bt\sin 2t + 2C\cos 2t - 2D\sin 2t \\ &= -2Bt\sin 2t + 2At\cos 2t + (A-2D)\sin 2t + (B+2C)\cos 2t\end{aligned}$$

and differentiating again gives

$$\begin{aligned}\ddot{x}_p(t) &= -2B\sin 2t - 4Bt\cos 2t + 2A\cos 2t - 4At\sin 2t + \\ &\quad 2\,(A-2D)\cos 2t - 2\,(B+2C)\sin 2t \\ &= -4At\sin 2t - 4Bt\cos 2t - 4\,(B+C)\sin 2t + 4\,(A-D)\cos 2t.\end{aligned}$$

Substituting into the differential equation gives

$$\begin{aligned}[-4At\sin 2t - 4Bt\cos 2t - 4\,(B+C)\sin 2t + 4\,(A-D)\cos 2t] &\quad + \\ [-2Bt\sin 2t + 2At\cos 2t + (A-2D)\sin 2t + (B+2C)\cos 2t] &\quad + \\ 4\,[At\sin 2t + Bt\cos 2t + C\sin 2t + D\cos 2t] &\quad = \quad t\sin 2t\end{aligned}$$

and equating the coefficients of $t\sin 2t$, $t\cos 2t$, $\sin 2t$ and $\cos 2t$ respectively, gives the following set of equations

$$\begin{aligned}-4A - 2B + 4A &= 1 \\ -4B + 2A + 4B &= 0 \\ -4\,(B+C) + (A-2D) + 4C &= 0 \\ 4\,(A-D) + (B+2C) + 4D &= 0.\end{aligned}$$

From the first two equations, $A = 0$ and $B = -\frac{1}{2}$. Substituting this into the third equation gives

$$2 - 2D = 0$$

so $D = 1$. From the last equation, $C = \frac{1}{4}$. Hence

$$x_p(t) = -\frac{1}{2}t\cos 2t + +\frac{1}{4}\sin 2t + \cos 2t$$

■

and the general solution is

$$\begin{aligned}x(t) &= x_h(t) + x_p(t) \\ &= e^{-\frac{1}{2}t}\left(c_1\sin\frac{\sqrt{15}}{2}t + c_2\cos\frac{\sqrt{15}}{2}t\right) - \frac{1}{2}t\cos 2t + \frac{1}{4}\sin 2t + \cos 2t.\end{aligned}$$

3.4.2 Method of Variation of Parameters for Constant or Variable Coefficient Equations

Recall in section 2.3.2 the method of variation of parameters was used to find solutions to ordinary, first order, linear, inhomogeneous differential equations (either constant or variable coefficients). The same approach may be used in the case of second order equations; however, due to the second order nature of the problem, the computations involved become a bit more algebraically complex. Nevertheless, proceed, as before, and consider the ordinary, second order, linear, inhomogeneous differential equation

$$\ddot{x}(t) + p(t)\dot{x}(t) + q(t)x(t) = f(t) \tag{3.12}$$

and assume a particular solution of the form

$$x_p(t) = \mu_1(t)x_1(t) + \mu_2(t)x_2(t) \tag{3.13}$$

where $x_1(t)$ and $x_2(t)$ are homogeneous solutions to equation 3.12. The approach is hopefully obvious: substitute $x_p(t)$ into equation 3.12 to see if equations for $\mu_1(t)$ and $\mu_2(t)$ may be obtained. So, proceeding thusly, and dropping the explicit dependence on t

$$\begin{aligned}
\dot{x}_p &= \dot{\mu}_1 x_1 + \mu_1 \dot{x}_1 + \dot{\mu}_2 x_2 + \mu_2 \dot{x}_2 \\
\ddot{x}_p &= \ddot{\mu}_1 x_1 + 2\dot{\mu}_1 \dot{x}_1 + \mu_1 \ddot{x}_1 + \ddot{\mu}_2 x_2 + 2\dot{\mu}_2 \dot{x}_2 + \mu_2 \ddot{x}_2
\end{aligned}$$

and substituting into equation 3.12 gives

$$\begin{aligned}
\ddot{x} + p\dot{x} + qx &= \left(\ddot{\mu}_1 x_1 + 2\dot{\mu}_1 \dot{x}_1 + \mu_1 \ddot{x}_1 + \ddot{\mu}_2 x_2 + 2\dot{\mu}_2 \dot{x}_2 + \mu_2 \ddot{x}_2\right) \\
&+ p\left(\dot{\mu}_1 x_1 + \mu_1 \dot{x}_1 + \dot{\mu}_2 x_2 + \mu_2 \dot{x}_2\right) \\
&+ q\left(\mu_1 x_1 + \mu_2 x_2\right) \\
&= f.
\end{aligned}$$

Rearranging a bit gives

$$\begin{aligned}
\ddot{x} + p\dot{x} + qx &= \mu_1\left(\ddot{x}_1 + p\dot{x}_1 + qx_1\right) \\
&+ \mu_2\left(\ddot{x}_2 + p\dot{x}_2 + qx_2\right) \\
&+ \left(\ddot{\mu}_1 x_1 + 2\dot{\mu}_1 \dot{x}_1 + \ddot{\mu}_2 x_2 + 2\dot{\mu}_2 \dot{x}_2\right) + p\left(\dot{\mu}_1 x_1 + \dot{\mu}_2 x_2\right) \\
&= f,
\end{aligned}$$

and noting that since x_1 and x_2 the terms in the parentheses multiplying μ_1 and μ_2 in the first two lines are zero, the equation reduces to

$$\left(\ddot{\mu}_1 x_1 + 2\dot{\mu}_1 \dot{x}_1 + \ddot{\mu}_2 x_2 + 2\dot{\mu}_2 \dot{x}_2\right) + p\left(\dot{\mu}_1 x_1 + \dot{\mu}_2 x_2\right) = f. \tag{3.14}$$

At this point, there is one equation for two unknown functions, $\mu_1(t)$ and $\mu_2(t)$; furthermore, it is second order, so at first glance it may seem not much progress has been made since one second order equation (equation 3.12) has been with another one (equation 3.14). However, since it is one equation with two unknowns, the system is under-determined, and we have the freedom to choose another independent equation. So, let us try to make the term in the left set of parentheses zero. Note that if we choose (with much foresight)

$$\dot{\mu}_1 x_1 + \dot{\mu}_2 x_2 = 0$$

then its derivative must also be zero, so

$$\ddot{\mu}_1 x_1 + \dot{\mu}_1 \dot{x}_1 + \ddot{\mu}_2 x_2 x_2 + \dot{\mu}_2 \dot{x}_2 = 0.$$

In light of this, equation 3.14 reduces to

$$\dot{\mu}_1 \dot{x}_1 + \dot{\mu}_2 \dot{x}_2 = f.$$

Collecting what is left together, determining μ_1 and μ_2 amounts to finding the functions that satisfy

$$\begin{aligned} \dot{\mu}_1 x_1 + \dot{\mu}_2 x_2 &= 0 \\ \dot{\mu}_1 \dot{x}_1 + \dot{\mu}_2 \dot{x}_2 &= f, \end{aligned}$$

which gives

$$\begin{aligned} \dot{\mu}_1(t) &= -\frac{x_2(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)} \qquad (3.15) \\ \dot{\mu}_2(t) &= \frac{x_1(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)}. \qquad (3.16) \end{aligned}$$

Thus, if $x_1(t)$ and $x_2(t)$ are known, everything on the right hand sides of the above equations are known and $\mu_1(t)$ and $\mu_2(t)$ may be determined by direct integration. Hence

$$\begin{aligned} \mu_1(t) &= -\int \frac{x_2(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)} dt + c_1 \\ \mu_2(t) &= \int \frac{x_1(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)} dt + c_2, \end{aligned}$$

where c_1 and c_2 are the integration constants and are arbitrary. Substituting this into the original assumed form of the solution, Equation 3.13 gives

$$\begin{aligned} x_p(t) &= -x_1(t) \left(\int \frac{x_2(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)} dt + c_1 \right) \\ &+ x_2(t) \int \left(\frac{x_1(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_1(t)} dt + c_2 \right). \end{aligned}$$

Note that

1. since the denominator in each integrand must be nonzero, $x_1(t)$ and $x_2(t)$ must be linearly independent; and,
2. since $x_p(t)$ has a linear combination of the two homogeneous solutions contained in it, it is actually the complete solution.

Hence the final answer is

$$\begin{aligned} x(t) &= c_1 x_1(t) + c_2 x_2(t) \qquad (3.17) \\ &- x_1(t) \int \frac{x_2(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)} dt \\ &+ x_2(t) \int \frac{x_1(t) f(t)}{x_1(t)\dot{x}_2(t) - \dot{x}_1(t) x_2(t)} dt. \end{aligned}$$

To illustrate the use of the method, consider a couple examples.

Example 3.4.4 Find the general solution

$$\begin{aligned} \ddot{x}+x &= \frac{1}{\cos t} \\ x(0) &= 0 \\ \dot{x}(0) &= 0. \end{aligned}$$

Note that the method of undetermined coefficients cannot be used for this problem since the inhomogeneous term is not of the appropriate form. For the homogeneous solution, there are complex roots, $\lambda = \pm i$; hence,

$$\begin{aligned} x_1(t) &= \cos t \\ x_2(t) &= \sin t. \end{aligned}$$

Note that $W(x1, x2) = 1$. Substituting into Equation 3.17 gives

$$\begin{aligned} x(t) &= c_1 \cos t + c_2 \sin t - \cos t \int \frac{\sin t}{\cos t} dt + \sin t \int \frac{\cos t}{\cos t} dt \\ &= c_1 \cos t + c_2 \sin t + \cos t \ln(\cos t) + t \sin t. \end{aligned}$$

Evaluating the initial conditions gives

$$x(0) = 0 \qquad \Longleftrightarrow \qquad c_1 = 0$$

and

$$\dot{x}(0) = 0 \qquad \Longleftrightarrow \qquad c_2 = 0.$$

■

It is worth observing that probably most engineering differential equations are amenable to the method of undetermined coefficients. To show that variation of parameters works for such equations as well, the following example repeats example 3.4.2 using variation of parameters.

Example 3.4.5 Solve

$$\begin{aligned} \ddot{x}+x &= \cos t \\ x(0) &= 0 \\ \dot{x}(0) &= 0. \end{aligned}$$

The homogeneous solutions are

$$\begin{aligned} x_1(t) &= \cos t \\ x_2(t) &= \sin t. \end{aligned}$$

A quick computation shows that $W(x_1, x_2) = 1$. Hence

$$x(t) = c_1 \cos t + c_2 \sin t - \cos t \int \sin t \cos t dt + \sin t \int \cos^2 t dt.$$

Aside 3.4.6 As a quick reminder of what you should already know from calculus, these integrals will be worked out in detail. For the first one, note that if $u = \sin t$ then $\frac{du}{dt} = \cos t$; hence, by substitution

$$\begin{aligned} \int \sin t \cos t dt &= \int u \frac{du}{dt} dt \\ &= \int \frac{d}{dt}\left(\frac{u^2}{2}\right) dt \\ &= \frac{u^2}{2} + c \\ &= \frac{1}{2}\sin^2 t + c. \end{aligned}$$

Of course, mentally most people just “cancel” the dt terms on the right hand side of the first line and skip right to $\int u du$, which is the familiar substitution rule, in the above process.

For the second integral, integrating by parts[1] gives

$$\begin{aligned} \int \cos^2 t dt &= \cos t \sin t + c + \int \sin^2 t dt \\ &= \cos t \sin t + c \int \left(1 - \cos^2 t\right) dt \\ &= \cos t \sin t + c + t - \int \cos^2 t dt, \end{aligned}$$

hence

$$\int \cos^2 t dt = \frac{t}{2} + \frac{\cos t \sin t}{2} + c$$

◇

So, returning to the example, the general solution is

$$\begin{aligned} x(t) &= c_1 \cos t + c_2 \sin t - \frac{1}{2}\cos t \sin^2 t + \frac{1}{2} t \sin t + \frac{1}{2}\sin^2 t \cos t \\ &= c_1 \cos t + c_2 \sin t + \frac{1}{2} t \sin t. \end{aligned}$$

Applying the initial conditions

$$\begin{aligned} x(0) = 0 &\iff c_1 = 0 \\ \dot{x}(0) = 0 &\iff c_2 = 0. \end{aligned}$$

[1] To remember integration by parts, simply integrate the product rule, *i.e.,*

$$\int \left(\frac{d}{dt} uv\right) dt = \int \left(u\frac{dv}{dt} + v\frac{du}{dt}\right) dt$$

which, following the substitution rules gives the usual formula

$$\int u dv = uv - \int v du.$$

Hence

$$x(t) = \frac{1}{2} t \sin t,$$

which, thankfully, is the same answer as before. ■

3.5 Stability

3.6 Summary

1. For ordinary, second order, linear, constant coefficient, homogeneous differential equations, solutions are of the form $e^{\lambda t}$. Substituting this into the differential equation gives the characteristic equation, which will have either distinct and real roots, a pair of complex conjugate roots or repeated roots. When the equation is in canonical form

   $$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 0$$

   the roots are

   $$\begin{aligned} \lambda_1 &= -\zeta\omega_n + \omega_n\sqrt{\zeta^2 - 1} \\ \lambda_2 &= -\zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1}. \end{aligned}$$

   (a) If the roots are real and distinct, then $\zeta > 1$ and the general solution is

   $$x(t) = c_1 e^{\left(-\zeta\omega_n + \omega_n\sqrt{\zeta^1 - 1}\right)t} + c_2 e^{\left(-\zeta\omega_n - \omega_n\sqrt{\zeta^1 - 1}\right)t}.$$

   (b) If the roots are a complex conjugate pair, then $0 < \zeta < 1$ and the general solution is

   $$x_1(t) = c_1 e^{-\zeta\omega_n t}\cos\omega_d t + c_2 e^{-\zeta\omega_n t}\sin\omega_d t,$$

   where $\omega_d = \omega_n\sqrt{1 - \zeta^2}$.

   (c) If the roots are repeated, then $\zeta = 1$ and the general solution is

   $$x(t) = c_1 e^{-\omega_n t} + c_2 t e^{-\omega_n t}.$$

2. For ordinary, second order, linear, constant coefficient, inhomogeneous differential equations use

   (a) undetermined coefficients if the inhomogeneous term is sums or products of polynomials, sines, cosines or exponentials; or,

   (b) variation of parameters if the inhomogeneous term is not of that form.

   Variation of parameters works for any form of inhomogeneous term, but is generally more difficult than undetermined coefficients. For both methods, the homogeneous solution is also needed.

3. For ordinary, second order, linear, variable coefficient, inhomogeneous differential equations, the method of variation of parameters works. However, two linearly independent homogeneous solutions are required for the method, and at least at this point, you do not have any method to find them!

3.7 Exercises

Problem 3.1 Determine the solution to

$$\begin{aligned}\ddot{x}+4x &= t^2+3e^t\\ x(0) &= 0\\ \dot{x}(0) &= 0.\end{aligned}$$

Problem 3.2 Determine the general solution to

$$\begin{aligned}\frac{x}{t}+6t+(\ln t-2)\,\dot{x} &= 0\\ t &> 0.\end{aligned}$$

Problem 3.3 Determine the solution to

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

Problem 3.4 Determine the solution to

$$\begin{aligned}6\ddot{x}-5\dot{x}+x &= 0\\ x(0) &= 4\\ \dot{x}(0) &= 0.\end{aligned}$$

Problem 3.5 Determine the general solution to

$$\ddot{x}-2t\dot{x}+x=\sec t.$$

Problem 3.6 Determine the solution to

$$\begin{aligned}\ddot{x}+4\dot{x}+5x &= 0\\ x(0) &= 1\\ \dot{x}(0) &= 0.\end{aligned}$$

Problem 3.7 Determine the solution to

$$\begin{aligned}\frac{\dot{x}}{x^2} &= 1-2t\\ x(0) &= -\frac{1}{6}.\end{aligned}$$

Problem 3.8 Determine the general solution to

$$25\ddot{x} - 20\dot{x} + 4x = 0.$$

Problem 3.9 Assume that $x_1(t)$ and $x_2(t)$ are (individually) solutions to the following ordinary, second order differential equations. For which of the following is the linear combination

$$x(t) = c_1 x_1(t) + c_2 x_2(t)$$

also a solution?

1. $$\ddot{x} + 5\dot{x} + 4x = 0,$$
2. $$\ddot{x} + \sin t\dot{x} + 4x = 0,$$
3. $$\ddot{x} + 4\dot{x}x = 0,$$
4. $$\ddot{x} + 5\dot{x} + 4x = t.$$

What are the differences between the equations for which $x(t)$ is a solution and $x(t)$ is not a solution?

Problem 3.10 Prove the following theorem regarding the principle of superposition for ordinary, linear, nth order, homogeneous differential equations.

Theorem 3.7.1 *Let the functions $x_1(t), \ldots, x_n(t)$ each satisfy the ordinary, nth order, linear, homogeneous differential equation*

$$f_n(t)\frac{d^n x}{dt^n} + f_{n-1}(t)\frac{d^{n-1}x}{dt^{n-1}} + \cdots + f_1(t)\frac{dx}{dt} + f_0(t)x = 0. \tag{3.18}$$

Then any linear combination of $x_1(t), \ldots, x_n(t)$, i.e.,

$$x(t) = c_1 x_1(t) + \cdots + c_n x_n(t),$$

also satisfies equation 3.18.

Problem 3.11 In the case of repeated roots of the characteristic equation

$$\lambda^2 + 2\zeta\omega_n\zeta\lambda + \omega_n^2 = 0,$$

prove that if the roots are repeated,

1. the value of the root is $\lambda = -\omega_n$; and,

2. the two solutions

$$\begin{aligned} x_1(t) &= e^{-\omega t} \\ x_2(t) &= te^{-\omega t} \end{aligned}$$

   are linearly independent.

Problem 3.12 Table 3.1 contains 27 differential equations. Tables 3.2 and 3.3 each contain plots, each of which illustrates three solutions. Match each equation with the corresponding plot. It is possible to do this by solving only six equations! Write your answers in the form of "The solution to Equation 3 C is plot x in Figure 5 B because..."

Problem 3.13 For each of the second order differential equations listed in Problem 1.1, determine which, if any, of the following solution methods apply based upon what has been covered in this book so far.

1. Assuming exponential solutions

2. Undetermined coefficients.

3. Variation of parameters.

4. Using the fact that the equation is separable.

5. Using the fact that the equation is exact.

6. Determining an approximate numerical solution.

It may be the case that no method, one method or more than one method may apply.

Problem 3.14 Plot the solution to

$$\begin{aligned} \ddot{x} + \dot{x} + x &= \cos 2.8t + \cos 3.0t \\ x(0) &= 2 \\ \dot{x}(0) &= 5. \end{aligned}$$

You may use any method you want, including writing a computer program to determine an approximate numerical solution, but plot the whole solution, not just the steady state solution. Explain the various features of the problem, namely

1. what is happening between 0 and 10 seconds; and

2. what is happening between 10 and 60 seconds.

	A	B	C
1	$\ddot{x} + 8\dot{x} + 4x = \sin t$ $x(0) = 1$ $\dot{x}(0) = 0$	$\dot{x} = -5x$ $x(0) = 1$	$\ddot{x} + \dot{x} + 4x = 0$ $x(0) = 1$ $\dot{x}(0) = 0$
2	$\dot{x} + 3x = 1$ $x(0) = 1$	$\ddot{x} + x = 0$ $x(0) = 1$ $\dot{x}(0) = 0$	$\ddot{x} + \frac{1}{2}\dot{x} + 4x = 0$ $x(0) = 1$ $\dot{x}(0) = 0$
3	$x\dot{x}e^{2t} - t = 0$ $x(0) = 1$	$\ddot{x} + 8\dot{x} + 4x = \sin 2t$ $x(0) = 1$ $\dot{x}(0) = 0$	$\dot{x} + x = 1$ $x(0) = 1$
4	$\dot{x} - 0.1x = 0$ $x(0) = 1$	$\dot{x} = -5x + 1$ $x(0) = 1$	$\dot{x} + (t - 0.1)x = 0$ $x(0) = 1$
5	$\ddot{x} + 4x = \sin t$ $x(0) = 1$ $\dot{x}(0) = 0$	$\dot{x} + (t - 1)x = 0$ $x(0) = 1$	$x\dot{x}e^{3t} - t = 0$ $x(0) = 1$
6	$\dot{x} = 0.5x$ $x(0) = 1$	$\ddot{x} + 4x = \sin 2t$ $x(0) = 1$ $\dot{x}(0) = 0$	$\ddot{x} + 3x = 0$ $x(0) = 1$ $\dot{x}(0) = 0$
7	$\dot{x} - x = 0$ $x(0) = 1$	$\dot{x} + x = 0$ $x(0) = 1$	$\ddot{x} + 2\dot{x} + 4x = 0$ $x(0) = 1$ $\dot{x}(0) = 0$
8	$x\dot{x}e^{t} - t = 0$ $x(0) = 1$	$\ddot{x} + 4x = \sin 1.9t$ $x(0) = 1$ $\dot{x}(0) = 0$	$\ddot{x} + 2x = 0$ $x(0) = 1$ $\dot{x}(0) = 0$
9	$\dot{x} + (t - 0.5)x = 0$ $x(0) = 1$	$\dot{x} + 3x = 0$ $x(0) = 1$	$\ddot{x} + 0.2\dot{x} - x + x^3$ $= 0.3\sin t$ $x(0) = 1$ $\dot{x}(0) = 0$

Table 3.1. Differential equations for Problem 3.12.

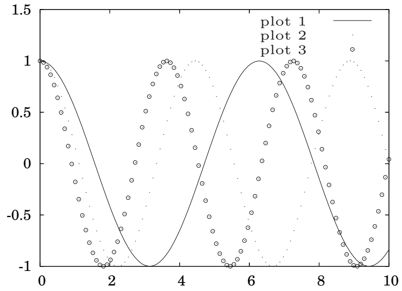
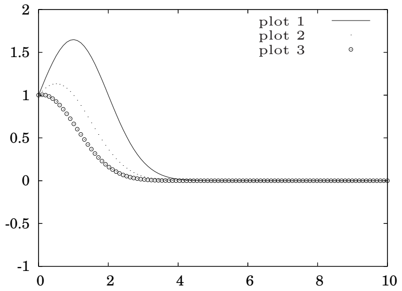
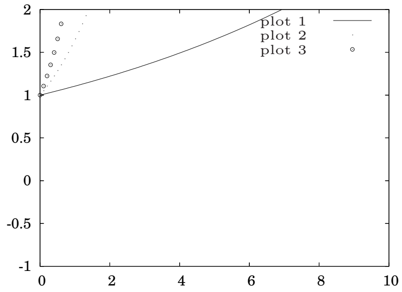
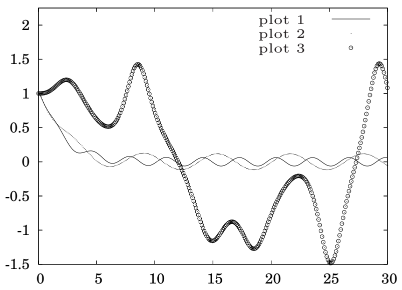
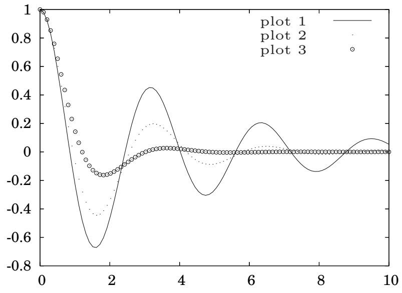
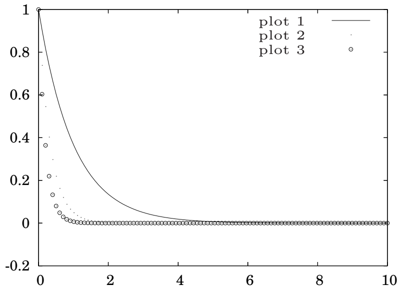
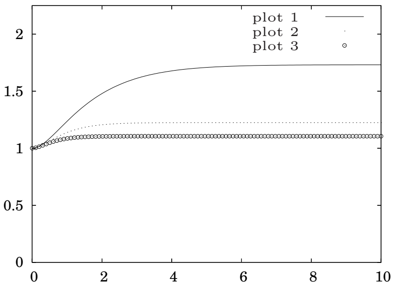
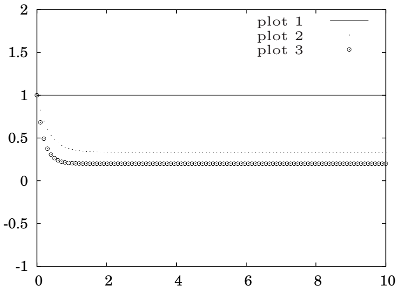
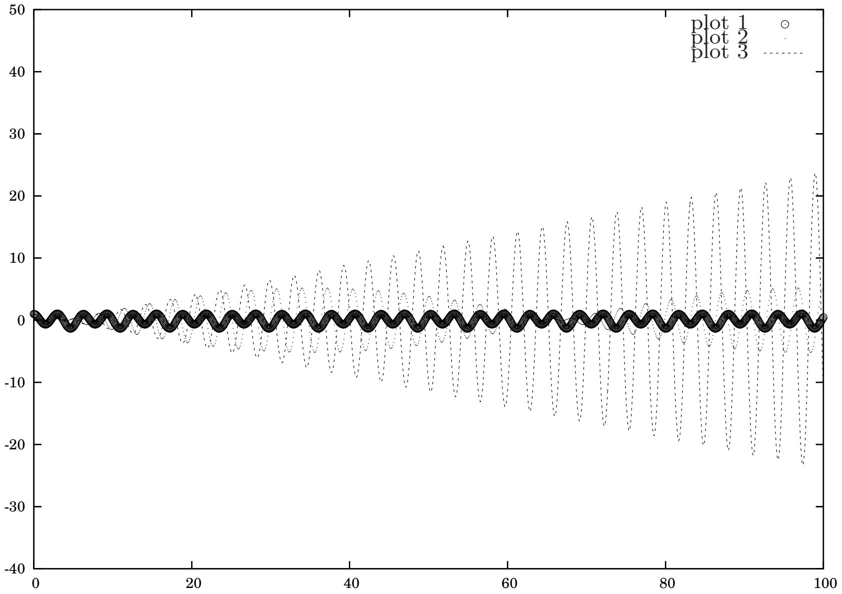
	A	B
1		
2		
3		

Table 3.2. Solution graphs for Problem 3.12.

<table>
<tr><th></th><th>A</th><th>B</th></tr>
<tr><td>4</td><td>


</td><td>


</td></tr>
<tr><th></th><th colspan="2">A</th></tr>
<tr><td>5</td><td colspan="2">


</td></tr>
</table>

Table 3.3. Solution graphs for Problem 3.12.

Chapter 4

Single Degree of Freedom Vibrations

This chapter presents applications of second order, ordinary, constant coefficient differential equations. The primary applications in mechanical engineering and related fields is that of vibrations analysis. Additionally, since a second order system is a canonical system for the design of some feedback controllers, a review of the response characteristics of second order systems to step inputs is included.

The study of single degree of freedom vibrations considers the analysis of problems of the type illustrated in Figure 4.1 and described by

$$m\ddot{x} + b\dot{x} + kx = f(t), \tag{4.1}$$

where $f(t)$ is an applied force. The term "single" refers to the fact that they system has only one degree of freedom. This is in contrast with multiple degree of freedom system, an example of which is illustrated in Figure 6.1. This type of problem is generally categorized according to whether it is

- *free* or *forced*; or,
- *damped* or *undamped*.

Sections 4.1 through 4.4 consider each of the four possible permutations of these cases.

This chapter is a complete study and analysis of the solutions to

$$m\ddot{x} + b\dot{x} + kx = f(t) \quad \Longleftrightarrow \quad \ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = \frac{f(t)}{m}. \tag{4.2}$$

The system is *free* if it is unforced, *i.e.,* $f(t) = 0$; otherwise it is *forced*. The system is *undamped* if $b = 0$ (equivalently $\zeta = 0$); otherwise, it is damped. While somewhat scattered throughout the example problems in Chapter 3, the quantities of major importance in this chapter that have already been introduced include the following:

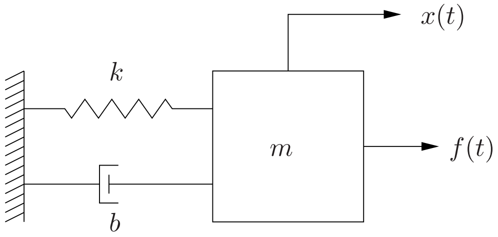


Figure 4.1. Mechanical system described by equation 4.1.

1. the *natural frequency*: $\omega_n = \sqrt{\frac{k}{m}} > 0$;
2. the *damping ratio*: $\zeta = \frac{b}{2\sqrt{km}} > 0$;
3. the *damped natural frequency*: $\omega_d = \omega_n \sqrt{1-\zeta^2}$ (only relevant for $0 < \zeta < 1$).

4.1 Free, undamped oscillations

This problem has been completely solved in section 3.4.1 and particularly in example 3.4.1. The results will be summarized here, so only the results and an analysis and interpretation of the results will be presented here. Free and undamped implies that in Figure 4.1 $b = 0$ and $f(t) = 0$, or equivalently, that the system is as illustrated in Figure 4.2, so the equation of motion reduces to

$$m\ddot{x} + kx = 0 \qquad \Longleftrightarrow \qquad \ddot{x} + \omega_n^2 x = 0,$$

which, as presented previously, has a general solution

$$x(t) = c_1 \cos \omega_n t + c_2 \sin \omega_n t.$$

If the initial conditions are specified as

$$\begin{aligned} x(0) &= x_0 \\ \dot{x}(0) &= \dot{x}_0, \end{aligned}$$

then the solution is

$$x(t) = x_0 \cos \omega_n t + \frac{\dot{x}_0}{\omega_n} \sin \omega_n t. \tag{4.3}$$

While this equation is relatively simple to interpret and plot, it can be made even simpler to analyze if the sine and cosine terms are combined. In particular, equate the solution with a single, phase shifted cosine function

$$\begin{aligned} x_0 \cos \omega_n t + \frac{\dot{x}_0}{\omega_n} \sin \omega_n t &= c \cos(\omega_n t + \phi) \\ &= c\left(\cos\phi \cos \omega_n t - \sin\phi \sin \omega_n t\right), \end{aligned}$$

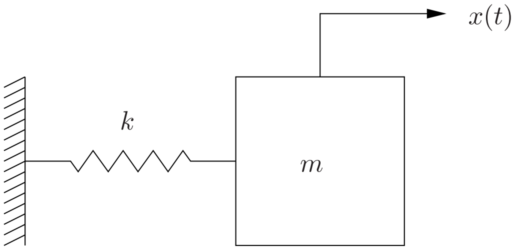


Figure 4.2. Mechanical system with solution described by equation 4.4.

and solving for c and ϕ by equating the coefficients of the $\cos \omega_n t$ and $\sin \omega_n t$ terms gives

$$\begin{aligned} c &= \sqrt{x_0^2 + \left(\frac{\dot{x}_0}{\omega_n}\right)^2} \\ \phi &= \tan^{-1}\left(-\frac{\dot{x}_0}{\omega_n x_0}\right), \end{aligned}$$

so an equivalent representation of the solution is

$$\begin{aligned} x(t) &= x_0 \cos \omega_n t + \frac{\dot{x}_0}{\omega_n} \sin \omega_n t \\ &= \sqrt{x_0^2 + \left(\frac{\dot{x}_0}{\omega_n}\right)^2} \cos\left(\omega_n t + \phi\right). \end{aligned} \tag{4.4}$$

A numerical example perhaps may be enlightening.

Example 4.1.1 Figure 4.3 is a plot of the solution to

$$\begin{aligned} \ddot{x} + \omega_n x &= 0 \\ x(0) &= 1 \\ \dot{x}(0) &= 1 \end{aligned}$$

for $\omega_n = 1, 2, 3, 4$ and 5.

As is obvious from the form of the solution in example 4.1.1, the solution is a cosine with a constant amplitude. As the natural frequency increases, the frequency of the response increases. Also due to the $\frac{\dot{x}_0}{\omega_n}$ term in the amplitude of the response, as ω_n increases, the amplitude of the response decreases. ■

From the example and an analysis of the form of the solution in equation 4.4, one may conclude the following regarding the response of an undamped, free, single degree of freedom system.

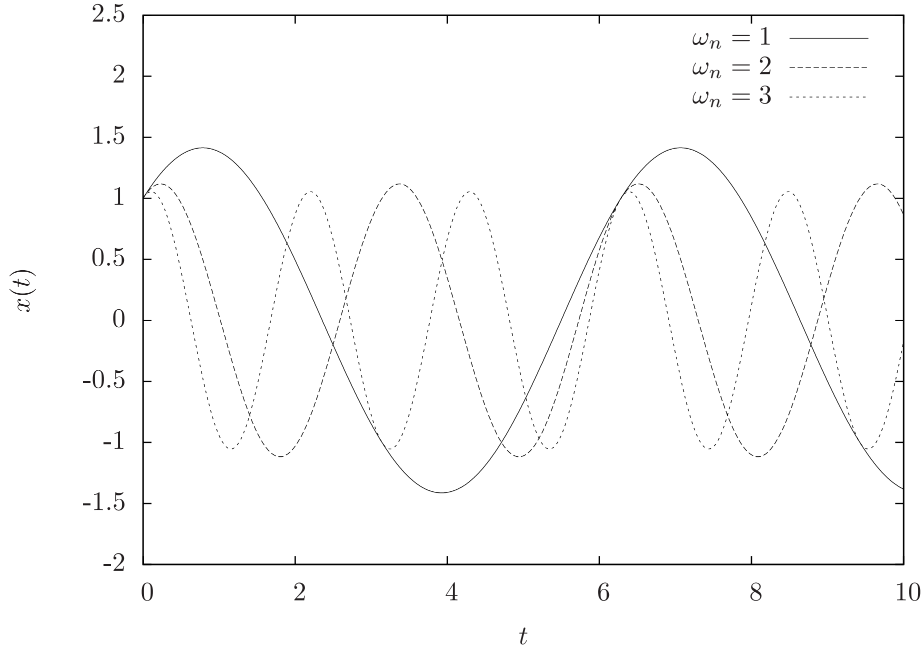


Figure 4.3. Solutions to system in Example 4.1.1.

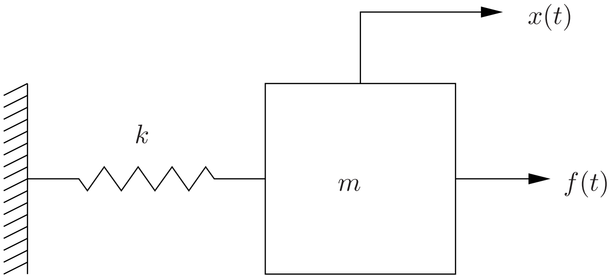


Figure 4.4. Single degree of freedom, undamped, forced oscillator.

1. If ω_n increases, the frequency of the response will increase.
2. If k increases, the frequency of the response will increase.
3. If m increases, the frequency of the response will decrease.
4. If $|x_0|$ increases, the magnitude of the response will increase.
5. If $|\dot{x}_0|$ increases, the magnitude of the response will increase.
6. If $\dot{x}_0 \neq 0$ and ω_n increases, the magnitude of the response will decrease.

4.2 Harmonically Forced, undamped vibrations

Now the problem considered in the previous section will be modified to add a forcing function acting on the mass as is illustrated in Figure 4.4. The most common scenario is the case when the forcing function, $f(t)$ is a harmonic function, *i.e.,* sines, cosines or combinations thereof.

Consider the case when $f(t) = F\cos\omega t$, *i.e.,* a harmonic function of magnitude F and frequency ω. Note that there are now multiple frequencies; namely, the natural frequency, $\omega_n = \sqrt{\frac{k}{m}}$ and the frequency of the forcing function, ω. In general, they are not the same and care must be taken to observe the subscript or absence thereof. The equation of motion for this system is

$$m\ddot{x} + kx = F\cos\omega t \qquad \Longleftrightarrow \qquad \ddot{x} + \omega_n^2 x = \frac{F}{m}\cos\omega t. \tag{4.5}$$

Clearly, this is an ordinary, second order, constant coefficient, linear, inhomogeneous differential equation; furthermore, due to the form of the inhomogeneous term, the method of undetermined coefficients is probably most expedient solution method. From section 4.1, the homogeneous solution is

$$x_h(t) = c_1\cos\omega_n t + c_2\sin\omega_n t.$$

Hence, for undetermined coefficients, assume

$$x_p(t) = A\cos\omega t + B\sin\omega t,$$

as long as $\omega \neq \omega_n$*!* The special case where $\omega = \omega_n$, which requires

$$x_p(t) = t\left(A\cos\omega t + B\sin\omega t\right),$$

will be considered subsequently.

Differentiating x_p and substituting gives

$$\begin{aligned} A &= \frac{F}{m\left(\omega_n^2 - \omega^2\right)} \\ B &= 0, \end{aligned}$$

so a general solution to equation 4.5 is

$$x(t) = c_1\cos\omega_n t + c_2\sin\omega_n t + \frac{F}{m\left(\omega_n^2 - \omega^2\right)}\cos\omega t. \tag{4.6}$$

If the initial conditions are specified as

$$\begin{aligned} x(0) &= x_0 \\ \dot{x}(0) &= \dot{x}_0, \end{aligned}$$

then a quick calculation gives

$$\begin{aligned} c_1 &= x_0 - \frac{F}{m\left(\omega_n^2 - \omega^2\right)} \\ c_2 &= \frac{\dot{x}_0}{\omega_n}, \end{aligned}$$

and hence the solution to the initial value problem is

$$x(t) = \left(x_0 - \frac{F}{m\left(\omega_n^2 - \omega^2\right)}\right)\cos\omega_n t + \frac{\dot{x}_0}{\omega_n}\sin\omega_n t + \frac{F}{m\left(\omega_n^2 - \omega^2\right)}\cos\omega t. \tag{4.7}$$

To put equation 4.7 into a form more amenable to analysis, recall the definition of the *static deflection* which is the amount the spring would displace under the load of a static force of magnitude F, particularly,

$$\delta = \frac{F}{k}.$$

Using this, and defining the frequency ratio as

$$r = \frac{\omega}{\omega_n}$$

a simple manipulation gives

$$x(t) = \left(x_0 - \frac{\delta}{1 - r^2}\right)\cos\omega_n t + \frac{\dot{x}}{\omega_n}\sin\omega_n t + \frac{\delta}{1 - r^2}\cos\omega t. \tag{4.8}$$

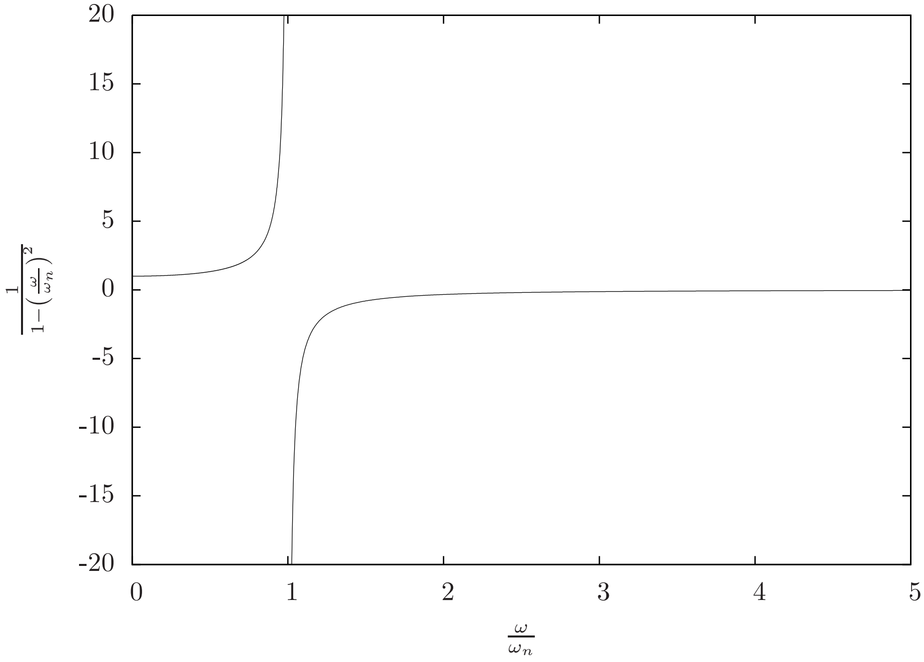


Figure 4.5. Magnification factor versus frequency ratio.

In light of this solution, the term *magnification factor* for

$$M = \frac{1}{1-\left(\frac{\omega}{\omega_n}\right)^2} = \frac{1}{1-r^2}$$

makes sense since it is the amount by which the static deflection is either amplified or attenuated in the solution. A plot of the magnification factor versus frequency ratio is illustrated in Figure 4.5. Note that the case where $r = 1$ is seemingly problematic; however, recall that is the case where $\omega = \omega_n$, which has a different solution. Also observe that for frequency ratios greater than one, the magnification ratio is negative, which represents the fact that the particular solution is out of phase with the forcing function.

Note that the solution, in equation 4.8 depends upon

1. the natural frequency, ω_n;
2. the forcing frequency, ω;
3. the static deflection, δ; and,
4. the initial conditions, x_0 and $\dot{x}_0$.

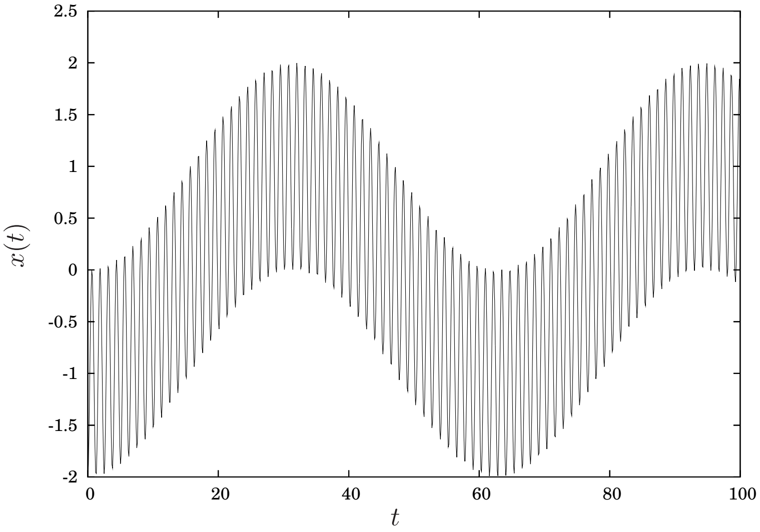


Figure 4.6. Harmonically forced, undamped solution where $\omega \ll \omega_n$.

Note, however, that the initial conditions as well as the static deflection simply scale individual terms of the solution. Therefore, the most interesting feature of the solution is its dependence on the forcing and natural frequencies, which will be explored in the following example.

Example 4.2.1 Plot the solution for

$$\begin{aligned} \ddot{x} + \omega_n^2 x &= \frac{F}{m}\cos\omega t \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

where $\omega \ll \omega_n$, *i.e.,* the forcing frequency is much smaller than the natural frequency. With zero initial conditions, the solution is

$$x(t) = \frac{\delta}{1-r^2}\left(\cos\omega t - \cos\omega_n t\right)$$

and if $\omega \ll \omega_n$, $r \approx 0$; hence,

$$x(t) \approx \delta\left(\cos\omega t - \cos\omega_n t\right).$$

Thus, the solution will vary in magnitude between 0 and 2δ depending upon whether ω and ω_n are in phase or out of phase.

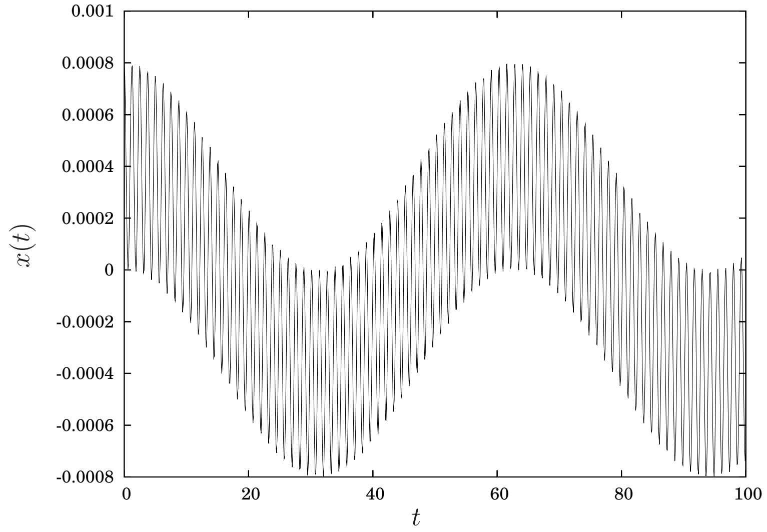


Figure 4.7. Harmonically forced, undamped solution where $\omega_n \ll \omega$.

A plot of the solution where $\delta = 1$, $\omega = 0.1$ and $\omega_n = 5$ is illustrated in Figure 4.6. Note that because the two frequencies are well separated, the solution is clearly the superposition of two cosine functions, one relatively fast and the other relatively slow. ■

Example 4.2.2 Now considering the other extreme where $\omega \gg \omega_n$, *i.e.,* the system is forced at a frequency that is much greater than the natural frequency. In this case, the frequency ratio will become very large and the coefficient of the solution

$$\frac{\delta}{1-r^2} \approx \frac{-\delta}{r^2}$$

will be very small.

A plot of the solution where $\delta = 1$, $\omega = 5$ and $\omega_n = 0.1$ is illustrated in Figure 4.7. At first glance this appears similar to the response when $\omega \ll \omega_n$; however, note the scale on the graph. The response is still the sum of two cosine functions but the magnitude of the response is much smaller than in the case where $\omega \ll \omega_n$ ■

Example 4.2.3 Yet another interesting feature of this solution is apparent when one considers the relative phase between the forcing function, $F \cos \omega t$

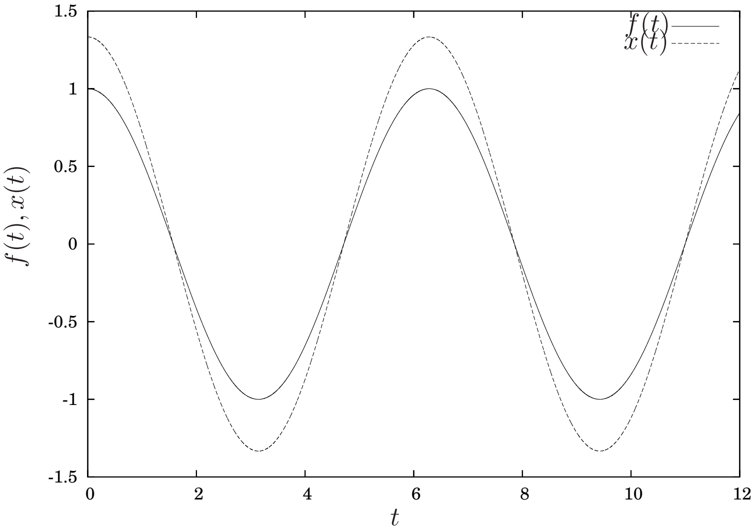


Figure 4.8. Forcing function and particular solution in phase for $\omega < \omega_n$.

and the response of the system. Just for the fun of it, let us assume that

$$\begin{aligned} x(0) &= -\frac{\delta}{1-r^2} \\ \dot{x}(0) &= 0. \end{aligned}$$

The initial conditions were picked so that the terms in the solution due to the homogeneous solutions are zero and the complete solution is the same as the particular solution; namely,

$$x(t) = \frac{\delta}{1-r^2} \cos \omega t.$$

Recall that the forcing function is

$$f(t) = F \cos \omega t.$$

The response, $x(t)$, and forcing function, $f(t)$, are plotted together for the two cases where $\omega < \omega_n$ ($r = 0.5$) and $\omega > \omega_n$ ($r = 1.5$) in Figures 4.8 and 4.9, respectively. In both figures, $\delta = 1$.

The interesting feature of these solutions is that the response of the system is in phase with the forcing function when $\omega < \omega_n$ and out of phase with the forcing function when $\omega > \omega_n$. The latter is the somewhat counter-intuitive case when the force is always directed in the opposite direction of the velocity of the mass. ■

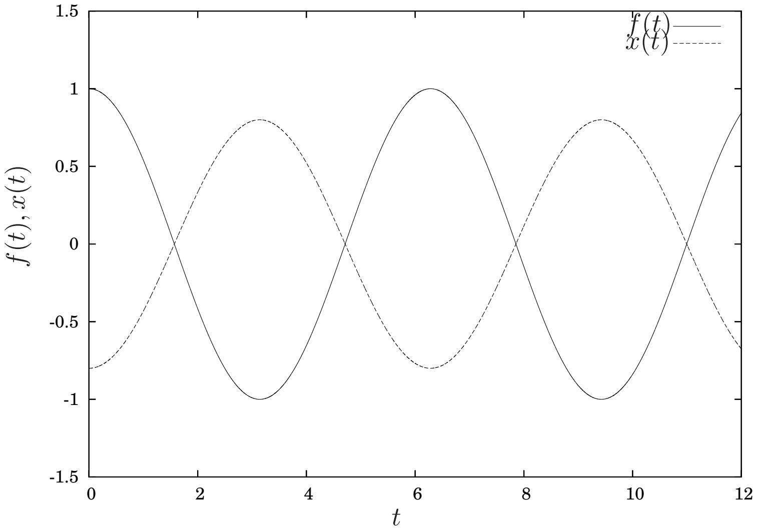


Figure 4.9. Forcing function and particular solution in phase for $\omega > \omega_n$.

4.2.1 Resonance

This section deals with the case where the forcing frequency and natural frequency are equal. This is known as *resonance* and a quick look at Figure 4.5 would give the impression that unbounded solutions are a possibility, which is the case. Additionally, resonance corresponds to the physically intuitive situation wherein a system is forced at the frequency at which it is most amenable.

For

$$\begin{aligned} \ddot{x} + \omega_n^2 x &= \frac{F}{m}\cos\omega t \qquad (4.9) \\ x(0) &= x_0 \\ \dot{x}(0) &= \dot{x}_0 \end{aligned}$$

where $\omega = \omega_n$, it is clearly the case that the assumed form of the particular solution is the same as the homogeneous solutions since

$$x_h(t) = c_1 \cos\omega_n t + c_2 \sin\omega_n t.$$

So, the correct assumption is

$$x_p(t) = t\left(A\cos\omega_n t + B\sin\omega_n t\right).$$

Skipping the mundane details of substituting and equating coefficients, the solution is

$$x(t) = x_0 \cos\omega_n t + \frac{\dot{x}_0}{\omega_n}\sin\omega_n t + \frac{\delta\omega_n t}{2}\sin\omega_n t. \qquad (4.10)$$

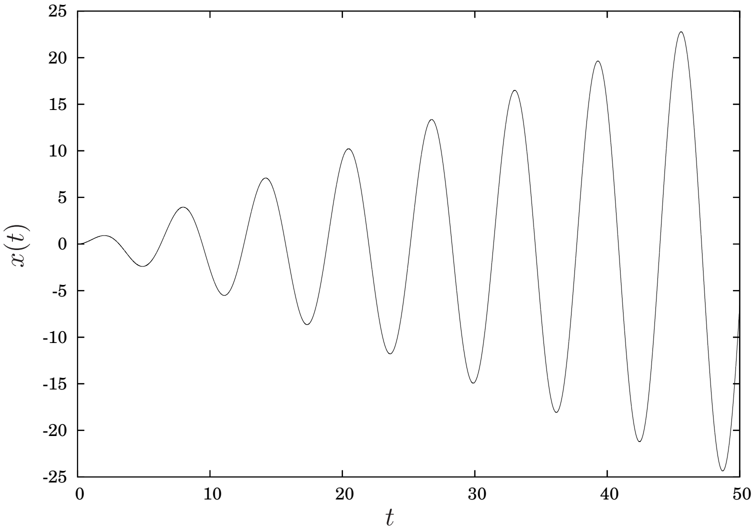


Figure 4.10. Solution for example 4.2.4.

Since the part of the solution that is the particular solution (the second $\sin\omega_n t$ term) is multiplied by t, it grows linearly in time. A specific example follows, but the general point that the solution grows with time is the fundamentally important point regarding resonance.

Example 4.2.4 Solve

$$\begin{aligned} \ddot{x}+x &= \cos t \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

and plot the solution versus time.

Since this equation is exactly of the form of equation 4.9 with $\omega_n = F = m = 1$, simply substituting those values into equation 4.10 gives the solution

$$x(t) = \frac{t}{2}\sin t,$$

which is plotted in Figure 4.10 ■

4.2.2 Near Resonance

Obviously in physical situations, it is impossible to exactly have $\omega = \omega_n$, so the question regarding the nature of the solution when $\omega \approx \omega_n$ and its relationship to the resonance solution naturally arises.

Consider

$$\begin{aligned} \ddot{x} + \omega_n^2 x &= \frac{F}{m}\cos\omega t \\ x(0) &= x_0 \\ \dot{x}(0) &= \dot{x}_0 \end{aligned} \tag{4.11}$$

where $\omega \approx \omega_n$. Since $\omega \neq \omega_n$, the solution is not from equation 4.10 but rather is from equation 4.7,

$$x(t) = \left(x_0 - \frac{F}{m\left(\omega_n^2 - \omega^2\right)} \right) \cos\omega_n t + \frac{\dot{x}_0}{\omega_n}\sin\omega_n t + \frac{F}{m\left(\omega_n^2 - \omega^2\right)}\cos\omega t.$$

If $\omega \approx \omega_n$, then the two coefficients with $(\omega_n - \omega)$ in the denominator will be very large. Rewriting the solution by grouping those two terms gives

$$x(t) = x_0\cos\omega_n t + \frac{\dot{x}_0}{\omega_n}\sin\omega_n t + \frac{F}{m\left(\omega_n^2 - \omega^2\right)}\left(\cos\omega t - \cos\omega_n t\right). \tag{4.12}$$

Note that the terms in this solution and in the resonance solution in equation 4.10 that depend on the initial conditions are identical. In the resonance case, the solution grows large because of the t term multiplying the $\sin\omega t$ function in the solution. In the near resonance case, the solution grows large because of the large coefficient, and the "growth" of the solution comes about because of the $\cos\omega t$ and $\cos\omega_n t$ terms shifting out of phase as t increases. To illustrate this fact, consider the following example.

Example 4.2.5 Solve

$$\begin{aligned} \ddot{x} + x &= \cos 1.05t \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

and plot the solution versus time.

The solution is of the form of equation 4.12 with $\omega_n = 1$, $\frac{F}{m} = 1$ and $\omega = 1.05$, and is given by substitution into equation 4.12

$$x(t) = \frac{1}{\left(1 - (1.05)^2\right)}\left(\cos 1.05t - \cos t\right).$$

A plot of this solution for $0 < t < 50$ is illustrated in Figure 4.11. Note that, at least to the extent possible by casual observation, it appears to be the same as the solution illustrated for resonance in Figure 4.10.

Plotting the solution for a longer period of time, $0 < t < 500$, as is illustrated in Figure 4.12, highlights the main difference. Since the solution grows because the cosine terms slowly go out of phase as time increases, they eventually must go back in phase, resulting in a decrease in magnitude of the solution. ■

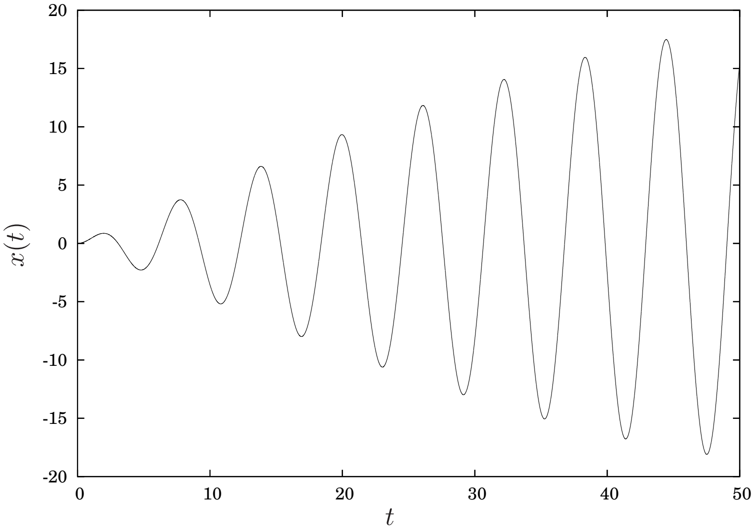


Figure 4.11. Solution for example 4.2.5 for $0 < t < 50$.

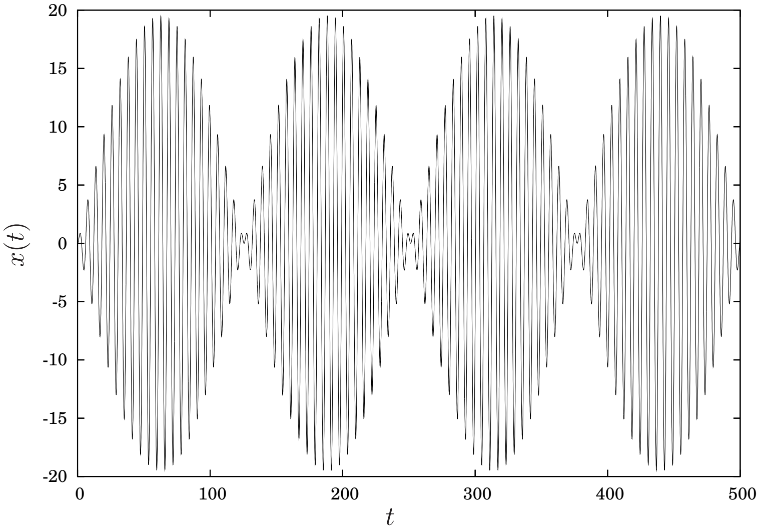


Figure 4.12. Solution for example 4.2.5 for $0 < t < 500$.

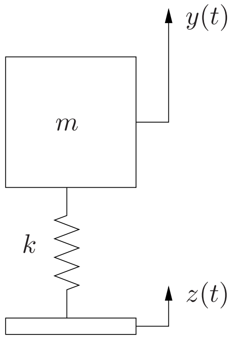


Figure 4.13. Undamped vibrating base system.

4.2.3 Vibrating Base

Now consider the problem of a coupled by a spring to a vibrating base, as is illustrated in Figure 4.13. In this problem the base of the system, illustrated by the thin bar, moves with a prescribed motion, $z(t)$. The focus of the analysis is on the resulting motion of the mass, $y(t)$ with particular emphasis on the dependence of this motion on the system parameters, m and k as well as the nature of the base motion, $z(t)$.

Using Newton's law, the equation of motion for this system is

$$m\ddot{y} + ky = kz(t).$$

or

$$\ddot{y} + \omega_n^2 y = \omega_n^2 z(t).$$

Thus, the only variables of concern in the problem is the natural frequency and the nature of $z(t)$. For simplicity, assume that $z(t)$ is harmonic, particularly, $z(t) = Z\cos\omega t$, so that

$$\ddot{y} + \omega_n^2 y = Z\omega_n^2 \cos\omega t.$$

Clearly, the homogeneous solution is

$$x_h(t) = c_1 \cos\omega_n t + c_2 \sin\omega_n t.$$

Assuming that $\omega \neq \omega_n$ and

$$x_p(t) = A\cos\omega t + B\sin\omega t$$

substituting and equating coefficients gives

$$x_p(t) = \frac{Z\omega_n^2}{\omega_n^2 - \omega^2}\cos\omega t$$

so that

$$x(t) = c_1 \cos \omega_n t + c_2 \sin \omega_n t + \frac{Z}{1-r^2} \cos \omega t,$$

where, as before,

$$r = \frac{\omega}{\omega_n}.$$

Since the coefficient of the part of the solution which is the particular solution is exactly of the form of equation 4.8, the analysis is exactly the same as the undamped, forced oscillation case, but where the static deflection is replaced by the magnitude of the base motion. In this case, the magnification factor,

$$M = \frac{1}{1-r^2}$$

has an even more direct interpretation in that it is the magnification of the base motion in the response of the mass motion. Referring back to Figure 4.5, M has a value of 1 at $r = 0$, increases to an unbounded value at $r = 1$, decreases to $M = 1$ at $r = \sqrt(2)$ and asymptotically approaches zero as r gets large. Hence, the resonance analysis is similar to that of the simple forced case. Additionally, the smallest magnification occurs at very high frequencies.

4.3 Free, damped vibrations

This section considers the case of damped oscillations with no forcing function, *i.e.,* the solution to

$$\begin{aligned} \ddot{x} + 2\zeta\omega_n \dot{x} + \omega_n^2 x &= 0 && (4.13) \\ x(0) &= x_0 && (4.14) \\ \dot{x}(0) &= \dot{x}_0. && (4.15) \end{aligned}$$

Since this is a constant coefficient, linear, homogeneous, second order ordinary differential equation, it has exponential solutions. The resulting characteristic equation is

$$\lambda^2 + 2\zeta\omega_n\lambda + \omega_n^2 = 0$$

with roots

$$\lambda = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}.$$

The nature of the solution will clearly depend upon whether ζ is less than one, equal to one or greater than one.

4.3.1 Damping ratio greater than one

In this case, the solution is

$$x(t) = c_1 e^{\left(-\zeta\omega_n + \omega_n\sqrt{\zeta^2-1}\right)t} + c_2 e^{\left(-\zeta\omega_n - \omega_n\sqrt{\zeta^2-1}\right)t}$$

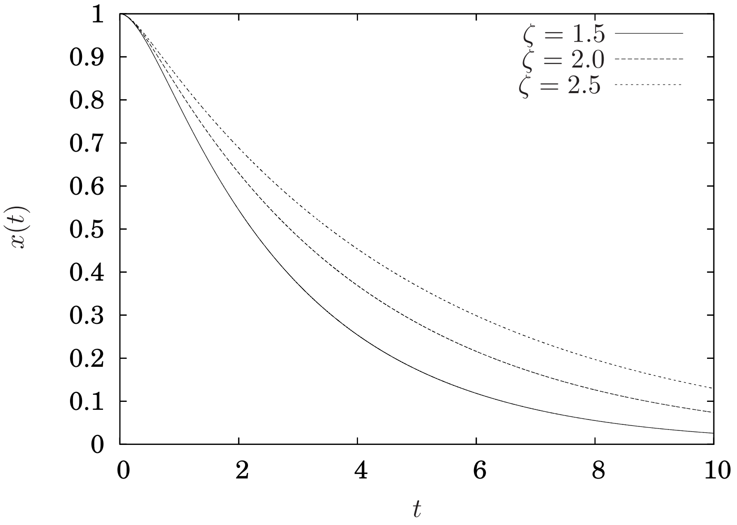


Figure 4.14. Solution of second order system for various values of ζ.

and evaluating the initial conditions gives

$$x(0) = c_1 + c_2 = x_0$$

and

$$\dot{x}(0) = \left(-\zeta\omega_n + \omega_n\sqrt{\zeta^2 - 1}\right)c_1 + \left(-\zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1}\right)c_2 = \dot{x}_0$$

which gives

$$\begin{aligned} x(t) &= \frac{\dot{x}_0 + x_0\left(\zeta\omega_n + \sqrt{\zeta^2 - 1}\right)}{2\sqrt{\zeta^2 - 1}} e^{\left(-\zeta\omega_n + \omega_n\sqrt{\zeta^2-1}\right)t} \\ &- \frac{\dot{x}_0 + \zeta\omega_n x_0 - x_0\sqrt{\zeta^2 - 1}}{2\sqrt{\zeta^2 - 1}} e^{\left(-\zeta\omega_n - \omega_n\sqrt{\zeta^2-1}\right)t}. \end{aligned}$$

Figure 4.14 illustrates the response for $\omega_n = 1$, $x(0) = 1$ and $\dot{x}(0) = 0$ for various values of ζ.

4.3.2 Damping ratio equal to one

When the damping ratio is equal to one there are repeated roots of the characteristic equation

$$\lambda = -\omega_n$$

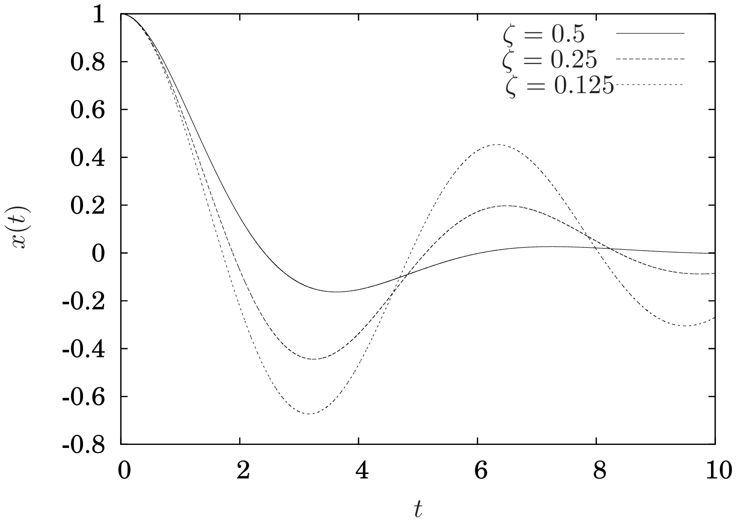


Figure 4.15. Solution of second order system for various values of ζ.

so the general solution to the homogeneous equation is

$$x(t) = c_1 e^{-\omega_n t} + c_2 t e^{-\omega_n t}.$$

4.3.3 Damping ratio less than one

When the damping ratio is less than one, the characteristic equation has complex roots

$$\lambda = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1} = -\zeta\omega_n \pm i\omega_n\sqrt{1-\zeta^2}$$

so the general solution to the differential equation is

$$x(t) = e^{-\zeta\omega_n t}\left(c_1 \cos\omega_n\sqrt{1-\zeta^2}t + c_2 \sin\omega_n\sqrt{1-\zeta^2}t\right). \tag{4.16}$$

Figure 4.15 illustrates the response for $\omega_n = 1$, $x(0) = 1$ and $\dot{x}(0) = 0$ for various values of ζ.

4.4 Harmonically forced, damped vibrations

In this section we consider the system illustrated in Figure 4.1 with the equation of motion given by Equation 4.2 where the applied force, $f(t)$ is assumed to be *harmonic*, *i.e.,* it is some combination of sine and cosine functions. For the rest of this section we will assume the forcing function is of the form

$$f(t) = F\cos\omega t.$$

Confirming the details that the solution for a sine function of a combination of sines and cosines is left as an exercise.

For the system

$$\ddot{x} + 2\zeta\omega_n x + \omega_n^2 x = \frac{F}{m}\cos\omega t \tag{4.17}$$

Section 4.3 provides all the possible cases for the homogeneous solution. Since the inhomogeneous term is of the class of function for which the method of undetermined coefficients is appropriate, we can choose to use either undetermined coefficients from Section 3.4.1 or variation of parameters from Section 3.4.2. Because it is more transparent, we will use undetermined coefficients.

Assuming a particular solution of the form

$$\ddot{x}_p(t) = A\cos\omega t + B\sin\omega t$$

gives

$$\begin{array}{rcl}\dot{x}_p & = & -A\omega\sin\omega t + B\omega\cos\omega t\\ \ddot{x}_p & = & -A\omega^2\cos\omega t - B\omega^2\sin\omega t\end{array}$$

and substituting into Equation 4.17 gives

$$\begin{aligned}&\left(-A\omega^2\cos\omega t - B\omega^2\sin\omega t\right) + \\ &\qquad 2\zeta\omega_n\left(-A\omega\sin\omega t + B\omega\cos\omega t\right) + \omega_n^2\left(A\cos\omega t + B\sin\omega t\right) = \frac{F}{m}\cos\omega t.\end{aligned}$$

A bit of algebra gives A and B so that

$$x_p(t) = \frac{F}{m}\left(\frac{\omega_n^2 - \omega^2}{\left(\omega_n^2 - \omega^2\right)^2 + \left(2\zeta\omega\omega_n\right)^2}\cos\omega t + \frac{2\zeta\omega\omega_n}{\left(\omega_n^2 - \omega^2\right)^2 + \left(2\zeta\omega\omega_n\right)^2}\sin\omega t\right). \tag{4.18}$$

Observe that as long as $\zeta \neq 0$ the solution given by Equation 4.18 is correct, even in the case of resonance when $\omega = \omega_n$. Furthermore, when $\zeta = 0$ this reduces to the undamped forced solution as long as $\omega \neq \omega_n$.

Since the solution in Equation 4.18 is in the form of a linear combination of a cosine and sine function an analysis of the effect of the nature of the response as a function of the forcing frequency, ω and the damping ratio, ζ is not straightforward. Hence, we will convert the solution to the form of a single trigonometric function with a phase shift, *e.g.*,

$$x_p(t) = c\cos\left(\omega t + \phi\right)$$

where the magnitude of the response, c and the phase shift, ϕ must be determined. Note that, reverting back to the coefficients A and B

$$\begin{aligned} x_p(t) &= A\cos\omega t + B\sin\omega t \\ &= \sqrt{A^2+B^2}\left(\frac{A}{\sqrt{A^2+B^2}}\cos\omega t + \frac{B}{\sqrt{A^2+B^2}}\sin\omega t\right). \end{aligned}$$

Since the coefficients of the sine and cosine terms must have values in the interval $[-1, 1]$, we can write this as

$$\begin{aligned} x_p(t) &= c\left(\cos\phi\cos\omega t - \sin\phi\sin\omega t\right) \\ &= c\cos\left(\omega t + \phi\right), \end{aligned}$$

where

$$\begin{aligned} c &= \sqrt{A^2+B^2} \\ \phi &= \tan^{-1}\left(-\frac{B}{A}\right), \end{aligned}$$

or, using the actual expressions for A and B

$$\begin{aligned} c &= \frac{F}{m}\sqrt{\frac{1}{\left(\omega_n^2-\omega^2\right)^2 + \left(2\zeta\omega_n\omega\right)^2}} \\ \phi &= \tan^{-1}\left(-\frac{2\zeta\omega\omega_n}{\omega_n^2-\omega^2}\right). \end{aligned}$$

So, that bit of work resulted in

$$x_p(t) = \frac{F}{m}\sqrt{\frac{1}{\left(\omega_n^2-\omega^2\right)^2 + \left(2\zeta\omega_n\omega\right)^2}}\cos\left(\omega t+\phi\right)$$

where ϕ is given as above.

A final step, that may not be obvious *a priori* is to factor an ω_n^2 out of the denominator of c, which gives

$$x_p(t) = \frac{F}{\omega_n^2 m}\sqrt{\frac{1}{\left(1-\frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2}}\cos\left(\omega t+\phi\right).$$

Note that

$$\frac{F}{\omega_n^2 m} = \frac{F}{k}$$

from the definition of the natural frequency and that furthermore, the quantity $\frac{F}{k}$ is the amount that the spring would be deflected under a static force of magnitude F. Hence we will call $\delta = \frac{F}{k}$ the *static deflection* and (finally) write

$$x_p(t) = \delta\sqrt{\frac{1}{\left(1-\frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2}}\cos\left(\omega t+\phi\right)$$

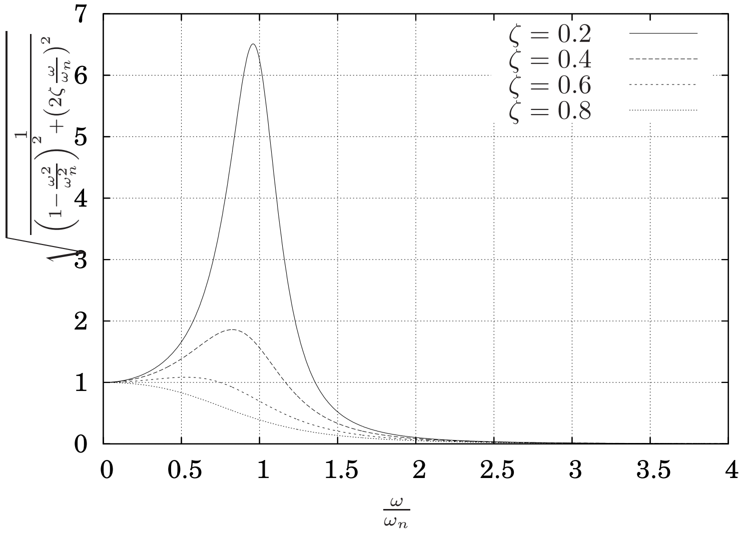


Figure 4.16. Magnification of static deflection for various damping ratios *versus* frequency ratio.

which is a complete description of the response in terms of the static deflection, the ratio of the forcing frequency to the natural frequency and the damping ratio. The quantity

$$M = \sqrt{\frac{1}{\left(1-\frac{\omega^2}{\omega_n^2}\right)^2+\left(2\zeta\frac{\omega}{\omega_n}\right)^2}}$$

can be interpreted to represent the amount that the static deflection is amplified or attenuated in the response of the system due to the frequency of the forcing function. The phase shift can similarly be expressed in terms of the frequency ratio by dividing the numerator and denominator by ω_n^2 giving

$$\phi = \tan^{-1}\left(-\frac{2\zeta\frac{\omega}{\omega_n}}{1-\frac{\omega^2}{\omega_n^2}}\right).$$

We may gain insight into the nature of the response by considering the nature of the dependence of M and ϕ on the frequency ratio $\frac{\omega}{\omega_n}$ and the damping ratio, ζ. A plot of M as a function of the frequency ratio for different damping ratios is illustrated in Figure 4.16 and a plot of the phase shift, ϕ as a function of the frequency ratio for different damping ratios is illustrated in Figure 4.17.

Up to now this section has considered only the *particular* solution to Equation 4.17. However, as long as the damping ratio is not zero, the homogeneous solution will decay and the above analysis of M and ϕ are appropriate to con-

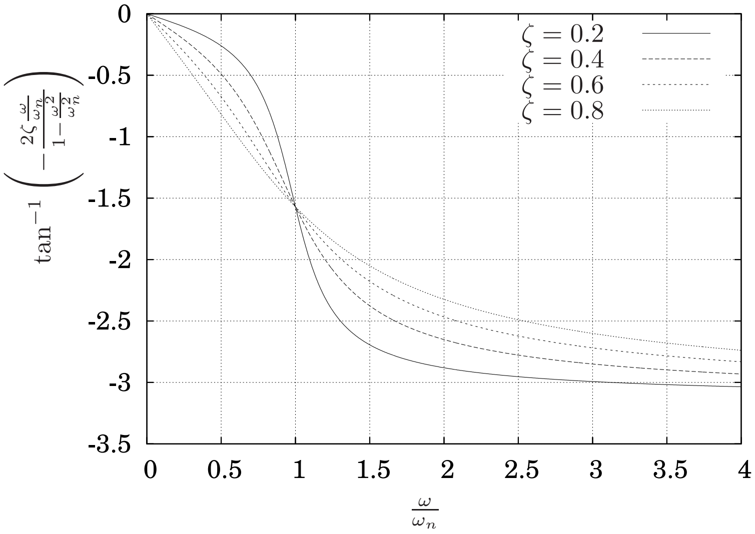


Figure 4.17. Phase shift between response and forcing function for various damping ratios *versus* frequency ratio.

sider for the *steady state* solution, *i.e.,* after the transient response represented by the homogeneous solution has become negligible.

4.4.1 Resonance

Resonance when the damping ratio is greater than zero does not require a different solution method. If damping is light, however, the magnitude of the response may be large; however, unlike the undamped case, it does not grow unbounded.

4.4.2 Vibrating Base

In this section we will consider an example that is illustrative of the operation of a suspension system.

Example 4.4.1 Consider the system illustrated in Figure 4.18. Assume that a vehicle is driving over a road with constant velocity, v and that the surface of the road is sinusoidal with wavelength λ and height, h. Determine the magnitude of the steady state motion of the car body (the mass) as well as the force transmitted to the mass as a function of the velocity of the vehicle. For this problem we will assume there is no gravity and that x is measured from the unstretched position of the spring. It is left as an exercise

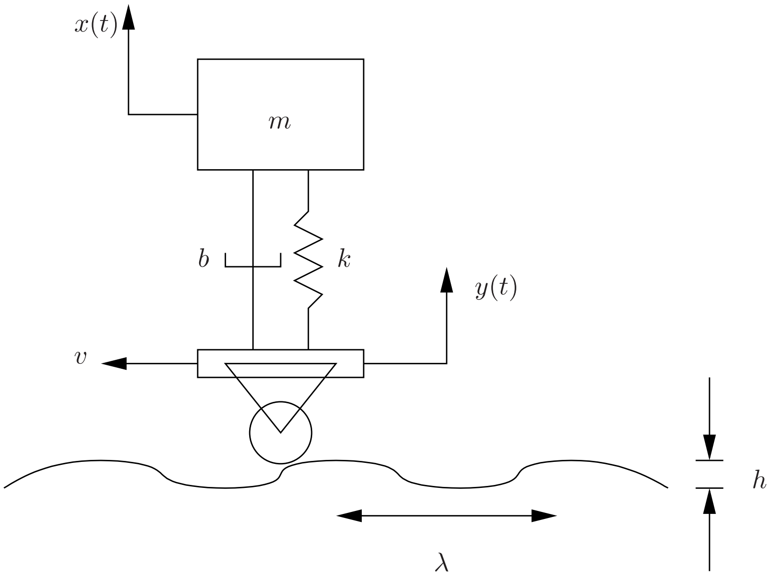


Figure 4.18. Model suspension system.

to show that if there is gravity if x is measured from the equilibrium position (the amount of static deflection), the equations of motion are unchanged.

The first task is to determine the vertical motion of the wheel. Since the velocity of the wheel is v, the horizontal position of the wheel at time t is given by vt. Since the wavelength of the oscillations of the road surface is λ, that means that the argument to the since function will need to go from zero to 2π in the amount of time it takes the vehicle to travel the distance λ. Hence, the time to travel λ is $T = \frac{\lambda}{v}$ is the period. Hence, the vertical motion of the wheel is given by

$$y(t) = h \sin\left(\frac{2\pi}{T}t + \hat{\phi}\right) = h \sin\left(\frac{2\pi v t}{\lambda} + \hat{\phi}\right) \tag{4.19}$$

where $\hat{\phi}$ is some unknown phase angle (we do not know where on the road the car was at $t = 0$, and we will see subsequently that if all we care about is the steady state behavior, then it does not matter).

To use Newton's law to derive the equations of motion of the system, we must draw a free body diagram of the mass, as illustrated in Figure 4.19. The only forces acting on the mass are from the damper and spring in the suspension. The force of the spring is proportional to the amount it is compressed, which in this case will by $y(t) - x(t)$. Similarly, the force from the damper will be proportional to the rate at which it is being compressed, which is $\dot{y}(t) - \dot{x}(t)$.

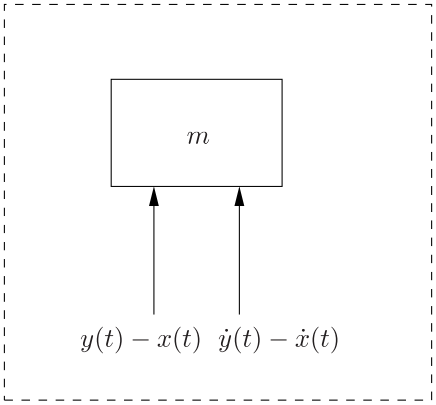


Figure 4.19. Free body diagram for mass in suspension problem.

Hence, using Newton's law, we have

$$\sum forces = ma \qquad \Longrightarrow \qquad b(\dot{y}-\dot{x})+k(y-x)=m\ddot{x},$$

or, substituting for $y(t)$ from equation 4.19 and rearranging gives

$$m\ddot{x}+b\dot{x}+kx=\frac{2\pi vhb}{\lambda}\cos\left(\frac{2\pi vt}{\lambda}+\hat{\phi}\right)+kh\sin\left(\frac{2\pi vt}{\lambda}+\hat{\phi}\right).$$

Following the procedure used several times previously, we can rewrite the right hand side to transform the equation into

$$m\ddot{x}+b\dot{x}+kx=\sqrt{\left(\frac{2\pi vhb}{\lambda}\right)^2+(kh)^2}\cos\left(\frac{2\pi vt}{\lambda}+\hat{\phi}+\overline{\phi}\right),$$

where

$$\overline{\phi}=\tan^{-1}\left(-\frac{2\pi vb}{\lambda k}\right).$$

Dividing both sides by m and letting $\phi=\hat{\phi}+\overline{\phi}$ gives

$$\ddot{x}+2\zeta\omega_n\dot{x}+\omega_n^2x=\sqrt{\left(\frac{2\pi vhb}{\lambda m}\right)^2+\left(\frac{kh}{m}\right)^2}\cos\left(\frac{2\pi vt}{\lambda}+\phi\right).$$

Finally, to simplify writing it, let

$$\omega=\frac{2\pi v}{\lambda}$$

(so ω is just proportional to v) which gives

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = \frac{h}{m}\sqrt{(\omega b)^2 + k^2}\cos(\omega t + \phi).$$

It may be tempting to think that we have already solved this problem since this looks a lot like equation 4.17; however, there is one critical distinction. In equation 4.17 the coefficient of the forcing term was constant. In this problem ω appears in the coefficient of the forcing term. Let us see what effect, if any, this has.

Assuming a particular solution of the form

$$x_p(t) = A\cos(\omega t + \phi) + B\sin(\omega t + \phi))$$

and doing all the usual work gives

$$\begin{aligned} A &= \frac{h}{m}\sqrt{(\omega b)^2 + k^2}\frac{\omega_n^2 - \omega^2}{(\omega_n^2 - \omega^2)^2 + (2\zeta\omega_n\omega)^2} \\ B &= \frac{h}{m}\sqrt{(\omega b)^2 + k^2}\frac{2\zeta\omega_n\omega}{(\omega_n^2 - \omega^2)^2 + (2\zeta\omega_n\omega)^2}. \end{aligned}$$

Examining the coefficients of the fractions and noting the k's and m's, one might be inclined to try to convert those to ω_n's and get the b term expressed somehow as ζ. In fact,

$$\begin{aligned} \frac{h}{m}\sqrt{(\omega b)^2 + k^2} &= h\sqrt{\left(\frac{\omega b}{m}\right)^2 + \left(\frac{k}{m}\right)^2} \\ &= h\sqrt{(2\zeta\omega_n\omega)^2 + (\omega_n^2)^2} \\ &= h\omega_n^2\sqrt{\left(2\zeta\frac{\omega}{\omega_n}\right)^2 + 1} \end{aligned}$$

Since we did it before, we might as well do it again. Dividing the numerator of both A and B by ω_n^2 and the denominator by ω_n^4, and while we are at it, computing $\sqrt{A^2 + B^2}$ gives the final answer

$$x_p(t) = h\sqrt{\frac{1 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2}{\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2}}\cos\left(\omega t + \hat{\phi} + \psi\right),$$

where

$$\psi = \tan^{-1}\left(-\frac{2\zeta\frac{\omega}{\omega_n}}{1 - \left(\frac{\omega^2}{\omega_n^2}\right)}\right).$$

Note that the magnitude of the variation in the road height, h is scaled by the term in the square root. In other words, the magnitude of the

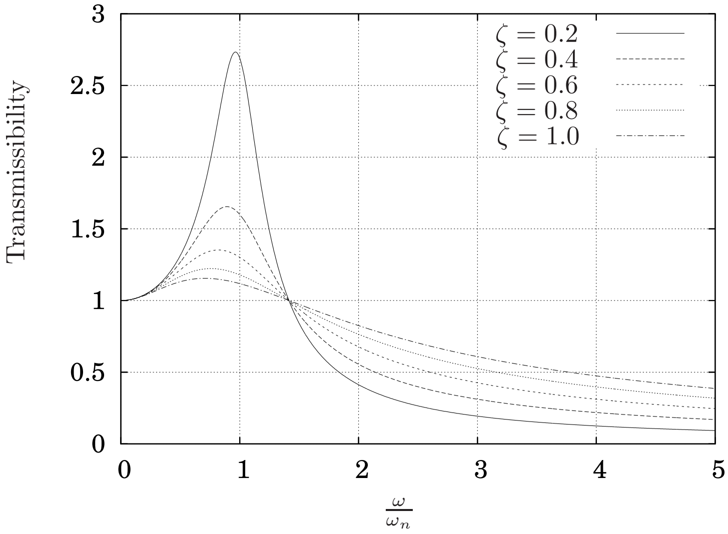


Figure 4.20. Transmissibility as a function of frequency ratio and damping ratio.

oscillation of the mass is the magnitude of the oscillation of the road times a factor that we will call the *transmissibility*. The transmissibility tells how much the oscillation of the road is transmitted to result in an oscillation of the mass. Plotting the transmissibility as a function of the frequency ratio for various damping ratios is probably a idea, so it appears in Figure 4.20.

Note that Figures 4.16 and 4.20 are not identical. In particular, in the latter case all the curves have the value of one at a frequency ratio of $\sqrt{2}$. Also, for high frequency ratios, corresponding to high velocities, a low damping ratio is preferable. This is in contrast to the magnitude factor for a an applied force where it is always the case that a larger damping ratio produces a smaller magnitude response, as should be apparent from Figure 4.16. ■

4.5 Exercises

Several of the following exercises refer to the mass-spring-damper system illustrated in Figure 4.21. Unless otherwise indicated, assume that there is no gravity and that $x(t) = 0$ at the unstretched position of the spring.

Problem 4.1 Consider the system illustrated in Figure 4.21. Assume that $b = 0$ and use either undetermined coefficients or variation of parameters to

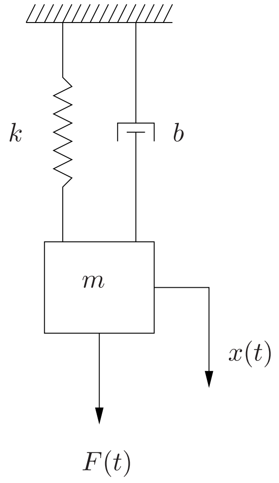


Figure 4.21. Mass-spring-damper system.

determine the solution to

$$\begin{aligned}\ddot{x}+\omega_n^2 x &= \frac{F}{m}\sin\omega t\\ x(0) &= x_0\\ \dot{x}(0) &= \dot{x}_0.\end{aligned}$$

Does it matter whether or not $\omega=\omega_n$? If so, be sure to consider both cases.

Problem 4.2 Determine an approximate numerical solution to the system in Problem 4.1 for the case where

$$\begin{aligned}m &= 1\\ k &= 4\\ F &= 1\end{aligned}$$

and $\omega=1.99$ or $\omega=2.0$. Plot the solution for each case on the same graph and explain any significant phenomenal that you observe.

Problem 4.3 Consider the system illustrated in Figure 4.21.

1. Determine the solution when $b\neq 0$ and $F(t)=F\sin\omega t$, $x(0)=x_0$ and $\dot{x}(0)=\dot{x}_0$. Does it matter if $0<\zeta<1$, $\zeta=1$ and $\zeta>1$? If so, be sure to determine the solution for each case.

2. Recall that for the case of undamped, forced oscillations, it was necessary to determine a separate form of the solution in the case of

resonance. Is the form of the solution determined in part 1 the same if $\omega = \omega_n$ or $\omega = \omega_d$? If so, be sure to determine those solutions as well.

3. Determine the magnification factor for the steady-state solution in part 1 and plot it for $\zeta = 0.2, 0.4, 0.6$ and 0.8 versus $\frac{\omega}{\omega_n}$.

4. Determine the phase shift between the forcing function and the steady state response in part 1 and plot it for $\zeta = 0.2, 0.4, 0.6$ and 0.8. Be sure to indicate what form you assumed for the particular solution, *e.g.,* $\cos(\omega t + \phi)$ or $\sin(\omega t - \phi)$, *etc.,* since the phase may be different depending on the form of the solution you used.

Problem 4.4 Use the figures you plotted for Problem 4.3 to determine good approximations to the steady-state solutions for

$$\begin{aligned}
\ddot{x} + 2\dot{x} + 25x &= 3\sin 2t \\
\ddot{x} + 2\dot{x} + 25x &= 3\sin 5t \\
\ddot{x} + 2\dot{x} + 25x &= 3\sin 10t \\
\ddot{x} + 4\dot{x} + 25x &= 3\sin 10t \\
\ddot{x} + 4\dot{x} + 49x &= 3\sin 10t \\
\ddot{x} + 2\dot{x} + 36x &= 6\sin 20t.
\end{aligned}$$

Problem 4.5 Consider the system illustrated in Figure 4.21. If $0 < \zeta < 1$ and

$$F(t) = F_c \cos\omega_c t + F_s \sin\omega_s t$$

is it possible to combine your answer from Problem 4.3 and the solution in Equation 4.18 to obtain the steady state solution, or is it necessary to work out the whole thing again? In either case, provide the answer and justify it.

Problem 4.6 Consider

$$\ddot{x} + 4\dot{x} + 16x = \cos 4t + \cos 4.2t.$$

If we are only interested in the steady state response, is it valid to write

$$x_{ss} = \delta_1 M_1 \cos(\omega t + \phi_1) + \delta_2 M_2 \cos(\omega t + \phi_2)$$

where M_1, M_2, ϕ_1 and ϕ_2 are determined from the appropriate graphs? How would you determine δ_1 and δ_2? Demonstrate whether or not it works by picking some initial conditions and writing a computer program to determine an approximate numerical solution and comparing it to the combination of the approximate steady state solutions determined from the graphs.

Problem 4.7 Consider the system illustrated in Figure 4.21 and let

$$\begin{aligned} m &= 1 \\ b &= 1 \\ k &= 1 \\ F(t) &= 3\cos 2t. \end{aligned}$$

Use Figures 4.16 and 4.17 to determine a good approximation for the steady state response of the system. Will the magnitude of the steady state response increase or decrease if the forcing frequency $\omega = 2$ is increased?

Problem 4.8 Write a computer program to determine an approximate numerical solution to the system in Problem 4.4 with $x(0) = 0$ and $\dot{x}(0) = 0$. Plot the approximate solution as well as the solution determined in Problem 4.4 and compare the results. Explain any significant differences.

Problem 4.9 Consider the system illustrated in Figure 4.21 and assume that there *is* gravity.

1. Determine the equation of motion for the system when $x = 0$ at the unstretched position of the spring.

2. Determine the equation of motion for the system when $x = 0$ at the equilibrium position. In other words, $x = 0$ at the position when the spring is stretched by an amount due to the weight of the mass.

Problem 4.10 Consider the system illustrated in Figure 4.21 with $F(t) = 0$, (damped, unforced). Let

$$\begin{aligned} \omega_n &= 1 \\ x(0) &= 1 \\ \dot{x}(0) &= 0 \end{aligned}$$

and plot the solution for $\zeta = 0.0, 0.2, 0.4, 0.6, 0.8$ and 1.0 for $t = 0$ to $t = 10$. Plot all the solutions on the same plot.

Problem 4.11 Write a computer program to determine an approximate numerical solution for

$$m\ddot{x} + b\dot{x} + kx = F\sin\omega t$$

when

$$\begin{aligned} \omega_n &= 2 \\ \zeta &= 0.3 \\ m &= 1 \\ \omega &= 1.5 \\ F &= 5 \\ x(0) &= 1 \\ \dot{x}(0) &= 1. \end{aligned}$$

1. On the same plot, plot the numerical solution and the solution for these values substituted into the closed-form solution from Problem 4.3.

2. Vary the step size for the numerical solution to determine the largest step size that gives a reasonable approximation to the exact solution.

3. Use the figures from Problem 4.3 to determine a good approximation for the steady state solution and plot the solution on the same graph. At approximately what time does the transient solution decay sufficiently so that the steady state solution is approximately equal to the exact solution?

Problem 4.12 Figure 4.20 plots the magnitude of the steady state oscillation of a mass subjected to a vibrating base. For some applications, such as an automotive suspension, the magnitude of the response is not the critical factor, but rather the net force to which the mass is subjected.

1. Determine an expression for the force to which the mass illustrated in Figure 4.18 is subjected.

2. Manipulate the expression for the force so that it is in the form of

   $$f = -khM_f \cos\left(\omega t + \phi\right).$$

   Explain the interpretation of the term M_f. Plot M_f as a function of $\frac{\omega}{\omega_n}$ for various damping ratios.

Problem 4.13 Use Figure 4.20 to determine the magnitude of the motion of the mass in Figure 4.18 if

$$\begin{aligned} k &= 2 \\ m &= 2 \\ b &= 1 \\ h &= 0.25. \end{aligned}$$

Plot the magnitude of the motion *versus* ω.

Chapter 5

Ordinary Second Order Linear Variable Coefficient Equations

This chapter presents the the use of power series solutions to differential equations. The primary use of such solutions will be for variable coefficient, linear, ordinary differential equations. The approach will be to assume a solution of the form

$$x(t) = a_0 + a_1 (t - t_0) + a_2 (t - t_0)^2 + a_3 (t - t_0)^3 + \cdots ,$$

and then substitute it into the differential equation to see if we can determine the coefficients. Of course, in engineering where the solution must be evaluated, at most only a finite number of terms in the series may be used unless the whole series converges to some elementary function in an identifiable manner. So, to be useful, we will want to know how to check the following.

1. To what extent does the series represent the actual solution to the differential equation?

2. To what extent is a truncacted series including only the first n terms a good approximation to the solution?

3. If we include more terms in a trucated series, are we guaranteed to obtain an better approximation?

4. Is there a way to determine the answer to some of the above questions before solving for the coefficients?

5. Is there a way to determine the answer to some of the above questions after solving for the coefficients?

5.1 Motivational Example

Example 5.1.1 Consider the linear, second order, ordinary, homogeneous, constant coefficient differential equation

$$\begin{aligned} \ddot{x} + 4x &= 0 \qquad (5.1) \\ x(0) &= 1 \\ \dot{x}(0) &= 1 \end{aligned}$$

Instead of making the usual assumption that this equation has exponential solutions, let us assume a power series form for the solution. In particular, assume

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 + a_6 t^6 + \cdots . \qquad (5.2)$$

Observe that this is a particularly convenient form since

$$\begin{aligned} a_0 &= x(0) \\ a_1 &= \dot{x}(0). \end{aligned}$$

We will proceed as usual: substitute the assumed form of the solution into the differential equation to see if we get anything useful. So, differentiating the solution givves

$$\dot{x}(t) = a_1 + 2a_2 t + 3a_3 t^2 + 4a_4 t^3 + 5a_5 t^4 + 6a_6 t^5 \cdots$$

and

$$\ddot{x}(t) = 2a_2 + 6a_3 t + 12a_4 t^2 + 20a_5 t^3 + 30a_6 t^4 \cdots .$$

Substituting into Equation 5.1 gives

$$\begin{aligned} &\left(2a_2 + 6a_3 t + 12a_4 t^2 + 20a_5 t^3 + 30a_6 t^4 + \cdots\right) \\ &+ \quad 4\left(a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 + a_6 t^6 + \cdots\right) = 0. \end{aligned}$$

Since this equality must hold for all t, we may equate the coefficients of t on each side of the equation, which is particularly easy since the right hand side is zero. Hence

$$\begin{aligned} 2a_2 + 4a_0 &= 0 \\ 6a_3 + 4a_1 &= 0 \\ 12a_4 + 4a_2 &= 0 \\ 20a_5 + 4a_3 &= 0 \\ 30a_6 + 4a_4 &= 0 \\ &\vdots \end{aligned}$$

and therefore through t^6,

$$x(t) = a_0 + a_1 t - 2a_0 t^2 - \frac{2}{3}a_1 t^3 + \frac{2}{3}a_0 t^4 + \frac{2}{15}a_1 t^5 - \frac{4}{45}a_0 t^6 + \cdots .$$

Substituting for the initial conditions gives

$$x(t) = 1 + t - 2t^2 - \frac{2}{3}t^3 + \frac{2}{3}t^4 + \frac{2}{15}t^5 - \frac{4}{45}t^6 + \cdots . \quad ■$$

It will not always be the case, but for the example just completed, it is possible to find an expression for every term in the power series.

Example 5.1.2 Consider again

$$\ddot{x} + 4x = 0 \tag{5.3}$$

and assume the same power series solution. Note, however, that we may express it in more general terms using

$$\begin{aligned} x(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 + a_6 t^6 + \cdots & (5.4) \\ &= \sum_{n=0}^{\infty} a_n t^n. & (5.5) \end{aligned}$$

Differentiating Equation 5.5 gives

$$\begin{aligned} \dot{x}(t) &= a_1 + 2a_2 t + 3a_3 t^2 + 4a_4 t^3 + 5a_5 t^4 + 6a_6 t^5 \cdots \\ &= \sum_{n=1}^{\infty} n a_n t^{n-1} \end{aligned}$$

and

$$\begin{aligned} \ddot{x}(t) &= 2a_2 + 6a_3 t + 12a_4 t^2 + 20a_5 t^3 + 30a_6 t^4 \cdots \\ &= \sum_{n=2}^{\infty} n\,(n-1)\,a_n t^{n-2}. \end{aligned}$$

Substituting the series expressions into Equation 5.3 gives

$$\sum_{n=2}^{\infty} n\,(n-1)\,a_n t^{n-2} + 4\sum_{n=0}^{\infty} a_n t^n = 0. \tag{5.6}$$

As before we want to equate powers of t, which would be easiest if there were just one sum in the expression. We may rewrite the first sum as

$$\sum_{n=2}^{\infty} n\,(n-1)\,a_n t^{n-2} = \sum_{n=0}^{\infty} (n+2)\,(n+1)\,a_{n+2} t^n$$

which simply shifted the index of the sum to start at zero instead of two.

So, now Equation 5.6 is of the form

$$\begin{aligned}
&\sum_{n=0}^{\infty}(n+2)(n+1)a_{n+2}t^n+4\sum_{n=0}^{\infty}a_nt^n \\
&\quad= \sum_{n=0}^{\infty}\left((n+2)(n+1)a_{n+2}t^n+4a_nt^n\right) \\
&\quad= \sum_{n=0}^{\infty}\left((n+2)(n+1)a_{n+2}+4a_n\right)t^n \\
&\quad= 0.
\end{aligned}$$

Thus,

$$(n+2)(n+1)a_{n+2}+4a_n=0$$

or

$$a_{n+2}=-\frac{4}{(n+2)(n+1)}a_n.$$

From this expression, all the even coefficients may be determined from a_0 and all the odd coefficients may be obtained from a_1. Observe that if $n=0$,

$$\begin{aligned}
a_2 &= -\frac{4}{(2)(1)}a_0 \\
&= -2a_0
\end{aligned}$$

for $n=2$,

$$\begin{aligned}
a_4 &= -\frac{4}{(4)(3)}a_2 \\
&= \left(-\frac{4}{(4)(3)}\right)\left(-\frac{4}{(2)(1)}\right)a_0 \\
&= \frac{2}{3}a_0
\end{aligned}$$

and for any even n

$$\begin{aligned}
a_n &= -\frac{4}{n(n-1)}a_{n-2} \\
&= \left(-\frac{4}{n(n-1)}\right)\left(-\frac{4}{(n-2)(n-3)}\right)\cdots\left(-\frac{4}{(2)(1)}\right)a_0 \\
&= \frac{(-4)^{\frac{n}{2}}}{n!}. \qquad (5.7)
\end{aligned}$$

The $\frac{n}{2}$ power is an integer since n is even. Recall that these coefficients are for the terms in the series in Equation 5.5. Since it will have to be combined with the odd terms, it will be much more conveneient for an expression where $n = 0, 1, 2, 3, \ldots$ instead of having to restrict it to be even. So, replacing n with $2n$ in Equation 5.7 gives

$$a_{2n} = \frac{(-4)^n}{(2n)!}$$

for $n = 0, 1, 2, \ldots$.

For the odd terms, for $n = 1$

$$\begin{aligned} a_3 &= -\frac{4}{(3)(2)} a_1 \\ &= -\frac{2}{3} a_1 \end{aligned}$$

and for $n = 3$

$$\begin{aligned} a_5 &= -\frac{4}{(5)(4)} a_3 \\ &= \left(-\frac{4}{(5)(4)}\right)\left(-\frac{4}{(3)(2)}\right) a_1 \\ &= \frac{2}{15} a_1 \end{aligned}$$

and for any odd n

$$\begin{aligned} a_n &= -\frac{4}{n(n-1)} a_{n-2} \\ &= \left(-\frac{4}{n(n-1)}\right)\left(-\frac{4}{(n-2)(n-3)}\right)\cdots\left(-\frac{4}{(3)(2)}\right) a_1. \end{aligned}$$

Since a_3 had one term multiplying a_1 and a_5 had two terms multiplying a_1, there are $\frac{n-1}{2}$ terms for a general odd n, or

$$a_n = \frac{(-4)^{\frac{n-1}{2}}}{n!} a_1$$

for $n = 3, 5, 7, \ldots$. Again, to have the series indexed by $n = 0, 1, 2, 3, \ldots$, change n to $2n + 1$ to give

$$a_{2n+1} = \frac{-4^n}{(2n+1)!} a_1$$

for $n = 0, 1, 2, \ldots$.

Clearly, if we only consider a finite number of terms in the series, the truncated series will only approximate the real solution. If we consider the case where

$$\begin{aligned} x(0) &= 1 \\ \dot{x}(0) &= 1, \end{aligned}$$

then we have

$$x(t) = \cos 2t + \frac{1}{2}\sin 2t \tag{5.8}$$

for the "usual" solution, and, in series form after an easy computation that shows that $a_0 = x(0)$ and $a_1 = \dot{x}(0)$,

$$x(t) = \sum_{n=0}^{\infty} \left(\frac{(-1)^n \, 2^{2n}}{(2n)!} t^{2n} + \frac{(-1)^n \, 2^{2n}}{(2n+1)!} t^{2n+1} \right). \tag{5.9}$$

Figure 5.1 illustrated the exact solution along with truncated series solutions including the first 10, 15 and 20 terms.

To make the final connection between the two forms of the solution in Equations 5.8 and 5.9, factor a $\frac{1}{2}$ out of the second term to obtain

$$x(t) = \sum_{n=0}^{\infty} \left(\frac{(-1)^n \, 2^{2n}}{(2n)!} t^{2n} + \left(\frac{1}{2}\right) \frac{(-1)^n \, 2^{2n+1}}{(2n+1)!} t^{2n+1} \right).$$

The first term is the Taylor series for $\cos 2t$ and the second term is the Taylor series for $\sin 2t$. ■

Of course, Example 5.1.1 solved an equation that we already knew how to solve; furthermore, it was even more work than assuming exponential solutions. The real utility to series methods is in the case of variable coefficient problems. If we were to return to Example 5.1.1, none of the steps in the process assumed that the coefficients were constant, and, in fact, it turns out not to be necessary.

5.2 Convergence

A very important properties of a series is its convergence properties. If we are to obatin series representations of solutions of differential equations, we must be able to determine in what manner they converge. A general analysis of the convergence properties of a series requires use of the property of *analyticity.* Some readers of this text may have the required background while others may not. Hence, instead of presenting general theorems to allow us to determine *a priori* that a series solution is converngent (or its interval of convergence), we will use a convergence test on each solution we obtain. This is much less efficient than having the theorems handy; on the other hand, it is more straight-forward.

The test for convergence of a series that we will primarily use is the *ratio test*

Theorem 5.2.1 *For the series* $\sum_{n=0}^{\infty} a_n$, *if* $a_n \neq 0$ *for* $n \geq 1$, *suppose*

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = r.$$

If

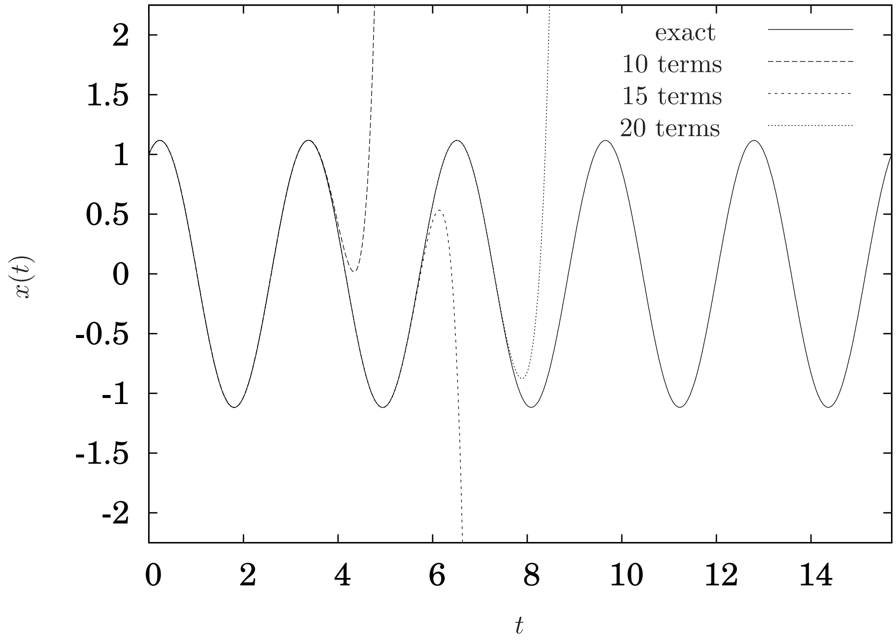


Figure 5.1. Exact and truncated series solutions for Example 5.1.1.

1. *If $r < 1$, then the series converges absolutely.*
2. *If $r > 1$, then the series diverges.*
3. *If $r = 1$, no conclusion can be drawn from this test alone.*

In the case at hand when the series is a power series of the form $\sum_{n=1}^{\infty} a_n t^n$, we have

$$\lim_{n\to\infty} \left| \frac{a_{n+1} t^{n+1}}{a_n t^n} \right| = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| |t| \, .$$

So, a corollarly to Theorem 5.2.1 is as follows.

Corollary 5.2.2 *If*

$$t_c = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|$$

the power series $\sum_{n=1}^{\infty} a_n t^n$ converges for $t \in (-t_c, t_c)$.

5.3 Series Solutions about an Ordinary Point

In Section 1.5.3 a linear differential equation was defined by Equation 1.6 to be of the form

$$f_n(t)\frac{d^n x}{dt^n}(t) + f_{n-1}(t)\frac{d^{n-1} x}{dt^{n-1}}(t) + \cdots + f_1(t)\frac{dx}{dt}(t) + f_0(t)x(t) = g(t).$$

While the methods generalize to higher order, we will restrict our attention to second order equations of the form

$$f_2(t)\frac{d^2 x}{dt^2}(t) + f_1(t)\frac{dx}{dt}(t) + f_0(t)x(t) = g(t), \tag{5.10}$$

because a surprizingly large number of important differential equations in engineering fall into this category.

Examining Equation 5.10, one may reasonably (and correctly) conclude that points where $f_2(t) = 0$ are problematic. This is intuitively because the order of the equation changes at those points. This section will consider solutions of Equation 5.10 for values of t where $f_2(t) \neq 0$. Unfortunately we can not simply ignore the case where $f_2(t)$ is zero. The following section consders that case.

Definition 5.3.1 For the second order, linear, ordinary differential equation

$$f_2(t)\frac{d^2 x}{dt^2}(t) + f_1(t)\frac{dx}{dt}(t) + f_0(t)x(t) = g(t),$$

where $f_2(t)$, $f_1(t)$ and $f_0(t)$ are analytic, a point t where $f_2(t) \neq 0$ is called an *ordinary point.* ◇

Example 5.3.2 Determine a series solution to Airy's equation

$$\frac{d^2x}{dt^2}(t) - tx(t) = 0.$$

Assume

$$x(t) = \sum_{n=0}^{\infty} a_n t^n.$$

As before

$$\begin{aligned} \frac{d^2x}{dt^2}(t) &= \sum_{n=2}^{\infty} n(n-1) a_n t^{n-1} \\ &= \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} t^n \end{aligned}$$

and substituting into the differential equation gives

$$\begin{aligned} \left(\sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} t^n\right) - t\left(\sum_{n=0}^{\infty} a_n t^n\right) &= \\ \left(\sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} t^n\right) - \left(\sum_{n=0}^{\infty} a_n t^{n+1}\right) &= 0. \end{aligned}$$

In order to equate powers of t, shift the index of summation in the second sum by one, *i.e.,*

$$\sum_{n=0}^{\infty} a_n t^{n+1} = \sum_{n=1}^{\infty} a_{n-1} t^n,$$

so we have

$$\left(\sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} t^n\right) - \sum_{n=1}^{\infty} a_{n-1} t^n = 0.$$

The two series have the same powers in t, but start at different values of n. To handle this, simply write the first term of the first series by itself, *i.e.,*

$$\begin{aligned} (2)(1) a_2 t^0 + \left(\sum_{n=1}^{\infty} (n+2)(n+1) a_{n+2} t^n\right) - \sum_{n=1}^{\infty} a_{n-1} t^n &= \\ 2a_2 + \sum_{n=1}^{\infty} \left[(n+2)(n+1) a_{n+2} - a_{n-1}\right] t^n &= 0. \end{aligned}$$

So, equating powers of t gives

$$a_2 = 0$$

and

$$a_{n+2} = \frac{1}{(n+2)(n+1)} a_{n-1}. \tag{5.11}$$

Since $a_2 = 0$, by Equation 5.11 gives that

$$a_5 = a_8 = a_{11} = a_{14} = \cdots = 0.$$

For $a_1, a_4, a_7, \ldots$, observe that

$$\begin{aligned}
a_4 &= \frac{1}{(4)(3)} a_1 \\
a_7 &= \frac{1}{(7)(6)} a_4 \\
&= \frac{1}{(7)(6)(4)(3)} a_1 \\
a_{11} &= \frac{1}{(10)(9)} a_7 \\
&= \frac{1}{(10)(9)(7)(6)(4)(3)} a_1
\end{aligned}$$

or, in general for $n = 1, 2, 3, \ldots$

$$a_{3n+1} = \frac{1}{(3n+1)(3n)(3n-2)(3n-3)\cdots(4)(3)} a_1.$$

Similarly, for $n = 1, 2, 3, \ldots,$

$$a_{3n} = \frac{1}{(3n)(3n-1)(3n-3)(3n-4)\cdots(3)(2)} a_0. \qquad \blacksquare$$

Substituting the cofficients into the solution gives

$$\begin{aligned}
x(t) &= a_0 \left(1 + \sum_{n=1}^{\infty} \frac{1}{(3n)(3n-1)(3n-3)(3n-4)\cdots(3)(2)} t^{3n} \right) + \\
&\quad a_1 \left(t + \frac{1}{(3n+1)(3n)(3n-2)(3n-3)\cdots(4)(3)} t^{3n+1} \right)
\end{aligned}$$

5.4 Series Solutions about a Singular Point

5.5 A Collection of Famous Series Solutions

5.5.1 Bessel's equation

Bessel's equation is

$$t^2 \frac{d^2 x}{dt^2} + r \frac{dx}{dt} + \left(t^2 - m^2\right) x = 0. \tag{5.12}$$

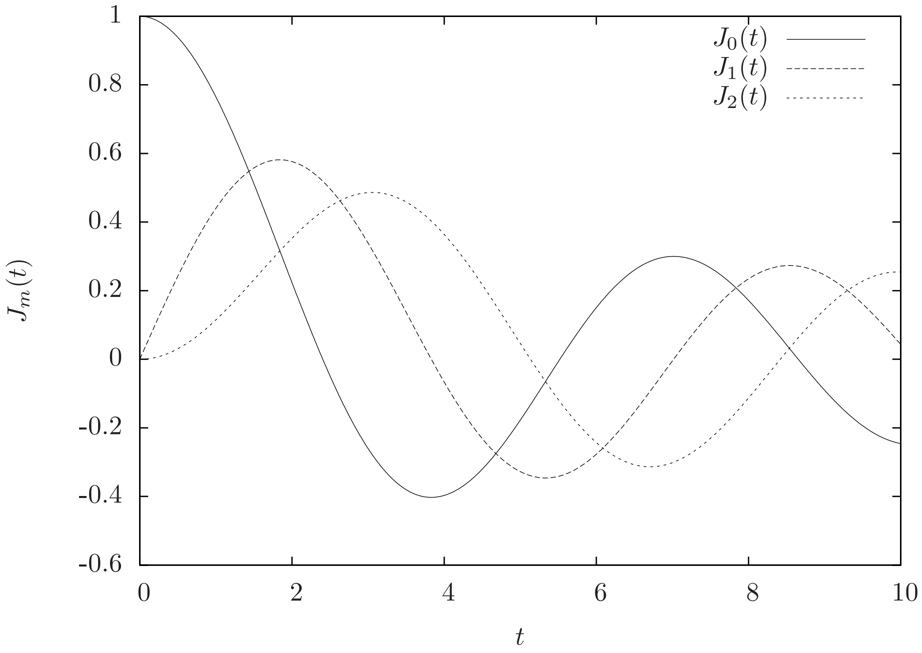


Figure 5.2. Bessel functions of the first kind.

The parameter m may be a real or complex number, but in the special case where it is an integer, it is called the *order* of the equation. Since this is a second order equation, two linearly independent solutions are necessary to determine a general solution. In the case when m is an integer, the two solutions are given by

$$J_m(t) = \sum_{n=1}^{\infty} \frac{(-1)^n}{2^{2n+m} n!\,(n+m)!} t^{2n+m}$$

and

$$Y_m(t) = \frac{J_r(t)\cos(m\pi) - J_{-m}(t)}{\sin(m\pi)}.$$

The function $J_m(t)$ is called the *Bessel function of the first kind* and the function $Y_m(t)$ is called the *Bessel function of the second kind*. Figure 5.2 illustrates $J_m(t)$ for various integer orders and Figure 5.3 illustrates $Y_m(t)$ for various integer orders. In addition to being the linearly independent solutions to Equation 5.12, these two functions have some additional remarkable properties which will be explored in Chapter 12.

It will turn out to be handy to have the values at which $J_m(t)$ is zero. Table 5.1 tabulates them for $J_m(t)$ and Table 5.2 tabulates them for $Y_m(t)$.

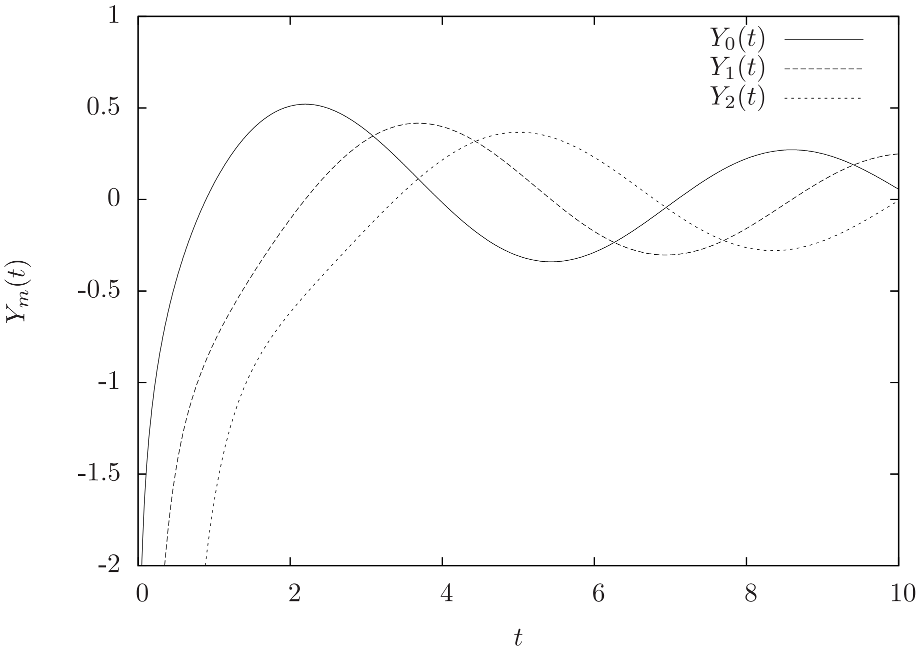


Figure 5.3. Bessel functions of the secon kind.

Order										
0	2.40483	5.52008	8.65373	11.7915	14.9309	18.0711	21.2116	24.3525	27.4935	30.6346
1	3.83171	7.01559	10.1735	13.3237	16.4706	19.6159	22.7601	25.9037	29.0468	32.1897
2	5.13562	8.41724	11.6198	14.796	17.9598	21.117	24.2701	27.4206	30.5692	33.7165
3	6.38016	9.76102	13.0152	16.2235	19.4094	22.5827	25.7482	28.9084	32.0649	35.2187
4	7.58834	11.0647	14.3725	17.616	20.8269	24.019	27.1991	30.371	33.5371	36.699
5	8.77148	12.3386	15.7002	18.9801	22.2178	25.4303	28.6266	31.8117	34.9888	38.1599
6	9.93611	13.5893	17.0038	20.3208	23.5861	26.8202	30.0337	33.233	36.422	39.6032
7	11.0864	14.8213	18.2876	21.6415	24.9349	28.1912	31.4228	34.6371	37.8387	41.0308
8	12.2251	16.0378	19.5545	22.9452	26.2668	29.5457	32.7958	36.0256	39.2404	42.4439
9	13.3543	17.2412	20.807	24.2339	27.5837	30.8854	34.1544	37.4001	40.6286	43.8438
10	14.4755	18.4335	22.047	25.5095	28.8874	32.2119	35.4999	38.7618	42.0042	45.2316

Table 5.1. Table of zeros of the Bessel function of the first kind.

Order										
0	0.893577	3.95768	7.08605	10.2223	13.3611	16.5009	19.6413	22.782	25.923	29.064
1	2.19714	5.42968	8.59601	11.7492	14.8974	18.0434	21.1881	24.3319	27.4753	30.6183
2	3.38424	6.79381	10.0235	13.21	16.379	19.539	22.694	25.8456	28.9951	32.143
3	4.52702	8.09755	11.3965	14.6231	17.8185	20.9973	24.1662	27.3288	30.487	33.642
4	5.64515	9.36162	12.7301	15.9996	19.2244	22.4248	25.6103	28.7859	31.9547	35.1185
5	6.74718	10.5972	14.0338	17.3471	20.6029	23.8265	27.0301	30.2203	33.4011	36.575
6	7.83774	11.811	15.3136	18.6707	21.9583	25.2062	28.429	31.6349	34.8286	38.0135
7	8.91961	13.0077	16.5739	19.9743	23.294	26.5668	29.8095	33.0318	36.2393	39.4358
8	9.99463	14.1904	17.8179	21.2609	24.6126	27.9105	31.1737	34.4129	37.6346	40.8434
9	11.0641	15.3613	19.0479	22.5328	25.9162	29.2394	32.5233	35.7797	39.0162	42.2376
10	12.1289	16.5223	20.266	23.7917	27.2066	30.555	33.8597	37.1336	40.3851	43.6195

Table 5.2. Table of zeros of the Bessel function of the second kind.

Chapter 6

Systems of Ordinary First Order Linear Constant Coefficient Equations

6.1 Introduction

So far this book has considered the theory and applications of first and second order differential equations. This chapter considers nth order differential equations, or equivalently, systems of n first order differential equations. As will become readily apparent, the theoretical basis for solving such systems relies heavily upon matrix algebra theory.

The types of equations considered in this chapter are systems of of n first order ordinary differential equations of the form

$$\begin{aligned} \dot{x}_1(t) &= f_1\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) && (6.1)\\ \dot{x}_2(t) &= f_2\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) && (6.2)\\ &\vdots && (6.3)\\ \dot{x}_n(t) &= f_n\left(x_1(t), x_2(t), \ldots, x_n(t), t\right). && (6.4) \end{aligned}$$

In vector form, this is equivalent to

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \frac{d}{dt}\begin{bmatrix} f_1\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) \\ f_2\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) \\ \vdots \\ f_n\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) \end{bmatrix} \tag{6.5}$$

which will often be written more concisely as

$$\dot{\xi}\left(t\right) = f\left(\xi\left(t\right), t\right) \tag{6.6}$$

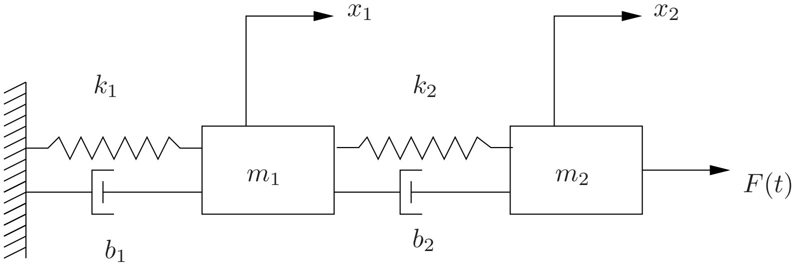


Figure 6.1. Two degree of freedom mass-spring-damper system.

where

$$\xi(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

and

$$f\left(\xi\left(t\right), t\right) = \begin{bmatrix} f_1\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) \\ f_2\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) \\ \vdots \\ f_n\left(x_1(t), x_2(t), \ldots, x_n(t), t\right) \end{bmatrix}$$

6.2 Motivational Example

Consider the mass-spring-damper system illustrated in Figure 6.1. While this is the simplified prototypical system that we will consider, it also is representative of a much larger class of useful engineering systems such as automobile suspensions and civil structures. As is the usual case, assume that x_1 and x_2 are ~~absolute~~ the displacements of m_1 and m_2 respectively measured ~~from~~ in an inertial coordinate system where the values of x_1 and x_2 are zero when the springs are unstretched. ~~the equilibrium configuration of the system. If there is no gravity, then x_1 and x_2 will be measured from the position of the masses when the springs are unstretched, and if there is gravity, then they will be measured from the position of the masses when the springs are statically compressed or extended by the weight of the masses.~~

Considering a free body diagram for each mass illustrated in Figure 6.2 and

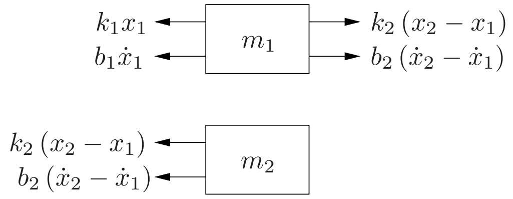


Figure 6.2. Free body diagrams for masses in Figure 6.1.

applying Newton's law gives

$$\begin{aligned} m_1 \ddot{x}_1 &= -b_1 \dot{x}_1 - k_1 x_1 + k_2 (x_2 - x_1) + b_2 (\dot{x}_2 - \dot{x}_1) \\ m_2 \ddot{x}_2 &= -k_2 (x_2 - x_1) - b_2 (\dot{x}_2 - \dot{x}_1) + F(t), \end{aligned}$$

and rearranging into the standard form of descending order of derivatives gives

$$\begin{aligned} m_1 \ddot{x}_1 + (b_1 + b_2) \dot{x}_1 - b_2 \dot{x}_2 + (k_1 + k_2) x_1 - k_2 x_2 &= 0 \\ m_2 \ddot{x}_2 - b_2 \dot{x}_1 + b_2 \dot{x}_2 - k_2 x_1 + k_2 x_2 &= F(t). \end{aligned} \tag{6.7}$$

These equations are *coupled* since x_1 appears in the x_2 equation and *vice-versa*. One's first inclination may be to try to solve one equation for one of either x_1 or x_2 and substitute into the other, but such an approach is impossible since the equations involve the derivatives of the variables as well.

An insightful extrapolation of the method considered in Chapter 3 might lead one to attempt to solve the homogeneous problem first followed by some method for the particular solutions; indeed, this is fundamentally the approach we will utilize. In fact, for the homogeneous case ($F(t) = 0$), *i.e.,*

$$\begin{aligned} m_1 \ddot{x}_1 + (b_1 + b_2) \dot{x}_1 - b_2 \dot{x}_2 + (k_1 + k_2) x_1 - k_2 x_2 &= 0 \\ m_2 \ddot{x}_2 - b_2 \dot{x}_1 + b_2 \dot{x}_2 - k_2 x_1 + k_2 x_2 &= 0, \end{aligned}$$

a good guess may be assume

$$\begin{aligned} x_1(t) &= e^{\lambda_1 t} \\ x_2(t) &= e^{\lambda_2 t}, \end{aligned} \tag{6.8}$$

and substitute. This will actually work and is essentially the approach we will take. ~~but as we consider higher and higher order systems, *e.g.*, systems like in Figure 6.1 but with more masses, presenting algebra will become somewhat cumbersome. In order to consider the problem more concisely (and more elegantly) resorting to matrix algebra is the typical approach. For the mathematically inclined, the abstraction is nice because it still presents the essence of the problem; however, for those less mathematically inclined it can be problematic.~~ In order to unify the solution method with one that is applicable to other types of problems in addition to vibrations problems with multiple masses, we will first confert the system into an equivalent system of first order equations and the result will be expressed in matrix form. The key concept to keep in mind is that behind all the matrix theory presented, the basic approach for the homogeneous problem is still to simply consider solutions of the form of Equation 6.9.

The general approach to solve systems of this type is to first convert the system into a system of first order equations. This is illustrated by the following example.

Example 6.2.1 Let

$$\begin{aligned} \xi_1 &= x_1 \\ \xi_2 &= \dot{x}_1 \\ \xi_3 &= x_2 \\ \xi_4 &= \dot{x}_2. \end{aligned}$$

Then

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} = \begin{bmatrix} \xi_2 \\ \frac{-b_1\xi_2 - k_1\xi_1 + k_2\xi_3 - k_2\xi_1 + b_2\xi_4 - b_2\xi_2}{m_1} \\ \xi_3 \\ \frac{-k_2\xi_3 + k_2\xi_1 - b_2\xi_4 + b_4\xi_2}{m_2}. \end{bmatrix} \tag{6.9}$$

Since this equation is linear in the ξ_i's, it can be expressed as

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_1+k_2}{m_1} & -\frac{b_1+b_2}{m_1} & \frac{k_2}{m_1} & \frac{b_2}{m_1} \\ 0 & 0 & 0 & 1 \\ \frac{k_2}{m_2} & \frac{b_2}{m_2} & -\frac{k_2}{m_2} & -\frac{b_2}{m_2} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix}.$$

If we let

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix}$$

and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_1+k_2}{m_1} & -\frac{b_1+b_2}{m_1} & \frac{k_2}{m_1} & \frac{b_2}{m_1} \\ 0 & 0 & 0 & 1 \\ \frac{k_2}{m_2} & \frac{b_2}{m_2} & -\frac{k_2}{m_2} & -\frac{b_2}{m_2} \end{bmatrix}, \tag{6.10}$$

then this whole system can be expressed simply as

$$\dot{\xi} = A\xi. \tag{6.11}$$

Clearly, the way to solve this equation hinges on the property of the matrix A. Exploiting the properties of A to solve this equation is our task at hand.■

Now, considering a general first order matrix differential equation of the form

$$\dot{\xi} = A\xi \tag{6.12}$$

the question arises as to the nature of the solution. Motivated by the results from Chapters 2 and 3, consider the possibility of a solution of the form

$$\xi(t) = \hat{\xi}e^{\lambda t},$$

where $\hat{\xi}$ is a constant vector. In full detail,

$$\xi(t) = \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \vdots \\ \xi_n(t) \end{bmatrix} = \begin{bmatrix} \hat{\xi}_1 \\ \hat{\xi}_2 \\ \vdots \\ \hat{\xi}_n \end{bmatrix} e^{\lambda t} = \begin{bmatrix} \hat{\xi}_1 e^{\lambda t} \\ \hat{\xi}_2 e^{\lambda t} \\ \vdots \\ \hat{\xi}_n e^{\lambda t} \end{bmatrix}.$$

Substituting this into Equation 6.12 gives

$$\lambda \begin{bmatrix} \hat{\xi}_1 e^{\lambda t} \\ \hat{\xi}_2 e^{\lambda t} \\ \vdots \\ \hat{\xi}_n e^{\lambda t} \end{bmatrix} = A \begin{bmatrix} \hat{\xi}_1 e^{\lambda t} \\ \hat{\xi}_2 e^{\lambda t} \\ \vdots \\ \hat{\xi}_n e^{\lambda t} \end{bmatrix}.$$

Inserting an identity matrix gives

$$\lambda \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \hat{\xi}_1 e^{\lambda t} \\ \hat{\xi}_2 e^{\lambda t} \\ \vdots \\ \hat{\xi}_n e^{\lambda t} \end{bmatrix} = A \begin{bmatrix} \hat{\xi}_1 e^{\lambda t} \\ \hat{\xi}_2 e^{\lambda t} \\ \vdots \\ \hat{\xi}_n e^{\lambda t} \end{bmatrix},$$

which can be rearranged to give

$$(A - \lambda I)\,\hat{\xi} = 0, \tag{6.13}$$

where the exponentials are canceled since they are never zero. Recall from linear algebra, that the values for λ that satisfy Equation 6.13 are the eigenvalues of the matrix A and the $\hat{\xi}$ that satisfy it are the corresponding eigenvectors of A. More importantly, what this shows is that solutions to Equation 6.12 are the product of the eigenvectors and exponentials of the eigenvalues of A.

6.3 Converting Ordinary Differential Equations of Order Greater than One or Systems of Equations of Order Greater than One into Systems of First Order Ordinary Differential Equations

Systems of first order differential equations may arise naturally, but they are often the result of converting equations of another form into that form. Rather than present a procedure to convert a system of ordinary differential equations to a system of ordinary first order differential equations, a rather involved example should suffice.

Example 6.3.1 Consider the system of three ordinary differential equations

$$\begin{aligned}\frac{d^3x_1}{dt^3}+\frac{1}{t}x_2 &= \cos t \\ \frac{dx_2}{dt} &= 3 \\ \frac{d^2x_3}{dt^2}+\frac{dx_3}{dt} &= x_1(t)+x_2(t).\end{aligned} \tag{6.14}$$

Note that the highest order derivative of x_1 is three, of x_2 is one and of x_3 is two. If we let

$$\begin{aligned}\xi_1 &= x_1 \\ \xi_2 &= \frac{dx_1}{dt} \\ \xi_3 &= \frac{d^2x_1}{dt^2} \\ \xi_4 &= x_2 \\ \xi_5 &= x_3 \\ \xi_6 &= \frac{dx_3}{dt}\end{aligned} \tag{6.15}$$

$$\tag{6.16}$$

then system of equations in Equation 6.14 is equivalent to

$$\frac{d}{dt}\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\\ \xi_4\\ \xi_5\\ \xi_6\end{bmatrix}=\begin{bmatrix}\xi_2\\ \xi_3\\ \cos(t)-\frac{1}{t}\xi_4\\ 3\\ \xi_5\\ \xi_1+\xi_4-\xi_6\end{bmatrix}.$$

■

The $\dot{\xi}_1$, $\dot{\xi}_2$, $\dot{\xi}_4$ and $\dot{\xi}_5$ components follow from the definitions in Equation 6.15 and the $\dot{\xi}_3$, $\dot{\xi}_4$ and $\dot{\xi}_6$ components are determined by solving the original three differential equations in Equation 6.14 for $\frac{d^3x_1}{dt^3}$, $\frac{dx_2}{dt}$ and $\frac{d^2x_3}{dt^2}$, respectively

6.4 Review of Linear Algebra

Moved to Appendix B.

6.5 Summary So Far

1. Systems of first order differential equations of the form

   $$\dot{\xi} = A\xi \qquad \xi \in \mathbb{R}^n, \qquad A \in \mathbb{R}^{n\times n}$$

   arise naturally in engineering problems with coupled elements.

2. The system is *homogeneous* since

   $$\dot{\xi} = A\xi \qquad \dot{\xi} - A\xi = 0$$

   and each homogeneous solution is of the form

   $$\xi_h(t) = \hat{\xi}^i e^{\lambda_i t}$$

   where $\hat{\xi}^i$ and λ_i is the ith eigenvector and eigenvalue of the matrix A.

3. In general A has n eigenvalue/eigenvector pairs $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\hat{\xi}^1, \ldots, \hat{\xi}^n\}$, (except possibly, as will be considered later, when A has repeated eigenvalues).

4. The general solution to $\dot{\xi} = A\xi$ is a linear combination of n homogeneous solutions

   $$\xi(t) = c_1\hat{\xi}^1 e^{\lambda_1 t} + \cdots + c_n \xi^n e^{\lambda_n t},$$

   and the coefficients c_i may be used to satisfy specified initial conditions.

6.6 Distinct Eigenvalues

The case where the matrix A has distinct eigenvalues is the easiest and will be considered first. It is basically a straight-forward application of what has been covered up to this point. First, a critically important theorem.

Theorem 6.6.1 *Let $A \in \mathbb{R}^{n\times n}$. If A has n distinct, real eigenvalues, then it has a set of n linearly independent eigenvectors.*

Proof Let $\lambda_1, \ldots, \lambda_n$ denote the distinct eigenvalues of A, *i.e.,* $\lambda_i \neq \lambda_j$ if $i \neq j$ and let $\hat{\xi}^1, \ldots, \hat{\xi}^n$ denote the corresponding eigenvectors. To show that the eigenvectors are linearly independent it suffices to show that

$$\alpha_1\hat{\xi}^1 + \alpha_2\hat{\xi}^2 + \cdots + \alpha_n\hat{\xi}^n = 0 \qquad \Longleftrightarrow \qquad \alpha_i = 0 \quad \forall i,$$

i.e., that is there is no linear combination of the eigenvectors that is zero.

First consider finding α_1 and α_2 such that

$$\alpha_1\hat{\xi}^1 + \alpha_2\hat{\xi}^2 = 0. \tag{6.17}$$

Multiply both sides of this equation by $(A - \lambda_2 I)$ (note it is a specific eigenvalue, λ_2)

$$\begin{aligned}
\alpha_1 \left(A - \lambda_2 I\right) \hat{\xi}^1 + \alpha_2 \left(A - \lambda_2 I\right) \hat{\xi}^2 &= 0 \\
\alpha_1 \left(A\hat{\xi}^1 - \lambda_2\hat{\xi}^1\right) + 0 &= 0 \\
\alpha_1 \left(\lambda_1\hat{\xi}^1 - \lambda_2\hat{\xi}^1\right) &= 0 \\
\alpha_1 \left(\lambda_1 - \lambda_2\right) \hat{\xi}^1 &= 0.
\end{aligned}$$

Since $\lambda_1 \neq \lambda_2$ and $\hat{\xi}^1 \neq 0$, then $\alpha_1 = 0$. Hence by equation 6.17, $\alpha_2 = 0$ and hence, by definition, the set $\left\{\hat{\xi}^1, \hat{\xi}^2\right\}$ is linearly independent.

Now proceed by induction and assume the set $\left\{\hat{\xi}^1, \hat{\xi}^2, \ldots, \hat{\xi}^i\right\}$ is linearly independent and consider

$$\alpha_1\hat{\xi}^1 + \alpha_2\hat{\xi}^2 + \cdots + \alpha_i\hat{\xi}^i + \alpha_{i+1}\hat{\xi}^{i+1} = 0. \tag{6.18}$$

Multiplying both sides of the equation by $(A - \lambda_{i+1} I)$ gives

$$\alpha_1 \left(\lambda_1 - \lambda_{i+1}\right) \hat{\xi}^1 + \alpha_2 \left(\lambda_2 - \lambda_{i+1}\right) \hat{\xi}^2 + \cdots + \alpha_i \left(\lambda_i - \lambda_{i+1}\right) \hat{\xi}^i + 0 = 0.$$

Since the set $\left\{\hat{\xi}^1, \hat{\xi}^2, \ldots, \hat{\xi}^i\right\}$ is linearly independent, then

$$\alpha_i = \alpha_2 = \cdots = \alpha_i = 0 \qquad \square$$

and hence by equation 6.18, $\alpha_{i+1} = 0$. Hence the set $\left\{\hat{\xi}^1, \hat{\xi}^2, \ldots, \hat{\xi}^i, \hat{\xi}^{i+1}\right\}$ is linearly independent.

Hence, by induction, the set $\left\{\hat{\xi}^1, \hat{\xi}^2, \ldots, \hat{\xi}^n\right\}$ is linearly independent.

6.6.1 Solution Technique for $\dot{\xi} = A\xi$

The general solution to $\dot{\xi} = A\xi$ is a linear combination of n homogeneous solutions

$$\xi(t) = c_1\hat{\xi}^1 e^{\lambda_1 t} + \cdots + c_n \xi^n e^{\lambda_n t},$$

and the coefficients c_i may be used to satisfy specified initial conditions. Since the eigenvectors are linearly independent, any initial condition may be satisfied with the appropriate coefficients, c_i's. In particular, for a specified $\xi(0)$

$$\begin{aligned}
\xi(0) &= c_1\hat{\xi}^1 + \cdots + c_n\hat{\xi}^n \\
&= \left[\begin{array}{ccc} \hat{\xi}^1 & \cdots & \hat{\xi}^n \end{array}\right] \left[\begin{array}{c} c_1 \\ \vdots \\ c_n \end{array}\right].
\end{aligned}$$

Thus the coefficients can most concisely be expressed as

$$\begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \hat{\xi}^1 & \cdots & \hat{\xi}^n \end{bmatrix}^{-1} \xi(0),$$

although, as illustrated in the examples below, it will usually be easiest just to solve for the coefficients using row reduction methods.

Example 6.6.2 Find the homogeneous solutions to

$$\dot{\xi} = A\xi \qquad \text{where} \qquad A = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}. \tag{6.19}$$

Aside 6.6.3 Note that the system in Equation 6.19 is exactly equivalent to the following two systems:

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$$

and

$$\begin{aligned} \dot{\xi}_1 &= \xi_1 + 2\xi_2 \\ \dot{\xi}_2 &= \xi_1. \end{aligned}$$

If this is not readily apparent by inspection, some time should be invested in verifying this fact. $\diamond$

As determined previously, the homogeneous solutions of Equation 6.19 can be computed by determining the eigenvalues and eigenvectors of A. Thus

$$\det\left(A - \lambda I\right) = \begin{vmatrix} 1-\lambda & 2 \\ 1 & -\lambda \end{vmatrix} = (1-\lambda)\,\lambda - 2 = \lambda^2 - \lambda - 2 - 0,$$

so the eigenvalues are

$$\begin{aligned} \lambda_1 &= 2 \\ \lambda_2 &= -1. \end{aligned}$$

Substituting each eigenvalue into $(A - \lambda I)\,\xi = 0$ gives

$$\begin{aligned} \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad &\Longrightarrow \qquad \hat{\xi}^1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad &\Longrightarrow \qquad \hat{\xi}^2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \end{aligned}$$

Thus

$$\begin{aligned}\xi_1(t) &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{2t} \\ \xi_2(t) &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t}\end{aligned}$$

both satisfy $\dot{\xi} = A\xi$. ■

From the above example, since each of the two solutions are homogeneous solutions, any linear combination of them also satisfies the differential equation, *i.e.,* the general solution,

$$\xi(t) = c_1 \hat{\xi}^1 e^{\lambda_1 t} + c_2 \hat{\xi}^2 e^{\lambda_2 t}$$

also satisfies $\dot{\xi} = A\xi$. If the problem were an initial value problem, then the coefficients c_1 and c_2 could be used to satisfy the initial condition.

Example 6.6.4 Returning to Example 6.6.2 determine the solution to

$$\dot{\xi} = A\xi$$

where

$$\xi(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The general solution to Equation 6.19 is

$$\xi(t) = c_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t}.$$

Substituting $t = 0$ and the initial condition gives

$$\xi(0) = c_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

which may be rearranged as

$$\begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

or in augmented matrix form as

$$\left[\begin{array}{cc|c} 2 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right].$$

Multiplying the second row by 2 and subtracting the first row from it gives

$$\left[\begin{array}{cc|c} 2 & 1 & 1 \\ 0 & -3 & -1 \end{array}\right].$$

~~which gives~~ Hence

$$\begin{aligned} c_1 &= \frac{1}{3} \\ c_2 &= \frac{1}{3}, \end{aligned}$$

so

$$\xi(t) = \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \end{bmatrix} e^{2t} + \begin{bmatrix} \frac{1}{3} \\ -\frac{1}{3} \end{bmatrix} e^{-t}$$

is the solution to the initial value problem. ■

Next are a few useful theorems that sometimes allow for some computational shortcuts. It turns out that when the matrix A is symmetric, its eigenvalues and eigenvectors have especially nice properties. First, however, we generalize the notion of a *symmetric* matrix to the complex case and the corresponding properties of a *Hermitian* matrix.

Definition 6.6.5 Hermitian Matrix Let $A \in \mathbb{C}^{n\times n}$. Let $A^* = \overline{A}^T$, *i.e.,* A^* denotes the matrix where A is transposed an all the elements are changed to their complex conjugates. A is *Hermitian* if $A = A^*$. ⋄

Note the following:

1. the notation $A \in \mathbb{C}^{n\times n}$ simply means that A is n by n with complex numbers for elements; and,

2. in particular, if A is real and symmetric, *i.e.,* $A \in \mathbb{R}^{n\times n}$ and $A = A^T$ it is Hermitian.

Theorem 6.6.6 *If $A \in \mathbb{C}^{n\times n}$ is Hermitian, i.e., $A = A^*$, then*

1. *all the eigenvalues of A are real;*

2. *A has n linearly independent eigenvectors, regardless of the multiplicity of any eigenvalue; and,*

3. *eigenvectors corresponding to different eigenvalues are orthogonal.*

Proof 1. Assume $A = A^*$. Since

$$A\hat{\xi}_i = \lambda_i\hat{\xi}_i \quad \Longrightarrow \hat{\xi}_i^* A\hat{\xi}_i = \lambda_i\hat{\xi}_i^*\hat{\xi}_i$$

the eigenvalue may be expressed as

$$\lambda_i = \frac{\hat{\xi}_i^* A\hat{\xi}_i}{\hat{\xi}_i^*\hat{\xi}_i}.$$

Note that the notation $\hat{\xi}_i^* \hat{\xi}_i$ is simply the dot product between the vector $\hat{\xi}_i$ and its complex conjugate. Then

$$\lambda_i^* = \left(\frac{\hat{\xi}_i^* A \hat{\xi}_i}{\hat{\xi}_i^* \hat{\xi}_i} \right)^* = \frac{\left(\hat{\xi}_i^* A \hat{\xi}_i\right)^*}{\left(\hat{\xi}_i^* \hat{\xi}_i\right)^*} = \frac{\hat{\xi}_i^* A^* \hat{\xi}_i}{\hat{\xi}_i^* \hat{\xi}_i} = \frac{\hat{\xi}_i^* A \hat{\xi}_i}{\hat{\xi}_i^* \hat{\xi}_i} = \lambda_i.$$

Since $\lambda_i = \lambda_i^*$, it must be real.

2. The proof of this part is beyond the scope of this book.

3. Let $\hat{\xi}_i$ be the eigenvector associated with eigenvalue λ_i and $\hat{\xi}_j$ be the eigenvector associated with eigenvalue λ_j. Since A is Hermetian,

$$\begin{aligned} \left(A\hat{\xi}_i\right)^* \hat{\xi}_j &= \hat{\xi}_i^* A^* \hat{\xi}_j \\ &= \hat{\xi}_i^* A \hat{\xi}_j \\ &= \lambda_j \hat{\xi}_i^* \hat{\xi}_j. \end{aligned}$$

But we also have

$$\left(A\hat{\xi}_i\right)^* \hat{\xi}_j = \lambda_i^* \hat{\xi}_i^* \hat{\xi}_j.$$

Since these are equal, and $\lambda_i \neq \lambda_j$, then $\hat{\xi}_i^* \hat{\xi}_j = 0$, *i.e.*, they are orthogonal. □

6.7 Complex Eigenvalues

Example 6.7.1 Again consider the mass-spring-damper system illustrated in Figure 6.1. Let

$$\begin{aligned} m_1 &= 1 \\ m_2 &= 1 \\ k_1 &= 10 \\ k_2 &= 1 \\ b_1 &= 0.1 \\ b_2 &= 0.1. \end{aligned}$$

The damping has been decreased greatly compared to the example for distinct real roots in Section 6.6, so oscillatory solutions should be expected. Substituting these values into the A matrix in Equation 6.10 gives

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -11 & -0.2 & 1 & 0.1 \\ 0 & 0 & 0 & 1 \\ 1 & 0.1 & -1 & -0.1 \end{bmatrix}$$

which has eigenvalues

$$\begin{aligned}
\lambda_1 &= -0.1093 + 3.3285i \\
\lambda_2 &= -0.1093 - 3.3285i \\
\lambda_3 &= -0.0407 + 0.9487i \\
\lambda_4 &= -0.0407 - 0.9487i,
\end{aligned}$$

and corresponding eigenvectors

$$\hat{\xi}_1 = \begin{bmatrix} -0.0094 - 0.2859i \\ 0.9527 \\ -0.0074 + 0.0287i \\ -0.0946 - 0.0278i \end{bmatrix} \qquad \hat{\xi}_2 = \begin{bmatrix} -0.0094 + 0.2859i \\ 0.9527 \\ -0.0074 - 0.0287i \\ -0.0946 + 0.0278i \end{bmatrix}$$

$$\hat{\xi}_3 = \begin{bmatrix} 0.0713 + 0.0060i \\ -0.0086 + 0.0674i \\ 0.7216 \\ -0.0294 + 0.6846i \end{bmatrix} \qquad \hat{\xi}_4 = \begin{bmatrix} 0.0713 - 0.0060i \\ -0.0086 - 0.0674i \\ 0.7216 \\ -0.0294 - 0.6846i \end{bmatrix}.$$

Observe that the eigenvalues occur in complex conjugate pairs. This should be obviously expected since eigenvalues are the roots of a polynomial. Less obvious, but probably not surprising is that the eigenvectors also occur in complex conjugate pairs. The reason this is true is given by the proof of the following. ■

Proposition 6.7.2 *If $A \in \mathbb{R}^{n \times n}$ and two eigenvalues of A are such that $\lambda_i = \overline{\lambda_j}$, then if $\hat{\xi}_i$ is the eigenvector corresponding to λ_i, $\overline{\hat{\lambda}}_i$ is an eigenvector corresponding to λ_j.*

Proof Eigenvector $\hat{\xi}_i$ satisfies

$$(A - \lambda_i I)\,\hat{\xi}_i = 0.$$

Taking the complex conjugate of both sides gives

$$\begin{aligned}
\overline{(A - \lambda_i I)\,\hat{\xi}_i} &= \overline{0} \\
\left(A - \overline{\lambda_i} I\right)\overline{\hat{\xi}_i} = 0 \\
(A - \lambda_j I)\,\overline{\hat{\xi}_i} = 0.
\end{aligned}$$

Thus we make take $\hat{\xi}_j = \overline{\hat{\xi}_i}$. □

To solve the initial value problem

$$\dot{\xi} = A\xi \qquad \xi(0) = \xi_0$$

we *may* to proceed as before and simply write the general solution

$$\xi(t) = c_1 \hat{\xi}^1 e^{\lambda_1 t} + \cdots c_1 \hat{\xi}^n e^{\lambda_n t},$$

substitute $t = 0$

$$\xi(0) = c_1\hat{\xi}^1 + \cdots c_1\hat{\xi}^n,$$

and solve for the unknown coefficients, c_i. The following example illustrates that fact. In order to make it computationally simple, however, a simple 2×2 system is considered rather than the 4×4 oscillation problem.

Example 6.7.3 Solve

$$\dot{\xi} = A\xi \qquad \xi(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

where

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

Computing the eigenvalues gives

$$\det\left(A - \lambda I\right) = \left(1 - \lambda\right)^2 + 4 = 0 \qquad \Longrightarrow \qquad \lambda = 1 \pm 2i.$$

For $\lambda_1 = 1 + 2i$

$$\begin{bmatrix} -2i & -2 \\ 2 & -2i \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Longrightarrow \qquad \hat{\xi}^1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -i \end{bmatrix}$$

and for $\lambda_2 = 1 - 2i$

$$\begin{bmatrix} 2i & -2 \\ 2 & 2i \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Longrightarrow \qquad \hat{\xi}^1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ i \end{bmatrix}$$

So the general solution is

$$\xi(t) = c_1\begin{bmatrix} 1 \\ -i \end{bmatrix}e^{(1+2i)t} + c_2\begin{bmatrix} 1 \\ i \end{bmatrix}e^{(1-2i)t}$$

and at $t = 0$,

$$\xi(0) = c_1\begin{bmatrix} 1 \\ -i \end{bmatrix} + c_2\begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Either solving for c_1 and c_2 by inverting the matrix or by eliminating one coefficient from one equation and substituting into the other gives

$$\begin{aligned} c_1 &= \frac{1}{2} + \frac{1}{2}i \\ c_2 &= \frac{1}{2} - \frac{1}{2}i. \end{aligned}$$

Finally, substituting c_1 and c_2 into the general solution gives

$$\xi(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}i \\ \frac{1}{2} - \frac{1}{2}i \end{bmatrix}e^{(1+2i)t} + \begin{bmatrix} \frac{1}{2} - \frac{1}{2}i \\ \frac{1}{2} + \frac{1}{2}i \end{bmatrix}e^{(1-2i)t}. \tag{6.20}$$

This is the correct answer, however it is somewhat dissatisfying in that it is complex; whereas, the matrix A and the initial conditions were all real. ~~Quite a bit more manipulation using Euler's formula eliminates this minor problem and yields~~ If the complex exponentials are expanded using Euler's formula, then

$$\xi(t) = \begin{bmatrix} \cos 2t - \sin 2t \\ \cos 2t + \sin 2t \end{bmatrix} e^{t} \tag{6.21}$$

is obtained. Interestingly, the complex parts of Equation 6.20 are identically zero; although, it is certainly difficult to see that without all the work to convert from Equation 6.20 to Equation 6.21. ■

The preceding example illustrates that the general solution may still be correctly expressed as a linear combination of the eigenvalues times the exponential of the corresponding eigenvectors. However,

1. the solution may not "naturally" result in a purely real expression for ξ, which is what is expected;

2. further, and perhaps arduous manipulation may be necessary to determine the form of the solution that is purely real;

3. many computations involving complex numbers, requiring four operations for multiplication and two operations for addition, are involved in computing the solution;

4. the fact that the eigenvalues and eigenvectors occur in complex conjugate pairs was not exploited at all.

In order to make the computations less burdensome, an alternative approach which is analogous to the approach in the case of second order system with complex roots is utilized. Fundamentally, the "shortcut" to this approach is based upon the conjugate nature of the eigenvalues and eigenvectors.

Consider a pair of complex conjugate eigenvalues and eigenvectors, denoted by

$$\begin{aligned} \lambda_1 &= \mu + i\omega \\ \lambda_2 &= \mu - i\omega \end{aligned}$$

and

$$\begin{aligned} \hat{\xi}^1 &= \mathbf{a} + i\mathbf{b} \\ \hat{\xi}^2 &= \mathbf{a} - i\mathbf{b}. \end{aligned}$$

Note that $\mathbf{a}$ and $\mathbf{b}$ are *vectors* in $\mathbb{R}^n$.

The general solution is

$$\xi(t) = c_1\hat{\xi}^1 e^{\lambda_1 t} + c_2\hat{\xi}^2 e^{\lambda_2 t} + \cdots .$$

Substituting for the components of λ_1, λ_2, $\hat{\xi}^1$ and $\hat{\xi}^2$ and using Euler's formula gives

$$\begin{aligned}
\xi(t) &= c_1\hat{\xi}^1 e^{\lambda_1 t} + c_2\hat{\xi}^2 e^{\lambda_2 t} + \cdots \\
&= c_1\left(\mathbf{a}+i\mathbf{b}\right)e^{(\mu+i\omega)t} + c_2\left(\mathbf{a}-i\mathbf{b}\right)e^{(\mu-i\omega)t} + \cdots \\
&= c_1\left(\mathbf{a}+i\mathbf{b}\right)e^{\mu t}\left(\cos\omega t + i\sin\omega t\right) + c_2\left(\mathbf{a}-i\mathbf{b}\right)e^{\mu t}\left(\cos\omega t - i\sin\omega t\right) + \cdots \\
&= e^{\mu t}\left[c_1\mathbf{a}\cos\omega t - c_1\mathbf{b}\sin\omega t + ic_1\mathbf{a}\sin\omega t + ic_1\mathbf{b}\cos\omega t + \right. \\
&\quad \left. c_2\mathbf{a}\cos\omega t - c_2\mathbf{b}\sin\omega t - ic_2\mathbf{b}\cos\omega t - ic_2\mathbf{a}\sin\omega t\right] + \cdots \\
&= e^{\mu t}\left[\left(c_1+c_2\right)\mathbf{a}\cos\omega t - \left(c_1+c_2\right)\mathbf{b}\sin\omega t\right] + \\
&\quad e^{\mu t} i\left[\left(c_1-c_2\right)\mathbf{a}\sin\omega t + \left(c_1-c_2\right)\mathbf{b}\cos\omega t\right] + \cdots .
\end{aligned}$$

Let

$$\begin{aligned}
k_1 &= c_1 + c_2 \\
k_2 &= i\left(c_1 - c_2\right)
\end{aligned}$$

and substituting into $\xi(t)$ gives

$$\xi(t) = k_1 e^{\mu t}\left(\mathbf{a}\cos\omega t - \mathbf{b}\sin\omega t\right) + k_2 e^{\mu t}\left(\mathbf{a}\sin\omega t + \mathbf{b}\cos\omega t\right) + \cdots . \qquad (6.22)$$

Example 6.7.4 Returning to the mass-spring-damper system in Example 6.7.1, observe that we have

$$\begin{aligned}
\mu_1 &= -0.1093 \\
\omega_1 &= 3.3285 \\
\mu_2 &= -0.0407 \\
\omega_2 &= 0.9487
\end{aligned}$$

and

$$\mathbf{a}_1 = \begin{bmatrix} -0.0094 \\ 0.9527 \\ -0.0074 \\ -0.0946 \end{bmatrix} \qquad \mathbf{b}_1 = \begin{bmatrix} -0.2859 \\ 0 \\ 0.0287 \\ -0.0278 \end{bmatrix}$$

$$\mathbf{a}_2 = \begin{bmatrix} 0.0713 \\ -0.0086 \\ 0.7216 \\ -0.0294 \end{bmatrix} \qquad \mathbf{b}_2 = \begin{bmatrix} 0.0060 \\ 0.0674 \\ 0 \\ 0.6846 \end{bmatrix}.$$

■

From equation 6.22, the general solution is of the form

$$\begin{aligned}
\xi(t) &= k_1 e^{\mu_1 t}\left(\mathbf{a}_1\cos\omega_1 t - \mathbf{b}_1\sin\omega_1 t\right) + k_2 e^{\mu_1 t}\left(\mathbf{a}_1\sin\omega_1 t + \mathbf{b}_1\cos\omega_1 t\right) \\
&+ k_3 e^{\mu_2 t}\left(\mathbf{a}_2\cos\omega_2 t - \mathbf{b}_2\sin\omega_2 t\right) + k_4 e^{\mu_2 t}\left(\mathbf{a}_2\sin\omega_2 t + \mathbf{b}_2\cos\omega_2 t\right),
\end{aligned}$$

or substituting all the numerical values

$$\begin{aligned}
\xi(t) &= k_1 e^{-0.1093t}\left(\begin{bmatrix} -0.0094 \\ 0.9527 \\ -0.0074 \\ -0.0946 \end{bmatrix}\cos 3.3285t - \begin{bmatrix} -0.2859 \\ 0 \\ 0.0287 \\ -0.0278 \end{bmatrix}\sin 3.3285t\right) \\
&+ k_2 e^{-0.1093t}\left(\begin{bmatrix} -0.0094 \\ 0.9527 \\ -0.0074 \\ -0.0946 \end{bmatrix}\sin 3.3285t + \begin{bmatrix} -0.2859 \\ 0 \\ 0.0287 \\ -0.0278 \end{bmatrix}\cos 3.3285t\right) \\
&+ k_3 e^{-0.0407t}\left(\begin{bmatrix} 0.0713 \\ -0.0086 \\ 0.7216 \\ -0.0294 \end{bmatrix}\cos 0.9487t - \begin{bmatrix} 0.0060 \\ 0.0674 \\ 0 \\ 0.6846 \end{bmatrix}\sin 0.9487t\right) \\
&+ k_4 e^{-0.0407t}\left(\begin{bmatrix} 0.0713 \\ -0.0086 \\ 0.7216 \\ -0.0294 \end{bmatrix}\sin 0.9487t + \begin{bmatrix} 0.0060 \\ 0.0674 \\ 0 \\ 0.6846 \end{bmatrix}\cos 0.9487t\right).
\end{aligned}$$

Example 6.7.5 Returning to Example 6.7.3,

$$\lambda_1 = 1 + 2i$$

and

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ -i \end{bmatrix}.$$

Hence

$$\begin{aligned} \mu &= 1 \\ \omega &= 1 \end{aligned}$$

and

$$\begin{aligned} \mathbf{a} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ \mathbf{b} &= \begin{bmatrix} 0 \\ -1 \end{bmatrix} \end{aligned}$$

Substituting into Equation 6.22 gives

$$\begin{aligned} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} &= k_1 e^t\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\cos 2t - \begin{bmatrix} 0 \\ -1 \end{bmatrix}\sin 2t\right) + \\ & k_2\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\sin 2t + \begin{bmatrix} 0 \\ -1 \end{bmatrix}\cos 2t\right). \end{aligned}$$

The initial condition is

$$\xi(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Substituting $t = 0$ into the solution and equating it to the intial condition gives

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} = k_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + k_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

which gives

$$\begin{aligned} k_1 &= 1 \\ k_2 &= -1. \end{aligned}$$

Hence,

$$\begin{aligned} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} &= e^t \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} \cos 2t - \begin{bmatrix} 0 \\ -1 \end{bmatrix} \sin 2t \right) - \\ &\quad \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} \sin 2t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} \cos 2t \right) \\ &= e^t \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cos 2t + \begin{bmatrix} -1 \\ 1 \end{bmatrix} \sin 2t \right) \end{aligned}$$

which is the same as Equation 6.21. ■

This next example contains one real eigenvalue and one complex conjugate pair of eigenvalues.

Example 6.7.6 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -7 & 0 & 8 \\ 0 & -2 & 0 \\ -4 & 0 & 1 \end{bmatrix}.$$

Computing

$$\det(A - \lambda I) = \begin{vmatrix} -7-\lambda & 0 & 8 \\ 0 & -2-\lambda & 0 \\ -4 & 0 & 1-\lambda \end{vmatrix}$$

by a cofactor expansion across the second row gives

$$\begin{aligned} -1\,(-2-\lambda)\,[(-7-\lambda)\,(1-\lambda) + 32] &= \\ (2+\lambda)\left(\lambda^2 + 6\lambda + 25\right) &= 0. \end{aligned}$$

Hence,

$$\lambda_1 = -2$$

and

$$\begin{aligned}\lambda &= \frac{-6 \pm \sqrt{36-100}}{2} \\ &= -3 \pm 4i.\end{aligned}$$

For $\lambda_1 = -2$

$$(A + 2I)\,\hat{\xi} = 0$$

is computed by

$$\left[\begin{array}{ccc|c} -5 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 \\ -4 & 0 & 3 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -5 & 0 & 8 & 0 \\ 0 & 0 & -\frac{17}{5} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

which gives

$$\hat{\xi}^1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

For $\lambda_2 = -3 + 4i$,

$$(A + (3-4i)\,I)\,\hat{\xi} = 0$$

is computed by

$$\left[\begin{array}{ccc|c} -4-4i & 0 & 8 & 0 \\ 0 & 1-4i & 0 & 0 \\ -4 & 0 & 4-4i & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -4-4i & 0 & 8 & 0 \\ 0 & 1-4i & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

which was obtained by dividing the first row by $1+i$ and subtracting the resut from the third row. If we let $\hat{\xi}^2_3 = 1$, then

$$\hat{\xi}^2 = \begin{bmatrix} 1-i \\ 0 \\ 1 \end{bmatrix}.$$

Since both the eigenvalues and eigenvectors must occur in complex conjugate pairs, for $\lambda_3 = -3 - 4i$

$$\hat{\xi}^3 = \begin{bmatrix} 1+i \\ 0 \\ 1 \end{bmatrix}.$$

Using the second eigenvalue $\mu = -3$, $\omega = 4$,

$$\mathbf{a} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}.$$

■

Hence,

$$\begin{aligned}\xi(t) &= c_1e^{-2t}\begin{bmatrix}0\\1\\0\end{bmatrix} + c_2e^{-3t}\left(\begin{bmatrix}1\\0\\1\end{bmatrix}\cos 4t - \begin{bmatrix}-1\\0\\0\end{bmatrix}\sin 4t\right) + \\ &\quad c_3e^{-3t}\left(\begin{bmatrix}1\\0\\1\end{bmatrix}\sin 4t + \begin{bmatrix}-1\\0\\0\end{bmatrix}\cos 4t\right).\end{aligned}$$

6.8 Repeated Eigenvalues

The case where some of the eigenvalues are repeated is the most complicated. This is because when there are repeated eigenvalues, there may or may not be a complete set of linearly independent eigenvectors associated with the repeated eigenvalue. The next set of examples illustrates this fact.

Example 6.8.1 Consider $\dot{\xi} = A\xi$ where

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

Computing the eigenvalues gives

$$(2-\lambda)^2 = 0 \qquad \Longrightarrow \qquad \lambda = 2.$$

Computing the eigenvectors,

$$\begin{bmatrix}0 & 1\\0 & 0\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \qquad \Longrightarrow \qquad \hat{\xi} = \begin{bmatrix}1\\0\end{bmatrix}.$$ ■

In the preceding example, the eigenvalue $\lambda = 2$ was repeated. It may not be surprising that there also is only one eigenvector, $\hat{\xi}$ as well. However, things are not so simple. Consider the following example.

Example 6.8.2 Consider $\dot{\xi} = A\xi$ where

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

Computing the eigenvalues gives

$$(2-\lambda)^2 = 0 \qquad \Longrightarrow \qquad \lambda = 2,$$

which is exactly the same as before. Now computing the eigenvectors,

$$\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}.$$

In this case, however, we have that

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \hat{\xi}^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

both satisfy the eigenvector equation and are linearly independent. ■

These two examples illustrate the fact that when there are n repeated eigenvalues, there may or may not be n linearly independent eigenvectors. This is problematic in that to use the approach utilized so far to solve $\dot{\xi} = A\xi$ we need n linearly independent eigenvectors in order to obtain a general solution.

First we address the practical computational matter of determining how many linearly independent eigenvectors are associated with a repeated eigenvalue. Then we delineate the solution techniques for each case.

6.8.1 Geometric and Algebraic Multiplicities

The number of times that an eigenvalue is repeated is called its *algebraic multiplicity.* Similarly, the number of linearly independent eigenvectors associated with an eigenvalue is called its *geometric multiplicity.* Clearly, the former is an algebraic concept and the latter a geometric one as is clear from the following more general mathematical definitions of the two terms.

Definition 6.8.3 (Algebraic Multiplicity) Let $A \in \mathbb{R}^{n\times n}$ and let

$$\det (A - \lambda I) = \sum_{i=1}^{m} (\lambda - \lambda_i)^{k_i}$$

where each λ_i is distinct. Note that $\sum_{i=1}^{m} k_i = n$. The number k_i is the *algebraic multiplicity* of eigenvalue λ_i. ⋄

Example 6.8.4 Returning to example B.1.23, for

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

we determined that

$$\det (A - 2I) = (1 - \lambda)(2 - \lambda)^3 .$$

Hence the algebraic multiplicity of $\lambda = 1$ is one and the algebraic multiplicity of $\lambda = 2$ is three. ■

Definition 6.8.5 (Geometric Multiplicity) Let $A \in \mathbb{R}^{n\times n}$. The dimension of the null space of $(A - \lambda_i I)$ is the *geometric multiplicity* of eigenvalue λ_i. ⋄

The definition of geometric multiplicity should make sense. Since the definition of an eigenvector is a nonzero vector, $\hat{\xi}$ satisfying

$$(A - \lambda I)\,\hat{\xi} = 0,$$

and the null space of a matrix is simply all the vectors that, when multiplied into the matrix produce the zero vector, the number of linearly independent vectors that produce the zero vector is simply the dimension of the null space.

First we will consider a matrix with distinct eigenvalues to illustrate the concept of the dimension of the null space of $(A - \lambda I)$ being the number of linearly independent eigenvectors associated with an eigenvalue as well as the simple procedural aspect of computing it.

Example 6.8.6 Determine all the linearly independent eigenvectors of

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & -2 & 4 \end{bmatrix}.$$

The characteristic equation is

$$\begin{vmatrix} (1-\lambda) & 0 & 1 \\ 0 & (1-\lambda) & 1 \\ 0 & -2 & (4-\lambda) \end{vmatrix} = \lambda^3 - 6\lambda^2 + 11\lambda - 6 = 0,$$

so the eigenvalues are

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_2 &= 2 \\ \lambda_3 &= 3. \end{aligned}$$

Since the eigenvalues are distinct, by Theorem 6.6.1, each should have one linearly independent eigenvector associated with it and $\dim\left(\mathcal{N}\left(A - \lambda_i I\right)\right) = 1$ for each λ_i.

In detail, for $\lambda_1 = 1$ the associated eigenvalue satisfies

$$(A - \lambda_1 I)\,\hat{\xi}^1 = (A - I)\,\hat{\xi}^1 = 0.$$

The augmented matrix is

$$\left[\begin{array}{ccc|c} 1-\lambda & 0 & 1 & 0 \\ 0 & 1-\lambda & 1 & 0 \\ 0 & -2 & 4-\lambda & 0 \end{array}\right]. \tag{6.23}$$

Substituting $\lambda_1 = 1$ and making a couple elementary row manipulations yields

$$\left[\begin{array}{ccc|c} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -2 & 3 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} 0 & -2 & 3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} 0 & -2 & 3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

The last augmented matrix has one row of zeros, indicating that the dimension of its null space is one, so there is one linearly independent eigenvector associated with $\lambda_1 = 1$.. From the second row, the third component of $\hat{\xi}^1$ clearly must be zero. Using this fact and noting the first row indicates that the second component must also be zero. Finally, the first component of $\hat{\xi}^1$ is clearly arbitrary. Thus, the eigenvector must be

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Similarly, substituting $\lambda_2 = 2$ into Equation 6.23 gives

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -2 & 2 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Picking the third component of $\hat{\xi}^2$ to be one, we have

$$\hat{\xi}^2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Finally, for $\lambda_3 = 3$

$$\left[\begin{array}{ccc|c} -2 & 0 & 1 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & -2 & 1 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -2 & 0 & 1 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

This time picking the third component of $\hat{\xi}^3$ to be 2 gives

$$\hat{\xi}^3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}.$$

■

Now consider an example with repeated eigenvalues.

Example 6.8.7 Determine the eigenvalues and eigenvectors of

$$A = \begin{bmatrix} 0 & 1 & 1 \\ -4 & 5 & 1 \\ -5 & 1 & 5 \end{bmatrix}.$$

The characteristic equation is

$$\lambda^3 - 10\lambda^2 + 32\lambda - 32 = 0,$$

so the eigenvalues are

$$\begin{aligned} \lambda_1 &= 2 \\ \lambda_2 &= 4 \\ \lambda_3 &= 4. \end{aligned}$$

For $\lambda_1 = 2$

$$\left[\begin{array}{ccc|c} -2 & 1 & 1 & 0 \\ -4 & 3 & 1 & 0 \\ -4 & 1 & 3 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -2 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -2 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Since there is one row of zeros, there is one linearly independent eigenvalue associated with $\lambda_1 = 2$, which is expected since it is not repeated. Picking the third component of $\hat{\xi}^1$ to be one,

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Now, for $\lambda_2 = 4$

$$\left[\begin{array}{ccc|c} -4 & 1 & 1 & 0 \\ -4 & 1 & 1 & 0 \\ -4 & 1 & 1 & 0 \end{array}\right] \iff \left[\begin{array}{ccc|c} -4 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Since there are two rows of zeros, there are two linearly independent eigenvectors associated with $\lambda_2 = 4$. Picking the third component of $\hat{\xi}^2$ to be 4 and the second component to be zero, we have

$$\hat{\xi}^2 = \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix}.$$

Since there are two rows of zeros, we can find another solution to the equations. To determine one, we pick another combination of variables with the only restriction that it cannot be a scaled version of two of the components of $\hat{\xi}^2$. Picking the third component to be zero and the second component to be 4 gives

$$\hat{\xi}^3 = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}.$$

The fact that there were two rows of zeros in upper triangular form of the augmented matrix indicates that the dimension of the null space of $(A - 4I)$ was two. Thus, we were able to determine two linearly independent eigenvectors associated with the repeated eigenvalue. ■

Finally, just to complete the picture, the following is an example of an eigenvalue with algebraic multiplicity two but a geometric multiplicity of one.

Example 6.8.8 Returning to the matrix from Example 6.8.1 with

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

we computed previously that $\lambda = 2$ was the only eigenvalue and that it had an algebraic multiplicity of two. Constructing the augmented matrix for $A - 2I$ gives

$$\left[\begin{array}{cc|c} 0 & 1 & 1 \\ 0 & 0 & 0 \end{array}\right].$$

Since there is one row of zeros, the geometric multiplicity is one. Clearly the first component of the eigenvector is arbitrary and the second component must be zero. Thus, for example

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

■

Finally, after this rather extensive detour into the realm of the nature of repeated eigenvalues and the computational details of computing the associated eigenvectors, we return to the main task at hand which is to solve $\dot{\xi} = A\xi$.

6.8.2 Homogeneous Solutions with Repeated Eigenvalues

Equal Algebraic and Geometric Multiplicities

This is the case for which to hope because the solution technique is identical to the case of distinct eigenvalues. Even if there are repeated eigenvalues, the general solution is simply

$$\xi(t) = c_1 \hat{\xi}^1 e^{\lambda_1 t} + c_2 \hat{\xi}^2 e^{\lambda_2 t} + \cdots + c_n \hat{\xi}^n e^{\lambda_n t}.$$

This is, in fact, the general solution. Since the set of eigenvectors is linearly independent, it will always be possible to solve for the coefficients for a specified initial condition regardless of the fact that some of the eigenvalues are repeated.

Repeated Complex Eigenvalues

The statement immediately preceding this is still correct, even if there are complex conjugate eigenvalues and even if some of the repeated eigenvalues are complex conjugates. In the first case where the repeated eigenvalues are real, the more convenient form of the solution will be to simply convert the complex conjugate eigenvalue and eigenvector pairs to the real and imaginary components and express the two homogeneous solutions corresponding to the complex conjugate pair in terms of the real functions given in Equation 6.22.

Algebraic Multiplicity Greater than the Geometric Multiplicity

The case where the geometric multiplicity of an eigenvalue is less than its algebraic multiplicity is much more interesting, but unfortunately, requires a bit more work. In this case, if we simply compute eigenvectors, we will have a set of homogeneous solutions of the form

$$\xi_h(t) = \hat{\xi}^i e^{\lambda_i t},$$

but we will not have n linearly independent eigenvalues, so the partial general solution will be of the form

$$\xi(t) = c_1\hat{\xi}^1 e^{\lambda_1 t} + c_2\hat{\xi}^2 e^{\lambda_2 t} + \cdots + c_m\hat{\xi}^m e^{\lambda_m t},$$

where $m < n$. In this case, it will not be possible to compute coefficients, c_i to satisfy any set of initial conditions since there is not a full set of linearly independent eigenvectors.

Recall from Chapter 3 that in the case of repeated roots, the approach was to multiply the one homogeneous solution by the independent variable, t and add it to the first solution. The following two examples illustrate that fact, but also then goes to make a connection to the matrix approach that is the subject of this chapter.

Example 6.8.9 Find the general solution to

$$\ddot{x} + 4\dot{x} + 4x = 0. \tag{6.24}$$

Assuming $x(t) = e^{\lambda t}$ and substituting gives

$$\begin{aligned} \lambda^2 + 4\lambda + 4 &= 0 \qquad (6.25)\\ (\lambda + 2)^2 &= 0. \end{aligned}$$

So, $\lambda = 2$ is the solution. Hence, $x_h(t) = e^{-2t}$ is a homogeneous solution. Since there is no other root to the characteristic equation, the approach (which was fully detailed in Chapter 3) is to assume a second homogeneous solution of the form $x_h(t) = te^{-2t}$. The fact that this a second homogeneous solution can be verified by substituting it into Equation 6.24 and the fact that it is linearly independent can be verified by computing the Wronskian. Thus the general solution to Equation 6.24 is

$$x(t) = c_1 e^{-2t} + c_2 te^{-2t}. \tag{6.26}$$

■

Example 6.8.10 Consider the same equation as in Equation 6.24, but first convert it into a system of two first order equations. The equivalent system is

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix} x \qquad \text{where} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix}.$$

Computing the eigenvalues for the matrix in the preceding equation gives

$$\begin{vmatrix} -\lambda & 1 \\ -4 & -4-\lambda \end{vmatrix} = \lambda^2 + 4\lambda + 4 = (\lambda + 2)^2 = 0.$$

It is no coincidence that the characteristic equation for the eigenvalue problem is exactly the same as Equation 6.25. Thus, the only distinction is one of nomenclature: there are "repeated eigenvalues" instead of "repeated roots." Now computing the eigenvectors corresponding to $\lambda_1 = -2$ gives

$$\left[\begin{array}{cc|c} 2 & 1 & 0 \\ -4 & -2 & 0 \end{array}\right] \quad \Longleftrightarrow \quad \left[\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right].$$

Thus, there is one linearly independent eigenvector,

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

The goal is to obviously construct a solution that is equivalent to the general solution in Equation 6.26. Differentiating Equation 6.26 gives

$$\dot{x}(t) = -2c_1 e^{-2t} + c_2 e^{-2t} - 2c_2 t e^{-2t},$$

or in vector form

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} x \\ \dot{x} \end{bmatrix} &= \frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \\ &= c_1 \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-2t} + c_2 \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{-2t} + t \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-2t} \right) \\ &= c_1 \hat{\xi}^1 e^{\lambda_1 t} + c_2 \left(\hat{\xi}^2 e^{\lambda_1 t} + t \hat{\xi}^1 e^{\lambda_1 t} \right). \end{aligned}$$

Clearly, in the notation of the last line in the above equation, $\hat{\xi}_1^1$ is simply the eigenvector that we already computed. The question is how to compute the other vector, $\hat{\xi}_1^2$. Note that the superscripts for the $\hat{\xi}$'s are indices, not powers.

Recall, that the whole business regarding eigenvalues and eigenvectors came about by simply *assuming* solutions of the form $\xi(t) = \hat{\xi} e^{\lambda t}$. Substituting this into $\dot{\xi} = A\xi$ then indicated that $\hat{\xi}$ had to be an eigenvector and λ had to be an eigenvalue. The approach now is pretty obvious: substitute the assumed form of the second homogeneous solution

$$\xi_h(t) = \left(\hat{\xi}^2 + t\hat{\xi}^1 \right) e^{\lambda t}$$

to verify first, that $\hat{\xi}_1^1$ indeed satisfies the eigenvector equation (so that the fact that they are the same in this example is not a coincidence) and second, to determine what sort of equation $\hat{\xi}_1^2$ must satisfy. Differentiating and substituting gives

$$\lambda \left(\hat{\xi}^2 + t\hat{\xi}^1 \right) e^{\lambda t} + \hat{\xi}^1 e^{\lambda t} = A \left(\hat{\xi}^2 + t\hat{\xi}^1 \right) e^{\lambda t}.$$

Since this must hold for all t, the coefficients of the different powers of t must be equal. Therefore, collecting terms multiplying the same powers of t gives

$$\begin{aligned} t^0 &: \lambda \left(\hat{\xi}^2 + \hat{\xi}^1 \right) e^{\lambda t} = A\hat{\xi}^2 e^{\lambda t} \\ t^1 &: \lambda \hat{\xi}^1 e^{\lambda t} = A\hat{\xi}^1 e^{\lambda t}. \end{aligned}$$

Since $e^{\lambda t}$ is never zero we have the following two equations

$$\begin{aligned}(A-\lambda I)\,\hat{\xi}^1 &= 0\\ (A-\lambda I)\,\hat{\xi}^2 &= \hat{\xi}^1.\end{aligned}$$

The first equation has already been solved, so

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

For the second equation we have

$$\left[\begin{array}{cc|c} 2 & 1 & 1 \\ -4 & -2 & -2 \end{array}\right] \quad \Longleftrightarrow \quad \left[\begin{array}{cc|c} 2 & 1 & 1 \\ 0 & 0 & 0 \end{array}\right].$$

Clearly, as with eigenvectors, the solution is determined only up to an arbitrary scaling constant. In this case, clearly, the vector

$$\hat{\xi}^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

satisfies the equation for $\hat{\xi}^2$. ■

The task now is to generalize the approach used in the above example to systems of n equations where the multiplicity of a repeated eigenvalue may be greater than 2.

Now consider the general case of

$$\dot{\xi} = A\xi \qquad A \in \mathbb{R}^{n\times n},$$

and assume that the algebraic multiplicity of eigenvalue λ_i is m but that the geometric multiplicity is less than m. Motivated by the above example, clearly the approach is to multiply exponential solutions by t to obtain additional linearly independent solutions. In the example, since the system was second order, the highest power of t in the general solution was 1; however, in the case where the algebraic multiplicity is greater than 2, additional powers of t may be necessary. Therefore, let us propose the following homogeneous solution corresponding to eigenvalue λ_i with algebraic multiplicity m

$$\xi(t) = \left(\hat{\xi}^m + t\hat{\xi}^{m-1} + \frac{t^2}{2!}\hat{\xi}^{m-2} + \cdots + \frac{t^{m-1}}{(m-1)!}\hat{\xi}^1\right) e^{\lambda_i t}. \tag{6.27}$$

Differentiating this proposed solution gives

$$\begin{aligned}\dot{\xi}(t) &= \lambda_i \left(\hat{\xi}^m + t\hat{\xi}^{m-1} + \frac{t^2}{2!}\hat{\xi}^{m-2} + \cdots + \frac{t^{m-1}}{(m-1)!}\hat{\xi}^1\right) e^{\lambda_i t}\\ &+ \left(\hat{\xi}^{m-1} + t\hat{\xi}^{m-2} + \cdots + \frac{t^{m-2}}{(m-2)!}\hat{\xi}^1\right) e^{\lambda_i t}. \end{aligned}\tag{6.28}$$

Also,

$$A\xi(t) = A\left(\hat{\xi}^m + t\hat{\xi}^{m-1} + t^2\hat{\xi}^{m-2} + \cdots + t^{m-1}\hat{\xi}^1\right)e^{\lambda_i t}. \tag{6.29}$$

Since $e^{\lambda_i t}$ is never zero it can be canceled from both equations and since $\dot{\xi} = At$ must hold for all t, each the terms for each power of t in Equations 6.28 and 6.29, which gives

$$\begin{array}{rcl} t^0 & : & \lambda_i\hat{\xi}^m + \hat{\xi}^{m-1} = A\hat{\xi}^m \\ t^1 & : & \lambda_i\hat{\xi}^{m-1} + \hat{\xi}^{m-2} = A\hat{\xi}^{m-1} \\ t^2 & : & \lambda_i\hat{\xi}^{m-2} + \hat{\xi}^{m-2} = A\hat{\xi}^{m-2} \\ \vdots & & \vdots \\ t^{m-1} & : & \lambda_i\hat{\xi}^1 = A\hat{\xi}^1. \end{array}$$

so, the following sequence is obtained

$$\begin{array}{rcl} (A - \lambda_i I)\,\hat{\xi}^1 & = & 0 \\ (A - \lambda_i I)\,\hat{\xi}^2 & = & \hat{\xi}^1 \\ (A - \lambda_i I)\,\hat{\xi}^3 & = & \hat{\xi}^2 \\ & \vdots & \\ (A - \lambda_i I)\,\hat{\xi}^m & = & \hat{\xi}^{m-1} \end{array} \tag{6.30}$$

The first equation is simply the equation for a regular eigenvalue. The vectors $\hat{\xi}^2$ through $\hat{\xi}^m$ are called *generalized eigenvectors* and are determined by sequentially solving the second through mth equations.

Note that if the second line of Equation 6.30 is multiplied on the left by $(A - \lambda_i I)$ then

$$(A - \lambda_i I)(A - \lambda_i I)\,\hat{\xi}^2 = (A - \lambda_i I)\,\hat{\xi}^1,$$

but since

$$(A - \lambda_i I)\,\hat{\xi}^1 = 0$$

then

$$(A - \lambda_i I)^2\,\hat{\xi}^2 = 0.$$

Similarly, multiplying the jth line in Equation 6.30 by $(A - \lambda_i I)^j$ where $1 < j < m$ gives

$$(A - \lambda_i I)^j\,\hat{\xi}^j = 0.$$

Further note that

$$(A - \lambda_i I)^m\,\hat{\xi}^j = (A - \lambda_i I)^{m-j}\,(A - \lambda_i I)^j\,\hat{\xi}^j = 0.$$

Hence, *all* the eigenvectors and generalized eigenvectors associated with λ_i are in the null space of $(A - \lambda_i I)^m$, which motivates the following definition.

Definition 6.8.11 (Generalized Eigenspace) The null space of $(A - \lambda_i I)^m$ is the *generalized eigenspace of A associated with λ_i.* ⋄

The following theorem assures us that the dimension of the generalized eigenspace associated with λ_i is the same as the algebraic multiplicity of λ_i. This fact is necessary in order to ensure that enough generalized eigenvectors exist to generate a full set of linearly independent homogeneous solutions to construct a general solution.

Theorem 6.8.12 *The dimension of the generalized eigenspace of A associated with λ_i is equal to the algebraic multiplicity of the eigenvalue λ_i, i.e., if the algebraic multiplicity of the eigenvalue λ_i is m, then*

$$\dim\left(\mathcal{N}\left(A-\lambda_i I\right)^m\right) = m.$$

Proof The reader is referred to [7] and [9]. □

The following theorem gives the form of the homogeneous solution for *any* vector in generalized eigenspace of λ_i.

Theorem 6.8.13 *For $A \in \mathbb{R}^{n\times n}$ and λ_i an eigenvector of A with algebraic multiplicity m, if*

$$\left(A-\lambda_i I\right)^m \hat{\xi} = 0,$$

then

$$\xi(t) = \left(\hat{\xi} + t\left(A-\lambda_i I\right)\hat{\xi} + \frac{t^2}{2!}\left(A-\lambda_i I\right)^2\hat{\xi} + \cdots + \frac{t^{m-1}}{(m-1)!}\left(A-\lambda_i I\right)^{m-1}\hat{\xi}\right)e^{\lambda_i t} \tag{6.31}$$

satisfies

$$\dot{\xi} = A\xi.$$

Proof This is by direct computation. Simply differentiate $\xi_h(t)$ and substitute into $\dot{\xi} = A\xi$. □

So, finally we have the following solution technique for $\dot{\xi} = A\xi$, for $A \in \mathbb{R}^{n\times n}$ where λ_i has an algebraic multiplicity of m.

1. For the non-repeated ed eigenvalues, λ_j, the corresponding homogeneous solution is $\xi_h(t) = \hat{\xi}^{j}_{\underline{j}} e^{\lambda_j t}$. If two of these eigenvalues are a complex conjugate pair, then converting the homogeneous solution to sines and cosines as outlined in Section 6.7 is preferable.

2. For each repeated λ_i

   (a) Find all m $\hat{\xi}$ in the generalized eigenspace of λ_i, *i.e.*,

   $$\left(A-\lambda_i I\right)^m \hat{\xi} = 0.$$

   These $\hat{\xi}$ may be regular eigenvectors, generalized eigenvectors or linear combinations thereof.

(b) The homogeneous solution corresponding to each $\hat{\xi}^{i}_{\underline{i}}$ is

$$\xi(t) = \left(\hat{\xi} + t\,(A - \lambda_i I)\,\hat{\xi} + \frac{t^2}{2!}\,(A - \lambda_i I)^2\,\hat{\xi} + \cdots + \frac{t^{m-1}}{(m-1)!}\,(A - \lambda_i I)^{m-1}\,\hat{\xi} \right) e^{\lambda_i t}.$$

A few examples will help illustrate the approach.

Example 6.8.14 Determine the general solution to $\dot{\xi} = A\xi$ where

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

Since the matrix is triangular, the eigenvalues are the values along the diagonal. Thus

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_2 &= 2 \\ \lambda_3 &= 2 \\ \lambda_4 &= 2. \end{aligned}$$

Thus, $\lambda = 2$ is an eigenvalue with algebraic multiplicity of 4. For $\lambda_1 = 1$, the eigenvector is

$$(A - \lambda_1 I)\,\hat{\xi}^1 = 0 \quad \Longleftrightarrow \quad \left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 2 & 0 \end{array}\right] \quad \Longleftrightarrow \quad \hat{\xi}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

For $\lambda_2 = \lambda_3 = \lambda_4 = 2$ we need to find all three vectors that satisfy $(A - 2I)^3\,\hat{\xi} = 0$, so we must compute

$$(A - 2I)^3 = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Hence we need to find three solutions to

$$\left[\begin{array}{cccc|c} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

The free components are obviously the second, third and fourth components. Hence we have

$$\hat{\xi}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\xi}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \hat{\xi}^4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus, the general solution is

$$\begin{aligned}
\xi(t) &= c_1\hat{\xi}_1 e^{\lambda_1 t} \\
&+ c_2\left(\hat{\xi}^2 + t\,(A-2I)\,\hat{\xi}^2 + \frac{t^2}{2!}\,(A-2I)^2\,\hat{\xi}^2\right)e^{\lambda_2 t} \\
&+ c_3\left(\hat{\xi}^3 + t\,(A-2I)\,\hat{\xi}^3 + \frac{t^2}{2!}\,(A-2I)^2\,\hat{\xi}^3\right)e^{\lambda_2 t} \\
&+ c_4\left(\hat{\xi}^4 + t\,(A-2I)\,\hat{\xi}^4 + \frac{t^2}{2!}\,(A-2I)^2\,\hat{\xi}^4\right)e^{\lambda_2 t}.
\end{aligned}$$

Since we need them in the answer, observe that

$$(A-2I) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad (A-2I)^2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and hence

$$\begin{aligned}
(A-2I)\,\hat{\xi}^2 &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & \qquad (A-2I)^2\,\hat{\xi}^2 &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
(A-2I)\,\hat{\xi}^3 &= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} & \qquad (A-2I)^2\,\hat{\xi}^3 &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
(A-2I)\,\hat{\xi}^4 &= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \qquad (A-2I)^2\,\hat{\xi}^4 &= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.
\end{aligned}$$

So, finally, the general solution is

$$\begin{aligned}
&= c_1\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} e^t + c_2\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_3\left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)e^{2t} + \\
&c_4\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + t\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \frac{t^2}{2}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)e^{2t}.
\end{aligned}$$

That this is a solution may be verified by directly substituting this into the original differential equation. ■

Example 6.8.15 Determine the general solution to

$$\dot{\xi} = a\xi$$

where

$$A = \left[\begin{array}{ccc} 3 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{array} \right].$$

Computing

$$\left| \begin{array}{ccc} 3-\lambda & -1 & 0 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{array} \right| = 0$$

using a cofactor expansion across the third row gives

$$\begin{aligned} (2-\lambda)\left[(3-\lambda)(1-\lambda)+1\right] &= \\ (2-\lambda)\left[\lambda^2 - 4\lambda + 4\right] &= \\ (2-\lambda)\left[(2-\lambda)^2\right] &= 0. \end{aligned}$$

Hence, $\lambda = 2$ has an algebraic multiplicity of three.

Next we must determine the vectors that span the null space of $(A - \lambda I)^3$. Substituting $\lambda = 2$ gives

$$(A - 2I) = \left[\begin{array}{ccc} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{array} \right]$$

and an simple calculation shows that

$$(A - 2I)^2 = (A - 2I)^3 = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right],$$

so we may choose any three vectors that span $\mathbb{R}^3$. Just for fun, we will choose

$$\hat{\xi}^1 = \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right] \quad \hat{\xi}^2 = \left[\begin{array}{c} 1 \\ 1 \\ 0 \end{array} \right] \quad \hat{\xi}^3 = \left[\begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right].$$

■

All that is left to do is to substitute into Equation 6.31, which gives

$$\begin{aligned}
\xi(t) &= c_1 e^{2t}\left(\begin{bmatrix}1\\0\\0\end{bmatrix} + t\begin{bmatrix}1 & -1 & 0\\1 & -1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\0\\0\end{bmatrix}\right) + \\
& \quad c_2 e^{2t}\left(\begin{bmatrix}1\\1\\0\end{bmatrix} + t\begin{bmatrix}1 & -1 & 0\\1 & -1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\1\\0\end{bmatrix}\right) + \\
& \quad c_3 e^{2t}\left(\begin{bmatrix}1\\1\\1\end{bmatrix} + t\begin{bmatrix}1 & -1 & 0\\1 & -1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}\right) + \\
&= c_1 e^{2t}\left(\begin{bmatrix}1\\0\\0\end{bmatrix} + t\begin{bmatrix}1\\1\\0\end{bmatrix}\right) + c_2 e^{2t}\begin{bmatrix}1\\1\\0\end{bmatrix} + c_3 e^{2t}\begin{bmatrix}1\\1\\1\end{bmatrix}.
\end{aligned}$$

Just to complete the picture, let us repeat the previous example, but choose the usual basis for $\mathbb{R}^3$ instead.

Example 6.8.16 Returning to Example 6.8.15, choose

$$\hat{\xi}^1 = \begin{bmatrix}1\\0\\0\end{bmatrix} \quad \hat{\xi}^2 = \begin{bmatrix}0\\1\\0\end{bmatrix} \quad \hat{\xi}^3 = \begin{bmatrix}0\\0\\1\end{bmatrix}.$$

Substituting into Equation 6.31 gives

$$\begin{aligned}
\xi(t) &= c_1 e^{2t}\left(\begin{bmatrix}1\\0\\0\end{bmatrix} + t\begin{bmatrix}1 & -1 & 0\\1 & -1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\0\\0\end{bmatrix}\right) + \\
& \quad c_2 e^{2t}\left(\begin{bmatrix}0\\1\\0\end{bmatrix} + t\begin{bmatrix}1 & -1 & 0\\1 & -1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\1\\0\end{bmatrix}\right) + \\
& \quad c_3 e^{2t}\left(\begin{bmatrix}0\\0\\1\end{bmatrix} + t\begin{bmatrix}1 & -1 & 0\\1 & -1 & 0\\0 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\0\\1\end{bmatrix}\right) + \\
&= c_1 e^{2t}\left(\begin{bmatrix}1\\0\\0\end{bmatrix} + t\begin{bmatrix}1\\1\\0\end{bmatrix}\right) + c_1 e^{2t}\left(\begin{bmatrix}0\\1\\0\end{bmatrix} + t\begin{bmatrix}-1\\-1\\0\end{bmatrix}\right) + c_3 e^{2t}\begin{bmatrix}0\\0\\1\end{bmatrix}.
\end{aligned}$$

This answer may appear to be different from the answer in Example 6.8.15, however, if we let

$$\begin{aligned}
k_1 &= c_1 \\
k_2 &= c_1 + c_2 \\
k_3 &= c_1 + c_2 + c_3
\end{aligned}$$

the answer is

$$\xi(t) = k_1 e^{2t} \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right) + k_2 e^{2t} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + k_3 e^{2t} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

which is the same. ■

6.9 Stability

6.10 Diagonalization and Jordan Normal Form

6.11 Applications of Homogeneous Systems of First Order Equations

This section has been moved to Chapter 7, Section 7.3.

6.12 Nonhomogeneous Systems of First Order Equations

Now we consider how to solve systems of the type

$$\dot{\xi} = A\xi + g(t),$$

where

$$\begin{array}{rcl} A & \in & \mathbb{R}^{n\times n} \\ \xi & \in & \mathbb{R}^n \\ g(t) & \in & \mathbb{R}^n, \end{array}$$

or in detail

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} + \begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_n(t) \end{bmatrix}. \qquad (6.32)$$

First consider a mechanical example that gives rise to equations of this nature.

Example 6.12.1 As an example of a type of system that is modeled by such a set of equations, consider again the system illustrated in Figure 6.1, but unlike before we will not assume that $F(t) = 0$. As before, if

$$\begin{aligned} \xi_1 &= x_1 \\ \xi_2 &= \dot{x}_1 \\ \xi_3 &= x_2 \\ \xi_4 &= \dot{x}_2 \end{aligned}$$

then the equations of motion given in Equation 6.7 are equivalent to

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_1+k_2}{m_1} & -\frac{b_1+b_2}{m_1} & \frac{k_2}{m_1} & \frac{b_2}{m_1} \\ 0 & 0 & 1 & 0 \\ \frac{k_2}{m_2} & \frac{b_2}{m_2} & -\frac{k_2}{m_2} & -\frac{b_2}{m_2} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{F(t)}{m_2} \end{bmatrix}.$$

■

The following three methods are appropriate for solving nonhomogeneous systems of first order linear ordinary differential equations.

6.12.1 Diagonalization and Jordan Canonical Form

The fundamental idea underlying this approach is to convert the system of coupled first order equations into decoupled equations. What this means mathematically will be apparent shortly, but the consequence of this approach is unlike the system in Equation 6.32 where the entire system must be solved at once, each equation (or row) can be solved individually, or one at a time. First we need to investigate the concept of converting a matrix to *diagonal* form.

For a system of the form

$$\dot{\xi} = A\xi + g(t) \tag{6.33}$$

we first consider the easier case where A has a full set of n linearly independent eigenvectors, $\hat{\xi}_1, \ldots, \hat{\xi}_n$, and define the matrix T as the matrix with the eigenvectors of A as its columns, *i.e.*,

$$T = \left[\begin{array}{ccc} \hat{\xi}_1 & \hat{\xi}_2 \cdots & \hat{\xi}_n \end{array}\right].$$

Since the definition of an eigenvector is

$$A\hat{\xi}_i = \lambda_i \hat{\xi}_i$$

then

$$AT = \left[\begin{array}{cccc} \lambda_1\hat{\xi}_1 & \lambda_2\hat{\xi}_2 & \cdots & \lambda_n\hat{\xi}_n \end{array}\right].$$

Now, since we assumed that $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_n$ were linearly independent, then T is invertible. Note that by definition

$$T^{-1}T = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Considering this equation column by column, we have

$$T^{-1}\hat{\xi}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad T^{-1}\hat{\xi}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots \quad T^{-1}\hat{\xi}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Also, since $A\hat{\xi}_i = \lambda_i \hat{\xi}_i$

$$T^{-1}A\hat{\xi}_1 = T^{-1}\lambda_1\hat{\xi}_1 = \lambda_i T^{-1}\hat{\xi}_1 = \lambda_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \lambda_1 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$T^{-1}A\hat{\xi}_2 = T^{-1}\lambda_2\hat{\xi}_2 = \lambda_i T^{-1}\hat{\xi}_2 = \lambda_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda_2 1 \\ \vdots \\ 0 \end{bmatrix},$$

and so forth until

$$T^{-1}A\hat{\xi}_n = T^{-1}\lambda_n\hat{\xi}_n = \lambda_i T^{-1}\hat{\xi}_n = \lambda_n \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \lambda_n \end{bmatrix}.$$

Finally, putting it all together gives the important relation

$$T^{-1}AT = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Now, using this, again consider Equation 6.33 and let

$$\xi = T\psi$$

where the columns of T are the eigenvectors of A as before. Note that since T is a constant matrix,

$$\dot{\xi} = T\dot{\psi}.$$

Substituting into Equation 6.33 gives

$$T\dot{\psi} = AT\psi + g(t),$$

or

$$\dot{\psi} = T^{-1}AT\psi + T^{-1}g(t).$$

In detail, this looks like

$$\begin{aligned}
\frac{d}{dt}\begin{bmatrix}\psi_1\\ \psi_2\\ \psi_3\\ \vdots\\ \psi_n\end{bmatrix} &= \begin{bmatrix}\lambda_1 & 0 & 0 & \cdots & 0\\ 0 & \lambda_2 & 0 & \cdots & 0\\ 0 & 0 & \lambda_3 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \lambda_n\end{bmatrix}\begin{bmatrix}\psi_1\\ \psi_2\\ \psi_3\\ \vdots\\ \psi_n\end{bmatrix} + T^{-1}\begin{bmatrix}g_1(t)\\ g_2(t)\\ g_3(t)\\ \vdots\\ g_n(t)\end{bmatrix} \qquad (6.34)\\
&= \begin{bmatrix}\lambda_1\psi_1\\ \lambda_2\psi_2\\ \lambda_3\psi_3\\ \vdots\\ \lambda_n\psi_n\end{bmatrix} + T^{-1}\begin{bmatrix}g_1(t)\\ g_2(t)\\ g_3(t)\\ \vdots\\ g_n(t)\end{bmatrix}\\
&= \begin{bmatrix}\lambda_1\psi_1\\ \lambda_2\psi_2\\ \lambda_3\psi_3\\ \vdots\\ \lambda_n\psi_n\end{bmatrix} + \begin{bmatrix}h_1(t)\\ h_2(t)\\ h_3(t)\\ \vdots\\ h_n(t)\end{bmatrix}
\end{aligned}$$

where

$$h(t) = T^{-1}g(t).$$

The significance of Equation 6.34 is that each of the ψ_i equations are *decoupled* and in the form of

$$\dot{\psi}_i = \lambda_i\psi + h_i(t).$$

Hence, each can be solved independently using the appropriate method from Chapter 2. For example, using an integrating factor

$$\begin{aligned}
\frac{d}{dt}\psi_i - \lambda_i\psi_i &= h_i(t)\\
e^{-\lambda_i t}\left(\frac{d}{dt}\psi_i - \lambda_i\psi_i\right) &= e^{-\lambda_i t}h_i(t)\\
\frac{d}{dt}\left(e^{-\lambda_i t}\psi_i\right) &= e^{-\lambda_i t}h_i(t).
\end{aligned}$$

Hence, integrating both sides gives

$$\begin{aligned}\int_0^t \frac{d}{d\tau}\left(e^{-\lambda_i\tau}\psi_i(\tau)\right) &= \int_0^t e^{-\lambda_i\tau}h_i(\tau)d\tau \\ e^{-\lambda_i t}\psi_i(t)-\psi_i(0) &= \int_0^t e^{-\lambda_i\tau}h_i(\tau)d\tau.\end{aligned}$$

Hence

$$\psi_i(t) = e^{\lambda_i t}\int_0^t e^{-\lambda_i\tau}h(\tau)d\tau + \psi_i(0)e^{\lambda_i t},$$

if the initial condition is specified or

$$\psi_i(t) = e^{\lambda_i t}\int_0^t e^{-\lambda_i\tau}h(\tau)d\tau + ce^{\lambda_i t},$$

if the general solution is desired.

After solving all the $\psi_i(t)$ equations, the solution for the ξ variables is simply computed using the original equation

$$\xi = T\psi.$$

Example 6.12.2 Determine the general solution to

$$\frac{d}{dt}\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\ 2 & 1 & -1\\ -8 & -5 & -3\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ \cos t\end{bmatrix}.$$

Computing the eigenvalues and eigenvectors gives

$$\begin{aligned}\lambda_1 &= -2\\ \lambda_2 &= -1\\ \lambda_3 &= 2\end{aligned}$$

and

$$\hat{\xi}_1 = \begin{bmatrix}-4\\ 5\\ 7\end{bmatrix}, \quad \hat{\xi}_2 = \begin{bmatrix}-3\\ 4\\ 2\end{bmatrix}, \quad \hat{\xi}_3 = \begin{bmatrix}0\\ -1\\ 1\end{bmatrix}.$$

Thus

$$T = \begin{bmatrix}-4 & -3 & 0\\ 5 & 4 & -1\\ 7 & 2 & 1\end{bmatrix}$$

and

$$T^{-1} = \begin{bmatrix}\frac{1}{2} & \frac{1}{4} & \frac{1}{4}\\ -1 & -\frac{1}{3} & -\frac{1}{3}\\ -\frac{3}{2} & -\frac{13}{12} & -\frac{1}{12}\end{bmatrix}.$$

Computing $T^{-1}AT$ and $T^{-1}g(t)$ gives the following equations for ψ

$$\frac{d}{dt}\begin{bmatrix}\psi_1\\\psi_2\\\psi_3\end{bmatrix}=\begin{bmatrix}-2&0&0\\0&-1&0\\0&0&2\end{bmatrix}\begin{bmatrix}\psi_1\\\psi_2\\\psi_3\end{bmatrix}+\begin{bmatrix}\frac{1}{4}\cos t\\-\frac{1}{3}\cos t\\-\frac{1}{12}\cos t\end{bmatrix}$$

or as individual equations

$$\begin{aligned}\dot{\psi}_1&=-2\psi_1+\frac{1}{4}\cos t\\\dot{\psi}_2&=-\psi_2-\frac{1}{3}\cos t\\\dot{\psi}_3&=2\psi_3-\frac{1}{12}\cos t.\end{aligned}$$

The solutions to these equations are

$$\begin{aligned}\psi_1&=e^{-2t}\int_0^t e^{2t}\frac{1}{4}\cos\tau d\tau+\psi_1(0)e^{-2t}\\\psi_2&=-e^{-t}\int_0^t e^{t}\frac{1}{3}\cos\tau d\tau+\psi_2(0)e^{-t}\\\psi_3&=-e^{2t}\int_0^t e^{-2t}\frac{1}{12}\cos\tau d\tau+\psi_3(0)e^{2t},\end{aligned}$$

or

$$\begin{aligned}\psi_1(t)&=c_1e^{-2t}+\frac{1}{10}\cos t+\frac{1}{20}\sin t\\\psi_2(t)&=c_2e^{-t}-\frac{1}{6}\cos t-\frac{1}{6}\sin t\\\psi_3(t)&=c_3e^{2t}+\frac{1}{30}\cos t-\frac{1}{60}\sin t.\end{aligned}$$

The final solution is computed by determining

$$\xi=T\psi,$$

■

~~which is a bit too messy to write out in detail.~~

which is

$$\begin{aligned}\begin{bmatrix}\xi_1\\\xi_2\\\xi_3\end{bmatrix}&=T=\begin{bmatrix}-4&-3&0\\5&4&-1\\7&2&1\end{bmatrix}\begin{bmatrix}c_1e^{-2t}+\frac{1}{10}\cos t+\frac{1}{20}\sin t\\c_2e^{-t}-\frac{1}{6}\cos t-\frac{1}{6}\sin t\\c_3e^{2t}+\frac{1}{30}\cos t-\frac{1}{60}\sin t\end{bmatrix}\\&=\begin{bmatrix}-4\psi_1(t)-3\psi_2(t)\\5\psi_1(t)+4\psi_2(t)-\psi_3(t)\\7\psi_1(t)+2\psi_2(t)+\psi_3(t)\end{bmatrix},\end{aligned}$$

which gives

$$\begin{aligned}
\xi_1(t) &= -4c_1e^{-2t} - \frac{2}{5}\cos t - \frac{1}{5}\sin t - 3c_2e^{-t} + \frac{1}{2}\cos t + \frac{1}{2}\sin t \\
\xi_2(t) &= 5c_1e^{-2t} + \frac{1}{2}\cos t + \frac{1}{4}\sin t + 4c_2e^{-t} - \frac{2}{3}\cos t - \frac{2}{3}\sin t - \\
& \quad c_3e^{2t} - \frac{1}{30}\cos t + \frac{1}{60}\sin t \\
\xi_3(t) &= 7c_1e^{-2t} + \frac{7}{10}\cos t + \frac{7}{20}\sin t + 2c_2e^{-t} - \frac{1}{3}\cos t - \frac{1}{3}\sin t + \\
& \quad c_3e^{2t} + \frac{1}{30}\cos t - \frac{1}{60}\sin t.
\end{aligned}$$

6.12.2 Undetermined Coefficients

Recall that the method of undetermined coefficients from Section 3.4.1 was based upon the fact that derivatives of functions of the form

1. $\sin\omega t$ and $\cos\omega t$,
2. $e^{\alpha t}$,
3. $\alpha_0 t^n + \alpha_1 t^{n-1} + \alpha_2 t^{n-2} + \cdots + \alpha_{n-1}t + \alpha_n$, and
4. products and sums of them,

are exactly the same set of functions. Thus when the nonhomogeneous term contains function of this type, the particular solution of an ordinary differential equation will be a general combination of the same type of functions. There are two slight complications or variations that are necessary distinguish the approach for systems of first order equations from one scalar second order system.

General Form of Particular Solution

The first complication is that even though the nonhomogeneous term may appear in one component of the differential equation, the form of the solution must have undetermined coefficients for all of the components. In a general functional description, if the all the nonhomogeneous terms that appear in the vector $g(t)$ would require a particular solution of the form

$$x_p(t) = af_1(t) + bf_2(t) + cf_3(t) + \cdots$$

in the scalar (first or second order) case, then in the case of

$$\dot{\xi} = A\xi + g(t), \qquad \xi \in \mathbb{R}^n,$$

then the assumed form of the solution will be

$$\xi_p(t) = af_1(t) + bf_2(t) + cf_3(t) + \cdots$$

where $a, b, c, \ldots \in \mathbb{R}^n$, *i.e.,* the coefficients are *vectors.* The following example illustrates this point.

Example 6.12.3 Find the general solution to

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} \xi_1 \\ xi_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \cos 4t \end{bmatrix}.$$

In the scalar case, the assumed form of the solution would simply be $x_p(t) = a\cos 4t + b\sin 4t$, so for this problem we assume

$$\xi_p(t) = a\cos 4t + b\sin 4t = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\cos 4t + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\sin 4t.$$

The rest of the procedure is exactly as before. Substitute the assumed form of the particular solution into the differential equations and equate the coefficients of different functions of t. Thus,

$$\dot{\xi}_p(t) = -4a\sin 4t + 4b\cos 4t,$$

and substituting gives

$$\begin{bmatrix} -4a_1\sin 4t + 4b_1\cos 4t \\ -4a_2\sin 4t + 4b_2\cos 4t \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} a_1\cos 4t + b_1\sin 4t \\ a_2\cos 4t + b_2\sin 4t \end{bmatrix} + \begin{bmatrix} 0 \\ \cos 4t \end{bmatrix}.$$

Since this must be true for all time, the coefficients of the sine and cosine terms in each equation must be equal. Thus, the coefficients are determined by the following four equations:

$$\begin{aligned} \text{sine term, first equation} \quad &\Longrightarrow \quad -4a_1 = 2b_1 + b_2 \\ \text{cosine term, first equation} \quad &\Longrightarrow \quad 4b_1 = 2a_1 + a_2 \\ \text{sine term, second equation} \quad &\Longrightarrow \quad -4a_2 = 3b_2 \\ \text{sine term, first equation} \quad &\Longrightarrow \quad 4b_2 = 3a_2 + 1. \end{aligned}$$

Solving these gives

$$\begin{aligned} a_1 &= -\frac{1}{50} \\ a_2 &= -\frac{3}{25} \\ b_1 &= -\frac{1}{25} \\ b_2 &= \frac{4}{25}. \end{aligned}$$

Thus the particular solution is

$$\xi_p(t) = \begin{bmatrix} -\frac{1}{50} \\ -\frac{3}{25} \end{bmatrix}\cos 4t + \begin{bmatrix} -\frac{1}{25} \\ -\frac{4}{25} \end{bmatrix}\sin 4t.$$

To compute the general solution, the homogeneous solution, *i.e.,* the solution to

$$\dot{\xi} = A\xi$$

is needed. A simple computation shows that the eigenvalues and eigenvectors of A are

$$\lambda_1 = 3, \quad \hat{\xi}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \lambda_2 = 2, \quad \hat{\xi}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Thus, the general solution is

$$\xi(t) = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{3t} + c_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{2t} + \begin{bmatrix} -\frac{1}{50} \\ -\frac{3}{25} \end{bmatrix} \cos 4t + \begin{bmatrix} -\frac{1}{25} \\ -\frac{4}{25} \end{bmatrix} \sin 4t.$$ ■

In the previous example, note that the sine and cosine terms appear in *both* components of the solution even though the nonhomogeneous term contains $\cos 4t$ only in the second term. This is due to the fact that the equations are *coupled*, and the effect of the nonhomogeneity is not limited to the line in which it appears.

Equivalent Homogeneous Solution and Nonhomogeneous Term

The second complication is when the nonhomogeneous term is the exponential of an eigenvalue of the matrix A. When confronted with this problem in Chapter 3, the approach was to multiply the assumed form of the particular solution by the dependent variable. The approach for nonhomogeneous systems of first order equations with equivalent homogeneous solutions and nonhomogeneous terms is similar, but with a slight twist, as the following examples illustrate.

The first example is the second order scalar case, which is included to help you recall the procedure from Chapter 3.

Example 6.12.4 (Review problem from Chapter 3) Determine the general solution to

$$\ddot{x} + 4x = \cos 2t. \tag{6.35}$$

Assuming a homogeneous solution of the form

$$x_h(t) = e^{\lambda t}$$

and substituting gives

$$\lambda^2 + 4 = 0 \qquad \Longrightarrow \qquad \lambda = \pm 2i.$$

For the particular solution, we are first inclined to assume a solution of the form

$$x_p(t) = a\cos 2t + b\sin 2t.$$

One that is observant and experienced in dealing with undetermined coefficients will immediately recognize that this will not work since it is actually a homogeneous solution. When $x_p(t)$ of this form is substituted into Equation 6.35 it will disappear leaving nothing to equate to the nonhomogeneous term since it is actually a solution of the homogeneous equation. In detail,

$$\ddot{x}_p(t) = -4a\cos 2t - 4b\sin 2t,$$

and substituting gives

$$\begin{aligned} -4a\cos 2t - 4b\sin 2t + 4\left(a\cos 2t + b\sin 2t\right) &= \cos 2t \\ 0 &= \cos 2t. \end{aligned}$$

The 0 on the left hand side of the previous equation is *guaranteed* to occur since $x_p(t)$ happens to satisfy

$$\ddot{x} + 4x = 0.$$

Recall, that the correct form to assume for the particular solution in this case would be

$$x_p(t) = t\left(a\cos 2t + b\sin 2t\right).$$

Then,

$$\ddot{x}_p(t) = -4\left((at-b)\cos 2t + (a+bt)\sin 2t\right)$$

and substituting and equating coefficients gives

$$\begin{aligned} -4\left((at-b)\cos 2t + (a+bt)\sin 2t\right) &+ \\ 4t\left(a\cos 2t + b\sin 2t\right) &= \cos 2t. \end{aligned}$$

Since this must be true for all t, the coefficients of $\sin 2t$, $t\sin 2t$, $\cos 2t$ and $t\cos 2t$ must be equal. Thus

$$\begin{aligned} -4a &= 0 \\ -4b + 4b &= 0 \\ 4b &= 1 \\ -4a + 4a &= 0 \end{aligned}$$

respectively. From this we obtain

$$\begin{aligned} a &= 0 \\ b &= \frac{1}{4}, \end{aligned}$$

and hence

$$x_p(t) = \frac{1}{4}t\sin 2t.$$

■

The analogous situation for a system of first order equations is when the nonhomogeneous term includes the exponential of one of the eigenvalues of the matrix A.

Example 6.12.5 (Wrong approach number 1) Determine the general solution to

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}.$$

An easy computation shows that the eigenvalues and corresponding eigenvectors of A are

$$\lambda_1 = 2 \quad \hat{\xi}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \lambda_2 = 3 \quad \hat{\xi}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Since the nonhomogeneous term contains e^{3t} which is precisely the exponential of an eigenvalue of A, we should expect to run into trouble equating coefficients. Trying it anyway gives

$$\xi_p(t) = ae^{3t} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} e^{3t}.$$

Thus

$$\dot{\xi}_p(t) = 3ae^{3t},$$

and substituting into the differential equation gives

$$3\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} e^{3t} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} e^{3t} + \begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}.$$

Equating coefficients of e^{3t} in each equation gives

$$\begin{aligned} 3a_1 &= 2a_1 + a_2 \\ 3a_2 &= 3a_2 + 1. \end{aligned}$$

Since there is no value for a_2 that can satisfy the second equation, there is no solution, and hence, no way to determine the undetermined coefficients. It is left as a homework problem to see that exactly the same thing happens if the eigenvalue is purely imaginary (complex) and the nonhomogeneous term contains a sine or cosine at the same frequency. ■

Since the correct approach in Chapter 3 was to simply multiply the assumed form of the solution by the independent variable, t, one may assume that the same approach works in this case as well. Unfortunately, as the following example illustrates, it does not work.

Example 6.12.6 (Wrong approach number 2) Again consider

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}.$$

Since the nonhomogeneous term contains e^{3t} which is precisely the exponential of an eigenvalue of A, we should expect to run into trouble equating coefficients. Thus assume

$$\xi_p(t) = ate^{3t} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} te^{3t}.$$

Thus

$$\dot{\xi}_p(t) = 3ate^{3t} + ae^{3t}$$

and substituting into the differential equation gives

$$\left(3t\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right)e^{3t} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}te^{3t} + \begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}.$$

Equating coefficients of e^{3t} and te^{3t} in each equation gives

$$\begin{aligned} a_1 &= 0 \\ 3a_1 &= 2a_1 + a_2 \\ a_2 &= 1 \\ 3a_2 &= 3a_2. \end{aligned}$$

Again, there is no solution. ■

The following example elaborates upon the reason why the simple approach of only multiplying by the independent variable t does not work.

Example 6.12.7 Determine the general solution to

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} e^{2t} \\ e^{2t} \end{bmatrix}.$$

These equations are decoupled, so we can immediately see (or compute)

$$\begin{aligned} \xi_{1_h} &= e^{3t} \\ \xi_{2_h} &= e^{2t}. \end{aligned}$$

Since the homogeneous solution for ξ_2 is the same as the nonhomogeneous term, clearly assuming e^{2t} will be problematic. Thus, we need a term of the form te^{2t} in the assumed form of the particular solution for ξ_3. However, a term of the form e^{2t} in the particular solution is exactly what is needed for the first line since the homogeneous solution for ξ_1 contains e^{3t}, not e^{2t}. Thus, assuming

$$\xi_p(t) = ae^{2t}$$

will not work because of the ξ_2 component, and

$$\xi_p(t) = ate^{2t}$$

will not work because of the ξ_1 component. A solution containing *both* terms is necessary. ■

Unfortunately, there is still one final twist to this whole affair. Since it is necessary to assume a particular solution that is the sum of the independent variable, t times the homogeneous solution and the homogeneous solution itself, there will not be a unique particular solution. This is because of the the term in the homogeneous solution that is not multiplied by the independent variable in the assumed form of the particular solution can be combined with the homogeneous solution in an arbitrary manner. This (along with the correct approach) is illustrated by the following example.

Example 6.12.8 (Right approach) Again consider

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}.$$

Observing that $\lambda = 3$ is an eigenvalue of A we assume

$$\xi_p(t) = ate^{3t} + be^{3t} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} te^{3t} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} e^{3t}.$$

Thus

$$\dot{\xi}_p(t) = 3ate^{3t} + ae^{3t} + 3be^{3t}$$

and substituting into the differential equation gives

$$\left(3t\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + 3\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right) e^{3t} =$$
$$\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} t + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right) e^{3t} + \begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}.$$

Equating coefficients of e^{3t} and te^{3t} in each equation gives

$$\begin{aligned} a_1 + 3b_1 &= 2b_1 + b_2 \\ 3a_1 &= 2a_1 + a_2 \\ a_2 + 3b_2 &= 3b_2 + 1 \\ 3a_2 &= 3a_2. \end{aligned}$$

Simplifying these equations gives only three independent equations

$$\begin{aligned} a_1 + b_1 &= b_2 \\ a_1 &= a_2 \\ a_1 &= 1. \end{aligned}$$

The reason there are less than four equations, and hence no unique solution, is because the vector b in the assumed form of the solution must be an eigenvector of A and hence can be combined in any linear way with one of the homogeneous solutions. One solution to the above three equations is

$$\begin{aligned} a_1 &= 1 \\ a_2 &= 1 \\ b_1 &= 0 \\ b_2 &= 1, \end{aligned}$$

and hence

$$\xi_p(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{3t} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{3t}. \tag{6.36}$$

This particular solution is *not* unique. Indeed,

$$\begin{aligned} a_1 &= 1 \\ a_2 &= 1 \\ b_1 &= -1 \\ b_2 &= 0, \end{aligned}$$

also work giving

$$\xi_p(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{3t} + \begin{bmatrix} -1 \\ 0 \end{bmatrix} e^{3t}. \tag{6.37}$$

The reason both particular solutions work is that when they are combined with the homogeneous solution, they yield the same solution. In particular, from Example 6.12.5 we can write the homogeneous solution as

$$\xi_h(t) = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{2t} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{3t}.$$

Then the general solution using the particular solution from Equation 6.36 gives

$$\xi(t) = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{2t} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{3t} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{3t} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{3t},$$

and the general solution using the particular solution from Equation 6.37 gives

$$\xi(t) = \hat{c}_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{2t} + \hat{c}_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{3t} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{3t} + \begin{bmatrix} -1 \\ 0 \end{bmatrix} e^{3t}.$$

For c_2 from the first equation and $\hat{c}_2$ from the second equation, if $\hat{c}_2 = c_2 + 1$ the equations are identical. ■

6.12.3 Variation of Parameters

With all the complications involved in the method of undetermined coefficients, one may be hesitant to even venture into the realm of variation of parameters since, at least in Chapter 3 the derivation was rather complicated. Thankfully, in the case of nonhomogeneous systems of first order equations, variation of parameters is even more straightforward than in the scalar second order case.

Given

$$\dot{\xi} = A\xi + g(t) \tag{6.38}$$

where

$$\begin{aligned} A &\in \mathbb{R}^{n\times n} \\ \xi &\in \mathbb{R}^n \\ g(t) &\in \mathbb{R}^n \end{aligned}$$

assume that $\xi_{1_h}, \xi_{2_h}, \ldots, \xi_{n_h}$ are n linearly independent homogeneous solutions to Equation 6.38, *i.e.*, they satisfy

$$\dot{\xi}_{i_h} = A\xi_{i_h}.$$

Because it is useful subsequently, we first construct and define a matrix, $\Xi(t)$ where the columns of $\Xi(t)$ are the homogeneous solutions, $\xi_{i_h}(t)$.

Definition 6.12.9 (Fundamental Matrix Solution) Let $\xi_{1_h}, \xi_{2_h}, \ldots, \xi_{n_h}$ satisfy

$$\dot{\xi}_{i_h} = A\xi_{i_h}.$$

The *fundamental matrix solution* is the matrix

$$\Xi(t) = \left[\begin{array}{cccc} \xi_{1_h}(t) & \xi_{2_h}(t) & \cdots & \xi_{n_h}(t) \end{array}\right],$$

i.e., the columns of $\Xi(t)$ are the homogeneous solutions. ◇

Example 6.12.10 Consider the general solution to $\dot{\xi} = A\xi$ where

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

■

Skipping the details the general solution was

$$\begin{aligned} \xi(t) &= c_1\hat{\xi}_1 e^{\lambda_1 t} + c_2\hat{\xi}_2 e^{\lambda_1 t} + c_3\hat{\xi}_3^1 e^{\lambda_3 t} + c_4\left(\hat{\xi}_3^2 + t\hat{\xi}_3^1\right)e^{\lambda_3 t} \\ &= c_1\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_2\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_3\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} e^{3t} + \\ &\quad\; c_4\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + t\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\right)e^{3t}. \end{aligned}$$

Since each term that is multiplied by a constant, c_i is a homogeneous solution simply construct a matrix with each one as a column to construct the fundamental matrix solution

$$\begin{aligned} \Xi(t) &= \left[\begin{array}{cccc} \hat{\xi}_1 e^{\lambda_1 t} & \hat{\xi}_2 e^{\lambda_2 t} & \hat{\xi}_3^1 e^{\lambda_3 t} & \left(\hat{\xi}_3^2 + t\hat{\xi}_3^1\right)e^{\lambda_3 t} \end{array}\right] \\ &= \left[\begin{array}{cccc} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} e^{2t} & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} e^{2t} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} e^{3t} & \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + t\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\right)e^{3t} \end{array}\right] \\ &= \begin{bmatrix} e^{2t} & 0 & 0 & 0 \\ 0 & e^{2t} & 0 & 0 \\ 0 & 0 & e^{3t} & te^{3t} \\ 0 & 0 & 0 & e^{3t} \end{bmatrix}. \end{aligned}$$

The fundamental matrix solution has one important property that will be used in the derivation of the variation of parameters solution; namely, the whole matrix satisfies the homogeneous equation. In other words, if $\Xi(t)$ is the fundamental matrix solution to

$$\dot{\xi} = A\xi$$

then

$$\dot{\Xi} = A\Xi.$$

This is true since each column of $\Xi(t)$ is a homogeneous solution and is illustrated by the following example.

Example 6.12.11 From Example 6.12.10 we have

$$\Xi(t) = \begin{bmatrix} e^{2t} & 0 & 0 & 0 \\ 0 & e^{2t} & 0 & 0 \\ 0 & 0 & e^{3t} & te^{3t} \\ 0 & 0 & 0 & e^{3t} \end{bmatrix}$$

so

$$\begin{aligned} \dot{\Xi}(t) &= \begin{bmatrix} 2e^{2t} & 0 & 0 & 0 \\ 0 & 2e^{2t} & 0 & 0 \\ 0 & 0 & 3e^{3t} & 3te^{3t} + e^{3t} \\ 0 & 0 & 0 & 3e^{3t} \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} e^{2t} & 0 & 0 & 0 \\ 0 & e^{2t} & 0 & 0 \\ 0 & 0 & e^{3t} & te^{3t} \\ 0 & 0 & 0 & e^{3t} \end{bmatrix}. \end{aligned}$$

Thus $\dot{\Xi} = A\Xi$. ■

Similar to the approach for second order equations, the approach to find the particular solution for a nonhomogeneous system of first order equations is to assume that the particular solution is of the form of

$$\xi_p(t) = \Xi(t)u(t)$$

where $u(t)$ is a vector of unknown functions. To determine $u(t)$, simply substitute into Equation 6.38. First note that (dropping the explicit dependence on t)

$$\dot{\xi}_p = \dot{\Xi}u + \Xi\dot{u}.$$

Substituting into Equation 6.38 gives

$$\dot{\Xi}u + \Xi\dot{u} = A\Xi u + g.$$

Since

$$\dot{\Xi} = A\Xi \qquad \Longrightarrow \dot{\Xi}u = A\Xi u$$

so

$$\Xi\dot{u} = g.$$

Since Ξ contains n linearly independent solutions, it is invertible and hence

$$\dot{u} = \Xi^{-1}g \qquad \Longrightarrow \qquad u(t) = \int_{t_0}^{t} \Xi^{-1}(\tau)g(\tau)d\tau.$$

Substituting into the assumed form of the particular solution gives a complete expression for the particular solution as

$$\xi_p(t) = \Xi\int_{t_0}^{t} \Xi^{-1}(\tau)g(\tau)d\tau.$$

Note that to even compute the particular solution we need the fundamental matrix which contains a full set of homogeneous solutions. Since any linear combination of the homogeneous solutions can be expressed as

$$c_1\xi_{1_h} + c_2\xi_{2_h} + \cdots + c_n\xi_{n_h} = \Xi(t)c$$

where

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

the general solution to Equation 6.38 is

$$\xi(t) = \Xi(t)c + \Xi(t)\int_{t_0}^{t} \Xi^{-1}(\tau)g(\tau)d\tau. \tag{6.39}$$

Finally, if the initial conditions, $\xi(t_0)$ are specified, then

$$\xi(t_0) = \Xi(t_0)c$$

since the integral with the same upper and lower limits is zero. Hence

$$c = \Xi^{-1}(0)\xi(t_0)$$

and substituting into the general solution gives the entire answer as

$$\xi(t) = \Xi(t)\Xi^{-1}(t_0)\xi(t_0) + \Xi(t)\int_{t_0}^{t} \Xi^{-1}(\tau)g(\tau)d\tau. \tag{6.40}$$

An example illustrates the straightforward application of this method.

Example 6.12.12 Solve

$$\frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} e^{-4t} \\ 0 \end{bmatrix}.$$

A simple computation determines the eigenvalues and eigenvectors for the matrix as

$$\lambda_1 = -4 \quad \hat{\xi}_1 = \left[\begin{array}{c} -1 \\ 1 \end{array} \right] \qquad \lambda_2 = -2 \quad \hat{\xi}_2 = \left[\begin{array}{c} 1 \\ 1 \end{array} \right],$$

thus

$$\Xi(t) = \left[\begin{array}{cc} -e^{-4t} & e^{-2t} \\ e^{-4t} & e^{-2t} \end{array} \right].$$

A simple computation determines that

$$\Xi^{-1}(t) = \frac{1}{2} \left[\begin{array}{cc} -e^{4t} & e^{4t} \\ e^{2t} & e^{2t} \end{array} \right],$$

and

$$\Xi^{-1}(t) g(t) = \left[\begin{array}{c} -\frac{1}{2} \\ \frac{1}{2} e^{-2t} \end{array} \right].$$

Assuming that $t_0 = 0$,

$$\begin{aligned} \int_0^t \Xi^{-1}(\tau) g(\tau) d\tau &= \int_0^t \left[\begin{array}{c} -\frac{1}{2} \\ \frac{1}{2} e^{-2\tau} \end{array} \right] d\tau \\ &= \left[\begin{array}{c} -\frac{1}{2}\tau \\ \frac{1}{4}\left(1 - e^{-2t}\right) \end{array} \right]. \end{aligned}$$

Then

$$\Xi(t) \int_0^t \Xi^{-1}(\tau) g(\tau) d\tau = \left[\begin{array}{c} \frac{1}{4}\left(e^{-2t} + 2te^{-4t} - e^{-4t}\right) \\ \frac{1}{4}\left(e^{-2t} - 2te^{-4t} - e^{-4t}\right) \end{array} \right]. \qquad \blacksquare$$

So finally we have

$$\begin{aligned} \xi(t) &= \Xi(t) c + \Xi(t) \int_{t_0}^t \Xi^{-1}(\tau) g(\tau) d\tau \\ &= c_1 \left[\begin{array}{c} -e^{-4t} \\ e^{-4t} \end{array} \right] + c_2 \left[\begin{array}{c} e^{-2t} \\ e^{-2t} \end{array} \right] + \left[\begin{array}{c} \frac{1}{4}\left(e^{-2t} + 2te^{-4t} - e^{-4t}\right) \\ \frac{1}{4}\left(e^{-2t} - 2te^{-4t} - e^{-4t}\right) \end{array} \right]. \end{aligned}$$

6.13 Applications of Nonhomogeneous Systems of Equations

6.14 Exercises

It is possible to complete all of these exercises by hand.

Problem 6.1 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -4 & 4 & 0 & 0 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & -2 & 3 \end{bmatrix}.$$

Problem 6.2 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} 6 & -4 \\ 0 & 2 \end{bmatrix}.$$

Determine the solution if

$$\xi(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Problem 6.3 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

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

Determine the solution if

$$\xi(0) = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}.$$

Problem 6.4 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

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

Determine the solution if

$$\xi(0) = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}.$$

Problem 6.5 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -8 & 7 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Determine the solution if

$$\xi(0) = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}.$$

Problem 6.6 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -\frac{7}{2} & \frac{15}{2} & -3 \\ -\frac{3}{2} & -\frac{1}{2} & 3 \\ 0 & 0 & 1 \end{bmatrix}.$$

Determine the solution if

$$\xi(0) = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

Problem 6.7 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -1 & -4 \\ 4 & -1 \end{bmatrix}.$$

Problem 6.8 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -1 & -3 & 2 \\ 1 & -5 & 5 \\ 3 & -3 & -2 \end{bmatrix}.$$

Problem 6.9 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 9 \\ 0 & 2 & 1 & 4 \\ 0 & -4 & 0 & 14 \end{bmatrix}.$$

Problem 6.10 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} 11 & 0 & 17 \\ 0 & -6 & 0 \\ -2 & 0 & 1 \end{bmatrix}.$$

Problem 6.11 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -5 & 1 & 0 & 0 \\ -1 & -3 & 0 & 0 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & 2 & -5 \end{bmatrix}.$$

Problem 6.12 Determine the general solution to

$$\dot{\xi} = A\xi$$

where

$$A = \begin{bmatrix} -5 & 0 & 0 & 0 & 0 \\ 0 & -3 & 2 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & 0 & -5 & 1 \\ 0 & 0 & 0 & -1 & -7 \end{bmatrix}.$$

Problem 6.13 Determine the solution to

$$\dot{\xi} = A\xi + g(t)$$

where

$$A = \begin{bmatrix} -3 & 1 & 0 \\ 0 & -2 & 0 \\ 1 & 1 & -4 \end{bmatrix}$$

and

$$g(t) = \begin{bmatrix} 0 \\ 0 \\ \cos t \end{bmatrix}.$$

Chapter 7

Applications of Systems of First Order Equations

7.1 Introduction

7.2 Linearization of Nonlinear Systems

7.3 Multi-Degree of Freedom Vibrations

~~Classical Normal Modes of Vibration~~

Consider the system illustrated in Figure 7.1. We will first analyze this system using the approach from classical vibrations theory and then relate it to the material covered previously in this chapter.

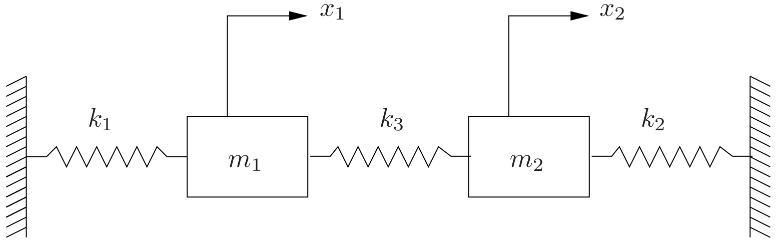


Figure 7.1. Two degree of freedom mass-spring-damper system.

A simple analysis of the free body diagrams for the two masses yields the following equations of motion

$$\begin{aligned} m\ddot{x}_1 + (k_1 + k_3)\, x_1 - k_3 x_2 &= 0 \\ m\ddot{x}_2 + (k_2 + k_3)\, x_2 - k_3 x_2 &= 0. \end{aligned} \tag{7.1}$$

Classical Approach

The classical approach is simply to assume (perhaps based upon some intuitive insight into the problem) the form of the solutions for masses one and two. For present purposes, assume

$$\begin{aligned} x_1(t) &= a_1 \cos \omega t \\ x_2(t) &= a_2 \cos \omega t. \end{aligned}$$

Note the assume form of the solution is very restrictive; in particular, it will at best only be valid when $\dot{x}_1(0) = \dot{x}_2(0) = 0$; furthermore, it assumes the frequency of oscillation of the two masses must be the same. Regardless, let us proceed to substitute these solutions into the equations of motion. Upon doing so we obtain

$$\begin{aligned} \left[-m_1 a_1 \omega^2 + (k_1 + k_3)\, a_1 - k_3 a_2\right] \sin \omega t &= 0 \\ \left[-m_2 a_2 \omega^2 + (k_2 + k_3)\, a_2 - k_3 a_1\right] \sin \omega t &= 0. \end{aligned}$$

Since this must be true for all t, the terms in brackets must be zero, which gives

$$\begin{aligned} \frac{a_1}{a_2} &= \frac{-k_3}{m_1 \omega^2 - k_1 - k_3} \\ \frac{a_1}{a_2} &= \frac{m_2 \omega^2 - k_2 - k_3}{-k_3}. \end{aligned}$$

Since these must be equal

$$\frac{-k_3}{m_1 \omega^2 - k_1 - k_3} = \frac{m_2 \omega^2 - k_2 - k_3}{-k_3},$$

which gives

$$\omega^4 + \left(\frac{k_1 + k_3}{m_1} + \frac{k_2 + k_3}{m_2}\right)\omega^2 + \frac{k_1 k_2 + k_2 k_3 + k_1 k_3}{m_1 m_2} = 0.$$

Note this is a quartic equation in ω but due to the absence of the odd powers of ω it may be considered a quadratic equation in ω^2. Although it is not necessary, to simplify things a bit, assume

$$\begin{aligned} k_1 = k_2 &= k \\ m_1 = m_2 &= m. \end{aligned} \tag{7.2}$$

Using these values

$$\begin{aligned}\frac{a_1}{a_2} &= \frac{-k_3}{m\omega^2 - k - k_3} \qquad (7.3)\\ \frac{a_1}{a_2} &= \frac{m\omega^2 - k - k_3}{-k_3},\end{aligned}$$

and

$$\omega^4 + \left(2\frac{k+k_3}{m}\right)\omega^2 + \frac{k\,(k+2k_3)}{m^2} = 0.$$

This has roots

$$\omega^2 = \frac{k+k_3}{m} \pm \sqrt{\left(\frac{k+k_3}{m}\right)^2 - \frac{k\,(k+2k_3)}{m^2}},$$

so

$$\begin{aligned}\omega^2 &= \frac{k}{m} \qquad \text{or}\\ &= \frac{k+2k_3}{m}.\end{aligned}$$

Substituting these values into Equation 7.3 gives

$$\begin{aligned}\frac{a_1}{a_2} &= 1\\ \frac{a_1}{a_2} &= -1,\end{aligned}$$

for each of the two values of ω^2 respectively.

The interpretation of these two pairs of values for ω^2 and $\frac{a_1}{a_2}$ is straightforward. Considering

$$\begin{aligned}\omega^2 &= \frac{k}{m}\\ \frac{a_1}{a_2} &= 1\end{aligned}$$

the two solutions are

$$\begin{aligned}x_1(t) &= a\cos\sqrt{\frac{k}{m}}t\\ x_2(t) &= a\cos\sqrt{\frac{k}{m}}t\end{aligned}$$

where $a_1 = a_2 = a$. Thus, the two masses move with the same frequency, in the same direction with the same magnitude of oscillation, as is schematically illustrated in Figure 7.2.

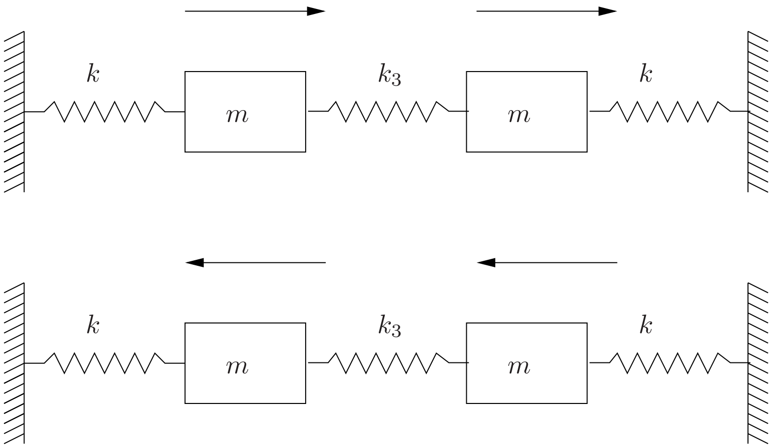


Figure 7.2. Mode one oscillations.

A similarly straight-forward analysis for the second solution shows that

$$\begin{aligned} x_1(t) &= a\cos\sqrt{\frac{k+2k_3}{m}}t \\ x_2(t) &= -a\cos\sqrt{\frac{k+2k_3}{m}}t, \end{aligned}$$

where the masses move in opposite directions, as is illustrated in Figure 7.3.

Since the system is linear, the principle of superposition applies; hence, any solution starting with zero initial velocities may be written as a combination of the two modes of oscillation

$$\begin{aligned} x_1(t) &= a\cos\sqrt{k}mt + b\cos\sqrt{k+2k_3}mt \\ x_2(t) &= a\cos\sqrt{k}mt - b\cos\sqrt{k+2k_3}mt. \end{aligned}$$

A similarly straightforward analysis starting with assumed solutions of the form

$$\begin{aligned} x_1(t) &= a_1\cos\omega t + c_1\sin\omega t \\ x_2(t) &= a_2\cos\omega t + c_2\sin\omega t \end{aligned}$$

would yield the same solutions for ω^2 and the same conditions on the relationship between the coefficients b_1 and b_2. Since the same conditions apply for b_1 and b_2, the same interpretation of the two modes applies for systems with initial velocities.

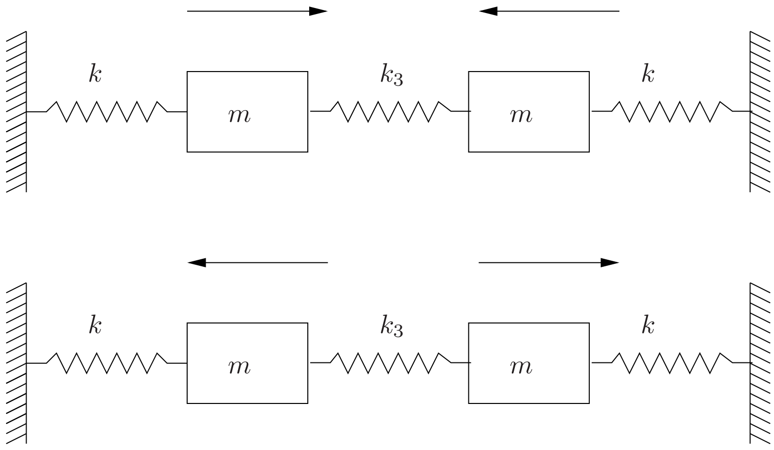


Figure 7.3. Mode two oscillations.

Hence any solution, including solutions with nonzero initial velocities, may be represented as

$$\begin{aligned} x_1(t) &= a\cos\sqrt{\frac{k}{m}}t + b\cos\sqrt{\frac{k+2k_3}{m}}t + c\sin\sqrt{\frac{k}{m}}t + d\sin\sqrt{\frac{k+2k_3}{m}}t \\ x_2(t) &= a\cos\sqrt{\frac{k}{m}}t - b\cos\sqrt{\frac{k+2k_3}{m}}t + c\sin\sqrt{\frac{k}{m}}t - d\sin\sqrt{\frac{k+2k_3}{m}}t, \end{aligned}$$

where the coefficients a, b, c and d depend upon the initial conditions.

Eigenvalue/Eigenvector Approach

Considering the equations of motion for the system illustrated in Figure 7.1, which are given by Equation 7.1, if

$$\begin{aligned} \xi_1 &= x_1 \\ \xi_2 &= \dot{x}_1 \\ \xi_3 &= x_2 \\ \xi_4 &= \dot{x}_2, \end{aligned}$$

and the simplifications given in Equation 7.2 hold, then

$$\dot{\xi} = \frac{d}{dt}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k+k_3}{m} & 0 & \frac{k_3}{m} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{k_3}{m} & 0 & -\frac{k+k_3}{m} & 0 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{bmatrix} = A\xi.$$

The eigenvalues of A are determined by the cofactor expansion

$$
\begin{aligned}
|A-\lambda I| &= \begin{vmatrix} -\lambda & 1 & 0 & 0 \\ -\frac{k+k_3}{m} & -\lambda & \frac{k_3}{m} & 0 \\ 0 & 0 & -\lambda & 1 \\ \frac{k_3}{m} & 0 & -\frac{k+k_3}{m} & -\lambda \end{vmatrix} \\
&= -\lambda \begin{vmatrix} -\lambda & \frac{k_3}{m} & 0 \\ 0 & -\lambda & 1 \\ 0 & -\frac{k+k_3}{m} & -\lambda \end{vmatrix} + (-1) \begin{vmatrix} -\frac{k+k_3}{m} & \frac{k_3}{m} & 0 \\ 0 & -\lambda & 1 \\ \frac{k_3}{m} & -\frac{k+k_3}{m} & -\lambda \end{vmatrix} \\
&= \lambda^4 + 2\frac{k+k_3}{m}\lambda^2 + \left(\frac{k+k_3}{m}\right)^2 - \left(\frac{k_3}{m}\right)^2 \\
&= 0.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\lambda_1 &= i\sqrt{\frac{k}{m}} \\
\lambda_2 &= -i\sqrt{\frac{k}{m}} \\
\lambda_3 &= i\sqrt{\frac{k+2k_3}{m}} \\
\lambda_4 &= -i\sqrt{\frac{k+2k_3}{m}}.
\end{aligned}
$$

Now computing the eigenvectors gives

$$(A-\lambda_1 I)\,\hat{\xi}_1 = 0 \quad \Longleftrightarrow \quad \left[\begin{array}{cccc|c} -i\sqrt{\frac{k}{m}} & 1 & 0 & 0 & 0 \\ -\frac{k+k_3}{m} & -i\sqrt{\frac{k}{m}} & \frac{k_3}{m} & 0 & 0 \\ 0 & 0 & -i\sqrt{\frac{k}{m}} & 1 & 0 \\ \frac{k_3}{m} & 0 & -\frac{k+k_3}{m} & -i\sqrt{\frac{k}{m}} & 0 \end{array}\right]$$

$$\text{multiply first row by } -\frac{\frac{k+k_3}{m}}{i\sqrt{\frac{k}{m}}} \text{ and add to second row} \quad \Longrightarrow \quad \left[\begin{array}{cccc|c} -i\sqrt{\frac{k}{m}} & 1 & 0 & 0 & 0 \\ 0 & i\frac{k_3}{\sqrt{km}} & \frac{k_3}{m} & 0 & 0 \\ 0 & 0 & -i\sqrt{\frac{k}{m}} & 1 & 0 \\ \frac{k_3}{m} & 0 & -\frac{k+k_3}{m} & -i\sqrt{\frac{k}{m}} & 0 \end{array}\right]$$

$$\text{multiply first row by } \frac{\frac{k_3}{m}}{i\sqrt{\frac{k}{m}}} \text{ and add to fourth row} \quad \Longrightarrow \quad \left[\begin{array}{cccc|c} -i\sqrt{\frac{k}{m}} & 1 & 0 & 0 & 0 \\ 0 & i\sqrt{\frac{k_3}{km}} & \frac{k_3}{m} & 0 & 0 \\ 0 & 0 & -i\sqrt{\frac{k}{m}} & 1 & 0 \\ 0 & -i\sqrt{\frac{k_3}{km}} & -\frac{k+k_3}{m} & -i\sqrt{\frac{k}{m}} & 0 \end{array}\right]$$

$$\text{add second row to fourth row} \quad \Longrightarrow \quad \left[\begin{array}{cccc|c} -i\sqrt{\frac{k}{m}} & 1 & 0 & 0 & 0 \\ 0 & -i\sqrt{\frac{k_3}{km}} & \frac{k_3}{m} & 0 & 0 \\ 0 & 0 & -i\sqrt{\frac{k}{m}} & 1 & 0 \\ 0 & 0 & -\frac{k}{m} & -i\sqrt{\frac{k}{m}} & 0 \end{array}\right]$$

$$\text{multiply third row by } -\frac{\frac{k}{m}}{i\sqrt{\frac{k}{m}}} \text{ and add to fourth row} \quad \Longrightarrow \quad \left[\begin{array}{cccc|c} -i\sqrt{\frac{k}{m}} & 1 & 0 & 0 & 0 \\ 0 & -i\sqrt{\frac{k_3}{km}} & \frac{k_3}{m} & 0 & 0 \\ 0 & 0 & -i\sqrt{\frac{k}{m}} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Thus,

$$\hat{\xi}_1 = \begin{bmatrix} -i \\ \sqrt{\frac{k}{m}} \\ -i \\ \sqrt{\frac{k}{m}} \end{bmatrix}.$$

Similar computations show that

$$\hat{\xi}_2 = \begin{bmatrix} i \\ \sqrt{\frac{k}{m}} \\ i \\ \sqrt{\frac{k}{m}} \end{bmatrix} \qquad \hat{\xi}_3 = \begin{bmatrix} 1 \\ i\sqrt{\frac{k+2k_3}{m}} \\ -1 \\ -i\sqrt{\frac{k+2k_3}{m}} \end{bmatrix} \qquad \hat{\xi}_4 = \begin{bmatrix} 1 \\ -i\sqrt{\frac{k+2k_3}{m}} \\ -1 \\ i\sqrt{\frac{k+2k_3}{m}} \end{bmatrix}.$$

The important point of this example is two fold:

1. the eigenvalues are exactly the same as the frequencies computed using the classical method; and,
2. the eigenvectors reflect the relative magnitude conditions as well; *i.e.*, in particular
   (a) the first and third components of $\hat{\xi}_1$ and $\hat{\xi}_2$ are identical, which is a consequence of the fact that $\frac{a_1}{a_2} = 1$ in the case where the frequency is $\sqrt{\frac{k}{m}}$; and,

   (b) the first and third components of $\hat{\xi}_3$ and $\hat{\xi}_4$ have the same magnitude but opposite sign, which is a consequence of the fact that $\frac{a_1}{a_2} = -1$ in the case where the frequency is $\sqrt{\frac{k+2k_3}{m}}$.

7.4 Undamped Structural Dynamics

Consider the structure illustrated in Figure **??**. It models a building with n floors. Assume that the ith floor of the building has a mass of m_i and that the mass of the floors is much greater than the mass of the supporting columns, so that the mass of the columns may be neglected.

7.5 Introduction to “Modern” Control

7.5.1 Controllability and observability

7.5.2 Pole placement

7.5.3 The linear quadratic regulator

7.6 Exercises

Problem 7.1 Determine the equations of motion for the system illustrated in Figure 7.4.

1. Write the equations as a set of two, second order ordinary differential equations.
2. Write the equations as a set of four, first order ordinary differential equations of the form
$$\dot{\xi} = A\xi$$
   where $\xi \in \mathbb{R}^4$ and $A \in \mathbb{R}^{4\times 4}$.

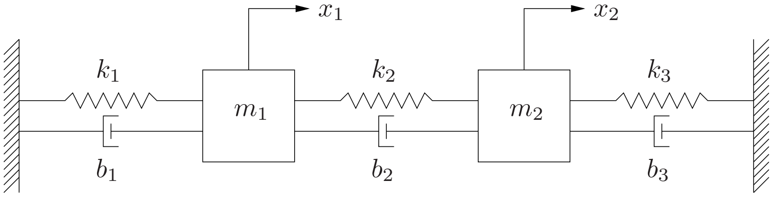


Figure 7.4. Two mass sysetm for Problem 7.1.

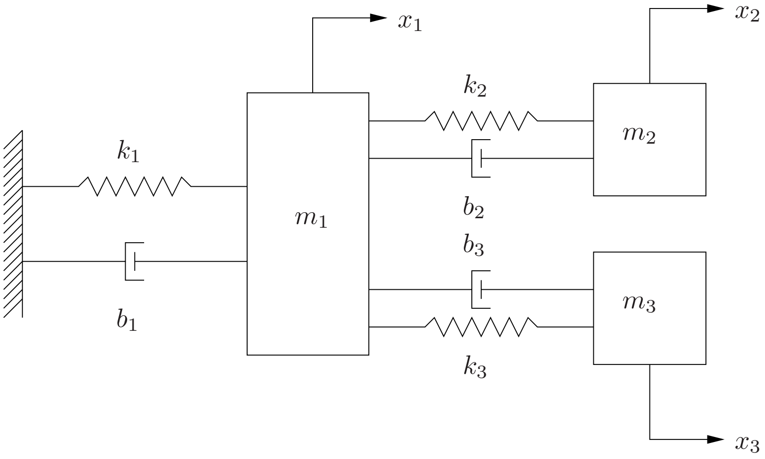


Figure 7.5. Three mass sysetm for Problem 7.2.

Problem 7.2 Determine the equations of motion for the system illustrated in Figure 7.5.

1. Write the equations as a set of three, second order ordinary differential equations.

2. Write the equations as a set of six, first order ordinary differential equations of the form
$$\dot{\xi} = A\xi$$
where $\xi \in \mathbb{R}^6$ and $A \in \mathbb{R}^{6\times 6}$.

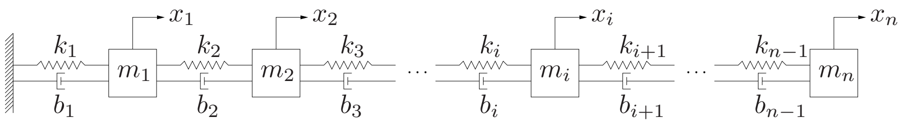


Figure 7.6. Sysetm with n masses for Problem 7.3.

Problem 7.3 Determine the equations of motion for the system illustrated in Figure 7.6.

1. Write the second order differential equation which is the equation of motion for masses 1, 2, i, and n.

2. Write the equations in the form

$$\dot{\xi} = A\xi$$

   where $\xi \in \mathbb{R}^{2n}$ and $A \in \mathbb{R}^{2n\times 2n}$. Since n is not specificed, it is acceptable for the matrix A to contain ellipses.

Problem 7.4 Consider the system illustrated in Figure 7.7.

1. Determine the equations of motion if x_1 and x_2 are measured from the unstretched position of the springs.

2. Determine the equations of motion if x_1 and x_2 are measured from the equilibrium position of the masses.

Problem 7.5 Consider the system with 10 masses illustrated in Figure 7.8. Assume that all the masses have a mass of one and all the spring have a spring constant of one except the spring between the second to last and last mass which has a spring constant of five. Assume the system starts with zero initial conditions.

1. Determine the equations of motion for the system and convert them to the form

$$\dot{\xi} = A\xi + g(t).$$

   Compute the eigenvalues and eigenvectors of the matrix A. You may use a computer program to do this computation.

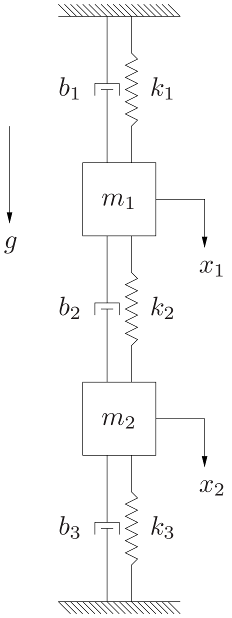


Figure 7.7. Two mass sysetm for Problem 7.4.

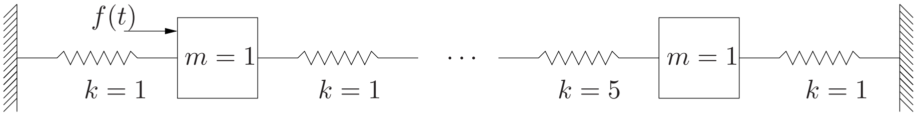


Figure 7.8. 10 mass system for Probelm 7.5.

2. Write a computer program to determine an approximate numerical solution for the system when

$$f(t) = \sin \omega t$$

for the cases where

$$\begin{aligned} \omega &= 0.25 \\ \omega &= 1.00 \\ \omega &= 1.97. \end{aligned}$$

Compare the response of the system for the three different frequencies and explain any significant differences. Relate these differences to the eigenvalues and eigenvectors of A.

Problem 7.6 Consider the structure illustrated in Figure **??**. Assume that all the masses have a mass of one and all the spring have a spring constant of one. Assume the system starts with zero initial conditions.

1. Determine the equations of motion for the system and convert them to the form

$$\dot{\xi} = A\xi + g(t).$$

Compute the eigenvalues and eigenvectors of the matrix A. You may use a computer program to do this computation.

2. Write a computer program to determine an approximate numerical solution for the system when

$$f(t) = \sin \omega t$$

for the cases where

$$\begin{aligned} \omega &= 0.25 \\ \omega &= 1.00 \\ \omega &= 1.97. \end{aligned}$$

Compare the response of the system for the three different frequencies and explain any significant differences. Relate these differences to the eigenvalues and eigenvectors of A.

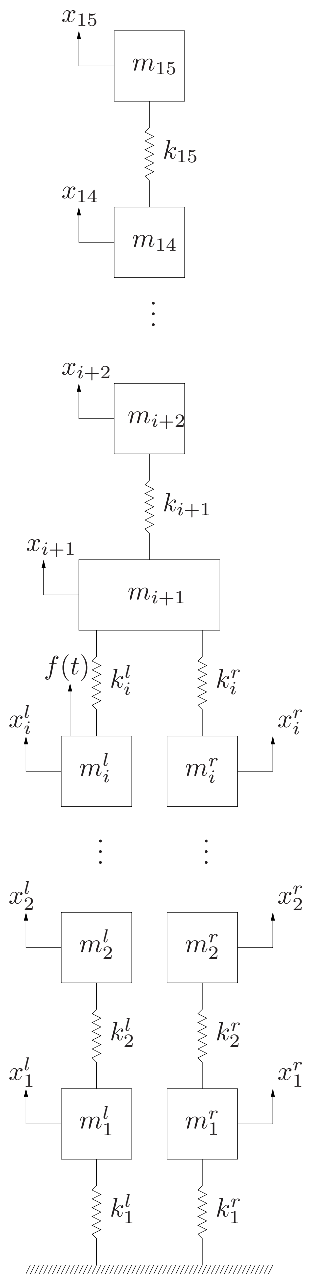


Figure 7.9. Structure for Probelm 7.6.

Chapter 8

The Laplace Transform

The Laplace transform is an integral transformation that converts solving ordinary differential equations into solving a system of *algebraic* equations. Various types of integral transform methods exist, but due to its central role in control theory, this text will focus on Laplace transforms.

8.1 Motivational Example

Integral transform methods are sufficiently abstract that it may be useful to demonstrate their utility up front. The steps involved with the following example will not be the obvious ones to the uninitiated, but nonetheless are intended to illustrate that

1. it is a way to solve linear, constant coefficient ordinary differential equations; and,

2. if one can tolerate the "overhead" of computing the transforms, it converts solving a differential equation into *algebra*.

Example 8.1.1 Consider

$$\begin{aligned} \dot{x} + 2x &= 6e^{4t} \\ x(0) &= 2. \end{aligned}$$

Let us start the exercise by stating two facts, both of which have some unstated assumptions that will be addressed subsequently.

Fact 8.1.2

$$\int_0^\infty e^{at}e^{-st}dt = \frac{1}{s+a}. \qquad \diamond$$

Fact 8.1.3

$$\int_0^\infty \frac{dx(t)}{dt} e^{-st} dt = s \int_0^\infty x(t) e^{-st} dt - x(0).$$

◇

Both of these facts can be verified by simply evaluating the integrals.

Returning to the problem at hand, multiply each side of the differential equation by e^{-st} and integrate from 0 to ∞

$$\begin{aligned}\int_0^\infty e^{-st}\left(\frac{dx(t)}{dt} + 2x(t)\right) dt &= \int_0^\infty \frac{dx(t)}{dt} e^{-st} dt + 2\int_0^\infty x(t) e^{-st} dt \\ &= 6\int_0^\infty e^{4t} e^{-st} dt.\end{aligned}$$

Clearly, the whole point of the exercise is to find $x(t)$, so there is not too much that can be done with the right hand side of the first equation except to get rid of the derivative of $x(t)$ in the first integral by making use of fact 8.1.3. Also, since we do not know what $x(t)$ is, for the time being, let

$$X(s) = \int_0^\infty x(t) e^{-st} dt.$$

Note that the second equation can be evaluated using fact 8.1.2, so

$$6\int_0^\infty e^{4t} e^{-st} dt = \frac{6}{s-4}.$$

Substituting these into the original differential equation gives

$$sX(s) - x(0) + 2X(s) = \frac{6}{s-4} \tag{8.1}$$

and substituting for $x(0)$ and solving for $X(s)$ gives

$$X(s) = \frac{1}{s+2}\left(\frac{6}{s-4} + 2\right) = \frac{2s-2}{(s-4)(s+2)} = \frac{1}{s+2} + \frac{1}{s-4}. \tag{8.2}$$

Referring to fact 8.1.2 it is clear that the right hand side of this equation is simply the same transform (multiply by e^{-st} and integrate) that was originally used on the differential equation of the sum of two exponentials. Hence, it is reasonable to assume that

$$x(t) = e^{-2t} + e^{4t}$$

is the solution to the differential equation. A quick substitution shows that indeed it satisfies the differential equation as well as the initial condition. ■

There will be a few important details added subsequently, but for purposes of this example take note that the integral

$$F(s) = \int_0^\infty f(t) e^{-st} dt$$

is called the Laplace transform of $f(t)$.

Observe the following

1. Much like in fact 8.1.2, for a given function, $f(t)$, the Laplace transform only needs to be computed one time. Hence, tables of Laplace transforms may be compiled that essentially eliminate the need for actually evaluating the integrals most of the time.

2. Once the equation was fully transformed, which is represented in equation 8.1, solving for $X(s)$ *was simply algebra!*

3. Converting from $X(s)$ back to $x(t)$ was simply a matter of determining which functions transformed to $x(t)$. Hence, this step too, can frequently be handled by tables.

4. The initial condition was handled automatically.

So, it is clearly justified to conclude that as long as the work involved in appropriately transforming the differential equation and then untransforming it at the end is not to great, this is a handy way to solve at least some types of differential equations. The general manner in which to do this will be outlined subsequently. However, before that a short review of a related concept, Fourier transforms, is in order.

8.2 Fourier Transforms

This chapter starts with a brief description of the Fourier transform. This is not strictly necessary for the Laplace transform material that follows, but since many students may already be familiar with it and it is a bit easier to understand than the Laplace transform it will be included here.

First, recall the definition of an improper integral

$$\int_a^\infty f(t)dt = \lim_{b\to\infty}\int_a^b f(t)dt.$$

Similarly for the lower limit of integration

$$\int_{-\infty}^b f(t)dt = \lim_{a\to-\infty}\int_a^b f(t)dt,$$

and

$$\int_{-\infty}^\infty f(t)dt = \lim_{a\to-\infty}\lim_{b\to\infty}\int_a^b f(t)dt.$$

Definition 8.2.1 For a function, $f(t)$, the *Fourier transform* is given by

$$\mathcal{F}(\omega) = \int_{-\infty}^\infty f(t)e^{i\omega t}dt$$

if the integral converges. ⋄

Note, by Euler's formula, the Fourier transform may also be written as

$$\mathcal{F}(\omega) = \int_{-\infty}^{\infty} f(t)\left(\cos \omega t + i \sin \omega t\right) dt.$$

Using this expression, the usual interpretation of the Fourier transform as providing the "frequency content" of the signal, $f(t)$ is obvious. For a given ω, the cosine and sine functions will be in phase with the components of the signal of $f(t)$ which have the same frequency and thus integrate to some non-zero value. For a given ω if there is no component of the signal $f(t)$ with that frequency, the integral will be zero. The relative contribution of the real and imaginary components of the transform will give the phase of a given frequency in the signal $f(t)$.

Just for completeness, the inverse Fourier transform is given by

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{F}(\omega) e^{i\omega t} d\omega.$$

8.3 Laplace Transforms

This section defines the Laplace transform and considers some of its properties.

Definition 8.3.1 Define the *Laplace transform* of a function $f(t)$ to be

$$F(s) = \int_{0^-}^{\infty} f(t) e^{-st} dt,$$

where $s \in \mathbb{C}$, *i.e., s* is a complex number. $\diamond$

First, we will clarify some notation. With respect to the limits of integration of the Laplace transform, define an integral of a function with lower limit 0^- and upper limit ∞ to be

$$\int_{0^-}^{\infty} f(t) dt = \lim_{\epsilon \downarrow 0} \int_{-\epsilon}^{\infty} f(t) dt,$$

where the notation $\lim_{\epsilon \downarrow 0}$ means that the limit approaches 0 from the right, *i.e.,* positive values or "from above" so the lower limit of integration approaches zero from below. The reason for having the lower limit be 0^- instead of simply 0 is because if something interesting, such as an impulse, occurs exactly at $t = 0$, having the lower limit equal to 0 is ambiguous as to whether or not that effect is included in the integral.

Second, with respect to the variable s, since it is a complex number it has a real and imaginary part. If it is denoted by $s = \sigma + i\omega$, then the Laplace transform becomes

$$F(s) = \int_{0^-}^{\infty} f(t) e^{-\sigma t}\left(\cos \omega t + i \sin \omega t\right) dt.$$

So, one way to interpret the Laplace transform is that it is similar to the Fourier transform in that it provides some information about the frequency content of $f(t)$, but has, for positive values of σ a multiplicative decaying exponential term.

Since the Laplace transform is a transform, we will frequently use an operator notation to represent it. If we are considering the function $f(t)$, the Laplace transform will be denoted by $\mathcal{L}$, *i.e.,*

$$F(s) = \mathcal{L}\left(f(t)\right) = \int_{0^-}^{\infty} e^{-st} f(t) dt.$$

The fundamental concept to keep in mind regarding the transformation is that it transforms the function from the time domain, t to the frequency domain, s.

The Laplace transform has an inverse. This is important because it guarantees that there is one and only one $F(s)$ corresponding to $\mathcal{L}\left(f(t)\right)$, so if we use the Laplace transform of a function to solve a differential equation, it will correspond to the unique solution.

Definition 8.3.2 The inverse Laplace transform is given by

$$f(t) = \mathcal{L}^{-1}\left(F(s)\right) = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} F(s) e^{st} ds$$

where σ is a real number such that $F(s)$ converges. Typically this will require that σ be larger than the real part of all values of s for which the denominator of $F(s)$ is equal to zero. ⋄

As will be clear subsequently, the values of s for which the denominator and numerator of $F(s)$ are zero provide almost all the essential information we need regarding the properties of the time domain function $f(t) = \mathcal{L}^{-1}\left(F(s)\right)$. For example, referring back to example 8.1.1, observe that the values for which the denominator of $X(s)$ in equation 8.2 is equal to zero are $s = -2$ and $s = 4$. It is no coincidence that these are exactly the values of the coefficients in the exponents of the time domain answer

$$x(t) = e^{-2t} + e^{4t}.$$

Because we will refer to these values frequently, they are given names.

Definition 8.3.3 The values of s for which the denominator of $F(s)$ is equal to zero are called the *poles of* $F(s)$. ⋄

Definition 8.3.4 The values of s for which the numerator of $F(s)$ is equal to zero are called the *zeros of* $F(s)$. ⋄

Example 8.3.5 The frequency domain function

$$F(s) = \frac{s+2}{(s-3)(s+10)}$$

has one zero at $s = -2$ and two poles, one at $s = 3$ and one at $s = -10$. ■

Of course, the poles and zeros may occur at values where s is complex, as is illustrated by the following example.

Example 8.3.6 The function

$$F(s) = \frac{s+a}{(s+a)^2 + b^2}$$

has a zero at $s = -a$ and two poles that comprise a complex-conjugate pair at $s = -a \pm ib$. ■

8.3.1 The Laplace Transform of Some Common Functions

This section will compute the Laplace transform of some functions common in engineering. Unless otherwise stated, we will make the following assumption for all computations regarding Laplace transforms.

Assumption 8.3.7 *In this book, whenever a Laplace transform or inverse Laplace transform is computed, the values for s are assumed to be such that all the required integrals converge.*

Example 8.3.8 The Laplace transform of $f(t) = e^{at}$ is

$$\begin{aligned}
\mathcal{L}\left(e^{at}\right) &= \int_{0^-}^{\infty} e^{-st} e^{at} dt \\
&= \int_{0^-}^{\infty} e^{(a-s)t} \\
&= \left.\frac{1}{a-s} e^{(a-s)t}\right|_0^\infty \\
&= \frac{1}{a-s}(0-1) \\
&= \frac{1}{s-a}.
\end{aligned}$$

Hence

$$\mathcal{L}\left(e^{at}\right) = \frac{1}{s-a}.$$

■

With regard to Assumption 8.3.7, note that the upper limit of integration only converges if the real part of s is greater than a. While this is a mathematical necessity, it is one that fortunately rarely concerns us in application and, consistent with the assumption, we will assume that the values of s are appropriately restricted. Henceforth, unless it is necessary, we will implicitly assume whatever restriction is necessary for convergence and only explicitly deal with it if it is necessary.

Example 8.3.9 Compute the Laplace transform of $f(t) = \sin \omega t$. We want to evaluate

$$\mathcal{L}\left(\sin \omega t\right) = \int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt.$$

Integrating once by parts gives

$$\begin{aligned} \int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt &= \left(-\frac{1}{\omega} e^{-st} \cos\left(\omega t\right)\right)\Bigg|_{0^-}^{\infty} - \frac{s}{\omega} \int_{0^-}^{\infty} e^{-st} \cos\left(\omega t\right) dt \\ &= \frac{1}{\omega} - \frac{s}{\omega} \int_{0^-}^{\infty} e^{-st} \cos\left(\omega t\right) dt. \qquad (8.3) \end{aligned}$$

Integrating the last term by parts gives

$$\begin{aligned} \int_{0^-}^{\infty} e^{-st} \cos \omega t dt &= \left(\frac{1}{\omega} e^{-st} \sin\left(\omega t\right)\right)\Bigg|_{0^-}^{\infty} + \frac{s}{\omega} \int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt \\ &= \frac{s}{\omega} \int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt. \end{aligned}$$

Substituting this into equation 8.3 gives

$$\int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt = \frac{1}{\omega} - \frac{s^2}{\omega^2} \int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt,$$

and solving for the original integral, $\mathcal{L}\left(\sin \omega t\right)$ gives

$$\int_{0^-}^{\infty} e^{-st} \sin\left(\omega t\right) dt = \frac{\frac{1}{\omega}}{1 + \frac{s^2}{\omega^2}} = \frac{\omega}{\omega^2 + s^2}.$$

■

One set of functions that may appear in differential equations for which the Laplace transform is particularly useful are those with discontinuities. So next we consider step functions and impulses.

Definition 8.3.10 The function

$$f(t) = \mathbb{1}(t) = \left\{ \begin{array}{ll} 0 & t < 0 \\ 1 & t \geq 0 \end{array} \right.$$

is called the *step function.* ⋄

The step function is illustrated in Figure 8.1. It will be useful in two ways. First, it is very common in controls since it represents the situation when some control command is activated, *i.e.*, at $t = 0$ the control command switches from "off" to "on." Second, it will allow us to easily piece together some discontinuous functions.

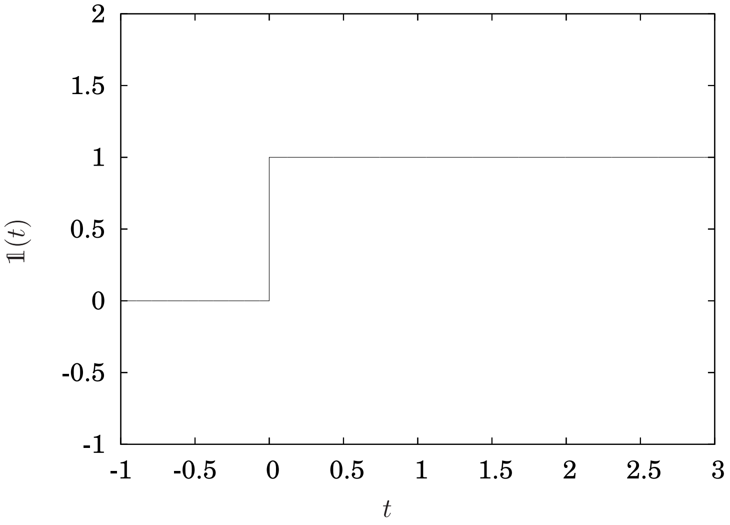


Figure 8.1. The step function, $\mathbb{1}(t)$.

Example 8.3.11 Compute the Laplace transform of the step function. Evaluating the transform gives

$$\begin{aligned}\int_{0^-}^{\infty} e^{-st}\mathbb{1}(t)dt &= \int_{0^-}^{\infty} e^{-st}dt \\ &= \frac{1}{-s}\left(e^{-st}\right)\Big|_{0^-}^{\infty} \\ &= \frac{1}{-s}(0-1) \\ &= \frac{1}{s}.\end{aligned}$$

We will occasionally need step functions where the discontinuity does not occur at zero. Note that the function $\mathbb{1}(t-\tau)$ will have the discontinuity occur at time $t=\tau$. A plot of $\mathbb{1}(t-1.5)$ is illustrated in Figure 8.2. Note that the proper interpretation of $\mathbb{1}(t-\tau)$ is that the function $1(t)$ is shifted by an amount τ. We will consider time shifts of arbitrary functions in Section 8.3.2.

Example 8.3.12 Compute the Laplace transform of $f(t)=\mathbb{1}(t-\tau)$. As-

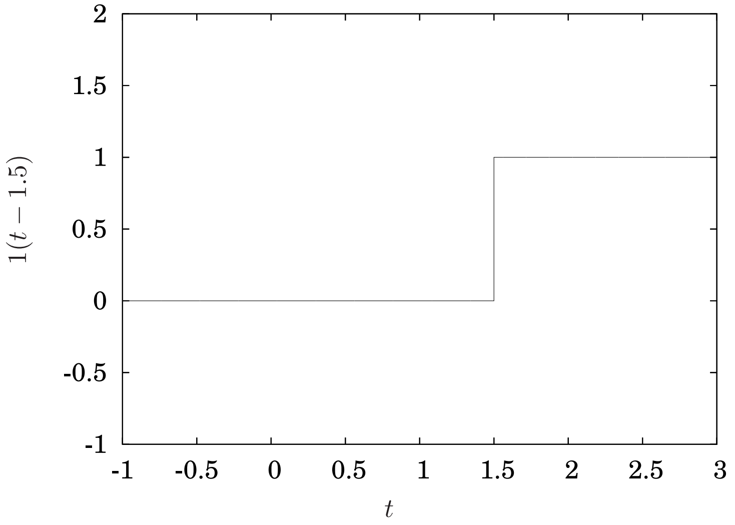


Figure 8.2. The step function, $\mathbb{1}(t - 1.5)$.

suming $\tau \geq 0$ substituting into the definition of the Laplace transform gives

$$\begin{aligned}
\int_{0^-}^{\infty} e^{-st}\mathbb{1}\left(t-\tau\right)dt &= \int_{0^-}^{\tau} 0e^{-st}dt + \int_{\tau}^{\infty} 1e^{-st}dt \\
&= \frac{1}{-s}\left(e^{-st}\right)\Big|_{\tau}^{\infty} \\
&= \frac{1}{-s}\left(0 - e^{s\tau}\right) \\
&= e^{-s\tau}\frac{1}{s}.
\end{aligned}$$

Another object that is elegantly handled by Laplace transform is the impulse. It provides a manner to model, for example, extremely large forces that occur over a very short period of time. An example of an impulse would be the force exerted by a bat on a ball.

Definition 8.3.13 Consider the function

$$\delta_\epsilon(t) = \begin{cases} \frac{1}{2\epsilon} & |t| \leq \epsilon \\ 0 & |t| > \epsilon \end{cases}$$

which is illustrated in Figure 8.3 for various values of ϵ. Define

$$\delta(t) = \lim_{\epsilon \to 0} \delta_\epsilon(t).$$

$\diamond$

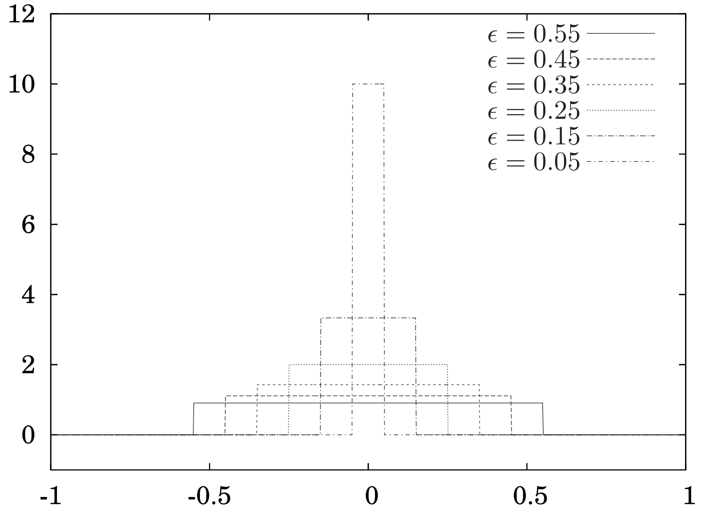


Figure 8.3. Series of functions leading to the definition of an impulse.

Note that while $\delta(t)$[1] is zero everywhere except at the origin, it still satisfies

$$\int_{-\infty}^{\infty} \delta(t)dt = 1,$$

and then furthermore for any function, $f(t)$,

$$\int_{-\infty}^{\infty} \delta(t)f(t)dt = f(0),$$

and similarly if shifted

$$\int_{-\infty}^{\infty} \delta(t-\tau)f(t)dt = f(\tau).$$

Example 8.3.14 Compute the Laplace transform of $f(t) = \delta(t)$. Substituting into the definition of the Laplace transform gives

$$\int_{0^-}^{\infty} e^{-st}\delta(t)dt = e^0 = 1.$$

■

[1] While $\delta(t)$ is commonly called the *delta function* or *Dirac delta function*, it is *not* a function. This is because it is zero everywhere except precisely at the point where we care about it, which is $t = 0$. A reader interested in pursuing the matter further is referred to [19].

$f(t), t \geq 0$	$F(s)$
$\delta(t)$	1
$\mathbb{1}(t)$	$\frac{1}{s}$
t	$\frac{1}{s^2}$
t^2	$\frac{2!}{s^3}$
t^3	$\frac{3!}{s^4}$
t^m	$\frac{m!}{s^{m+1}}$
e^{-at}	$\frac{1}{s+a}$
te^{-at}	$\frac{1}{(s+a)^2}$
$\frac{1}{2!}t^2e^{-at}$	$\frac{1}{(s+a)^3}$
$\frac{1}{(m-1)!}t^{m-1}e^{-at}$	$\frac{1}{(s+a)^m}$
$1 - e^{-at}$	$\frac{a}{s(s+a)}$
$\frac{1}{a}(at - 1 + e^{-at})$	$\frac{a}{s^2(s+a)}$
$e^{-at} - e^{-bt}$	$\frac{b-a}{(s+a)(s+b)}$
$(1 - at)e^{-at}$	$\frac{s}{(s+a)^2}$
$1 - e^{-at}(1 + at)$	$\frac{a^2}{s(s+a)^2}$
$be^{-bt} - ae^{-at}$	$\frac{(b-a)s}{(s+a)(s+b)}$
$\sin at$	$\frac{a}{s^2+a^2}$
$\cos at$	$\frac{s}{s^2+a^2}$
$e^{-at}\cos bt$	$\frac{s+a}{(s+a)^2+b^2}$
$e^{-at}\sin bt$	$\frac{b}{(s+a)^2+b^2}$
$t \sin at$	$\frac{2as}{(s^2+a^2)^2}$
$t \cos at$	$\frac{s^2-a^2}{(s^2+a^2)^2}$
$1 - e^{at}\left(\cos bt + \frac{a}{b}\sin bt\right)$	$\frac{a^2+b^2}{s[(s+a)^2+b^2]}$

Table 8.1. Table of Laplace transform pairs.

The reason that the lower limit of the integral in the definition of the Laplace transform is 0^- is so that it is clear whether or not to include impulses that occur at $t = 0$. Since the impulse has zero width, if the lower limit were simply 0, then it whether or not the impulse is included in the integral would be ambiguous.

Example 8.3.15 Compute the Laplace transform of $f(t) = \delta(t - \tau)$. Substituting into the definition of the Laplace transform gives

$$\int_{0^-}^{\infty} e^{-st}\delta(t - \tau)dt = e^{-s\tau}.$$

■

Table 8.1 summarizes the Laplace transform of some common functions in engineering.

8.3.2 Properties of the Laplace Transform

It will be useful to study the definition of the Laplace transform to determine some of its generic properties that we may exploit when using it. The first property we will consider is how the derivative of a function acts under a Laplace transform. It turns out that it is very simple and extremely useful. It is simple in that the Laplace transform transform of a derivative of a function is algebraically related to the Laplace transform of the function itself. In particular, it is simply multiplication of $F(s)$ by s. So, in the frequency domain, differentiation by t is replaced by multiplication by s. This is also its utility in that the Laplace transform then transforms differential equations into algebraic equations.

Theorem 8.3.16 *If the Laplace transform of a function, $f(t)$ is $\mathcal{L}\left(f(t)\right) = F(s)$, then*

$$\mathcal{L}\left(\frac{df(t)}{dt}\right) = sF(s) - f(0).$$

Proof The proof is simply evaluating the integral by as follows

$$\begin{aligned}\int_{0^-}^{\infty} \frac{df(t)}{dt} e^{-st} dt &= \left(e^{-st} f(t)\right)\Big|_{0^-}^{\infty} + s\int_{0^-}^{\infty} f(t)e^{-st}dt \\ &= \left(e^{-st} f(t)\right)\Big|_{0^-}^{\infty} + sF(s) \\ &= (0 - f(0)) + sF(s) \\ &= sF(s) - f(0).\end{aligned}$$

The second property we consider is called the *Final Value Theorem.* It is useful because it will allow us to determine the steady state values of a solution to a differential equation without having to compute the inverse Laplace transform.

Theorem 8.3.17 *If all the poles of $sF(s)$ are in the left half of the complex plane, then*

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} sF(s). \tag{8.4}$$

Proof Consider

$$\begin{aligned}\lim_{s\to 0}\int_{0^-}^{\infty} e^{-st}\frac{df(t)}{dt}dt &= \int_{0^-}^{\infty}\left(\lim_{s\to 0} e^{-st}\frac{df(t)}{dt}\right)dt \\ &= \int_{0^-}^{\infty}\frac{df(t)}{dt}dt \\ &= \lim_{t\to\infty} f(t) - f(0). \end{aligned} \tag{8.5}$$

Also, by theorem 8.3.16

$$\lim_{s\to 0}\int_{0^-}^{\infty} e^{-st}\frac{df(t)}{dt}dt = \lim_{s\to 0}\left(sF(s) - f(0)\right). \tag{8.6}$$

Setting equations 8.5 to 8.6 gives

$$\lim_{t\to\infty} f(t) - f(0) = \lim_{s\to 0}\left(sF(s) - f(0)\right)$$

or

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0}\left(sF(s)\right).$$

□

Another very useful property of Laplace transforms is that the shifts in time have a very simple form.

Theorem 8.3.18 *If $\mathcal{L}\left(f(t)\right) = F(s)$, then the Laplace transform of a function shifted in time satisfies*

$$\mathcal{L}\left(f(t-\tau)\mathbb{1}(t-\tau)\right) = e^{-s\tau}F(s)$$

for $\tau \geq 0$.

PROOF The proof is based upon a simple change of variable. If we let $\hat{t} = t-\tau$, then

$$\begin{aligned}
\mathcal{L}\left(f\left(t-\tau\right)\mathbb{1}(t-\tau)\right) &= \int_{0^-}^{\infty} e^{-st} f\left(t-\tau\right)\mathbb{1}(t-\tau)dt \\
&= \int_{-\tau^-}^{\infty} e^{-s\left(\hat{t}+\tau\right)} f\left(\hat{t}\right)\mathbb{1}\left(\hat{t}\right)d\hat{t} \\
&= e^{-s\tau}\int_{-\tau^-}^{\infty} e^{-s\hat{t}} f\left(\hat{t}\right)\mathbb{1}\left(\hat{t}\right)d\hat{t} \\
&= e^{-s\tau}\left(\int_{-\tau^-}^{0^-} e^{-s\hat{t}} f\left(\hat{t}\right)\mathbb{1}\left(\hat{t}\right)d\hat{t} + \int_{0^-}^{\infty} e^{-s\hat{t}} f\left(\hat{t}\right)\mathbb{1}\left(\hat{t}\right)d\hat{t}\right) \\
&= e^{-s\tau}\int_{0^-}^{\infty} e^{-s\hat{t}} f\left(\hat{t}\right)\mathbb{1}\left(\hat{t}\right)d\hat{t} \\
&= e^{-s\tau}\int_{0^-}^{\infty} e^{-s\hat{t}} f\left(\hat{t}\right)d\hat{t} \\
&= e^{-s\tau}F(s).
\end{aligned}$$

The proper interpretation of Theorem 8.3.18 takes some care, especially with respect to the step function appearing in it. Figure 8.4 illustrates a function as well as that function shifted by an amount $\tau \approx 0.75$. Since the lower limit of the Laplace transform is $t = 0^-$, the values for $f(t)$ for $t < 0$ and the values do not affect the Laplace transform. Mathematically, $\mathcal{L}\left(f(t)\mathbb{1}(t)\right) = \mathcal{L}\left(f(t)\right)$.

When $f(t)$ is shifted by a positive τ, then we need to either account for the part of $f(t)$ shifted into positive times, or exclude it. If we want to include it, then we must reevaluate the integral in the transform, because $F(s)$ only contains information about $f(t)$ for positive time and $e^{-s\tau}$ does not depend on $f(t)$, and hence contains no information regarding $f(t)$. If we do want to use $F(s)$ and not evaluate the integral, then we must exclude the part of $f(t)$ shifted into positive time. This is accomplished by multiplying $f(t-\tau)$ by $\mathbb{1}(t-\tau)$ since the step function will be zero for $t < \tau$, which corresponds exactly to the part of $f(t)$ which $F(s)$ does not represent.

So, the functions to which Theorem 8.3.18 applies are illustrated in Figure 8.5. The portion of $f(t)$ for $t \geq 0$ is shifted by an amount τ, but for $t < \tau$,

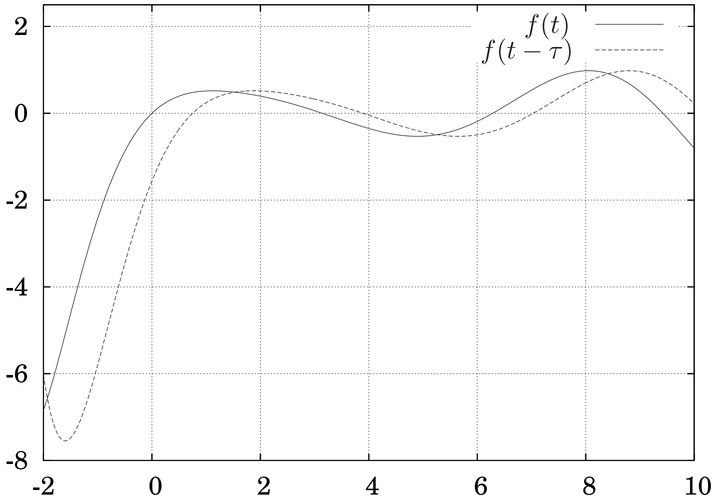


Figure 8.4. A function, $f(t)$ compared to $f(t-\tau)$.

the shifted function must be zero. This fact appears in the proof of Theorem 8.3.18 in the line where the integral with lower limit τ^- and upper limit 0^- is evaluated to zero.

Finally we will consider units. From the definition of the Laplace transform of a function, $f(t)$,

$$F(s) = \int_{0^-}^{\infty} f(t)e^{-st}dt,$$

since t has units of seconds, $F(s)$ will have the units of $f(t)$ times seconds. Since the exponent of e must be dimensionless, s must have units of one divided by seconds. Of course, there is the possibility for much confusion here since we are using the symbol s to represent the argument of the function $F(s)$ as well as for the units of time, which is seconds.

Example 8.3.19 Let $x(t)$ denote the position of something, with units m. Then

$$X(s) = \mathcal{L}\left\{x(t)\right\}$$

will have units m · s. ■

The derivative works as expected as an operator.

Example 8.3.20 Let $x(t)$ denote the position of something, with units m. Then

$$sX(s) - x(0) = \mathcal{L}\left\{\dot{x}(t)\right\}$$

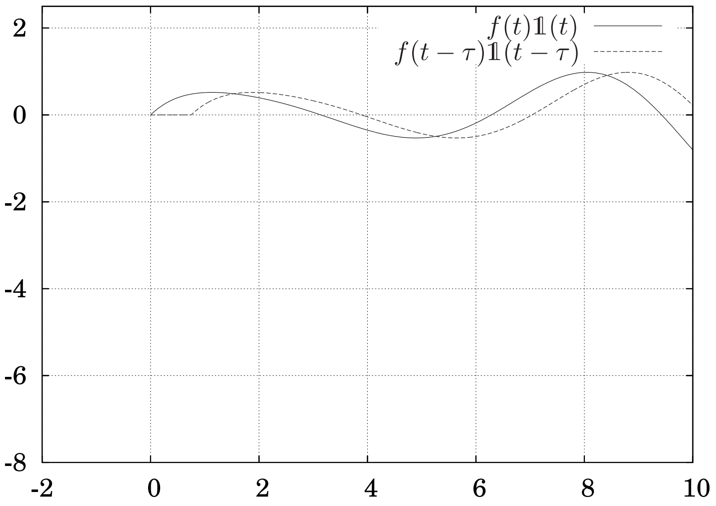


Figure 8.5. A function, $f(t)\mathbb{1}(t)$ compared to $f(t-\tau)\mathbb{1}(t-\tau)$ for which $\mathcal{L}(f(t)) = \mathcal{L}(f(t-\tau)\mathbb{1}(t-\tau))$ and Theorem 8.3.18 properly applies.

Name	Time Function	Laplace Transform
Transform pair	$f(t)$	$F(s)$
Superposition	$\alpha f_1(t) + \beta f_2(t)$	$\alpha F_1(s) + \beta F_2(s)$
Differentiation	$\frac{d^m}{dt^m} f(t)$	$s^m F(s) - s^{m-1} f(0) - s^{m-2} \dot{f}(0) -$ $\cdots - s \frac{d^{m-2}}{dt^{m-2}} f(0) - \frac{d^{m-1}}{dt^{m-1}} f(0)$
Time delay ($\tau \geq 0$)	$f(t-\tau)\mathbb{1}(t-\tau)$	$F(s)e^{-s\tau}$
Time scaling	$f(at)$	$\frac{1}{\lvert a \rvert} F\left(\frac{s}{a}\right)$
Frequency shift	$e^{-at} f(t)$	$F(s+a)$
Integration	$\int f(\xi) d\xi$	$\frac{1}{s} F(s)$
Convolution	$f_1(t) * f_2(t)$	$F_1(s) F_2(s)$
Initial Value Theorem	$f(0^+)$	$\lim_{s \to \infty} sF(s)$
Final Value Theorem	$\lim_{t \to \infty} f(t)$	$\lim_{s \to 0} sF(s)$
Time product	$f_1(t) f_2(t)$	$\frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F_1(\xi) F_s(s-\xi) d\xi$
Multiplication by time	$tf(t)$	$-\frac{d}{ds} F(s)$

Table 8.2. Properties of the Laplace transform.

will have units $\frac{\mathrm{m\cdot s}}{\mathrm{s}} = \mathrm{m}$, and

$$s^2X(s) - sx(0) - \dot{x}(0) = \mathcal{L}\left\{\ddot{x}(t)\right\}$$

will have units $\frac{\mathrm{m\cdot s}}{\mathrm{s}^2} = \frac{\mathrm{m}}{\mathrm{s}}$. ■

Just as $\frac{d}{dt}$ alters the units of $x(t)$ by dividing by s, the manner in which the Laplace transform of a derivative works is by dividing the units of $\mathcal{L}\left\{x(t)\right\}$ by s.

8.4 Solving Initial Value Problems

Laplace transforms may be used to solve initial value problems for linear, constant coefficient ordinary differential equations. There are two attributes worth noting. First, there is no need to separate the solution method into homogeneous and particular solutions. Second, the method works particularly well for system where the inhomogeneous term is discontinuous. In such a case the methods from Chapters 2 and 3 would require that we "piece together" solutions, which would amount to evaluating the constants in the homogeneous solution each time there is a discontinuity in the inhomogeneous term.

We will illustrate the means to use Laplace transforms to solve initial value problems with a few examples. The procedure is the same as in example 8.1.1, which is to Laplace transform the entire equation, algebraically solve for the dependent variable and then determine inverse Laplace transform to find the time domain function for the dependent variable.

Example 8.4.1 Find the solution to

$$\begin{aligned} \ddot{x} + 4\dot{x} + 13x &= 20\cos 5t - 12\sin 5t \\ x(0) &= 1 \\ \dot{x}(0) &= 15. \end{aligned}$$

Laplace transforming the equation gives

$$\begin{aligned} \left(s^2X(s) - sx(0) - \dot{x}(0)\right) + 4\left(sX(s) - x(0)\right) + 13X(s) &= \\ 20\frac{s}{s^2+25} \quad - \quad 12\frac{5}{s^2+25}. \end{aligned}$$

Substituting the initial conditions gives

$$\left(s^2X(s) - s - 15\right) + 4\left(sX(s) - 1\right) + 13X(s) = 20\frac{s}{s^2+25} - 12\frac{5}{s^2+25}.$$

Rearranging some gives

$$X(s)\left(s^2 + 4s + 13\right) = \frac{20s - 60}{s^2+25} + s + 19,$$

or

$$X(s) = \frac{20s-60}{(s^2+25)(s^2+4s+13)} + \frac{s+19}{s^2+4s+13}.$$

Now we want to covert the right hand side into a combination of terms that appear in Table 8.1. Attempting to factor the denominator $s^2+4s+13$ will show that it has the complex roots, $s=-2\pm 3i$, and is, by completing the square, equivalent to $(s+2)^2+9$, which is of the form of a denominator in table. So

$$X(s) = \frac{20s-60}{(s^2+25)\left((s+2)^2+9\right)} + \frac{s+19}{(s+2)^2+9}.$$

A partial fraction expansion[2], of the first term gives

$$\begin{aligned} X(s) &= \frac{as+b}{s^2+25} + \frac{cs+d}{(s+2)^2+9} + \frac{s+19}{(s+2)^2+9} \\ &= \frac{(as+b)\left(s^2+4s+13\right)+(cs+d)\left(s^2+25\right)}{(s^2+25)\left((s+2)^2+9\right)} + \frac{s+19}{(s+2)^2+9}. \end{aligned}$$

Equating numerators in the first term gives

$$\begin{aligned} (a+c)s^3+(4a+b+d)s^2+(13a+4b+25c)s &+ \\ (13b+25d) &= 20s-60 \end{aligned}$$

and some tedious algebra gives

$$\begin{aligned} a &= 0 \\ b &= 5 \\ c &= 0 \\ d &= -5. \end{aligned}$$

So,

$$\begin{aligned} X(s) &= \frac{5}{s^2+25} - \frac{5}{(s+2)^2+9} + \frac{s+19}{(s+2)^2+9} \\ &= \frac{5}{s^2+25} + \frac{s+14}{(s+2)^2+9}. \end{aligned}$$

Referring to the table, we want either $s+2$ or 3 in the numerator of the second term, so we split the second term into two terms as follows

$$\begin{aligned} X(s) &= \frac{5}{s^2+25} + \frac{s+2}{(s+2)^2+9} + \frac{12}{(s+2)^2+9} \\ &= \frac{5}{s^2+25} + \frac{s+2}{(s+2)^2+9} + 4\frac{3}{(s+2)^2+9}. \end{aligned}$$

[2] Readers not familiar with partial fractions are referred to Appendix A.3.

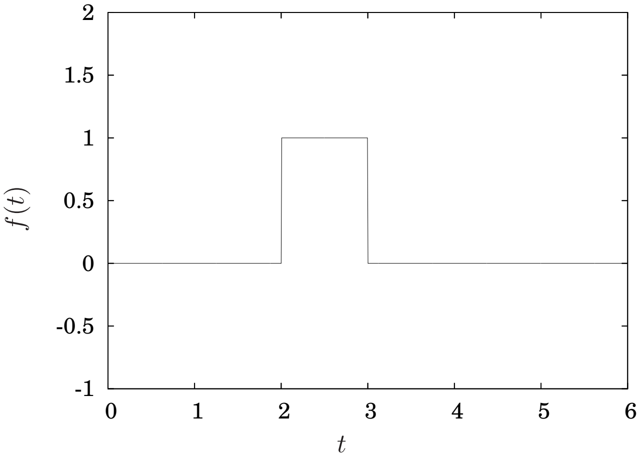


Figure 8.6. Function for Example 8.4.2.

Now all the terms are entries in Table 8.1 and the solution is

$$x(t) = \sin 5t + e^{-2t}\cos 3t + 4e^{-2t}\sin 3t. \qquad \blacksquare$$

8.4.1 Solving Differential Equations with Discontinuous Forcing

Step functions and time shifts may be combined in useful ways to easily evaluate differential equations that have inhomogeneous terms with discontinuities. The means to effectively do this will be presented as a series of examples.

Example 8.4.2 Determine the solution to

$$\begin{aligned} \dot{x} + x &= f(t) \\ x(0) &= 0 \end{aligned} \tag{8.7}$$

where

$$f(t) = \begin{cases} 1 & 2 \leq t < 3 \\ 0 & \text{otherwise} \end{cases}$$

The function $f(t)$ is illustrated in Figure 8.6.

For purposes of using the tools at our disposal to solve this differential equation, the critical observation is that we may write

$$f(t) = \mathbb{1}(t-2) - \mathbb{1}(t-3),$$

which is illustrated in Figure 8.7.

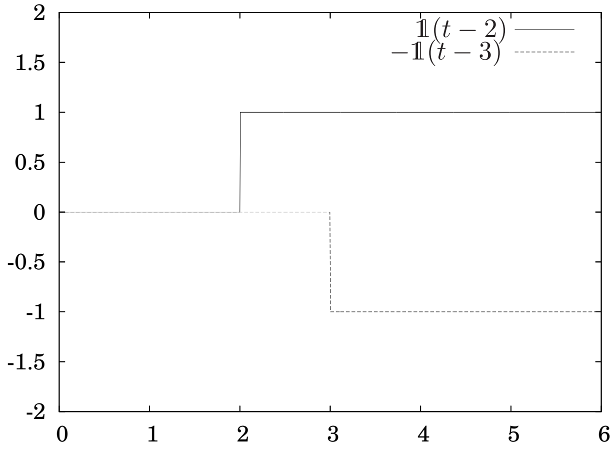


Figure 8.7. Two step function combined to give $f(t)$ in Figure 8.6 from Example 8.4.2.

So, now we simply Laplace transform

$$\begin{aligned} \dot{x} + x &= \mathbb{1}(t-2) - \mathbb{1}(t-3) \\ x(0) &= 0 \end{aligned}$$

to get

$$sX(s) + X(s) = \frac{e^{-2s}}{s} - \frac{e^{-3s}}{s}$$

and solving for $X(s)$ gives

$$X(s) = \frac{1}{s\,(s+1)} \left(e^{-2s} - e^{-3s}\right).$$

If needed we could use partial fractions to convert the fraction into terms appearing in a table; however, in this case the term itself is in Table 8.1. In particular

$$\mathcal{L}^{-1}\left(\frac{1}{s\,(s+1)}\right) = 1 - e^{-t}.$$

Hence,

$$\begin{aligned} X(s) &= \mathcal{L}\left(1-e^{-t}\right)\left(e^{-2s} - e^{-3s}\right) \\ &= e^{-2s}\mathcal{L}\left(1-e^{-t}\right) - e^{-3s}\mathcal{L}\left(1-e^{-t}\right). \end{aligned}$$

So, referring to Theorem 8.3.18 (or the corresponding entry in Table 8.2), each term that is multiplied by $e^{-\tau s}$ must have t shifted by τ, *and* must be

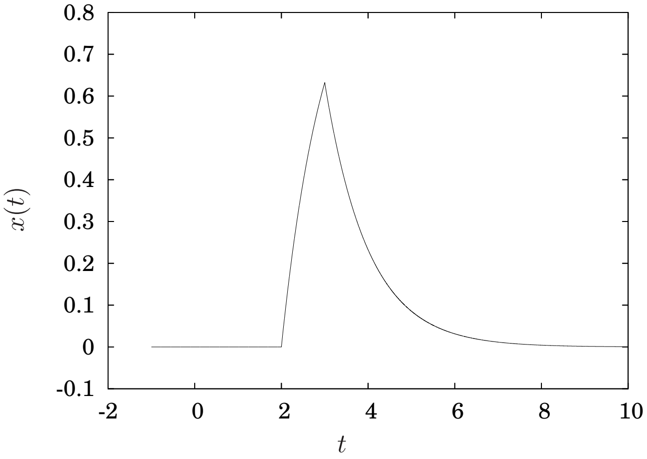


Figure 8.8. Solution for Example 8.4.2.

multiplied by $\mathbb{1}(t-\tau)$. Hence

$$x(t) = \left(1 - e^{-(t-2)}\right)\mathbb{1}(t-2) - \left(1 - e^{-(t-3)}\right)\mathbb{1}(t-3), \tag{8.8}$$

is the solution to equation 8.7. A plot of Equation 8.8 is illustrated in Figure 8.8. Written in another form this solution is

$$x(t) = \begin{cases} 0 & t < 2 \\ 1 - e^{-(t-2)} & 2 \leq t < 3 \\ e^{-(t-3)} - e^{-(t-2)} & t \geq 3 \end{cases} .$$

■

At this point we can recognize that if we are able to piece together step functions to be either one (or negative one) for specific ranges in time, then we can use such a structure to multiply other functions to have them appear for only a limited period of time. The next example illustrates that fact.

Example 8.4.3 Find the solution to

$$\begin{aligned} \dot{x} + x &= \begin{cases} 0 & t < 1 \\ 3t^2 & 1 \leq t < 2 \\ 0 & t \leq 2 \end{cases} \\ x(0) &= 0. \end{aligned}$$

We can write the inhomogeneous term as a combination of step functions as

$$\begin{aligned} \dot{x} + x &= 3t^2\left[\mathbb{1}\left(t-1\right) - \mathbb{1}\left(t-2\right)\right] \\ &= 3t^2\mathbb{1}\left(t-1\right) - 3t^2\mathbb{1}\left(t-2\right). \end{aligned}$$

If we denote $f(t) = t^2$, neither of the two terms on the right hand side are in the appropriate form to use Theorem 8.3.18. For the first one, we need

$$\begin{aligned} f(t-1) &= (t-1)^2 \\ &= t^2 - 2t + 1 \end{aligned}$$

and for the second one we need

$$\begin{aligned} f(t-2) &= (t-2)^2 \\ &= t^2 - 4t + 4. \end{aligned}$$

So, to make the equation amenable for use by Theorem 8.3.18, write

$$\begin{aligned} \dot{x} + x &= 3\left[t^2 \mathbb{1}\,(t-1) - t^2 \mathbb{1}\,(t-2)\right] \\ &= 3\left[\left((t-1)^2 + 2t - 1\right) \mathbb{1}\,(t-1) - \left((t-2)^2 + 4t - 4\right) \mathbb{1}\,(t-2)\right] \\ &= 3\left[(t-1)^2 \mathbb{1}(t-1) - (t-2)^2 \mathbb{1}(t-2) + \right. \\ &\quad \left.(2t-1)\,\mathbb{1}(t-1) - (4t-4)\,\mathbb{1}(t-2)\right]. \end{aligned}$$

The first two terms may make use of Theorem 8.3.18, but now we need to take care of the terms that were added, *i.e.,* the $2t-1$ and $4t-4$ terms. So, write

$$2t - 1 = 2(t-1) + 1$$

and

$$4t - 4 = 4(t-2) + 4$$

and substituting gives

$$\begin{aligned} \dot{x} + x &= \left[(t-1)^2 \mathbb{1}(t-1) - (t-2)^2 \mathbb{1}(t-2) + \right. \\ &\quad \left.(2(t-1)+1)\,\mathbb{1}(t-1) - (4(t-2)+4)\,\mathbb{1}(t-2)\right]. \end{aligned}$$

Let us consider this term by term using the relationship

$$\mathcal{L}\left(f(t-\tau)\mathbb{1}(t-\tau)\right) = e^{-\tau s}\mathcal{L}\left(f(t)\right).$$

1. For the first term

$$\begin{aligned} \mathcal{L}\left((t-1)^2\,\mathbb{1}(t-1)\right) &= e^{-s}\mathcal{L}\left(t^2\right) \\ &= e^{-s}\frac{2}{s^3}. \end{aligned}$$

2. For the second term

$$\begin{aligned} \mathcal{L}\left((t-2)^2\,\mathbb{1}\,(t-2)\right) &= e^{-2s}\mathcal{L}\left(t^2\right) \\ &= e^{-2s}\frac{2}{s^3}. \end{aligned}$$

3. For the third term

$$\begin{aligned}\mathcal{L}\left(\left(2\left(t-1\right)+1\right)\mathbb{1}\left(t-1\right)\right) &= e^{-s}\mathcal{L}\left(2t+1\right)\\ &= e^{-s}\left(\frac{2}{s^2}+\frac{1}{s}\right).\end{aligned}$$

4. For the last term

$$\begin{aligned}\mathcal{L}\left(\left(4(t-2)+4\right)\mathbb{1}(t-2)\right) &= e^{-2s}\mathcal{L}\left(4t+4\right)\\ &= e^{-2s}\left(\frac{4}{s^2}+\frac{4}{s}\right).\end{aligned}$$

Taking the Laplace transform of the entire equation gives

$$sX(s)+X(s)=e^{-s}\frac{2}{s^3}-e^{-2s}\frac{2}{s^3}+e^{-s}\left(\frac{2}{s^2}+\frac{1}{s}\right)-e^{-2s}\left(\frac{4}{s^2}+\frac{4}{s}\right).$$

So

$$\begin{aligned}X(s) &= e^{-s}\left(\frac{2}{s^3\left(s+1\right)}+\frac{2}{s^2\left(s+1\right)}+\frac{1}{s\left(s+1\right)}\right)-\\ &\quad e^{-2s}\left(\frac{2}{s^3\left(s+1\right)}+\frac{4}{s^2\left(s+1\right)}+\frac{4}{s\left(s+1\right)}\right).\end{aligned}$$

From Table 8.1 we can find the inverse Laplace transform of the second two terms

$$\begin{aligned}\mathcal{L}^{-1}\left(\frac{1}{s\left(s+1\right)}\right) &= 1-e^{-t}\\ \mathcal{L}^{-1}\left(\frac{1}{s^2\left(s+1\right)}\right) &= t-1+e^{-t}.\end{aligned}$$

It is left as an exercise to show that

$$\mathcal{L}^{-1}\left(\frac{1}{s^3\left(s+1\right)}\right)=\frac{1}{2}t^2-t+1-e^{-t}.$$

■

Finally, remembering to replace t by $t-1$ or $t-2$ depending on whether the Laplace transform is multiplied by e^{-s} or e^{-2s} respectively,

$$\begin{aligned}x(t) &= 2\left(\frac{1}{2}(t-1)^2-(t-1)+1-e^{-(t-1)}\right)\mathbb{1}\left(t-1\right)+\\ &\quad 2\left((t-1)-1+e^{-(t-1)}\right)\mathbb{1}\left(t-1\right)+\\ &\quad \left(1-e^{-(t-1)}\right)\mathbb{1}\left(t-1\right)-\\ &\quad 2\left(\frac{1}{2}(t-2)^2-(t-2)+1-e^{-(t-2)}\right)\mathbb{1}\left(t-2\right)-\\ &\quad 4\left((t-2)-1+e^{-(t-2)}\right)\mathbb{1}\left(t-2\right)-\\ &\quad 4\left(1-e^{-(t-2)}\right)\mathbb{1}\left(t-2\right),\end{aligned}$$

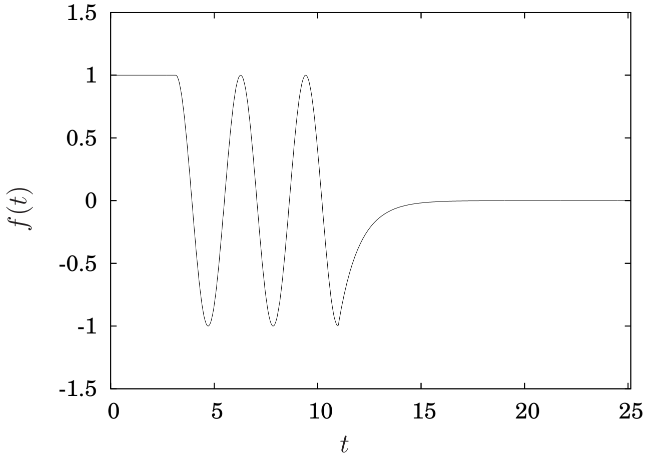


Figure 8.9. Inhomogeneous term for Equation 8.9 in Example 8.4.4.

which simplifies to

$$\begin{aligned} x(t) &= \left[(t-1)^2 + 1 - e^{-(t-1)}\right] \mathbb{1}\,(t-1) - \\ &\quad \left[(t-2)^2 + 2\,(t-2) + 2 - 2e^{-(t-2)}\right] \mathbb{1}\,(t-2)\,. \end{aligned}$$

Finally, another example involving some trigonometric functions.

Example 8.4.4 Find the solution to

$$\begin{aligned} \dot{x} + 2x &= f(t) \\ x(0) &= 1 \end{aligned} \qquad (8.9)$$

where

$$f(t) = \begin{cases} 1 & t < \pi \\ \cos 2t & \pi \leq t < \frac{7\pi}{2} \\ -e^{-\left(t-\frac{7\pi}{2}\right)} & t > \frac{7\pi}{2} \end{cases}$$

This function is illustrated in Figure 8.9.

To express $f(t)$ in a manner that is convenient to Laplace transform, we may write $f(t)$ as the sum of three functions, $f(t) = f_1(t) + f_2(t) + f_3(t)$

where

$$\begin{aligned} f_1(t) &= \begin{cases} 1 & 0 \leq t < \pi \\ 0 & \text{otherwise} \end{cases} \\ f_2(t) &= \begin{cases} \cos 2t & \pi \leq t < \frac{7\pi}{2} \\ 0 & \text{otherwise} \end{cases} \\ f_3(t) &= \begin{cases} -e^{-\left(t-\frac{7\pi}{2}\right)} & t \geq \frac{7\pi}{2} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Each of these functions may be written as a single expression using step functions as

$$\begin{aligned} f_1(t) &= \mathbb{1}\left(t\right) - \mathbb{1}\left(t-\pi\right) \\ f_2(t) &= \mathbb{1}\left(t-\pi\right)\cos 2t - \mathbb{1}\left(t-\frac{7\pi}{2}\right)\cos 2t \\ f_3(t) &= -\mathbb{1}\left(t-\frac{7\pi}{2}\right)e^{-\left(t-\frac{7\pi}{2}\right)}. \end{aligned}$$

The second function, $f_2(t)$ is not in a form that will allow us to use Theorem 8.3.18 since the argument to the step functions and the cosine function do not match. What we need is to convert $\cos 2t$ to a function of $t-\pi$ and $t-\frac{7\pi}{2}$ for each of the step functions. Observing that

$$\begin{aligned} \cos\left(2\left(t-\pi\right)\right) &= \cos 2t \\ \cos\left(2\left(t-\frac{7\pi}{2}\right)\right) &= -\cos 2t \end{aligned}$$

we then have

$$f_2(t) = \mathbb{1}\left(t-\pi\right)\cos 2\left(t-\pi\right) + \mathbb{1}\left(t-\frac{7\pi}{2}\right)\cos 2\left(t-\frac{7\pi}{2}\right).$$

So,

$$\mathcal{L}\left(f(t)\right) = \left(1-e^{-\pi s}\right)\frac{1}{s} + \left(e^{-\pi s} + e^{-\frac{7\pi}{2}s}\right)\frac{s}{s^2+4} - e^{-\frac{7\pi}{2}s}\frac{1}{s+1}$$

Laplace transforming Equation 8.9 gives

$$\begin{aligned} &\left(sX(s)-1\right) + 2X(s) = \mathcal{L}\left(f(t)\right) \\ &= \left(1-e^{-\pi s}\right)\frac{1}{s} + \left(e^{-\pi s} + e^{-\frac{7\pi}{2}s}\right)\frac{s}{s^2+4} - e^{-\frac{7\pi}{2}s}\frac{1}{s+1} \end{aligned}$$

and solving for $X(s)$ gives

$$\begin{aligned} &X(s) = \\ &\left(\left(1-e^{-\pi s}\right)\frac{1}{s} + \left(e^{-\pi s} + e^{-\frac{7\pi}{2}s}\right)\frac{s}{s^2+4} - e^{-\frac{7\pi}{2}s}\frac{1}{s+1} + 1\right)\frac{1}{s+2}. \end{aligned}$$

Considering the inverse Laplace transform term-by-term gives

1. Rearranging the first term
$$\left(1-e^{-\pi s}\right)\frac{1}{s\,(s+2)}=\frac{1}{2}\left(1-e^{-\pi s}\right)\frac{2}{s\,(s+2)}$$
so
$$\begin{aligned}&\mathcal{L}^{-1}\left(\frac{1}{2}\left(1-e^{-\pi s}\right)\frac{1}{s\,(s+2)}\right)=\\&\quad\frac{1}{2}\left[\left(1-e^{-2t}\right)\mathbb{1}\,(t)-\left(1-e^{-2(t-\pi)}\right)\mathbb{1}\,(t-\pi)\right].\end{aligned}$$

2. The product in the second term needs to be expanded as
$$\begin{aligned}\frac{s}{s^2+4}\frac{1}{s+2}&=\frac{as+b}{s^2+4}+\frac{c}{s+2}\\&=\frac{(a+c)\,s^2+(2a+b)\,s+(2b+4c)}{\left(s^2+4\right)(s+2)},\end{aligned}$$
Equating numerators gives
$$(a+c)\,s^2+(2a+b)\,s+(2b+4c)=s.$$
Since this must be true for arbitrary s, the coefficients of different powers of s must be equal, so
$$\begin{aligned}a+c&=0\\2a+b&=1\\2b+4c&=0,\end{aligned}$$
and solving for a, b and c and substituting gives
$$\begin{aligned}\frac{s}{s^2+4}\frac{1}{s+2}&=\frac{\frac{1}{4}s+\frac{1}{2}}{s^2+4}+\frac{-\frac{1}{4}}{s+2}\\&=\frac{1}{4}\left(\frac{s+2}{s^2+4}-\frac{1}{s+2}\right)\\&=\frac{1}{4}\left(\frac{s}{s^2+4}+\frac{2}{s^2+4}-\frac{1}{s+2}\right),\end{aligned}$$
where each term appears in Table 8.1. Hence
$$\begin{aligned}&\mathcal{L}^{-1}\left(\left(e^{-\pi s}+e^{-\frac{7\pi}{2}s}\right)\frac{s}{s^2+4}\frac{1}{s+2}\right)\\&\quad=\mathcal{L}^{-1}\left(\frac{1}{4}\left(e^{-\pi s}+e^{-\frac{7\pi}{2}s}\right)\left(\frac{s}{s^2+4}+\frac{2}{s^2+4}-\frac{1}{s+2}\right)\right)\\&\quad=\frac{1}{4}\Bigg[\mathbb{1}\,(t-\pi)\left(\cos 2\,(t-\pi)+\sin 2\,(t-\pi)-e^{-2(t-\pi)}\right)+\\&\qquad\mathbb{1}\left(t-\frac{7\pi}{2}\right)\left(\cos 2\left(t-\frac{7\pi}{2}\right)+\sin 2\left(t-\frac{7\pi}{2}\right)-e^{-2\left(t-\frac{7\pi}{s}\right)}\right)\Bigg].\end{aligned}$$

3. The product in the next term can be expanded as

$$\begin{aligned}\frac{1}{s+1}\frac{1}{s+2} &= \frac{a}{s+1}+\frac{b}{s+2}\\ &= \frac{a\,(s+2)+b\,(s+1)}{(s+1)\,(s+2)}\\ &= \frac{(a+b)\,s+(2a+b)}{(s+1)\,(s+2)}.\end{aligned}$$

   Equating powers of s in the numerator gives

$$\frac{1}{s+1}\frac{1}{s+2}=\frac{1}{s+1}-\frac{1}{s+2}$$

   both of which are in Table 8.1. Hence

$$\mathcal{L}^{-1}\left(-e^{-\frac{7\pi}{2}s}\frac{1}{s+1}\frac{1}{s+2}\right)=\mathbb{1}\left(t-\frac{7\pi}{s}\right)\left(e^{-\left(t-\frac{7\pi}{2}\right)}-e^{-2\left(t-\frac{7\pi}{2}\right)}\right).$$

4. Finally, the last term gives

$$\mathcal{L}^{-1}\left(\frac{1}{s+2}\right)=e^{-2t}.$$

The entire solutions is, of course, the sum of these four terms and is

$$\begin{aligned}x(t) = {} & \frac{1}{2}\left[\left(1-e^{-2t}\right)\mathbb{1}\,(t)-\left(1-e^{-2(t-\pi)}\right)\mathbb{1}\,(t-\pi)\right]+\\ & \frac{1}{4}\left[\mathbb{1}\,(t-\pi)\left(\cos 2\,(t-\pi)+\sin 2\,(t-\pi)-e^{-2(t-\pi)}\right)+\right.\\ & \left.\mathbb{1}\left(t-\frac{7\pi}{2}\right)\left(\cos 2\left(t-\frac{7\pi}{2}\right)+\sin 2\left(t-\frac{7\pi}{2}\right)-e^{-2\left(t-\frac{7\pi}{s}\right)}\right)\right]\\ & -\mathbb{1}\left(t-\frac{7\pi}{s}\right)\left(e^{-\left(t-\frac{7\pi}{2}\right)}-e^{-2\left(t-\frac{7\pi}{2}\right)}\right)+e^{-2t}.\end{aligned}$$

We can check the solution by evaluating it in each of the regions in which $f(t)$ has a different form. In particular

1. For $0 \le t < \pi$,

$$x(t)=\frac{1}{2}\left(1-e^{-2t}\right)+e^{-2t} \tag{8.10}$$

   and

$$\dot{x}(t)=-e^{-2t}. \tag{8.11}$$

   Hence, substituting into Equation 8.9 gives

$$\dot{x}+2x=-e^{-2t}+2\left(\frac{1}{2}\left(1-e^{-2t}\right)+e^{-2t}\right)=1.$$

   Also checking the initial condition gives

$$x(0)=\frac{1}{2}\left(1-e^{0}\right)+e^{0}=1.$$

2. For $\pi \leq t < \frac{7\pi}{2}$, $x(t)$ is the same as in equation 8.10 with the addition of the terms multiplied by $\mathbb{1}\,(t-\pi)$,

$$\begin{aligned} x(t) &= \frac{1}{2}\left(1-e^{-2t}\right)+e^{-2t}-\frac{1}{2}\left(1-e^{-2(t-\pi)}\right)+ \\ &\quad \frac{1}{4}\left(\cos 2\,(t-\pi)+\sin 2\,(t-\pi)-e^{-2(t-\pi)}\right) \qquad (8.12) \end{aligned}$$

and

$$\begin{aligned} \dot{x}(t) = &-\frac{1}{2}e^{-2t}-e^{-2(t-\pi)} \\ &+\frac{1}{2}\left(-\sin\left(2\,(t-\pi)\right)+\cos\left(2\,(t-\pi)\right)-e^{-2(t-\pi)}\right). \end{aligned}$$

Substituting into Equation 8.9 gives

$$\dot{x}+2x=\cos\left(2\,(t-\pi)\right)=\cos 2t.$$

Also, the solution the solutions in Equations 8.10 and 8.12 must match at $t=\pi$. Substituting $t=\pi$ into Equation 8.10 gives

$$x(\pi)=\frac{1}{2}\left(1-e^{-2\pi}\right)+e^{-2\pi}=\frac{1}{2}\left(1+e^{-2\pi}\right).$$

Substituting $t=\pi$ into Equation 8.12 gives

$$\begin{aligned} x(\pi) &= \frac{1}{2}\left(1-e^{-2\pi}\right)+e^{-2\pi}-\frac{1}{2}\left(1+e^{-2(\pi-\pi)}\right)+ \\ &\quad \frac{1}{4}\left(\cos 2\,(\pi-\pi)+\sin 2\,(\pi-\pi)-e^{-2(\pi-\pi)}\right) \\ &= \frac{1}{2}\left(1-e^{-2\pi}\right)+\frac{1}{4}\,(1+0-1) \\ &= \frac{1}{2}\left(1-e^{-2\pi}\right), \end{aligned}$$

so the two solutions match at $t=\pi$.

3. Verifying for $t \geq \frac{7\pi}{2}$ is left as an exercise. ■

8.5 Transfer Functions

The notion of a transfer function is particularly useful in engineering since it is a concise representation of the relationship between the input and output of a system with all the intermediate variables implicitly represented. In order to determine transfer functions in engineering a student must have basic abilities to model engineering components. If that is not something that comes naturally, perhaps a review of the material from Section 1.9 would be useful before proceeding. A simple example will help illustrate the concept of a transfer function.

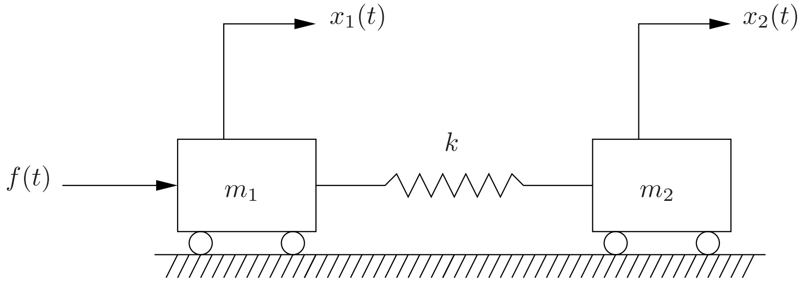


Figure 8.10. System to control for example 8.5.1.

Example 8.5.1 Consider the task of controlling the system illustrated in Figure 8.10. What is desired is to control the position of mass 2 with the input force, $f(t)$. Exactly how to control it will be addressed subsequently. Now we consider the task of determining a convenient way to model it.

The equations of motion are simple to determine:

$$\begin{aligned} m_1\ddot{x}_1 &= k\left(x_2 - x_1\right) + f(t) && (8.13)\\ m_2\ddot{x}_2 &= k\left(x_1 - x_2\right). && (8.14)\end{aligned}$$

Clearly these are coupled and, in the present form it will be impossible to determine $x_2(t)$ without simultaneously solving for $x_1(t)$. The same is true if we represent it as a system of first order equations by setting

$$\begin{aligned} \xi_1 &= x_1\\ \xi_2 &= \dot{x}_1\\ \xi_3 &= x_2\\ \xi_4 &= \dot{x}_2 \end{aligned}$$

which gives

$$\frac{d}{dt}\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\\ \xi_4\end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0\\ -\frac{k}{m_1} & 0 & \frac{k}{m_1} & 0\\ 0 & 0 & 0 & 1\\ \frac{k}{m_2} & 0 & -\frac{k}{m_2} & 0\end{bmatrix} + \begin{bmatrix}0\\ \frac{f(t)}{m_1}\\ 0\\ 0\end{bmatrix}.$$

Solving these equations is no problem, however it would be especially convenient if we could have a more concise representation of the relationship between the input, $f(t)$ and the output, $x_2(t)$. Recalling that a main feature of Laplace transforms is that, once transformed, solving the differential equations is reduced to algebra, if we Laplace transform the equations of motion, it may be possible to algebraically eliminate the intermediate variable(s).

Assuming that the initial conditions are all zero, *i.e.,*

$$\begin{aligned} x_1(0) &= 0 \\ \dot{x}_1(0) &= 0 \\ x_2(0) &= 0 \\ \dot{x}_2(0) &= 0 \end{aligned}$$

taking the Laplace transform of equations 8.13 and 8.14 gives

$$\begin{aligned} m_1 s^2 X_1(s) &= k\left(X_2(s) - X_1(s)\right) + F(s) \qquad (8.15) \\ m_2 s^2 X_2(s) &= k\left(X_1(s) - X_2(s)\right). \qquad (8.16) \end{aligned}$$

These are two equations that are linear in three functions, $X_1(s)$, $X_2(s)$ and $F(s)$. Hence, we may use one of the equations to eliminate one of the functions. Since we are interested in the relationship between the input force, $f(t)$ and the position of mass 2, $x_2(t)$, it makes sense to solve one equation for $X_1(s)$ and substitute into the other equation. Solving Equation 8.16 for $X_1(s)$ gives

$$X_1(s) = \frac{m_2 s^2 + k}{k} X_2(s).$$

Substituting this into the Equation 8.15 and rearranging gives

$$X_2(s) = \frac{k}{s^2\left(m_1 m_2 s^2 + k\left(m_1 + m_2\right)\right)} F(s). \qquad (8.17)$$

Since it directly relates the effect of the input force on the position of the output mass, we will call the function

$$\frac{X_2(s)}{F(s)} = \frac{k}{s^2\left(m_1 m_2 s^2 + k\left(m_1 + m_2\right)\right)}$$

the *transfer function* from the input $F(s)$ to the output $X_2(s)$.

Observe the following about equation 8.17.

1. This is a concise relationship between the input force and the position of mass two. In fact, the variable representing the position of mass one does not explicitly appear in the equation at all.

2. Given an input force, $f(t)$, we could compute its Laplace transform, $F(s) = \mathcal{L}\left(f(t)\right)$, substitute $F(s)$ into equation 8.17 and, in principle, compute the inverse Laplace transform of $X_2(s)$ to find the motion of $x_2(t)$.

3. While mass one does not explicitly appear in the equation, it is implicitly in the equation in the terms in the denominator of the transfer function. In fact, it should be obvious that it cannot be eliminated. After all, the only way mass two moves is by the force accelerating mass one, and mass one's motion affecting mass two through the spring. ■

In light of the usefulness of the formulation of the relationship between the force and position of the mass represented by Equation 8.17, we may define a transfer function in the following manner.

Definition 8.5.2 A *transfer function* is the ratio of the Laplace transform of the output to the input of some system assuming all the initial conditions are zero. ⋄

What exactly is the *input* and *output* of a system depends on the problem and either must be stated or should be clear from the context of the problem. Subsequently it will be apparent that the output of one system may be the input to another. For example, the output of a motor which may be the torque or position of the motor shaft, is the input to whatever it is driving.

As will be clear subsequently, the denominator of the transfer function is of particular importance.

Definition 8.5.3 Let

$$G(s)=\frac{N(s)}{D(s)}$$

be a transfer function. The equation

$$D(s)=0,$$

i.e., setting the denominator equal to zero is called the *characteristic equation*.⋄

A property regarding a transfer function that will be assumed throughout the rest of this text is that the order of the polynomial in the denominator is greater than the order of the polynomial in the numerator. Such a transfer function is called *proper*.

Now, we will make the problem more complicated by replacing the general forcing function in Example 8.5.1, $f(t)$, with something more realistic.

> **Example 8.5.4** Consider the same system as in Example 8.5.1 but where the force is generated by a belt attached to a pulley attached to a d.c. motor which is driven by an electric circuit, as illustrated in Figure 8.11 and 8.12. In Figure 8.11, the first mass is attached to a belt that driven by a pulley. The pulley on the left is attached to a d.c. motor that is driven by the circuit illustrated in Figure 8.12. The pulley on the right is identical to the pulley on the left except it is not driven and is free to rotate. Each pulley has a radius r and moment of inertia of J about its center. Assume that belt is light so that its mass may be ignored and that it does not slip on the pulleys. The motor circuit is comprised of an ideal current source, a resistor and a d.c. motor attached to the output. The d.c. motor has a torque constant of k_τ and a back e.m.f. constant of k_e. We wish to determine the transfer function from the input current to the circuit to the position of mass 2.
>
> The Laplace transform of the differential equations for the two masses are given in Example 8.5.1 in Equations 8.15 and 8.16. So what is left is to model the belt and pulley system as well as the circuit. Free body diagrams

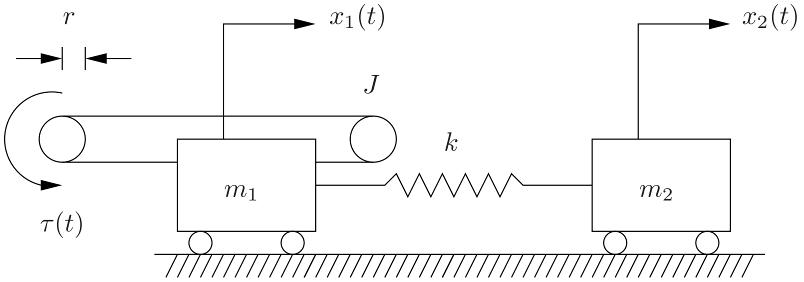


Figure 8.11. System to control for example 8.5.4.

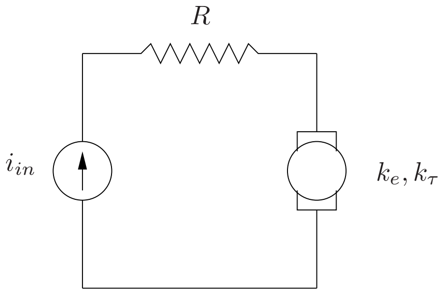


Figure 8.12. Motor driving circuit for Example 8.5.4.

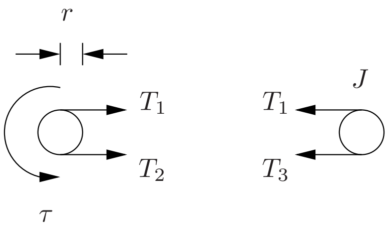


Figure 8.13. Free body diagrams of the pulleys from Example 8.5.4.

of the two pulleys are illustrated in Figure 8.13. Since the bottom portion of the belt is attached to the mass, if the mass is accelerating the tension on each side of the mass must be different. Since there is no mechanical component between the pulleys on the top, the tension in the top belt is constant. Denote the tension in the top portion of the belt by $T_1(t)$, and $T_2(t)$ and $T_3(t)$ on the bottom of the belt to the left and right of the mass respectively.

If we denote the angular position of both pulleys by θ, since the belt does not slip, θ is related to the position of the mass by $r\theta = x_1$. Newton's law on the right pulley gives

$$J\frac{\ddot{x}_1}{r} = r\left(T_1 - T_3\right)$$

and on the left pulley gives

$$J\frac{\ddot{x}_1}{r} = r\left(T_2 - T_1\right) + \tau.$$

The force on mass 2 is

$$f = T_2 - T_3.$$

Taking the Laplace transform of these three equations with zero initial conditions gives

$$\begin{aligned} Js^2X_1(s) &= r^2\left(T_1(s) - T_3(s)\right) \\ Js^2X_1(s) &= r^2\left(T_2(s) - T_1(s)\right) + rT(s) \\ F(s) &= T_2(s) - T_3(s). \end{aligned}$$

Adding the first two equations gives

$$2Js^2X_1(s) = r^2\left(T_2(s) - T_3(s)\right) + rT(s)$$

and using the last equation

$$2Js^2X_1(s) = r^2F(s) + rT(s) \tag{8.18}$$

Since the circuit has a current source, the torque produced by the motor is

$$\tau = k_\tau i \qquad \Longleftrightarrow \qquad T(s) = k_\tau I(s).$$

Substituting into Equation 8.18 gives

$$2Js^2 X_1(s) = r^2 F(s) + rk_\tau I(s),$$

and eliminating $X_1(s)$ and $F(s)$ from this equation and Equations 8.15 and 8.16 gives

$$\frac{X_2(s)}{I_{in}(s)} = \frac{k_\tau kr}{s^2\left[2J\left(m_2 s^2 + k\right) - r^2\left(k\left(m_1 + m_2\right) + m_1 m_2 s^2\right)\right]}. \qquad \blacksquare$$

Let us consider one more example which probably qualifies as rocket science.

Example 8.5.5 Consider the rocket illustrated in Figure 8.14. The velocity of the center of mass (com) of the rocket is at an angle θ_r with respect to the axis of symmetry of the rocket body. The point through which all aerodynamic forces may be resolved is call the *center of pressure* (cop). The component of the aerodynamic force along the axis of symmetry of the rocket body is called the *drag* and the component orthogonal to the drag is called the *lift*. The lift force will be denoted by f_l. The mass moment of inertia of the rocket about its center of mass will be denoted by J_r. Assume the distance between the center of mass and center of pressure is l_1 and the distance between the center of mass and the location of the rocket nozzle is l_2.

The rocket is controlled by *thrust vectoring*, which means that the nozzle of the rocket engine is gimballed and can pivot. The thrust of the rocket engine is denoted by f_t and the angle of the nozzle with respect to the center-line of the rocket body is denoted by θ_n.

In this problem we will be concerned with the *angle of attack* of the rocket, *i.e.,* the angle between the direction it is pointing and its velocity. This is rocket is *unstable* since the center of pressure is above the center of mass. This would typically be considered a poor design; however, if we want the rocket to be highly maneuverable, then perhaps it is a good feature. The problem is to find the transfer function from the nozzle angle to the pitch angle of the rocket. To simplify the analysis, we will assume that the velocity of the rocket is constant.

Basic aerodynamics provides a formula for the lift force, which is

$$f_l = C_l \frac{\rho \|v\|^2 A}{2}$$

where C_l is the coefficient of lift, ρ is the density of the air, $\|v\|$ is the magnitude of the velocity of the rocket and A is reference area, which is a function of the lateral area of the rocket exposed to the sideways flow due

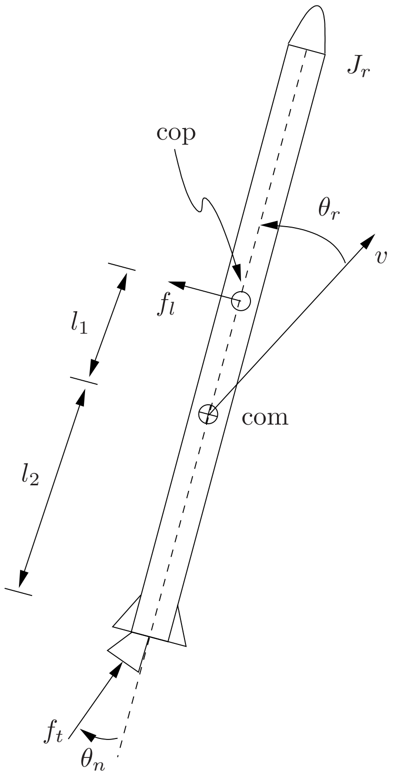


Figure 8.14. Rocket for Example 8.5.5.

to a non-zero angle of attack.. We will assume that A is proportional to the angle of attack, θ_r, so

$$A = A_{ref} \sin \theta_r$$

and then

$$f_l = \frac{1}{2} \rho C_l \|v\|^2 A_{ref} \sin \theta_r$$

and for $\theta_r \ll 1$, then

$$f_l \approx \frac{1}{2} \rho C_l \|v\|^2 A_{ref} \theta_r.$$

Assuming $\theta_r \ll 1$ is reasonable; otherwise the rocket would essentially be flying "sideways."

Newton's law for the rotation of the rocket body gives

$$J_r \ddot{\theta}_r = f_t l_2 \sin \theta_n + \frac{1}{2} \rho C_l \|v\|^2 A_{ref} l_1 \theta_r.$$

For small θ_n,

$$J_r \ddot{\theta}_r = f_t l_2 \theta_n + \frac{1}{2} \rho C_l \|v\|^2 A_{ref} l_1 \theta_r.$$

Taking the Laplace transform and assuming zero initial conditions gives

$$J_r s^2 \Theta_r(s) = f_t l_2 \Theta_n(s) + \frac{1}{2} \rho C_l \|v\|^2 A_{ref} l_1 \Theta_r(s) \tag{8.19}$$

and solving for the transfer function gives

$$\frac{\Theta_r(s)}{\Theta_n(s)} = \frac{f_t l_2}{J_r s^2 - \frac{1}{2} \rho C_l \|v\|^2 A_{ref} l_1}.$$

■

Let us extend the example now to include some sort of actuation. We will assume that the thrust vectoring is achieved by attaching a d.c. motor to the axis of rotation of the rocket engine nozzle. For very large rocket engines, such as the main engines launch vehicles, the actuation for the thrust vectoring is achieved by hydraulic systems. For smaller systems, such as the maneuvering thrusters for the space shuttle orbiter, the actuation is achieved by d.c. servor motors.[3]

Example 8.5.6 Figure 8.15 is a schematic of the nozzle actuation system. The nozzle has moment of inertia J_n about its pivot point and there are two springs with spring constant $\frac{k_n}{2}$ attached to the nozzle a length l_3 from the pivot point. A dc motor with torque constant K_τ and back emf constant k_e attached to the pivot point that rotates the nozzle and provides a torque τ. The circuit driving the motor is illustrated in Figure 8.16. Find the transfer function from the input voltage to the circuit to the angle of the nozzle, and then find the transfer function from the input voltage to the angle of attack of the rocket. Assume that the overall rotation of the nozzle is small.

[3] A servo motor is a unit where the angle of the motor is controlled. A signal to the servo motor, typically a *pulse width modulated* signal indicates what the angle of the shaft of the motor should be, and internal feedback control circuitry controls the output angle of the shaft so that is accomplished. The means to to this will be covered when we consider feedback in the following sections.

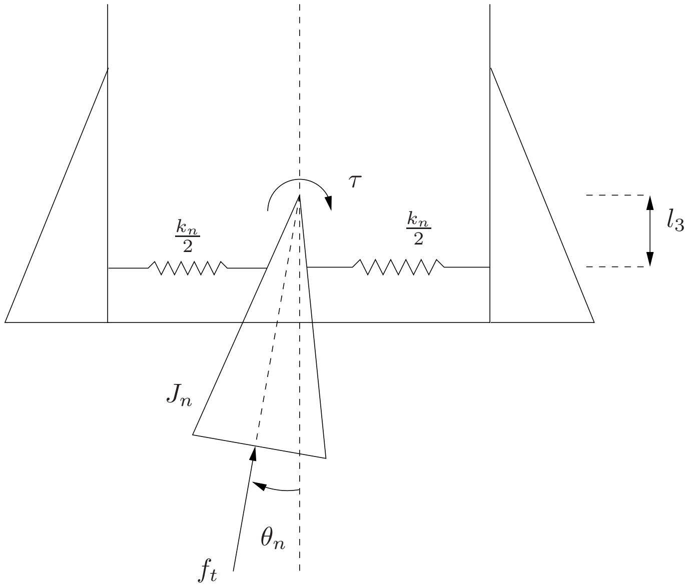


Figure 8.15. Close-up of vectored thrust rocket nozzle for Example 8.5.6.

If $\theta_n \ll 1$, then the restoring torque about the pivot point due to the displacement of the springs will be approximately

$$\tau_s = l_3^2 k \sin\theta_n \approx l_3^2 k_n \theta_n.$$

The only torques about the pivot point are τ_s from the springs and τ from the dc motor. Hence, Newton's law about the pivot point is

$$\begin{aligned} J_n\ddot{\theta}_n &= \tau - \tau_s \\ &= \tau - l_3^2 k_n \theta_n. \end{aligned}$$

Computing the Laplace transform with zero initial conditions gives

$$s^2 J_n \Theta_n(s) = T(s) - l_3^2 k_n \Theta_n(s)$$

where $T(s) = \mathcal{L}\left(\tau(t)\right)$ so the transfer function from the motor torque to the nozzle angle is

$$\frac{\Theta_n(s)}{T(s)} = \frac{1}{J_n s^2 + l_3^2 k_n}. \tag{8.20}$$

Returning to Equation 8.19, the torque required to pivot the nozzle has an equal and opposite effect on the rocket body. In particular, Equation 8.19 is now

$$J_r s^2 \Theta_r(s) = f_t l_2 \Theta_n(s) + \frac{C_l \|v\|^2 A_{ref}}{2} l_1 \Theta_r(s) + T(s). \tag{8.21}$$

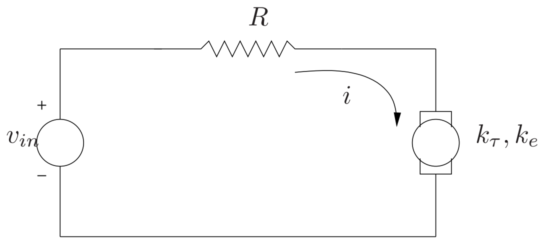


Figure 8.16. Actuator circuit for vector thrust nozzle in Example 8.5.6.

Now considering the circuit, Kirchhoff's voltage law around the circuit gives

$$v_{in} = iR + k_e \dot{\theta}_n$$

or

$$V_{in}(s) = I(s)R + sk_e\Theta_n(s) \tag{8.22}$$

and the torque property of the motor gives

$$\tau = ik_\tau$$

or

$$T(s) = k_\tau I(s). \tag{8.23}$$

So we have four equations, 8.20, 8.21, 8.22 and 8.23 and five variables, $\Theta_r(s)$, $\Theta_n(s)$, $T(s)$, $V_{in}(s)$ and $I(s)$. A few lines of algebra gives

$$\frac{\Theta_r(s)}{V_{in}(s)} = \frac{k_\tau \left(J_n s^2 + \left(k_n l_3^2 + f_t l_2\right)\right)}{\left(J_r s^2 - \frac{C_l \|v\|^2 A_{ref}}{2} l_1\right)\left(J_n R s^2 + k_e k_\tau s + k_n l_3^2 R\right)}.$$

This expression is rather complicated, but it is not surprising: the effect of a voltage through a circuit with a motor attached to a nozzle that directs the angle of attack of a rocket will not necessarily be very simple. ■

Example 8.5.7 Determine the transfer function from the input voltage, v_{in} to the output voltage, v_{out} for the circuit illustrated in Figure 8.17.

Let v_{L_1}, v_{C_2}, *etc.*, denote the voltage drops across inductor L_1, capacitor C_2, *etc.* and assume that positive voltage drops are in the direction of the arrows for the currents. Kirchhoff's voltage law gives ~~three equations of the form~~

~~There is something wrong with this example. I think the math is right, but the circuit seems weird. With an ideal voltage source, neither C_1 nor L_1 appear in the transfer function, which makes sense—but why have them in the circuit then?~~

$$\begin{aligned}
v_{in} &= v_{C_1}\\
v_{in} &= v_{L_1}\\
v_{in} &= v_{C_3} + v_{L_3} + v_{L_2}\\
v_{out} &= v_{C_2}\\
v_{out} &= v_{L_2}.
\end{aligned}$$

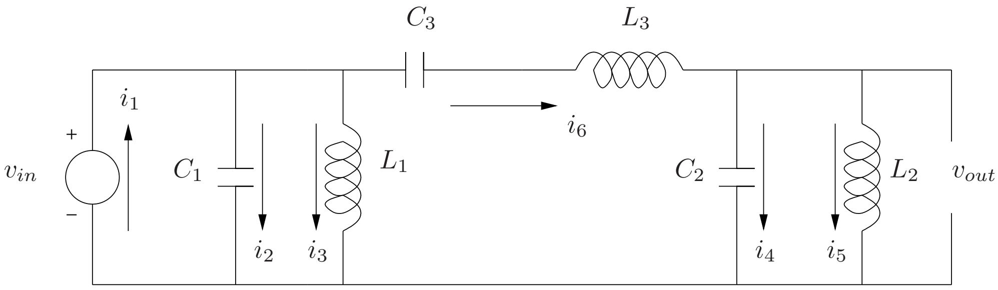


Figure 8.17. Circuit for Example 8.5.7.

Kirchhoff's current law gives two equations at the top and bottom nodes in the center of the circuit of the form

$$\begin{aligned} i_1 &= i_2 + i_3 + i_6 \\ i_6 &= i_4 + i_5. \end{aligned}$$

Note that even though the top part of vertical inductor and capacitor pairs do not meet at a point, since there is no component between them they meet at a node.

The inductor and capacitors are described by

$$\begin{aligned} v_{L_j} &= L_j \frac{di_{L_j}}{dt} \\ i_{C_j} &= C_j \frac{dv_{C_j}}{dt}. \end{aligned}$$

respectively. Laplace transforming and solving for the voltage across a capacitor gives

$$V_{C_j}(s) = \frac{I_{C_j}(s)}{sC_j}.$$

Laplace transforming the voltage equations and substituting for the component laws gives

$$\begin{aligned} V_{in}(s) &= \frac{I_2(s)}{sC_1} \\ V_{in}(s) &= sL_1 I_3(s) \\ V_{in}(s) &= \frac{I_6(s)}{sC_3} + sL_3 I_6(s) + sL_2 I_5(s) \\ V_{out}(s) &= \frac{I_4(s)}{sC_2} \\ V_{out}(s) &= sL_2 I_5(s). \end{aligned}$$

These five voltage equations along with the two current equations gives seven equations. The variables are the input and output voltages and the six currents. So, we have the right number of equations and variables to eliminate the six currents to find the transfer function from the input voltage to the output voltage. Doing so gives

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{C_3 L_2 s^2}{C_2 C_3 L_2 L_3 s^4 + (C_2 L_2 + C_3 L_2 + C_3 L_3) s^2 + 1}. \qquad ■$$

May systems have more than one input and more than one output. Even for control systems where we want to control a single variable with one input, there will often be external disturbances. The following example illustrates this fact.

Example 8.5.8 Consider the mechanical system illustrated in Figure 8.18, which is the same as the system in Example 8.5.1 except now an external disturbance force, $d(t)$ is acting on the second mass. The equations of motion for each mass are

$$\begin{aligned} m_1 \ddot{x}_1(t) &= k\left(x_2(t) - x_1(t)\right) + f(t) \\ m_2 \ddot{x}_2(t) &= k\left(x_1(t) - x_2(t)\right) - d(t), \end{aligned}$$

so

$$\begin{aligned} \left(m_1 s^2 + k\right) X_1(s) &= k X_2(s) + F(s) \\ \left(m_2 s^2 + k\right) X_2(s) &= k X_1(s) + D(s). \end{aligned}$$

Eliminating $X_1(s)$ gives

$$X_2(s) = \frac{k}{(m_1 s^2 + k)(m_2 s^2 + k) - k^2} F(s) + \frac{m_1 s^2 + k}{(m_1 s^2 + k)(m_2 s^2 + k) - k^2} D(s).$$

The term multiplying $F(s)$ is the transfer function from $F(s)$ to $X_2(s)$ and the term multiplying $D(s)$ is the transfer function from $D(s)$ to $X_2(s)$. Clearly, both from the equation as well as intuition, the response $x_2(t)$ will be a linear combination of the two terms. A lot of the purpose of controls is to specify $f(t)$ as a function of either or both $x_1(t)$ and $x_2(t)$ so that $x_2(t)$ maintains a desired value, regardless of the disturbance, $d(t)$. ■

8.6 Block Diagram Representation and Algebra

Block diagrams are a graphical means to represent transfer functions and feedback control systems. They are particularly convenient because they represent

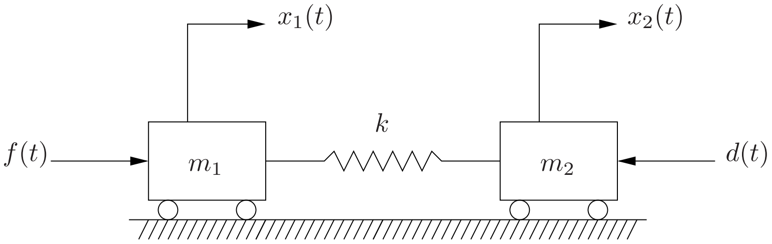


Figure 8.18. System to control for example 8.5.8.

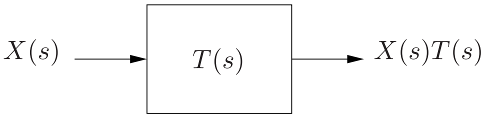


Figure 8.19. A block with an input and output arrow.

feedback in a visually intuitive manner, the various components are often isolated and the overall representation is simpler. The salient point to keep in mind is that they are simply an alternative representation, and that this alternative representation is as rigorous as the algebraic representation.

Block diagrams are comprised of four types of components.

1. A *block* represents a transfer function describing the relationship between some input and output. It is usually graphically represented by rectangle. The output is equal to the input times the transfer function inside the block.

2. *Arrows* represent signals, which are the Laplace transform of some time domain function. Arrows directed into blocks represent input signals and arrows directed out of blocks represent output signals from that transfer function. A block with an input and output arrow is illustrated in Figure 8.19.

3. *Comparators* add or subtract multiple signals, as is illustrated in Figure 8.20. The sign associated with any signal is indicated near the corresponding arrow where it enters the comparator.

4. *Branch points* distribute a signal concurrently to multiple arrows, as is illustrated in Figure 8.21. They do not “split” or “divide” the signal.

Since the elements of a block diagram are defined with mathematical precision it is important to keep in mind that they are an *exact* representation of a

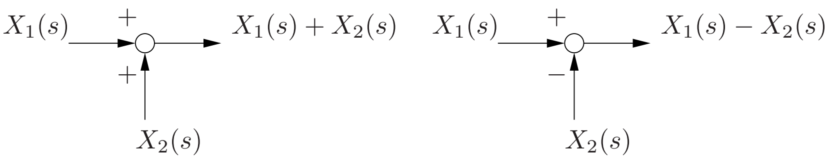


Figure 8.20. A block diagram comparator.

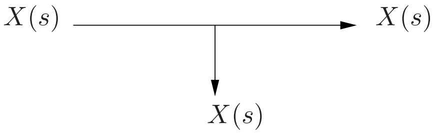


Figure 8.21. A block diagram branch point.

system. In other words, there is a one-to-one correspondence between a block diagram representation and an equation that represents the differential equation governing the system. This notion will be highlighted by a series of examples, including some that refer back to the robot arm PID control examples from Section 9.2.

Example 8.6.1 Consider the block diagram in Figure 8.22. Using the rules for the block diagram representation of transfer functions we will verify that this is the same representation as determined in Example 9.2.3.

The signal coming out of the first comparator is the error, $E(s) = \Theta_d(s) - \Theta(t)$. Then it is multiplied by the proportional gain to give the torque, $T(s) = k_p\,(\Theta_d(s) - \Theta(s))$. Figure 8.23 illustrates the same block diagram with these two signals labeled. Then it it added to the gravity term

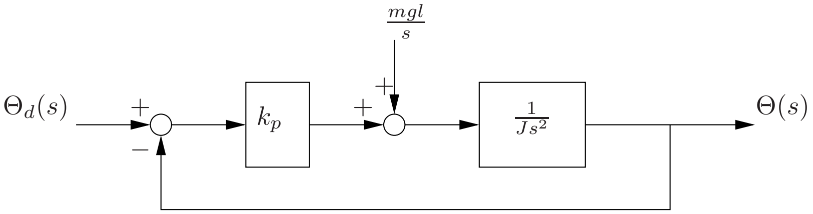


Figure 8.22. Block diagram for proportional control of a robot arm in Example 8.6.1.

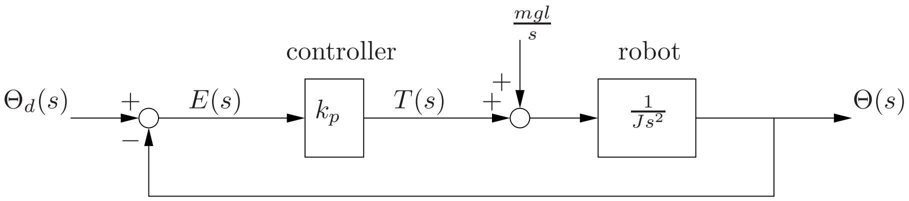


Figure 8.23. Block diagram for proportional control of a robot arm in Example 8.6.1.

and finally multiplied by the robot dynamics to give $\Theta(s)$. Mathematically,

$$\Theta(s) = \left[k_p\left(\Theta_d(s) - \Theta(s)\right) - \frac{mgl}{s}\right]\frac{1}{Js^2}.$$

Solving for the arm angle gives

$$\Theta(s) = \frac{k_p \Theta_d s - mgl}{s\left(Js^2 + k_p\right)},$$

is the same as Equation 9.5. ■

Note that other than simply providing an alternative representation Equation 9.5, the block diagram in Figure 8.22 also represents the interconnected nature of the system more explicitly and more naturally. The algebraic representation of a transfer function does not necessarily provide an indication of the relationship between components; whereas, the block diagram provides this information explicitly.

Since components of a block diagram have explicit algebraic meaning, just as it is possible to algebraically manipulate an equation, it is possible to algebraically manipulate a block diagram. All of these are relatively straight-forward and a few examples should help elucidate the concept.

Example 8.6.2 A branch point carries a signal concurrently along multiple arrows. If a signal is multiplied by a transfer function in a block before a branch point, the arrows out of the branch point are both multiplied by the transfer function inside the block. In order to move a branch point from the output side of a block to the input side, both arrows must then have the transfer function inside a block so that they carry the same signal. This is represented in Figure 8.24.

Similarly, the manner in which to move a branch point from the input side of a block to the output side of the block is illustrated in Figure 8.25.

Example 8.6.3 The previous example illustrated how to move a branch point to another side of a block. Mathematically it represents the algebraic property of distribution. The algebraic property that multiplication

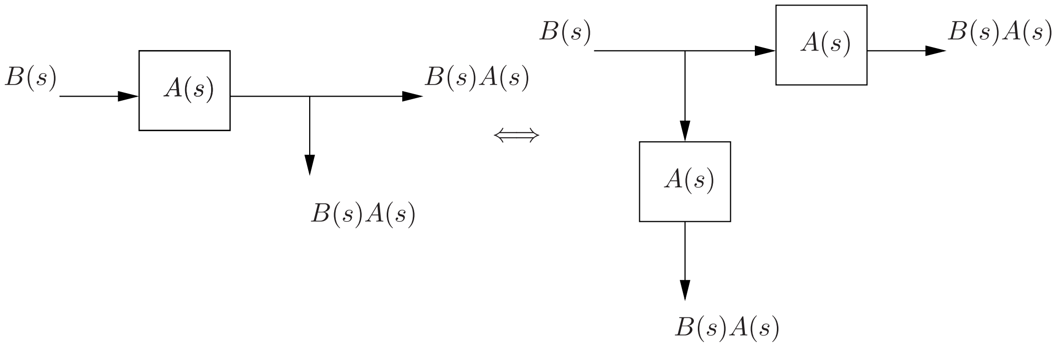


Figure 8.24. Moving a branch point to the input side of a block.

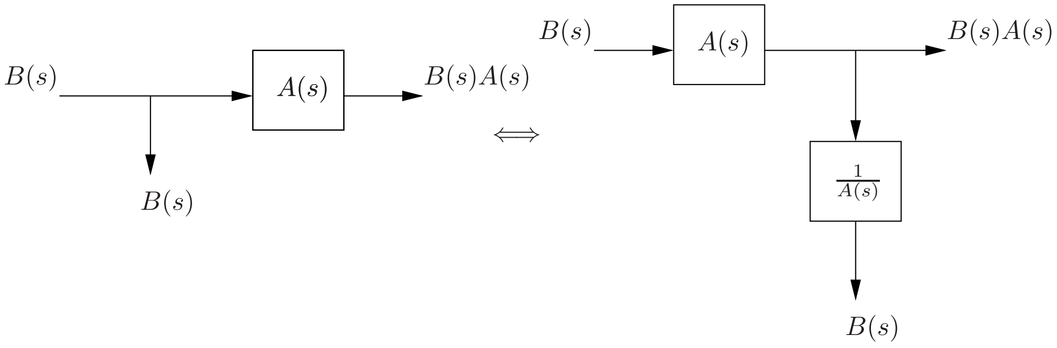


Figure 8.25. Moving a branch point to the output side of a block. ■

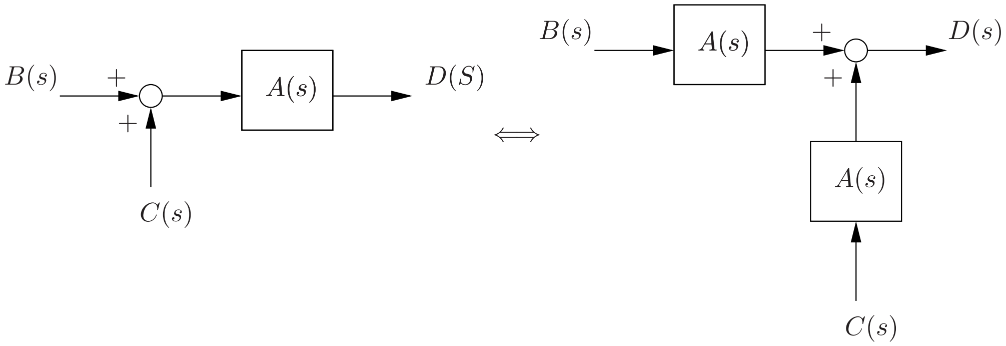


Figure 8.26. Equivalent block diagrams representing the fact that multiplication distributes over addition.

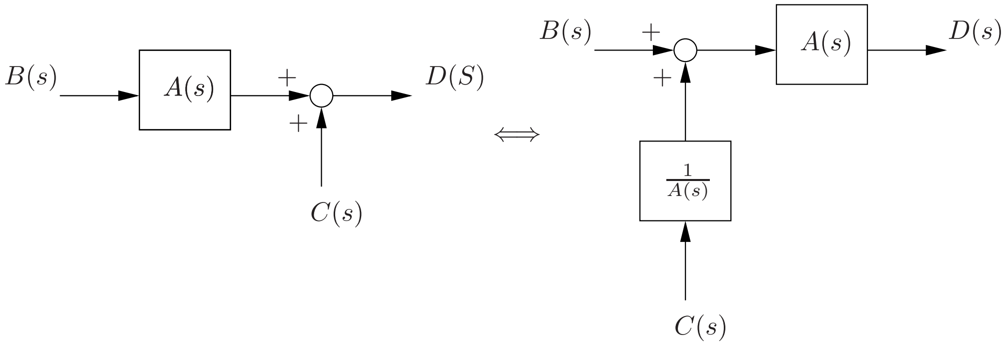


Figure 8.27. Equivalent block diagrams.

distributes over is represented by the equality

$$D(s) = (B(s) + C(s))\, A(s) = B(s)A(s) + C(s)A(s).$$

In a block diagram, it is represented by the fact that the two block diagrams in Figure 8.26 are equivalent.

Similarly, the relationship

$$D(s) = B(s)A(s) + C(s) = \left(B(s) + \frac{C(s)}{A(s)} \right) A(s)$$

is represented in Figure 8.27.

So, we now have a rule to move a comparator to either side of a block. If a comparator is moved to the output side of a block, each arrow entering the comparator must multiply the block. If a comparator is moved to the

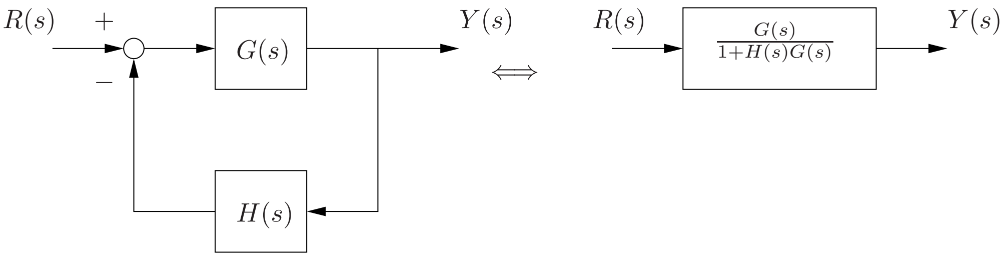


Figure 8.28. Feedback transfer function.

input side of the block, the arrow that originally did not multiply the block must have a block that inverts the multiplication of the block. ■

The next example illustrates what is perhaps the most important block diagram manipulation that we will commonly utilize.

Example 8.6.4 Consider the feedback system illustrated on the left in Figure 8.28. We will show that it is equivalent to the block diagram on the right.

To show these are equivalent, write

$$Y(s) = (R(s) - H(s)Y(s))\,G(s)$$

and solve for $Y(s)$, which gives

$$Y(s) = \frac{G(s)}{1 + H(s)G(s)}R(s).$$

■

As the next example shows, the order of branch points may be switched as long as there is no component between them. However, in general switching the order of a comparator and branch point will require some care.

Example 8.6.5 The two block diagrams in Figure 8.29 are equivalent.

Switching a comparator and branch point in a similar manner results in a block diagram that is generally *not* equivalent, as is illustrated in Figure 8.30.

These and a few other manipulations are summarized in Table 8.3.

The canonical form for a feedback block diagram is the form on the right in Figure 8.28, where there is one *feedforward* block leading from the input to output and one *feedback* block. This form is convenient because it is natural minimal representation for a feedback system, and many analysis and design methods in controls start with this canonical form. In particular, the root locus design method in Section 9.9 and the frequency response methods from Section 9.11 both start with this canonical form.

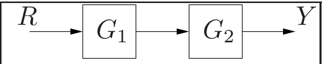
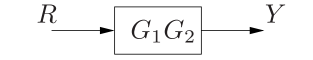
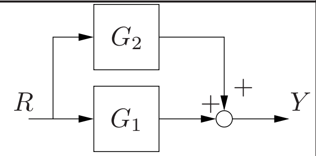
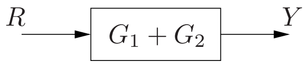
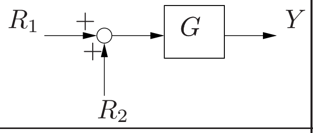
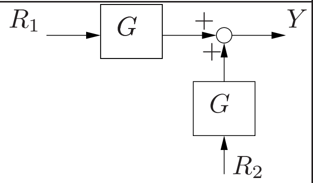
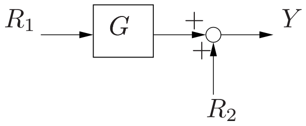
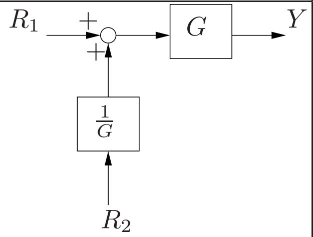
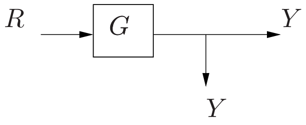
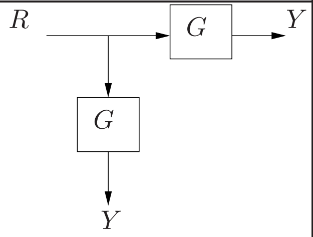
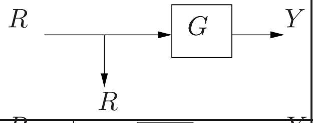
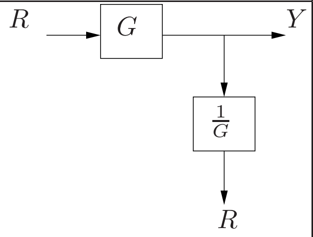
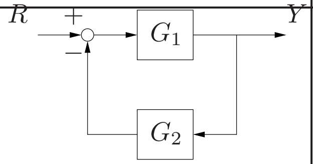
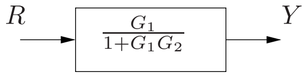
cascade elements		
parallel elements		
moving comparator		
moving comparator		
moving branch point		
moving branch point		
eliminating feedback loop		

Table 8.3. Summary of block diagram algebraic manipulations.

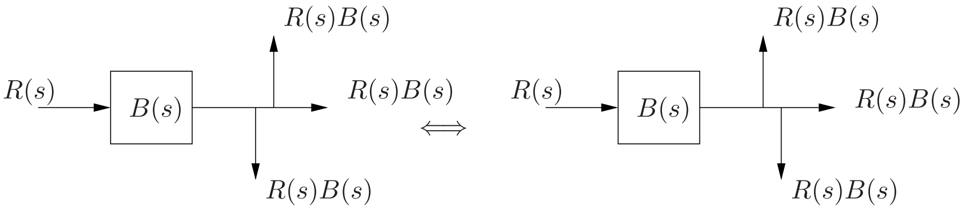


Figure 8.29. Switching the order of branch points in a block diagram.

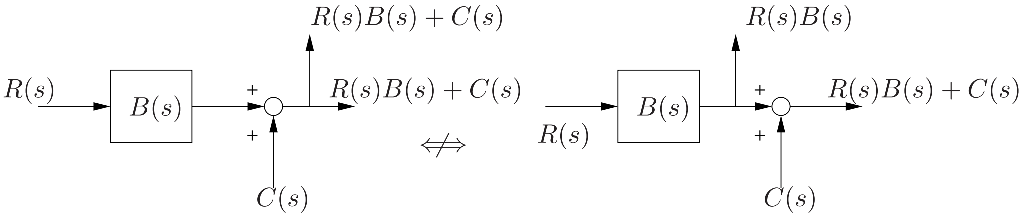


Figure 8.30. Switching the order of a branch point and comparator in a block diagram. ■

Just as a sequence of algebraic steps may be used to simplify an complicated algebraic expression, a sequence of corresponding manipulations in a block diagram may be used to determine an alternative block diagram. According to [14], a good recipe for simplifying block diagrams is the following.

1. Combine cascade blocks.
2. Combine parallel blocks.
3. Eliminate interior feedback loops.
4. Shift comparators to the left.
5. Shift branch points to the right.
6. Iterate until canonical form is obtained.

The following example illustrates block diagram manipulations for a reasonably complicated block diagram.

Example 8.6.6 Consider the block diagram illustrated in Figure 8.31. Determine the transfer function from the input to the output.

In Figure 8.32, the block diagram has been modified by switching moving the branch point that was between the comparator and block containing $C(s)$ to the output side of $C(s)$. The the block containing the transfer

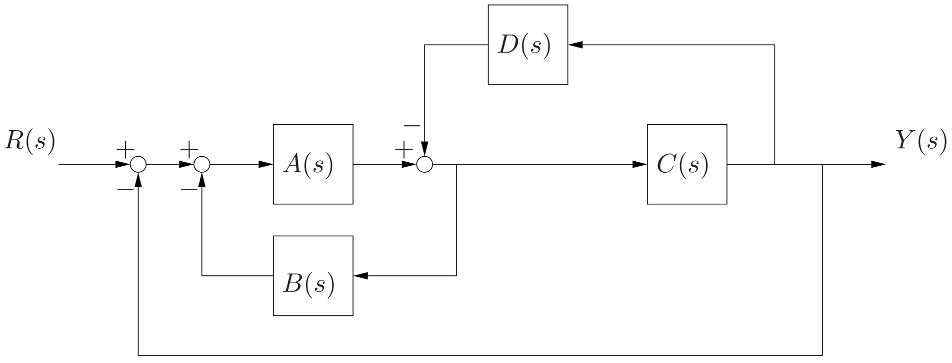


Figure 8.31. Block diagram for Example 8.6.6.

function $B(s)$ was modified by dividing by $C(s)$. Also, since the order of adjacent branch points does not matter, the branch point was moved to be the middle of the three on the right side of Figure 8.32. Now the result from Example 8.6.4 may be used to simplify the portion outlined by the dotted box. The result is illustrated in Figure 8.33.

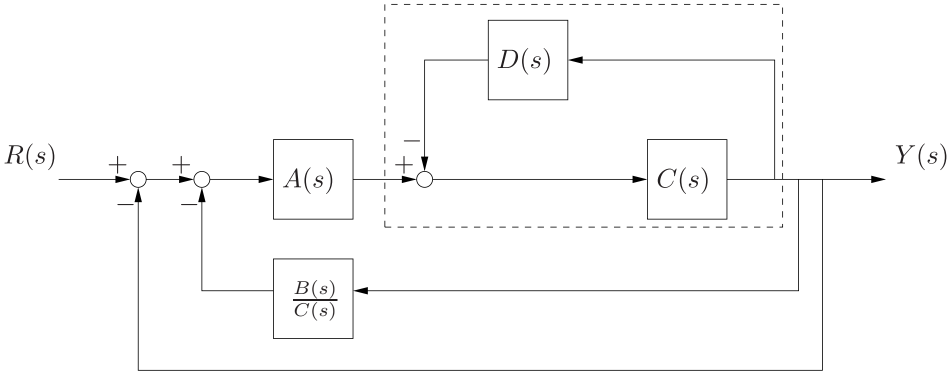


Figure 8.32. Block diagram for Example 8.6.6.

Since the two blocks in the top are adjacent, they are simply multiplied. Hence, they may be combined as is illustrated in Figure 8.34. After combining them, the portion of the block diagram in the dotted line is exactly of the form from Example 8.6.4. The simplified result is illustrated in

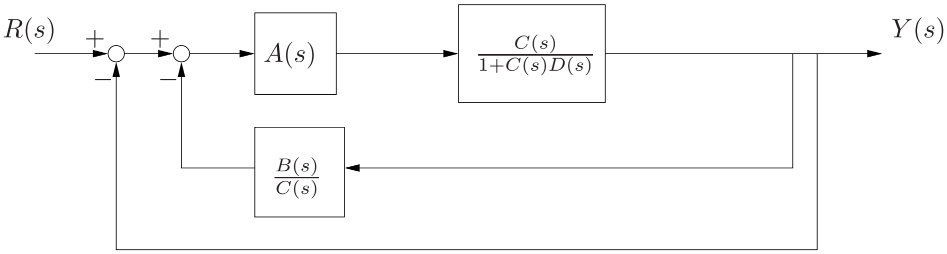


Figure 8.33. Block diagram for Example 8.6.6.

Figure 8.35 after the simplification of

$$\frac{\frac{A(s)C(s)}{1+C(s)D(s)}}{1+\frac{B(s)}{C(s)}\frac{A(s)C(s)}{1+C(s)D(s)}} = \frac{A(s)C(s)}{1+C(s)D(s)+A(s)B(s)}.$$

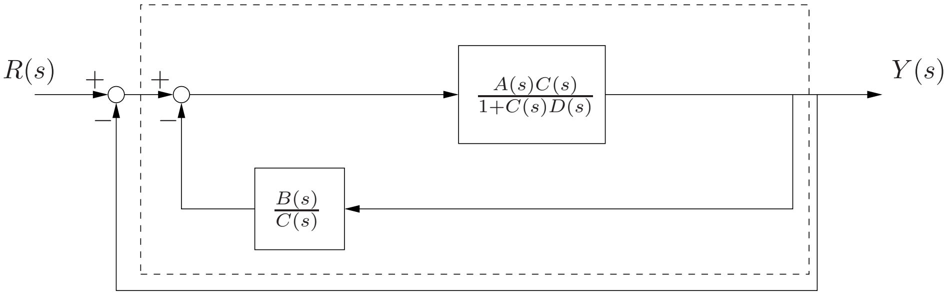


Figure 8.34. Block diagram for Example 8.6.6.

Finally, all of Figure 8.35 is of the form of the feedback loop from Example 8.6.4, so this would be the usual stopping point for this problem. Just for completeness we will take it one step further and reduce it to one block with one transfer function which, after some simplification, is illustrated in Figure 8.36. ■

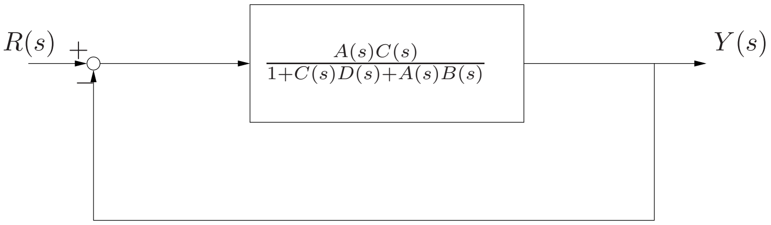


Figure 8.35. Block diagram for Example 8.6.6.

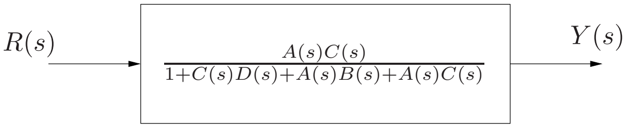


Figure 8.36. Block diagram for Example 8.6.6.

8.7 Computational Tools

8.7.1 Newton's method

8.7.2 Matlab

Four functions are particularly useful and will be highlighted here. Since we are dealing with polynomials in s, a function that multiplies polynomials will be handy, which is what `conv()` does. The function `pzmap()` computes and plots the poles and zeros of a transfer function. The functions `step()` and `impulse()` computes and plots an approximate numerical solution for the step and impulse responses, respectively.

The `conv()` function takes two vectors as arguments. The elements of the vectors are the coefficients of the powers of s in a polynomial. For example, $\left(s^2+3s+5\right)\left(7s^4+11s^2\right)$ is computed as follows:

```
>> conv([1 3 5],[7 0 11 0])
ans =
     7    21    46    33    55     0
```

which tells us that

$$\left(s^2+3s+5\right)\left(7s^4+11s^2\right)=7s^5+21s^4+46s^3+33s^2+55s.$$

Note that the 0's are necessary, both in the vectors entered into `conv()` as well as in the answer, to determine to what power of s the coefficient belongs.

The `step()` function computes an approximate numerical solution to the step response of a transfer function. If $G(s)$ is a transfer function, then the step response is given by

$$y(t) = \mathcal{L}^{-1}\left(G(s)\frac{1}{s}\right).$$

In its simplest implementation, the arguments to `step()` are vectors whose components are the coefficients of the polynomials in s in the numerator and denominator of $G(s)$, respectively. For example, to compute and plot the step response for

$$G(s) = \frac{s+2}{s^2+5s+10}$$

which is

$$y(t) = \mathcal{L}^{-1}\left(G(s)\frac{1}{s}\right)$$

enter

```
>> step([1 2],[1 5 10])
```

at the command prompt. If there is a need to record the response, enter

```
>> [y,t] = step([1 2],[1 5 10])
```

and then the vector `y` would contain the step response, and each element of `y` would correspond to the time contained in the corresponding element of `t`.

The `impulse()` function is the same as `step()` except it determines a numerical solution for the impulse response. The `pzmap()` function takes the input in the same format as `step()` and `impulse()`, but it plots the location of the poles and zeros of the transfer function. This is useful for transfer functions with polynomials that are of higher order than can be factored by hand.

8.7.3 Octave

The syntax for Octave is very similar to matlab, with the only exception that the `step()` and `impulse()` functions require that the transfer function be designated as such with the `tf()` function.

The `conv()` function takes two vectors as arguments. The elements of the vectors are the coefficients of the powers of s in a polynomial. For example, $\left(s^2+3s+5\right)\left(7s^4+11s^2\right)$ is computed as follows:

```
octave:> conv([1 3 5],[7 0 11 0])
ans =
     7    21    46    33    55     0
```

which tells us that

$$\left(s^2+3s+5\right)\left(7s^4+11s^2\right) = 7s^5+21s^4+46s^3+33s^2+55s.$$

Note that the 0's are necessary, both in the vectors entered into `conv()` as well as in the answer, to determine to what power of s the coefficient belongs.

The `step()` function computes an approximate numerical solution to the step response of a transfer function. If $G(s)$ is a transfer function, then the step response is given by

$$y(t) = \mathcal{L}^{-1}\left(G(s)\frac{1}{s}\right).$$

In its simplest implementation, the arguments to `step()` are vectors whose components are the coefficients of the polynomials in s in the numerator and denominator of $G(s)$, respectively. For example, to compute and plot the step response for

$$G(s) = \frac{s+2}{s^2+5s+10}$$

which is

$$y(t) = \mathcal{L}^{-1}\left(G(s)\frac{1}{s}\right)$$

enter

```
octave:> step(tf([1 2],[1 5 10]))
```

at the command prompt. The output to this function is illustrated in Figure 8.37. If there is a need to record the response, enter

```
octave:> [y,t] = step(tf([1 2],[1 5 10]))
```

and then the vector `y` would contain the step response, and each element of `y` would correspond to the time contained in the corresponding element of `t`.

The `impulse()` function is the same as `step()` except it determines a numerical solution for the impulse response. The `pzmap()` function takes the input in the same format as `step()` and `impulse()`, but it plots the location of the poles and zeros of the transfer function. This is useful for transfer functions with polynomials that are of higher order than can be factored by hand.

8.8 Exercises

Problem 8.1 Determine the inverse Laplace transform of

$$F(s) = \frac{a}{s^3\,(s+a)}.$$

Problem 8.2 Solve

$$\begin{aligned} \ddot{x} + 4x &= \cos 5t \\ x(0) &= 1 \\ \dot{x}(0) &= 1 \end{aligned}$$

using Laplace transforms.

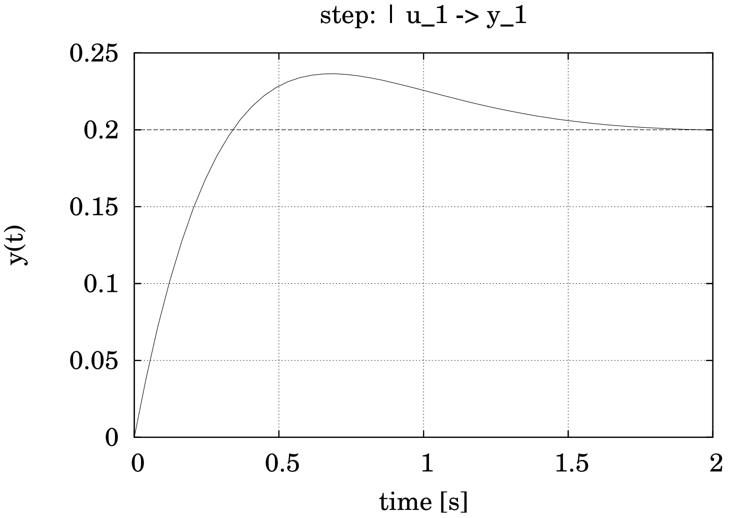


Figure 8.37. The step response of $G(s) = \frac{s+2}{s^2+5s+10}$ produced by the Octave command `step(tf([1 2],[1 5 10]))`.

Problem 8.3 Determine the solution to

$$\begin{aligned} \ddot{x} + 16x &= 0 \\ x(0) &= 1 \\ \dot{x}(0) &= 0 \end{aligned}$$

using Laplace transforms.

Problem 8.4 Determine the solution to

$$\begin{aligned} \ddot{x} + 16x &= 0 \\ x(0) &= 1 \\ \dot{x}(0) &= 1 \end{aligned}$$

using Laplace transforms.

Problem 8.5 Determine the solution to

$$\begin{aligned} \ddot{x} + 5\dot{x} + 6x &= 0 \\ x(0) &= 2 \\ \dot{x}(0) &= -5 \end{aligned}$$

using Laplace transforms.

Problem 8.6 Determine the solution to

$$\begin{aligned} \ddot{x} &= 0 \\ x(0) &= 0 \\ \dot{x}(0) &= 1 \end{aligned}$$

using Laplace transforms.

Problem 8.7 Determine the solution to

$$\begin{aligned} \ddot{x} + 2\dot{x} + 5x &= 6\cos 3t - 4\sin 3t \\ x(0) &= 0 \\ \dot{x}(0) &= 5 \end{aligned}$$

using Laplace transforms.

Problem 8.8 Determine the solution to

$$\begin{aligned} \ddot{x} + 16x &= \delta(t) \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

using Laplace transforms.

Problem 8.9 Determine the solution to

$$\begin{aligned} \ddot{x} + 16x &= \delta(t-2) \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

using Laplace transforms. Plot the solution.

Problem 8.10 Determine the solution to

$$\begin{aligned} \ddot{x} + 9x &= \mathbb{1}(t) \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

using Laplace transforms.

Problem 8.11 Determine the solution to

$$\begin{aligned} \ddot{x} + 9x &= \mathbb{1}(t-3) \\ x(0) &= 1 \\ \dot{x}(0) &= 0. \end{aligned}$$

Plot the solution.

Problem 8.12 Determine the solution to

$$\begin{aligned} \ddot{x} + 4x &= \begin{cases} \cos t, & 0 \leq t < \pi \\ 0, & \pi \leq t \end{cases} \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

using Laplace transforms. Plot your answer. Compare your answer with an approximate numerical solution for the differential equation obtained by writing a computer program or using a computer package such as Matlab.

Problem 8.13 Determine the solution to

$$\begin{aligned} \ddot{x} + 25x &= \begin{cases} t, & t \leq t < 1 \\ \cos(t-1), & 1 \leq t \end{cases} \\ x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

using Laplace transforms. Plot your answer. Compare your answer with an approximate numerical solution for the differential equation obtained by writing a computer program or using a computer package such as Matlab.

Problem 8.14 Solve

$$\begin{aligned} \dot{x} - 5x &= \begin{cases} 0 & t < 3 \\ t & 3 \leq t < 4 \\ 0 & 4 \leq t \end{cases} \\ x(0) &= 0. \end{aligned}$$

Problem 8.15 This problem is going to find the transfer function for a loudspeaker.

From physics, if a wire of length l carries a current of i amperes and is arranged at a right angle to a magnetic field of strength B Tesla, then the force (in Newtons) on the wire is at a right angle to the plane of the wire and magnetic field and has a magnitude

$$f = Bli. \tag{8.24}$$

In a speaker, the wire is usually coiled to fit a longer length in a small space.

This is illustrated schematically in Figure 8.38. A current, i through the coil, c causes a force, f on the mass (which, in this example, is the magnet) in the direction shown with a magnitude given by Equation 8.24.

1. Find the transfer function from the current through the speaker coil, i to the location of the mass, x.

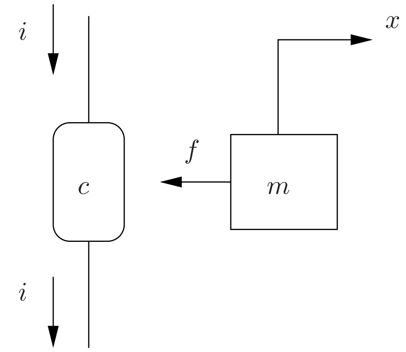


Figure 8.38. Speaker model for Problem 8.15.

2. Now we will attach a high pass filter to the speaker. The circuit is illustrated in Figure 8.39. An analysis of the properties of high pass filters is presented in Chapter 10.

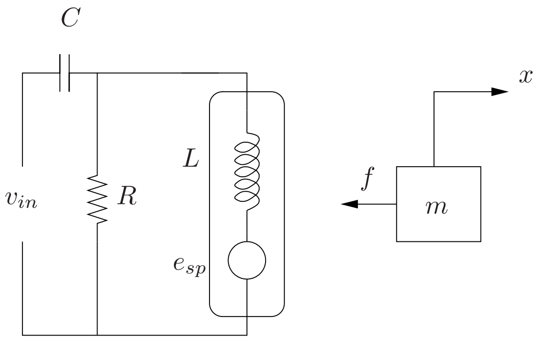


Figure 8.39. Speaker model for Problem 8.15.

Everything in the circuit should be obvious except the circle labeled e_{sp}. Just like a d.c. motor, there is a voltage drop across the speaker due to the speaker moving. It is given by

$$e_{sp} = Bl\dot{x}.$$

Find the transfer function from v_{in} to the position of the speaker, x.

Problem 8.16 Solve

$$\begin{aligned}\ddot{x}+9x &= \cos 2t\\ x(0) &= 1\\ \dot{x}(0) &= 1\end{aligned}$$

using Laplace transforms.

Problem 8.17 Consider the inverted pendulum illustrated in Figure 8.40.

1. Determime the equation of motion for the system.

2. Determine the best linear approximation for the equation of motion for small θ.

3. Using the linear approximation, determine the transfer function from the input torque, τ to the angle of the pendulum, θ.

4. Determine the transfer function from the input torque to the angular velocity of the pendulum.

5. Assume the torque is produced by a dc motor that is driven by the circuit illustrated in Figure 8.41. Determine the transfer function from the input voltage to the circuit to the pendulum angle, θ.

6. Assume the torque is produced by a dc motor that is driven by the circuit illustrated in Figure 8.41. Determine the transfer function from the input voltage to the circuit to the pendulum angle angular velocity.

Problem 8.18 Consider the system illustrated in Figure 8.42. A pulley with a mass moment of inertia J_1 and r_1 is subjected to a torque, τ. A light belt connects the first pulley to a second pulley with an inner and outer spool. The mass moment of inertia of the pulley is J_2. The inner spool has radius r_1 and the outer spool has a radius r_2. A belt around the outer spool of the second pully is attached to a third pulley and mass. The third pulley has radius r_2 and mass moment of inertia J_3. The mass has a mass m and is attached to a linear spring with spring constant k. The variable $x(t)$ represents the displacement of the mass with respect to an inertial coordinate frame. The variables $\theta_1(t)$, $\theta_2(t)$ and $\theta_3(t)$ represent the angular displacements of the pulleys.

1. Determine the transfer function from $\tau(t)$ to $x(t)$.

2. Assume the torque, τ is imposed on the first pulley by a dc motor

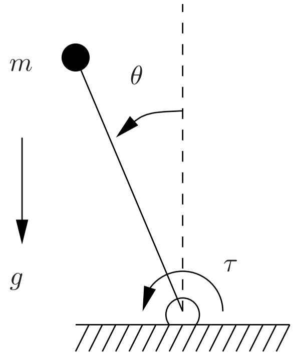


Figure 8.40. System for Problem 8.17.

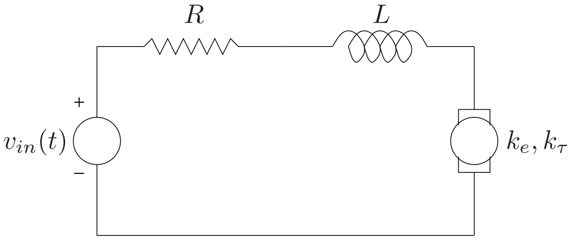


Figure 8.41. Motor circuit for Problem 8.17.

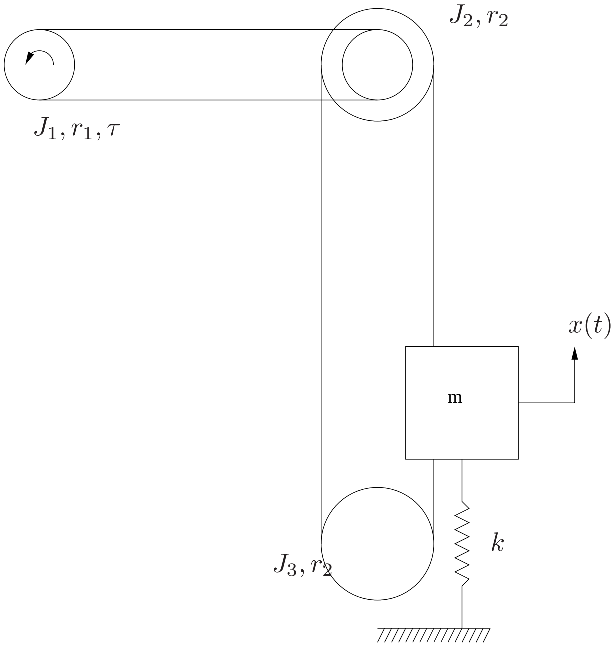


Figure 8.42. System for Problem 8.18.

driven by the circuit illustrated in Figure 8.43. Find the transfer function from the input voltage of the circuit, $v_{in}(t)$ to $x(t)$.

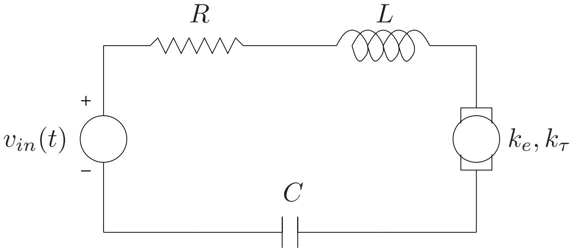


Figure 8.43. Motor circuit for Problem 8.18.

Chapter 9

Basic Control Theory: Analysis

9.1 Introduction

The exploitation of *feedback* was fundamental to many engineering breakthroughs of the 20th century. While feedback was certainly manifested well before that, such as in Watt's steam engine governor [], it was the need for and development of feedback amplifiers in the first half of the century that drove the development of the theory and analysis that made the use of feedback of general utility.

The utility of feedback has several aspects:

1. it may stabilize an otherwise unstable system;

2. it may improve the performance of a system;

3. it may make a system operate similarly regardless of variability in the components or operating conditions; and,

4. it may increase the bandwidth of the response of a system.

This chapter provides an detailed introduction to classical control theory. In order to develop some intuition regarding feedback and because it is ubiquitous, Section 9.2 presents an introduction to proportional–derivative–integral control. It is intended as an introduction to this very common control methodology and intended also as an introduction to the concept of feedback. Section 9.4 provides the definition of various quantities that are commonly used to specify desired control system behavior. Section 8.6 considers block diagrams, which are a graphical representation of the differential equations describing control systems.

The most critical section is Section 9.6 which discusses how system response is a function of the location in the complex plane of the poles of a transfer function. Understanding of this material is critical for understanding the root locus design method, which is in Section 9.9. Finally, Section 9.11 presents the frequency response analysis and design methods.

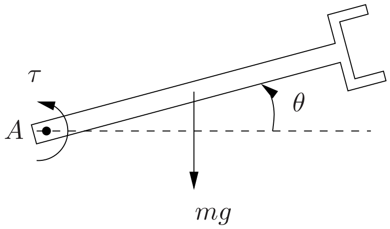


Figure 9.1. Robot arm mechanism.

9.2 PID Control

It is accurate to say that the vast majority of feedback control, particularly of mechanical systems, in industry is the so-called *proportional plus integral plus derivative (PID)* control. While designing PID controllers is usually somewhat *ad hoc*, this section will be devoted to the analysis of the features of these controllers as well as presenting a few "rules of thumb" with respect to designing them. The approach will be by way of an exhaustive example.

Example 9.2.1 Consider the simple "robot arm" illustrated in Figure 9.1. The arm is a rigid link constrained to rotate about the fixed point A. The arm has a moment of inertia, J and a center of mass located at a length, l (not shown) from the point A. The arm has a mass m and is subjected to gravity. The robot is fitted with a sensor that is able to determine the angle θ, which is measured from the horizontal position as indicated. Finally, a motor provides a torque, τ about the point A.

The purpose of *feedback control* is to determine a *control law* which makes the arm move to a desired angle, say θ_d and stay there despite any variable forces that may be applied to the arm (say by manipulating different objects of different masses). The idea of *feedback* is that the sensor measures θ which is then used (fed back) to determine a good value for the torque, τ.

Using Newton's law, the equation of motion for the system is

$$J\ddot{\theta} = \tau - mgl\cos\theta.$$

This is an ordinary, second order, nonlinear, constant coefficient, inhomogeneous differential equation. In order to make it much more amenable to analysis, we will assume that $\theta \ll 1$ so that $\cos\theta \approx 1$. In such a case, then the equation of motion is

$$J\ddot{\theta} = \tau - mgl. \tag{9.1}$$

■

9.2.1 Proportional control

The idea of proportional control is simple and has an obvious intuitive appeal: have the control input be proportional to the error in the system. How this would specifically be implemented in the robot arm and its efficacy is considered in the following examples. The first example, Example 9.2.2 uses the techniques from Chapter 3 to solve the resulting equations. In order to partially motivate the use of frequency domain analysis tools from Chapter 8 in control theory and the use of a transfer function, developed subsequently in section 8.5, Example 9.2.3 presents the same analysis, but using Laplace transform tools.

Example 9.2.2 Returning to the system in example 9.2.1, using proportional control would be to specify that

$$\tau(t) = k_p \left(\theta_d(t) - \theta(t)\right), \tag{9.2}$$

where, as stated previously, $\theta_d(t)$ is the desired position of the arm at time t. Thus, the torque, τ is proportional to the error, $\theta_d(t) - \theta(t)$. The proportionality constant, k_p is called the *proportional gain.*

Depending upon the system, sometimes proportional controls suffices. However, in the case at hand, it is straightforward to illustrate that the approach has several drawbacks. Substituting the control law from equation 9.3 into the (linearized) equation of motion from equation 9.1 gives

$$J\ddot{\theta} = k_p \left(\theta_d - \theta\right) - mgl$$

or

$$J\ddot{\theta} + k_p\theta = k_p\theta_d - mgl, \tag{9.3}$$

which is an ordinary, second order, constant coefficient, linear, inhomogeneous differential equation. The rest of this example will analyze this system using the tools and methods from Chapter 3. Obviously the homogeneous solution is

$$\theta_h(t) = c_1 \cos\left(\sqrt{\frac{k_p}{J}}t\right) + c_2 \sin\left(\sqrt{\frac{k_p}{J}}t\right)$$

and the particular solution depends upon the form of $\theta_d(t)$.

In order to precede with the analysis, let θ_d be a specified constant. In that case,

$$\theta_p(t) = \theta_d - \frac{mgl}{k_p},$$

and the general solution is

$$\theta(t) = c_1 \cos\left(\sqrt{\frac{k_p}{J}}t\right) + c_2 \sin\left(\sqrt{\frac{k_p}{J}}t\right) + \theta_d - \frac{mgl}{k_p}.$$

In order to precede further and plot some solutions, let us specify some numerical values for the initial conditions and all parameter values except

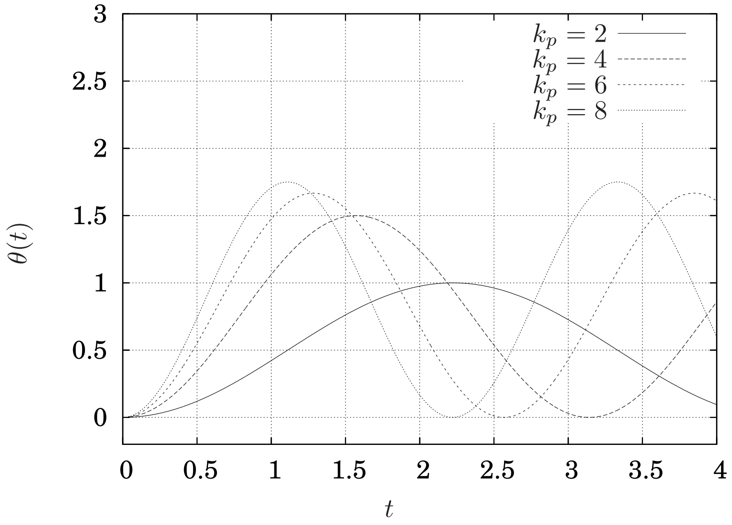


Figure 9.2. Response of robot arm under proportional control.

for k_p; namely,

$$\begin{aligned} J &= 1 \\ mgl &= 1 \\ \theta(0) &= 0 \\ \dot{\theta}(0) &= 0 \\ \theta_d &= 1 \end{aligned}$$

in which case

$$\theta(t) = \left(\frac{1}{k_p} - 1\right) \cos \sqrt{k_p} t + 1 - \frac{1}{k_p}. \tag{9.4}$$

While $\theta_d = 1$ violates the assumption that θ is small, because the Equation 9.2 is linear, the nature of the solutions will be qualitatively the same as the case when the assumption is satisfied. In other words, due to linearity, the shape of the response will be the same regardless of whether the desired value is one or 0.01. The value of one is used simply to have the equations in a somewhat "normalized" form.

A plot of the movement of the robot arm for various values of k_p is illustrated in Figure 9.2. Note that with proportional control

1. the solutions are oscillatory and are not decaying;

2. as k_p increases the frequency of oscillation increases;

3. as k_p increases the average value of the oscillation approaches $\theta_d = 1$; and,

4. as k_p increases, the earliest time at which $\theta = \theta_d$ decreases.

Clearly, using proportional control for this example is not adequate if we desire that the robot arm approach θ_d and not oscillate about it. ■

Now, the same analysis is repeated using Laplace transforms.

Example 9.2.3 Referring back to example 9.2.2, the equation of motion for proportional feedback is

$$J\ddot{\theta} + k_p\theta = k_p\theta_d - mgl.$$

Assuming zero initial conditions, the Laplace transform of the above equation is

$$Js^2\Theta(s) + k_p\Theta(s) = k_p\Theta_d(s) - \frac{mgl}{s}.$$

and

$$\Theta(s) = \frac{k_p s\Theta_d - mgl}{Js^2 + k_p}. \tag{9.5}$$

Assuming, as before, $mgl = J = 1$ and $\theta_d = 1$ so $\Theta_d(s) = \frac{1}{s}$, then

$$\Theta(s) = \frac{k_p - 1}{s\left(s^2 + k_p\right)}. \tag{9.6}$$

The inverse Laplace transform for 9.6 is exactly the same as equation 9.4, and for various k_p values must give the same response curves as are illustrated in Figure 9.2 ■

9.2.2 Proportional plus derivative control

The idea of proportional plus derivative control is that, in contrast to proportional control, the control law should also reflect the derivative of the error. The intuition is that while the error may be positive or negative, how large the control input should be should also depend upon whether the error is increasing or decreasing.

Referring to Figure 9.2, the idea is that, for example, for the case of $k_p = 8$ and $0 < t < 0.6$ where the error, $\theta_d - \theta > 0$, since the error is decreasing, reducing τ relative to what it is for just proportional control should reduce the amount by which the response "overshoots" during the time interval from approximately $0.6 < t < 1.6$.

Example 9.2.4 Returning to the system in examples 9.2.1 and 9.2.2, using proportional plus derivative control (PD control) would be to specify that

$$\tau(t) = k_p\left(\theta_d(t) - \theta(t)\right) + k_d\left(\dot{\theta}_d(t) - \dot{\theta}(t)\right),$$

where, as stated previously, $\theta_d(t)$ is the desired position of the arm at time t. Thus, the torque, τ is not simply proportional to the error, but also includes a term proportional to the derivative of the error. The proportionality constant for the derivative term, k_d is called the *derivative gain*.

Substituting this into the equation of motion and rearranging gives

$$J\ddot{\theta} + k_d\dot{\theta} + k_p\theta = k_p\theta_d + k_d\dot{\theta}_d - mgl. \tag{9.7}$$

In this case, the homogeneous solution is

$$\theta_h = e^{-\frac{k_d}{2J}t}\left(c_1\cos\left(\frac{\sqrt{4k_pJ - k_d^2}}{2J}t\right) + c_2\sin\left(\frac{\sqrt{4k_pJ - k_d^2}}{2J}t\right)\right)$$

as long as $k_d^2 < 4Jk_p$. Note that the oscillations due to the homogeneous solution decay with time as long as $k_p, k_d, J > 0$, so this potentially improves the performance over proportional control since the continued oscillations present in proportional control will decay with derivative control. Thus, the steady state solution depends only upon the form of the particular solution, which, of course, depends upon the exact nature of $\theta_d(t)$.

In order to continue the analysis as before, let us consider the case where θ_d is a constant. In that case, since $\dot{\theta}_d = 0$, the particular solution is the same the proportional control case in example 9.2.2 and hence

$$\begin{aligned} \theta(t) &= e^{-\frac{k_d}{2J}t}\left(c_1\cos\left(\frac{\sqrt{4k_pJ - k_d^2}}{2J}t\right) + c_2\sin\left(\frac{\sqrt{4k_pJ - k_d^2}}{2J}t\right)\right) + \\ &\quad \theta_d - \frac{mgl}{k_p}. \end{aligned}$$

Clearly, since the homogeneous solution decays for positive k_d, k_p and J, the steady state solution is

$$\theta_{ss}(t) = \theta_d - \frac{mgl}{k_p}.$$

Before plotting some solutions with numerical values, note for the steady state response, a very large k_p is desirable since it makes $\theta_{ss} \to \theta_d$. Also, if k_d increases, any oscillations should decay more quickly.

To plot a solution, let

$$\begin{aligned} J &= 1 \\ mgl &= 1 \\ \theta(0) &= 0 \\ \dot{\theta}(0) &= 0 \\ \theta_d &= 1 \end{aligned}$$

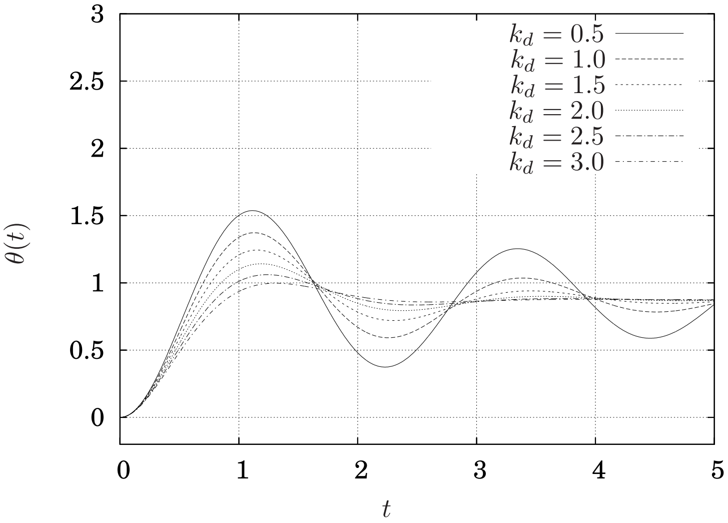


Figure 9.3. Response of robot arm under PD control with fixed $k_p = 8.0$ and various k_d.

in which case

$$\begin{aligned}\theta(t) \quad = \quad & e^{-\frac{k_d}{2}t}\left[\left(\frac{1}{k_p}-1\right)\cos\left(\frac{\sqrt{4k_p-k_d^2}}{2}t\right)+\right.\\ & \left.\left(\frac{k_d\,(1-k_p)}{k_p\sqrt{4kp-k_d^2}}\right)\sin\left(\frac{\sqrt{4k_p-k_d^2}}{2}t\right)\right]+1-\frac{1}{k_p}. \qquad (9.8)\end{aligned}$$

Figure 9.3 illustrates the response for a fixed $k_p = 8.0$ and various k_d values. Note that as k_d is increased, the oscillations decay more quickly and the value of the first maximum (near $t = 1.0$) is decreased. Also note that changing k_d does not affect the steady state value of $\theta(t)$ and that the steady state error is nonzero.

Figure 9.4 illustrates the response for a fixed $k_d = 1.0$ and various k_p values. Note that as k_p is increased, the final steady state error decreases (recall $\theta_d = 1$), but that the initial overshoot is greatly increased and the frequency of oscillation increases. ■

Now, the same analysis is repeated using Laplace transforms.

Example 9.2.5 Picking up from equation 9.7 from example 9.2.4, the equation of motion for proportional plus derivative control is

$$J\ddot{\theta} + k_d\dot{\theta} + k_p\theta = k_p\theta_d + k_d\dot{\theta}_d - mgl.$$

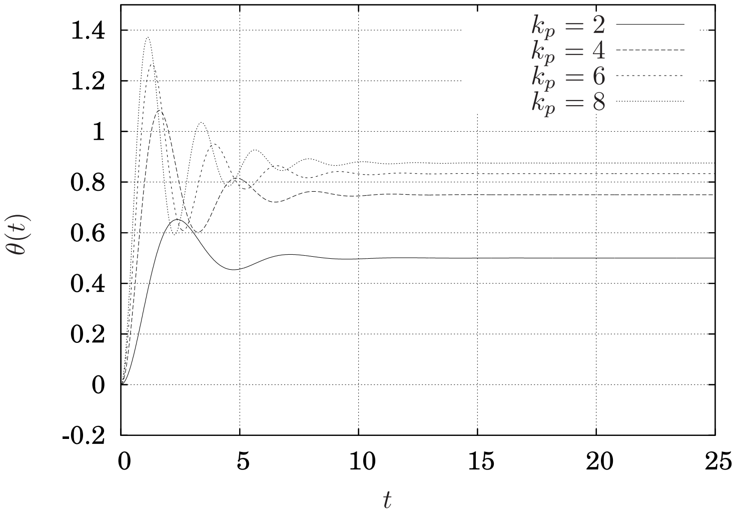


Figure 9.4. Response of robot arm under PD control with fixed $k_d = 1.0$ and various k_p.

Assuming zero initial conditions and Laplace transforming gives

$$Js^2\Theta(s) + k_d s\Theta(s) + k_p\Theta(s) = k_p\Theta_d(s) + k_d s\Theta_d(s) - \theta_d(0) - \frac{mgl}{s}$$

which gives

$$\Theta(s) = \frac{k_d\Theta_d(s)s^2 + k_p\Theta_d(s)s - s\theta_d(0) - mgl}{s\left(Js^2 + k_d s + k_p\right)}.$$

As before, the nature of the solution depends upon $\Theta_d(s)$. Assuming $\theta_d = J = mgl = 1$, then

$$\Theta_d(s) = \frac{1}{s}$$ ■

and

$$\Theta(s) \quad = \quad \frac{k_d s + k_p - 1}{s\left(s^2 + k_d s + k_p\right)},$$

and the inverse Laplace transform is the same as the solution from example 9.2.4 given in equation 9.8 and plotted in Figures 9.3 and 9.4 for various gain values.

Also, using Theorem 8.3.17,

$$\begin{aligned}\lim_{t\to\infty}\theta(t) &= \lim_{s\to 0} s\Theta(s)\\ &= \lim_{s\to 0} s\frac{k_p-1}{s\left(s^2+k_ds+k_p\right)}\\ &= 1-\frac{1}{k_p}.\end{aligned}$$

9.2.3 Proportional plus integral plus derivative control

With proportional plus derivative plus integral (PID) control, a third term is added to proportional plus derivative control that is, naturally, the integral of the error. In examples 9.2.2, 9.2.3, 9.2.4 and 9.2.5 there was always a steady state error, *i.e.,* $\lim_{t\to\infty}\theta(t)\neq\theta_d$. The idea behind integral control is that as time increases, if there is a consistent error the input to the system will increase with time to compensate for the error. The need for integral control in many problems is obvious considering the robot arm from these examples. In the case of both proportional and proportional plus derivative control, if there is no error, *i.e.,* $\theta=\theta_d$ then $\tau=0$. If the torque is zero, then there is nothing to offset the torque caused by gravity and the arm will rotate. The steady state value in the case of PD control is the angle at which the error is great enough to cause an error that will result in a torque that will offset the torque due to gravity.

The following example illustrates the the efficacy of integral control with respect to eliminating steady state error.

Example 9.2.6 Returning, yet again, to the system from example 9.2.1, adding integral control yields an expression for the torque of the form

$$\tau=k_p\left(\theta_d-\theta\right)+k_d\left(\dot{\theta}_d-\dot{\theta}\right)+k_i\int_0^t\theta_d(\hat{t})-\theta(\hat{t})d\hat{t}$$

so the equation of motion for the robot arm becomes

$$J\ddot{\theta}+k_d\dot{\theta}+k_p\theta=k_p\theta_d+k_d\dot{\theta}_d+k_i\int_0^t\theta_d(\hat{t})-\theta(\hat{t})d\hat{t}-mgl. \tag{9.9}$$

This is a second order integral-differential equation and there are various ways to handle the integral term.

1. One way to eliminate the integral is to differentiate the entire equation with respect to time as follows

$$J\dddot{\theta}+k_d\ddot{\theta}+k_p\dot{\theta}=k_p\dot{\theta}_d+k_d\ddot{\theta}_d+k_i\left(\theta_d-\theta\right)$$

or

$$J\dddot{\theta}+k_d\ddot{\theta}+k_p\dot{\theta}+k_i\theta=k_i\theta_d+k_p\dot{\theta}_d+k_d\ddot{\theta}_d.$$

Note that the solution to this equation requires an initial condition for $\ddot{\theta}$; however, that may be computed from equation 9.9 using $\theta(0)$ and $\dot{\theta}(0)$. In particular, if $\theta(0) = \dot{\theta}(0) = 0$

$$\ddot{\theta}(0) = k_p\theta_d(0) + k_d\dot{\theta}_d(0).$$

In order to proceed, assume, as before, $mgl = J = \theta_d = 1$ and $\theta(0) = \dot{\theta}(0) = 0$, which gives

$$\begin{aligned} \dddot{\theta} + k_d\ddot{\theta} + k_p\dot{\theta} + k_i\theta &= k_i \qquad (9.10)\\ \theta(0) &= 0\\ \dot{\theta}(0) &= 0\\ \ddot{\theta}(0) &= k_p\theta_d(0) = k_p. \end{aligned}$$

The particular solution is easy to obtain from the method of undetermined coefficients; namely,

$$\theta_p(t) = 1.$$

The homogeneous solution, on the other hand, depends upon the roots of the characteristic equation

$$\lambda^3 + k_d\lambda^2 + k_p\lambda + k_i = 0,$$

which clearly depends upon k_i, k_d and k_p.

In order to proceed, let $k_i = 32$, $k_d = 8$ and $k_p = 24$, in which case

$$\begin{aligned} \lambda_1 &= -6\\ \lambda_2 &= -2 - 2i\\ \lambda_3 &= -2 + 2i \end{aligned}$$

and

$$\theta_h(t) = c_1e^{-6t} + e^{-2t}\left(c_2\cos 2t + c_3\sin 2t\right),$$

and hence the general solution is

$$\theta(t) = c_1e^{-6t} + e^{-2t}\left(c_2\cos 2t + c_3\sin 2t\right) + 1.$$

Evaluating the initial conditions gives

$$\theta(t) = 1 + 2e^{-4t} + e^{-2t}\left(\sin 2t - 3\cos 2t\right).$$

Note, at $t \to \infty$, $\theta(t) \to 1 = \theta_d$, so this method does, indeed, eliminate the steady state error. Note also, though, that this solution is only valid for the specific values for $k_p = 24$, $k_d = 8$ and $k_i = 32$.

Figure 9.5 illustrate the response of the arm for fixed values of k_p, k_d and various k_i. Note that any nonzero value for k_i eliminates the

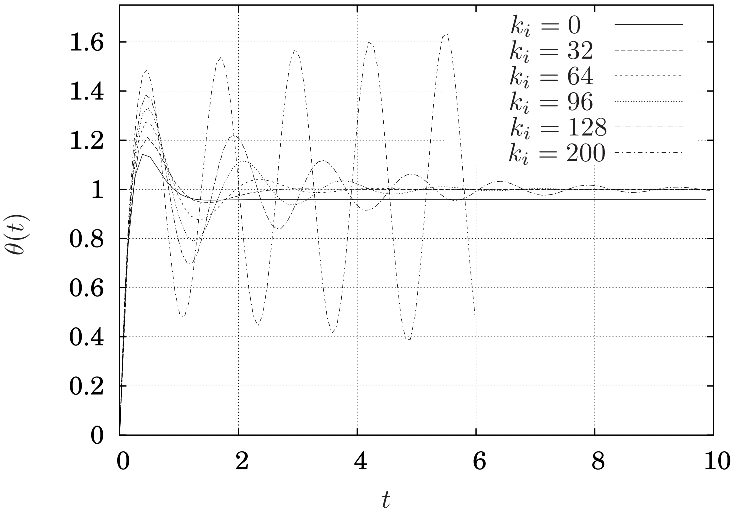


Figure 9.5. Response of robot arm under PID control for fixed $k_p = 24$ and $k_d = 8$ and various k_i.

steady state error; however, increasing k_i increases the magnitude and duration of the transient oscillations, and, if large enough, destabilizes the system ($k_i = 200$). The reason for this is that k_i is the coefficient of θ in Equation 9.10 and hence appears in the characteristic equation. As k_i gets large, one or more of the roots of the characteristic equation has a positive real part, which corresponds to an exponential with a positive coefficient.

2. An alternative to differentiating equation 9.9 is to convert the system into a coupled set of ordinary differential equations. The approach is mathematically equivalent to the preceding approach, but perhaps is more amenable to numerical analysis. Note that since

$$\frac{d}{dt}\int_0^t \theta_d(\hat{t}) - \theta(\hat{t})d\hat{t} = \theta_d(t) - \theta(t),$$

if we define a new variable, $\hat{I}$ (I for "integral"), then equation 9.9 is equivalent to the two ordinary differential equations

$$\begin{aligned} J\ddot{\theta} + k_d\dot{\theta} + k_p\theta &= k_p\theta_d + k_d\dot{\theta}_d + k_i\hat{I} \\ \dot{\hat{I}} &= \theta_d - \theta. \end{aligned}$$

If

$$\begin{aligned} x_1 &= \theta \\ x_2 &= \dot{\theta} \\ x_3 &= \hat{I}, \end{aligned}$$

then

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ \frac{k_p\theta_d + k_d\dot{\theta}_d + k_i x_3 - k_d x_2 - k_p x_1}{J} \\ \theta_d - x_1 \end{bmatrix},$$

which can be solved analytically using the methods from Chapter 6 or, perhaps more conveniently, can be solved numerically using the methods from Chapter 13. Regardless, the same results will be obtained as outlined above. ■

Example 9.2.7 finally, for completeness, we will determine the PID control equations using Laplace transforms. The equation of motion for the robot arm under PID control was given in Equation 9.9 and is

$$J\ddot{\theta} + k_d\dot{\theta} + k_p\theta = k_p\theta_d + k_d\dot{\theta}_d + k_i\int_0^t \theta_d(\hat{t}) - \theta(\hat{t})d\hat{t} - mgl.$$

Taking the Laplace transform with zero initial conditions gives

$$\left(Js^2 + k_d s + k_p + \frac{k_i}{s}\right)\Theta(s) = \left(k_p + k_d s + \frac{k_i}{s}\right)\Theta_d(s)$$ ■

or

$$\begin{aligned} \Theta(s) &= \frac{k_p + k_d s + \frac{k_i}{s}}{Js^2 + k_d s + k_p + \frac{k_i}{s}}\Theta_d(s) \\ &= \frac{k_d s^2 + k_p s + k_i}{Js^3 + k_d s^2 + k_p s + k_i}\Theta_d(s). \end{aligned}$$

Using Theorem 8.3.17 and assuming $\Theta_d(s) = \frac{1}{s}$, *i.e.*, θ_d is a unit step input

$$\begin{aligned} \lim_{t\to\infty}\theta(t) &= \lim_{s\to 0} s\Theta(s) \\ &= \lim_{s\to 0} s\frac{k_d s^2 + k_p s + k_i}{Js^3 + k_d s^2 + k_p s + k_i}\frac{1}{s} \\ &= 1. \end{aligned}$$

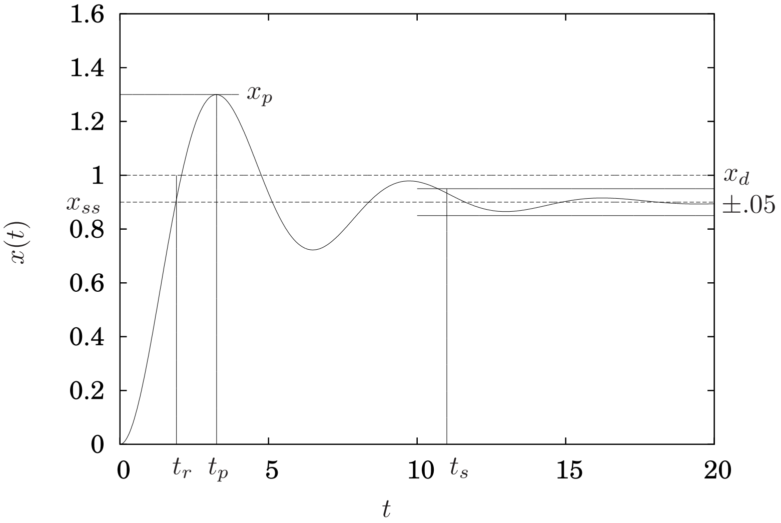


Figure 9.6. Time domain specifications definitions for a unit step input.

9.3 Sensitivity

9.4 Time Domain Specification

The qualitative discussions regarding the effect of altering controller gains in section 9.2 practically beg us to be more precise and quantitative about the nature of the response of a system. Consider a generic system response to a unit step input illustrated in Figure 9.6.

From the diagram, the following quantities are apparent.

1. The *rise time*, t_r, is the time at which the response is first equal to the magnitude of the input. For a unit step input, it is the time at which the response is first equal to one. If the system is overdamped, then the response may only asymptotically approach the desired value. In that case the rise time may be defined to be the time it takes to achieve 90% of the desired value. Unless otherwise specified, in this book the rise time will refer to the first definition.

2. The *peak time*, t_p, is the time at which the response reaches its maximum value.

3. The *settling time*, t_s, is the time after which the response always stays within a range of its steady state value. In Figure 9.6, this is illustrated

as 0.9 ± 0.05, but other ranges may be specified as a certain percentage, *e.g.,* "the 3% settling time."

4. The *maximum percentage overshoot*, O, is defined to be the percentage that the peak value, x_p exceeds the desired value, x_d, *i.e.,*

$$O = \frac{x_p - x_d}{x_d}.$$

Collectively these terms are referred to as the *transient response* since they describe how the system transitions from the initial conditions to the steady state behavior, but do not describe the steady state behavior. With regard to the steady state, the *steady state error* is the difference between the steady state value of the response and the desired value, *i.e.,*

$$e_{ss} = x_{ss} - x_d.$$

The tools used to determine the nature of the transient response are discussed subsequently in Section 9.6. The usual tool for the steady state error is Theorem 8.3.17, the Final Value Theorem.

These time domain specifications may be used to specify the manner in which a control system should respond. For example, it may be desired that a control surface on an airplane wing, say an aileron, respond with less than 1 second rise time, less than 1% overshoot and a settling time less than 3 seconds. As in many design problems, it may or may not be possible to meet all the specifications. Whether or not it is possible depends, among other things, upon the dynamics of the system and the nature of the actuation.

9.5 Block Diagram Representation and Algebra

Moved to Section 8.6.

9.6 Response *versus* Pole Location

This section considers the mathematical basis for the rest of this chapter. Understanding this section is critical for a fundamental understanding of what follows. The main concept is that *the nature of the response of a system is governed by the location in the complex plane of the poles of the transfer function describing the system.*

If we consider a generic transfer function, $G(s)$ and the relationship between the reference signal for the system, $R(s)$ and the output, $Y(s)$, we have

$$Y(s) = G(s)R(s).$$

If we were to solve this for the time domain response of the system, $y(t)$, we would need to know the input, $R(s)$ and then would compute a partial fraction expansion of $G(s)R(s)$ to algebraically manipulate the expression to be a combination of terms that appear in a Laplace transform table.

Example 9.6.1 To solve

$$Y(s) = -\frac{6s+5}{(s+3)\left(s^2+4\right)} \tag{9.11}$$

for $y(t)$, we would convert

$$\begin{aligned} Y(s) &= -\frac{6s+5}{(s+3)\left(s^2+4\right)} \\ &= \frac{c_1}{s+3} + \frac{c_2 s + c_3}{s^2+4} \\ &= \frac{1}{s+3} - \frac{s+3}{s^2+4} \\ &= \frac{1}{s+3} - \frac{s}{s^2+4} - \frac{3}{2}\frac{2}{s^2+4}, \end{aligned}$$

which corresponds to

$$y(t) = e^{-3t} - \cos 2t - \frac{3}{2}\sin 2t. \tag{9.12}$$

Now, if we look at the original transfer function in Equation 9.11, the poles (the values of s for which the denominator is equal to zero, see Definition 8.3.3), are $s=-3$ and $s=\pm 2i$. It is no coincidence that the solution in Equation 9.12 is a linear combination of an exponential with a -3 in the exponent, corresponding to the pole at $s=-3$ and sine and cosine functions with a frequency of 2, corresponding to the complex conjugate pair of poles at $s=\pm 2i$. ■

Because they are commonly used to characterize the nature of a transfer function, the time domain solution of the output for two specific inputs are given names.

Definition 9.6.2 For a transfer function, $G(s)$, input $R(s)$ and output $Y(s)$ which satisfy

$$Y(s) = G(s)R(s),$$

the *unit impulse response*, or simply the *impulse response* is the inverse Laplace transform of the output when the input is an impulse. Since $\mathcal{L}\left(\delta(t)\right) = 1$, the impulse response is given by the inverse Laplace transform of the transfer function

$$y_\delta(t) = \mathcal{L}^{-1}\left(G(s)\right).$$

◇

Definition 9.6.3 For a transfer function, $G(s)$, input $R(s)$ and output $Y(s)$ which satisfy

$$Y(s) = G(s)R(s),$$

the *unit step response*, or simply the *step response* is the inverse Laplace transform of the output when the input is a unit step function. Since $\mathcal{L}\left(\mathbb{1}(t)\right) = \frac{1}{s}$, the impulse response is given by

$$y_{\mathbb{1}}(t) = \mathcal{L}^{-1}\left(\frac{G(s)}{s}\right).$$

◇

A detailed study of Table 8.1, would make it clear that what differentiates the fundamental nature of the response of a system is the location of the poles. We will consider the various possible cases which depend upon whether the pole is real, zero, purely imaginary or complex.

9.6.1 Real poles

First we will consider the case where a transfer function has a pole that is real.

Consider

$$Y(s) = \frac{1}{s+p}R(s). \tag{9.13}$$

Note that $Y(s)$ has a pole at $s = -p$. Regardless of the nature of $R(s)$, a partial fraction expansion of Equation 9.13 will be of the form

$$Y(s) = \frac{c_1}{s+p} + \sum \hat{R}(s)$$

where $\sum \hat{R}(s)$ are the terms in the partial fraction expansion due to the input.

So, *regardless of the input*, if p is real, $y(t)$ will contain a term of the form e^{-pt}, it i.e.,

$$y(t) = c_1 e^{-pt} + \text{other terms.}$$

Hence, we have the following proposition.

Proposition 9.6.4 *If a transfer function has a pole that is a real, it will have an exponential term in the solution and that exponential term will decay to zero if $p < 0$ and will grow unbounded if $p > 0$.*

Example 9.6.5 Predict the unit step response of the two transfer functions

$$G_1(s) = \frac{2}{s+2}$$

and

$$G_2(s) = \frac{4}{s+4}$$

without actually computing the inverse Laplace transform.

We want to compare

$$\begin{aligned} Y_1(s) &= G_1(s)R(s) \\ &= \frac{2}{s+2}\frac{1}{s} \end{aligned}$$

and

$$\begin{aligned} Y_2(s) &= G_2(s)R(s) \\ &= \frac{4}{s+4}\frac{1}{s}. \end{aligned}$$

Using Theorem 8.3.17,

$$\begin{aligned}\lim_{t\to\infty} y_1(t) &= \lim_{s\to 0} s\frac{2}{s+2}\frac{1}{s}\\ &= 1\end{aligned}$$

and

$$\begin{aligned}\lim_{t\to\infty} y_2(t) &= \lim_{s\to 0} s\frac{4}{s+4}\frac{1}{s}\\ &= 1.\end{aligned}$$

The final value theorem may be applied to both of these since all the poles of both $sG_1(s)R(s)$ and $sG_2(s)R(s)$ are in the left half plane if $R(s)$ is a step function. If we were to compute them, the partial fraction expansions would be of the form

$$Y_1(s) = \frac{c_1}{s+2} + \frac{c_2}{s}$$

and

$$Y_2(s) = \frac{c_1}{s+4} + \frac{c_2}{s}.$$

Observe that in both cases the first term gives an exponential solution with a negative coefficient in the exponent and the second term gives a constant value. Since $G_1(s)$ has a pole at $s = -2$ and $G_2(s)$ has a pole at $s = -4$, as is illustrated in Figure 9.7, the exponential part of the solution in $Y_2(s)$ decays more quickly than the exponential part in $Y_1(s)$. Hence we may conclude that $y_2(t)$ converges more quickly to the steady state value than $y_1(t)$. This is verified by in Figure 9.8 which compares the two solutions. The association between pole locations, Figure 9.7, and the nature of the response, Figure 9.8, cannot be emphasized enough. In this particular example, if another system were compared that had a pole farther to the left, then its response would be even faster, and if another system had a pole farther to the right (but still less than zero), its response would be slower. If any system has a pole to the right of the imaginary axis, the solution will blow up, *i.e.,* the system will be unstable.

For comparison, the impulse response of two transfer functions with pole locations as the same points as before are illustrated in Figure 9.9. Again, the system with the pole farther to the left has a faster decaying transient response. ■

9.6.2 Poles at the origin

Now we will consider some poles at the origin.

Consider

$$Y(s) = \frac{1}{s}R(s). \tag{9.14}$$

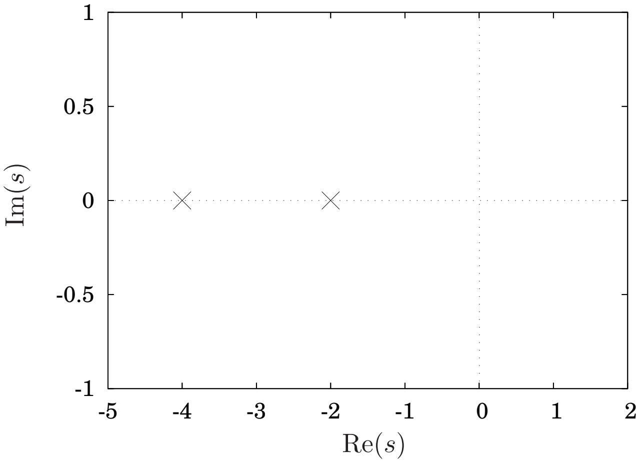


Figure 9.7. Pole locations for $G_1(s) = \frac{2}{s+2}$ and $G_2(s) = \frac{4}{s+4}$.

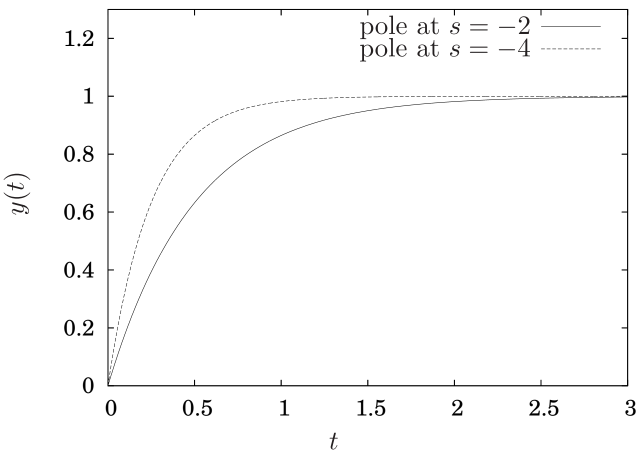


Figure 9.8. Step response for $G_1(s) = \frac{2}{s+2}$ and $G_2(s) = \frac{4}{s+4}$.

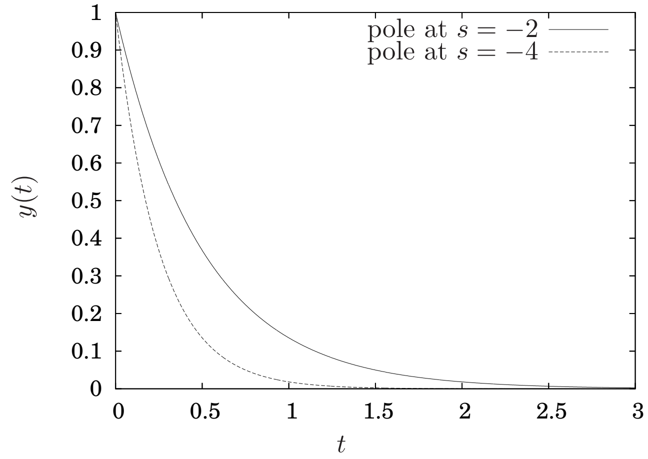


Figure 9.9. Impulse response for $G_1(s) = \frac{1}{s+2}$ and $G_2(s) = \frac{1}{s+4}$.

Note that $Y(s)$ has a pole at $s = 0$. Regardless of the nature of $R(s)$, a partial fraction expansion of Equation 9.14 will be of the form

$$Y(s) = \frac{c_0}{s} + \sum \hat{R}(s)$$

where $\sum \hat{R}(s)$ are the terms in the partial fraction expansion due to the input.

So, *regardless of the input*, $y(t)$ will contain a term of the form c_0, it i.e.,

$$y(t) = c_0 + \text{other terms}.$$

We may conclude from that, in general, if a transfer function has a pole at the origin it will have a constant term in the solution.

If the transfer function has multiple poles at the origin, *i.e.,*

$$Y(s) = \frac{1}{s^n} R(s)$$

then the partial fraction expansion will be of the form

$$\begin{aligned} Y(s) &= \frac{c_0 s^{n-1} + c_1 s^{n-2} + \cdots + c_{n-1}}{s^n} + \sum \hat{R}(s) \\ &= \frac{c_0}{s} + \frac{c_1}{s^2} + \cdots + \frac{c_{n-1}}{s^n} + \sum \hat{R}(s). \end{aligned}$$

So, *regardless of the input*, $y(t)$ will contain an $n - 1$th order polynomial in t, *i.e.,*

$$y(t) = c_0 + c_1 t + \frac{c_2}{2} t^2 + \cdots + \frac{c_{n-1}}{(n-1)!} t^{n-1} + \text{other terms}.$$

Hence, we have the following.

Proposition 9.6.6 *If a transfer function has multiple poles at the origin, it will have a polynomial term in the solution that has an order one less than the multiplicity of the pole at the origin.*

9.6.3 Purely imaginary poles

Now we will consider a complex conjugate pair of purely imaginary poles.

Consider

$$Y(s) = \frac{1}{s^2+\omega^2}R(s),$$

which as poles at $s = \pm i\omega$. The partial fraction expansion will be of the form

$$Y(s) = \frac{c_1 s}{s^2+\omega^2} + \frac{c_2}{\omega}\frac{\omega}{s^2+\omega^2} + \sum \hat{R}(s)$$

and the solution will be of the form

$$y(t) = c_1 \cos\omega t + \frac{c_2}{\omega}\sin\omega t + \text{other terms}.$$

So, we have shown the following.

Proposition 9.6.7 *If a transfer function has a purely imaginary complex conjugate pair of poles, it will have sine and cosine terms in the solution. If the magnitude of the purely imaginary pair of poles is increased, the frequency of oscillation will increase.*

Example 9.6.8 The poles of

$$G_1(s) = \frac{2}{s^2+4}$$

and

$$G_2(s) = \frac{8}{s^2+16}$$

are plotted in Figure 9.10. The corresponding step responses are plotted in Figure 9.11. Note as the poles move farther from the real axis, the frequency of oscillation increases. For comparison, the impulse response, when $R(s) = 1$, for both cases is illustrated in Figure 9.12. ■

9.6.4 Complex conjugate poles

Finally, the last case to consider is when a transfer function contains a complex conjugate pair of poles with nonzero real and imaginary parts.

Consider

$$Y(s) = \frac{1}{(s+a)^2+b^2}R(s)$$

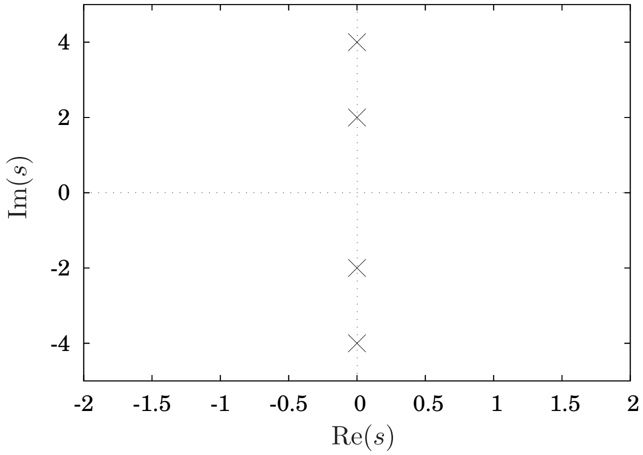


Figure 9.10. Pole locations for $G_1(s) = \frac{2}{s^2+4}$ and $G_2(s) = \frac{8}{s^2+16}$.

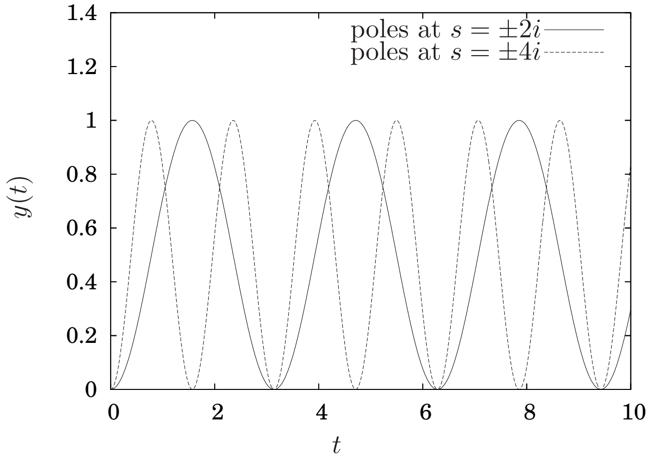


Figure 9.11. Step response for $G_1(s) = \frac{2}{s^2+4}$ and $G_2(s) = \frac{8}{s^2+16}$.

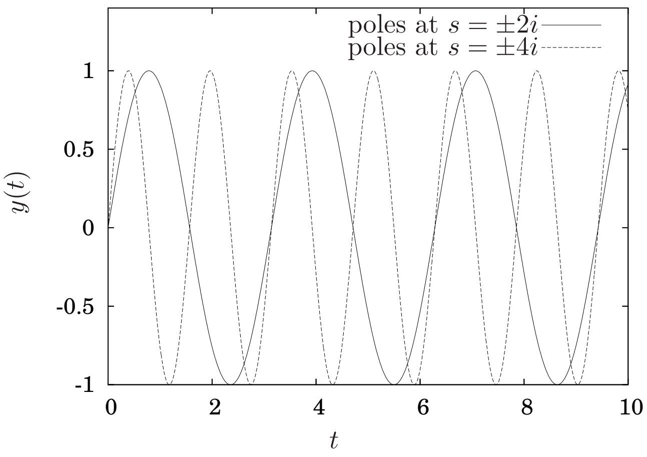


Figure 9.12. Impulse response for $G_1(s) = \frac{2}{s^2+4}$ and $G_2(s) = \frac{4}{s^2+16}$.

which has a complex conjugate pair of poles at $s = -a \pm -b$. The partial fraction expansion will be of the form

$$Y(s) = c_1 \frac{s+a}{(s+a)^2 + b^2} + c_1 c_2 \frac{b}{(s+a)^2 + b^2} + \sum \hat{R}(s)$$

and hence the solution will be of the form

$$y(t) = c_1 e^{-at} \cos bt + c_2 e^{-at} \sin bt + \text{other terms.}$$

This shows the following.

Proposition 9.6.9 *If a transfer function contains a complex conjugate pair of poles, it will have exponentially decaying or growing sinusoidal terms in the solution. Whether or not the terms are decaying or growing depend upon whether the real part of the pair of poles is negative or positive, respectively.*

Example 9.6.10 The poles of

$$\begin{aligned} G_1(s) &= \frac{1}{s^2 + 2s + 5} \\ &= \frac{1}{(s+1)^2 + 4} \end{aligned}$$

and

$$\begin{aligned} G_1(s) &= \frac{1}{s^2 + 2s + 10} \\ &= \frac{1}{(s+1)^2 + 9} \end{aligned}$$

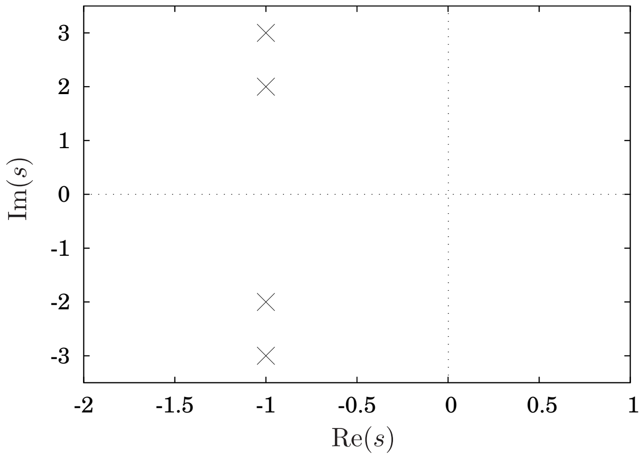


Figure 9.13. Pole locations for $G_1(s) = \frac{1}{(s+1)^2+4}$ and $G_2(s) = \frac{1}{(s+1)^2+9}$.

are plotted in Figure 9.13. The corresponding step responses are plotted in Figure 9.14 and the impulse responses are plotted in Figure 9.15. Because the analysis of the response is a bit more complicated due to the fact that both the real and imaginary parts of the poles need to be considered and also because the response of a second order system of this type is the basis for many control design methods, a complete analysis of a system with complex conjugate poles is discussed subsequently, in Section 9.8. ■

A summary of these results is illustrated in Figure 9.16. Any poles in the right half plane lead to instabilities. Complex conjugate purely imaginary poles contribute sinusoidal solutions. Poles at the origin contribute polynomial solutions in t. Negative real poles contribute to decaying exponential terms and complex conjugate poles with negative real part contribute decaying sinusoidal terms.

Also, because the real part of any pole corresponds exactly to the coefficient of time in an exponential, we may talk about "fast" and "slow" poles. In particular, for poles with negative real part, the farther the pole is to the left, the faster it decays. All poles with positive real part are unstable; however, the larger the magnitude of the real part of the positive pole, faster the instability grows. This is qualitatively summarized in Figure 9.17.

Based upon what we know so far, we can state the following important result.

Proposition 9.6.11 *Given a transfer function, $G(s)$, the impulse and step responses are stable if and only if all the poles of*

$$Y(s) = G(s)R(s)$$

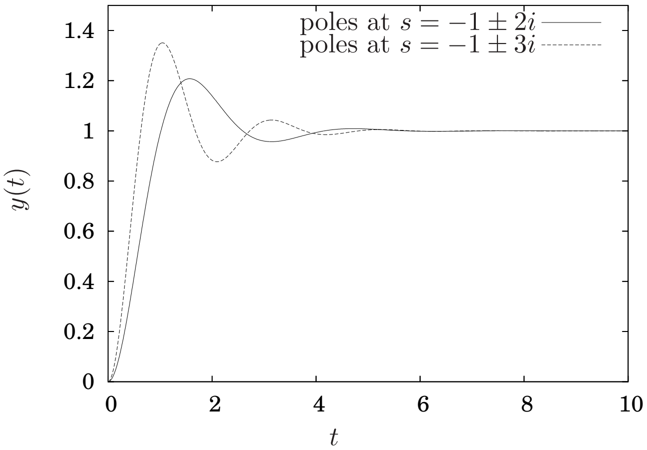


Figure 9.14. Step response for $G_1(s) = \frac{1}{(s+1)^2+4}$ and $G_2(s) = \frac{1}{(s+1)^2+9}$.

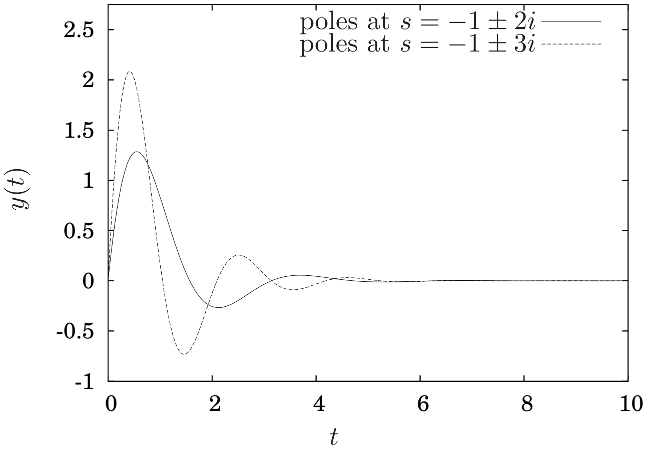


Figure 9.15. Impulse response for $G_1(s) = \frac{1}{(s+1)^2+4}$ and $G_2(s) = \frac{1}{(s+1)^2+9}$.

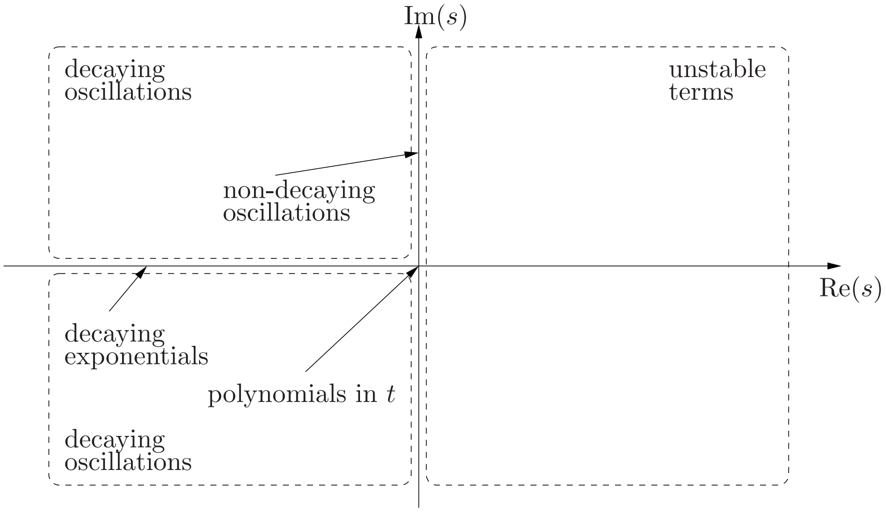


Figure 9.16. Contributions of poles at various locations to the response of a system.

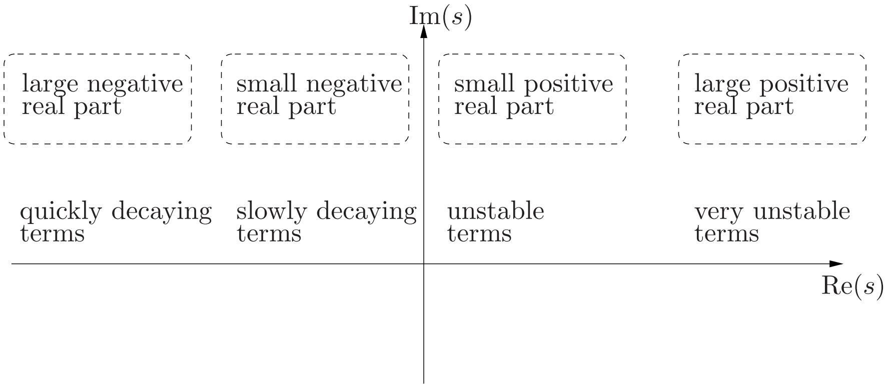


Figure 9.17. The effect of the magnitude of the real part of a pole on the nature of its contribution to the solution.

are in the left half plane where $R(s) = 1$ *or* $R(s) = \frac{1}{s}$ *in the case of the impulse and step responses respectively.*

Before we proceed with a detailed analysis of a prototypical type of response, which is the step response of a second order system, we need an important result very relevant to engineering design of control systems. It is the notion of *dominant poles* in a transfer function. The idea is that if a transfer function has a multiple left poles in the left half plane and none in the right half plane, then the poles far to the left will contribute very little to the solution because they decay so quickly. An example will hopefully elucidate this idea.

Example 9.6.12 Consider the step response of

$$G(s) = \frac{5}{\frac{1}{10}(s+10)\frac{1}{8}(s+8)\left((s+1)^2+4\right)}$$

i.e.,

$$\begin{aligned} y(t) &= \mathcal{L}^{-1}\left(G(s)\frac{1}{s}\right) \\ &= \mathcal{L}^{-1}\left(\frac{5}{\frac{1}{10}(s+10)\frac{1}{8}(s+8)\left((s+1)^2+4\right)}\right). \end{aligned}$$

Before we solve this, we observe that two poles are pretty far to the left and two are relatively close to the imaginary axis as is illustrated in Figure 9.18. Since the effect of the two far to the left should decay rapidly, the solution should be rather close to that if only the complex conjugate poles near the imaginary axis comprised the system, *i.e.,* the step response of

$$G_1(s) = \frac{5}{\left((s+1)^2+4\right)}$$

should be a good approximation to the step response to

$$G(s) = \frac{5}{\frac{1}{10}(s+10)\frac{1}{8}(s+8)\left((s+1)^2+4\right)}$$

Skipping all the gritty details, for

$$Y(s) = G(s)\frac{1}{s}$$

the time domain response is

$$y(t) = 1 + \frac{4}{17}e^{-10t} - \frac{25}{53}e^{-8t} - \frac{688}{901}e^{-t}\cos 2t - \frac{58}{53}e^{-t}\sin 2t$$

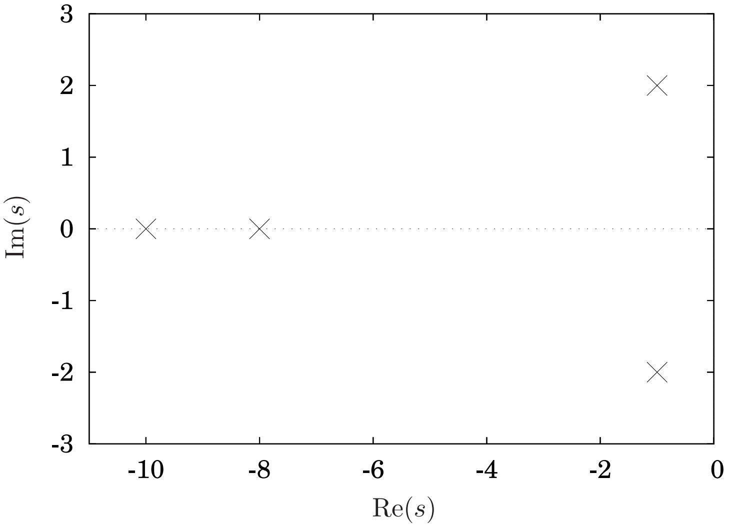


Figure 9.18. Pole locations for transfer function in Example 9.6.12. The two poles near the imaginary axis should dominate the response.

and for

$$Y_1(s) = G_1(s)\frac{1}{s}$$

the time domain response is

$$y_1(t) = 1 - e^{-t}\left(\cos t2 + \frac{1}{2}\sin 2t\right).$$

The two step responses are plotted in Figure 9.19. Clearly, the step response of $G_1(s)$ is a fairly good approximation of the step response of $G(s)$. The reason for this is because the e^{-10t} and e^{-8t} terms decay so rapidly. ∎

9.7 Stability

From Section 9.6 it is clear that if a transfer function has any poles in the right half complex plane then components of the solution will grow unbounded. In this section we will make the notion of stability a bit more precise and present a test to determine whether or not a transfer function has any right half plane poles. For a transfer function the notion of stability is a little more complicated than "the solutions do not blow up" since, the solution depends upon the input to the transfer function as well as possibly any initial conditions. So, the correct notion is that "bounded inputs result in bounded outputs."

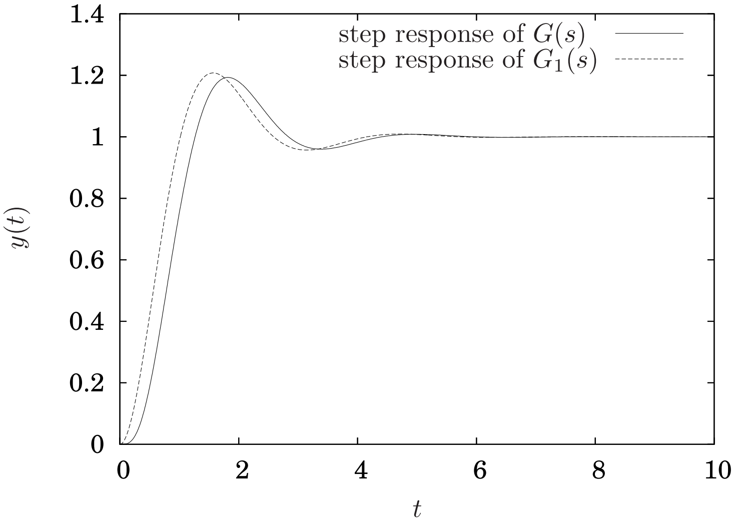


Figure 9.19. Comparison of step responses of the transfer function with four poles, $G(s)$, and the transfer function with only the two dominant poles, $G_1(s)$ from Example 9.6.12.

Definition 9.7.1 A transfer function is *bounded input bounded output stable* ("BIBO stable") if the output is bounded for every input that is bounded. Mathematically, if

$$\frac{Y(s)}{R(s)} = G(s)$$

is the transfer function and $y(t) = \mathcal{L}^{-1}(Y(s))$ and $r(t) = \mathcal{L}^{-1}(R(s))$ are the time domain functions describing the output and input of the transfer function respectively, then the transfer function is BIBO stable if and only if

$$|r(t)| \leq K_r \quad \forall t \qquad \Longrightarrow \qquad |y(t)| \leq K_y \quad \forall t$$

where K_r and K_y are some real constants. $\diamond$

It should be clear from our analysis in Section 9.6 that a necessary condition for stability is that there should be no poles of the transfer function in the right half complex plane. For high order polynomials, which are difficult to factor by hand, it will be useful to have a test which, while it does not give exactly what the poles of the transfer function are, at least provides some information about how many are in the left or right half complex plane.

There are various methods to do this[1], but the one typically covered in undergraduate controls courses is the so-called *Routh criterion.*

[1] The Hurwitz criterion, the Hermite criterion, the Liénard-Chipart criterion and the Kharitonov test, for example

The method is based upon constructing an array and examining the number of sign changes of numbers in the first column of the array. First we will define the array and then we will present the stability test.

Consider an nth order polynomial of the form

$$D(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n.$$

Our interest is the case when this is the denominator of a transfer function. The *Routh array* is constructed as follows.

$$\begin{array}{r|cccccc}
s^n & a_0 & a_2 & a_4 & a_6 & \cdots & 0 \\
s^{n-1} & a_1 & a_3 & a_5 & a_7 & \cdots & 0 \\
s^{n-2} & b_1 & b_2 & b_3 & \cdots & & 0 \\
s^{n-3} & c_1 & c_2 & c_3 & \cdots & 0 & 0 \\
s^{n-4} & d_1 & d_2 & d_3 & \cdots & 0 & 0 \\
\vdots & & & & & & \\
s^0 & e_1 & 0 & 0 & 0 & 0 & 0
\end{array}$$

where

$$\begin{aligned}
b_1 &= -\frac{\begin{vmatrix} a_0 & a_2 \\ a_1 & a_3 \end{vmatrix}}{a_1} \\
b_2 &= -\frac{\begin{vmatrix} a_0 & a_4 \\ a_1 & a_5 \end{vmatrix}}{a_1} \\
b_3 &= -\frac{\begin{vmatrix} a_0 & a_6 \\ a_1 & a_7 \end{vmatrix}}{a_1} \\
&\vdots
\end{aligned}$$

$$\begin{aligned}
c_1 &= -\frac{\begin{vmatrix} a_1 & a_3 \\ b_1 & b_2 \end{vmatrix}}{b_1} \\
c_2 &= -\frac{\begin{vmatrix} a_1 & a_5 \\ b_1 & b_3 \end{vmatrix}}{b_1} \\
c_3 &= -\frac{\begin{vmatrix} a_1 & a_7 \\ b_1 & b_4 \end{vmatrix}}{b_1} \\
&\vdots
\end{aligned}$$

$$\begin{aligned} d_1 &= -\frac{\begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}}{c_1} \\ d_2 &= -\frac{\begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix}}{c_1} \\ &\vdots \end{aligned}$$

Any term that is not defined is zero.

Finally, the point of all of this is the following theorem that will help us determine the stability of a transfer function.

Theorem 9.7.2 *The number of solutions to*

$$a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n = 0$$

that are in the right half plane is equal to the number of sign changes of the coefficients in the first column of the Routh array.

For our purposes, if the polynomial we use is the denominator of a transfer function, the number of sign changes in the first column of the Routh array is equal to the number of right half plane poles.

Before we present any examples, *sufficient condition* for stability will be presented as a corollary following from Routh's criterion.

Corollary 9.7.3 *If the coefficients in*

$$D(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n.$$

are not all of the same sign, then $D(s)$ will have at least one right half plane root.

The proof is left as an exercise.

Example 9.7.4 Determine the number of right half plane poles of the transfer function

$$G(s) = \frac{s+6}{s^4 + 7s^3 + 18s^2 + 22s + 12}.$$

This is, of course, equivalent to determining the number of solutions to

$$s^4 + 7s^3 + 18s^2 + 22s + 12 = 0 \tag{9.15}$$

that have a positive real part.

So, the start of the Routh array is

$$\begin{array}{c|cccc} s^4 & 1 & 18 & 12 & 0 \\ s^3 & 7 & 22 & 0 & 0 \end{array}.$$

Computing the next row,

$$\begin{aligned} b_1 &= \frac{(7)(18)-22}{7} = \frac{104}{7} \\ b_2 &= \frac{(7)(12)-0}{7} = 12 \end{aligned}$$

so the array is

$$\begin{array}{c|cccc} s^4 & 1 & 18 & 12 & 0 \\ s^3 & 7 & 22 & 0 & 0 \\ s^2 & \frac{104}{7} & 12 & 0 & 0 \end{array}.$$

Computing the next row,

$$\begin{aligned} c_1 &= \frac{22\frac{104}{7} - 7 * 12}{\frac{104}{7}} = \frac{425}{26} \\ c_2 &= 0 \end{aligned}$$

so the array is

$$\begin{array}{c|cccc} s^4 & 1 & 18 & 12 & 0 \\ s^3 & 7 & 22 & 0 & 0 \\ s^2 & 104 & 12 & 0 & 0 \\ s^1 & \frac{425}{26} & 0 & 0 & 0 \end{array}.$$

Finally,

$$d_1 = 12,$$

so the complete array is

$$\begin{array}{c|cccc} s^4 & 1 & 18 & 12 & 0 \\ s^3 & 7 & 22 & 0 & 0 \\ s^2 & 104 & 12 & 0 & 0 \\ s^1 & \frac{425}{26} & 0 & 0 & 0 \\ s^0 & 12 & & & \end{array}.$$

Since there are no sign changes in the first row, all the solutions to Equation 9.15 are in the right half plane. ■

There is a minor complication when a zero appears in the first column of the Routh array. In such a case simply replace the zero with a small positive variable, $0 < \epsilon \ll 1$ and proceed with the subsequent computations and analysis as usual.

Example 9.7.5 Determine the number of solutions to

$$s^4 + 4s^3 + s^2 + 4s + 5 = 0$$

with positive real part.

Constructing the Routh array, the first two rows are

$$\begin{array}{c|cccc} s^4 & 1 & 1 & 5 & 0 \\ s^3 & 4 & 4 & 0 & 0 \end{array}.$$

The first term in the third row will be

$$b_1 = -\frac{\begin{vmatrix} 1 & 1 \\ 4 & 4 \end{vmatrix}}{4} = 0.$$

Replacing this with $0 < \epsilon \ll 1$ and substituting for

$$b_2 = -\frac{\begin{vmatrix} 1 & 5 \\ 4 & 0 \end{vmatrix}}{4} = 5$$

gives

$$\begin{array}{c|cccc} s^4 & 1 & 1 & 5 & 0 \\ s^3 & 4 & 4 & 0 & 0 \\ s^2 & \epsilon & 5 & 0 & 0 \end{array}.$$

Proceeding to the next row gives

$$\begin{aligned} c_1 &= -\frac{\begin{vmatrix} 4 & 4 \\ \epsilon & 5 \end{vmatrix}}{\epsilon} \\ &= \frac{4\epsilon - 20}{\epsilon} \end{aligned}$$

so the array is

$$\begin{array}{c|cccc} s^4 & 1 & 1 & 5 & 0 \\ s^3 & 4 & 4 & 0 & 0 \\ s^2 & \epsilon & 5 & 0 & 0 \\ s^1 & \frac{4\epsilon-20}{\epsilon} & 0 & 0 & 0 \end{array}.$$

Finally, computing the last row gives

$$\begin{array}{c|cccc} s^4 & 1 & 1 & 5 & 0 \\ s^3 & 4 & 4 & 0 & 0 \\ s^2 & \epsilon & 5 & 0 & 0 \\ s^1 & \frac{4\epsilon-20}{\epsilon} & 0 & 0 & 0 \\ s^0 & 5 & 0 & 0 & 0 \end{array}.$$

■

Since ϵ is *small and positive* $4\epsilon - 20$ will be negative. Hence there are two sign changes in the first column and therefore there are two roots with positive real part.

This may be confirmed numerically where the poles are determined to be

$$\begin{aligned} p_1 &= -3.92219 \\ p_2 &= 0.40832 + 1.12184i \\ p_3 &= 0.40832 - 1.12184i \\ p_4 &= -0.89444. \end{aligned}$$

We can use this to accomplish a bit more than simply determine whether or not poles are in the right half plane. For example, we may determine ranges of parameter values for which a transfer function is stable.

Example 9.7.6 Determine the values of k for which

$$G(s) = \frac{k\frac{s+2}{s^2-2s+2}}{1+k\frac{s+2}{s^2-2s+2}}$$

is stable.

Simplifying the denominator gives

$$D(s) = s^2 + (k-2)\,s + (2+2k)\,.$$

Constructing the Routh array gives

$$\begin{array}{c|ccc} s^2 & 1 & 2+2k & 0 \\ s^1 & k-2 & 0 & 0 \end{array}\,.$$

Computing the last row gives the complete array

$$\begin{array}{c|ccc} s^2 & 1 & 2+2k & 0 \\ s^1 & k-2 & 0 & 0 \\ s^0 & 2k+2 & 0 & 0 \end{array}\,.$$

In order for there to be no sign change from the s^2 row to the s^1 row, we need that $k > 2$. In order for the first element s^0 to be greater than zero we need $k > -1$. In order to satisfy both, we need $k > 2$. ■

A final example will illustrate the obvious fact that there will not necessarily be *any* values for k to make some transfer functions stable.

Example 9.7.7 Determine the values of k for which

$$G(s) = \frac{k\frac{s-2}{s^2-2s+2}}{1+k\frac{s-2}{s^2-2s+2}}$$

is stable.

In this case the characteristic polynomial is

$$D(s) = s^2 + (k-2)\,s + (2-2k)\,,$$

and the Routh array is

$$\begin{array}{c|ccc} s^2 & 1 & 2-2k & 0 \\ s^1 & k-2 & 0 & 0 \\ s^0 & 2-2k & 0 & 0 \end{array}\,.$$

In order for the first element in the s^1 row to be positive we need that $k > 2$. In order for the first element in the s^0 row to be positive we need $k < 1$. There are no values of k that can satisfy both. ■

9.8 Response of a Second Order System

Recall the canonical form for a generic second order system from Section 3.3

$$m\ddot{x} + b\dot{x} + kx = f(t) \qquad \Longleftrightarrow \qquad \ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n x = \frac{f(t)}{m}$$

where

$$\begin{aligned} \omega_n &= \sqrt{\frac{k}{m}} \\ \zeta &= \frac{b}{2\sqrt{mk}}. \end{aligned}$$

Taking the Laplace transform with zero initial conditions gives

$$\begin{aligned} X(s) &= \frac{1}{s^2 + 2\zeta\omega_n s + \omega_n^2}\frac{F(s)}{m} \\ &= G(s)R(s). \end{aligned}$$

If $0 \leq \zeta < 1$, the poles of $G(s)$ are

$$\begin{aligned} s &= -\zeta\omega_n \pm i\omega_n\sqrt{1-\zeta^2} \\ &= \omega_n\left(-\zeta \pm i\sqrt{1-\zeta^2}\right) && (9.16) \\ &= -\zeta\omega_n \pm i\omega_d, && (9.17) \end{aligned}$$

where

$$\omega_d = \omega_n\sqrt{1-\zeta^2}.$$

Using the notation from Table 8.1, if the denominator of the transfer function contains a term of the form $\left((s+a)^2 + b^2\right)$, then the poles are located at

$$\begin{aligned} s &= -a \pm ib \\ &= \omega_n\left(-\zeta \pm i\sqrt{1-\zeta^2}\right). \end{aligned}$$

As is illustrated in Figure 9.20, the relationship between the pole location and the parameters in the canonical second order system are as follows.

1. The length of the vector from the origin to the pole is ω_n.

2. If the angle from the imaginary axis to the vector from the origin to the pole is denoted by θ, then
$$\zeta = \sin\theta.$$

3. The damped natural frequency, ω_d is the imaginary component of the pole,
$$\begin{aligned} \omega_d &= b \\ &= \omega_n\sqrt{1-\zeta^2}. \end{aligned}$$

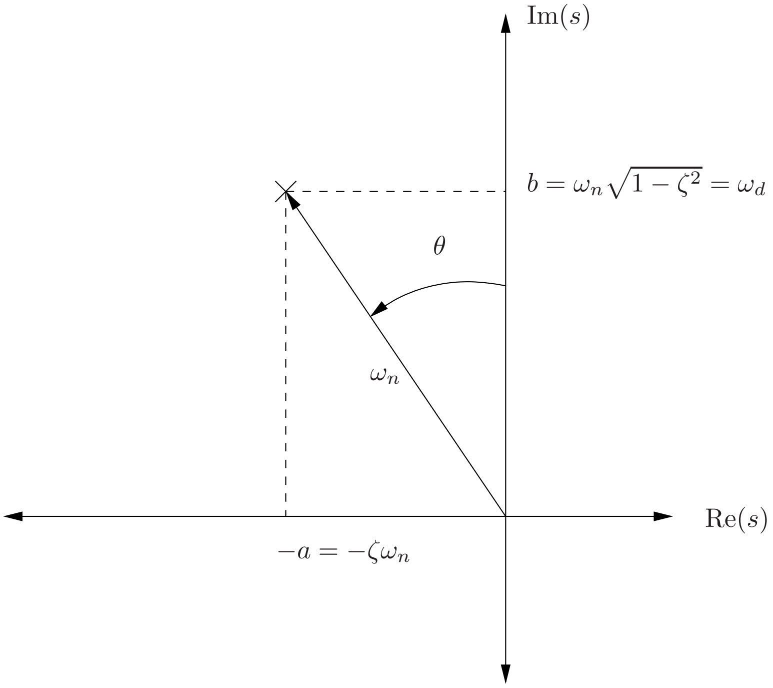


Figure 9.20. Relationship between the location of complex conjugate poles and ω_n, ζ and ω_d.

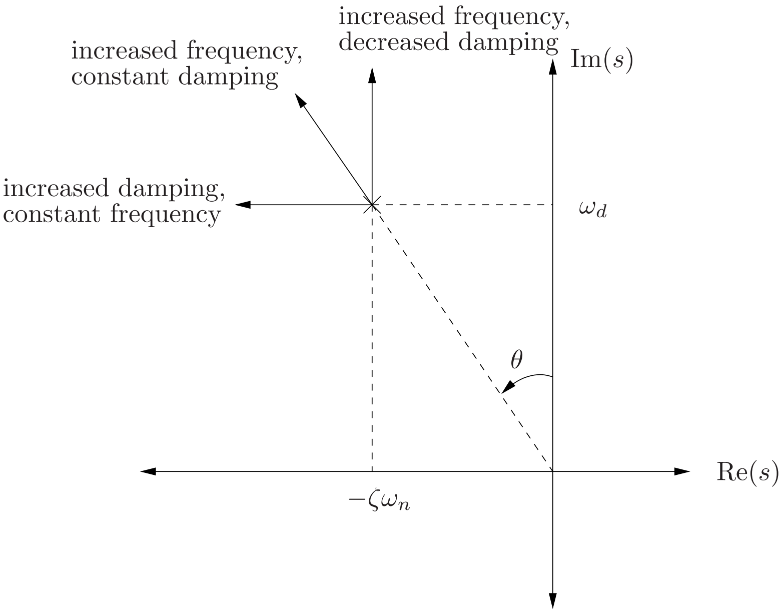


Figure 9.21. Effect of moving the location of a complex conjugate pole.

From the discussion in the previous section regarding the response when a transfer function contains a complex conjugate pole, we can deduce that the solution will have terms of the form $e^{-\zeta\omega_n t}\sin\omega_d t$ and $e^{-\zeta\omega_n t}\cos\omega_d t$. Hence, if the effects of moving the location of a complex conjugate pole with negative real part are as follows.

1. If the imaginary part of the pole is increased and the real part is held constant, then the frequency of the response will increase and the damping ratio will decrease.

2. If the real part of the pole is decreased and the imaginary part is held constant, then the damping ratio is increased and the frequency of the response will be constant.

3. If the angle between the imaginary axis and the vector from the origin to the pole is held constant and the the magnitude of the vector is increased, then the damping remains constant and the frequency of the response increases.

All three of these cases are illustrated in Figure 9.21.

9.8.1 Second order system step response

Now let us relate the location of the poles for a second order system to the time domain specifications defined in Section 9.4 for a unit step input.

Consider

$$G(s) = \frac{k}{s^2 + 2\zeta\omega_n s + \omega_n^2}.$$

For a step input to this transfer function, we will have

$$\begin{aligned}
Y(s) &= \frac{k}{s^2 + 2\zeta\omega_n s + \omega_n^2}\frac{1}{s} \\
&= \frac{k}{\omega_n^2}\left[-\frac{s + 2\zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} + \frac{1}{s}\right] \\
&= \frac{k}{\omega_n^2}\left[-\frac{s + 2\zeta\omega_n}{\left(s + \zeta\omega_n\right)^2 + \omega_d^2} + \frac{1}{s}\right] \\
&= \frac{k}{\omega_n^2}\left[-\frac{s + \zeta\omega_n}{\left(s + \zeta\omega_n\right)^2 + \omega_d^2} - \frac{\zeta\omega_n}{\omega_d}\frac{\omega_d}{\left(s + \zeta\omega_n\right)^2 + \omega_d^2} + \frac{1}{s}\right] \\
&= \frac{k}{\omega_n^2}\left[-\frac{s + \zeta\omega_n}{\left(s + \zeta\omega_n\right)^2 + \omega_d^2} - \frac{\zeta}{\sqrt{1-\zeta^2}}\frac{\omega_d}{\left(s + \zeta\omega_n\right)^2 + \omega_d^2} + \frac{1}{s}\right]
\end{aligned}$$

so

$$y(t) = \frac{k}{\omega_n^2}\left[-e^{-\zeta\omega_n t}\left(\cos\omega_d t + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\omega_d t\right) + 1\right]. \tag{9.18}$$

Plots of the step response for $k = \omega_n^2$ for various ζ and various ω_n are illustrated in Figure 9.22 and 9.23.

Steady state error

Since,

$$\lim_{t\to\infty} y(t) = \frac{k}{\omega_n^2}$$

the steady state error to a unit step input will be zero if $k = \omega_n^2$.

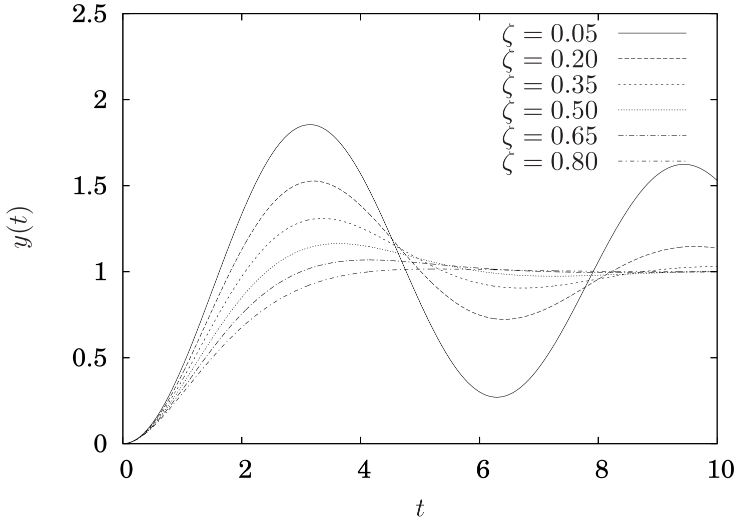


Figure 9.22. Step response of second order system with $\omega_n = 1$ and for various ζ.

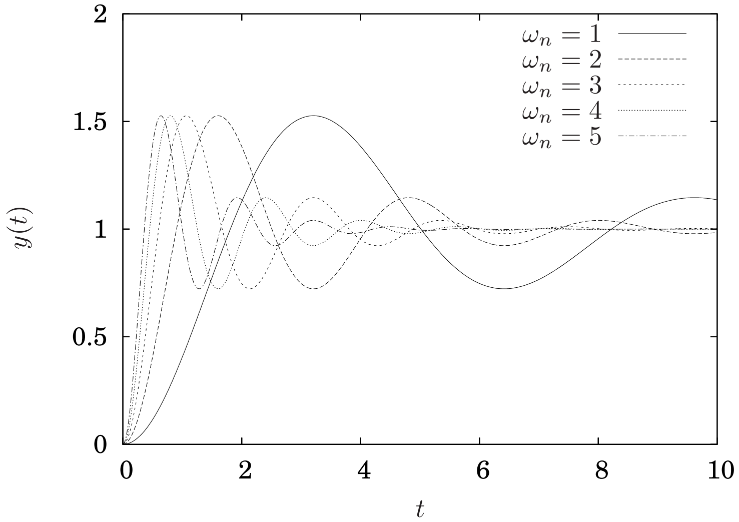


Figure 9.23. Step response of second order system with $\zeta = 0.2$ and for various ω_n.

Peak time

The peak time is determined by finding the time when the derivative of Equation 9.18 is zero. Hence

$$\begin{aligned}
\frac{d}{dt}y(t) &= \zeta\omega_n e^{-\zeta\omega_n t}\left(\cos\omega_d t + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\omega_d t\right) - \\
&\quad e^{-\zeta\omega_n t}\left(-\sin\omega_d t + \frac{\zeta\omega_d}{\sqrt{1-\zeta^2}}\cos\omega_d t\right) \\
&= \zeta\omega_n e^{-\zeta\omega_n t}\left(\cos\omega_d t + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\omega_d t\right) - \\
&\quad e^{-\zeta\omega_n t}\left(-\sin\omega_d t + \zeta\omega_n\cos\omega_d t\right) \\
&= \left(\frac{\zeta^2\omega_n}{\sqrt{1-\zeta^2}} + 1\right)\sin\omega_d t \\
&= 0.
\end{aligned}$$

The first positive time for which this is zero is the peak time and is

$$t_p = \frac{\pi}{\omega_d}.$$

Overshoot

The overshoot is determined by substituting the peak time into Equation 9.18:

$$\begin{aligned}
x_p &= y\left(\frac{\pi}{\omega_d}\right) \\
&= \frac{k}{\omega_n^2}\left[-e^{-\frac{\zeta\omega_n\pi}{\omega_d}}\left(\cos\pi + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\pi\right) + 1\right] \\
&= \frac{k}{\omega_n^2}\left(1 + e^{-\frac{\zeta\omega_n\pi}{\omega_d}}\right).
\end{aligned}$$

Hence, the *percentage* overshoot is given by the exponential term. Substituting for the definition of ω_d gives

$$O = e^{-\frac{\pi\zeta}{\sqrt{1-\zeta^2}}}. \tag{9.19}$$

Observe that the percentage overshoot depends upon the damping ratio only. A plot of O it versus ζ is given in Figure 9.24. Since the maximum overshoot is a function of only the damping ratio, there is a simple geometric interpretation for second order poles that will meet an overshoot specification, as is illustrated by the following example.

Example 9.8.1 Determine the region in the complex plane where the poles should be located in order for a second order system to have a maximum

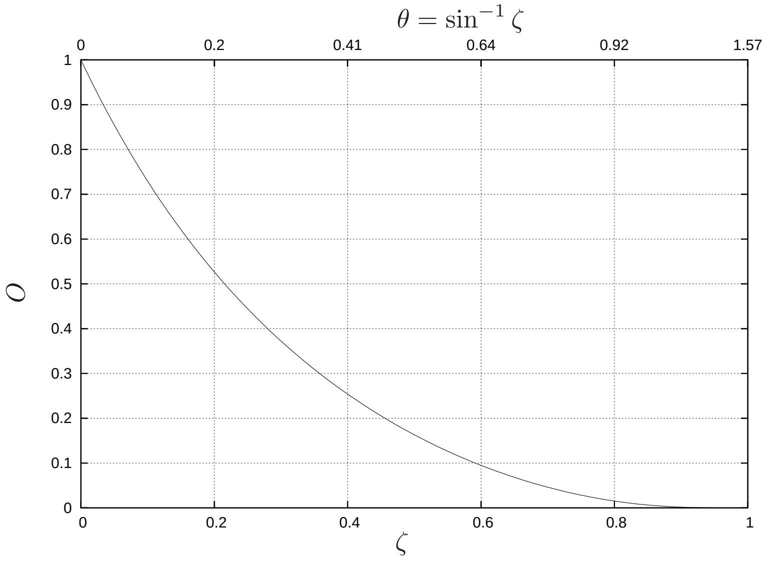


Figure 9.24. Percentage overshoot, O *versus* damping ratio, ζ and angle between imaginary axis and the pole, θ.

overshoot of less than 10%. Either referring to Figure 9.24 or solving Equation 9.19 gives

$$O < 0.1 \qquad \Longleftrightarrow \qquad \zeta > 0.6.$$

So, the region in the complex plane where a pair of second order poles must be located to satisfy this specification are illustrated in Figure 9.25. ■

Settling time

To determine the settling time, note that the rate at which the transient response decays is governed by the exponential term, $e^{-\zeta\omega_n t}$. Hence, for the 5% settling time,

$$0.05 = e^{-\zeta\omega_n t_s} \qquad \Longleftrightarrow \qquad t_s = -\frac{\ln(0.05)}{\zeta\omega_n} \approx \frac{3}{\zeta\omega_n}.$$

Since $\zeta\omega_n$ is the real component of the pole, the settling time is given by the distance from the imaginary axis. Similarly, if we were interested in the x% settling time, it would be given by

$$t_s = -\frac{\ln(\frac{x}{100})}{\zeta\omega_n}.$$

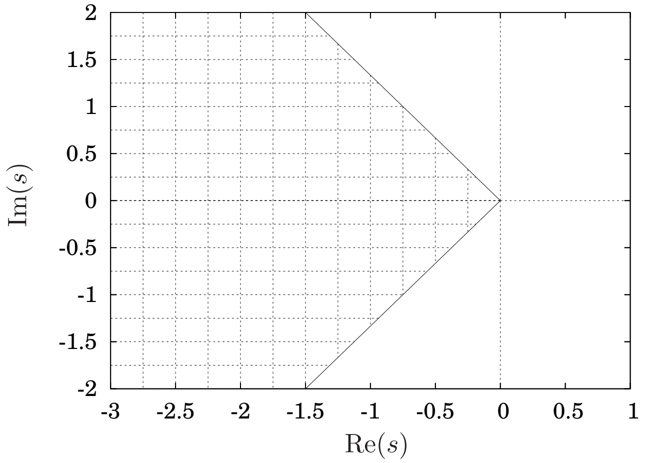


Figure 9.25. Hatched region corresponds to pole locations for a second order system with less than 10% overshoot.

Example 9.8.2 Determine the region in the complex plane where the poles should be located in order for a second order system to have a 2% settling time of less than 3 seconds. Thus

$$-\frac{\ln 0.02}{\zeta\omega_n} < 3$$

which gives

$$\zeta\omega_n > 1.3.$$

Since $\zeta\omega_n$ is the real component of the pole, the region in the complex plane where the 2% settling time is less than 3 seconds is illustrated in Figure 9.26. ■

Rise time

The solution will first equal its steady state value when

$$e^{-\zeta\omega_n t_r}\left(\cos\omega_d t_r + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\omega_d t_r\right) = 0,$$

which requires

$$\cos\omega_d t_r = -\frac{\zeta}{\sqrt{1-\zeta^2}}\sin\omega_d t_r \qquad \Longleftrightarrow \qquad \tan\omega_d t_r = -\frac{\sqrt{1-\zeta^2}}{\zeta}$$

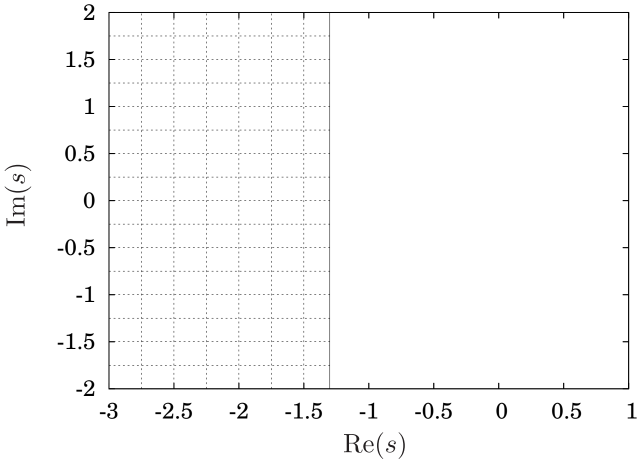


Figure 9.26. Hatched region corresponds to pole locations for a second order system with a settling time of less than 3 seconds.

or, solving for t_r

$$t_r = \frac{1}{\omega_d}\tan^{-1}\left(-\frac{\sqrt{1-\zeta^2}}{\zeta}\right).$$

This has an infinite number of solutions, but t_r will be the smallest positive time that satisfies this equation.

Given a specified rise time, t_r, then we require that

$$\omega_d \geq \frac{1}{t_r}\tan^{-1}\left(-\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \tag{9.20}$$

in order for the system response to be equal to or faster than the specification. This relationship is not as simple as the preceding ones for the overshoot and settling time since it does not reduce to a line in the complex plane. Since we are pursuing a geometric interpretation, a more useful relationship would be the one between the damped natural frequency, ω_d, and the angle from the imaginary axis to the pole, θ.

First observe that if $\zeta \ll 1$

$$\tan^{-1}\left(-\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \approx \frac{\pi}{2},$$

since $\frac{\pi}{2}$ is the first positive value for which satisfies $\tan^{-1} 0$, so we have

$$t_r \approx \frac{\pi}{2\omega_d}.$$

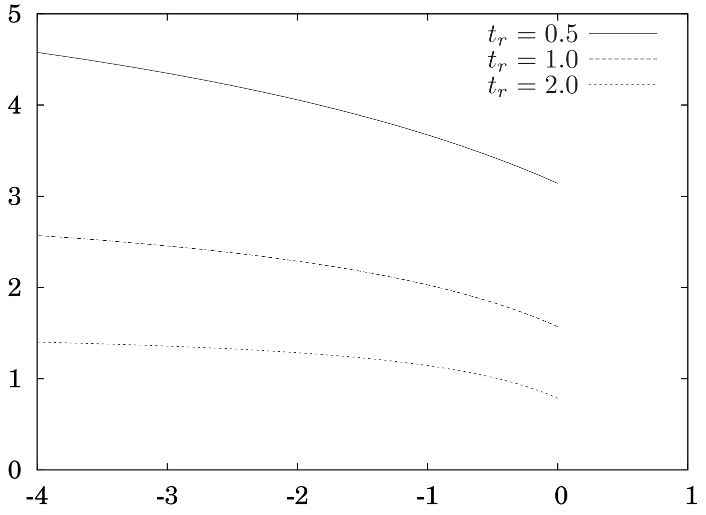


Figure 9.27. Lines in the complex plane of pole locations corresponding to constant rise times.

Conversely, if $\zeta \approx 1$, then

$$\tan^{-1}\left(-\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \approx \pi,$$

so

$$t_r \approx \frac{\pi}{\omega_d}.$$

So, we may use these as an approximation for small ζ and for $\zeta \approx 1$.

For intermediate values, since $\zeta = \sin\theta$ and keeping in mind we need to consider the first positive value, substituting into Equation 9.20 gives

$$\begin{aligned} \omega_d &\geq \frac{1}{t_r}\tan^{-1}\left(\frac{\sqrt{1-\sin^2\theta}}{\sin\theta}\right) \\ &\geq \frac{1}{t_r}\left(\theta + \frac{\pi}{2}\right) \end{aligned} \tag{9.21}$$

where the additive $\frac{\pi}{2}$ term is is needed to make the value the first positive solution for the inverse tangent function. A plot of ω_d which satisfy the equality in Equation 9.21 for $t_r = 1, 2$ and 3 is illustrated in Figure 9.27. Regions in the complex plane which satisfy the inequality will be above the curves.

Example 9.8.3 The region in the complex plane where the poles should be located in order for a second order system to have a rise time of less than 2 seconds would be above the bottom curve in Figure 9.27. ■

Remark 9.8.4 Different approximations for the rise time appear in different texts. For example [6] gives

$$t_r \approx \frac{1.8}{\omega_n},$$

and [13] gives

$$t_r \approx \frac{0.8 + 2.5\zeta}{\omega_n} \qquad \text{or} \qquad t_r \approx \frac{1 - 0.4167\zeta + 2.917\zeta^2}{\omega_n}.$$

One reason for the different formulae is because of different definitions of the rise time. For example, in [13] it is defined as the time to go from 10% to 90% of the final value. Another reason for the differences is the fact that rise time is a sort of fickle quantity, as is illustrated in Figure 9.28 and 9.29. In those figures, the poles were located along the bottom curve in Figure 9.27, corresponding to $t_r = 2$.

However, observe that if, instead of the quantity of interest being when the system first achieved the steady state value, it was instead 90% of the steady state value. Referring to Figure 9.29, it is clear that there is a drastic difference between the times at which each curve passes through $y(t) = 0.9$.

The reader is cautioned to either be careful about the definition of the rise time, and choose an appropriate one for a given situation, or to accept the somewhat gross approximate nature of the concept. It is necessary component for specifying control system performance, however. If all that were considered were the settling time and the overshoot, these specifications would be easily satisfied by a system with very slow performance. If the application requires a reasonable response time, then the rise time, or something similar, must be considered. ⋄

We can now combine these specifications to determine pole locations that satisfy more than one specification.

Example 9.8.5 To determine the region in the complex plane where the poles should be located in order for a second order system to have a 2% settling time of less than 3 seconds and a maximum overshoot less than 10% we can take the intersection of the regions in Figures 9.25 and 9.26. This region is illustrated in Figure 9.30. ■

Example 9.8.6 Returning to Example 9.8.5, assume that, in addition to the overshoot and settling time specifications, we require a rise time of less than 3 seconds. We could copy the 1 second curve from Figure 9.27. Alternatively, we could adopt a more conservative engineering approximation and use the approximation for a damping ratio near one. Hence, we will use

$$\omega_d \geq \frac{\pi}{t_r},$$

or, for this case,

$$\omega_d \geq \frac{\pi}{3}.$$

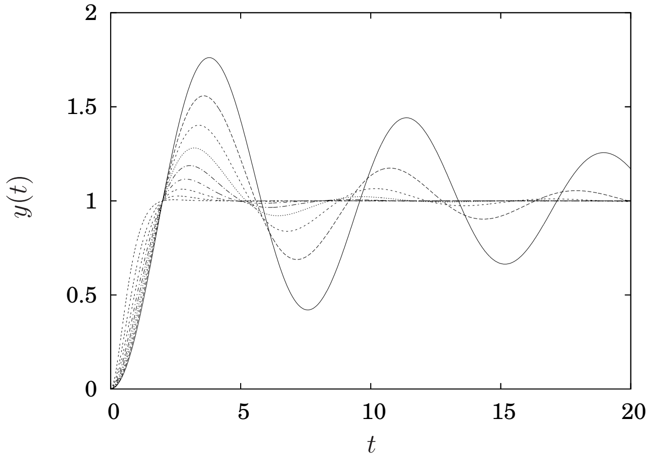


Figure 9.28. Step responses corresponding to $t_r = 2$.

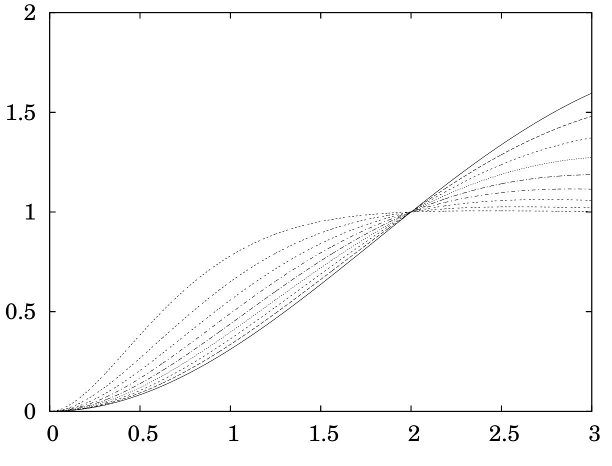


Figure 9.29. Step responses corresponding to $t_r = 2$.

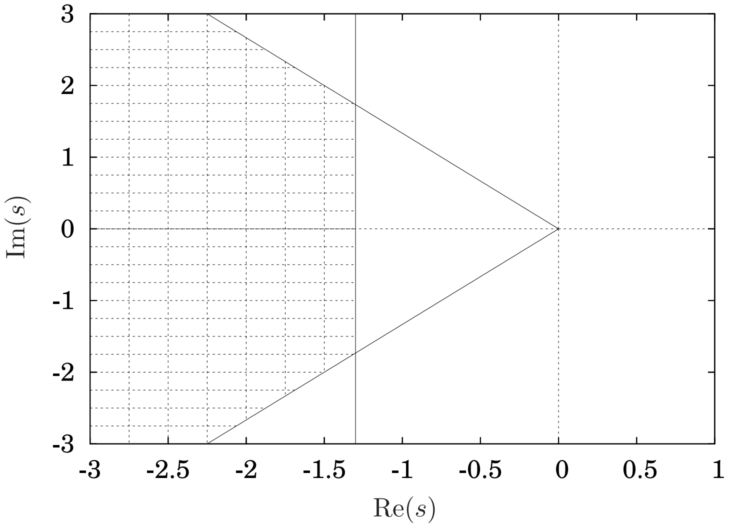


Figure 9.30. Hatched region corresponds to second order pole locations with a settling time less than 3 seconds and less than 10% overshoot.

The region satisfying all three specifications is illustrated in Figure 9.31.

As a check, the step response for a unit step input is illustrated in Figure 9.32, and it is apparent that all three specifications are satisfied. ■

9.8.2 Additional Poles and Zeros

At this point we have developed practically every possible way to consider the response of a second order system and it is completely characterized in terms of the time domain specifications of the step response, *etc.* Unfortunately, the world is not composed entirely of second order systems, so it will be useful to relate, when possible, the response of a system that is not second order to the second order response we know so well. We will consider several ways in which a system may deviate from a canonical second order system.

Additional poles far to the left

We considered this previously in Example 9.6.12). If there are poles relatively far to the left, their effect will decay very fast compared to poles near the imaginary axis and the system response will be dominated by the poles near the imaginary axis.

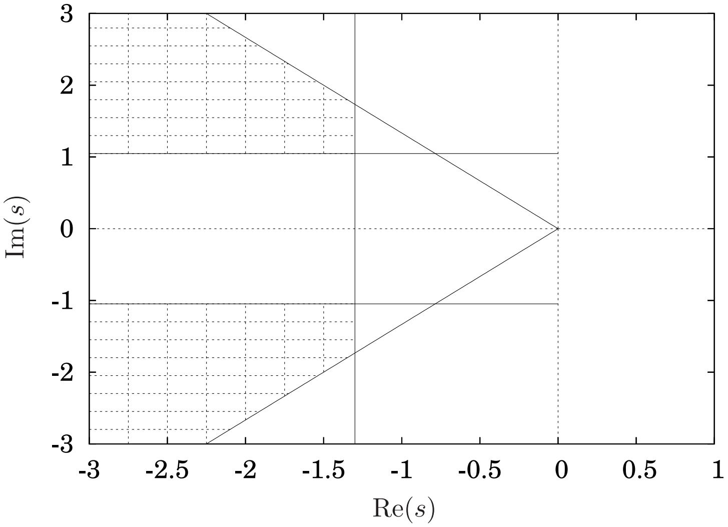


Figure 9.31. Hatched region corresponds to second order pole locations with a settling time less than 3 seconds, less than 10% overshoot and less than a 3 second rise time.

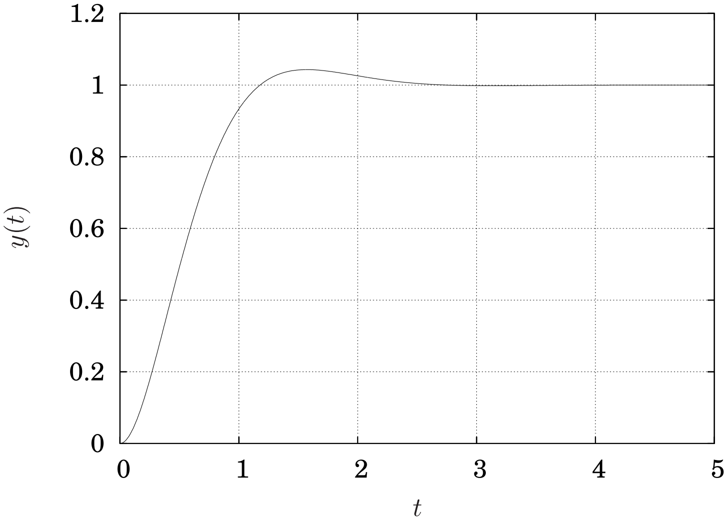


Figure 9.32. Step response of $G(s) = \frac{8}{(s+2)^2+2^2}$, which has poles within the region satisfying all three specifications in Example 9.8.6.

An additional real zero

If a transfer function has a complex conjugate pair of poles and one real zero, the effect of the zero on the response will depend on the location of the zero. This section will draw some conclusions based upon inference from an example. The analytical proof of the conclusions is left to Exercise 9.6.

Example 9.8.7 Consider

$$G(s) = \frac{5}{s^2 + 2s + 5}$$

and

$$G_1(s) = \frac{\frac{5}{r}\,(s + r)}{s^2 + 2s + 5}$$

where $r = 10, 1, -1$ and $= 10$, corresponding to the zero being far to the left, in the left half plane but near the imaginary axis, in the right half plane and near the imaginary axis and far to the right. The step responses are illustrated in Figure 9.33. ■

From this example we may infer the following general rules:

- if the zero is far from the imaginary axis, then it has little effect on the step response;
- if the zero is in the left half plane and close to the imaginary axis, it will decrease the rise time and increase the overshoot; and,
- if the zero is in the right half plane and close to the imaginary axis, it will increase the rise time, perhaps increase the overshoot and perhaps the system will initially move in the "wrong direction."

An additional real pole

Consider

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

and

$$G_1(s) = \frac{\omega_n^2}{\left(s^2 + 2\zeta\omega_n s + \omega_n^2\right)\frac{1}{r}\,(s + r)}.$$

From Equation 9.18, the partial fraction expansion is

$$Y(s) = -\frac{s + 2\zeta\omega_n}{\left(s + \zeta\omega_n\right)^2 + \omega_d^2} + \frac{1}{s} \tag{9.22}$$

and the step response is

$$y(t) = -e^{-\zeta\omega_n t}\left(\cos\omega_d t + \frac{\zeta}{\sqrt{1 - \zeta^2}}\sin\omega_d t\right) + 1.$$

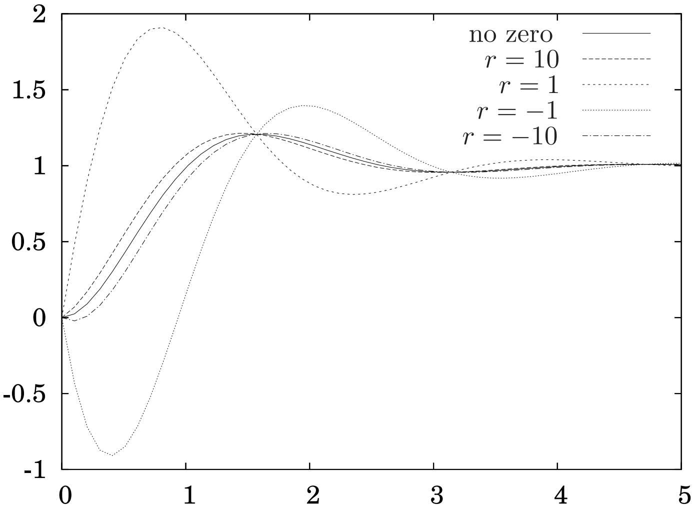


Figure 9.33. The effect of an additional real zero on the second order step response: the step response of $G_1(s) = \frac{\frac{5}{r}(s+r)}{s^2+2s+5}$ for various values of r.

Computing the partial fraction expansion for a step response gives

$$\begin{aligned} Y(s) &= \frac{\omega_n^2 r}{(s^2+2\zeta\omega_n s+\omega_n^2)(s+r)}\frac{1}{s} \\ &= -\left(\frac{\omega_n^2}{r^2-2\omega_n\zeta r+\omega_n^2}\right)\frac{1}{s+r}+\frac{1}{s} \\ &+ \left(\frac{-r^2(s+2\omega_n\zeta)-r\left(\omega_n^2-4\omega_n^2\zeta^2-2s\omega_n\zeta\right)}{(r^2-2\omega_n\zeta r+\omega_n^2)}\right)\frac{1}{(s^2+2\omega_n\zeta s+\omega_n^2)} \end{aligned} \tag{9.23}$$

As $r \to \infty$ we expect this to approach Equation 9.22, which it does. Because of the r^2 in the denominator of the first term, it approaches zero, and in the last term the r^2 terms in the numerator and denominator would dominate, giving the same second order term as in Equation 9.22. At least for one additional pole, this verifies our intuition that poles far to the left will have little effect on the response.

If r is positive and small, which corresponds to a pole close to the imaginary axis in the left half plane, then the exponential term will dominate the solution. Mathematically

$$\lim_{r\downarrow 0} Y(s) = -\frac{1}{s+r}+\frac{1}{s},$$

so for small r,

$$y(t) \approx 1 - e^{-rt},$$

which has no overshoot and infinite rise time. Since r is small, the exponential terms decays very slowly and hence $y(t)$ approaches the steady state value very slowly.

Conceptually interpolating between these to extremes we can conclude that if a second order system has an additional pole in the left half plane then

- if it is far to the left, it will have little effect on the response;
- if it is very close to the imaginary axis compared to the second order poles, it will dominate the response and the solution will slowly asymptotically approach the steady state value;
- if it is of the same order as the second order poles it should decrease both the rise time and overshoot; and,
- if the pole is anywhere in the right half plane, then the solution will be unstable.

Example 9.8.8 Consider

$$G(s) = \frac{5}{s^2 + 2s + 5}$$

and

$$G_1(s) = \frac{5r}{(s+r)\,(s^2 + 2s + 5)}$$

where $r = 10, 1, 0.1,$ and -1, corresponding to the pole being far to the left, in the left half plane of the same order of magnitude as the complex conjugate pair of poles, very near the imaginary axis and in the right half plane. The step responses are illustrated in Figure 9.34. ■

Poles and zeros close together

If a pole and zero are close together, algebraically they nearly cancel. It is natural to expect, then, that their effect in the solution would nearly cancel as well. In fact, this is true which we will demonstrate with one zero and one pole located near each other.

Consider

$$G(s) = \frac{\omega_n^2}{(s^2 + 2\zeta\omega_n s + \omega_n^2)} \frac{s + \hat{r}}{s + r}.$$

A partial fraction expansion gives

$$G(s) = \frac{c_1 s + c_2}{s^2 + 2\zeta\omega_n s + \omega_n^2} + \frac{c_3}{s + r}$$

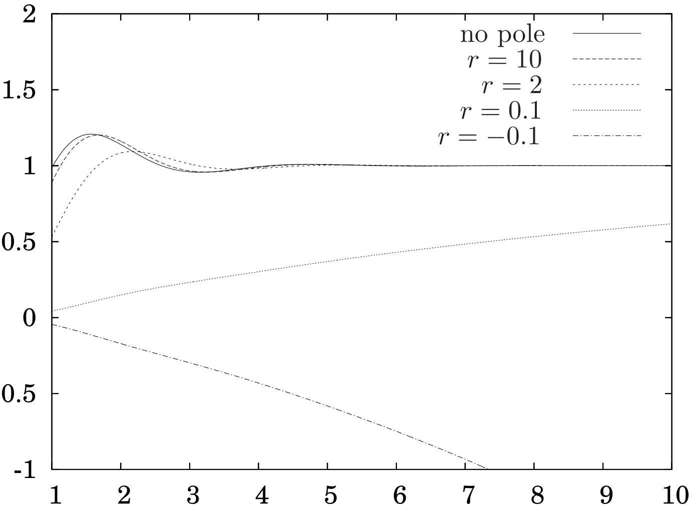


Figure 9.34. The effect of an additional real pole on the second order step response: the step response of $G_1(s) = \frac{5r}{(s+r)(s^2+2s+5)}$ for various values of r.

where

$$
\begin{aligned}
c_1 &= \frac{\omega_n^2\left(1-\frac{\hat{r}}{r}\right)}{r - 2\zeta\omega_n + \frac{\omega_n^2}{r}} \\
c_2 &= \frac{\omega_n^2}{r}\left[\hat{r} + \frac{\omega_n^2}{r}\left(\frac{1-\frac{\hat{r}}{r}}{r-2\zeta\omega_n+\frac{\omega_n^2}{r}}\right)\right] \\
c_3 &= -\frac{\omega_n^2\left(1-\frac{\hat{r}}{r}\right)}{r - 2\zeta\omega_n + \frac{\omega_n^2}{r}}.
\end{aligned}
$$

If the pole and zero are located at the same point, then $r = \hat{r}$ and $c_1 = c_3 = 0$ and $c_2 = \omega_n^2$, as we would expect.

Furthermore, if the pole and zero are close together, then $r \approx \hat{r}$, so $c_1, c_3 \ll 1$ and $c_2 \approx \omega_n^2$. The effect of the magnitude of the coefficients will depend upon whether the pole and zero are in the left or right half plane. If they are in the left half plane, then the coefficient of the exponential term will be small, so the solution will be approximately the same as the second order system. If it is in the right half plane, even though the coefficient is small, the exponential term will grow unbounded and the system will be unstable.

Example 9.8.9

$$G(s) = \frac{5}{s^2+2s+5},$$

$$G_1(s) = \frac{5}{s^2+2s+5}\frac{s+1}{\frac{1}{.95}\left(s+0.95\right)},$$

$$G_2(s) = \frac{5}{s^2+2s+5}\frac{s+1}{\frac{1}{2}\left(s+2\right)},$$

and

$$G_3(s) = \frac{5}{s^2+2s+5}\frac{s-1}{\frac{1}{.95}\left(s-0.95\right)}.$$

The first transfer function, $G(s)$ only has a complex conjugate pair of poles. The second, $G_1(s)$ has an additional pole at $s = -.95$ and an additional zero at $s = -1$. The third has a pole and zero that are not close together and finally, the fourth, $C_3(s)$ has a pole and zero that are close, but in the right half plane.

The step responses are illustrated in Figure 9.35. Observe that if the pole and zero are in the left half plane and are close together, then they almost cancel and the step response is much like that of $G(s)$. If they are not close together then they have a substantial effect on the response. If they are in the right half plane, then even if they are close together the system is unstable. ■

Remark 9.8.10 It is true that mathematically if there is a pole and zero in the right half plane that *exactly* cancel, then they will have no effect on the response of the system. However, for a real engineering system, if there is a pole in the right half plane attempting to cancel it with a zero will not work since it is impossible to characterize any real system exactly. ⋄

9.9 The Root Locus Design Method

The root locus design method is probably the most fundamental feedback control design methodology. This section develops the rules for constructing root locus plots and presents examples illustrating the utility of the method for control design.

9.9.1 Motivational Example

From Section 9.6 it should be clear that the nature of the response of a system is dictated by the pole locations of the transfer function which describes it. A *root locus plot* is a plot of how the poles of a transfer function change as some parameter in the system is varied. This is useful because it may give us a means to determine the value of such a parameter that gives the system some desired

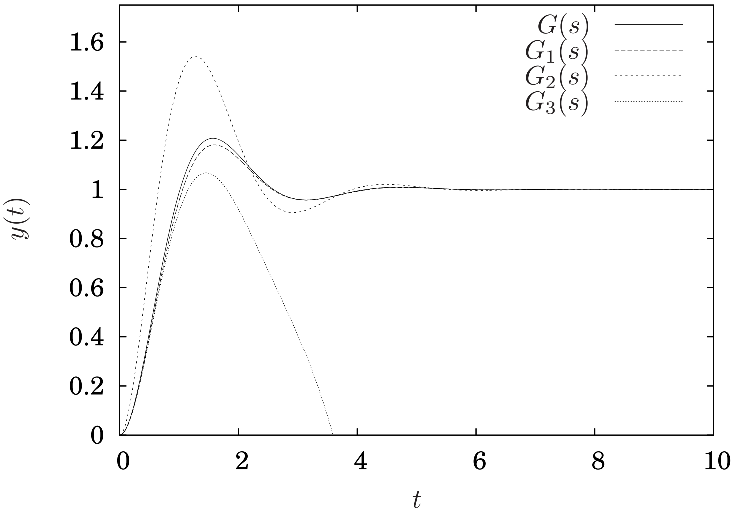


Figure 9.35. Effect of additional poles and zeros that are close together from Example 9.8.9.

response such as a specified rise time, maximum overshoot, settling time, *etc.* We will motivate this by a particular example and then in the following sections develop the rules that will allow us to basically sketch a root locus plot by hand.

Example 9.9.1 Consider the system illustrated in Figure 9.36 and assume the task is to control the position of the mass so that it stays at some desired location, x_d. Assume that there is some way to measure $x(t)$.

The equation of motion is

$$m\ddot{x} + b\dot{x} + kx = f(t).$$

To simplify the following equations, let $\hat{f}(t) = \frac{f(t)}{k}$ so the equation of motion

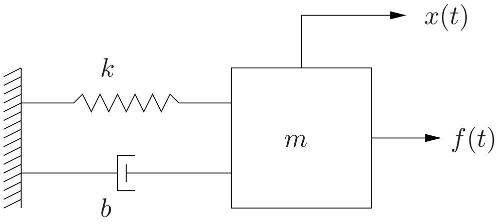


Figure 9.36. System for Example 9.9.1.

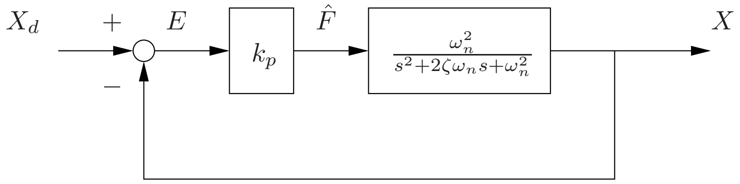


Figure 9.37. Block diagram for proportional control for Example 9.9.1.

can written

$$\ddot{x} + \frac{b}{m}\dot{x} + \frac{k}{m}x = \frac{k}{m}\hat{f}(t)$$

and the transfer function from the input force to the position of the mass is given by

$$\begin{aligned}\frac{X(s)}{\hat{F}(s)} &= \frac{\frac{k}{m}}{s^2 + \frac{b}{m}s + \frac{k}{m}} \\ &= \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}\end{aligned}$$

We will use proportional control so that

$$\hat{f}(t) = k_p\left(x_d - x(t)\right)$$

or

$$\hat{F}(s) = k_p\left(X_d(s) - X(s)\right).$$

A block diagram representation of the system with proportional control is illustrated in Figure 9.37. Included in the figure are labels for the error signal, $E(s)$ and the force, $\hat{F}(s)$. The transfer function from the desired position of the mass to the actual position is given by

$$\frac{X(s)}{X_d(s)} = \frac{k_p\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2\left(k_p + 1\right)}. \tag{9.24}$$

The nature of the transient response is easy to determine from the poles of Equation 9.24, which are simply given by the quadratic equation

$$p = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - (k_p + 1)}.$$

If $\omega_n = 1$ and $\zeta = 2$, then

$$p = -2 \pm \sqrt{4 - (k_p + 1)},$$

so if $k_p < 3$ the solutions will be exponentials and if $k_p > 3$ the solutions will be damped oscillations. A plot of the pole locations for various k_p is illustrated in Figure 9.38. Observe that for very small k_p there will be two real poles. One will be near the origin and the other will be near $s = -4$. As k_p increases the poles move toward each other along the real axis and at $k_p = 3$ they will both be at $s = -2$. Further increasing k_p will result in a complex conjugate poles. The real part of the poles for$k_p \geq 3$ is fixed at $s = -2$ and the imaginary part increases as k_p increases.

Before we solve for the step responses we can observe the following regarding the nature of the step response.

1. For $k_p \geq 3$ the settling time will not be changed by altering k_p. For $k_p < 3$ the settling time will be larger than for $k_p \geq 3$ since one of the poles will have a real part to the right of $s = -2$. Thus, the best we can do for settling time is at $k_p = 3$.

2. There will be no overshoot for $k_p \leq 3$ since the solutions will be exponentials. For $k_p > 3$ increasing k_p will increase the overshoot since it will decrease the angle between the pole and the imaginary axis.

3. For $k_p > 3$ the rise time will decrease as k_p increases. Note that it may be the case that it will be possible to satisfy either a rise time or a overshoot specification, but not both, since one gets worse with increasing k_p while the other gets better.

To verify our analysis, the corresponding step responses are illustrated in Figure 9.39. ■

At least in one respect our attempt to control the location of the mass in Example 9.9.1 is deficient since the steady state value of x depends on k_p and the steady state error is not zero. The way to remedy this should be obvious from Section 9.2, which is to add integral control. In the following example we will do just that. As expected, this will eliminate the steady state error. The larger point of the example, though is that once integral control is added, plotting the pole locations will not be easy since the denominator of the transfer function will be a third order polynomial, so a method to plot how the poles move as a parameter is varied that works for higher order polynomials is needed.

Example 9.9.2 In this example we will add integral control to try to control the location of the mass in Figure 9.36. Let

$$\hat{f}(t) = k_p \left(x_d - x(t)\right) + k_i \int_0^t x_d(\tau) - x(\tau) d\tau.$$

In general, it will be necessary to consider how altering both k_p and k_i affect the nature of the solution. Since we are considering how the system responds when one parameter is varied, we will fix the relationship between

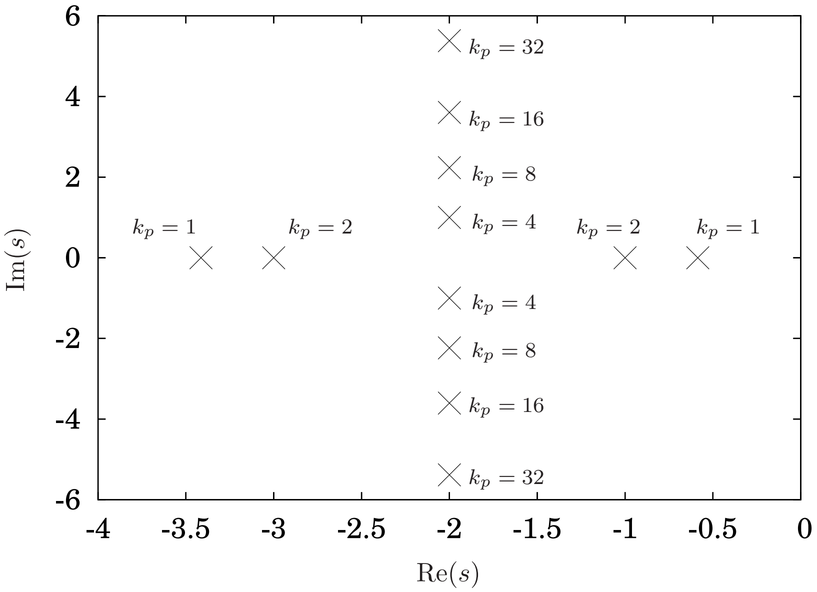


Figure 9.38. Pole locations for various k_p for Example 9.9.1.

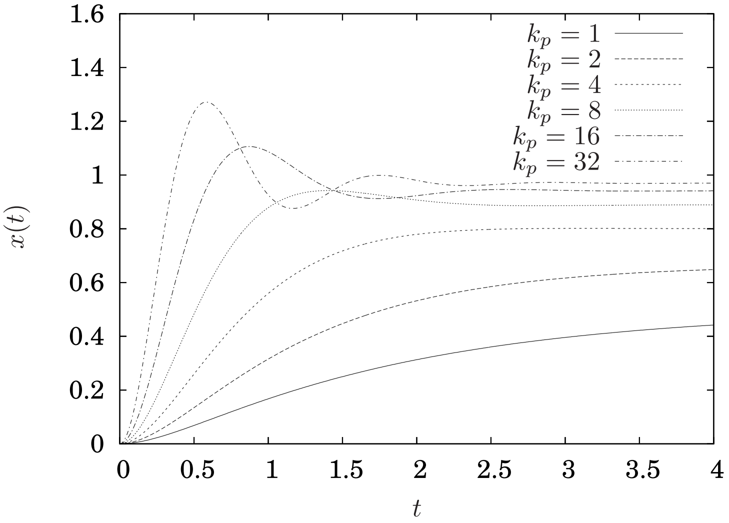


Figure 9.39. Step response for various k_p for Example 9.9.1.

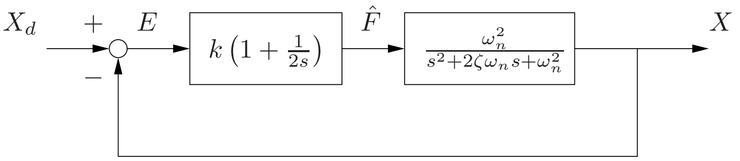

X_d
$+$
$-$
E
$k\left(1+\frac{1}{2s}\right)$
$\hat{F}$
$\frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}$
X

Figure 9.40. Block diagram for proportional plus integral control for Example 9.9.2.

k_p and k_i and if a satisfactory result is not obtained, change the relationship between them and start over.

Somewhat arbitrarily let $k_i = \frac{k_p}{2}$, so

$$\hat{f}(t) = k\left((x_d - x(t)) + \frac{1}{2}\int_0^t x_d(\tau) - x(\tau)d\tau\right)$$

where $k_p = k$ and $k_i = \frac{k}{2}$. The Laplace transform of the control law is

$$\begin{aligned} \hat{F}(s) &= k\left((X_d - X) + \frac{1}{2s}\right)(X_d - X) \\ &= k\left(1 + \frac{1}{2s}\right)(X_d - X) \end{aligned}$$

and the block diagram representation of this system is illustrated in Figure 9.40.

Using this control law, the transfer function, after a bit of algebra, is

$$\frac{X(s)}{X_d(s)} = \frac{k\omega_n^2\left(s+\frac{1}{2}\right)}{s^3 + 2\zeta\omega_n s^2 + \omega_n^2(1+k)s + \frac{k\omega_n^2}{2}}$$

and if $\omega_n = 1$ and $\zeta = 2$

$$\begin{aligned} \frac{X(s)}{X_d(s)} &= \frac{k\left(s+\frac{1}{2}\right)}{s^3 + 4s^2 + (1+k)s + \frac{k}{2}} \\ &= \frac{2ks + k}{2s^3 + 8s^2 + 2(1+k)s + k}. \end{aligned}$$

A critical point regarding the preceding two equations is that, in contrast to the system in Example 9.9.1, a tool as simple as the quadratic equation is not available to check how the poles of the transfer function vary as the parameter k is varied. Of course it may be done numerically, but having a tool available to do the analysis by hand is important to gain insight into such system. We will return to this example subsequently after we develop a method for doing that. ■

9.9.2 A Quick Review of Functions of a Complex Variable

A more detailed review of complex variable theory is contained in Appendix A. This section highlights the results necessary for sketching root locus plots.

Consider a transfer function of the form

$$G(s) = \frac{N(s)}{D(s)}.$$

Regardless of whether we can do it by hand, we may write the numerator and denominator in factored form. In particular, if we write

$$\begin{aligned} G(s) &= \frac{N(s)}{D(s)} \\ &= \frac{\prod_{i=1}^{n_z}(s - z_i)}{\prod_{i=1}^{n_p}(s - p_i)}, \end{aligned} \tag{9.25}$$

where the z_i are the zeros, the p_i are the poles, n_z is the number of zeros and n_p is the number of poles of $G(s)$.

Example 9.9.3 The transfer function

$$G(s) = \frac{s+3}{s^4 + 11s^3 + 40s^2 + 58s + 40} \tag{9.26}$$

may be written in the form

$$G(s) = \frac{s+3}{(s+4)(s+5)(s+(1+i))(s+(1-i))}. \tag{9.27}$$

■

A fundamental property of complex numbers is that they may be represented in a Cartesian manner, which is typically of the form

$$s = a + ib$$

where a is the real component and b is the imaginary component of s. This is a useful representation since the usual rules for multiplication hold as long as one considers $i = \sqrt{-1}$.

An alternative representation is in polar coordinates where s is represented by a magnitude and phase which are the usual Euclidean norm and angle if the number is plotted in its Cartesian coordinates. Referring to Figure 9.41, clearly if $s = a + ib$, then

$$\begin{aligned} r &= \sqrt{a^2 + b^2} \\ &= |s| \end{aligned}$$

and

$$\begin{aligned} \theta &= \tan^{-1}\left(\frac{b}{a}\right) \\ &= \angle s. \end{aligned}$$

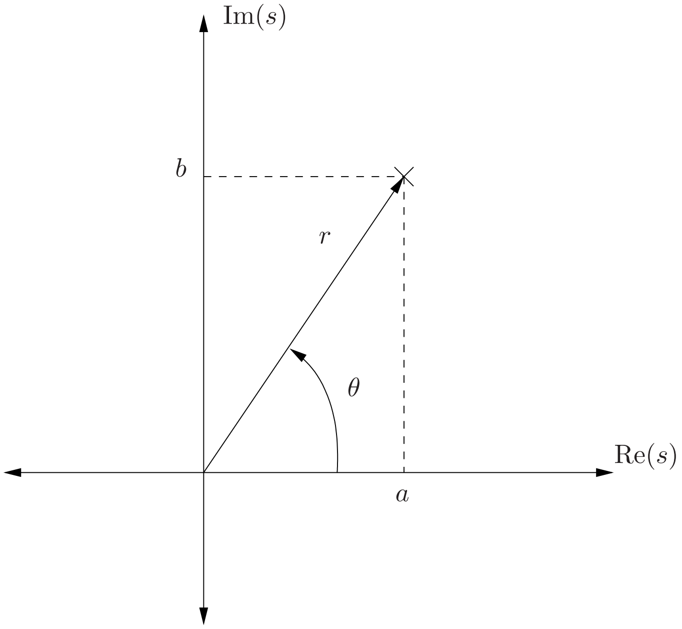


Figure 9.41. Cartesian, $s = a + ib$ and polar, $s = (r, \theta)$ forms of a complex number, s.

The Cartesian form is easy for addition and subtraction since if $s_1 = a_1 + ib_1$ and $s_2 = a_2 + ib_2$, then

$$s_1 + s_2 = (a_1 + a_2) + i\,(b_1 + b_2)\,.$$

However, multiplication is easier in polar form. In particular, if $s_1 = (r_1, \theta_1)$ and $s_2 = (r_2, \theta_2)$, then the product is

$$s_1 s_2 = (r_1 r_2, \theta_1 + \theta_2)$$

and the quotient is

$$\frac{s_1}{s_2} = \left(\frac{r_1}{r_2}, \theta_1 - \theta_2\right).$$

Proving these two results is simple and is left as an exercise.

The critical concept in this section is relating how to evaluate a transfer function in terms of the location in the complex plane of s and the location of its poles and zeros. Returning to Equation 9.25,

$$G(s) = \frac{\prod_{i=1}^{n_z}(s - z_i)}{\prod_{i=1}^{n_p}(s - p_i)} = \frac{(s - z_1)(s - z_2)\cdots(s - z_{n_z})}{(s - p_1)(s - p_2)\cdots\left(s - p_{n_p}\right)}.$$

note that the numerator is simply the product of the difference between s and all of the zeros of $G(s)$. Similarly, the denominator is the product of the difference between s and each of the poles of $G(s)$.

This concept is critical to understand the development which follows. If it is still not clear after the following example, the reader is strongly encouraged to fully understand it before proceeding to the next section.

Example 9.9.4 Returning to the transfer function from Example 9.9.3 let

$$G(s) = \frac{s+3}{(s+4)(s+5)(s+(1+i))(s+(1-i))}.$$

If we wish to determine $G(s)$ at a particular value for s the easiest thing would be just to substitute in into G. For example,

$$\begin{aligned} G(0) &= \frac{3}{(4)(5)(1+i)(1-i)} \\ &= \frac{3}{40} \end{aligned}$$

and

$$\begin{aligned} G(i) &= \frac{3+i}{(4+i)(5+i)(1+2i)(1)} \\ &= \frac{3+i}{1+47i} \\ &= \frac{3+i}{1+47i}\frac{1-47i}{1-47i} \\ &= \frac{-44+142i}{2210}. \end{aligned}$$

Note that in polar coordinates

$$G(0) = \left(\frac{3}{40}, 0\right)$$

and

$$\begin{aligned} G(i) &\approx (0.067267, -1.2278) \\ &= (0.067267, -70.346^\circ)\,. \end{aligned}$$

"Plugging and chugging" may be best to evaluate the Cartesian form of $G(s)$. In polar form there is a geometric interpretation as well. Figure 9.42 plots the poles and zeros of $G(s)$ and marks $s = i$ with a $+$. Now consider each term in the numerator and denominator of $G(s)$. Each is of the form $s - z$ or $s - p$ and one way to interpret $s - z$ or $s - p$ is that it is the vector from z or p respectively to the point s, as is illustrated in Figure 9.43.

So, an alternative way to evaluate $G(s)$ is to consider the vectors from all the zeros and poles of $G(s)$ to the point s. The magnitude of $G(s)$ will be the product of the magnitudes of all the vectors from the zeros of $G(s)$ to s divided by the product of the magnitudes of all the vectors from the poles of $G(s)$ to s. Mathematically,

$$\begin{aligned} |G(s)| &= \frac{\prod_{i=1}^{n_z}(s - z_i)}{\prod_{i=1}^{n_p}(s - p_i)} \\ &= \frac{|s - z_1|\,|s - z_2|\cdots|s - z_{n_z}|}{|s - p_1|\,|s - p_2|\cdots|s - p_{n_p}|}. \end{aligned}$$

In words, we may graphically measure the length of all the arrow from the poles and zeros of $G(s)$ to s and multiply them for zeros and divide by them for poles to determine the magnitude of $G(s)$.

Since the angle of complex numbers add when they are multiplied, then the angle of $G(s)$ is determined by summing the angles from all the zeros to s and by subtracting the angle from all the poles to s. Mathematically,

$$\begin{aligned} \angle G(s) &= \sum_{i=1}^{n_z} \angle(s - z_i) - \sum_{i=1}^{n_p} \angle(s - p_i) \\ &= \angle(s - z_1) + \angle(s - z_2) + \cdots + \angle(s - z_{n_z}) \\ &- \angle(s - p_1) - \angle(s - p_2) - \cdots - \angle\left(s - p_{n_p}\right). \end{aligned}$$

In words, we can measure the angle from all the zeros and poles to s and sum the angles from the zeros and subtract the angles from the poles.

Now evaluating $G(s)$ using this graphical interpretation

$$\begin{aligned} |G(i)| &= \frac{\left|\sqrt{10}\right|}{|1|\left|\sqrt{5}\right|\left|\sqrt{17}\right|\left|\sqrt{26}\right|} \\ &\approx 0.067267, \end{aligned}$$

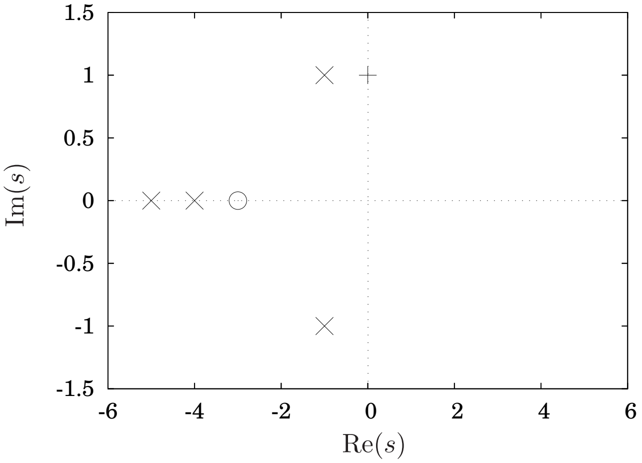


Figure 9.42. The poles and zeros of $G(s) = \frac{s+3}{(s+4)(s+5)(s+(1+i))(s+(1-i))}$ and the point $s = i$ (+) for Example 9.9.4.

and

$$\begin{aligned}\angle G(i) &= \tan^{-1}\left(\frac{1}{3}\right) - \tan^{-1}\left(\frac{0}{1}\right) - \tan^{-1}\left(\frac{2}{1}\right) - \tan^{-1}\left(\frac{1}{4}\right) - \tan^{-1}\left(\frac{1}{5}\right) \\ &\approx -70.346^\circ.\end{aligned}$$

Observe that with an even scale on the graph, a ruler and protractor, one could evaluate $G(s)$ with pretty decent accuracy. ■

In the next section we are going to be very concerned with values of s for which $G(s)$ is negative and real. In that case we want $\angle G(s) = \pm 180^\circ$, which is the angle for a negative real number.

9.9.3 Root Locus Plotting Rules

Consider the transfer function described by the block diagram illustrated in Figure 9.44. Because the feedback loop does not contain a block, it is called *unity feedback* . The transfer function for this system is

$$\frac{Y(s)}{R(s)} = \frac{kG(s)}{1+kG(s)}. \tag{9.28}$$

We wish to study how the poles of $\frac{kG(s)}{1+kG(s)}$ change as k is varied, so we need to know the poles change with k. A pole is a value of s which satisfied

$$1 + kG(s) = 0.$$

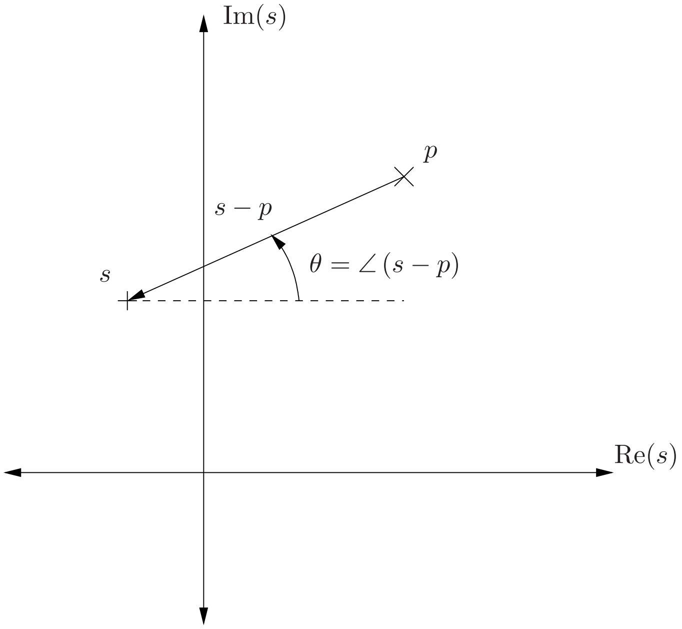


Figure 9.43. The vector $(s-p)$ from the point p to the point s.

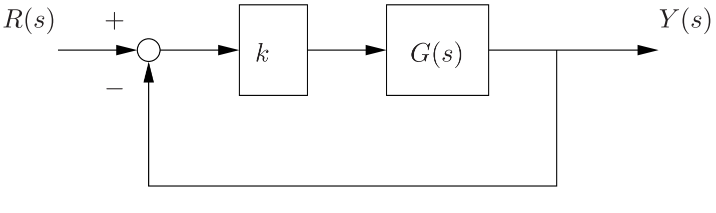


Figure 9.44. Unity feedback block diagram.

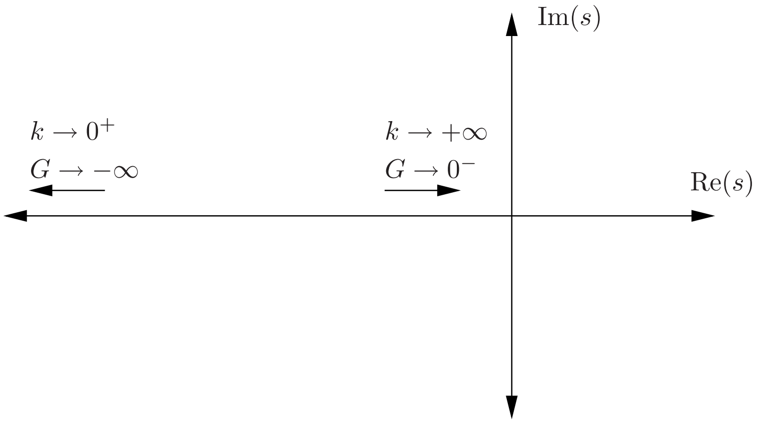


Figure 9.45. The relationship on the complex plane of k and $G(s)$ if $1 + kG(s) = 0$.

Hence, s is a pole if

$$G(s) = -\frac{1}{k}.$$

We will limit our attention to k values that are zero or real and positive. As k goes from 0 to $+\infty$, $-\frac{1}{k}$ will go from the $-\infty$ to the origin along the negative real axis as is illustrated in Figure 9.45.

So, if we can determine all s values for which $\angle G(s) = 180°$, we will have plotted all the solutions to $1 + kG(s) = 0$ for positive k values. Doing so we can determine how the poles of the transfer function in Equation 9.28 change with k, and hence will be able to determine properties of the response of Equation 9.28 such as the percent overshoot, settling time, rise time, *etc.* This is the *root locus plot* for the transfer function $\frac{Y(s)}{R(s)}$.

First let us consider the two limiting cases where $k = 0$ and $k \to +\infty$. In the case where $k = 0$, the only way for $1 + kG(s)$ to be zero is if $G(s)$ is unbounded. So we can state the following rule.

Rule 9.9.5 If the denominator of $G(s)$ is nth order, then At $k = 0$, one of each of the n branches of the root locus will start at one of the poles of $G(s)$. ⋄

As $k \to +\infty$, the only way for $1 + kG(s)$ to equal zero is for $G(s) \to 0$. Since we are considering only proper transfer functions where the order of the denominator is greater than the numerator, the only way for $|G(s)| \to \infty$ is for s to approach a pole. In contrast, there are two ways that $|G(s)|$ may approach 0. The first, obviously, is if s approaches a zero. Also, if s grows unbounded in any direction $G(s)$ will approach 0 since the transfer function is proper and order of the denominator is greater than the order of the numerator. So, we can state the second rule.

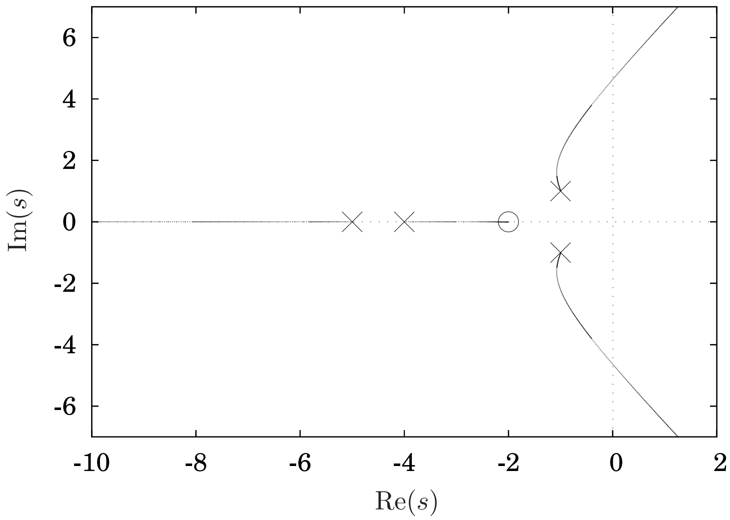


Figure 9.46. The root locus plot for Example 9.9.7. The poles of $G(s)$ are marked with a $\times$ and the zeros of $G(s)$ are marked with a $\circ$.

Rule 9.9.6 As $k \to +\infty$, the root locus either approaches a zero of $G(s)$ or grows unbounded. ⋄

An example may be useful at this point. At this point we only have two rules, so many of the features of the root locus unrelated to these two rules will not be obvious.

Example 9.9.7 Figure 9.46 illustrates the solutions of

$$1 + k \frac{s+3}{(s+4)\,(s+5)\,(s^2+2s+3)} = 0$$

as k goes from 0 to $+\infty$. As is clear from the figure, the branches of the root locus starts at each pole of $G(s)$ and one of them approaches the one zero of $G(s)$ while the other three grow unbounded as $k \to +\infty$. ■

Before we determine exactly how the solutions to $1+kG(s)$ grow unbounded as $k \to +\infty$, observe that the root locus is comprised of *branches*. For a transfer function with a characteristic equation of order n, the fundamental theorem of algebra requires that there be n solutions. Indeed, in the previous example, it appears that for any given value of k, there are four solutions and as k varies, these solutions move along continuous lines in the complex plane. The fact that these lines are continuous should make sense. If k is only slightly altered, then the n solutions will only be slightly altered as well. Hence, as k varies continuously from 0 to $+\infty$, the solutions to $1 + kG(s) = 0$ will vary continuously as well. Since the root locus starts at $k = 0$ at the poles of $G(s)$, each branch that corresponds to one of the solutions will start at one of the poles. Since we will need to refer back to it, we will restate this argument as a proposition.

Proposition 9.9.8 *The solutions to $1 + kG(s) = 0$ depend continuously on k.*

Referring to Figure 9.46 with is the root locus for Example 9.9.7, it is clear that the three branches that grow unbounded do so along specific asymptotes. Since $G(s)$ is the same in Example 9.9.7 as in Example 9.9.4 we can use the same labels for the poles as in Figure 9.42.

Specifically, the branch of the locus that leaves pole p_1 grows unbounded in a manner where both the real and imaginary parts of s approach $+\infty$. The branch that leaves p_2 has an imaginary part that grows to $-\infty$ whereas the real part approaches $+\infty$. This branch also appears symmetric to the first branch about the real axis. In fact this must be so since if a complex number is a root of a polynomial, its complex conjugate must also be a root. Finally, the third branch grows unbounded with a zero imaginary part and a real part that approaches $-\infty$.

In order to determine these asymptotes, consider a map of the poles and zeros of $G(s)$ that has a very large scale, such as illustrated in Figure 9.47 for the transfer function in the previous examples. If we desire to determine whether or not a point, indicated by a cross in the figure, we will use the fact that it will be on the root locus only if $\angle G(s) = \pm 180°$.

Recall that

$$\angle G(s) = \sum_{i=1}^{n_z} \angle\left(s - z_i\right) - \sum_{i=1}^{n_p} \angle\left(s - p_i\right).$$

For very large s, the angle from all the poles and zeros are approximately the same. Hence if θ is this angle then

$$\angle G(s) = \theta\left(n_z - n_p\right)$$

and s will be on the root locus if

$$\theta\left(n_z - n_p\right) = \pm 180°.$$

We can always add or subtract 360° from the angle. Doing so and solving for θ gives

$$\theta = \frac{(180° + n360°)}{(n_z - n_p)} \qquad n = 0, 1, \ldots \left(n_p - n_z\right),$$

where n_z is the number of zeros and n_p is the number of poles of $G(s)$. It looks like we have figured out another rule.

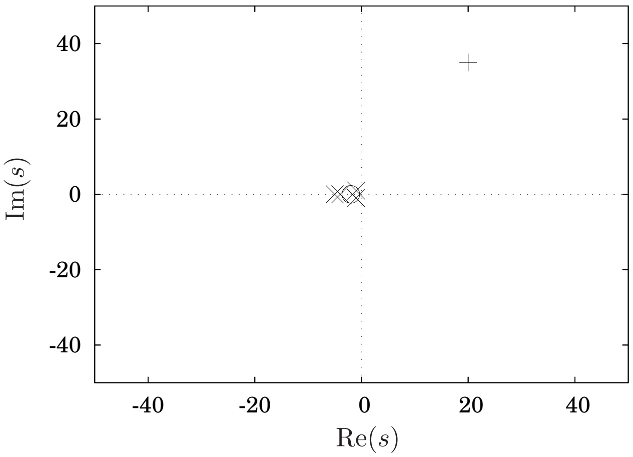


Figure 9.47. The poles and zeros of a transfer function with a large scale appear in a small cluster.

Rule 9.9.9 The branches of the root locus that grow unbounded do so along asymptotes with angles

$$\theta_n = \frac{(180^\circ + n360^\circ)}{(n_z - n_p)}$$

⋄

Let us verify this rule on the previous example.

Example 9.9.10 Since $G(s)$ from Example 9.9.7 had four poles and one zero, there asymptotes will have angles

$$\begin{aligned} \theta_0 &= -60^\circ \\ \theta_1 &= -180^\circ \\ \theta_2 &= -300^\circ \\ &= 60^\circ. \end{aligned}$$

Rule 9.9.9 gives and *angle* of the asymptote, but not the point at which the asymptotes intersect the real axis.

Rule 9.9.11 The asymptotes intersect the real axis at the point

$$s_{int} = \frac{\sum_{i=1}^{n_z} z_i - \sum_{i=1}^{n_p} p_i}{n_z - n_p}$$

⋄

A particularly easy set of points on the locus to plot are those on the real axis. The critical fact to consider is that for a point, s, on the real axis, the angle

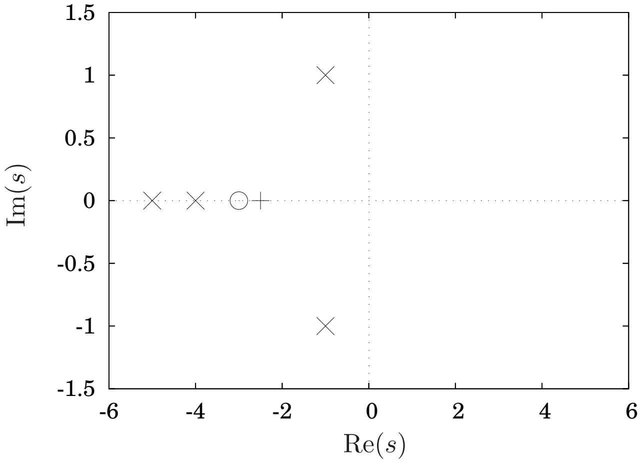


Figure 9.48. Evaluating $\angle G(s) = 0°$ by considering the geometry of s relative to the poles and zeros of $G(s)$ on the complex plane.

of the point will not be affected by either complex conjugate poles or zeros. In each case the contribution to the angle from each part of the complex conjugate pair will cancel. This is illustrated in the following example.

Example 9.9.12 Consider once again

$$1 + k\frac{s+3}{(s+5)\,(s+5)\,(s^2+2s+2)} = 0$$

and let $s = -2.5$. The poles and zeros are plotted in Figure 9.48 and the point s is illustrated by the cross. Observe that the contribution to the angle of $G(s)$ by the complex conjugate pair of poles in this example is $360°$, which is equivalent to $0°$. Hence, when evaluating

$$\angle G(s) = \sum_{i=1}^{n_z} \angle\,(s - z_i) - \sum_{i=1}^{n_p} \angle\,(s - p_i)$$

they do not matter. ■

From the preceding example it is hopefully obvious that *all* complex conjugate pairs of poles or zeros will contribute nothing to $\angle G(s)$. Hence, *for real s* only the poles and zeros on the real axis affect $\angle G(s)$. Furthermore, if s is real and the only poles and zeros that matter are real, all of the angles in the sum

$$\angle G(s) = \sum_{i=1}^{n_z} \angle\,(s - z_i) - \sum_{i=1}^{n_p} \angle\,(s - p_i)$$

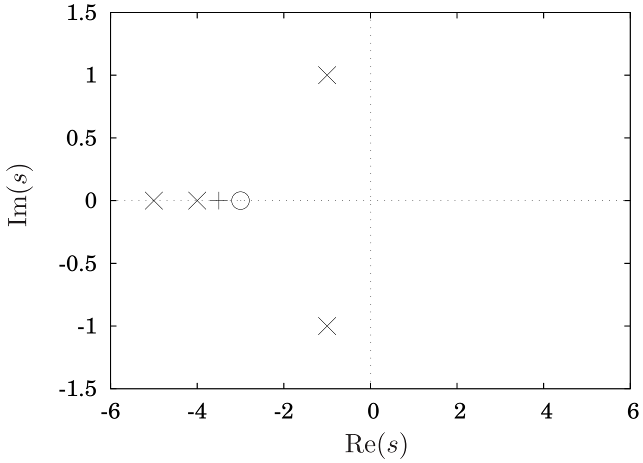


Figure 9.49. Evaluating $\angle G(s) = 180^\circ$ by considering the geometry of s relative to the poles and zeros of $G(s)$ on the complex plane.

will either be 0° or 180° depending upon whether the point s is to the right or left respectively of the pole or zero in question.

In fact, if s is to the right of *all* the real poles and zeros of $G(s)$, then $\angle G(s) = 0$ since the angle from each of them is zero. If s is decreased and crosses to the left of the first pole or zero, then $\angle G(s) = \pm 180^\circ$ where the sign of the angle depends upon whether or not it was a pole or a zero.

Example 9.9.13 Returning to the previous series of examples, if $s = -3.5$, as is illustrated in Figure 9.49, then $\angle G(s) = 180^\circ$. This should be clear from the figure since s is to the left of z_1, so $\angle (s - z_1) = 180^\circ$. Since s is to the right of p_2 and p_3, $\angle (s - p_2) = \angle (s - p_3) = 0$. ■

If we continue to decrease s so that it passes another pole or zero, then the angle of $G(s)$ will increase or decrease by 180°. Regardless $\angle G(s) = 0^\circ$ since it will either algebraically sum to zero or it will be 360°, which is equivalent to zero. Once it passes another one, $\angle G(s) = \pm 180^\circ$ and then when it passes the next, $\angle G(s) = 0$, *etc.* Hence, we have the following rule.

Rule 9.9.14 On the real axis, the root locus is to the left of an odd number of zeros and poles. ⋄

This rule certainly holds in the example case we have been considering if we refer back to Figure 9.46.

At this point we have considered every feature of the root locus in Figure 9.46 except one, which is the angle the locus departs the complex conjugate pair of poles. The loci appear to depart p_1 at approximately 90° and depart p_2 at approximately −90°. Instead of zooming out to consider a very large s like we did to find the asymptote angles, we will zoom in and consider a point s very close to a pole. We should be able to determine, for example, that a point s very close to p_1 must be at an angle approximately equal to 90° from p_1.

In fact, in order to determine the departure angle from a pole or zero, all we must do is consider a point very close to it. If s is very close to a pole, p_i, or zero, z_i, the angle from all the other poles and zeros to it is approximately the same as the angle from the other poles and zeros to the pole or zero to which s is close and all we must do is solve

$$\angle G(s) = \sum_{i=1}^{n_z} \angle (s - z_i) - \sum_{i=1}^{n_p} \angle (s - p_i)$$

for the term $(s - p_i)$ or $(s - z_i)$ and substitute p_i or z_i for s in all the terms except the one to which s is adjacent.

Rule 9.9.15 The angle at which a branch of the root locus leaves a pole, p_j is given by

$$\angle (s - p_j) = \sum_{i=1}^{n_z} \angle (p_j - z_i) - \sum_{i=1, i \neq j}^{n_p} \angle (p_j - p_i) - 180^\circ$$

and the angle which it approaches a zero, z_j is given by

$$\angle (s - z_i) = 180^\circ - \sum_{i=1, i \neq j}^{n_z} \angle (z_j - z_i) + \sum_{i=1, i \neq j}^{n_p} \angle (z_j - p_i) .$$

◇

Rule 9.9.15 works for real as well as complex poles and zeros. However there is no point in doing the computation for the real poles and zeros because Rule 9.9.14 gives the appropriate angle.

The final rule is addresses a feature not present in Figure 9.46, so we will present another example that will review a couple of the rules we know so far and introduce the need for the final rule.

Example 9.9.16 Consider

$$G(s) = \frac{s + 6}{(s + 1)(s + 3)}.$$

The poles and zeros of $G(s)$ are illustrated in Figure 9.50. The first rule will will apply is Rule 9.9.14, so we know the root locus will be between the two poles and then to the left of the zero. If we compute the asymptote angles, we will get only one asymptote at $\theta_0 = 180^\circ$, which coincides with the part to the left of the zero already completed by Rule 9.9.14. This part of the root locus is illustrated in Figure 9.51.

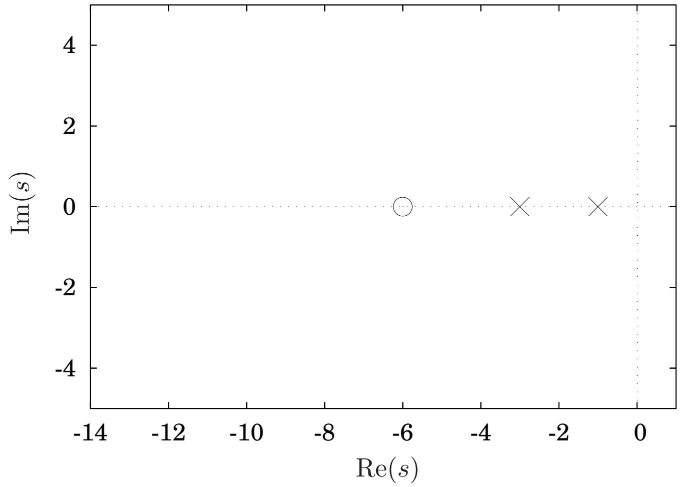


Figure 9.50. Poles and zeros of $G(s) = \frac{s+6}{(s+1)(s+3)}$ for Example 9.9.16.

Now, consider Rules 9.9.5 and 9.9.6, which require that the branches of the root locus start at the poles of $G(s)$ and end at either zeros of $G(s)$ or grow unbounded. So far the root locus does have branches that start at the poles and it does end at the zero and a does grow unbounded.

However, recall Proposition 9.9.8 which states that the root locus must be continuous. Therefore there must be a way that the branches that start from the poles are connected to the branches that go to the zero or infinity. They cannot connect along the real axis because between the middle pole and the zero $\angle G(s) = 0°$. Hence, the only way it may happen is that they "break away" from the real axis between the poles and "break in" to the real axis to the left of the zero. If you have not already peeked, the root locus, computed numerically, is illustrated in Figure 9.52.

This example is actually quite interesting. For small k, the poles are both real, then as k is increased they are a complex conjugate pair, and as k is even further increased they become real again. Intriguing. ■

First we will determine the rule to compute exactly where the root locus will break in and away from the real axis. Then we will present an argument as to why the curve between the break in and away points is a rather nice near-circle, as opposed to being, for example, very wavy between the break in and away points.

The rule for the break in and out points is simple. Since the root locus starts at the poles, the point at which the locus will break away corresponds to the maximum value that k attains on the real axis. Correspondingly, the branches

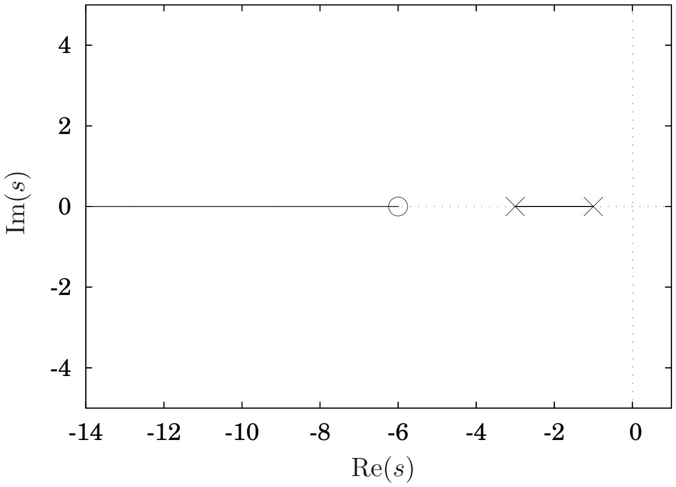


Figure 9.51. A partial root locus plot for Example 9.9.16.

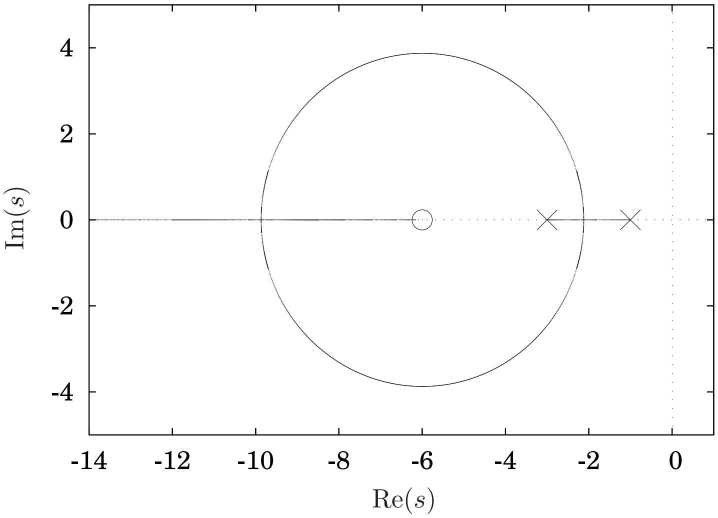


Figure 9.52. The root locus plot for $G(s) = \frac{s+6}{(s+1)(s+3)}$.

will break in at a minimum value for k. Hence the break away and break in points can be determined by solving $1 + kG(s) = 0$ for k and determining the points on the part of the locus on the real axis for which the derivative of $k = \frac{1}{G(s)}$ with respect to s is maximum.

Rule 9.9.17 For the part of the root locus on the real axis (determined by Rule 9.9.14), the locus will break in or break away at points where

$$\frac{d}{ds}\left(\frac{1}{G(s)}\right) = 0. \tag{9.29}$$

◇

Note that there may be other points at which the derivative in Equation 9.29 are zero, but if they are not on the real axis to the left of an odd number of poles and zeros they are not relevant. The reason these may occur include, for example, an extremum for k that corresponds to a negative value for k.

Example 9.9.18 Returning to Example 9.9.16, solving $1 + kG(s) = 0$ for k gives

$$k = \frac{(s+1)(s+3)}{s+6}.$$

Differentiating with respect to s gives

$$\begin{aligned}\frac{dk}{ds} &= \frac{(2s+4)(s+6) - (1)\left(s^2+4s+3\right)}{(s+6)^2} \\ &= \frac{s^2+12s+21}{(s+6)^2}.\end{aligned}$$

Hence,

$$\frac{dk}{ds} = 0 \qquad \Longleftrightarrow \qquad s = -6 \pm \sqrt{15}$$

or

$$s \approx -2.1270 \qquad \text{and} \qquad s \approx -9.8730,$$

which conforms to Figure 9.52. ■

Let us revisit the very first motivational example.

Example 9.9.19 The transfer function from Example 9.9.1 was

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2},$$

or using the numerical values $\omega_n = 1$ and $\zeta = 2$,

$$G(s) = \frac{1}{s^2+4s+1}.$$

We will reconstruct the whole root locus subsequently, but for now observe that solving $1 + kG(s) = 0$ for k gives

$$k = s^2 + 4s + 1$$

and hence

$$\frac{dk}{ds} = 0 \qquad \Longleftrightarrow \qquad s = -2,$$

which is where the poles break away from the real axis in Figure 9.38. ■

A summary of the root locus plotting rules, reordered in a manner that is most useful for sketching the root locus by hand appears in Table 9.1.

9.9.4 Examples

This section will present a few examples which will illustrate the application of the root locus plotting rules in Table 9.1.

Example 9.9.20 Let us return to the PI control problem from Example 9.9.2. In that problem the transfer function was expressed in the block diagram in Figure 9.40. If $\omega_n = 1$ and $\zeta = 2$, transfer function is

$$\frac{X(s)}{X_d(s)} = \frac{k\frac{1+\frac{1}{2s}}{s^2+4s+1}}{1+k\frac{1+\frac{1}{2s}}{s^2+4s+1}} \tag{9.30}$$

Hence, the transfer function to use in all the plotting rules is

$$\begin{aligned} G(s) &= \frac{1+\frac{1}{2s}}{s^2+4s+1} \\ &= \frac{s+\frac{1}{2}}{s\left(s^2+4s+1\right)}. \end{aligned}$$

Recall that when we added integral control in Example 9.9.2 we could not proceed any farther than determining the transfer function since the denominator was third order. Now, after all the work in the preceding section, we can accomplish what we wanted, which was to see how the poles of the transfer function in Equation 9.30 vary as the gain k is varied from 0 to $+\infty$.

Let us follow the steps exactly as they appear in Table 9.1.

1. $G(s)$ has a zero at $a = -\frac{1}{2}$ and three poles at $s = 0$, $s \approx -3.73205$ and $s \approx -0.26795$, all of which are easy to determine by hand. A plot of the poles and zeros for $G(s)$ appears in Figure 9.53.

2. Now filling in to the left of an odd numbers of zeros plus poles results in the partial root locus plot illustrated in Figure 9.54.

3. There are three poles and one zeros, so $n_z - n_p = -2$. Hence the two asymptote angles are

$$\begin{aligned} \theta_0 &= -90^\circ \\ \theta_1 &= 90^\circ, \end{aligned}$$

Rules to plot the solutions of $1 + kG(s) = 0$ for $k \in [0, \infty)$.

1. Plot the poles and zeros of $G(s)$. Each branch of the root locus starts at one of the poles. If $G(s)$ has n_p poles and n_z zeros, n_z of the branches will end at the zeros (Rules 9.9.5 and 9.9.6)

2. Draw the root locus on the real axis to the left of an odd number of poles plus zeros. (Rule 9.9.14)

3. Compute the asymptote angles using

$$\theta_n = \frac{(180^\circ + n360^\circ)}{(n_z - n_p)}$$

   Sketch the asymptotes, which intersect the real axis at

$$s_{int} = \frac{\sum_{i=1}^{n_z} z_i - \sum_{i=1}^{n_p} p_i}{n_z - n_p}.$$

   (Rules 9.9.9 and 9.9.11)

4. If $G(s)$ has any complex conjugate pairs of poles or zeros, compute the departure or arrival angles, respectively, by taking a point very close to one of them and computing the angle from the pole or zero that would be necessary to ensure $\angle G(s) = 180^\circ$. (Rule 9.9.15)

5. Compute the break away or break in points from the real axis, if any, by computing the values for which

$$\frac{d}{ds}\left(\frac{1}{G(s)}\right) = 0.$$

   (Rule 9.9.17)

6. Complete the root locus keeping in mind that the branch connecting two sections cannot be too complicated if the order of the numerator and denominator of $G(s)$ is not too large.

Table 9.1. Root locus plotting rules.

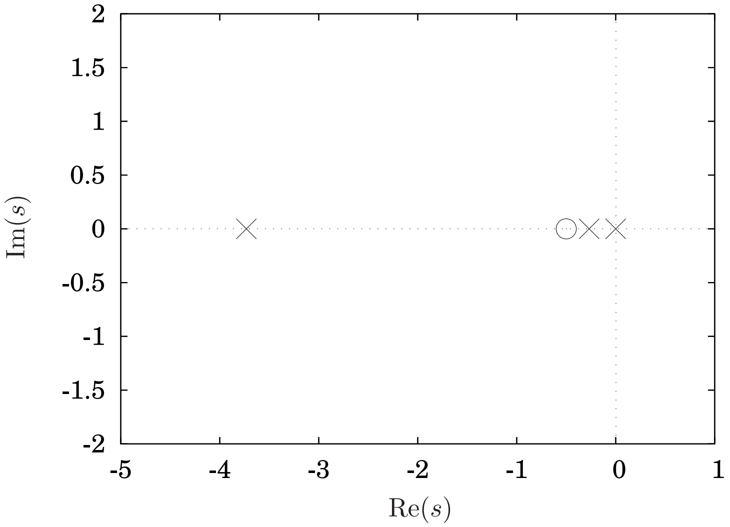


Figure 9.53. Partial root locus plot for $G(s) = \frac{s+\frac{1}{2}}{s(s^2+4s+1)}$ after step 1.

and the intersection with the real axis is at

$$\begin{aligned} s_{int} &= \frac{(-1/2) - (0 - 3.73205 - 0.26795)}{1-3} \\ &= -1.75. \end{aligned}$$

The asymptotes are sketched on the root locus diagram by the dashed lines in Figure 9.55.

4. Step 4 does not apply since there are no complex conjugate poles and zeros.

5. Differentiating $k = \frac{1}{G(s)}$ with respect to s gives

$$\begin{aligned} \frac{d}{ds}\left(\frac{1}{G(s)}\right) &= \frac{d}{ds}\frac{s\left(s^2+4s+1\right)}{s+\frac{1}{2}} \\ &= \frac{\left(3s^2+8s+1\right)\left(s+\frac{1}{2}\right) - \left(s^3+4s^2+s\right)}{\left(s+\frac{1}{s}\right)^2} \\ &= \frac{2s^3+\frac{11}{2}s^2+4s+\frac{1}{2}}{\left(s+\frac{1}{2}\right)^2}. \end{aligned}$$

Finding the zeros of the cubic polynomial in the numerator unfortunately is a bit hard to do by hand. Hence, we will just give the answer.

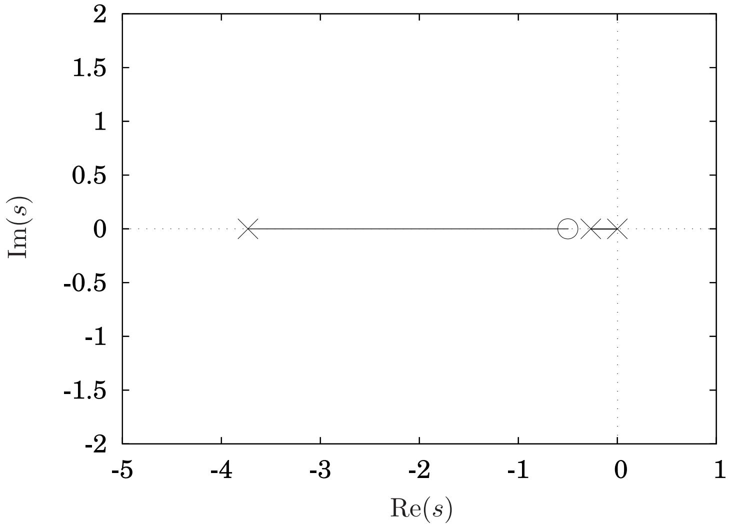


Figure 9.54. Partial root locus plot for $G(s) = \frac{s+\frac{1}{2}}{s(s^2+4s+1)}$ after step 2.

$\frac{dk}{ds} = 0$ at the values

$$\begin{aligned} s &= -1.59307 \\ s &= -1.00000 \\ s &= -0.15693. \end{aligned}$$

The root locus must break out from the point $s = -0.15693$ since it is between two poles. The other two points are also on the root locus on the real line, so one must be a break in point and one a break away point for the loci to grow unbounded along the asymptotes.

6. The completed root locus is illustrated in Figure 9.56.

Example 9.9.21 Sketch the root locus plot for

$$G(s) = \frac{s+3}{(s+1)(s+2)}.$$

1. The poles are at $s = -1$ and $s = -2$. There is one zero at $s = -3$.

2. The root locus on the real axis is to the left of an odd number of zeros plus poles, as is illustrated in Figure 9.57.

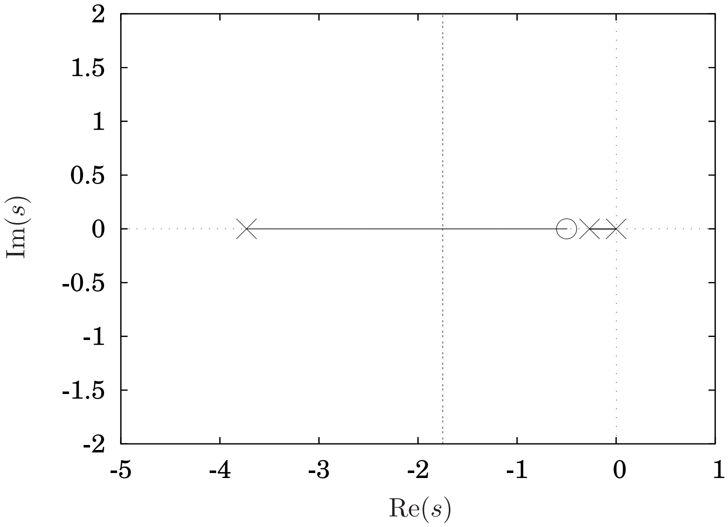


Figure 9.55. Partial root locus plot for $G(s) = \frac{s+\frac{1}{2}}{s(s^2+4s+1)}$ after step 3.

3. There are two poles and one zero, so the only asymptote is at $\theta = 180°$, which has already been plotted by the step dealing with the root locus on the real axis.
4. There are no complex conjugate poles or zeros of $G(s)$, so this step does not apply.
5. The break in and away points will be where

$$\frac{dk}{ds} = 0$$

or

$$\begin{aligned}\frac{dk}{ds} &= \frac{d}{ds}\frac{1}{G(s)} \\ &= \frac{d}{ds}\left(\frac{(s+1)(s+2)}{s+3}\right) \\ &= \frac{d}{ds}\left(\frac{s^2+3s+2}{s+3}\right) \\ &= \frac{(2s+3)(s+3)-\left(s^2+3s+2\right)}{(s+3)^2} \\ &= \frac{s^2+6s+7}{(s+3)^2},\end{aligned}$$

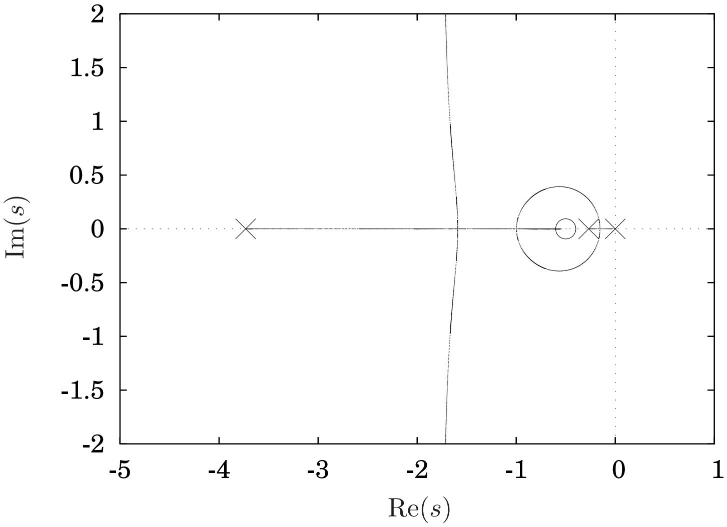


Figure 9.56. The completed root locus plot for $G(s) = \frac{s+\frac{1}{2}}{s(s^2+4s+1)}$. ■

which is equal to zero at $s \approx -4.4142$ and $s \approx -1.5858$. The first must be a break in point and the latter a break away point.

6. Since the root locus must break out between the poles and break in to the left of the zero, the root locus must be comprised of complex conjugate pairs between the two. Because we are sketching the roots of a relatively low order polynomial, the path between the break away and break in point must be relatively low order, *i.e.,* constant, or nearly constant, curvature, which is somewhat semi-circular. The completed root locus is illustrated in Figure 9.58. ■

Example 9.9.22 Sketch the root locus for

$$G(s) = \frac{s-1}{s^2+2s+5}.$$

1. There is a zero at $s = 1$ and two poles at $s = -1 \pm 2i$.

2. On the real axis, the root locus will be to the left of the zero, as is illustrated in Figure 9.9.22.

3. There are two poles and one zero, so the only asymptote is at 180°, which has already been completed by the rule for the locus on the real axis.

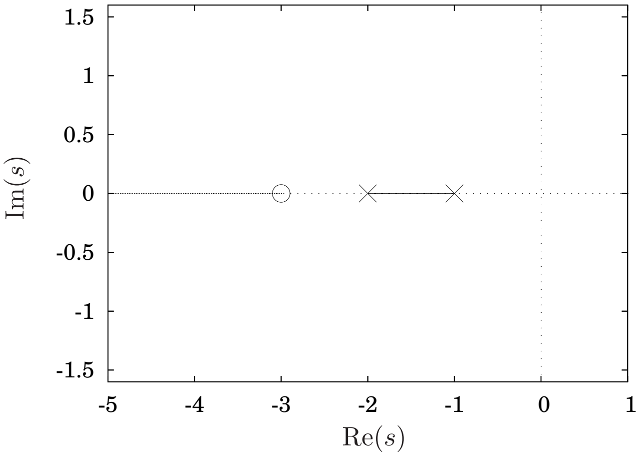


Figure 9.57. Partial root locus for Example 9.9.21.

.

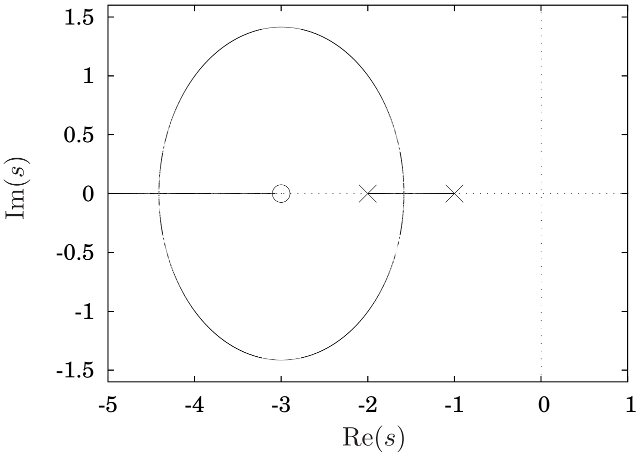


Figure 9.58. Root locus for Example 9.9.21.

.

4. Considering a point near the upper complex conjugate pole and determining $\angle G(s)$, we have

   $$135^\circ - 90^\circ - \theta = \pm 180^\circ$$

   which gives the departure angle as

   $$\theta = 225^\circ.$$

   The angle of departure for the bottom pole is symmetric, so is equal to 135°.

5. The locus must break in to the real axis as some point because it ends at the zero and along the asymptote going to $-\infty$. Computing

   $$\begin{aligned} \frac{dk}{ds} &= \frac{d}{ds}\left(\frac{1}{G(s)}\right) \\ &= \frac{d}{ds}\left(\frac{s^2+2s+5}{s-1}\right) \\ &= \frac{(2s+2)(s-1)-\left(s^2+2s+5\right)}{(s-1)^2} \\ &= \frac{s^2-2s-7}{(s-1)^2}, \end{aligned}$$

   which gives

   $$\frac{dk}{ds} = 0 \qquad \Longleftrightarrow \qquad s \approx 3.8284, -1.8284.$$

   Since the first solution is not on the root locus on the real axis, we may ignore it. The second solution gives the break in point.

6. The complete root locus plot is illustrated in Figure 9.60. ■

9.9.5 Determining the Gain

Once the root locus plot has been sketched, if it appears that it passes through a region where it will have the desired response characteristics, then it will be necessary to determine the gain (k) value that corresponds to a point on the locus in that region. Fortunately, this is relatively easy.

Since points on the root locus satisfy

$$1 + kG(s) = 0$$

then

$$\begin{aligned} |k| &= \left|-\frac{1}{G(s)}\right| \\ &= \frac{1}{|G(s)|}. \end{aligned}$$

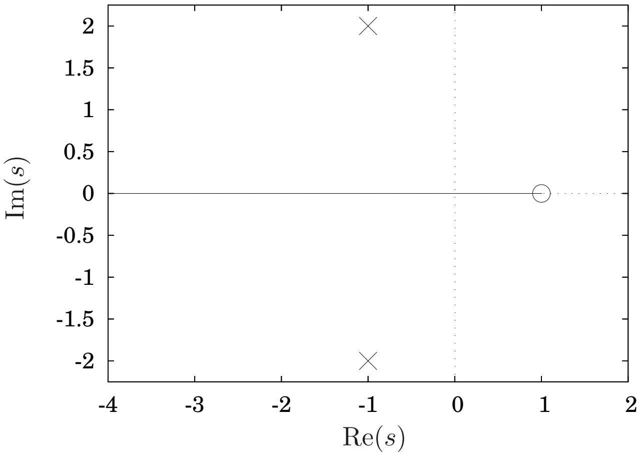


Figure 9.59. Partial root locus for Example 9.9.22.

.

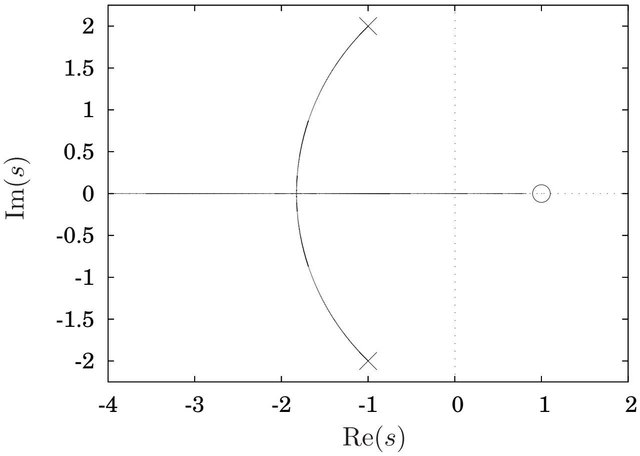


Figure 9.60. Root locus for Example 9.9.22.

.

Also, since

$$\begin{aligned}|G(s)| &= \left|\frac{\prod_{i=1}^{n_z}(s-z_i)}{\prod_{i=1}^{n_p}(s-p_i)}\right| \\ &= \frac{|s-z_1|\,|s-z_2|\cdots|s-z_{n_z}|}{|s-p_1|\,|s-p_2|\cdots|s-p_{n_p}|},\end{aligned}$$

we have

$$k = \frac{|s-p_1|\,|s-p_2|\cdots|s-p_{n_p}|}{|s-z_1|\,|s-z_2|\cdots|s-z_{n_z}|}. \tag{9.31}$$

In words, the value of k is simply the product of the distance from the point on the locus to all the poles divided by the product of the distance from the point on the locus to all the zeros.

Example 9.9.23 The root locus plot for

$$G(s) = \frac{s+2}{(s+1)(s+3)}$$ ■

is illustrated in Figure 9.58.

If we wish to determine the gain corresponding to the top and bottoms of the "circle" portion of the locus, we simply measure the distance from the two poles and zero to the point, as is illustrated in Figure 9.61 and substitute into Equation 9.31. The three distances are approximately, 2.4495, 1.7321 and 1.4142

$$\begin{aligned}k &\approx \frac{(2.4495)(1.7321)}{1.4142} \\ &\approx 3.\end{aligned}$$

9.9.6 Computational Tools

This section will discuss some very basic numerical methods to find roots of polynomials.

Newton's method

Matlab

The Matlab command to plot a root locus is `rlocus()`. As the reader probably expects, it takes the numerator and denominator of $G(s)$ as arguments and then plots the solutions to $1 + kG(s) = 0$ for $k \in [0)$

Example 9.9.24 To use Matlab to plot the root locus for the system in Example 9.9.16 where

$$G(s) = \frac{s+6}{(s+1)(s+3)}.$$ ■

enter

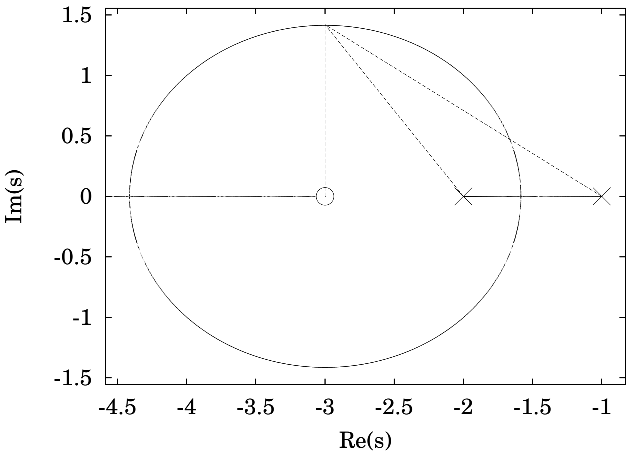


Figure 9.61. Measuring the distance from the poles and zeros of $G(s)$ to the point of interest to determine the gain for Example 9.9.23.

.

```
>> rlocus([1 6],conv([1 1],[1 3]))
```

or equivalently

```
>> rlocus([1 6],[1 4 3])
```

Octave

The Octave command to plot a root locus is `rlocus()`. As the reader probably expects, it takes the numerator and denominator of $G(s)$ as arguments and then plots the solutions to $1 + kG(s) = 0$ for $k \in [0)$

Example 9.9.25 To use Octave to plot the root locus for the system in Example 9.9.16 where

$$G(s) = \frac{s+6}{(s+1)(s+3)}.$$

■

enter

```
octave:> rlocus(tf([1 6],conv([1 1],[1 3])))
```

or equivalently

```
octave:> rlocus(tf([1 6],[1 4 3]))
```

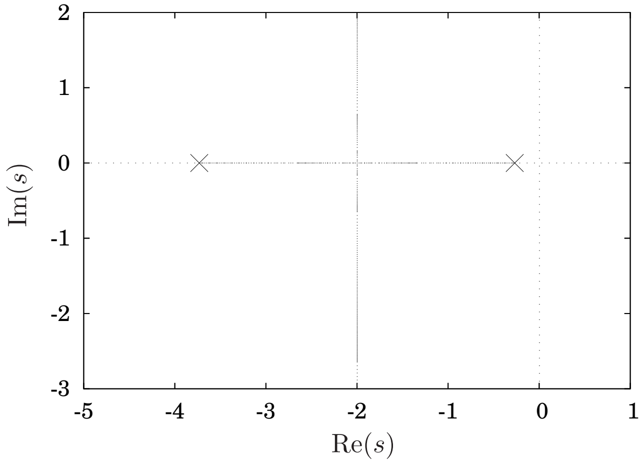


Figure 9.62. Root locus plot for the system from Example 9.10.1.

9.10 Controller Design Using the Root Locus Method

Root locus plots may be used to determine a "good" value for the feedback gain. Since all the tools necessary to accomplish this have already been developed, we will do this by way of a couple examples.

Example 9.10.1 Consider again the system illustrated in Figure 9.36 with the transfer function

$$\frac{X(s)}{\hat{F}(s)} = \frac{\frac{k}{m}}{s^2 + \frac{b}{m}s + \frac{k}{m}}.$$

If $m = k = 1$ and $b = 4$ then

$$\frac{X(s)}{\hat{F}(s)} = \frac{1}{s^2 + 4s + 1}.$$

The root locus plot is relatively simple. The open loop transfer function has two poles, one at $s \approx -3.73205$ and the other at $s \approx -0.26795$. On the real axis, the locus is between the two poles. The asymptote angles are $\pm 90°$ and the asymptotes intersect the real axis at $s = -2$. The break away point is at $s = -2$. The complete root locus plot is illustrated in Figure 9.62.

Focusing on the transient response for the moment, assume it is desired that the percentage overshoot be less than 10%. From Figure 9.24, the damping ratio must be greater than 0.6. Since $\sin^{-1} 0.6 \approx 37°$, we need that k be in the region between the lines of constant damping illustrated

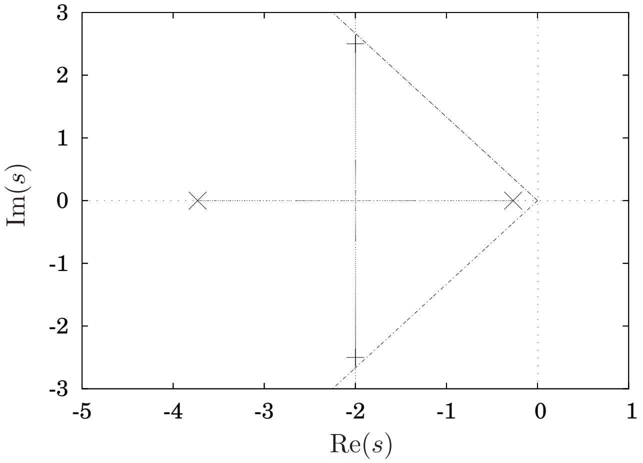


Figure 9.63. Pole locations which result in less than a 10% overshoot for Example 9.10.1.

in Figure 9.63. Picking the point $s = -2 \pm 2.5i$ to locate the poles of the closed loop transfer function, we may determine k from the distance from the two poles of the open loop transfer function. By Equation 9.31, we need to know the distance from all of the open loop poles and zeros to the desired pole location of the closed loop transfer function. Figure 9.64 indicates the two relevant distances, both of which are equal to $\sqrt{1.7^2 + 2.5^2} \approx 3.0414$. Hence we will use $k = 9.25$.

To verify the answer, we will compute the step response of the closed loop system using the computed gain. The closed loop transfer function is

$$\begin{aligned}\frac{Y(s)}{R(s)} &= \frac{k\frac{1}{s^2+4s+1}}{1+k\frac{1}{s^2+4s+1}} \\ &= \frac{k}{s^2+4s+1+k},\end{aligned}$$

and for $k = 4$,

$$\frac{Y(s)}{R(s)} = \frac{4}{s^2+4s+5}.$$

The step response is illustrated in Figure 9.65, and the overshoot appears to be slightly less than 10%.

Let us make the preceding example more difficult by adding a rise time specification to the problem as well.

Example 9.10.2 For the system in Example 9.10.1, in addition to requiring the closed loop step response to have less than a 10% overshoot, assume

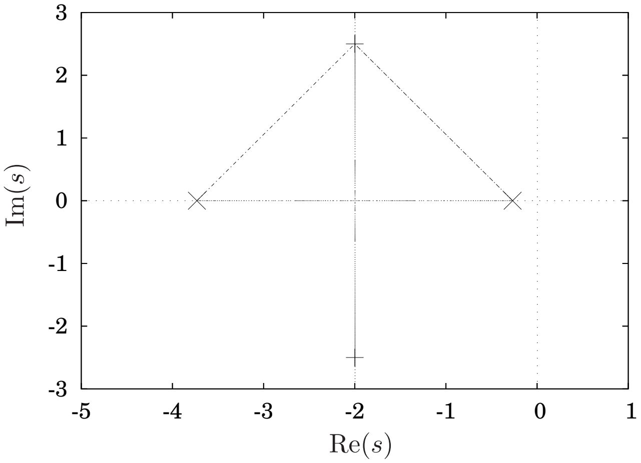


Figure 9.64. Distances to determine the gain for Example 9.10.1.

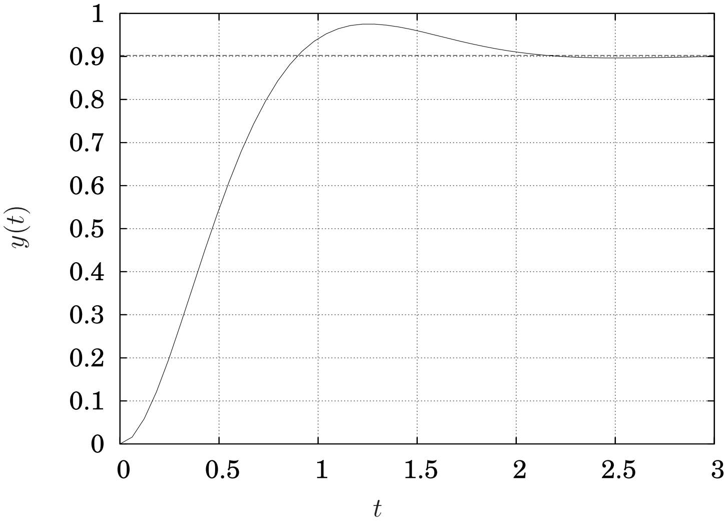


Figure 9.65. Step response demonstrating the desired overshoot for Example 9.10.1.

■

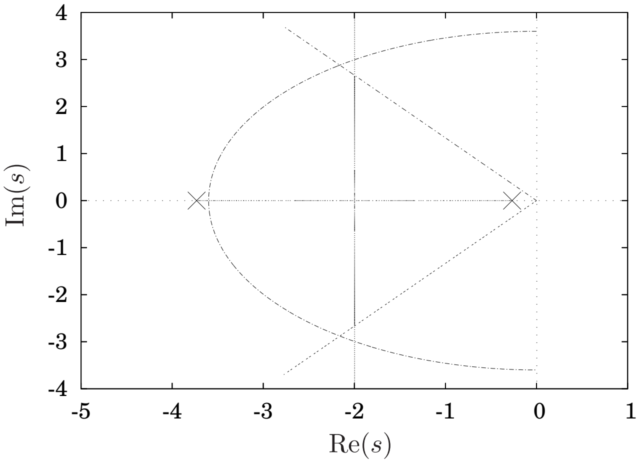


Figure 9.66. Complex plane regions satisfying the overshoot and rise time requirements for Example 9.10.2.

also that we desire the rise time to be less than 0.5 seconds. Use the approximation

$$t_r \approx \frac{1.8}{\omega_n}.$$

If we require

$$t_r \leq 0.5$$

then

$$\frac{1.8}{\omega_n} \leq 0.5 \qquad \Longrightarrow \qquad \omega_n \geq 3.6.$$

The region in the complex plane where the closed loop system will satisfy this requirement is outside the semi-circle illustrated in Figure 9.66. Also plotted are the lines corresponding to the damping ratio that satisfy the overshoot requirement.

Observe that it is impossible to choose a gain value that corresponds to a point on the root locus that is between the lines that indicate the overshoot specification and outside the semi-circle that indicates the rise time specification.

This is true even for the part of the root locus on the real axis. If you choose a point that is outside the semi-circle on the root locus on the real axis there is another pole on the branch that is coming from the other pole, which does not satisfy the specifications. For example, if we place a closed loop pole at $s = -3.65$, which seemingly satisfies the specification, this corresponds to a k value determined by

$$k \approx 0.2775.$$

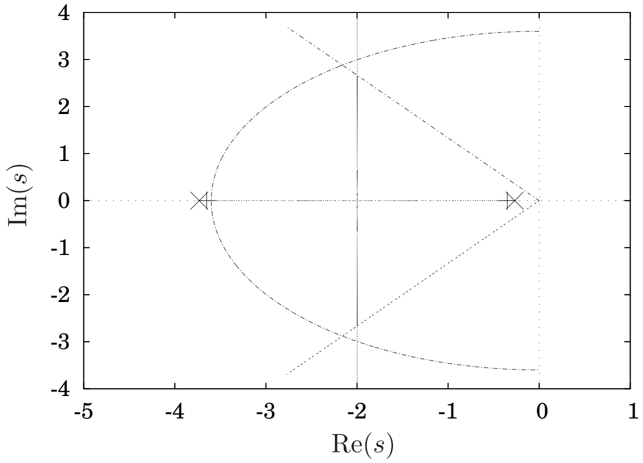


Figure 9.67. Closed-loop poles indicated by a + for $k = 0.2775$ for Example 9.10.2.

For $k = 0.2775$ the closed loop poles will be located at $s = -3.36$ and $s = -0.35$, as is illustrated in Figure 9.67, and the latter does not satisfy the rise time specification. Hoping it will anyway, we can compute the step response, which is illustrated in Figure 9.68. Clearly, it does not work. ■

The next example will consider how to use the root locus analysis to design a good controller to stabilize an unstable system.

Example 9.10.3 Consider the inverted pendulum illustrated in Figure 9.69. Assume the bar has a length, l, and is light with negligible inertia and that the mass moves under the influence of gravity and a torque, τ that is applied about the point of rotation of the pendulum. Assume that we require an overshoot less than 25% and a rise time less than 0.6 seconds.

Using Newton's second law for a planar system rotating about a point, the equation of motion is

$$ml^2\ddot{\theta} = mgl\sin\theta + \tau.$$

This is a nonlinear equation due to the $\sin\theta$ term. For small θ, $\sin\theta \approx \theta$, and making this approximation we have

$$ml^2\ddot{\theta} - mgl\theta = \tau.$$

For computational purposes, let $mgl = ml^2 = 1$. Using proportional feedback for τ, the transfer function from a specified desired angle, $\Theta_d(s)$ to the actual angle, $\Theta(s)$ is given illustrated in Figure 9.70 where the controller, $C(s) = k_p$.

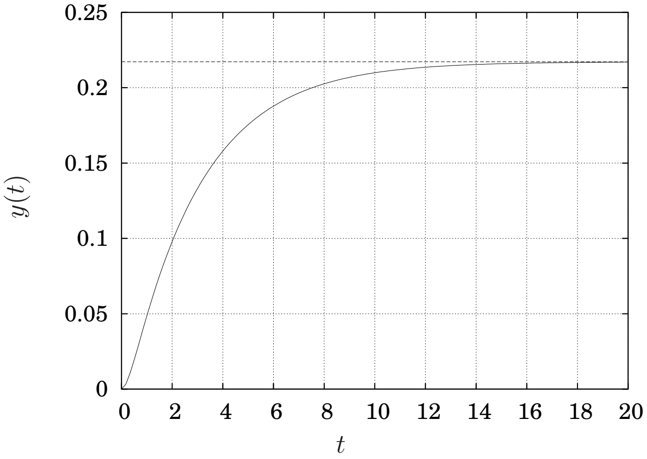


Figure 9.68. Closed-loop step response for $k = 0.2775$ for Example 9.10.2.

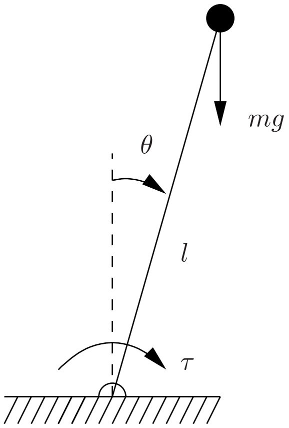


Figure 9.69. Inverted pendulum system for Example 9.10.3.

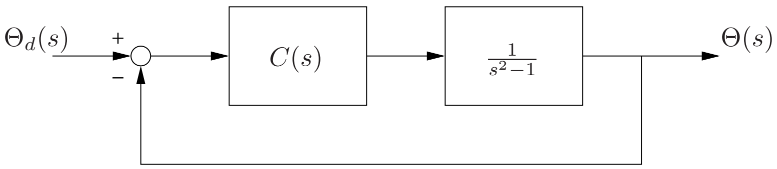


Figure 9.70. Block diagram for feedback control of the inverted pendulum in Example 9.10.3.

The root locus plot for the open loop transfer function

$$G(s) = \frac{1}{s^2 - 1}$$

is illustrated in Figure 9.71. From the root locus plot we can conclude that for $k_p \leq 1$, the system will be unstable and for $k_p > 1$ the system will be marginally stable. In other words, the linearized equation will have non-decaying sinusoidal solutions. The step responses of the linearized system with $k_p = 0.1$, $k_p = 1$ and $k_p = 2$ are illustrated in Figure 9.72.

In this case it will be impossible to meet the overshoot specification. If $k_p \leq 1$ the system is unstable and for $k_p > 1$ there is zero damping, independent of k_p.

For any real engineering system, predicting an exactly marginally stable response is impossible since any modeling errors will keep the system from either behaving in either an exactly linear manner or, for that matter, being exactly on the imaginary axis.

The obvious thing to do to add some extra stability, and hence to pull the branches of the root locus to the left, is to add some derivative control. If we specify

$$C(s) = k\left(\frac{1}{2}s + 1\right)$$

which fixes $k_d = \frac{1}{2}k_p$, then the open loop transfer function is

$$G(s) = \frac{\frac{1}{2}s + 1}{s^2 - 1}$$

is illustrated in Figure 9.73.

The regions in the complex plane where the overshoot and rise time specifications are met are illustrated in Figure 9.74. From Figure 9.24, an overshoot of less than 25% corresponds to a damping ratio of greater than 0.4, which corresponds to a pole location at an angle of $\sin^{-1}(0.4) \approx 25°$. Using the relationship

$$t_r \approx \frac{9}{5\omega_n}$$

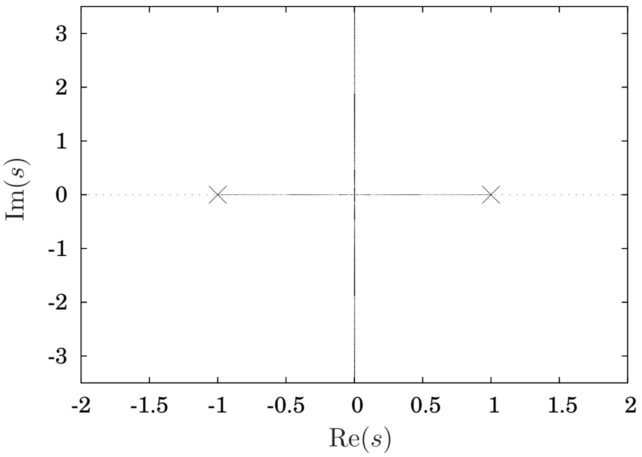


Figure 9.71. Root locus plot for linearized inverted pendulum for Example 9.10.3.

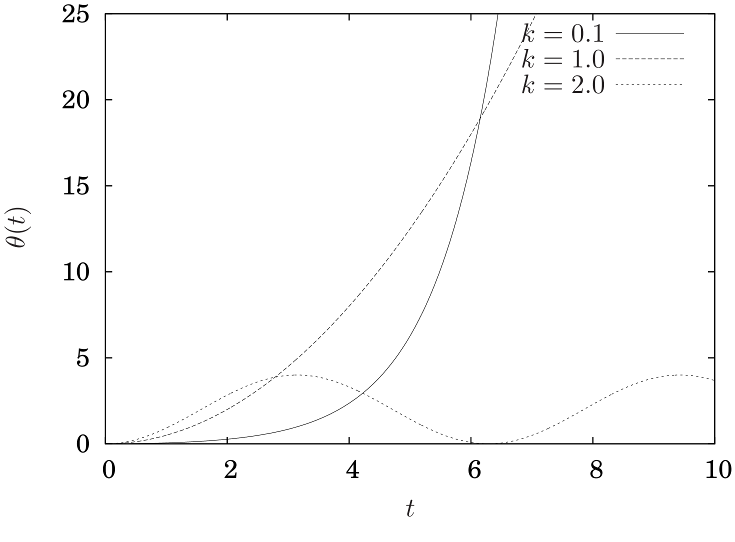


Figure 9.72. Step responses for various proportional gains for unity feedback for the linearized inverted pendulum in Example 9.10.3.

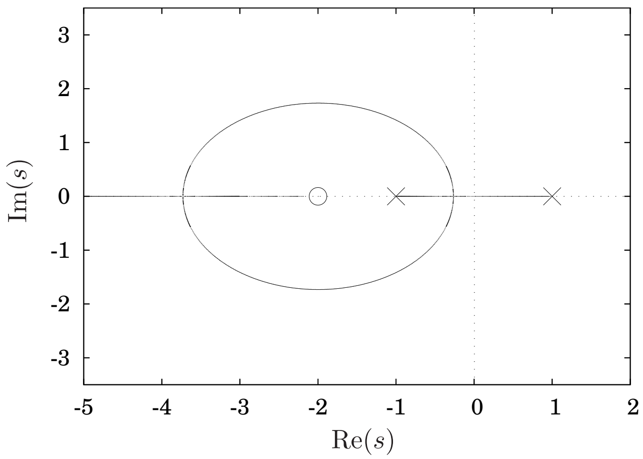


Figure 9.73. Root locus plot for linearized inverted pendulum for Example 9.10.3 with PD control.

a rise time less than 0.6 seconds requires a natural frequency greater than 3. Since the closed-loop poles start at the open loop poles, an analysis of the root locus plot shows that any gain that meets the rise time specification will also meet the overshoot specification. Using a rough approximation, if we desire to place the poles at $s \approx -3.5 \pm i$ then

$$\begin{aligned} k &\approx \frac{\sqrt{4.5^2+1^2}\sqrt{2.5^2+1^2}}{\sqrt{1.5^2+1^2}} \\ &= 12. \end{aligned}$$

A plot of the closed loop poles with $k = 12$ is illustrated in Figure 9.75. The closed loop step response for $k = 12$ is illustrated in Figure 9.76. ■

9.11 Frequency Response Analysis and Design

Frequency-response analysis of a system focuses upon analyzing the relationship between the input and output of a transfer function when the input is a purely sinusoidal signal. If an input to a transfer function is a pure sinusoid, $r(t) = \sin \omega t$ the output will be a sinusoid of the same frequency (see Exercise 9.14), but with a magnitude and phase shift that depend on the frequency of the input.

Example 9.11.1 Consider

$$\frac{R(s)}{Y(s)} = \frac{2}{s+2}$$

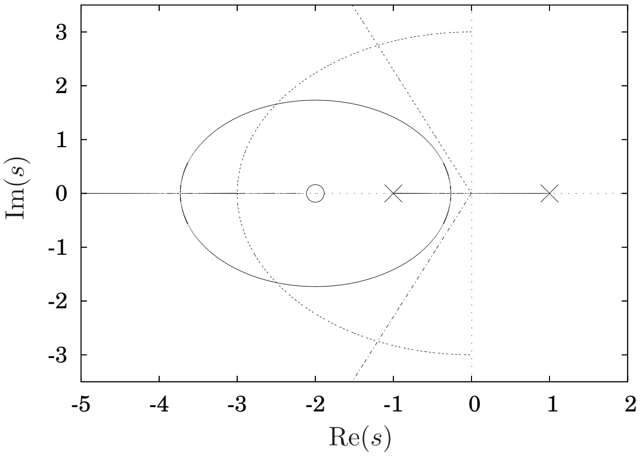


Figure 9.74. Pole locations satisfying the overshoot and rise time specifications for Example 9.10.3.

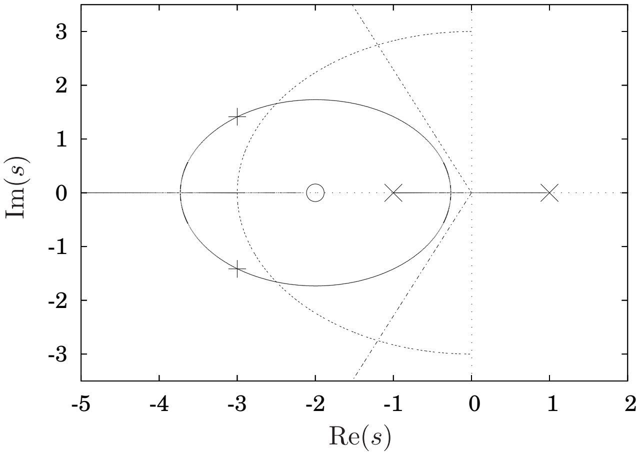


Figure 9.75. Closed loop pole locations for PD control with $k = 12$ for Example 9.10.3.

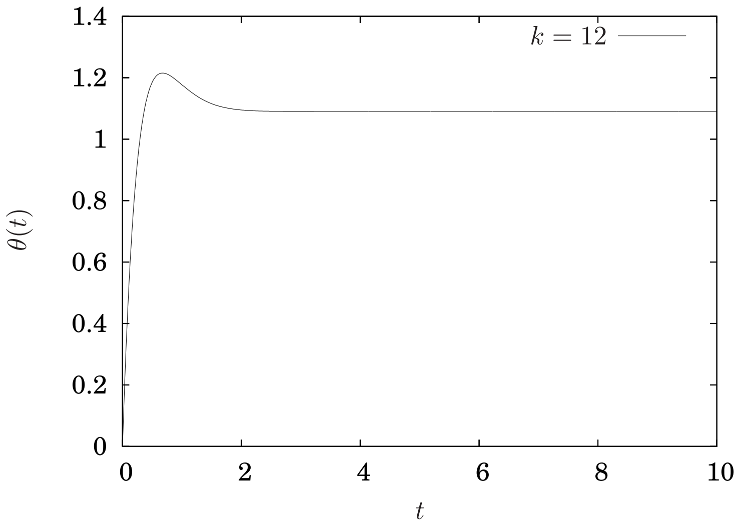


Figure 9.76. Closed loop step response with $k = 12$ and PD control for the linearized inverted pendulum in Example 9.10.3.

and two input signals

$$\begin{aligned} r_1(t) &= \sin t \\ r_2(t) &= \sin 3t, \end{aligned}$$

or

$$\begin{aligned} R_1(s) &= \frac{1}{s^2+1} \\ R_2(s) &= \frac{3}{s^2+9}. \end{aligned}$$

Solving either

$$\begin{aligned} y_1(t) &= \mathcal{L}^{-1}\left(\frac{2}{s+2}\frac{1}{s^2+1}\right) \\ y_2(t) &= \mathcal{L}^{-1}\left(\frac{2}{s+2}\frac{3}{s^2+9}\right) \end{aligned}$$

or

$$\begin{aligned} \dot{y}_1 + 2y &= 2\sin t \\ \dot{y}_2 + 2y &= 2\sin 3t \end{aligned}$$

(the latter with zero initial conditions) gives

$$\begin{aligned} y_1(t) &= \frac{2}{5}\left(e^{-2t} + 2\sin t - \cos t\right) \\ y_2(t) &= \frac{2}{13}\left(3e^{-2t} + 2\sin 3t - 3\cos 3t\right). \end{aligned}$$

Hence, for large t, the steady state solutions are

$$\begin{aligned} y_{1,ss}(t) &= \frac{2}{5}\left(2\sin t - \cos t\right) \\ y_{2,ss}(t) &= \frac{2}{13}\left(2\sin 3t - 3\cos 3t\right), \end{aligned}$$

or using the relationship

$$\sin\left(\omega t + \phi\right) = \cos\phi \sin\omega t + \sin\phi\cos\omega t$$

$$\begin{aligned} y_{1,ss}(t) &= \frac{2}{5}\sqrt{5}\sin\left(t + \phi_1\right), \quad \phi_1 = \tan^{-1}\frac{-1}{2} \\ y_{2,ss}(t) &= \frac{2}{13}\sqrt{13}\sin\left(3t + \phi_2\right), \quad \phi_2 = \tan^{-1}\frac{-3}{2}. \end{aligned}$$

Observe that the steady state solution is a sinusoid of the same frequency as the input, but the magnitude is scaled and there may be a phase shift. Foreshadowing what is to come, note that if we substitute $s = i\omega$ into $G(s)$, we get, for each case respectively,

$$\begin{aligned} G(i) &= \frac{2}{i+2} = \frac{4-2i}{5} \\ G(3i) &= \frac{2}{3i+2} = \frac{4-6i}{13}. \end{aligned}$$

The magnitude of these are

$$\begin{aligned} |G(i)| &= \frac{1}{5}\sqrt{20} = \frac{2}{\sqrt{5}} \\ |G(3i)| &= \frac{1}{13}\sqrt{52} = \frac{2}{\sqrt{13}}. \end{aligned}$$

So, it appears that if we simply substitute $s = i\omega$ into $G(s)$ and determine its magnitude, it gives the magnitude of the response.

Similarly, if we compute

$$\begin{aligned} \angle G(i) &= \tan^{-1}\left(\frac{-2}{4}\right) = \tan^{-1}\left(\frac{-1}{2}\right) \\ \angle G(3i) &= \tan^{-1}\left(\frac{-6}{4}\right) = \tan^{-1}\left(\frac{-3}{2}\right), \end{aligned}$$

it appears that the phase shift in the steady state response is given by the angle of $G(i\omega)$.

In fact, both of these are true in general, which is what we will prove next. ■

Next we will show that if a transfer function is stable, then if the input is $A\sin\omega t$ then the steady state solution will be scaled by $|G(i\omega)|$ and have a phase shift of $\phi = \tan^{-1}\left(\frac{\mathrm{Im}(i\omega)}{\mathrm{Re}(i\omega)}\right)$.

Proposition 9.11.2 *If all the poles of $G(s)$ are in the left half plane, then if*

$$y(t) = \mathcal{L}^{-1}\left(G(s)\frac{A\omega}{s^2+\omega^2}\right)$$

as t becomes large, the steady state solution is given by

$$y_{ss}(t) = A\left|G(i\omega)\right|\sin(\omega t + \phi)$$

where

$$\phi = \tan^{-1}\left(\frac{\mathrm{Im}(i\omega)}{\mathrm{Re}(i\omega)}\right).$$

Proof Let

$$G(s) = \frac{N(s)}{D(s)}$$

then one form of a partial fraction expansion will be

$$\begin{aligned} G(s)\frac{A\omega}{s^2+\omega^2} &= \frac{N(s)}{D(s)}\frac{A\omega}{s^2+\omega^2} \\ &= \frac{C_1(s)}{D(s)} + \frac{A\omega\left(c_1 s + c_2\right)}{s^2+\omega^2}. \end{aligned}$$

Using the method from Appendix A.3, to determine $C_2(s)$, multiply both sides of this equation by $\left(s^2+\omega^2\right)$ and take the limit as $s \to i\omega$, *i.e.*,

$$\lim_{s\to i\omega}\left(G(s)\frac{A\omega}{s^2+\omega^2}\left(s^2+\omega^2\right)\right) = \lim_{s\to i\omega}\left(\frac{AC_1(s)}{D(s)}\left(s^2+\omega^2\right) + \frac{A\omega\left(c_1 s + c_2\right)}{s^2+\omega^2}\left(s^2+\omega^2\right)\right)$$

which gives

$$c_1 i\omega + c_2 = G(i\omega),$$

so

$$\begin{aligned} c_1 &= \frac{1}{\omega}\mathrm{Im}\left(G(i\omega)\right) \\ c_2 &= \mathrm{Re}\left(G(i\omega)\right). \end{aligned}$$

Referring to Table 8.1, the c_1 term corresponds to the cosine component in the steady state solution and the c_2 term corresponds to the sine component.

Hence,

$$\begin{aligned} y_{ss}(t) &= A\left(\mathrm{Re}\left(G(i\omega)\right)\sin\omega t + \mathrm{Im}\left(G(i\omega)\right)\cos\omega t\right) \\ &= \sqrt{\left[\mathrm{Re}\left((G(i\omega)\right)\right]^2 + \left[\mathrm{Im}\left((G(i\omega)\right)\right]^2}\sin(\omega t + \phi) \\ &= \left|G(i\omega)\right|\sin(\omega t + \phi) \end{aligned}$$

where

$$\phi = \tan^{-1}\left(\frac{\mathrm{Im}\left(G\left(i\omega\right)\right)}{\mathrm{Re}\left(G\left(i\omega\right)\right)}\right).$$

□

It turns out that

1. it is relatively easy to sketch by hand $|G\left(i\omega\right)|$ and $\phi = \tan^{-1}\left(\frac{\mathrm{Im}(G(i\omega))}{\mathrm{Re}(G(i\omega))}\right)$, so it is not too difficult to obtain information about the steady state response of the system to sinusoidal inputs; and,
2. more importantly very useful information regarding the stability of a system under unity feedback and information on designing feedback controllers may be obtained by graphs of the magnitude and phase of the steady state response to a sinusoidal input.

This type of analysis is referred to as a *frequency response analysis* and is common in control theory, particularly in electrical engineering.

9.11.1 Bode Plots

A *Bode plot* is a log-log plot of the magnitude and angle of $G(i\omega)$ versus ω. It is conventional to plot the magnitude on a log scale and it is conventional to do so in decibels. For our purposes definition of a decibel is

$$\left|G\left(i\omega\right)\right|_{dB} = 20\log_{10}\left|G\left(i\omega\right)\right|.$$

Since the plot is on a logarithmic scale, we may construct the plot for individual components of a transfer function and them add them together to construct the overall plot. In order to do this, we need to have handy the plot for the typical components of a transfer function.

Example 9.11.3 Consider

$$G(s) = \frac{10}{s+10}.$$

A computer-generated Bode plot of $G(s)$ is illustrated in Figure 9.77.

Let us first rewrite $G\left(i\omega\right)$ in a way that will help with our analysis subsequently:

$$\begin{aligned} G\left(i\omega\right) &= \frac{10}{i\omega+10} \\ &= \frac{10}{i\omega+10}\frac{10-i\omega}{10-i\omega} \\ &= \frac{100-10i\omega}{100+\omega^2}. \end{aligned}$$

For $\omega \ll 10$,

$$G\left(i\omega\right) \approx 1,$$

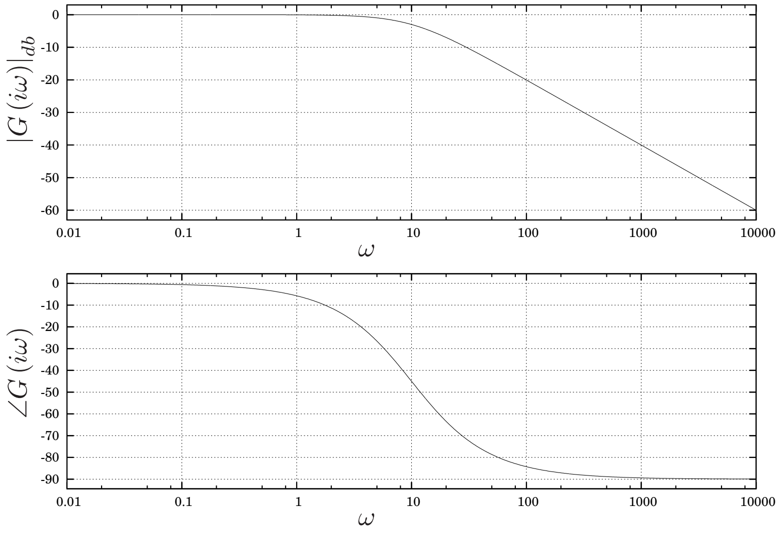


Figure 9.77. Bode plot for $G(s) = \frac{10}{s+10}$.

so

$$|G(i\omega)| \approx 1 = 0db$$

and

$$\angle G(i\omega) \approx 0^\circ.$$

For small frequencies, the Bode plot in Figure 9.77 corresponds to this.

For $\omega \gg 10$

$$G(i\omega) \approx -\frac{10i}{\omega}$$

so

$$|G(i\omega)| \approx \frac{10}{\omega}.$$

As ω increases, $|G(i\omega)|$ decreases, and in particular whenever ω increases by a factor of 10, $|G(i\omega)|$ decreases by a factor of 10. Since a decrease by a factor of 10 corresponds to a decrease of 20 dB, the slope of the magnitude curve at high ω should be -20 dB/decade (-20 dB for every increase in order of magnitude of ω). Also, for $\omega \gg 10$, $G(i\omega)$ is almost a purely negative imaginary number, so

$$\angle G(i\omega) \approx -90^\circ.$$

For large frequencies, the Bode plot in Figure 9.77 corresponds to this. ■

The beauty of logarithms is that multiplication is reduced to addition. We may exploit this fact when sketching Bode plots by sketching each term in a transfer function individually and then adding them.

Example 9.11.4 Sketch the Bode plot for

$$G(s) = \frac{100}{s\,(10s+1)}.$$

Explicitly writing all three terms we have

$$G(s) = 100 \cdot \frac{1}{s} \cdot \frac{1}{10s+1}$$

so

$$\begin{aligned} |G\,(i\omega)| &= \left| 100 \cdot \frac{1}{i\omega} \cdot \frac{1}{10\omega i + 1} \right| \\ &= |100| \cdot \left| \frac{1}{i\omega} \right| \cdot \left| \frac{1}{10\omega i + 1} \right| \end{aligned}$$

or, in decibels and making use of the fact that the logarithm of a product is the sum of the logarithms

$$|G\,(i\omega)|_{dB} = |100|_{dB} + \left| \frac{1}{i\omega} \right|_{dB} + \left| \frac{1}{10\omega i + 1} \right|_{dB}. \qquad (9.32)$$

Similarly, because in polar coordinates the angle of complex numbers add when complex numbers are multiplied we have

$$\begin{aligned} \angle G\,(i\omega) &= \angle \left(100 \cdot \frac{1}{i\omega} \cdot \frac{1}{10\omega i + 10} \right) \\ &= \angle 100 + \angle \frac{1}{i\omega} + \angle \frac{1}{10\omega i + 1}. \end{aligned} \qquad (9.33)$$

■

Considering each term in Equation 9.32.

9.11.2 Gain and Phase Margins

9.12 Controller Design using Bode Plots

9.13 Exercises

Problem 9.1 Find the transfer function for the circuit illustrated in Figure 9.78.

Problem 9.2 Show that if the transfer function has multiple poles at the origin, *i.e.*,

$$Y(s) = \frac{1}{s^n} R(s)$$

then *regardless of the input*, $y(t)$ will contain an $n-1$th order polynomial in t, *i.e.*,

$$y(t) = c_0 + c_1 t + \frac{c_2}{2} t^2 + \cdots + \frac{c_{n-1}}{(n-1)!} t^{n-1} + \text{other terms}.$$

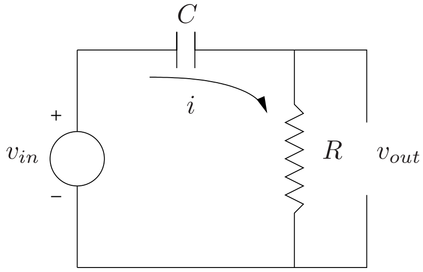


Figure 9.78. Circuit for problem 9.1.

Problem 9.3 Write a computer program that determines an approximate numerical solution to the equations of motion for the robot arm from Example 9.2.1. Assuming zero initial conditions and a small desired angle, use your program to verify the following "rules of thumb" for PID control for a step input.

Your program should be for the original nonlinear model, not the linearized one that we can solve analytically. The idea is to verify that what we determined analytically using the linearized version works for the nonlinear case as well as long as the desired angle of the robot arm is small.

1. For proportional control, i.e., $k_p > 0$, $k_d = 0$ and $k_i = 0$, the solutions are oscillatory, and increasing kp increases the frequency of oscillation (which decreases the rise time and peak time) but decreases the mean steady state error. The settling time is infinite. Hint pick a starting value of $k_p = 5$.

2. If derivative control is added to the proportional controller, *i.e.,* $k_p > 0$, $k_d > 0$ and $k_i = 0$, then

   (a) for small k_d the solutions are decaying oscillations;

   (b) increasing k_d decreases the settling time;

   (c) increasing k_d sufficiently eliminates the oscillatory behavior completely, resulting in an solution which exponentially decays to the final, steady state value;

   (d) increasing k_p decreases the final steady state error;

   (e) increasing k_p decreases the rise time.

   Hint pick a starting value of about $k_d = 0.5$.

3. Adding integral control (PID control)

   (a) eliminates the steady state error, even for small values of k_p,

   (b) increasing k_i generally increases the overshoot and settling time;

(c) increasing k_p decreases rise time, but may increase overshoot;

(d) increasing k_d increases damping and stability.

Hint: pick a starting value of about $k_i = 0.5$.

4. Choose a set of gain values from the above simulations that seems to work well. Use those for an attempt to have the desired angle be large. Does it still work well?

Problem 9.4 Verify the results in Figure 9.21 by using a computer to compute a numerical solution to the step response to

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

and by appropriately varying ω_n and ζ so that the poles move in the three directions indicated in the figure. Submit plots illustrating the pole locations and corresponding step responses and whether or not the change in the step response when the pole is moved in one of the three directions indicated are as predicted.

Problem 9.5 Plots of the poles and zeros of different transfer functions appear in Table 9.2. Match the plots of the pole and zero zero locations with the corresponding step responses in Table 9.3.

Problem 9.6 The step response of

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}.$$

is given by Equation 9.18. Now consider

$$G(s) = \frac{\frac{\omega_n^2}{r}(s + r)}{s^2 + 2\zeta\omega_n s + \omega_n^2}.$$

Compute the partial fraction expansion of the step response of $G(s)$, and using the resulting time function, explain why the rules for an additional real zero added to a second order system are true.

Problem 9.7 If $s_1 = a_1 + ib_1$ and $s_2 = a_2 + ib_2$, and

$$s_1 s_2 = (a_1 a_2 - b_1 b_2) + i\,(a_1 b_2 + a_2 b_1)$$

use the fact that

$$r = \sqrt{a^2 + b^2}$$

and

$$\theta = \tan^{-1}\left(\frac{b}{a}\right)$$

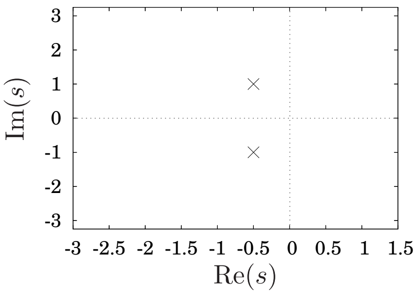
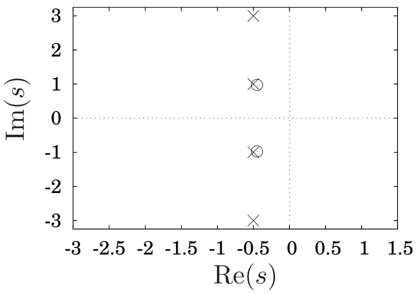
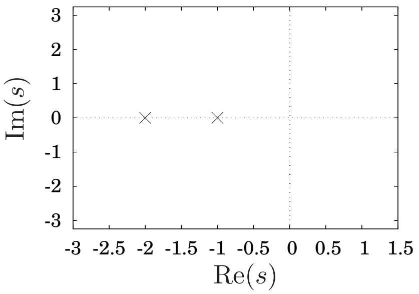
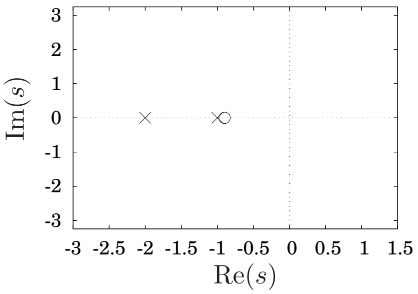
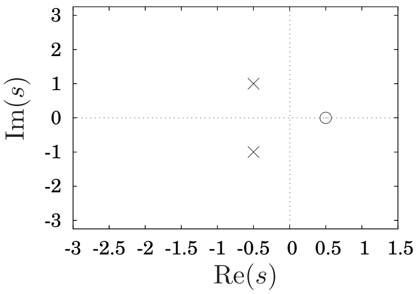
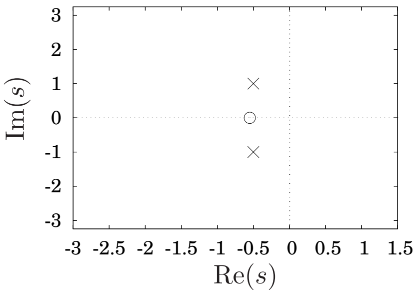
	A	B
1		
2		
3		

Table 9.2. Pole-zero maps for Problem 9.5.

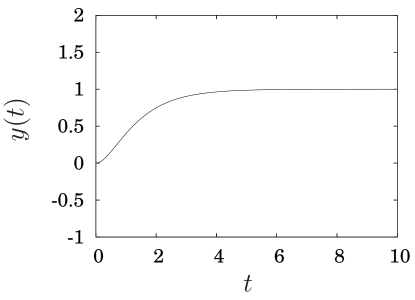
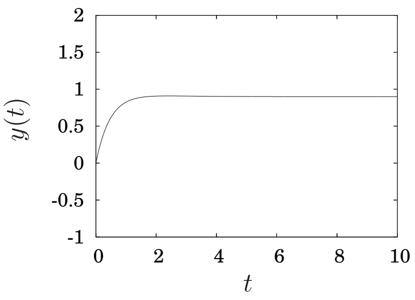
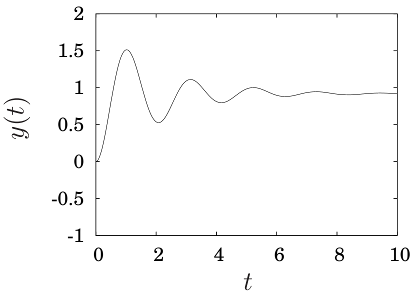
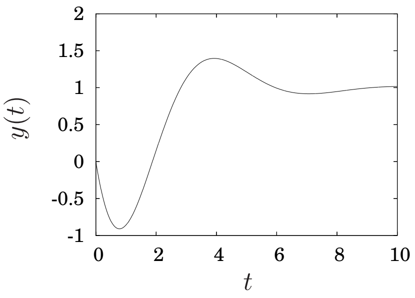
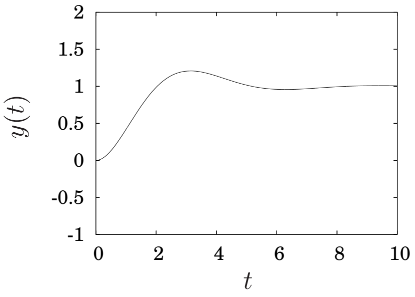
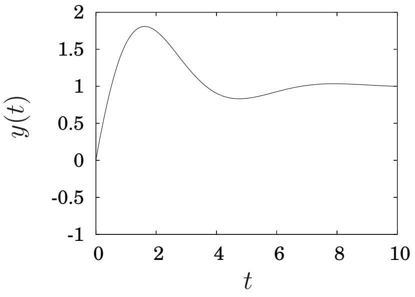
	A	B
1		
2		
3		

Table 9.3. Step responses for Problem 9.5.

to show that in polar coordinates

$$s_1 s_2 = (r_1 r_2, \theta_1 + \theta_2)$$

and

$$\frac{s_1}{s_2} = \left(\frac{r_1}{r_2}, \theta_1 - \theta_2\right).$$

Problem 9.8 The root locus plots we considered in Section 9.9 considered only the case where $k \in [0, +\infty)$ and is often called the 180° root locus. Determine each of the rules (Rule 9.9.5 through Rule 9.9.17) for the case where $k \in (-\infty, 0]$, called the 0° root locus.

Problem 9.9 Consider a DC motor connected to the circuit illustrated in Figure 9.79. Assume that the shaft of the motor has a moment of inertia J.

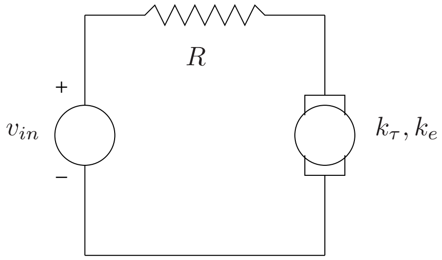


Figure 9.79. DC motor for Problem 9.9.

1. If the block diagram in Figure 9.80, represents this system, determine $G(s)$.

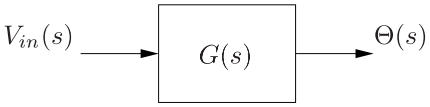


Figure 9.80. Block diagram for d.c. motor in Problem 9.9.

2. Consider the block diagram illustrated in Figure 9.81. Determine the transfer function, $C(s)$ in the controller block for
   (a) proportional control;
   (b) proportional plus derivative control;
   (c) proportional plus integral control; and

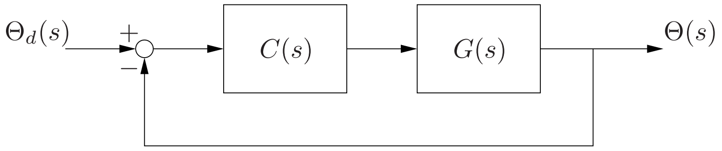


Figure 9.81. Feedback control loop for Problem 9.9.

(d) proportional plus derivative plus integral control.

3. Determine the transfer function from the desired angular position of the motor to the actual position, $\frac{\Theta_d(s)}{\Theta(s)}$ (do not substitute for $C(s)$ or $G(s)$).

4. If $\omega = \dot{\theta}$, determine the transfer function $\frac{\Omega(s)}{\Omega_d(s)}$.

5. If

$$\begin{aligned} k_e &= 1 \\ k_\tau &= 2 \\ R &= 3 \\ J &= 4 \end{aligned}$$

and we use proportional control, use the root locus plotting rules to sketch, by hand, the how the poles of $\frac{\Theta(s)}{\Theta_d(s)}$ vary as the proportional gain is varied from 0 to $+\infty$. Determine the approximate gain value, if any, that gives a damping ratio of approximately $\frac{1}{2}$.

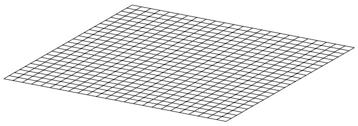
6. For the same parameter values as above, use PD control and fix the ratio between the proportional gain and the derivative gain to be $\frac{1}{2}$, *i.e.,*

$$\begin{aligned} v_{in} &= k_p\left(\theta_d - \theta\right) + k_d\left(\dot{\theta}_d - \dot{\theta}\right) \\ &= k\left[\left(\theta_d - \theta\right) + \frac{1}{2}\left(\dot{\theta}_d - \dot{\theta}\right)\right], \end{aligned}$$

sketch the root locus plot for the system. Discuss qualitatively what will happen to the rise time, the percentage overshoot and the settling time as k increases.

Problem 9.10 Consider

$$G(s) = \frac{4}{(s+1)(s+3)}.$$



1. Sketch the root locus plot for this transfer function
2. If this transfer function is placed in a feedback loop as illustrated in Figure 9.81, with $C(s) = k$, what will happen to the overshoot of the step response as k gets large? Explain your answer.
3. Determine the maximum value for k so that the percentage overshoot remains under 20%.

Problem 9.11 Consider

$$G(s) = \frac{s+5}{(s+1)(s+3)}.$$

1. Sketch the root locus plot for this transfer function
2. If this transfer function is placed in a feedback loop as illustrated in Figure 9.81, with $C(s) = k$, what will happen to the overshoot of the step response as k gets large? Explain your answer.

Problem 9.12 Consider

$$G(s) = \frac{1}{(s+1)(s+3)(s+5)}.$$

1. Sketch the root locus plot for this transfer function
2. What can you say about the stability of the response of the system under unity feedback as k gets large?

Problem 9.13 Prove Corollary 9.7.3.

Problem 9.14 Prove that if all the poles of a transfer function are in the left half plane, then the steady state response of the system with a sinusoidal input, $r(t) = \sin \omega t$ will be a sinusoid with the same frequency, but with a possibly a different magnitude and a phase shift, *i.e.*, $y(t) = m \sin(\omega t + \phi)$.

Problem 9.15 A minor complication occurs if a transfer function has two more poles or zeros at the same location. The root locus plot for

$$G(s) = \frac{s+3}{s^2(s+2)},$$

which has a double pole at the origin is illustrated in Figure 9.82 and the root locus for

$$G(s) = \frac{s+3}{s(s+2)^2},$$

which has a double pole at $s = -2$ is illustrated in Figure 9.83. Multiple poles at the same location are often, but not always, represented by slightly offset ×'s to indicate that there is more than one pole of $G(s)$ at that location.

Do all the rules summarized in Table 9.1 still apply? Explain your answer for each of the rules by specifically referring to the features of Figures 9.82 and 9.83.

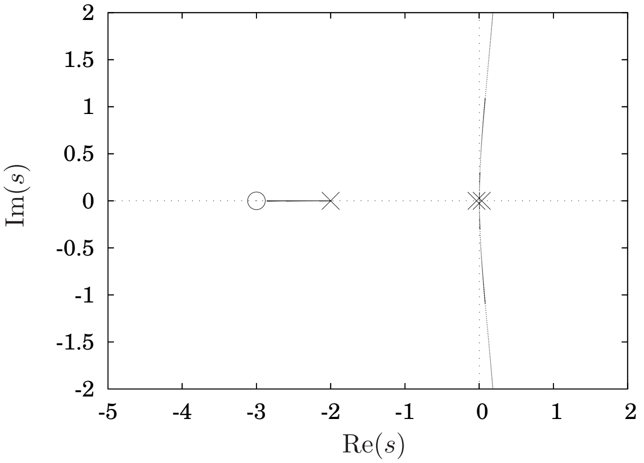


Figure 9.82. Root locus plot for $G(s) = \frac{s+3}{s^2(s+2)}$ for Problem 9.15.

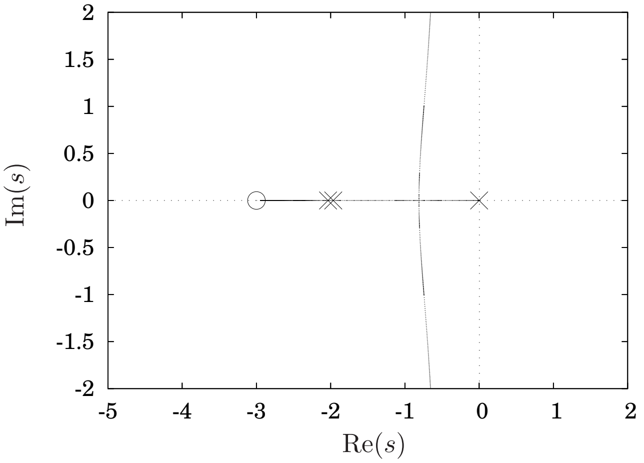


Figure 9.83. Root locus plot for $G(s) = \frac{s+3}{s(s+2)^2}$ for Problem 9.15.

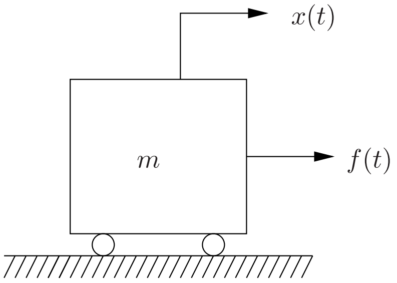


Figure 9.84. System for Problem 9.16.

Problem 9.16 In order to do this problem, you must understand how to deal with multiple poles in the same location, which was considered in Problem 9.15.

Consider the system illustrated in Figure 9.84.

1. Determine the transfer function from the applied force, $f(t)$ to the position of the mass, $x(t)$.
2. Sketch the root locus plot for this transfer functions if $m = 1$.
3. What does the root locus plot tell you about using proportional control to control the position of the mass? Specifically,
   (a) will it be stable, unstable or on the margin;
   (b) for the step response, by changing the proportional gain can you affect
      i. the rise time;
      ii. the settling time;
      iii. the percent overshoot?

Problem 9.17 Consider again the system illustrated in Figure 9.84.

1. Determine the transfer function from the applied force, $f(t)$ to the velocity of the mass, $\dot{x}(t)$.
2. Sketch the root locus plot for this transfer function.
3. Discuss the use of proportional control for this system. What characteristics of the response of the system can you affect by altering the proportional gain?

Problem 9.18 Sketch the Bode plot for

$$G(s) = \frac{100}{(s+10)(s+100)}.$$

Problem 9.19 Sketch the Bode plot for

$$G(s) = \frac{100}{(s+10)(s+100)(s+1000)}.$$

Problem 9.20 Sketch the Bode plot for

$$G(s) = \frac{s}{(s+10)(s+100)(s+1000)}.$$

Problem 9.21 Sketch the Bode plot for

$$G(s) = \frac{s+100}{(s+10)(s+10000)}.$$

Problem 9.22 Consider the low pass filter illustrated in Figure 9.85. Determine the transfer function from the input voltage to the output voltage. Sketch the Bode plot for this circuit if $R_{HP} = 100$ and $C_{HP} = 100$.

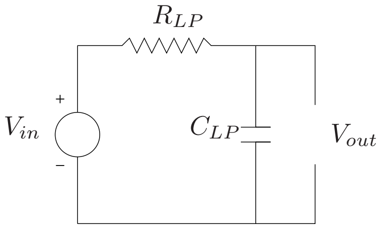


Figure 9.85. Low pass filter for Problem 9.22.

Problem 9.23 Consider the high pass filter illustrated in Figure 9.86. Determine the transfer function from the input voltage to the output voltage. Sketch the Bode plot for this circuit if $R_{HP} = 10$ and $C_{HP} = 10$.

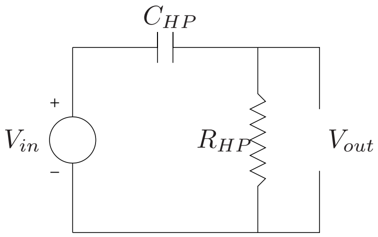
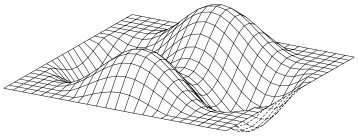


Figure 9.86. High pass filter for Problem 9.23.



Problem 9.24 Consider connecting a low pass filter and high pass filter together in series, as is illustrated in Figure 9.87.

1. Determine the transfer function from the input voltage to the output voltage. (Hint: if you did Problems 9.22 and 9.23, this is trivial).

2. If

$$\begin{aligned} R_{HP} &= 10 \\ R_{LP} &= 100 \\ C_{HP} &= 10 \\ C_{LP} &= 100 \end{aligned}$$

   sketch the Bode plot for this transfer function.

3. How would you modify the circuit to make the band pass region either narrower or wider? Do so and either sketch or use a computer package to generate the Bode diagram.

4. How would you make the transitions in the band pass filter sharper, *i.e.,* steeper transitions? Do so and either sketch or use a computer package to generate the Bode diagram.

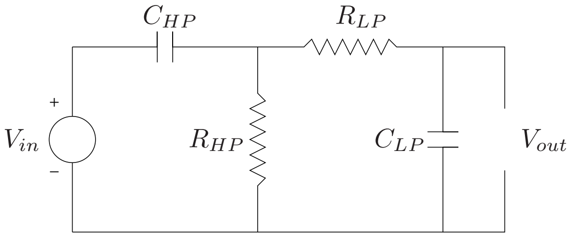


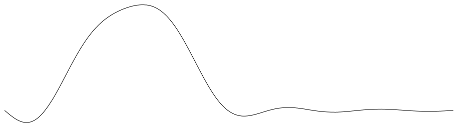
Figure 9.87. Band pass filter for Problem 9.24.

Problem 9.25 Consider

$$G(s) = \frac{1}{s\,(s+2)\,(s+4)}.$$

1. Sketch the root locus plot for this transfer function.

2. Determine the closed loop transfer function, *i.e.,* $\frac{Y(s)}{R(s)}$ if $G(s)$ is in the block diagram illustrated in Figure 9.88. Use the Routh array to determine the values for k for which the closed loop transfer function is stable.



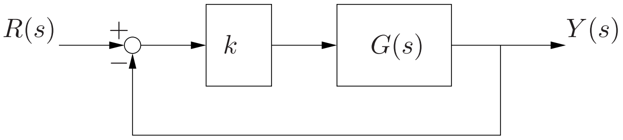


Figure 9.88. Closed-loop system for Problem 9.25.

3. Verify your computation from the previous step by using the root locus plot to determine the values for k for which the closed loop transfer function is stable.

4. Sketch the Bode diagram for gain values much smaller, equal to and much larger than the gain values you determined in the previous steps and determine the gain and phase margins in each case.

Problem 9.26 Sketch the root locus plot for

$$\frac{Y(s)}{R(s)} = \frac{kG(s)}{1 + kG(s)}$$

where

$$G(s) = \frac{1}{s^2 + 4s + 5}.$$

1. Indicate on your root locus plot the region on the complex plan where the the maximum percent overshoot for the step response for a complex conjugate pair of poles is less than 16%. Label any angles that you use in this determination.

2. Compute and then indicate on the root locus plot the region on the complex plan where the the rise time for the step response for a complex conjugate pair of poles is less than .65 seconds. Use the approximation

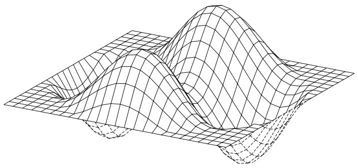
$$t_r \approx \frac{1.8}{\omega_n}.$$

Label any angles or distances that you used in this determination.

3. Use the root locus plot to determine the approximate range of values for the parameter k that satisfy both the rise time and overshoot specifications.

Problem 9.27 Sketch the root locus plot for

$$\frac{Y(s)}{R(s)} = \frac{kG(s)}{1 + kG(s)}$$



where

$$G(s) = \frac{1}{(s+3)\,(s^2+4s+5)}.$$

Be sure to include the details of all your computations.

1. Use your sketch on the previous page to determine the *approximate* value for k at which the root locus crosses the imaginary axis.

2. Use the Routh array to determine the *exact* value of k at which the root locus crosses the imaginary axis.

Problem 9.28 Sketch the Bode plot for

$$G(s) = \frac{20000}{(s+10)\,(s+1000)}.$$

Problem 9.29

1. Sketch the root locus plot for

$$G(s) = \frac{20}{s^2+s+10}.$$

2. Consider the phase lead compensator of the form

$$C(s) = \frac{s+10}{s+20}.$$

Sketch the root locus plot for $C(s)G(s)$ and explain why this phase lead compensator increases the stability of the system under unity feedback.

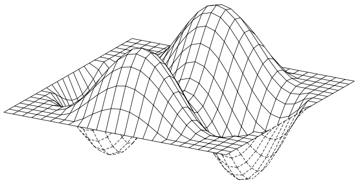
Problem 9.30 Use a partial fraction expansion to compute x(t) when

$$X(s) = \frac{4}{s^2+2s+4}\frac{1}{s}.$$

Use a partial fraction expansion to compute x(t) when

$$X(s) = \frac{4}{(s^2+2s+4)\,(s+20)}\frac{1}{s}.$$

Are the responses similar? Explain whether this was expected or unexpected.



Chapter 10

Basic Filter Theory

This chapter considers the characteristics of several types of filters. They will be employed in various manners as the basic building blocks in the subsequent chapter on controller design. This chapter is particularly straightforward: everything is a simple application of circuit analysis and frequency response.

To motivate the utility of filters, consider the following example.

Example 10.0.1 A common problem in the control of large structures such as aircraft and space vehicles is to control the rigid body mode of the structure while not exciting the elastic modes of the structure. As an example, consider the coupled mass system in Figure 10.1. We want to control the position of the center of mass of the system while affecting the relative position of the two masses as little as possible. Mathematically, we want to use $f(t)$ to control $\frac{m_1x_1(t)+m_2x_2(t)}{m_1+m_2}$ and simultaneously we want $x_1(t) - x_2(t)$ to remain constant. The force $d(t)$ is an external disturbance, which we will initially consider to be zero.

One approach to this problem would be a multi-loop feedback method where some balance is struck between controlling the rigid body mode of the system and suppressing the flexible mode. In this example we will

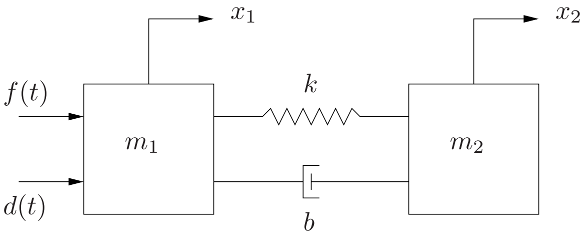


Figure 10.1. Controlled coupled mass system with a rigid body mode and a flexible mode.

approach the problem simply as one of controlling x_1 with feedback and designing the control law to minimize the effect on $x_1(t) - x_2(t)$ rather than simultaneously controlling the two.

The equations of motion for the two masses are

$$\begin{aligned} m_1 \ddot{x}_1 &= k\left(x_2 - x_1\right) + b\left(\dot{x}_2 - \dot{x}_1\right) + f(t) \\ m_2 \ddot{x}_2 &= k\left(x_1 - x_2\right) + b\left(\dot{x}_1 - \dot{x}_2\right). \end{aligned}$$

Adding the two equations gives

$$m_1 \ddot{x}_1 + m_2 \ddot{x}_2 = f(t).$$

If we let x_{com} denote the center of mass, then

$$x_{com}(t) = \frac{m_1 x_1(t) + m_2 x_2(t)}{m_1 + m_2}$$

and

$$\left(m_1 + m_2\right) \ddot{x}_{com} = f(t).$$

Subtracting the two equations gives

$$m_1 \ddot{x}_1 - m_2 \ddot{x}_2 = 2k\left(x_2 - x_1\right) + 2b\left(\dot{x}_2 - \dot{x}_1\right) + f(t).$$

So if we let

$$x_{flex} = \frac{x_1}{m_2} - \frac{x_2}{m_1}$$

then

$$m_1 m_2 \ddot{x}_{flex} =$$

■

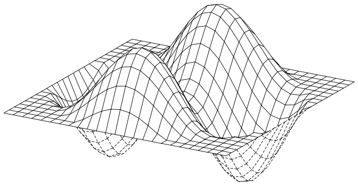
10.1 Low and High Pass Filters

Consider the circuit illustrated in Figure 10.2. For reasons that will be addressed subsequently, this is called a *low pass filter.* Kirchhoff's voltage law around the circuit gives

$$\begin{aligned} v_{in} &= v_R + v_C \\ &= v_r + v_{out}, \end{aligned}$$

where v_R and v_C are the voltage drops across the resistor and capacitor respectively. Since

$$\begin{aligned} v_R &= iR \\ i &= C\frac{dv_C}{dt} \end{aligned}$$



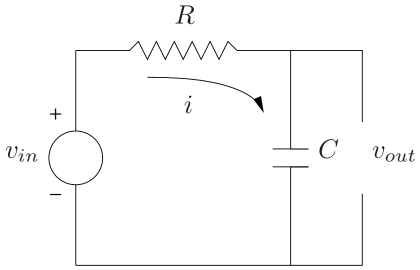


Figure 10.2. Low pass filter circuit.

in the frequency domain we have

$$\begin{aligned} V_R(s) &= I(s)R \\ I(s) &= CsV_C(s) \\ &= CsV_{out}(s) \end{aligned}$$

and substituting into the voltage equation gives

$$V_{in}(s) = (CRs + 1)\, V_{out}(s).$$

So the transfer function is

$$\frac{V_{out}}{V_{in}} = \frac{1}{CRs + 1}.$$

Clearly, this circuit has a pole at $s = -\frac{1}{CR}$. The frequency, $\omega = \frac{1}{CR}$ is called the *cutoff frequency*. For the case where $C = R = 10$, the Bode plot is illustrated in Figure 10.3. Frequencies below $\omega \approx 0.01$ are passed through the filter without any amplification or attenuation; in contrast, frequencies above $\omega \approx 0.01$, are attenuated.

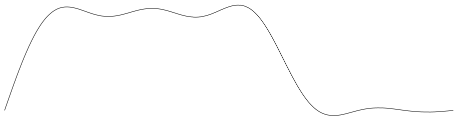
If the output voltage is measured across the resistor instead of the capacitor, the circuit is a high pass filter, which is illustrated in Figure 10.4. An easy circuit analysis gives the transfer function as

$$\frac{V_{out}}{V_{in}} = \frac{CRs}{CRs + 1}$$

and the frequency response is illustrated in Figure 10.5. Frequencies above $\omega \approx 0.01$ are passed through the filter without any amplification or attenuation; in contrast, frequencies below $\omega \approx 0.01$, are attenuated.

10.2 Band Pass Filters

The band pass filter may be constructed by placing low and high pass filter in series. If the cutoff frequency of If the cutoff of the low pass filter is higher



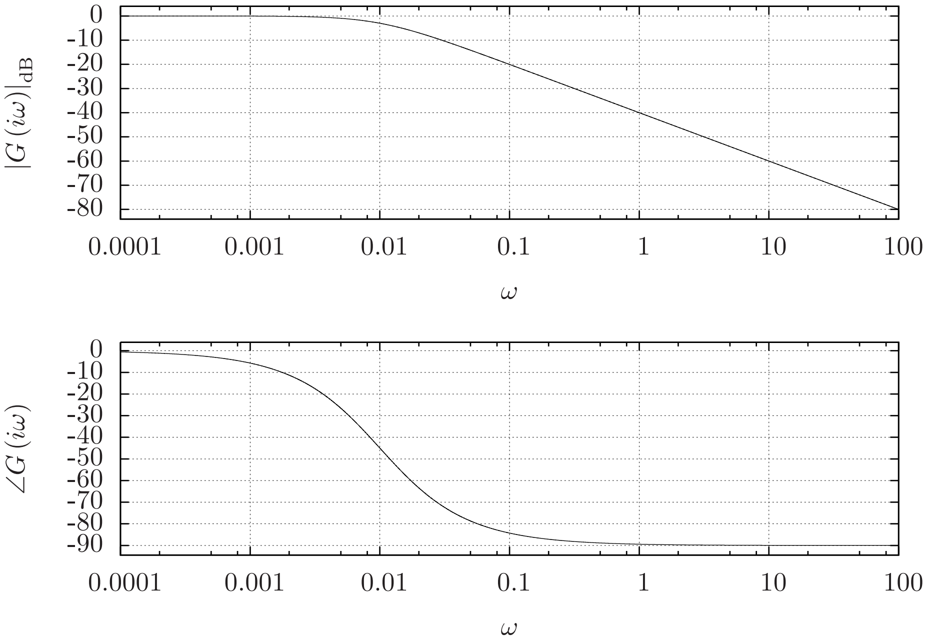


Figure 10.3. Frequency response of a low pass filter with $C = 10$ and $R = 10$.

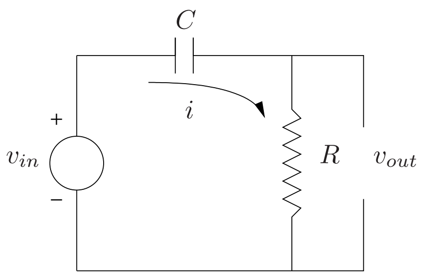
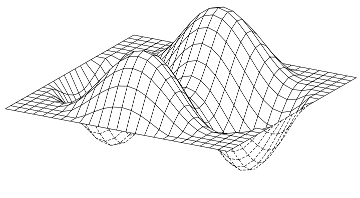


Figure 10.4. High pass filter circuit.



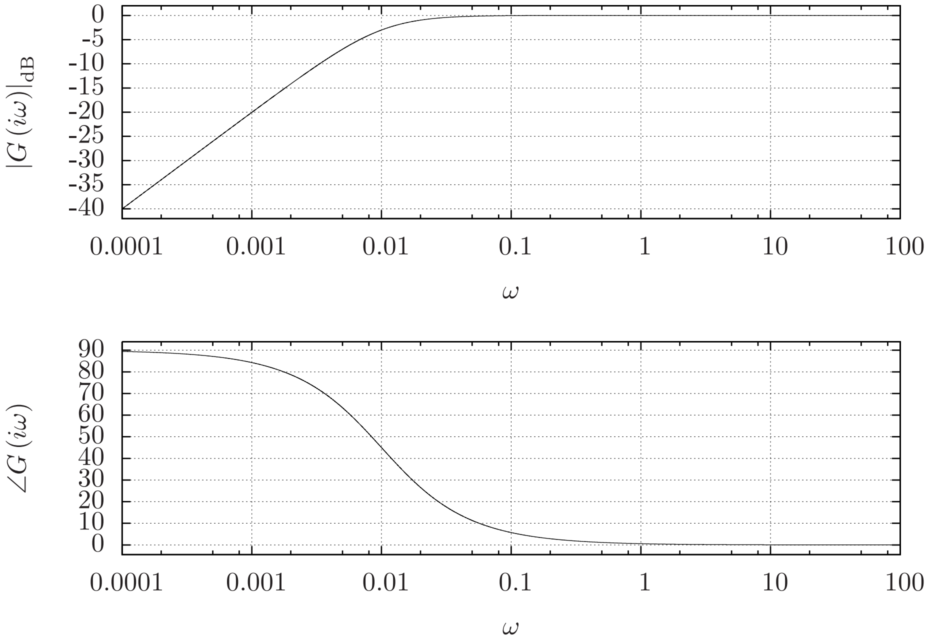


Figure 10.5. Frequency response of a high pass filter with $C = 10$ and $R = 10$.

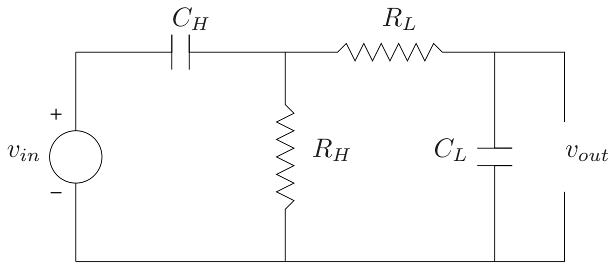


Figure 10.6. Band pass filter circuit.

than the high pass filter, a band pass filter results. The circuit in Figure 10.6 is simply a low and high pass filter placed in series, where R_H, C_H, R_L and C_L indicate the resistors and capacitors corresponding to the high and low pass filters, respectively. The transfer function is simply obtained by multiplying the transfer functions for the low and high pass filters,

$$\frac{V_{out}}{V_{in}} = \frac{1}{C_L R_L s + 1} \frac{C_H R_H s}{C_H R_H s + 1}.$$

The frequency response for when $R_1 = 1$ and $C_1 = 1$ (cutoff for the low pass filter $\omega = 1$) and $R_2 = 10$ and $C_2 = 10$ (cutoff for the high pass is $\omega = 0.01$) is illustrated in Figure 10.7. Frequencies between the two cutoffs are passed unattenuated whereas frequencies above the low pass filter cutoff and below the high pass cutoff are attenuated.

10.3 Notch Filters

If placing low pass and high pass filters in series results in a band pass filter, a reasonable guess to get a notch filter would be to place them in series with their outputs added, as is illustrated in Figure 10.8, where R_H, C_H, R_L and C_L indicate the resistors and capacitors corresponding to the high and low pass filters, respectively.

It is left as an exercise to show that the transfer function for the notch filter circuit in Figure 10.8 is given by

$$\frac{V_{out}}{V_{in}} = \frac{(C_L R_H + 1)(C_H R_L + 1)}{C_L C_H R_L R_H s^2 + (C_H R_L + C_L R_H + C_L R_L) s + 1}.$$

The frequency response for the notch filter is illustrated in Figure 10.9 where $C_L = 10$, $R_L = 10$, $C_H = 1$ and $R_H = 1$.

If the same filter is placed in series with itself, its general effect will be multiplied. For example, for the notch filter, if two of them placed in series, as is illustrated in Figure 10.10, and if $C_L = 10$, $R_L = 10$, $C_H = 1$ and $R_H = 1$, the

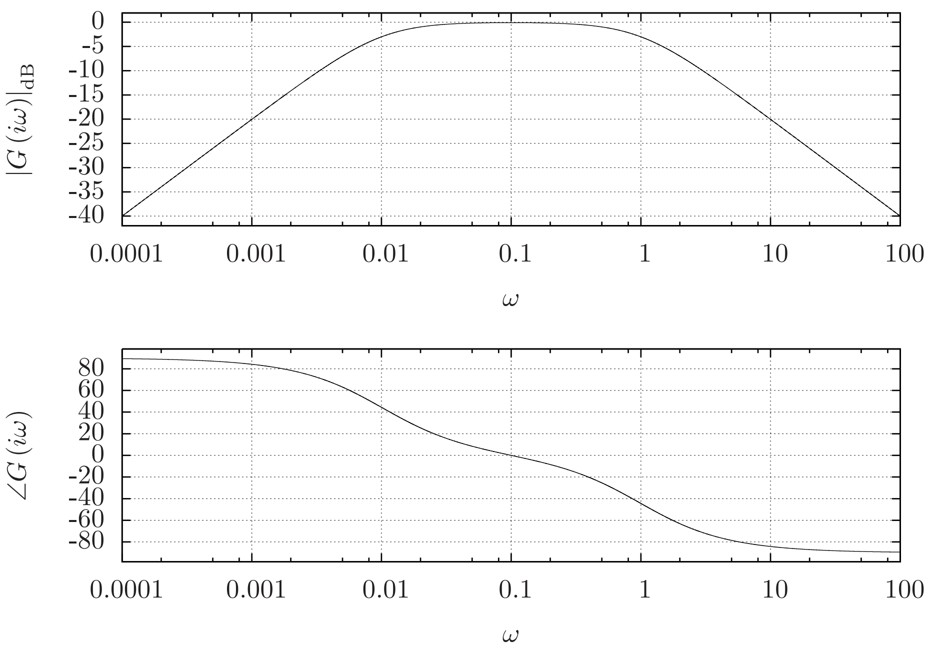


Figure 10.7. Frequency response of a notch filter with $C_1 = 10$, $R_1 = 10$ and $C_2 = 1$ and $R_2 = 1$.

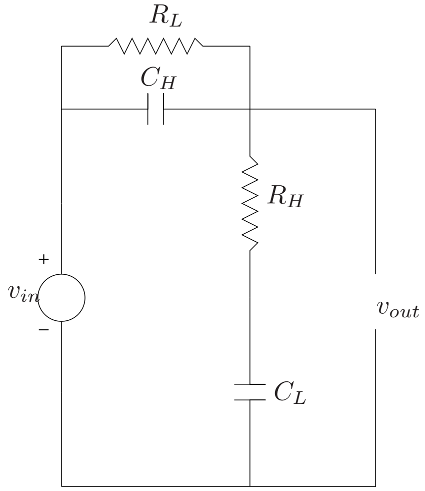


Figure 10.8. Notch filter circuit.

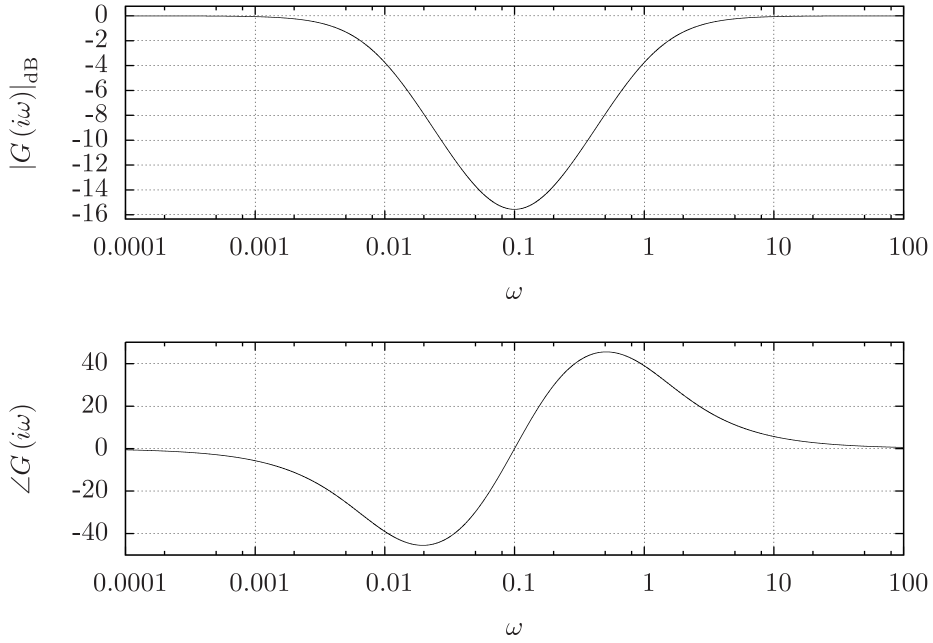


Figure 10.9. Frequency response of a notch filter with $C_1 = 10$, $R_1 = 10$ and $C_2 = 1$ and $R_2 = 1$.

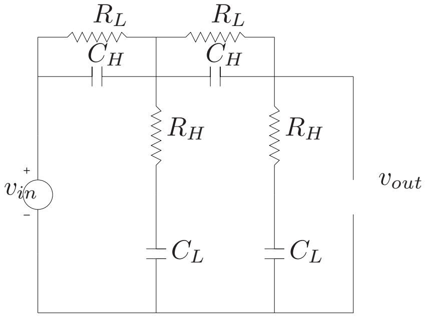


Figure 10.10. Notch filter circuit.

frequency response is illustrated in Figure 10.11. Note that when compared with the frequency response in Figure 10.9, the depth of the notch in Figure 10.11 is increased. The slope of the magnitude curves at the cutoff frequencies is also increases (which may be hard to casually see since the scales on the graphs are different).

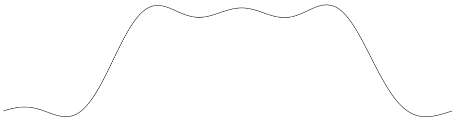
10.4 Phase Lead and Lag Filters

10.5 Exercises

Problem 10.1 Show that the transfer function for the notch filter illustrated in Figure 10.8 is given by

$$\frac{V_{out}}{V_{in}} = \frac{(C_L R_H + 1)(C_H R_L + 1)}{C_L C_H R_L R_H s^2 + (C_H R_L + C_L R_H + C_L R_L) s + 1}.$$

Problem 10.2 Place two band pass filters in series and plot the frequency response. Describe the effect of placing the filters in series compared with a single filter.



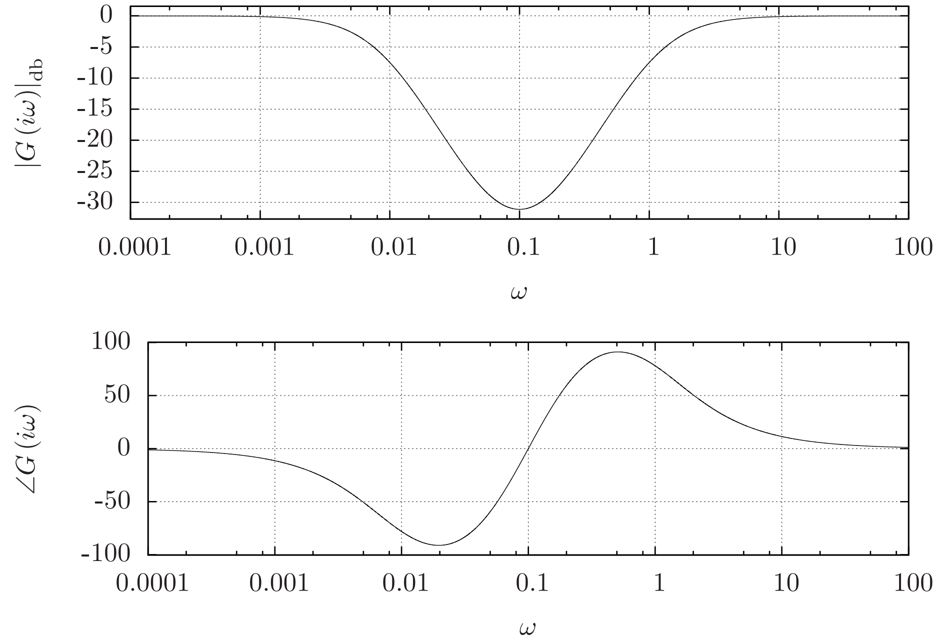
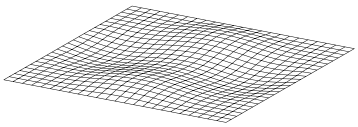


Figure 10.11. Frequency response of two notch filters in series with $C_1 = 10$, $R_1 = 10$ and $C_2 = 1$ and $R_2 = 1$.



Chapter 11

Basic Control Theory: Design

11.1 Types of Feedback Compensation

Various configurations are possible for feedback compensation. This book will particularly focus on compensation added in a feedback loop in the configurations illustrated in Figure 11.1 which is commonly called *cascade compensation*. In Figure 11.1, the block with $G_p(s)$ represents the plant dynamics, the output of which we desire to control. The block $G_c(s)$ is the compensator block, which must be designed based upon the performance specifications for the system. The block with $G_s(s)$ represents the sensor dynamics. In this text, this will often be idealized as the identity; however, in most applications the dynamics (or, at a minimum, the gain) of the sensors must be considered.

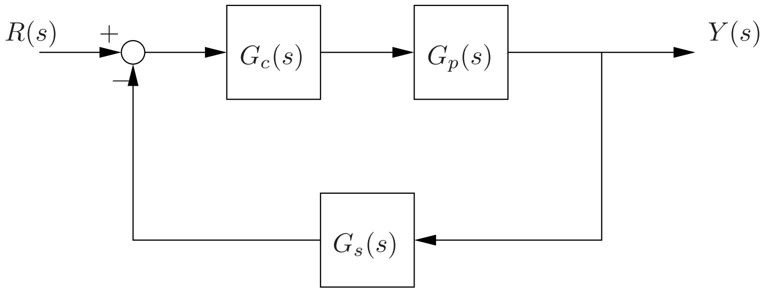


Figure 11.1. Cascade compensation configuration.

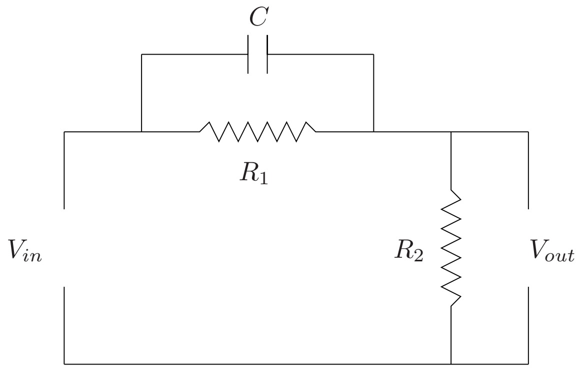


Figure 11.2. Lead compensator circuit.

11.2 Lead/Lag Compensation

Section 9.2 introduced the notion of PID control and presented the usual effects of each type of feedback on a second order system (*e.g.,* introducing or increasing the gain for derivative control increases damping). While the mathematical analysis is useful and the proper point to initially consider the tool, what was completely missing was the means by which one could actually implement it in a real engineering system.

Lead and lag filters are easy to implement with analog circuits and are hence economical and effective means for control. As will be demonstrated subsequently, a lead compensator is a means to approximate PD control and a lag compensator is a means to approximate PI control. Combined, obviously, results in an approximate manner to implement PID control.

11.2.1 Lead compensation

Consider the circuit illustrated in Figure 11.2. It is left as an exercise (Problem 11.1) to show that the transfer function for this circuit is

$$\frac{V_{out}}{V_{in}} = \left(\frac{R_2}{R_1 + R_2}\right) \frac{R_1 C s + 1}{\left(\frac{R_2}{R_1+R_2}\right) R_1 C s + 1}.$$

Example 11.2.1 Consider a phase lead circuit where

$$\begin{aligned} R_1 &= 10 \\ R_2 &= 0.1 \\ C &= 1. \end{aligned}$$

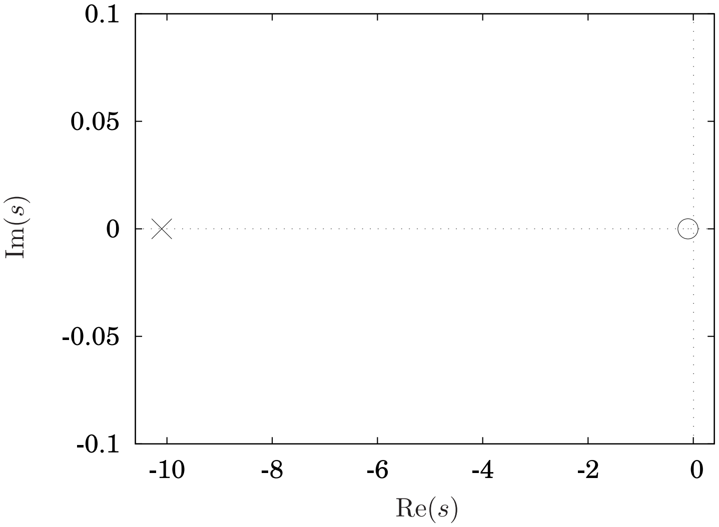


Figure 11.3. Pole and zero location for a lead compensator.

The pole and zero for this transfer function is illustrated in Figure 11.3. Observe that, for these parameter values, the zero is very close to the imaginary axis and the pole is relatively far to the left. If a zero is located very near the imaginary axis, the term in the transfer function is of the form

$$(s + \epsilon) \approx s$$

if $\epsilon \ll 1$. Because differentiation in the frequency domain is given by multiplication by s, this term approximates differentiation. Since the pole is relatively far to the left, we may treat its contribution as negligible. By this interpretation, then, if a lead compensator is added in cascade form, it acts much like term proportional to the derivative of the error, which, in accordance with our intuition from Section 9.2 generally acts in a stabilizing manner. While this example considered very specific parameter values, in fact, it is left as an exercise to show that the pole is to the left of the for all positive values for the circuit parameters and hence the interpretation of the effect of the circuit is general. ■

11.2.2 Lag compensation

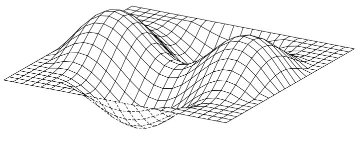
11.3 Tracking and System Type

11.4 Disturbance Rejection

11.5 Multi-Loop Feedback

11.6 Exercises

Problem 11.1 Determine the transfer function from the input voltage to the output voltage for the circuit illustrated in Figure 11.2. Show that for this circuit the pole and zero are in the left half plane and that the pole is always to the left of the zero.



Chapter 12

Partial Differential Equations

This chapter considers techniques for solving partial differential equations. The solution method that is considered in this book is the *separation of variables* separation of variables method. Partial differential equations can be categorized by type, similar to categorizing ordinary differential equations. However, in contrast to ordinary differential equations, the categorization will not affect the solution *method*, but rather be a reflection of the properties of the resulting solution, which itself, is a result of the underlying physics.

The outline of this chapter is to first present three common engineering problems that lead to different types of partial differential equations. As will be apparent, there are some broad commonalities with respect to the solution technique. An extension of the theory developed by the engineering problems will be investigated later in the chapter in section 12.7.

12.1 The Wave Equation

The so-called *wave equation* describes many different physical wave-like phenomena. It will be motivated and initially solved using the example of a vibrating string.

12.1.1 Derivation of the Wave Equation

Consider the elastic string illustrated in Figure 12.1. Let x describe the location along a straight line between the end points and u denote the displacement of the string. Let the length between the end points be L. The function u will be a function of the position along the string, x as well as time, *i.e.,* $u(x, t)$. Solving the wave equation will amount to determining the function, $u(x, t)$ that gives the displacement of the string at time t and location x. Let the tension in the string be denoted by τ and the mass per unit length be denoted by ρ.

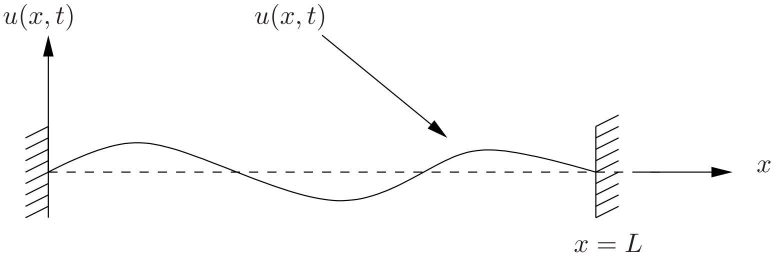


Figure 12.1. Vibrating string.

The string is assumed to elastic, which means it may have an internal tension, τ, but it requires no moment to bend it.

The derivation of the wave equation is simply using Newton's law on a infinitesimal segment of the string. Consider the small section illustrated in Figure 12.2. Newton's law on the element in the vertical direction gives

$$\rho\Delta x\frac{\partial^2 u\left(x+\frac{\Delta x}{2},t\right)}{\partial t^2} = \tau\left(x+\Delta x,t\right)\sin\left(\theta\left(x+\Delta x,t\right)\right) - \tau\left(x,t\right)\sin\left(\theta\left(x,t\right)\right). \tag{12.1}$$

Expanding each of the terms in a Taylor series individually gives

$$\begin{aligned}
\frac{\partial^2 u\left(x+\frac{\Delta x}{2},t\right)}{\partial t^2} &= \frac{\partial^2 u\left(x,t\right)}{\partial t^2} + \frac{\partial^3 u\left(x,t\right)}{\partial^2 t\partial x}\frac{\Delta x}{2} + \cdots \\
\tau(x+\Delta x,t) &= \tau(x,t) + \frac{\partial\tau(x,t)}{\partial x}\Delta x + \cdots \\
\sin\left(\theta(x+\Delta x,t)\right) &= \sin\left(\theta(x,t)\right) + \frac{d\sin(\theta)}{d\theta}\frac{\partial\theta(x,t)}{\partial x}\Delta x \\
&= \sin\left(\theta(x,t)\right) + \cos\left(\theta(x,t)\right)\frac{\partial\theta(x,t)}{\partial x}\Delta x
\end{aligned}$$

Substituting into equation 12.1 and keeping terms only up to Δx, *i.e.,* assuming $\Delta x \ll 1$ gives

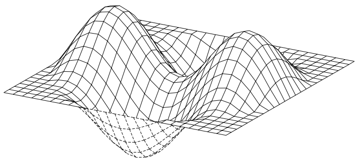
$$\rho\Delta x = \frac{\partial^2 u(x,t)}{\partial t^2} = \tau(x,t)\cos\left(\theta(x,t)\right)\frac{\partial\theta(x,t)}{\partial x}\Delta x + \sin\left(\theta(x,t)\right)\frac{\partial\tau(x,t)}{\partial x}\Delta x,$$

or

$$\rho\frac{\partial^2 u(x,t)}{\partial t^2} = \tau(x,t)\cos\left(\theta(x,t)\right)\frac{\partial\theta(x,t)}{\partial x} + \sin\left(\theta(x,t)\right)\frac{\partial\tau(x,t)}{\partial x}. \tag{12.2}$$

To proceed any further, we need some assumptions. Assume that the string only undergoes small displacements, *i.e.*, $u(x,t) \ll 1$ and furthermore that the slope of the string is small, *i.e.*, $\frac{\partial u}{\partial x} \ll 1$. This would imply immediately that

$$\begin{aligned}
\sin\left(\theta(x,t)\right) &\approx \theta(x,t) \\
\cos\left(\theta(x,t)\right) &\approx 1.
\end{aligned}$$



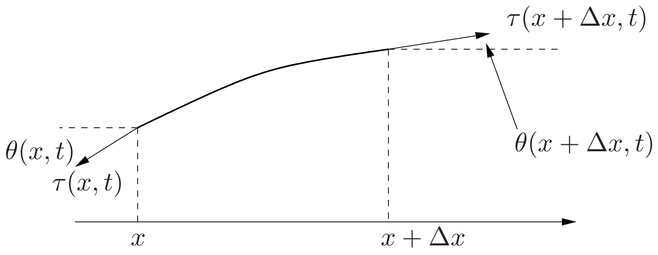


Figure 12.2. Infinitesimal element of the string.

Also, express the tension in the string as

$$\tau(x,t) = \tau + \hat{\tau}(x,t)$$

where τ is a constant and is the tension in the string when it is still $(u(x,t) = 0)$. For small motions, it will be the case that $\hat{\tau}(x,t) \ll 1$ and $\frac{\partial \tau(x,t)}{\partial x} \ll 1$.

Since both terms in the second term of the sum on the right hand side of equation 12.2 are small, then

$$\rho \frac{\partial^2 u(x,t)}{\partial t^2} = \tau \frac{\partial \theta(x,t)}{\partial x}. \tag{12.3}$$

Finally, for small $u(x,t)$ and small $\theta(x,t)$,

$$\theta(x,t) \approx \tan(\theta(x,t)) = \frac{\partial u(x,t)}{\partial x}$$

so

$$\frac{\rho}{\tau}\frac{\partial^2 u(x,t)}{\partial t^2} = \frac{\partial^2 u(x,t)}{\partial x^2}. \tag{12.4}$$

12.1.2 Boundary Conditions

In general, the wave equation is of the form

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}.$$

Analogous to ordinary differential equation, in order to solve this equation conditions on $u(x,t)$ at the initial time for the problem, usually $t = 0$ *as well as* conditions on $u(x,t)$ on the physical boundaries of the problem must be specified. The latter are normally called *boundary conditions* and play a fundamental role in the solution of the problem.

To proceed, assume the simplest case: the ends of the string are fixed, *i.e.*,

$$u(0,t) = u(L,t) = 0.$$

Also, assume that the initial shape and velocity of the string is known, *i.e.*,

$$\begin{aligned} u(x,0) &= f(x) \\ \left.\frac{du}{dt}\right|_{(x,0)} &= g(x), \end{aligned}$$

so the function $f(x)$ is the initial shape profile of the string and $g(x)$ is the initial velocity profile.

12.1.3 Separation of Variables

The basic idea behind the method of *separation of variables* is that the solution to the wave equation can be expressed in the form

$$u(x,t) = X(x)T(t),$$

i.e., the solution is the product of two functions where one of the functions only depends on the spatial variable, x and the other function only depends on the temporal variable, t. Subsequently in section 12.7 it will be shown that this assumption is not some sort of wild guess, but rather is theoretically well-grounded. Regardless, at this point there is no harm in assuming it is true, substituting it into the differential equation and see what pops out. Note that due to the assumed form of $u(x,t)$

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= X(x)\frac{d^2T(t)}{dt^2} = X(x)T''(t) \\ \frac{\partial^2 u}{\partial x^2} &= \frac{d^2X(x)}{dx^2}T(t) = X''(x)T(t). \end{aligned}$$

So, substituting into the wave equation gives

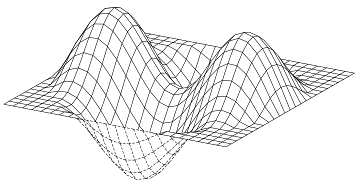
$$X(x)T'' = \alpha^2 X''(x)T(t).$$

or

$$\frac{X''(x)}{X(x)} = \frac{1}{\alpha^2}\frac{T''(t)}{T(t)}. \tag{12.5}$$

The critical feature of equation 12.5 is that the left hand side depends only upon x, the right hand side depends only upon t and they are equal. The only way a non-trivial function of x can equal a non-trivial function of t for arbitrary x and t is for both sides to be equal to a constant. Note, this does *not* mean $X(x)$ is a constant and $T(t)$ is a constant; rather, the *ratios* $\frac{X''(x)}{X(x)}$ and $\frac{T''(t)}{T(t)}$ must be constant. That constant is denoted by $-\lambda$, and is called an *eigenvalue*. Thus

$$\frac{X''(x)}{X(x)} = \frac{1}{\alpha^2}\frac{T''(t)}{T(t)} = -\lambda$$



which are actually *two* equations

$$\begin{aligned}\frac{d^2X(x)}{dx^2}+\lambda X(x) &= 0\\ \frac{d^2T(t)}{dt^2}+\alpha^2\lambda T(t) &= 0.\end{aligned}$$

The general solutions to these two equations are obvious from inspection (hopefully):

$$X(x) = c_1\sin\sqrt{\lambda}x + c_2\cos\sqrt{\lambda}x \tag{12.6}$$

$$T(t) = c_3\sin\alpha\sqrt{\lambda}t + c_4\cos\alpha\sqrt{\lambda}t. \tag{12.7}$$

This solution assumes that $\lambda > 0$. It is left as an exercise to show that if $\lambda \leq 0$ the boundary conditions can not be satisfied.

Now consider the boundary conditions

$$u(0,t) = u(L,t) = 0.$$

Substituting $u(x,t) = X(x)T(t)$ gives

$$X(0)T(t) = X(L)T(t) = 0$$

which gives

$$\begin{aligned}X(0) &= 0\\ X(L) &= 0.\end{aligned}$$

To satisfy the first boundary condition, $c_2 = 0$ in equation 12.6. To satisfy the second, *either*

$$c_1 = 0$$

or

$$\lambda = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, \ldots.$$

Note that $c_1 = 0$ leads to the trivial solution ($u(x,t) = 0$) which will not be able to satisfy the initial shape and velocity profiles. Note also that *there are an infinite number of solutions*, one for each $n = 1, 2, \ldots$.

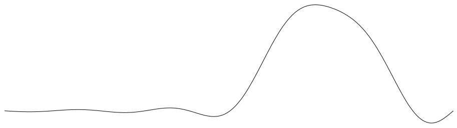
Hence, proceed with the assumption for $\lambda = \frac{n^2\pi^2}{L^2}$. Substituting into $u(x,t) = X(x)T(t)$ gives

$$u_n(x,t) = c_1\sin\frac{n\pi x}{L}\left(c_{3,n}\sin\frac{\alpha n\pi t}{L} + c_{4,n}\cos\frac{\alpha n\pi t}{L}\right) \quad n = 1, 2, \ldots$$

or

$$u_n(x,t) = \sin\frac{n\pi x}{L}\left(a_n\sin\frac{\alpha n\pi t}{L} + b_n\cos\frac{\alpha n\pi t}{L}\right), \quad n = 1, 2, \ldots \tag{12.8}$$

where the constant were combined into a_n and b_n.



Observe the following very important point. *Any* of the $u_n(x,t)$ satisfies the wave equation as well as the two boundary conditions, as do any linear combination of the $u_n(x,t)$.

The last task is to satisfy the initial conditions, which were

$$\begin{aligned} u(x,0) &= f(x) \\ \left.\frac{du}{dt}\right|_{(x,0)} &= g(x). \end{aligned}$$

This may seem like an impossible task at first, but perhaps the availability of an infinite number of solutions will be of some help. In fact, let us just go for it and try to combine all the infinite number of solutions together in the form

$$u(x,t) = \sum_{n=1}^{\infty} \sin\frac{n\pi x}{L}\left(a_n \sin\frac{\alpha n\pi t}{L} + b_n \cos\frac{\alpha n\pi t}{L}\right). \tag{12.9}$$

Note that in this form, the initial conditions are

$$u(x,0) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L} = f(x) \tag{12.10}$$

$$u'(x,0) = \sum_{n=1}^{\infty} a_n \frac{\alpha n\pi}{L} \sin\frac{n\pi x}{L} = g(x). \tag{12.11}$$

Finally, the last bit of trickery is to multiply each side of equation 12.10 by $\sin\frac{m\pi x}{L}$ and integrate from 0 to L

$$\int_0^L \sin\frac{m\pi x}{L}\left(\sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}\right)dx = \int_0^L \sin\frac{m\pi x}{L} f(x)dx \tag{12.12}$$

and note that every single term on the left hand side of the equation is zero except for $m = n$, which nicely kills off all but one of the infinite number of terms in the series.[1] Hence

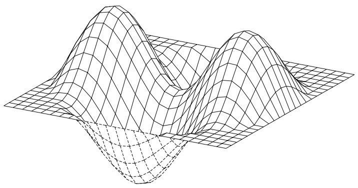
$$b_n \int_0^L \left(\sin\frac{n\pi x}{L}\right)^2 dx = \int_0^L \sin\frac{n\pi x}{L} f(x)dx,$$

or

$$b_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}dx. \tag{12.13}$$

Using these values for b_n, equation 12.10 is called the *Fourier sine series* for $f(x)$. The following example illustrates the computations involved in computing the Fourier sine series as well as gives an indication of the convergence properties of such a series.

[1] The detailed computation for this is in Appendix C.1.



Example 12.1.1 Let $L = 3$ and

$$f(x) = \left\{ \begin{array}{ll} x & x < 1 \\ \frac{3-x}{2} & 1 \leq x \leq 3 \end{array} \right. . \tag{12.14}$$

Computing the Fourier coefficients,

$$\begin{aligned}
b_n &= \frac{2}{3} \int_0^3 f(x) \sin \frac{n\pi x}{3} dx \\
&= \frac{2}{3} \left[\int_0^1 x \sin \frac{n\pi x}{3} dx + \int_1^3 \frac{3-x}{2} \sin \frac{n\pi x}{3} \right] \\
&= \frac{2}{3} \left[-\frac{3x}{n\pi} \cos \frac{n\pi x}{3} \Big|_0^1 + \int_0^1 \frac{3}{n\pi} \cos \frac{n\pi x}{3} dx + \frac{3}{2} \int_1^3 \sin \frac{n\pi x}{3} dx + \right. \\
&\quad \left. \frac{3x}{2n\pi} \cos \frac{n\pi x}{3} \Big|_1^3 - \frac{1}{2} \int_1^3 \frac{3}{n\pi} \cos \frac{n\pi x}{3} dx \right] \\
&= \frac{2}{3} \left[-\frac{3x}{n\pi} \cos \frac{n\pi x}{3} \Big|_0^1 + \frac{9}{n^2\pi^2} \sin \frac{n\pi x}{3} \Big|_0^1 - \frac{9}{2n\pi} \cos \frac{n\pi x}{3} \Big|_1^3 + \right. \\
&\quad \left. \frac{3x}{2n\pi} \cos \frac{n\pi x}{3} \Big|_1^3 - \frac{9}{2n^2\pi^2} \sin \frac{n\pi x}{3} \Big|_1^3 \right] \\
&= \frac{2}{3} \left[-\frac{3}{n\pi} \cos \frac{n\pi}{3} + \frac{9}{n^2\pi^2} \sin \frac{n\pi}{3} - \frac{9}{2n\pi} \left(\cos n\pi - \cos \frac{n\pi}{3} \right) + \right. \\
&\quad \left. \left(\frac{9}{2n\pi} \cos n\pi - \frac{3}{2n\pi} \cos \frac{n\pi}{3} \right) + \frac{9}{2n^2\pi^2} \sin \frac{n\pi}{3} \right] \\
&= \frac{9}{n^2\pi^2} \sin \frac{n\pi}{3}.
\end{aligned}$$

Figure 12.3 is a plot of the first four terms in the Fourier series; namely,

$$\begin{aligned}
f_1(x) &= \frac{9}{1^2\pi^2} \sin \frac{1\pi}{3} \sin \frac{1\pi x}{3} = \frac{9\sqrt{3}}{2\pi^2} \sin \frac{\pi x}{3} \\
f_2(x) &= \frac{9}{2^2\pi^2} \sin \frac{2\pi}{3} \sin \frac{2\pi x}{3} = \frac{9\sqrt{3}}{8\pi^2} \sin \frac{2\pi x}{3} \\
f_3(x) &= \frac{9}{3^2\pi^2} \sin \frac{3\pi}{3} \sin \frac{3\pi x}{3} = 0 \\
f_4(x) &= \frac{9}{4^2\pi^2} \sin \frac{4\pi}{3} \sin \frac{4\pi x}{3} = -\frac{9\sqrt{3}}{32\pi^2} \sin \frac{4\pi x}{3}
\end{aligned}$$

Figure 12.4 illustrates the sum of the first three, 10 and 20 components. Note that the curve converges to $f(x)$ as the number of components increases. ■

Now return to the solution for the vibrating string in equation 12.9. The b_n coefficients have already been determined by equation 12.13 and the a_n coefficients are computed similarly, *i.e.*, multiply each side of equation 12.11 by

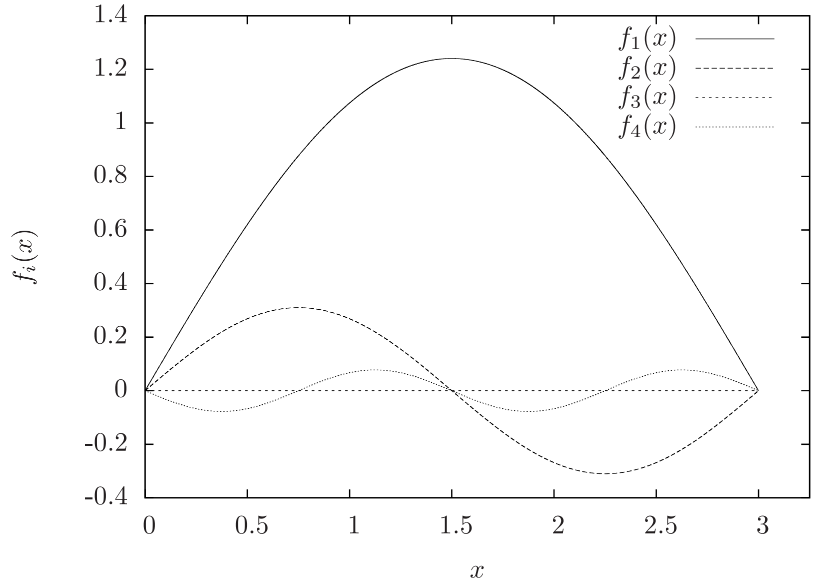


Figure 12.3. First four Fourier sine components of equation 12.14.

$\sin \frac{m\pi x}{L}$ and integrate from 0 to L which gives

$$a_n = \frac{2}{\alpha n\pi} \int_0^L g(x) \sin \frac{n\pi x}{L} dx. \tag{12.15}$$

Summary of the solution to the vibrating string problem

For small displacements $u(x,t)$, the vibration of a string of length L fixed at each end point is given by

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}$$

with

$$\begin{aligned} u(0,t) &= 0 \\ u(L,t) &= 0 \end{aligned}$$

as boundary conditions and

$$\begin{aligned} u(x,0) &= f(x) \\ \left.\frac{du}{dt}\right|_{(x,0)} &= g(x), \end{aligned}$$

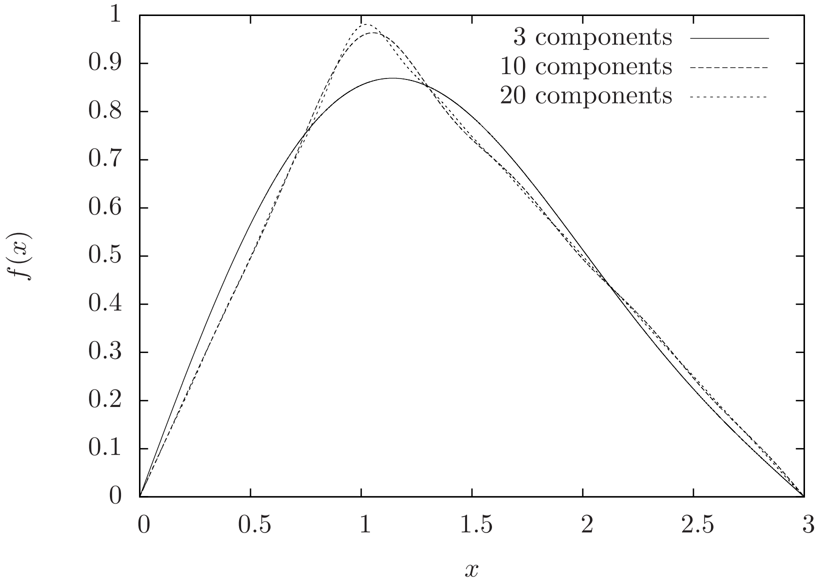


Figure 12.4. Truncated Fourier sine series converging to $f(x)$ from equation 12.14.

as initial conditions, where $f(x)$ and $g(x)$ are the initial shape of the string and initial velocity profile, respectively.

From the preceding analysis, the solution for the vibrating string problem is

$$u(x,t) = \sum_{n=1}^{\infty} \left[\sin \frac{n\pi x}{L} \left(a_n \sin \frac{\alpha n \pi t}{L} + b_n \cos \frac{\alpha n \pi t}{L} \right) \right]$$

where

$$\begin{aligned} \alpha &= \sqrt{\frac{T}{\rho}} \\ a_n &= \frac{2}{\alpha n \pi} \int_0^L g(x) \sin \frac{n\pi x}{L} dx \\ b_n &= \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx. \end{aligned}$$

Example 12.1.2 Solve

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}$$

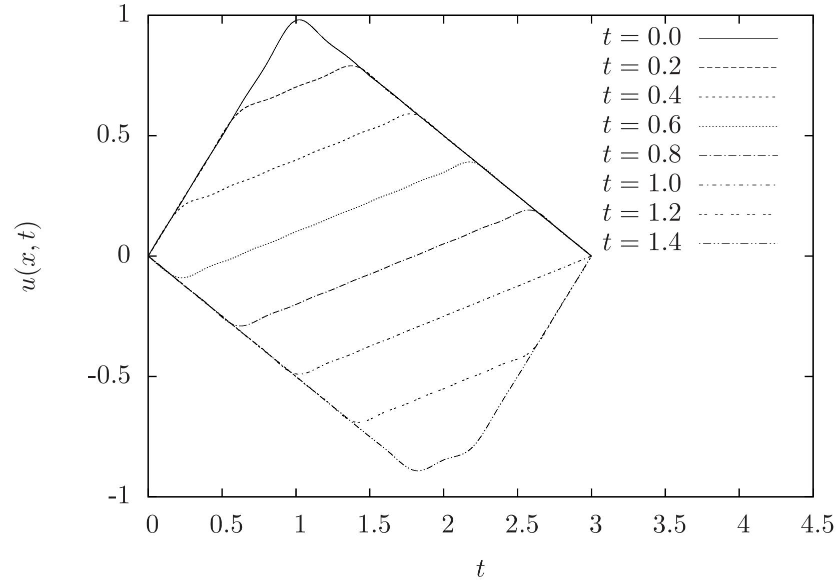


Figure 12.5. Response of a plucked string from example 12.1.2.

where $L = 3$ and $\alpha = 2$ subjected to the boundary conditions

$$\begin{aligned} u(0,t) &= 0 \\ u(L,t) &= 0 \end{aligned}$$

and initial conditions

$$\begin{aligned} u(x,0) &= \begin{cases} x & x < 1 \\ \frac{3-x}{2} & 1 \le x \le 3 \end{cases} \\ u'(x,0) &= 0. \end{aligned}$$

This represents a string plucked $\frac{1}{3}$ of the way along its length with zero initial velocity.

All the computations for this problem have already been carried out. Substituting for b_n from example 12.1.1 and $a_n = 0$ (since the initial velocity is zero) into equation 12.9 gives

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{3}\cos\frac{2n\pi t}{3} = \sum_{n=1}^{\infty}\left(\frac{9}{n^2\pi^2}\sin\frac{n\pi}{3}\right)\sin\frac{n\pi x}{3}\cos\frac{2n\pi t}{3}.$$

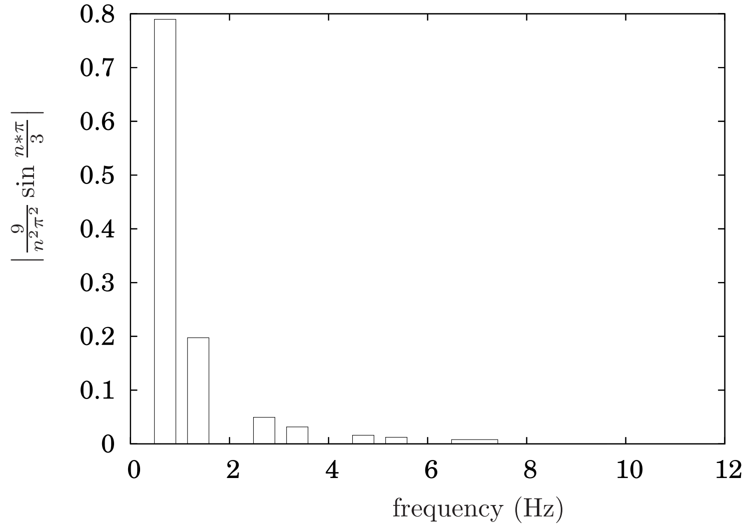


Figure 12.6. Spectrum for plucked string in example 12.1.2.

A plot of the motion of the string for various t values including the first 20 terms in the Fourier series is illustrated in Figure 12.5. A plot of the magnitude of the coefficient, b_n *versus* frequency, $\frac{\alpha n\pi}{L}$ is illustrated in Figure 12.6. This is called the *spectrum* of the response and is an illustration of the contribution of each mode to the overall response of the system. ■

Example 12.1.3 Consider the same string as in example 12.1.2 but instead of having the string plucked, like a guitar or banjo, consider it being impacted by a small hammer over a small segment of its length, like a piano. Thus, solve

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}$$

where $L = 3$ and $\alpha = 2$ subjected to the boundary conditions

$$\begin{aligned} u(0,t) &= 0 \\ u(L,t) &= 0 \end{aligned}$$

and initial conditions

$$\begin{aligned} u(x,0) &= 0 \\ u'(x,0) &= \begin{cases} 0 & 0 < x \le \frac{3}{4} \\ 1 & \frac{3}{4} < x \le 1 \\ 0 & 1 < x \le 3 \end{cases} . \end{aligned}$$

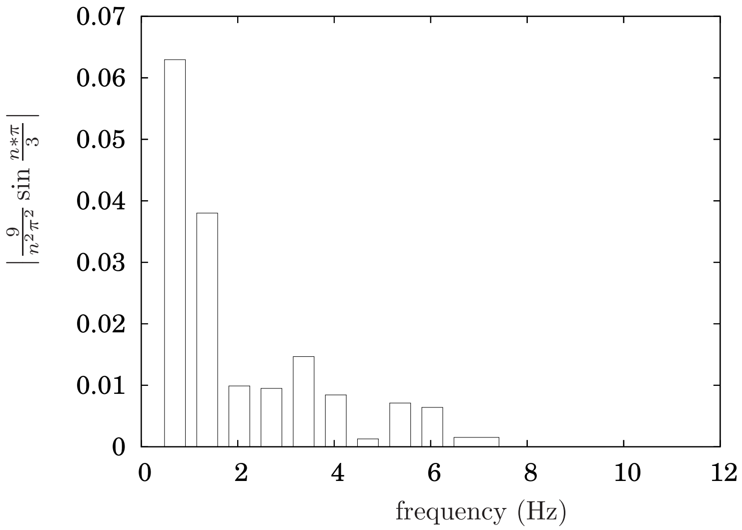


Figure 12.7. Spectrum for impacted string in example 12.1.3.

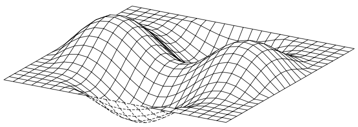
Substituting into equation 12.15

$$\begin{aligned} a_n &= \frac{1}{n\pi}\int_0^3 g(x)\sin\frac{n\pi x}{3}dx \\ &= \frac{1}{n\pi}\int_{\frac{3}{4}}^{1}\sin\frac{n\pi x}{3}dx \\ &= -\frac{3}{n^2\pi^2}\cos\frac{n\pi x}{3}\bigg|_{\frac{3}{4}}^{1} \\ &= \frac{3}{n^2\pi^2}\left(\cos\frac{n\pi}{4}-\cos\frac{n\pi}{3}\right). \end{aligned}$$

Hence,

$$u(x,t)=\sum_{n=1}^{\infty}\frac{3}{n^2\pi^2}\left(\cos\frac{n\pi}{4}-\cos\frac{n\pi}{3}\right)\sin\frac{n\pi x}{3}\sin\frac{2n\pi t}{3}$$

A plot of the magnitude of the coefficient, b_n *versus* frequency, $\frac{\alpha n\pi}{L}$ is illustrated in Figure 12.7. This is called the *spectrum* of the response and is an illustration of the contribution of each mode to the overall response of the system. Note that the relative contributions of the harmonics in this example are different than from the plucked example (example 12.1.2). This explains why a plucked and struck sting (say on a guitar) do not sound the same, even if they are the same note. ■



12.1.4 The Wave Equation with Dispersion

12.2 Fourier Series

Motivated by our apparent ability to use an infinite series of sine and cosine functions to match any initial conditions for the wave equation defined on the length of the string (equations 12.13 and 12.15), we will now consider the general problem of representing an arbitrary periodic function as a trigonometric series.

Motivated by the form of the solution to the wave equation, consider the series

$$f(x) = \sum_{n=0}^{\infty} \left[a_n \sin \frac{n\pi x}{L} + b_n \cos \frac{n\pi x}{L} \right]. \tag{12.16}$$

The question to consider is under what conditions will we be able to compute the infinite number of coefficients a_n and b_n so that this series converges to a specified function? There are a variety of reasons to pursue this, not the least of which are

- we may be forced to represent a function in this manner, as was the case for satisfying the initial conditions for the wave equation; and,

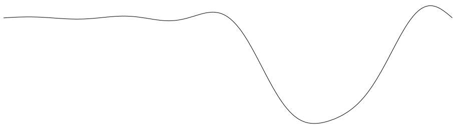
- even though it is an infinite series, sine and cosine functions are generally pretty easy to deal with, so, in the right context, it may be worth the effort to represent some given function as a trigonometric series of this nature because it may save us work elsewhere.

An example of the second case is considered in Section 12.2.6.

12.2.1 Periodic functions

As an initial observation, it is worth noting that because of the periodic nature of the trigonometric functions, it will probably not be possible to represent *any* function by a series of the form of equation 12.16. In particular, observe that

$$\begin{aligned}
f(x+2L) &= \sum_{n=0}^{\infty} \left[a_n \sin \frac{n\pi (x+2L)}{L} + b_n \cos \frac{n\pi (x+2L)}{L} \right] \\
&= \sum_{n=0}^{\infty} \left[a_n \sin \left(\frac{n\pi x}{L} + 2n\pi \right) + b_n \cos \left(\frac{n\pi x}{L} + 2n\pi \right) \right] \\
&= \sum_{n=0}^{\infty} \left(a_n \sin \frac{n\pi x}{L} + b_n \cos \frac{n\pi x}{L} \right) \\
&= f(x).
\end{aligned}$$



Of course, mathematically what this represents is that the series repeats itself over every interval of $2L$. Observe that similarly

$$\begin{aligned} f(x) &= f(x+2L) \\ &= f(x+4L) \\ &= f(x+6L) \\ &\vdots \\ &= f(x+2mL) \end{aligned}$$

where m is a natural number (positive integer). Motivated by this we define a periodic function as follows.

Definition 12.2.1 A function $f(x)$ is *periodic with period* T if T is the smallest number such that $f(x) = f(x+T)$. ⋄

Having defined a periodic function and observed that the series we are considering is periodic with period $2L$, it is obvious to conclude that the class of functions for which the series will converge *must be periodic.* In the case of the wave equation and other partial differential equations that will be considered subsequently, the initial shape of the string was not periodic; however, we were only interested in its shape over the length of the string. If we had plotted the Fourier series for the initial condition outside the domain of $x = 0$ to $x = L$ we would have observed that, in fact, the function was periodic, but we were only interested in it over the length of one half of its period.

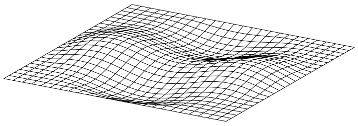
If we wish to consider the properties of the series for general periodic functions, since the length L was one half of the period, we could substitute $L = \frac{T}{2}$ in the sine and cosine functions in the series to put it in the form

$$f(x) = \sum_{n=0}^{\infty} \left[a_n \sin \frac{2n\pi x}{T} + b_n \cos \frac{2n\pi x}{T} \right] \tag{12.17}$$

that is in terms of the period, T rather than the length L.

12.2.2 Inner products

Not surprisingly, the "trick" (equation 12.12) that allowed us to compute the infinite number of coefficients in the Fourier series will be used in a similar manner here. However, instead of simply considering it to be a trick, whose only redeeming feature is one of mathematical manipulation, we will investigate things a bit further to see that, in fact, this "trick" is nothing more than using the usual dot product to project one vector onto another. In the rest of this section we will consider the generic properties of the dot product and its geometric interpretation which includes the important concept of orthogonality.



The dot product

Recall from vector algebra that the dot product between two vectors is defined as

$$\mathbf{x}\cdot\mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = x_1y_1 + x_2y_2 + \cdots + x_ny_n = \sum_{i-1}^{n} x_iy_i.$$

So, in words, the dot product is simply the sum of the product of all of the corresponding components of the vectors **x** and **y**.

To generalize this idea to functions, first note that, loosely speaking, one may think of a function as a vector by "sampling" its values at various points (perhaps an infinite number of points) along its domain, *i.e.,*

$$f(x) = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \end{bmatrix}.$$

Now, considering the dot product between two functions $f(x)$ and $g(x)$ over an interval of $-L < x < L$ and taking the values of each at e we may write

$$\begin{bmatrix} f(-L) \\ f(-L+dx) \\ f(-L+2dx) \\ \vdots \\ f(L) \end{bmatrix} \cdot \begin{bmatrix} g(-L) \\ g(-L+dx) \\ g(-L+2dx) \\ \vdots \\ g(L) \end{bmatrix} = \sum_{n=0}^{\frac{2L}{dx}} f(-L+ndx)g(-L+ndx).$$

Now clearly our goal is going to be to take the limit as $dx \to 0$; however in this limit this sum will typically not converge for nonzero $f(x)$ and $g(x)$ since it will be the infinite sum of finite values. However, if we modify it slightly by multiplying the product of f and g by dx, and taking the limit as $dx \to 0$ we have

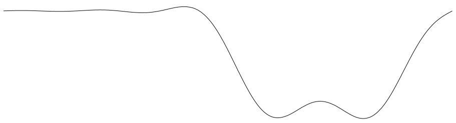
$$\lim_{dx\to 0} \sum_{n=0}^{\frac{2L}{dx}} f(-L+ndx)g(-L+ndx)dx = \int_{-L}^{L} f(x)g(x)dx.$$

Motivated by this we define the inner product between two periodic functions with period $2L$.

Definition 12.2.2 Let $f(x)$ and $g(x)$ be periodic functions with period $2L$. The *inner product of* f *and* g, denote by $\langle f, g\rangle$ is

$$\langle f, g\rangle = \int_{-L}^{L} f(x)g(x)dx.$$

◇



With this definition, it is clear that all the usual properties of the dot product generalize to this inner product. In particular

1. $\langle f_1 + f_2, g\rangle = \langle f_1, g\rangle + \langle f_2, g\rangle$;
2. $\langle \alpha f, g\rangle = \alpha\langle f, g\rangle$;
3. $\langle f, g\rangle = \langle g, f\rangle$ (for *real* f and g); and,
4. $\langle f, f\rangle \neq 0$ unless $f = 0$.

In addition to the usual properties of a dot product holding for the generalization of the inner product to functions, the main intuitive idea also holds: *the inner product gives a measure of the degree of "alignment" the functions.*

Example 12.2.3 Consider the three functions

$$\begin{aligned} f_1(x) &= \sin x \\ f_2(x) &= \sin 2x \\ f_3(x) &= \begin{cases} x & 0 \le x \le \frac{\pi}{2} \\ \pi - x & \frac{\pi}{2} < x \le \frac{3\pi}{2} \\ x - 2\pi & \frac{3\pi}{2} < x \le 2\pi \end{cases} \end{aligned}$$

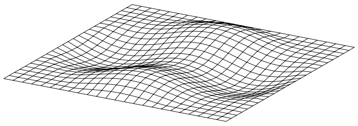
These three functions are plotted in Figure 12.8. Observe that $f_1(x)$ and $f_3(s)$ are well-aligned over the interval; whereas, $f_2(x)$ is not aligned with $f_1(x)$ or $f_3(x)$. In fact, careful inspection of Figure 12.8 will make it clear that for every value for x where $f_2(x)$ and the other two functions have a the same sign, there is a point where they have the same absolute values, but opposite signs. Thus, if the interpretation of the inner product is that it is a measure of alignment of the functions, we would expect that $\langle f_1(x), f_3(x)\rangle$ would be positive and that both $\langle f_2(x), f_1(x)\rangle$ and $\langle f_2(x), f_3(x)\rangle$ would be zero.

Computing the three inner products on the interval $[0, 2\pi]$ gives

$$\begin{aligned} \langle f_1(x), f_2(x)\rangle &= \int_0^{2\pi} (\sin x)(\sin 2x)\, dx \\ &= 0 \end{aligned}$$

by Proposition C.1.1. Computing

$$\begin{aligned} \langle f_1(x), f_3(x)\rangle &= \int_0^{2\pi} (\sin x)\, f_3(x) dx \\ &= \int_0^{\frac{\pi}{2}} x \sin x dx + \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} (\pi - x)(\sin x)\, dx + \\ &\quad \int_{\frac{3\pi}{2}}^{2\pi} (x - 2\pi)(\sin x)\, dx \\ &= 4, \end{aligned}$$



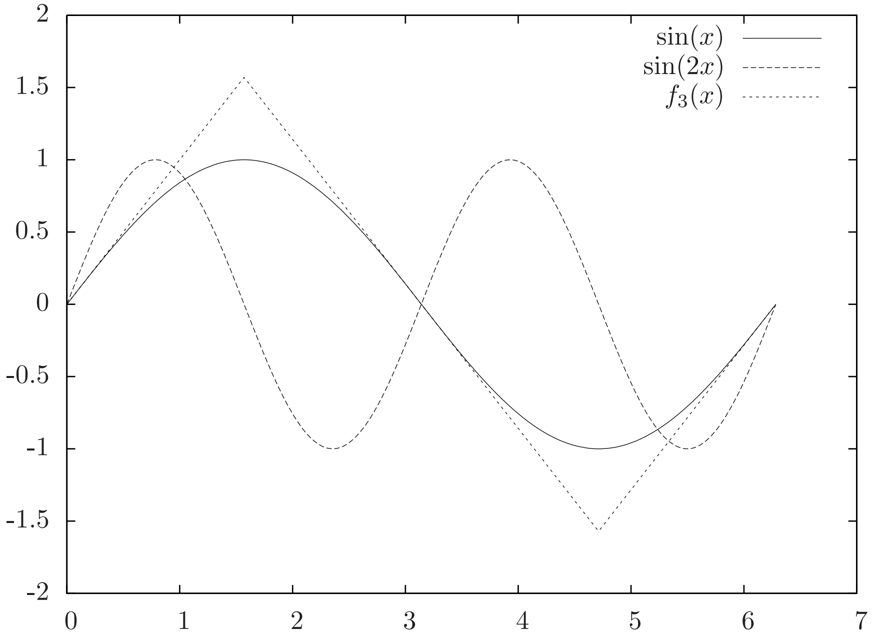


Figure 12.8. Three functions from Example 12.2.3. ■

which makes sense that it is nonzero since, by Figure 12.8, the functions are somewhat aligned. Also,

$$\begin{aligned}
\langle f_2(x), f_3(x)\rangle &= \int_0^{2\pi} f_3(x)\sin 2x dx \\
&= \int_0^{\frac{\pi}{2}} x\sin 2x dx + \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} (\pi - x)(\sin 2x)\, dx + \\
&\quad \int_{\frac{3\pi}{2}}^{2\pi} (x-2\pi)(\sin 2x)\, dx \\
&= 0.
\end{aligned}$$

12.2.3 Orthogonality

Recall from vector algebra that two vectors are orthogonal if their dot product is zero. In Euclidean space this corresponds to the angle between the vectors being 90°. Given two vectors with varying orientation, the dot product will be maximum when they are perfectly aligned (colinear and pointing in the same direction) and zero when they are perfectly "unaligned," *i.e.,* orthogonal. Using a similar notion, we will define two functions to be orthogonal when their inner product is zero, *i.e.,* the functions f and g are orthogonal if $\langle f, g\rangle = 0$.

For the present case, the most important class of functions that are orthogonal are trigonometric and have already been used. In particular

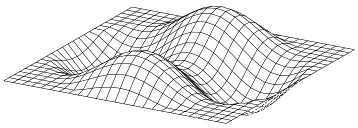
$$\begin{aligned}
\langle \sin\frac{n\pi x}{L}, \sin\frac{n\pi x}{L}\rangle &= \int_{-L}^{L} \sin\frac{n\pi x}{L}\sin\frac{m\pi x}{L} dx = \begin{cases} 0 & m \neq n \\ L & m = n \end{cases} \\
\langle \cos\frac{n\pi x}{L}, \cos\frac{n\pi x}{L}\rangle &= \int_{-L}^{L} \cos\frac{n\pi x}{L}\cos\frac{m\pi x}{L} dx = \begin{cases} 0 & m \neq n \\ L & m = n \end{cases} \\
\langle \sin\frac{n\pi x}{L}, \cos\frac{n\pi x}{L}\rangle &= \int_{-L}^{L} \sin\frac{n\pi x}{L}\cos\frac{m\pi x}{L} dx = 0 \qquad \forall m, n.
\end{aligned}$$

While a full investigation is beyond the scope of this text, it is worth noting that there are other sets of orthogonal functions as well, including Legendre polynomials, Hermite polynomials, Chebyshev polynomials and Bessell functions.

12.2.4 The general Fourier series

Given a function, $f(x)$, with period $T = 2L$, we now have all the tools to be able to express it as a Fourier series of the forjm

$$f(x) = \sum_{n=0}^{\infty} a_n \sin\frac{n\pi x}{L} + b_n \cos\frac{n\pi x}{L}.$$



To find the coefficients, multiply by $\sin\frac{m\pi x}{L}$ for the a_n and multiply by $\cos\frac{m\pi x}{L}$ for the b_n and integrate from $-L$ to L with respect to x. Due to the orthogonality of the sine and cosine functions, all the terms in the series will vanish except for one of them, which will allow us to solve for the coefficient. In particular,

$$\int_{-L}^{L} \sin\frac{m\pi x}{L}\left(\sum_{n=0}^{\infty} a_n \sin\frac{n\pi x}{L} + b_n \cos\frac{n\pi x}{L}\right) dx = \int_{-L} Lf(x)\sin\frac{m\pi x}{L} dx,$$

which gives

$$La_m = \int_{-L}^{L} f(x)\sin\frac{m\pi x}{L} dx$$

or

$$a_m = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{m\pi x}{L} dx.$$

Similarly,

$$b_m = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{m\pi x}{L} dx.$$

Note that since $\sin 0 = 0$, a_0 will always be equal to zero (which is why all the Fourier series for the wave equation started at $n = 1$). However, the same is not true for b_0, which will have to be evaluated for each series. In particular,

$$\begin{aligned}\int_{-L}^{L} \cos\frac{0\pi x}{L}\left(\sum_{n=0}^{\infty} a_n \sin\frac{n\pi x}{L} + b_n \cos\frac{n\pi x}{L}\right) dx &= \int_{-L}^{L} b_0 dx \\ &= 2Lb_0 \\ &= \int_{-L}^{L} f(x)dx,\end{aligned}$$

which gives

$$b_0 = \frac{1}{2L}\int_{-L}^{L} f(x)dx.$$

Note that this is "off" by a factor of two compared to all the other coefficients. Hence, it is conventional to write

$$f(x) = \frac{b_0}{2} + \sum_{n=1}^{\infty} a_n \sin\frac{n\pi x}{L} + b_n \cos\frac{n\pi x}{L} \tag{12.18}$$

where

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L} dx, \qquad n = 1, 2, 3, \ldots \tag{12.19}$$

and

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L} dx, \qquad n = 0, 1, 2, 3, \ldots \tag{12.20}$$

which allows us to use the same forumula for b_0 as the rest of the cosine coefficients.

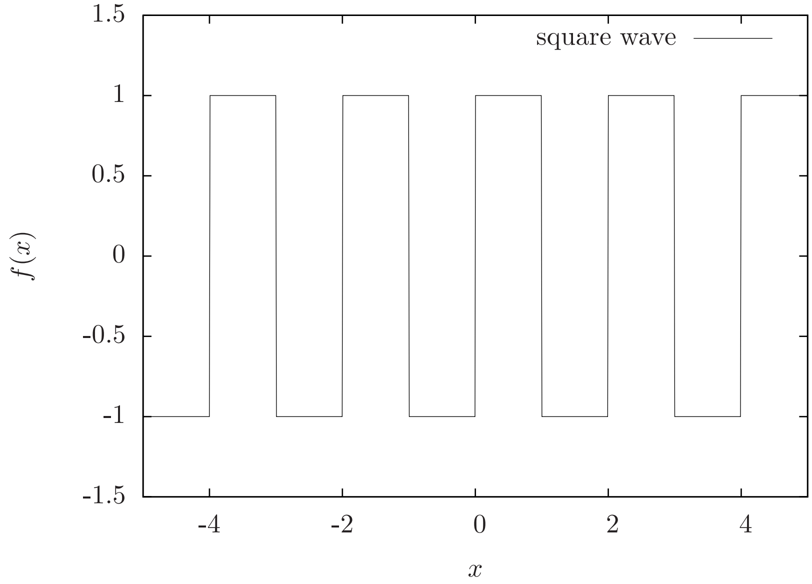


Figure 12.9. Square wave function for Example 12.2.5.

Remark 12.2.4 Since all the functions involved are periodic, the integrals in Equations 12.19 and 12.20 may have any limits as long as the difference between the upper and lower limit is $T = 2L$. ⋄

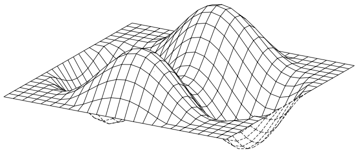
12.2.5 Examples of Fourier series

A few examples will be helpful at this point.

Example 12.2.5 Determine the Fourier series representation for the square wave function, given by

$$f(x) = \begin{cases} 1 & 0 < x \leq 1 \\ -1 & 1 < x \leq 2 \end{cases}$$

for $x \in (0, 2]$ and by $f(x+2) = f(x)$ for other $x \notin (0, 2]$, which is illustrated in Figure 12.9.



Computing the Fourier coefficients,

$$\begin{aligned}
a_n &= \int_0^2 f(x)\sin\frac{2n\pi x}{2}dx \\
&= \int_0^1 (1)\sin(n\pi x)\,dx + \int_1^2 (-1)\sin(n\pi x)\,dx \\
&= \frac{1}{n\pi}\left[-\cos(n\pi x)|_0^1 - -\cos(n\pi x)|_1^2\right] \\
&= \frac{1}{n\pi}\left[-\cos(n\pi) + 1 + \cos(2n\pi) - \cos(n\pi)\right] \\
&= \frac{2}{n\pi}\left[1 - \cos(n\pi)\right].
\end{aligned}$$

and

$$\begin{aligned}
b_n &= \int_0^2 f(x)\cos\frac{2n\pi x}{2}dx \\
&= \int_0^1 (1)\cos(n\pi x)\,dx + \int_1^2 (-1)\cos(n\pi x)\,dx \\
&= \frac{1}{n\pi}\left[\sin(n\pi x)|_0^1 - \sin(n\pi x)|_1^2\right] \\
&= 0.
\end{aligned}$$

Also,

$$\begin{aligned}
b_0 &= \int_0^2 f(x)dx \\
&= \int_0^1 1dx - \int_1^2 dx \\
&= x|_0^1 - x|_1^2 \\
&= 0.
\end{aligned}$$

Hence,

$$f(x) = \sum_{n=1}^{\infty} \frac{2}{n\pi}\left(1 - \cos(n\pi)\right)\sin(n\pi x).$$

Plots comparing the exact square wave to the first five, 10 and 50 terms, respectively, are illustrated in Figures 12.10 through 12.12. ■

Example 12.2.6 Determine the Fourier series representation for the sawtooth wave function given by

$$f(x) = \frac{x}{2}$$

for $x \in (0, 2]$ and $f(x+2) = f(x)$. This function is illustrated in Figure 12.13.

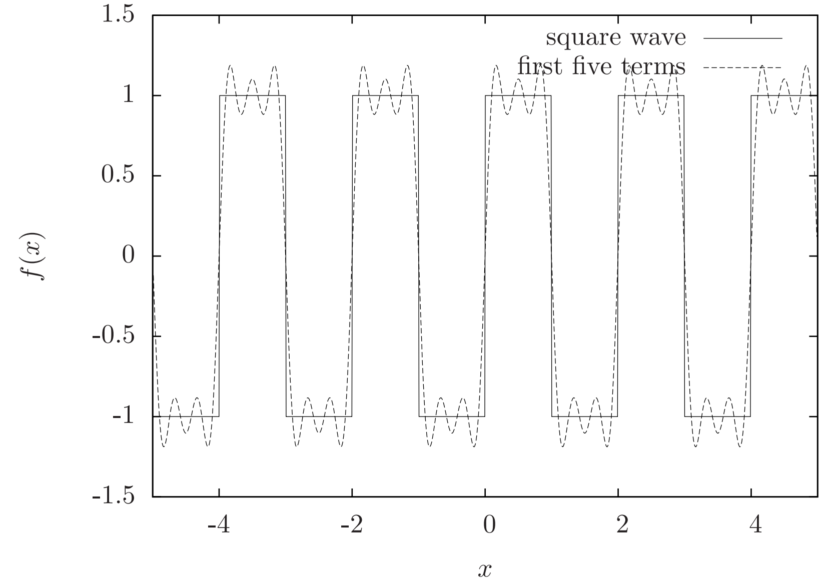


Figure 12.10. The first five terms in the Fourier series for the square wave in Example 12.2.5.

Computing the Fourier coefficients and noting that $T = 2$ so $L = 1$

$$\begin{aligned} a_n &= \frac{1}{1}\int_0^2 f(x)\sin\frac{n\pi x}{1}dx \\ &= \int_0^2 \frac{x}{2}\sin(n\pi x)\,dx \\ &= -\frac{\cos 2n\pi}{n\pi} \end{aligned}$$

and

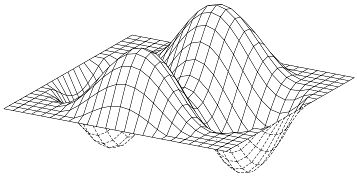
$$\begin{aligned} b_n &= \frac{1}{1}\int_0^2 f(x)\cos\frac{n\pi x}{1}dx \\ &= 0 \end{aligned}$$

for $n \neq 0$. For $n = 0$,

$$\begin{aligned} b_0 &= \frac{1}{1}\int_0^2 \frac{x}{2}dx \\ &= 1. \end{aligned}$$

Hence

$$f(x) = \frac{1}{2} + \sum_{n=1}^{\infty} -\frac{\cos(2n\pi)}{n\pi}\sin(n\pi x)\,.$$



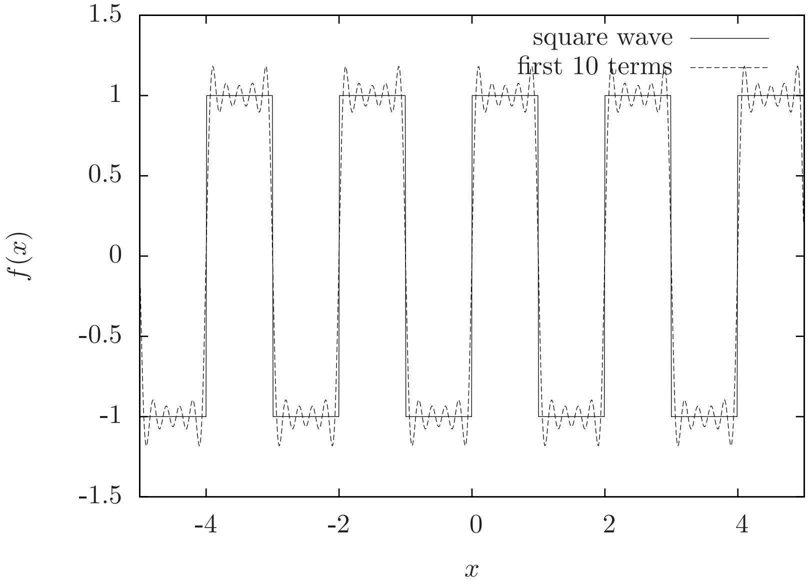


Figure 12.11. The first 10 terms in the Fourier series for the square wave in Example 12.2.5.

A plot of the first five and 10 terms of the series is illustrated in Figure 12.14.

Example 12.2.7 Compute the Fourier series for the function

$$f(x) = \begin{cases} x, & 0 < x \leq 1 \\ 1, & 1 < x \leq 2 \end{cases}$$

where $f(x+2) = f(x)$.

The function is periodic with period $T = 2$; hence, $L = 1$. The coeffi-

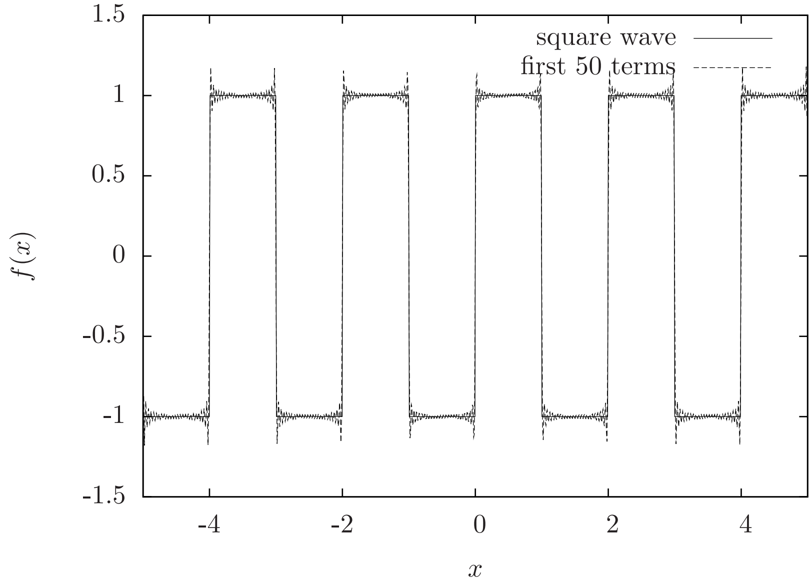
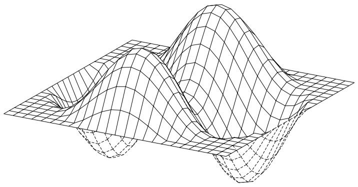


Figure 12.12. The first 50 terms in the Fourier series for the square wave in Example 12.2.5.

cients are given by

$$\begin{aligned}
a_n &= \frac{1}{1}\int_{-1}^{1} f(x)\sin\frac{n\pi x}{1}dx \\
&= \int_{-1}^{0} 1\sin(n\pi x)\,dx + \int_{0}^{1} x\sin(n\pi x)\,dx \\
&= -\frac{1}{n\pi}\cos(n\pi x)\bigg|_{-1}^{0} - \left(\frac{1}{n\pi}x\cos(n\pi x)\right)\bigg|_{0}^{1} + \frac{1}{n\pi}\int_{0}^{1}\cos(n\pi x)\,dx \\
&= -\frac{1}{n\pi}\cos(n\pi x)\bigg|_{-1}^{0} - \left(\frac{1}{n\pi}x\cos(n\pi x)\right)\bigg|_{0}^{1} + \frac{1}{n^2\pi^2}\sin(n\pi x)\bigg|_{0}^{1} \\
&= -\frac{1}{n\pi}(1-\cos(-n\pi)) - \frac{1}{n\pi}(\cos(n\pi)-0) + \frac{1}{n^2\pi^2}(0-0) \\
&= -\frac{1}{n\pi}
\end{aligned}$$



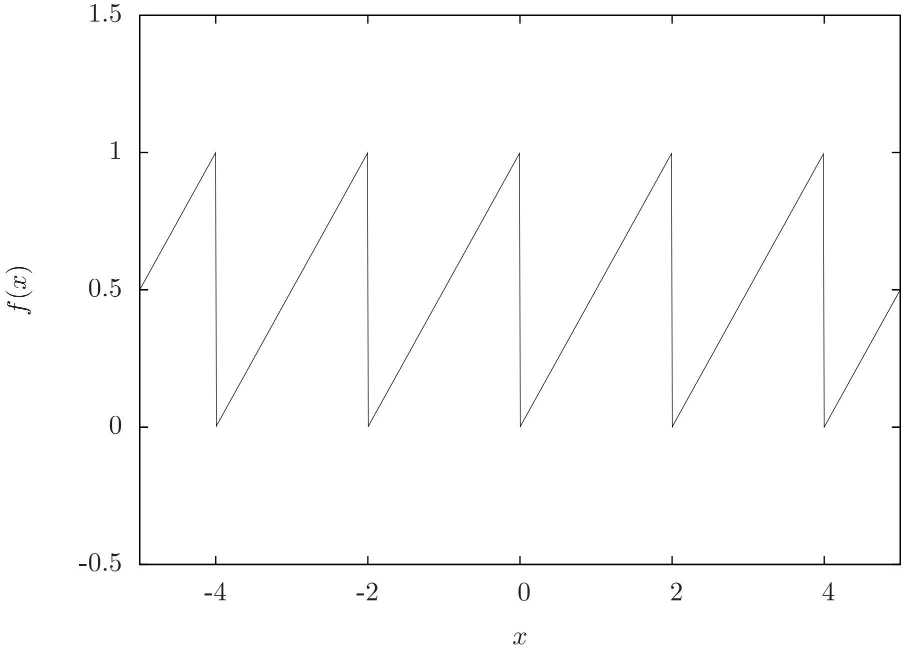


Figure 12.13. Sawtooth wave for Example 12.2.6.

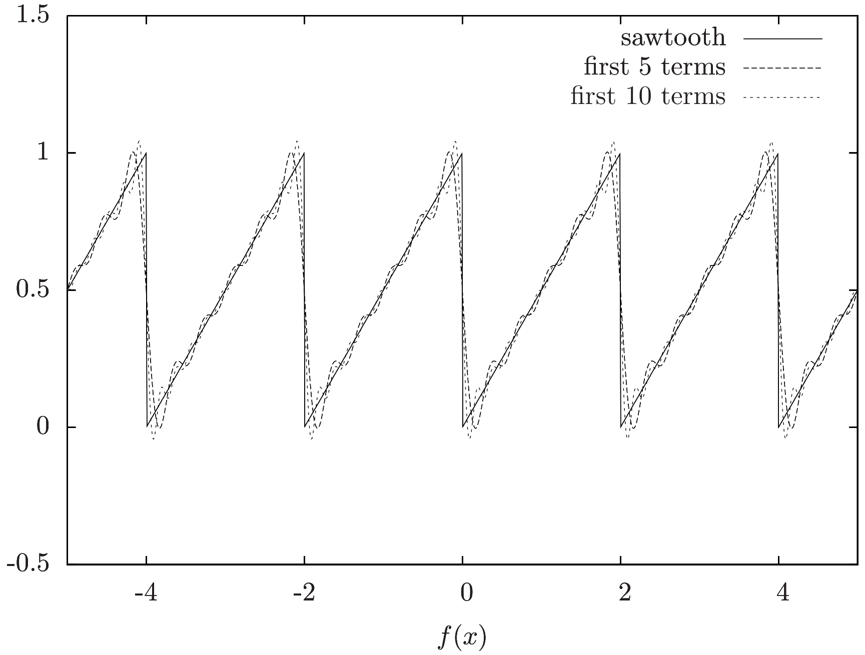


Figure 12.14. First five and 10 terms in the Fourier series for the sawtooth function in Example 12.2.6. ■

and

$$\begin{aligned}
b_n &= \frac{1}{1}\int_{-1}^{1} f(x)\cos\frac{n\pi x}{1}dx \\
&= \int_{-1}^{0} 1\cos(n\pi x)\,dx + \int_{0}^{1} x\cos(n\pi x)\,dx \\
&= \frac{1}{n\pi}\sin(n\pi x)\Big|_{-1}^{0} + \frac{1}{n\pi}x\sin(n\pi x)\Big|_{0}^{1} - \frac{1}{n\pi}\int_{0}^{1}\sin(n\pi x)\,dx \\
&= \frac{1}{n\pi}\sin(n\pi x)\Big|_{-1}^{0} + \frac{1}{n\pi}x\sin(n\pi x)\Big|_{0}^{1} + \frac{1}{n^2\pi^2}\cos(n\pi x)\Big|_{0}^{1} \\
&= \frac{1}{n^2\pi^2}\left(\cos(n\pi) - 1\right).
\end{aligned}$$

The b_0 coefficient must be computed separately,

$$\begin{aligned}
b_0 &= \int_{-1}^{0}(1)(1)\,dx + \int_{0}^{1} xdx \\
&= x|_{-1}^{0} + \frac{1}{2}x^2\Big|_{0}^{1} \\
&= 0 - (-1) + \frac{1}{2} \\
&= \frac{3}{2}.
\end{aligned}$$

Hence,

$$f(x) = \frac{3}{4} + \sum_{n=1}^{\infty}\left(-\frac{1}{n\pi}\sin(n\pi x) + \frac{\cos(n\pi) - 1}{n^2\pi^2}\cos(n\pi x)\right).$$

A plot of $f(x)$ as well as partial sums of the series including the first 10 and 20 terms of the series is illustrated in Figure 12.15. ■

12.2.6 Forced oscillations with discontinuous forcing functions

12.3 The Heat Equation

12.3.1 Derivation and interpretation

Hence, the one dimensional heat conduction equation is give by

$$\alpha^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}.$$

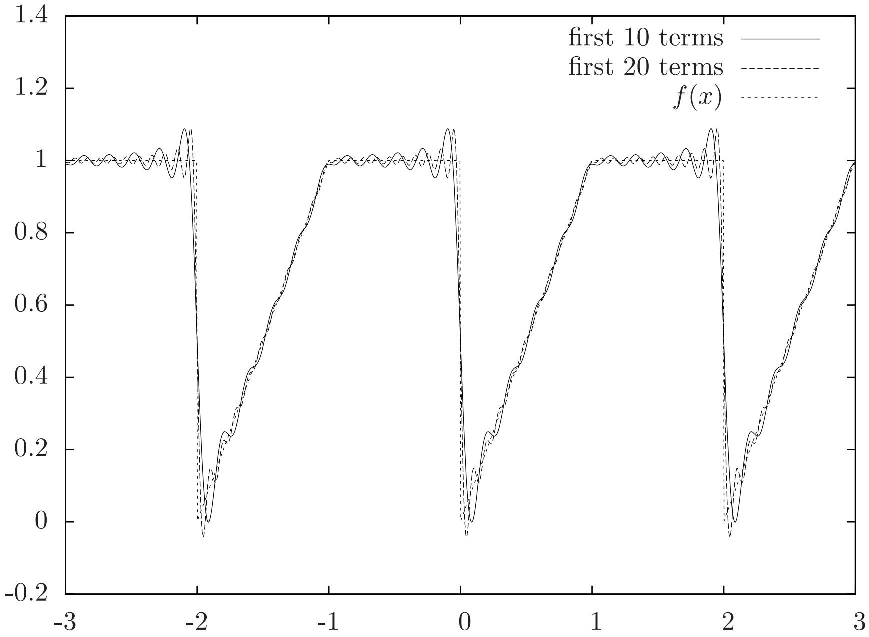


Figure 12.15. Fourier series for Example 12.2.7.

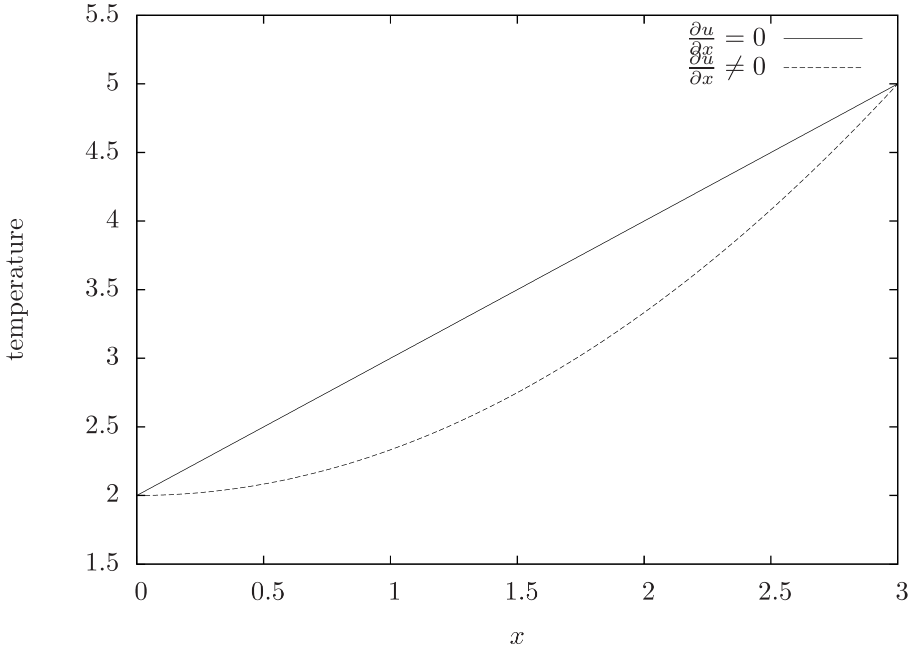


Figure 12.16. Head condution with zero and nonzero temperature curvature.

The mathematical interpreation of the heat equation is that the temperature at a given point, x will change if the curvature of the temperature profile is nonzero. Figure 12.16 contains two temperature profiles. In the one with no curvature, the rate of heat condution along the entire bar will be constant. In the case where the temperature profile has a nonzero, there will be a higher rate of heat conduction where the gradient is steep and a lower rate where it is less steep. Thus, for the curved temperature profile, there will be more heat condution from the left to the center than there will be from the center to the right boundary, the consequence of which will be that the temperature in the center of the bar will increase. In fact, as will be shown subsequently, the steady-state solution is the solution with no curvature, *i.e.,* a straight line. Hence, the curved solution will approach the straight solution as $t \to \infty$.

12.3.2 Solution to the heat equation with homogeneous boundary conditions

The usual approach to solve the heat equation is to solve it with *homogeneous boundary conditions* and then solve it with nonhomogeneous boundary conditions. Homogeneous boundary conditions are where boundary conditions are

$$u(0,t) = u(L,t) = 0,$$

i.e., the temperature at both ends is zero. This is not a very realistic situation, but we will use the solution for the homogeneous boundary conditions as part of the solution to the more realistic nonhomogeneous case.

The complete problem statement includes the differential equation, the boundary conditions as well as an initial temperature profile:

$$\begin{aligned} \alpha^2 \frac{\partial^2 u}{\partial x^2} &= \frac{\partial u}{\partial t} \\ u(0,t) &= 0 \\ u(L,t) &= 0 \\ u(x,0) &= f(x). \end{aligned}$$

The approach is exactly the same as for the wave equation. Assuming

$$u(x,t) = X(x)T(t)$$

and substituting into Equation 12.3.1 gives

$$\alpha^2 X''(x)T(t) = X(x)T'(t).$$

As before, this is separable, so

$$\frac{X''(x)}{X(x)} = \frac{1}{\alpha^2}\frac{T'(t)}{T(t)},$$

and since the left hand side is a function only of x and the right hand side is only a function of t, and x and t are independent, then each side must be equal to a constant. Hence,

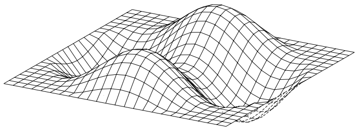
$$\frac{X''(x)}{X(x)} = \frac{1}{\alpha^2}\frac{T'(t)}{T(t)} = -\lambda$$

or

$$\begin{aligned} X''(x) + \lambda X(x) &= 0 \\ T'(t) + \alpha^2 \lambda T(t) &= 0. \end{aligned}$$

This is similar to the wave equation except that the equation for $T(t)$ is a first order equation instead of second order. This should make sense since we would not expect that the temperature profile in a bar would exhibit solutions that are oscillatory, which may the case for a second order equation.

We will proceed as before by applying the boundary conditions to determine λ, which will give an infinit number of solutions for $X(x)$. We may then use the infinite number of solutions to satisfy the initial temperature profile by using a Fourier series. In fact, the homogeneous boundary conditions give rise to exactly



ths same case as for the wave equation. In particular, the general solution for $X(x)$ is

$$X(x) = c_1 \sin \sqrt{\lambda x} + c_2 \cos \sqrt{\lambda x}$$

and the boundary conditions require

$$\begin{aligned} u(0,t) = 0 \quad &\Longrightarrow \quad X(0) = 0 \\ &\Longrightarrow \quad c_2 = 0, \end{aligned}$$

and

$$\begin{aligned} u(L,t) = 0 \quad &\Longrightarrow \quad X(L) = 0 \\ &\Longrightarrow \quad \lambda = \frac{n^2\pi^2}{L^2}, \quad n = 1, 2, 3, \ldots . \end{aligned}$$

So, we have

$$X(x) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{L}.$$

Since the general solution for $T(t)$ is then

$$\begin{aligned} T(t) &= e^{-\alpha^2 \lambda t} \\ &= e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \end{aligned}$$

the general solution to Equation 12.3.1 is

$$u(x,t) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{L} e^{-\left(\frac{\alpha n \pi}{L}\right)^2 t}. \tag{12.21}$$

Given some initial temperature profile, at $t = 0$ the exponential term is one and the initial profile may be satisfied by a Fourier series, *i.e.,*

$$\begin{aligned} u(x,0) &= \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{L} e^{-\left(\frac{\alpha^2 n \pi}{L}\right)^2 0} \\ &= \sum_{n=1} \infty c_n \sin \frac{n\pi x}{L} \\ &= f(x). \end{aligned}$$

The coefficients are determined by expoiting orthogonality as before. In particular, multiplying by $\sin \frac{m\pi x}{L}$ and integrating from 0 to L with respect to x gives

$$\int_0^L \sin \frac{m\pi x}{L} \sum_{n=1} \infty c_n \sin \frac{n\pi x}{L} dx = \int_0^L \sin \frac{m\pi x}{L} f(x) dx.$$

Since the sine functions are orthogonal except for the case where $n = m$, the infinite series reduces to one term, which gives

$$\int_0^L \sin \frac{m\pi x}{L} \sin \frac{m\pi x}{L} dx = \int_0^L \sin \frac{m\pi x}{L} f(x) dx.$$

Evaluating the integral on the left hand side gives what is exactly the same answer as before for the wave equation; namely,

$$c_n = \frac{2}{L}\int_0^L \sin\frac{n\pi x}{L} f(x)dx.$$

Hence, for homogeneous boundary conditions

$$u(0,t) = u(L,t) = 0$$

and initial condition

$$u(x,0) = f(x)$$

we have the general solution

$$u(x,t) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi x}{L}\right)\exp\left(-\left(\frac{\alpha n\pi}{L}\right)^2 t\right). \qquad (12.22)$$

Example 12.3.1 Determine the solution to

$$4\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}$$

with

$$u(0,t) = u(10,t) = 0$$

and

$$u(x,0) = \begin{cases} x & 0 < x \le 5 \\ 10 - x & 5 < x \le 10 \end{cases}$$

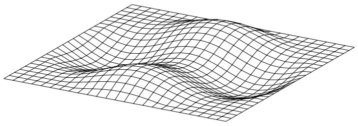
This is the case where $\alpha = 2$ and $L = 10$ and $u(x,0)$ is as illustrated in Figure 12.17, and the solution is simply given by substituting into Equation 12.22.

The only work is to determine the coeffiients in the Fourier series to satisfy the initial condition,

$$\begin{aligned} c_n &= \frac{2}{L}\int_0^L \sin\frac{n\pi x}{L} f(x)dx \\ &= \frac{2}{10}\left[\int_0^5 x\sin\frac{n\pi x}{10}dx + \int_5^{10}(10-x)\sin\frac{n\pi x}{L}dx.\right] \end{aligned}$$

Using the fact that

$$\int_a^b x\sin cx dx = -\frac{1}{c}x\cos cx\Big|_a^b + \frac{1}{c^2}\sin cx\Big|_a^b$$



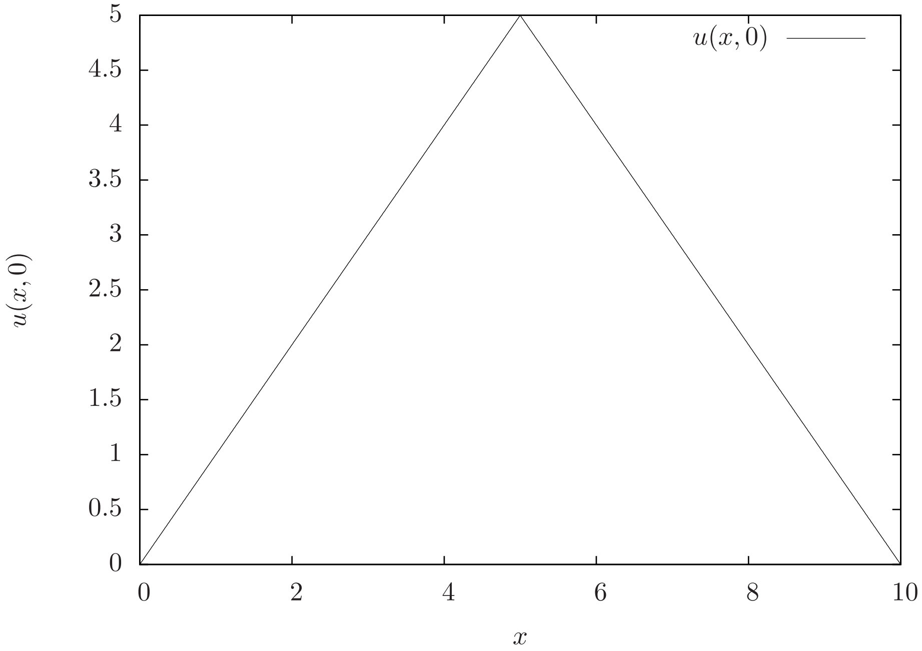


Figure 12.17. Initial temperature profile for Example 12.3.1.

the coeffients are

$$\begin{aligned}
c_n &= \frac{1}{5}\Bigg[-\frac{10}{n\pi} x \cos\frac{n\pi x}{10}\Bigg|_0^5 + \left(\frac{10}{n\pi}\right)^2 \sin\frac{n\pi x}{10}\Bigg|_0^5 + \\
&\quad -10\frac{10}{n\pi}\cos\frac{n\pi x}{10}\Bigg|_5^{10} + \\
&\quad \frac{10}{n\pi} x\cos\frac{n\pi x}{10}\Bigg|_5^{10} - \left(\frac{10}{n\pi}\right)^2 \sin\frac{n\pi x}{10}\Bigg|_5^{10}\Bigg] \\
&= \frac{2}{n\pi}\left[-5\cos\frac{n\pi}{2} - 0 + \frac{10}{n\pi}\left(\sin\frac{n\pi}{2} - 0\right)\right. \\
&\quad \left. -\,10\left(\cos n\pi - \cos\frac{n\pi}{2}\right) + 10\cos n\pi - 5\cos\frac{n\pi}{2} - \frac{10}{n\pi}\left(0 - \sin\frac{n\pi}{2}\right)\right] \\
&= \frac{40}{n^2\pi^2}\sin\frac{n\pi}{2}.
\end{aligned}$$

A plot of the solution for various times is illustrated in Figure 12.18. ■

12.3.3 Solution to the heat equation with inhomogeneous boundary conditions

Of course, the boundary conditions for the heat condution equation are seldom both zero. Let us now consider the case where

$$\begin{aligned}
u(0,t) &= T_1 \\
u(L,t) &= T_2.
\end{aligned}$$

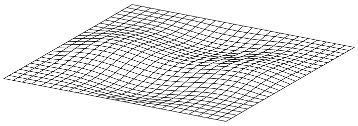
From section 12.3.1, the steady state solution will be

$$\begin{aligned}
\lim_{t\to\infty} u(x,t) &= \frac{T_2 - T_1}{L}x + T_1 \\
&= u_{ss}(x)
\end{aligned}$$

which is simply a straight line from $u(0,t) = T_1$ at $x = 0$ to $u(L,t) = T_2$ at $x = L$. Since there is no curvature, this solution will satisfy the heat equation since it is constant in time. It also satisfies the boundary conditions.

Since we have the solution to homogeneous boundary conditions from section 12.3.2 given by

$$u_h(x,t) = \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{L} e^{-\left(\frac{\alpha n\pi}{L}\right)^2 t}$$



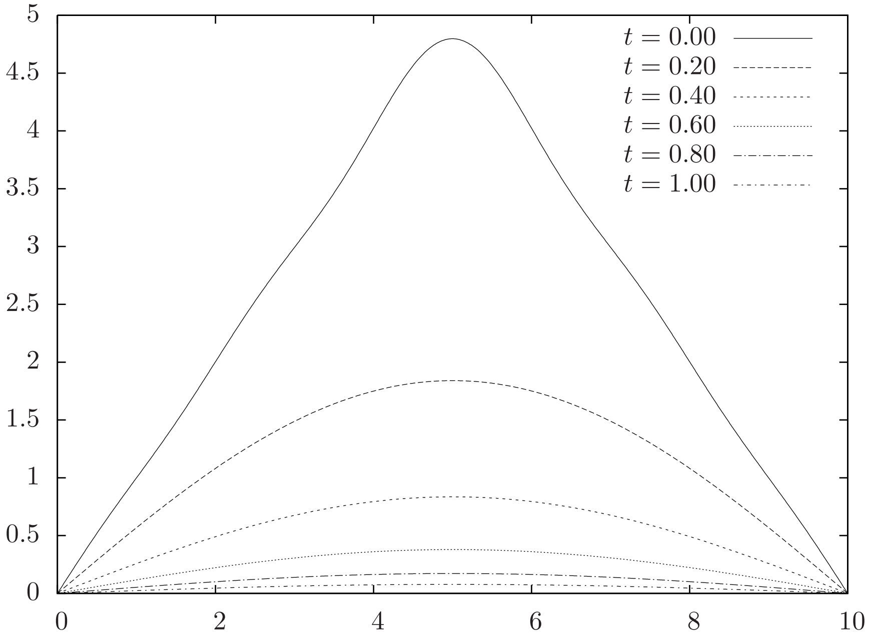


Figure 12.18. Solution for heat equation in Example 12.3.1 for various times.

and the steady-state solution from Equation 12.23, it makes sense to try to add them to find the complete solution. Along these lines, let

$$\begin{aligned}\hat{u}(x,t) &= u_h(x,t)+u_{ss}(x)\\ &= \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{L} e^{-\left(\frac{\alpha n\pi}{L}\right)^2 t} + \frac{T_2-T_1}{L}x+T_1.\end{aligned}$$

We need to

1. check that it satisfies the heat equation;
2. check that it satisfies the boundary conditions; and,
3. find equations for the c_n so that it satisfies the initial conditions.

For the coefficients, substituting $t=0$ into the general solution gives

$$\begin{aligned}u(x,0) &= \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{L} e^{-\left(\frac{\alpha n\pi}{L}\right)^2 0} + \frac{T_2-T_1}{L}x+T_1\\ &= \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{L} + \frac{T_2-T_1}{L}x+T_1\\ &= f(x).\end{aligned}$$

Thus, at $t=0$ we may write

$$\sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{L} = f(x) - \frac{T_2-T_1}{L}x - T_1.$$

If we let

$$\hat{f}(x) = f(x) - \frac{T_2-T_1}{L} - T_1$$

we may compute the Fourier coefficients in the usual manner:

$$c_n = \frac{2}{L}\int_0^L \hat{f}(x)\sin\frac{n\pi x}{L}dx.$$

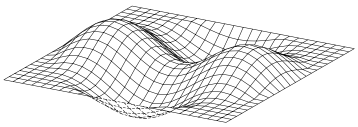
12.4 Laplace's Equation

Laplace's equation is

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \tag{12.23}$$

and represents the steady state temperature distribution in a rectangular domain. Various combinations of boundary conditions are possible and are fully explored in the exercises. In this section we will consider

$$\begin{aligned}u(0,y) &= 0\\ u(a,y) &= 0\\ u(x,0) &= 0\\ u(x,b) &= f(x).\end{aligned}$$



Assuming $u(x,y) = X(x)Y(y)$ and substituting into Equation 12.23 gives

$$X''(x)Y(y) + X(x)Y''(y) = 0 \qquad \Longleftrightarrow \qquad \frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)}.$$

Since the left hand side is only a function of x and the right hand side is only a function of y and x and y are independent variables, each side must be equal to the same constant, *i.e.,*

$$\frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)} = -\lambda.$$

At this point we do not know whether λ must be positive or negative. We will assume that it is real since it is a coefficient in the ordinary differential equations for $X(x)$ and $Y(y)$ and we are seeking real solution to Equation 12.23.

Based upon our experience with the wave and heat conduction equation, it is reasonable to expect that we will have to use a Fourier series to satisfy the boundary condition $u(x,b) = f(x)$. Hence, we will consider the $X(x)$ equation first. Specifically, we have

$$X'' + \lambda X(x) = 0 \tag{12.24}$$

with

$$\begin{aligned} X(0) &= 0 \\ X(a) &= 0. \end{aligned}$$

Regardless of the value of λ, the general solution to Equation 12.24 is

$$X(x) = c_1 e^{\sqrt{-\lambda}x} + c_2 e^{-\sqrt{-\lambda}x}.$$

The boundary condition at $x = 0$ gives

$$X(0) = c_1 + c_2 = 0 \qquad \Longleftrightarrow \qquad c_1 = -c_2.$$

Hence,

$$X(x) = c_1 \left(e^{\sqrt{-\lambda}x} - e^{\sqrt{-\lambda}x} \right).$$

The boundary condition at $x = a$ requires that

$$X(a) = c_1 \left(e^{\sqrt{-\lambda}a} - e^{-\sqrt{-\lambda}a} \right) = 0.$$

If $\lambda < 0$, then $\sqrt{-\lambda}$ is real. Hence, either $c_1 = 0$ or $e^{\sqrt{-\lambda}a} = e^{-\sqrt{-\lambda}a}$. If $c_1 = 0$, then $u(x,y) = 0$ and the solution can not satisfy the boundary condition at $y = b$ unless it happens that $f(x) = 0$. Furthermore, it is not possible for $e^{\sqrt{-\lambda}a} = e^{-\sqrt{-\lambda}a}$ if $\lambda < 0$. So, then either $\lambda = 0$ or $\lambda > 0$.

In the case where $\lambda = 0$, Equation 12.24 is of the form

$$X''(x) = 0$$

which has a general solution

$$X(x) = c_1 x + c_2.$$

Using the boundary conditions gives

$$\begin{aligned} X(0) &= 0 \qquad \Longrightarrow \qquad c_1 = 0 \\ X(a) &= 0 \qquad \Longrightarrow \qquad c_2 = 0. \end{aligned}$$

Again, unless $f(x) = 0$, this will not work.

Rapidly running out of options, consider the case where $\lambda > 0$. In that case

$$\begin{aligned} X(x) &= c_1 e^{\sqrt{-\lambda}x} + c_2 e^{-\sqrt{-\lambda}x} \\ &= c_1 e^{i\sqrt{\lambda}x} + c_2 e^{-i\sqrt{\lambda}x} \\ &= (c_1 + c_2)\cos\sqrt{\lambda}x + i\,(c_1 - c_2)\sin\sqrt{\lambda}x \\ &= \hat{c}_1 \cos\sqrt{\lambda}x + \hat{c}_2 \sin\sqrt{\lambda}x. \end{aligned}$$

Applying the boundary condition at $x = 0$ gives

$$X(0) = 0 \qquad \Longrightarrow \qquad \hat{c}_1 = 0.$$

At $x = a$ the boundary condition requires that either $\hat{c}_2 = 0$ or

$$\sin\sqrt{\lambda}a = n\pi \quad n = 1, 2, 3, \ldots.$$

As before if $\hat{c}_2 = 0$, then $u(x, y) = 0$ which can not satisfy the boundary condition at $y = b$ unless $f(x) = 0$. So, finally we have that

$$\lambda = \left(\frac{n\pi}{a}\right)^2.$$

Now, substituting the value of λ into the equation for $Y(y)$ gives

$$Y''(y) - \left(\frac{n\pi}{a}\right)^2 Y(y) = 0,$$

which has a general solution

$$Y(y) = k_1 e^{\frac{npiy}{a}} + k_2 e^{-\frac{n\pi y}{a}}.$$

Applying the boundary condition at $y = 0$ gives

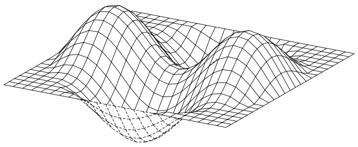
$$Y(0) = k_1 + k_2 = 0 \qquad \Longrightarrow \qquad k_1 = -k_2.$$

Hence,

$$Y(y) = k_1 \left(e^{\frac{n\pi y}{a}} - e^{-\frac{n\pi y}{a}}\right).$$

We have an infinite number of general solutions for $X(x)$ and one solution for $Y(y)$. Combining them gives

$$u(x, y) = \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{a}\left(e^{\frac{n\pi y}{a}} - e^{-\frac{n\pi y}{a}}\right).$$



To satisfy the boundary condition at $y = b$ we need that

$$u(x, b) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{a} \left(e^{\frac{n\pi b}{a}} - e^{-\frac{n\pi b}{a}} \right) = f(x).$$

At this point it is hopefully obvious what must be done: multiply by $\sin \frac{m\pi x}{a}$ and integrate from 0 to a with respect to x. Doing so gives

$$\int_0^a \left(\sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{a} \left(e^{\frac{n\pi b}{a}} - e^{-\frac{n\pi b}{a}} \right) \right) \sin \frac{m\pi x}{a} dx = \int_0^a f(x) \sin \frac{m\pi x}{a} dx.$$

Rearranging the left hand side gives

$$\sum_{n=1}^{\infty} c_n \left(e^{\frac{n\pi b}{a}} - e^{-\frac{n\pi b}{a}} \right) \int_0^a \sin \frac{n\pi x}{a} \sin \frac{m\pi x}{a} dx = \int_0^a f(x) \sin \frac{m\pi x}{a} dx,$$

and due to the orthogonality of the sine functions the only nonzero term in the infinite series is the case where $n = m$, so

$$c_m \left(e^{\frac{m\pi b}{a}} - e^{-\frac{m\pi b}{a}} \right) \frac{a}{2} = \int_0^a f(x) \sin \frac{m\pi x}{a} dx$$

or, finally,

$$c_n = \frac{2}{a} \frac{1}{e^{\frac{n\pi b}{a}} - e^{-\frac{n\pi b}{a}}} \int_0^a f(x) \sin \frac{n\pi x}{a} dx.$$

So, in summary, the solution to

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

with boundary conditions

$$\begin{aligned} u(0, y) &= 0 \\ u(a, y) &= 0 \\ u(x, 0) &= 0 \\ u(x, b) &= f(x). \end{aligned}$$

is

$$u(x, y) = \sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{a} \left(e^{\frac{n\pi y}{a}} - e^{-\frac{n\pi y}{a}} \right).$$

where

$$c_n = \frac{2}{a} \frac{1}{e^{\frac{n\pi b}{a}} - e^{-\frac{n\pi b}{a}}} \int_0^a f(x) \sin \frac{n\pi x}{a} dx.$$

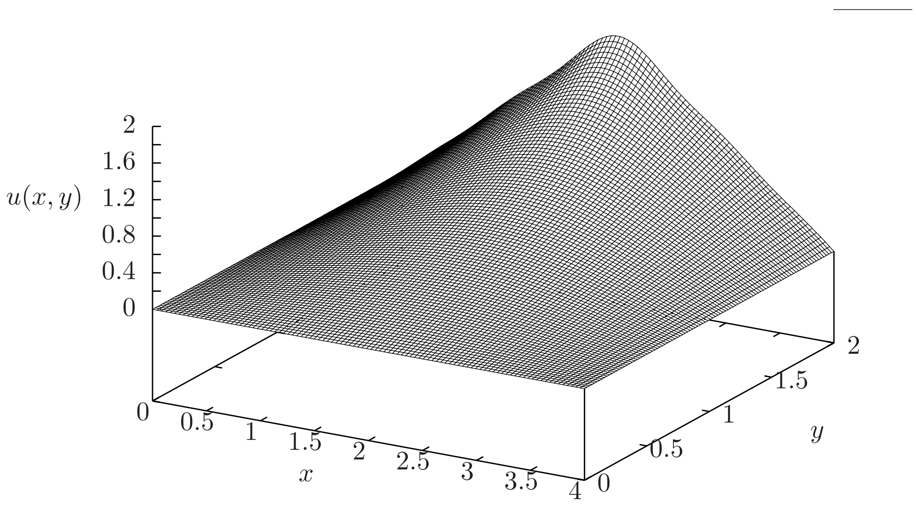


Figure 12.19. Approximate solution to Laplace's equation from Example 12.4.1 using a partial sum containing the first 10 terms.

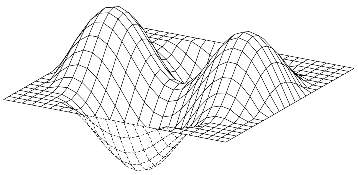
Example 12.4.1 Find the solution to Laplace's equation in a rectanglular domain where $a = 4$ and $b = 2$ and

$$u(x,2) = \begin{cases} x, & 0 < x \le 1 \\ 2 - x, & 1 < x \le 2. \end{cases}$$

Evaluating the integrals for the Fourier coefficients gives

$$c_n = \frac{64 e^{\frac{n\pi}{2}} \cos\left(\frac{n\pi}{4}\right) \sin^3\left(\frac{n\pi}{4}\right)}{\left(-1 + e^{n\pi}\right) n^2 \pi^2}$$

A plot of the solution containing the first 10 terms of the series is illustrated in Figure 12.19. ■



12.5 Vibrating Membranes

12.5.1 The Two Dimensional Wave Equation in Rectangular Coordinates

The two dimensional wave equation is given by

$$\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial y^2} = \frac{1}{\alpha^2} \frac{\partial^2 u}{\partial t^2}.$$

12.5.2 The Two Dimensional Wave Equation in Polar Coordinates

In polar coordinates, the two dimensional wave equation is

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = \frac{1}{\alpha^2} \frac{\partial^2 u}{\partial t^2} \tag{12.25}$$

with boundary condition

$$u(\hat{r}, \theta, t) = 0 \tag{12.26}$$

and initial conditions

$$\begin{aligned} u(r, \theta, 0) &= f(r, \theta) \\ \left.\frac{\partial u}{\partial t}\right|_{r,\theta,0} &= g(r, \theta). \end{aligned}$$

Assuming a solution of the form

$$u(r, \theta, t) = R(r)\Theta(\theta)T(t)$$

and substituting into Equation 12.25 gives

$$R''(r)\Theta(\theta)T(t) + \frac{1}{r}R'(r)\Theta(\theta)T(t) + \frac{1}{r^2}R(r)\Theta''(\theta)T(t) = \frac{1}{\alpha^2}R(r)\Theta(\theta)T''(t)$$

and dividing by $R(r)\Theta(\theta)T(t)$ gives

$$\frac{R''(r)}{R(r)} + \frac{1}{r} \frac{R'(r)}{R(r)} + \frac{1}{r^2} \frac{\Theta''(\theta)}{\Theta(\theta)} = \frac{1}{\alpha^2} \frac{T''(t)}{T(t)}.$$

Since the right side of the equation only depends on t and the left side depends only on r and θ, and all three variables are independent, both sides must be constant; hence,

$$\frac{R''(r)}{R(r)} + \frac{1}{r} \frac{R'(r)}{R(r)} + \frac{1}{r^2} \frac{\Theta''(\theta)}{\Theta(\theta)} = \frac{1}{\alpha^2} \frac{T''(t)}{T(t)} = -\lambda$$

where λ is a yet to be determined constant. Hence,

$$T''(t) + \alpha^2 \lambda T(t) = 0 \tag{12.27}$$

and

$$\frac{R''(r)}{R(r)}+\frac{1}{r}\frac{R'(r)}{R(r)}+\frac{1}{r^2}\frac{\Theta''(\theta)}{\Theta(\theta)}=-\lambda.$$

Multiplying by r^2 and rearranging gives

$$r^2\frac{R''(r)}{R(r)}+r\frac{R'(r)}{R(r)}+r^2\lambda=-\frac{\Theta''(\theta)}{\Theta(\theta)}.$$

Since the left side of this equation only depends on r and the right side only depends on θ and the variables are independent, these also must be equal to a constant, which is not necessarily the same as λ. Calling this constant γ, we have

$$r^2\frac{R''(r)}{R(r)}+r\frac{R'(r)}{R(r)}+r^2\lambda=-\frac{\Theta''(\theta)}{\Theta(\theta)}=\gamma.$$

Hence,

$$\Theta''(\theta)+\gamma\Theta(\theta)=0 \tag{12.28}$$

and

$$r^2R''(r)+rR'(r)+\left(r^2\lambda-\gamma\right)R(r)=0. \tag{12.29}$$

If we determine the solutions to Equations 12.27, 12.28 and 12.29, we will have a solution to Equation 12.25.

We will proceed has we have done before by finding the general solutions to the ordinary differential equations for $R(r)$, $\Theta(\theta)$ and $T(t)$ and applying the boundary conditions. While is appears that we only have one boundary condition given by Equation 12.26, there is also the fact that the solution for $\Theta(\theta)$ must be periodic, *i.e.,* $\Theta(\theta)=\Theta(\theta+2\pi)$. Thus, γ must be positive and

$$\Theta(\theta)=c_1\sin\sqrt{\gamma}\theta+c_2\cos\sqrt{\gamma}\theta.$$

In order for $\Theta\left(\theta+2\pi\right)=\Theta\left(\theta\right)$, $\sqrt{\gamma}$ must be an integer, or

$$\gamma=m^2,\qquad m=0,1,2,\ldots,$$

so

$$\Theta_n\left(\theta\right)=c_1\sin n\theta+c_2\cos n\theta,\qquad n=1,2,3,\ldots.$$

The ordinary differential equation for $R(r)$ is variable coefficient, so we must use a power series solution. Assuming

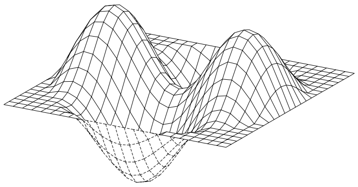
$$R(r)=r^k\sum_{n=0}^{\infty}a_nr^n=\sum_{n=0}^{\infty}a_nr^{n+k}$$

we have

$$R'(r)=(n+k)\sum_{n=0}^{\infty}a_nr^{n+k-1}$$

and

$$R''(r)=(n+k)\left(n+k-1\right)\sum_{n=0}^{\infty}a_nr^{n+k-2}.$$



Substituting into Equation 12.29 gives

$$(n+k)\,(n+k-1)\sum_{n=0}^{\infty} a_n r^{n+k} + (n+k)\sum_{n=0}^{\infty} a_n r^{n+k} + \left(r^2\lambda - \gamma\right)\sum_{n=0}^{\infty} a_n r^{n+k} = 0.$$

Collecting the first two terms and distributing the r^2 over the last sum gives

$$\left((n+k)^2 - \gamma\right)\sum_{n=0}^{\infty} a_n r^{n+k} + \lambda\sum_{n=0}^{\infty} a_n r^{n+k+2}.$$

Shifting the index of summation on the last sum gives

$$\left((n+k)^2 - \gamma\right)\sum_{n=0}^{\infty} a_n r^{n+k} + \lambda\sum_{n=2}^{\infty} a_{n-2} r^{n+k} = 0.$$

Hence,

$$a_n = \frac{\lambda}{(n+k)^2 - \gamma} a_{n-2}, \qquad n \geq 2$$

and

$$a_0\left((0+k)^2 - \gamma\right) = 0$$

and

$$a_1\left((1+k)^2 - \gamma\right) = 0.$$

Since $\gamma = m^2$ for $m = 0, 1, 2, \ldots,$, the a_0 equation requires that $k = \pm m$, for $m = 0, 1, 2, \ldots$. Then, if k makes the term multiplying a_0 zero, it can not make the term multiplying the a_1 term zero; hence, we need that $a_1 = 0$.

Define

$$J_m(\lambda r) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{2n+m} n!\,(m+n)!} (\lambda r)^{2n+m}.$$

complete...

So, finally we have

$$u(r,\theta,t) = \sum_{k=1}^{\infty}\sum_{m=0}^{\infty} J_m\left(\frac{\lambda_{m,k} r}{\hat{r}}\left(a_{m,k}\sin(m\theta) + \cos(m\theta)\right)\left(a\cos\sqrt{\lambda_{m,k}}t + b\cos\sqrt{\lambda_{m,k}t}\right)\right)$$

12.6 The Euler-Bernoulli Beam Equation

The *Euler-Bernoulli beam equation* is a partial differential equation describing small vibrations of beams. In contrast to strings, beams can support bending loads, which results in a higher order partial differential equation describing its motion.

12.6.1 Derivation of the Beam Equation

Consider the cantilever beam illustrated in Figure 12.20. Assume that the beam is subjected to a distributed load that may vary in time, $f(x,t)$, where the units of $f(x,t)$ are force per unit length. We will make the following assumptions about the manner in which the beam deflects.

Assumption 12.6.1 *1. Assume that the beam deflects in the vertical direction only and that the deflection of the beam in the vertical direction is small.*

2. Assume that the slope is also small.

3. Assume that any planar cross section of the beam remains planar when it is deflected.

Consider the coordinate axes illustrated in Figure 12.20 with the y–axis directed into the page. Since the beam deflects in the z–direction only, all deflections remain within the plane of the page. Define the *neutral plane* to be the plane before deformation whose length is not changed when the beam is deformed. In Figure 12.21 the top of the beam is extended and the bottom of the beam is compressed. The neutral plane is illustrated by a dashed line. Let $u(x,t)$ represent the deflection of the beam's neutral plane at location x at time t from its unloaded equilibrium position, as is illustrated in Figure 12.21.

Now we may restate the assumption that the deformations are small with the equation $u(x,t) \ll 1$ and the assumption that the slope is small by $\frac{\partial u}{\partial x}(x,t) \ll 1$. Having defined the neutral plane and coordinate axes we state another assumption.

Assumption 12.6.2

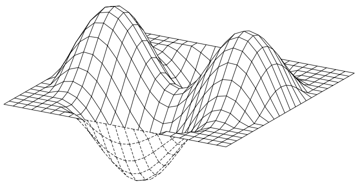
Assume that a cross section normal to the neutral plane does not change in height or width when the beam is deflected.

To derive the equation of motion, consider a small segment of the beam, as is illustrated in Figure 12.22. Let A be the cross sectional area of the beam and ρ the density. Newton's law in the vertical direction gives

$$\rho A \frac{\partial^2 u}{\partial t^2}(x,t)dx = V(x+dx,t) - V(x,t) + \frac{1}{2}\left(f(x,t) + f(x+dx,t)\right)dx, \tag{12.30}$$

where the total applied load is computed as the average of $f(x,t)$ and $f(x+dx,t)$ times the length of the segment, dx. Expanding $V(x+dx,t)$ in a Taylor series about x gives

$$V(x+dx,t) = V(x,t) + \frac{\partial V}{\partial x}(x,t)dx + \cdots$$



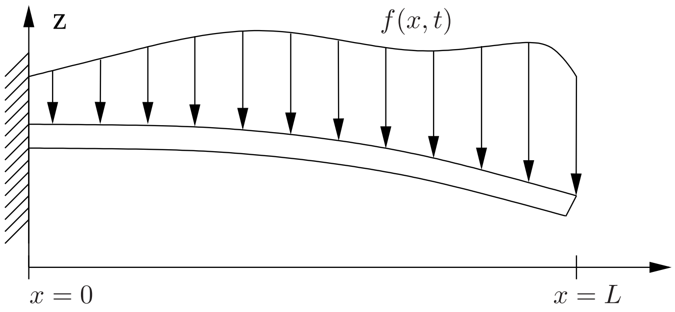


Figure 12.20. Loaded beam.

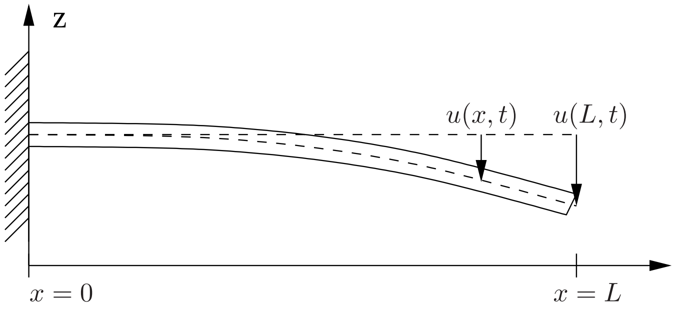


Figure 12.21. Deflected beam.

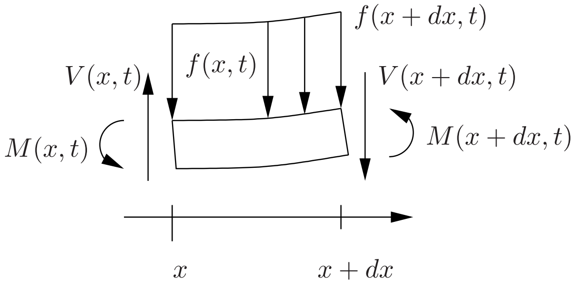


Figure 12.22. Small segment of a beam.

and similarly expanding $f(x+dx,t)$ gives

$$f(x+dx,t) = f(x,t) + \frac{\partial f}{\partial x}(x,t)dx + \cdots .$$

Keeping the higher order terms and substituting into Equation 12.30 gives

$$\rho A \frac{\partial^2 u}{\partial t^2}(x,t)dx = \frac{\partial V}{\partial x}dx + \frac{1}{2}\left(2f(x,t) + \frac{\partial f}{\partial x}dx\right)dx,$$

or

$$\rho A \frac{\partial^2 u}{\partial t^2}(x,t) = \frac{\partial V}{\partial x} + \frac{1}{2}\left(2f(x,t) + \frac{\partial f}{\partial x}\right)dx. \tag{12.31}$$

Taking the limit as $dx \to 0$ gives

$$\rho A \frac{\partial^2 u}{\partial t^2}(x,t) = \frac{\partial V}{\partial x}(x,t) + f(x,t). \tag{12.32}$$

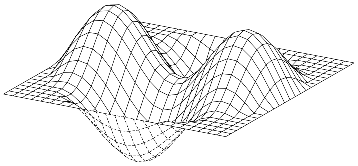
Since we are assuming the motion is only vertical, there is no angular acceleration,so the sum of the moments about any point must be zero. Computing the moments about the center of the right end of the beam segment in Figure 12.22 gives

$$M(x,t) - M(x+dx,t) - V(x,t)dx + \frac{1}{2}\left(f(x,t) + f(x+dx,t)\right)dx\frac{dx}{2} = 0,$$

where the moment due to the loading is approximated as the average load with an average moment arm of $\frac{dx}{2}$.

Using a Taylor series expansion M and f,

$$\begin{aligned} M(x+dx) &= M(x,t) + \frac{\partial M}{\partial x}(x,t)\,dx + \cdots \\ f(x+dx,t) &= f(x,t) + \frac{\partial f}{\partial x}(x,t)dx + \cdots \end{aligned}$$



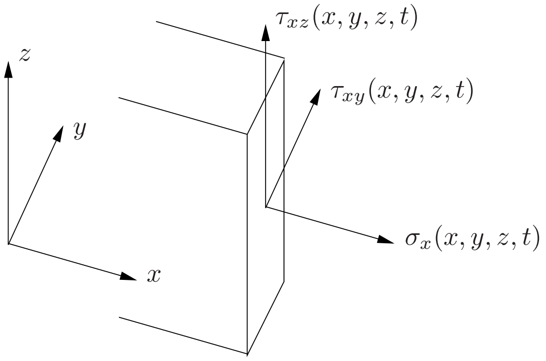


Figure 12.23. Cross section of a beam.

gives

$$-\frac{\partial M}{\partial x}(x,t)dx - V(x,t)dx + \frac{1}{2}\left(2f(x,t) + \frac{\partial f}{\partial x}(x,t)dx\right)\frac{dx^2}{2} = 0$$

or

$$-\frac{\partial M}{\partial x}(x,t) - V(x,t) + \frac{1}{2}\left(2f(x,t) + \frac{\partial f}{\partial x}(x,t)dx\right)\frac{dx}{2} = 0$$

Taking the limit as $dx \to 0$ gives

$$-\frac{\partial M}{\partial x}(x,t) = V(x,t), \tag{12.33}$$

or, substituting into Equation 12.31 gives

$$\rho A\frac{\partial^2 u}{\partial t^2}(x,t) = -\frac{\partial^2 M}{\partial x^2}(x,t) + f(x,t). \tag{12.34}$$

Now, consider a cross section of the beam, as is illustrated in Figure 12.23. The normal stress on the face of the cross section is denoted by $\sigma_x(x,y,z,t)$ and the two shear stresses are $\tau_{xy}(x,y,z,t)$ in the horizontal direction and $\tau_{xz}(x,y,z,t)$ in the vertical direction.

If we consider the normal stress over a small area of a cross section, as is illustrated in Figure 12.24, the moment due to the total force acting on that area is

$$dM = z\sigma_x(x,y,z,t)dA,$$

or integrating over the whole surface of the cross section

$$M(x,t) = \int\int z\sigma_x(x,y,z,t)dzdy, \tag{12.35}$$

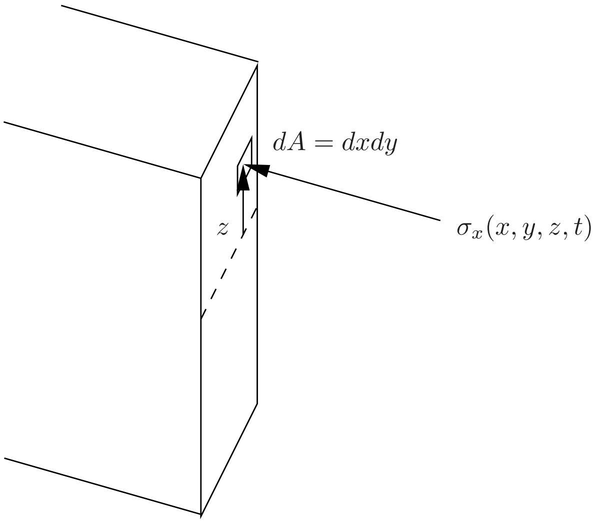


Figure 12.24. Moment due to normal stress over a small area.

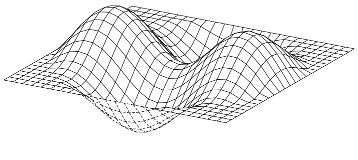
where the limits of integration are determined by the geometry of the cross section.

The basic constitutive law from solid mechanics is that normal stress and strain are related by

$$\sigma(x, y, z, t) = E\epsilon(x, y, z, t) \tag{12.36}$$

where E is the modulus of elasticity and has units of pascals, denoted by Pa where $1\text{Pa} = 1\frac{\text{N}}{\text{m}^2}$. Finally, to relate the strain to the deformation of the beam, consider the deflection of a small segment of the beam illustrated in Figure 12.25. Since the slope is small, $\theta(x, t) \approx \frac{\partial u}{\partial x}(x, t)$. Since strain is defined as the displacement per unit length, we have for the location z on the right face of the segment

$$\begin{aligned}
\epsilon_x(x, y, z, t) &= \frac{z\left(\sin\theta(x, t) - \sin\theta(x + dx, t)\right)}{dx} \\
&= \frac{z\left(\frac{\partial u}{\partial x}(x, t) - \frac{\partial u}{\partial x}(x + dx, t)\right)}{dx}(x, y, z, t) \\
&= \frac{z\left(\frac{\partial u}{\partial x}(x, t) - \left(\frac{\partial}{\partial x}(x, t) + \left(\frac{\partial u}{\partial x}(x, t)\right) dx\right)\right)}{dx}(x, t) \\
&= -z\frac{\partial^2 u}{\partial x^2}(x, t).
\end{aligned}$$



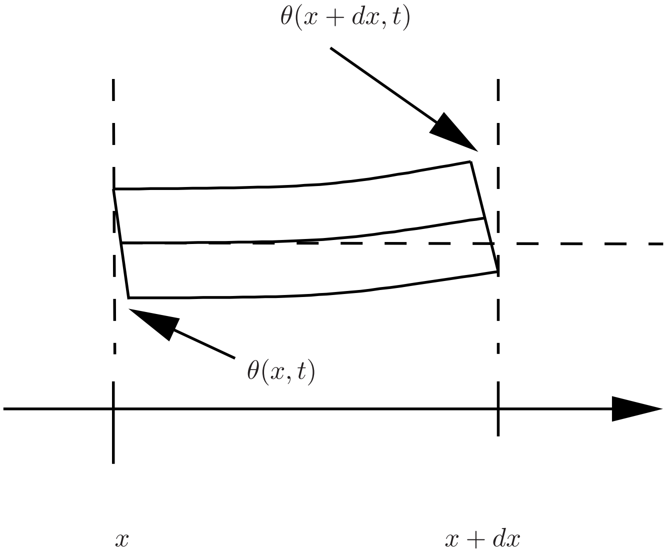


Figure 12.25. Strain relationship for small beam segment.

Substituting this into Equation 12.36, and using that in Equation **??** gives

$$\begin{aligned} M(x,t) &= \int\int -z^2 E \frac{\partial^2 u}{\partial x}(x,t)dydz \\ &= -E\frac{\partial^2 u}{\partial x^2}(x,t)\int\int z^2 dydz. \end{aligned}$$

Since the integral is the definition of the area moment of inertia, if we let

$$I(x) = \int\int z^2 dydz$$

we have

$$M(x,t) = -EI(x)\frac{\partial^2 u}{\partial x^2}(x,t)$$

and substituting this into Equation 12.34 gives

$$\rho A \frac{\partial^2 u}{\partial t^2}(x,t) = -\frac{\partial^2}{\partial x^2}\left(EI(x)\frac{\partial^2 u}{\partial x^2}(x,t)\right) + f(x,t).$$

Finally, if the cross section of the beam is uniform along its length, then we have

$$\rho A \frac{\partial^2 u}{\partial t^2}(x,t) = -EI\frac{\partial^4 u}{\partial x^4}(x,t) + f(x,t). \tag{12.37}$$

12.6.2 Solutions to the Beam Equation

Static deflection

Let us consider the case where

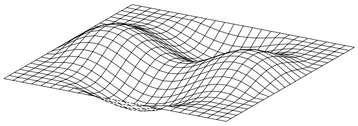
$$\frac{\partial^2 u}{\partial t^2}(x,t) = 0.$$

First, we will consider the case where the beam is cantilever and subjected to a static force at the end, as is illustrated in Figure 12.26. Since there is no acceleration and the solution does not depend upon time, the beam equation reduces to

$$EI\frac{d^4 u}{dx^4}(x) = 0. \tag{12.38}$$

There are four boundary conditions:

1. since the beam is fixed at zero, $u(0) = 0$;
2. since the beam is a cantilever beam, the slope zero is zero, $\frac{du}{dx}(0) = 0$;
3. since there is a point load at $x = L$, the shear force at $x = L$ must equal F, and using Equation 12.33 gives $EI\frac{d^3 u}{dx^2}(L) = F$; and,



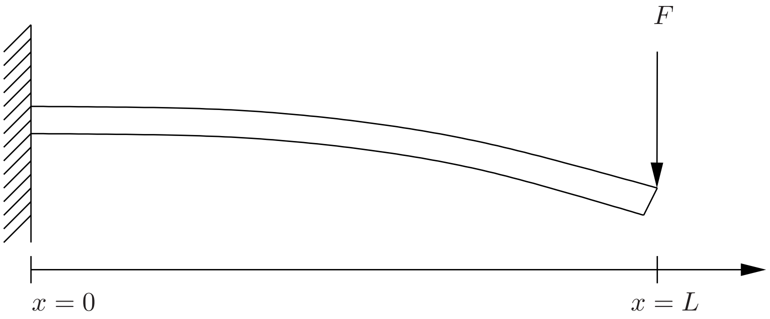


Figure 12.26. Cantilever beam subjected to a force at the end.

4. since there is no moment applied at $x = L$, $\frac{d^2u}{dx^2}(L) = 0$.

Clearly, the general solution to Equation 12.38 is a third order polynomial in x,

$$u(x) = c_1x^3 + c_2x^2 + c_3x + c_4.$$

Applying the boundary conditions gives

1. fixed at zero:
$$u(0) = 0 \qquad \Longrightarrow \qquad c_4 = 0;$$

2. cantilever:
$$\frac{du}{dx}(0) = 0 \qquad \Longrightarrow \qquad c_3 = 0;$$

3. shear at end:
$$6EIc_1 = F \qquad \Longrightarrow \qquad c_1 = \frac{F}{6EI}$$
and

4. no moment at end:
$$\frac{F}{6EI}6L + 2c_2 = 0 \qquad \Longrightarrow \qquad c_2 = -\frac{FL}{2EI}$$

Hence

$$u(x) = \frac{F}{6EI}x^3 - \frac{FL}{2EI}x^2$$

and at $x = L$, an applied force of F produces a displacement of

$$u(L) = \frac{FL^3}{6EI} - \frac{FL}{2EI}L^2 = -\frac{L^3}{3EI}F$$

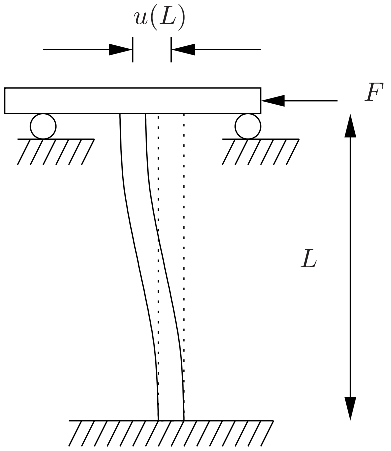


Figure 12.27. A deflecting column.

Since the displacement is proportional to the applied force and the proportionality constant is $\frac{L^3}{6EI}$, we can conclude that a cantilever spring will have a spring constant of

$$k = \frac{3EI}{L^3}.$$

In the case of a rectangular beam with width w and height h,

$$I = \int\int z^2 dz dy = \frac{1}{12} wh^3$$

so

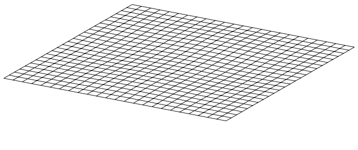
$$k = \frac{Ewh^3}{4L^3}.$$

Proving following two force deflection relationships is left as an exercise. The first is a cantilever beam which can not bend at either end, as is illustrated in Figure 12.27. For this system

$$k = \frac{12EI}{L^3},$$

or in the case of a rectangular cross section,

$$k = \frac{Ewh^3}{L^3},$$

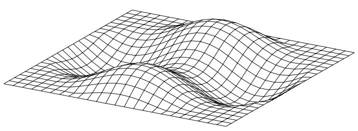


12.7 Sturm-Liouville Theory

Problem 12.1 Show that the the eigenvalue for the wave equation with boundary conditions

$$u(0,t) = u(L,t) = 0$$

must be positive.



Chapter 13

Numerical Methods

This chapter deals with numerical methods for determining approximate solutions for differential equations. It presents the derivation of the methods as well as analyses of the types of errors that are inherent in each method. Section 13.1 presents Euler's method with more mathematical rigor than was considered in Section 1.10. Section 13.2 presents a method based upon Taylor series, which, actually, is the basis for all the methods we consider. Section 13.3 presents the ubiquitous Runge-Kutta method. Section 13.4 considers the various types of errors inherent in each method. All the methods presented in Sections 13.1 through Section 13.3 work for a single, first order ordinary differential equation. Section 13.5 extends these methods to systems of coupled first order ordinary differential equations. Finally, Section 13.7 presents some basic techniques for determining approximate numerical solutions for partial differential equations.

13.1 Another Look at Euler's Method

In Section 1.10, Euler's method was derived as an approximation to the usual definition of the derivative. In particular, for a first order, ordinary differential equation of the form

$$\dot{x} = f(x,t) \tag{13.1}$$

the derivative with respect to time is approximated by

$$\dot{x}(t) \approx \frac{x(t+\Delta t) - x(t)}{\Delta t}$$

for $\Delta t \ll 1$. Consequently,

$$x(t+\Delta t) \approx x(t) + f(x(t),t)\Delta t \tag{13.2}$$

for $\Delta t \ll 1$.

In this section a slightly more sophisticated analysis will be undertaken that will allow for easy extensions to higher order methods and error analysis. In particular, the analysis will be based upon a Taylor series expansion of the form

$$x\left(t+\Delta t\right) = x(t) + \frac{dx(t)}{dt}\Delta t + \frac{1}{2!}\frac{d^2x(t)}{dt^2}\left(\Delta t\right)^2 + \frac{1}{3!}\frac{d^3x(t)}{dt^3}\left(\Delta t\right)^3 + \cdots . \quad (13.3)$$

Since the problem statement includes the fact that

$$\dot{x}(t) = f(x(t), t)$$

substituting this into equation 13.3 gives

$$x\left(t+\Delta t\right) = x(t) + f(x(t),t)\Delta t + \frac{1}{2!}\frac{df(x(t),t)}{dt}\left(\Delta t\right)^2 + \frac{1}{3!}\frac{d^2f(x(t),t)}{dt^2}\left(\Delta t\right)^3 + \cdots . \quad (13.4)$$

Clearly, Euler's method amounts to only using the first two terms in the series to approximate $x(t+\Delta t)$, and if $\Delta t \ll 1$, the *local truncation error* due to the fact that only a finite number of terms is used is proportional to $\left(\Delta t\right)^2$. In other words, if the time step is cut in half, the truncation error is reduced by $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$ and if Δt is reduced by an order of magnitude, the truncation error is reduced by $\left(\frac{1}{10}\right)^2 = \frac{1}{100}$.

The following example illustrates the method as well as the effect of the time step on the error.

Example 13.1.1 Use Euler's method to determine an approximate solution to

$$\begin{aligned} \dot{x} &= 5x \\ x(0) &= 1 \end{aligned}$$

for $0 < t \leq 2$.

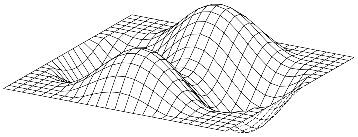
Note that the exact solution is easy to compute and is $x(t) = e^{5t}$. At a given time, the equation to compute the value of the solution at the next time step is given by

$$\begin{aligned} x(t+\Delta t) &= f(x(t),t)\Delta t \\ &= 5x(t)\Delta t. \end{aligned}$$

A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. A plot of the solutions for two time steps as well as the exact solution is illustrated in Figure 13.1.

The first few steps of the results of the computations for the case where $\Delta t = 0.1$ are

t	$x(t)$	e^{5t}
0.000000	1.000000	1.000000
0.100000	1.500000	1.648721
0.200000	2.250000	2.718282



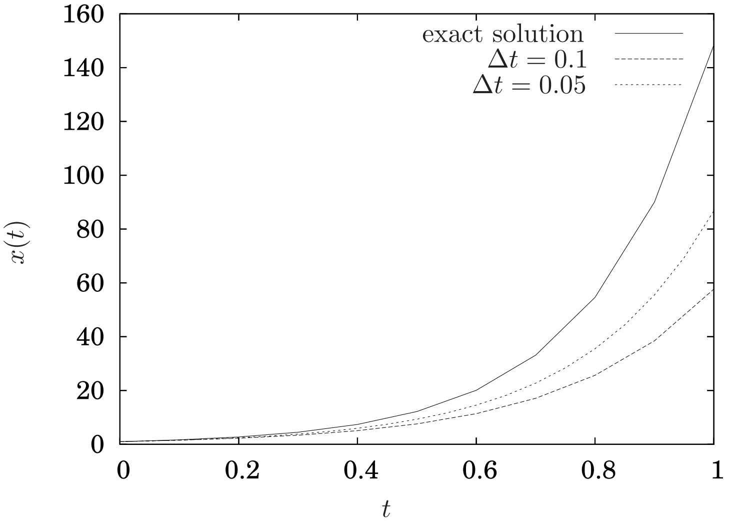


Figure 13.1. Solution to example 13.1.1 illustrating the fact that for Euler's method the accumulated global error is proportional to Δt.

and first few steps of the results of the computations for the case where $\Delta t = 0.05$ are

t	$x(t)$	e^{5t}
0.000000	1.000000	1.000000
0.050000	1.250000	1.284025
0.100000	1.562500	1.648721.

After the first step when $\Delta t = 0.1$, the error in the approximate solution is $1.648721 - 1.5000000 = 0.14872$. When $\Delta t = 0.05$ the error is $1.284025 - 1.250000 = 0.034025$. The critical observation is that when the time step was cut by a factor of two, after the first step of the algorithm the error was decreased by approximately a factor of four, illustrating the fact that the error in Euler's method is proportional to $(\Delta t)^2$.

However, referring to Figure 13.1 it appears that the *overall error*, or *global error* is not decreased by a factor of four, but rather simply cut in half, *i.e.,* it appears that the overall error is proportional to Δt. Upon a little reflection, the reason for this is obvious. The error at each time step may be decreased in proportion to $(\Delta t)^2$ but the number of time steps necessary to cover a specified time interval is inversely proportional to Δt. Specifically in this example, if Δt is reduced by a factor of two, the number of time steps necessary to go from $t = 0$ to $t = 1$ is doubled. Because of this, even though the error introduced at each time step is proportional to

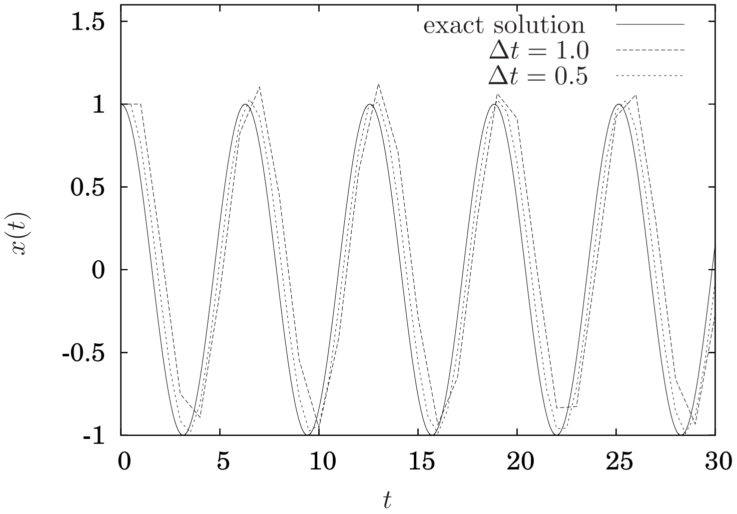


Figure 13.2. Solution to example 13.1.2 illustrating the fact that for Euler's method the overall accumulated error is proportional to Δt.

$(\Delta t)^2$, if the number of steps needed is proportional to $\frac{1}{\Delta t}$, the overall error will be proportional to Δt. ■

Another example will help flesh out the relationship between the changes in step size and the resultant error.

Example 13.1.2 Determine an approximate solution to

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

using Euler's method. Not too much thought (or even less work) gives the exact solution as $x(t) = -\cos t$. A plot of the approximate solutions for $\Delta t = 1.0$ and $\Delta t = 0.5$ as well as the exact solution are illustrated in Figure 13.2.

Note that, as was the case in the previous example, decreasing the time step by a factor of two generally decreases the overall error by a factor of two as well. In other words, the overall error is proportional to the time step. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

It appears that the above reasoning is correct. After a single time step, starting when the approximate and exact solutions are identical, Euler's method will produce an error proportional to $(\Delta t)^2$. However, because the number of time steps necessary to for the algorithm to complete a particular time interval is inversely proportional to the time step, the overall error, *i.e.,* the error that can be discerned by general observation, will typically be proportional to Δt.

13.2 Taylor Series Methods

If it is necessary to increase the accuracy of the approximate solution without the computational burden of an excessively small step size, the relatively obvious thing to do is to starting including higher order terms from the Taylor series expansion for $x(t+\Delta t)$ in equation 13.4. Upon initially considering this notion, it may appear to be a rather trivial exercise. While it is manageable to include the $(\Delta t)^2$ term, and even possibly the $(\Delta t)^3$ term, a quick review of multivariable calculus will illustrate that the complexity of such an endeavor quickly becomes rather burdensome. The reason for this is simply because the function f depends upon both x and t, but x also depends upon t, but determining exactly that dependence of x on t is the whole point of the problem, *i.e.,* determining $x(t)$.

Hence, dropping the explicit dependence of x on t

$$\frac{df(x,t)}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial t} = \frac{\partial f}{\partial x} f + \frac{\partial f}{\partial t}.$$

or, including the dependence

$$\begin{aligned} \frac{df(x(t),t)}{dt} &= \left. \left(\frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial t} \right) \right|_{(x(t),t)} \\ &= \left. \left(\frac{\partial f}{\partial x} f + \frac{\partial f}{\partial t} \right) \right|_{(x(t),t)}, \end{aligned}$$

where the notation

$$\left. \frac{\partial f}{\partial x} \right|_{(x(t),t)}$$

means, as usual, to compute the partial derivative of f with respect to x and then evaluate it at the values of $x(t)$ and t.

Again, the crux of the matter is that the problem statement specifies how $\dot{x}$ depends on x and t, but not how x depends on t. Thus, one cannot simply compute derivative of $f(x(t),t)$ with respect to t since $x(t)$ is not known; rather, one must compute resort to the chain rule as expressed in the equations above.

13.2.1 Second order Taylor series expansion

Returning to the Taylor series expansion and including all the terms through $(\Delta t)^2$ gives

$$\begin{aligned} x\left(t+\Delta t\right) &= x(t) + f(x(t),t)\Delta t + \frac{1}{2}\left.\frac{df}{dt}\right|_{(x(t),t)} (\Delta t)^2 + \cdots \\ &= x(t) + f(x(t),t)\Delta t + \frac{1}{2}\left.\left(\frac{\partial f}{\partial x} f + \frac{\partial f}{\partial t}\right)\right|_{(x(t),t)} (\Delta t)^2 + \cdots \end{aligned} \tag{13.5}$$

Hence, keeping all terms through $(\Delta t)^2$, which should produce a step truncation error proportional to $(\Delta t)^3$ and an overall error proportional to $(\Delta t)^2$ is given by

$$x\left(t+\Delta t\right) = x(t) + f(x(t),t)\Delta t + \frac{1}{2}\left.\left(\frac{\partial f}{\partial x} f + \frac{\partial f}{\partial t}\right)\right|_{(x(t),t)} (\Delta t)^2 .$$

Returning to example 13.1.1 illustrates the fact that the error is indeed as would be expected.

Example 13.2.1 Use a second order Taylor series expansion to determine an approximate solution to

$$\begin{aligned} \dot{x} &= 5x \\ x(0) &= 1 \end{aligned}$$

for $0 < t \leq 2$.

Since $f(x(t),t) = 5x(t)$,

$$\begin{aligned} \frac{\partial f}{\partial x} &= 5 \\ \frac{\partial f}{\partial t} &= 0. \end{aligned}$$

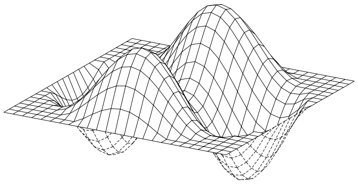
Hence,

$$x(t+\Delta t) = x(t) + 5x(t)\Delta t + \frac{25}{2}x(t)\left(\Delta t\right)^2 . \tag{13.6}$$

A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4.

The first few steps of the results of the computations for the case where $\Delta t = 0.1$ are

t	$x(t)$	e^{5t}
0.000000	1.000000	1.000000
0.100000	1.625000	1.648721
0.200000	2.640625	2.718282



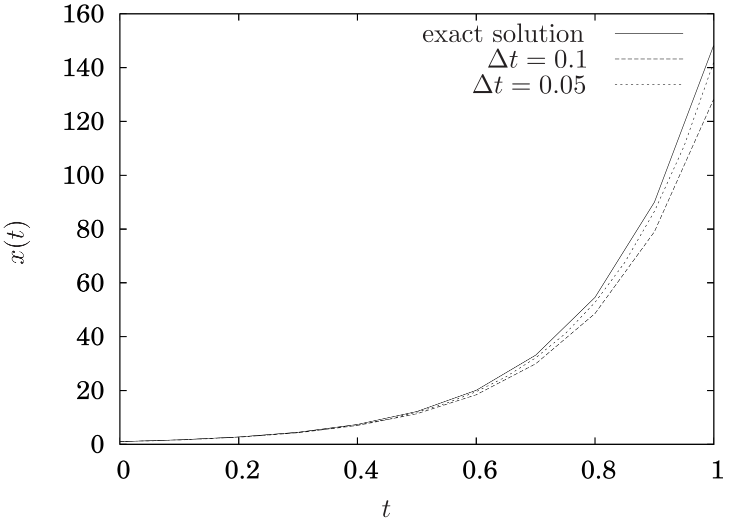


Figure 13.3. Solution to example 13.2.1. Note that for the second order Taylor series, the overall accumulated error is proportional to $(\Delta t)^2$.

and first few steps of the results of the computations for the case where $\Delta t = 0.05$ are

t	$x(t)$	e^{5t}
0.000000	1.000000	1.000000
0.050000	1.281250	1.284025
0.100000	1.641602	1.648721

A the two approximate solutions and the exact solution are illustrated in Figure 13.3.

Observe that after the first time step, the error for $\Delta t = 0.1$ is $1.648721 - 1.625000 = 0.023721$, and the error for $\Delta t = 0.05$ is $1.284025 - 1.281250 = 0.0027750$. Since $\frac{0.023721}{0.0027750} = 8.5 \approx 8$ it is clear that the error is proportional to $(\Delta t)^3$ since the step size was reduced by a factor of two and the error was reduced by a factor of eight.

With respect to the overall error, referring to Figure 13.3, it is clear that the overall error is proportional to $(\Delta t)^2$ since the $\Delta t = 0.05$ curve has approximately $\frac{1}{4}$ the error of the $\Delta t = 0.1$ curve. Observe also that for the case of $\Delta t = 0.1$ in Figures 13.1 and 13.3, the overall error decreases by an order of magnitude, which is consistent with the second order Taylor series in the latter case including the $(\Delta t)^2$ in the expansion. ■

Because the form of the partial derivatives in equation 13.5, an example with the function $f(x,t)$ that includes both x and t may be helpful.

Example 13.2.2 Determine an approximate solution to

$$\begin{aligned} \dot{x} &= -x^3 + \sin(tx) \\ x(0) &= 1 \end{aligned}$$

using a second order Taylor series expansion.

In this problem

$$f(x,t) = -x^3 + \sin(tx)$$

Hence,

$$\begin{aligned} \frac{\partial f}{\partial x} &= -3x^2 + t\cos(tx) \\ \frac{\partial f}{\partial t} &= x\cos(tx). \end{aligned}$$

Thus, the equation for $x(t+\Delta t)$ using a second order Taylor series expansion is

$$\begin{aligned} x(t+\Delta t) &= x(t) + f(x(t),t)\Delta t + \\ & \quad \frac{1}{2}\left.\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right)\right|_{(x(t),t)} (\Delta t)^2 \\ &= x(t) + \left(-x^3(t) + \sin(tx(t))\right)\Delta t + \\ & \quad \frac{1}{2}\left\{\left[-3x^2(t) + t\cos(tx(t))\right]\left[-x^3(t) + \sin(tx(t))\right] + \right. \\ & \quad \left. x(t)\cos(tx(t))\right\}(\Delta t)^2 . \end{aligned}$$

The solution is illustrated (along with another solution generated by another method) in Figure 13.6 for the cases where $\Delta t = 0.4$ and $\Delta t = 0.2$. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

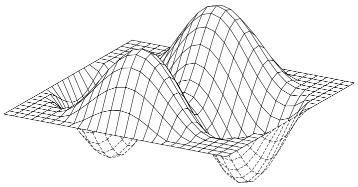
13.2.2 Third order Taylor series expansion

The obvious thing to do at this point to improve the accuracy of the method is to try to include the third order terms in the expansion. So, let's go for it. Starting with equation 13.4

$$x(t+\Delta t) = x(t) + f(x(t),t)\Delta t + \frac{1}{2!}\frac{df(x(t),t)}{dt}(\Delta t)^2 + \frac{1}{3!}\frac{d^2 f(x(t),t)}{dt^2}(\Delta t)^3 + \cdots . \tag{13.7}$$

As has already been stated, the dependence of f on x and t is specified, but the dependence of x on t. Hence, as in the case of the second order Taylor series, the chain rule must be used to expand the derivatives in terms of known quantities. In particular, as above

$$\frac{df}{dt} = \frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}. \tag{13.8}$$



So, to start to compute the next higher order term,

$$\begin{aligned}\frac{d^2 f}{dt^2} &= \frac{d}{dt}\frac{df}{dt}\\ &= \frac{d}{dt}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right)\\ &= \frac{d}{dt}\left(\frac{\partial f}{\partial x}\right)f + \frac{\partial f}{\partial x}\frac{d}{dt}(f) + \frac{d}{dt}\left(\frac{\partial f}{\partial t}\right), \qquad (13.9)\end{aligned}$$

where the last line is simply using the product rule for differentiating the $\frac{\partial f}{\partial x}f$ term. Recall that the need for the expansion in equation 13.8 was the fact that f depended on both x and t, but x also depended on t, but only the derivative of x with respect to t is known. Similarly, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial t}$ can depend on both x and t as well, so must be expanded similarly. Hence,

$$\begin{aligned}\frac{d}{dt}\left(\frac{\partial f}{\partial x}\right) &= \frac{\partial^2 f}{\partial x^2}f + \frac{\partial^2 f}{\partial x \partial t}\\ \frac{d}{dt}\left(\frac{\partial f}{\partial t}\right) &= \frac{\partial^2 f}{\partial x \partial t}f + \frac{\partial^2 f}{\partial t^2}.\end{aligned}$$

Using these two expressions as well as the one for $\frac{df}{dt}$ in equation 13.8 in equation 13.9 gives

$$\frac{d^2 f}{dt^2} = \left(\frac{\partial^2 f}{\partial x^2}f + \frac{\partial^2 f}{\partial x \partial t}\right)f + \frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right) + \frac{\partial^2 f}{\partial x \partial t}f + \frac{\partial^2 f}{\partial t^2}. \qquad (13.10)$$

Finally, substituting the terms from equations 13.10 and 13.8 gives

$$\begin{aligned}x(t+\Delta t) &= x(t) + f(x(t),t)\Delta t + \frac{1}{2}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right)\Bigg|_{(x(t),t)}(\Delta t)^2 +\\ &\quad \frac{1}{6}\left[\left(\frac{\partial^2 f}{\partial x^2}f + \frac{\partial^2 f}{\partial x \partial t}\right)f + \frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right) + \right.\\ &\quad \left.\frac{\partial^2 f}{\partial x \partial t}f + \frac{\partial^2 f}{\partial t^2}\right]\Bigg|_{(x(t),t)}(\Delta t)^3 \qquad (13.11)\end{aligned}$$

Remark 13.2.3

1. While it is theoretically possible to use equation 13.11, as a practical matter it would be quite arduous to correctly compute all the partial derivatives, products, *etc.*

2. If even greater accuracy is needed, including the fourth order terms in Δt will result in an absolutely huge expansion since every term in equation 13.11 depends on f, which will result in two partial derivative terms when expanded, as will all the terms that are products of two terms, which is every term except one. Clearly, an approach that gives higher order accuracy without the hassle of such computations would be useful. Hence, the next section. ⋄

Example 13.2.4 Use the first, second and third order Taylor series methods to determine an approximate numerical solution to

$$\begin{aligned}\dot{x} &= 10x(1-x)\\ x(-1) &= \frac{1}{1+e^{10}}\end{aligned}$$

and compare it to the exact solution, which is

$$x(t) = \frac{1}{1+e^{-10t}}.$$

For this problem,

$$\dot{x} = f(x,t) = 10x(1-x)$$

so for the second and third order methods we need to compute

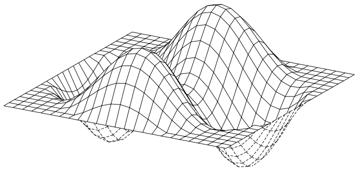
$$\begin{aligned}\frac{\partial f}{\partial x} &= 10-20x\\ \frac{\partial f}{\partial t} &= 0\\ \frac{\partial^2 f}{\partial x^2} &= -20\\ \frac{\partial^2 f}{\partial t^2} &= 0\\ \frac{\partial^2 f}{\partial t\partial x} &= 0.\end{aligned}$$

Thus, the equation for the first order method (or Euler's method) is

$$\begin{aligned}x(t+\Delta t) &= x(t)+f(x,t)\Delta t\\ &= x(t)+10x(1-x)\Delta t.\end{aligned}$$

The equation for the second order method is

$$\begin{aligned}x(t+\Delta t) &= x(t)+f(x,t)\Delta t+\frac{1}{2}\frac{df}{dt}(\Delta t)^2\\ &= x(t)+f(x,t)\Delta t+\frac{1}{2}\left[\frac{\partial f}{\partial x}f+\frac{\partial f}{\partial t}\right](\Delta t)^2\\ &= x(t)+(10x\,(1-x))\,\Delta t+\\ &\quad \frac{1}{2}\left[(10-20x)\,(10x\,(1-x))\right](\Delta t)^2.\end{aligned}$$



The equation for the third order method is

$$\begin{aligned}
x(t+\Delta t) &= x(t) + f(x,t)\Delta t + \frac{1}{2}\frac{df}{dt}(\Delta t)^2 + \frac{1}{6}\frac{d^2 f}{dt^2}(\Delta t)^3 \\
&= x(t) + f(x,t)\Delta t + \frac{1}{2}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right)(\Delta t)^2 + \\
&\quad \frac{1}{6}\left[\left(\frac{\partial^2 f}{\partial x^2}f + \frac{\partial^2}{\partial x \partial t}\right) + \frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right) + \right. \\
&\quad \left. \frac{\partial^2 f}{\partial x \partial t}f + \frac{\partial^2 f}{\partial t^2}\right](\Delta t)^3 \\
&= x(t) + (10x(1-x))\Delta t + \\
&\quad \frac{1}{2}\left[(10-20x)(10x(1-x))\right](\Delta t)^2 + \\
&\quad \frac{1}{6}\left[-20(10x(1-x)) + (10-20x)^2(10x(1-x))\right](\Delta t)^3 .
\end{aligned}$$

Clearly, as the order of the method increases, so does the complexity of the expression for $x(t+\Delta t)$. ■

13.3 The Runge-Kutta Method

The main idea behind the so-called Runge-Kutta methods is, instead of evaluating all the partial derivatives necessary in a straight-forward Taylor series computations, to approximate the derivatives to the same order of accuracy using combinations of the function $f(x,t)$ evaluated not only at $x(t)$ and t, but other x and t values as well.

Consider the function, $x(t)$ illustrated in Figure 13.4. The curve represents the unknown function $x(t)$. Assume that $x(t)$ is known exactly at two points, say at $t = 1.5$ and $t = 2.0$. Anyone with a background in first year calculus knows that the derivative of x with respect to t at $t = 1.5$ can be approximated by

$$\dot{x} \approx \frac{x(t+\Delta t) - x(t)}{\Delta t} = \frac{x(2.0) - x(1.5)}{0.5}.$$

In the figure, it is clear that the derivative of $x(t)$ at $t = 1.5$, which is the slope of the tangent line at that point, is approximately the same as the slope of the line connecting the values of $x(1.5)$ and $x(2.0)$ at times $t = 1.5$ and $t = 2.0$. Similarly, for that matter, the slope at $t = 2.0$ is approximately the same as well, as is the slope at any point between $t = 1.5$ and $t = 2.0$. Furthermore, the smaller the difference between the two points in time is, the better the approximation will be.

Now, consider the task of computing an approximation to the second derivative of x with respect to t. Since the second derivative is the derivative of the derivative, it will be necessary to have an approximate computation for the derivative at two values for t. Hence, assume that the exact values for $x(t)$ are

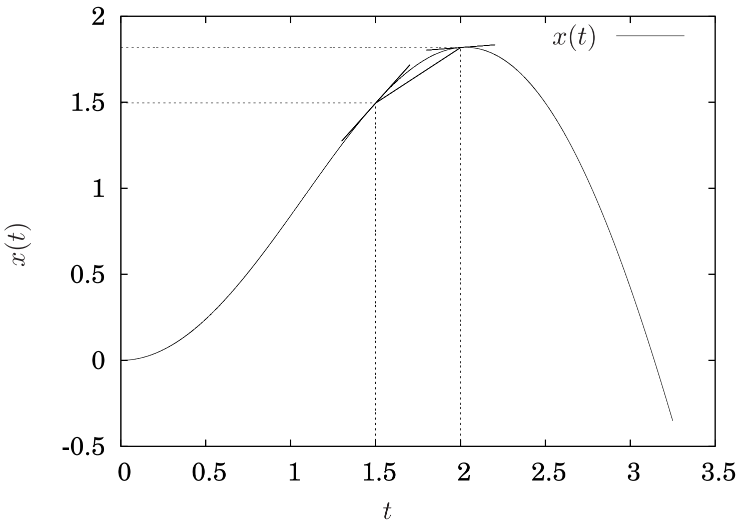


Figure 13.4. Approximating derivatives of a function $x(t)$ by computing the slope of a line connecting two points.

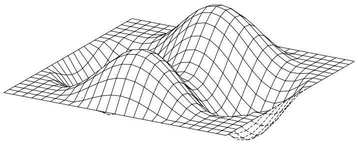
known for three points, say $t = 1.5$, $t = 2.0$ and $t = 2.5$, as is illustrated in Figure 13.5. The second derivative, then, is approximated by

$$\begin{aligned}\ddot{x} &\approx \frac{\dot{x}(t+\Delta t)-\dot{x}(t)}{\Delta t}\\ &\approx \frac{\dot{x}(2.0)-\dot{x}(1.5)}{0.5}\\ &\approx \frac{\left(\frac{x(2.5)-x(2.0)}{0.5}\right)-\left(\frac{x(2.0)-x(1.5)}{0.5}\right)}{0.5}\\ &= \frac{x(2.5)-2x(1.5)+x(1.5)}{(0.5)^2}\end{aligned}$$

where the second to last equation was obtained simply by substituting the equation for the approximate value of the derivative for each of $t = 1.5$ and $t = 2.0$. The main point is that the computation for an approximation for the second derivative of x with respect to t required that *three* points of $x(t)$ be known.

So, to summarize, in order to approximate the derivative of x we needed to evaluate $x(t)$ at two points in time. In order to approximate the second derivative, we needed to compute $x(t)$ at three points in time. Clearly, to compute an approximate for the nth derivative, we will need to evaluate $x(t)$ at $n+1$ points in time.

The main approach of the Runge-Kutta methods in this section is, in order to avoid all the complications associated with expanding the derivatives of $f(x(t), t)$



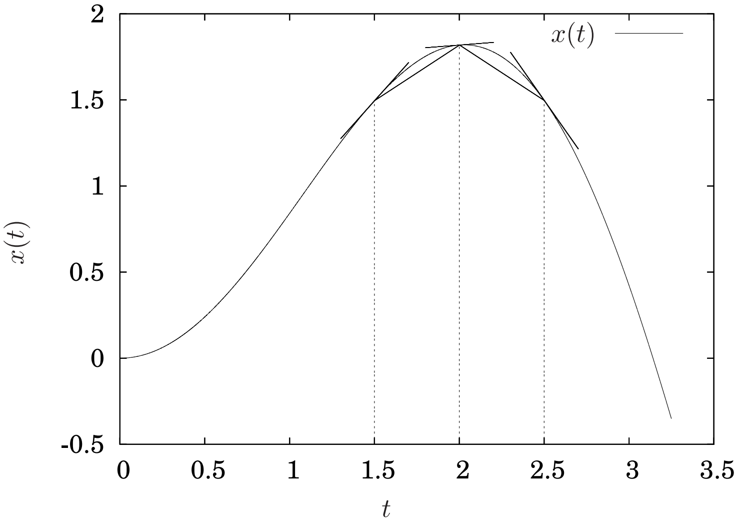


Figure 13.5. Approximating derivatives of a function $x(t)$ by computing the slope of a line connecting two points.

in a Taylor series, the higher order derivatives will be approximated by simply evaluating $f(x(t),t)$ at different x and t values to approximate the higher order terms in the Taylor series.

So, in the case of attempting to compute approximate solutions to

$$\dot{x} = f(x(t),t)$$

there is a slight twist, which is that the first derivative of x is already given by the problem; namely, $f(x(t),t)$. So, the picture gets a little more abstract because the approximate derivatives that we will be computing will not be for $x(t)$, but rather for $f(x(t),t)$, *i.e.,* we will approximate the terms in equation 13.4 instead of equation 13.3. The one final conceptual complication is that the whole point of the problem is to determine $x(t)$; hence, these approximations for derivatives are not simple to compute because the $x(t)$ to plug into $f(x(t),t)$ is not known.

The approach will ultimately be to approximate the $x(t+\Delta t)$ value that is used to evaluate the $f(x(t+\Delta t),t+\Delta t)$ values, that will be used to determine approximations to the derivatives of $f(x(t),t)$ that appear in the Taylor series expansion of $x(t+\Delta t)$ in order to compute an approximation for $x(t+\Delta t)$.

13.3.1 The first order Runge-Kutta method

Approximating

$$x\left(t+\Delta t\right) = x(t)+f(x(t),t)\Delta t+\frac{1}{2!}\frac{df(x(t),t)}{dt}\left(\Delta t\right)^2+\frac{1}{3!}\frac{d^2f(x(t),t)}{dt^2}\left(\Delta t\right)^3+\cdots.$$

through the Δt term requires no derivative computations for $f(x(t),t)$. Hence, it is just Euler's method,

$$x(t+\Delta t) \approx x(t) + f(x(t),t)\Delta t.$$

13.3.2 The second order Runge-Kutta method

The goal is to compute

$$\begin{aligned} x\,(t+\Delta t) &= x(t) + f(x(t),t)\Delta t + \frac{1}{2!}\frac{df(x(t),t)}{dt}\,(\Delta t)^2 + \cdots \\ &= x(t) + f(x(t),t)\Delta t + \frac{1}{2}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right)\bigg|_{(x(t),t)}(\Delta t)^2 + \cdots \end{aligned} \tag{13.12}$$

through the $(\Delta t)^2$ term without computing the derivatives of $f(x(t),t)$, but rather by evaluating $f(x(t),t)$ at different values of $x(t)$ and t to approximate those derivatives. With that in mind, consider the task of determining the values of $c_1,\ldots,c_4$ in the following

$$x(t+\Delta t) = x(t) + c_1 f(x(t),t)\Delta t + c_2 f\left(x(t) + c_3 f(x(t),t)\Delta t, t + c_4\Delta t\right)\Delta t \tag{13.13}$$

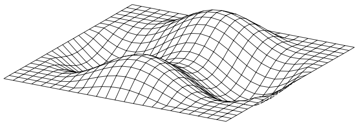
that will make it exactly equal to equation 13.12 up to the $(\Delta t)^2$ term. Careful scrutiny of the second f term will show that this is the term where $f(x(t),t)$ is evaluated at different values for $x(t)$ and t; namely, $x(t)+c_3 f(x(t),t)\Delta t$ for the x-value and $t+c_4\Delta t$ for the t-value.

Although it is understandable that by this point the reader may be inclined to quit Taylor series for life, the way to determine the c_3 and c_4 constants is, obviously, to expand $f(x(t)+c_2 f(x(t),t)\Delta t, t+c_4\Delta t)$ in a Taylor series. In particular,

$$\begin{aligned} &f(x(t)+c_3 f(x(t),t)\Delta t, t+c_4\Delta t) = \\ &\qquad f(x(t),t) + \frac{\partial f}{\partial x}\bigg|_{(x(t),t)}(c_3 f(x(t),t)\Delta t) + \frac{\partial f}{\partial t}\bigg|_{(x(t),t)} c_4\Delta t + \cdots . \end{aligned}$$

Substituting this into equation 13.13 gives

$$\begin{aligned} x(t+\Delta t) &= x(t) + c_1 f(x(t),t)\Delta t + \\ &\quad c_2\left(f(x(t),t) + \frac{\partial f}{\partial x}\bigg|_{(x(t),t)}(c_3 f(x(t),t)\Delta t) + \frac{\partial f}{\partial t}\bigg|_{(x(t),t)} c_4\Delta t + \cdots\right) \\ &= x(t) + (c_1+c_2)\,f(x(t),t)\Delta t + \\ &\quad \left(c_2c_3\left(\frac{\partial f}{\partial x}f\right)\bigg|_{(x(t),t)} + c_2c_4\,\frac{\partial f}{\partial t}\bigg|_{(x(t),t)}\right)(\Delta t)^2 + \cdots . \end{aligned} \tag{13.14}$$



Equating coefficients in equations 13.12 and 13.14 gives

$$\begin{aligned} c_1 + c_2 &= 1 \\ c_2 c_3 &= \frac{1}{2} \\ c_2 c_4 &= \frac{1}{2}. \end{aligned}$$

Clearly, there are multiple solutions, but

$$\begin{aligned} c_1 &= \frac{1}{2} \\ c_2 &= \frac{1}{2} \\ c_3 &= 1 \\ c_4 &= 1 \end{aligned}$$

is perhaps the most commonly used. Hence, substituting these values into equation 13.13 gives

$$x\left(t+\Delta t\right) \approx x(t) + \frac{1}{2}\left(f\left(x(t),t\right) + f\left(x(t) + f\left(x(t),t\right)\Delta t, t+\Delta t\right)\right)\Delta t, \quad (13.15)$$

which is known as either the *improved Euler* formula or *second order Runge-Kutta* formula.

The following example illustrates the fact that this approach gives the same approximate solution as the second order Taylor series method, but without the need to compute the derivatives of the function $f(x(t),t)$.

Example 13.3.1 Determine an approximate solution to

$$\begin{aligned} \dot{x} &= 5x \\ x(0) &= 1 \end{aligned}$$

for $0 < t \leq 2$ using the second order Runge-Kutta method.

Since $f(x(t),t) = 5x$, then

$$f(x(t) + f(x(t),t)\Delta t, t+\Delta t) = 5\left(x(t) + f(x(t),t)\Delta t\right).$$

Hence, the second order Runge-Kutta formula is

$$\begin{aligned} x(t+\Delta t) &= x(t) + \frac{\Delta t}{2}\left[5\left(x(t)\right) + 5\left(x(t) + f(x(t),t)\Delta t\right)\right] \\ &= x(t) + \frac{\Delta t}{2}\left[5\left(x(t)\right) + 5\left(x(t) + 5x(t)\Delta t\right)\right]. \\ &= x(t) + 5x(t)\Delta t + \frac{25}{2}x(t)\left(\Delta t\right)^2, \end{aligned}$$

which happens to be identical to equation 13.6. Hence, any computer program that computes an approximate solution will be the same as for example 13.2.1. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

Because of the rather complicated form of $f(x(t) + f(x(t), t)\Delta t, t + \Delta t)$, a slightly more complicated example is in order.

Example 13.3.2 Determine an approximate solution to

$$\begin{aligned} \dot{x} &= -x^3 + \sin(tx) \\ x(0) &= 1 \end{aligned}$$

using the second order Runge-Kutta formula (the improved Euler formula). This is the initial value problem from example 13.2.2. Since

$$f(x, t) = -x^3 + \sin(tx)$$

substituting $x(t) + f(x(t), t)\Delta t$ for x and $t + \Delta t$ for t gives

$$\begin{aligned} f(x(t) + f(x(t), t)\Delta t, t + \Delta t) &= -\left(x(t) + f(x(t), t)\Delta t\right)^3 + \\ &\quad \sin\left(\left(t + \Delta t\right)\left(x(t) + f(x(t), t)\Delta t\right)\right) \end{aligned}$$

(verify this yourself — it is *critical!*). Hence, substituting for $f(x(t), t)$ and $f(x(t) + f(x(t), t)\Delta t, x + \Delta t)$ into equation 13.15 gives

$$\begin{aligned} x(t + \Delta t) &\approx x(t) + \frac{\Delta t}{2}\left[\left(x^3(t) + \sin(tx(t))\right) + \right. \\ &\left.\Big(\left(x(t) + f(x(t), t)\Delta t\right)^3 + \sin\left(\left(t + \Delta t\right)\left(x(t) + f(x(t), t)\Delta t\right)\right)\Big)\right] \end{aligned}$$

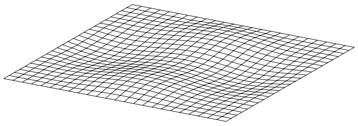
Figure 13.6 illustrates the approximate solution for the cases where $\Delta t = 0.2$ and $\Delta t = 0.4$. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

Comparison of second order Runge-Kutta and Taylor series methods

As is clear from the formulae in examples 13.2.2 and 13.3.2, the second order Taylor series method and the second order Runge-Kutta method do not result in exactly the same approximate solution. However, both methods are accurate to the same order. Figure 13.6 illustrates an accurate solution (generated with a very small time step) and solutions from the two second order approximate methods for two different time steps. Clearly, the two approximate solutions are not identical; however, they both demonstrated second order accuracy.

Interpretation of the second order Runge-Kutta formula

While the next two subsequent sections will present the results of exactly this same approach carried out to third and fourth order, respectively, this approach yields a rather easy interpretation beyond the fact that it is the result of the above mathematical manipulations.



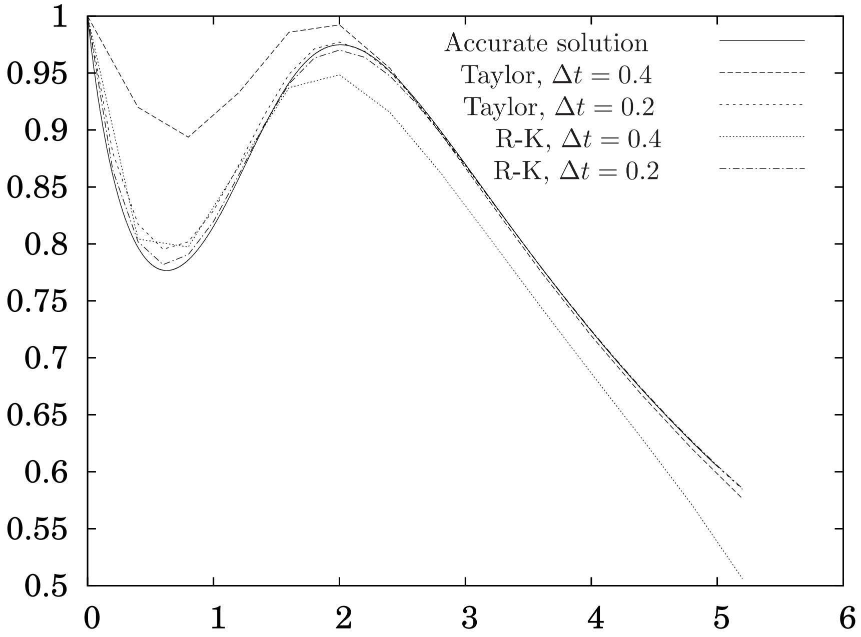


Figure 13.6. A comparison of the solutions from examples 13.2.2 and 13.3.2. The second order Taylor series and second order Runge-Kutta do not give the same approximate solutions; however, both methods have the same order of accuracy.

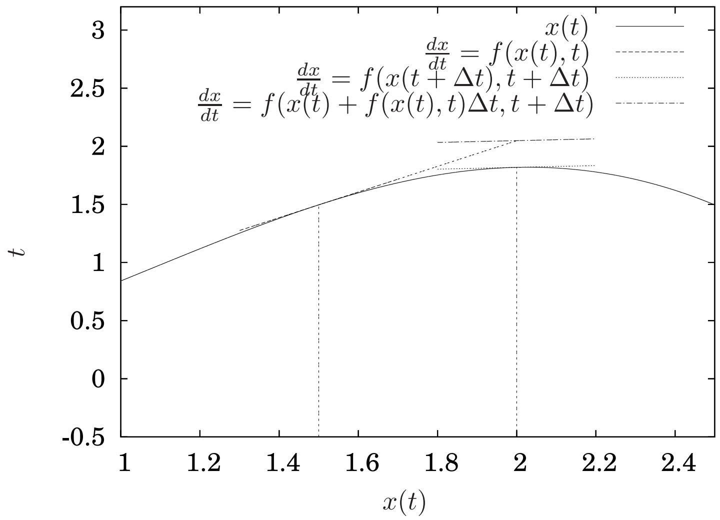


Figure 13.7. Interpretation of the improved Euler method. It uses the average of the slopes at the values at the beginning of the time step and at the end of the time step, but with an $x(t + \Delta t)$ value computed using a first order approximation.

One way to think of the second order Runge-Kutta formula is that it is simply Euler's method using the average of the slopes of $x(t)$ at the two endpoints of the time interval, *i.e.,* the average of the slope at $x(t)$ and the slope at $x(t+\Delta t)$. Mathematically, this formula would be

$$x(t + \Delta t) = x(t) + \frac{1}{2}\left[f(x(t), t) + f(x(t + \Delta t), t + \Delta t)\right]\Delta t.$$

However, the term $x(t+\Delta t)$ appears on both sides of the equation and is exactly the term that is unknown. Also, unless the function $f(x(t), t)$ is of a very special form, it will generally be impossible to solve this equation for $x(t + \Delta t)$. The idea is to replace the $x(t+\Delta t)$ term that is on the right hand side of the equation with an approximation for it; particularly, simply using Euler's formula for it on the right hand side. Hence,

$$x(t + \Delta t) = x(t) + \frac{1}{2}\left[f(x(t), t) + f(x(t) + f(x(t), t)\Delta t), t + \Delta t)\right]\Delta t.$$

Initially, this approach may intuitively be no better than Euler's method since Euler's method was used on the right hand side of the equation. However, since it was used in a term that is already multiplied by Δt, the overall order of that term will be $(\Delta t)^2$ and hence an order better in accuracy.

This is conceptually illustrated in Figure 13.7 which illustrates the same function, $x(t)$ that was illustrated in Figures 13.4 and 13.5. Figure 13.7 illustrates the same function, but plotted over a much shorter time interval. In this figure, $t = 1.5$, $\Delta t = 0.5$ and $t + \Delta t = 2.0$. The slope of $x(t)$ is known and is $f(x(t), t)$. The value of $x(t + \Delta t)$ is not known, and hence $f(x(t + \Delta t), t + \Delta t)$ cannot be directly computed. However, if Δt is small, then $x(t+\Delta t) \approx x(t) + f(x(t), t)\Delta t$ and also then

$$f(x(t + \Delta t), t + \Delta t) \approx f(x(t) + f(x(t), t)\Delta t, t + \Delta t),$$

which is illustrated graphically in Figure 13.7.

13.3.3 The third order Runge-Kutta method

The third order Runge-Kutta method (as well as the fourth order method in the following subsection) is derived in exactly the same manner as the second order Runge-Kutta method, except to third and fourth orders, respectively. Hence the goal is to compute equation 13.11 through the $(\Delta t)^3$ term without explicitly computing the derivatives of $f(x(t), t)$ but rather approximating those derivatives to third order by evaluating $f(x(t), t)$ at different x and t values. In particular, equating

$$\begin{aligned} x(t+\Delta t) \quad = \quad & x(t) + f(x(t), t)\Delta t + \frac{1}{2}\left(\frac{\partial f}{\partial x} f + \frac{\partial f}{\partial t}\right)\Bigg|_{(x(t),t)} (\Delta t)^2 + \\ & \frac{1}{6}\left[\left(\frac{\partial^2 f}{\partial x^2} f + \frac{\partial^2 f}{\partial x \partial t}\right) f + \frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial x} f + \frac{\partial f}{\partial t}\right) + \right. \\ & \quad \left. \frac{\partial^2 f}{\partial x \partial t} f + \frac{\partial^2 f}{\partial t^2}\right]\Bigg|_{(x(t),t)} (\Delta t)^3 + \cdots \end{aligned}$$

and

$$\begin{aligned} x(t+\Delta t) \quad = \quad & x(t) + [c_1 f + \\ & c_2 f\left(x + c_3 f \Delta t, t + c_4 \Delta t\right) + \\ & c_5 f(x + c_6 f \Delta t + c_7 f\left(x + c_8 f \Delta t, t + c_9 \Delta t\right)\Delta t, t + c_{10}\Delta t)]\,\Delta t \end{aligned} \tag{13.16}$$

(if no arguments to f are specified, it is evaluated at $(x(t), t)$) to third order gives

$$\begin{aligned}
c_1 &= \frac{1}{6} \\
c_2 &= \frac{2}{3} \\
c_3 &= \frac{1}{2} \\
c_4 &= \frac{1}{2} \\
c_5 &= \frac{1}{6} \\
c_6 &= -1 \\
c_7 &= 2 \\
c_8 &= \frac{1}{2} \\
c_9 &= \frac{1}{2} \\
c_{10} &= 1. \qquad (13.17)
\end{aligned}$$

A detailed derivation appears in Appendix C.2.1.

A more standard expression of this solution is

$$x(t + \Delta t) = x(t) + \frac{1}{6}\left(v_1 + 4v_2 + v_3\right) \qquad (13.18)$$

where

$$\begin{aligned}
v_1 &= f\left(x(t), t\right)\Delta t \qquad (13.19) \\
v_2 &= f\left(x(t) + \frac{1}{2}v_1, t + \frac{1}{2}\Delta t\right)\Delta t \\
v_3 &= f\left(x(t) + 2v_2 - v_1, t + \Delta t\right)\Delta t. \qquad (13.20)
\end{aligned}$$

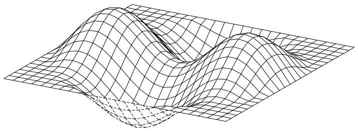
Example 13.3.3 Determine an approximate solution to

$$\begin{aligned}
\dot{x} &= -x^3 + \sin(tx) \\
x(0) &= 1
\end{aligned}$$

using the third order Runge-Kutta method.

This is simply a matter of substituting into equations 13.18 and 13.19 as follows:

$$x(t + \Delta t) = x(t) + \frac{1}{6}\left(v_1 + 4v_2 + v_3\right)$$



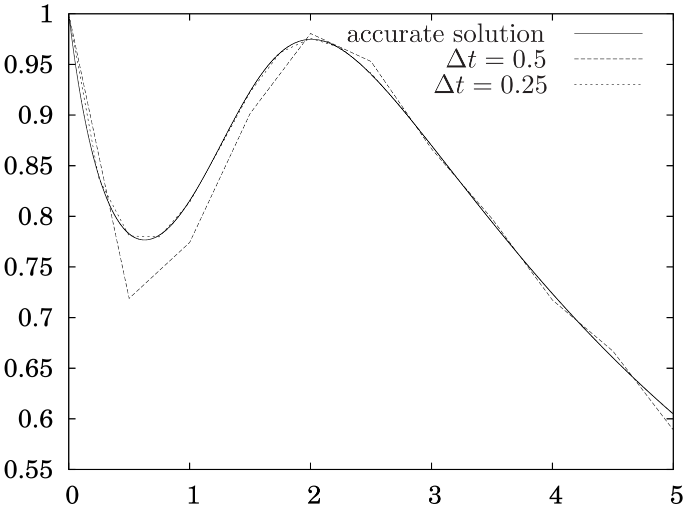


Figure 13.8. Accurate and approximate solutions for example 13.3.3. The third order Runge-Kutta has a local truncation error proportional to $(\Delta t)^4$ and an overall accumulated error proportional to $(\Delta t)^3$.

where

$$\begin{aligned} v_1 &= \left(-(x(t))^3 + \sin(tx(t))\right)\Delta t \\ v_2 &= \left(-\left(x(t)+\frac{1}{2}v_1\right)^3 + \sin\left(\left(t+\frac{1}{2}\Delta t\right)\left(x(t)+\frac{1}{2}v_1\right)\right)\right)\Delta t \\ v_3 &= \left(-(x(t)+v_2)^3 + \sin((t+\Delta t)(x(t)+v_2))\right)\Delta t. \end{aligned}$$

An accurate solution determined with a very small time step as well as approximate solutions for $\Delta t = 0.5$ and $\Delta t = 0.25$ are illustrated in Figure 13.8. Note the substantial increase in accuracy when the time step is cut by a factor of two. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

13.3.4 The fourth order Runge-Kutta method

Again, the idea is exactly the same as the previous Runge-Kutta derivations. The famous fourth order Runge-Kutta formula is

$$x(t+\Delta t) = x(t) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \tag{13.21}$$

where

$$\begin{aligned} k_1 &= f\left(x(t),t\right)\Delta t \\ k_2 &= f\left(x(t)+\frac{1}{2}k_1, t+\frac{1}{2}\Delta t\right)\Delta t \\ k_3 &= f\left(x(t)+\frac{1}{2}k_2, t+\frac{1}{2}\Delta t\right)\Delta t \\ k_4 &= f\left(x(t)+k_3, t+\Delta t\right)\Delta t. \end{aligned} \tag{13.22}$$

Example 13.3.4 Determine an approximate solution to

$$\begin{aligned} \dot{x} &= -x^3+\sin(tx) \\ x(0) &= 1 \end{aligned}$$

using the fourth order Runge-Kutta method.

This is simply a matter of substituting into equations 13.21 and 13.22 as follows:

$$x(t+\Delta t) = x(t)+\frac{1}{6}\left(k_1+2k_2+2k_3+k_4\right)$$

where

$$\begin{aligned} k_1 &= \left(-\left(x(t)\right)^3+\sin\left(tx(t)\right)\right)\Delta t \\ k_2 &= \left(-\left(x(t)+\frac{1}{2}k_1\right)^3+\sin\left(\left(t+\frac{1}{2}\Delta t\right)\left(x(t)+\frac{1}{2}k_1\right)\right)\right)\Delta t \\ k_3 &= \left(-\left(x(t)+\frac{1}{2}k_2\right)^3+\sin\left(\left(t+\frac{1}{2}\Delta t\right)\left(x(t)+\frac{1}{2}k_2\right)\right)\right)\Delta t \\ k_4 &= \left(-\left(x(t)+k_3\right)^3+\sin\left(\left(t+\Delta t\right)\left(x(t)+k_3\right)\right)\right)\Delta t. \end{aligned}$$

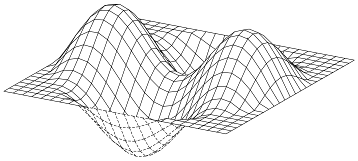
An accurate solution determined with a very small time step as well as approximate solutions for $\Delta t = 0.5$ and $\Delta t = 0.25$ are illustrated in Figure 13.9. Note the substantial increase in accuracy when the time step is cut by a factor of two and the generally better accuracy than the lower order methods for the same time steps. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

13.4 Error Analysis

13.4.1 Local truncation error

In each of the methods we have considered, we have explicitly accounted for a certain number of terms in the Taylor series expansion

$$x\left(t+\Delta t\right) = x(t)+f(x(t),t)\Delta t+\frac{1}{2!}\frac{df(x(t),t)}{dt}\left(\Delta t\right)^2+\frac{1}{3!}\frac{d^2f(x(t),t)}{dt^2}\left(\Delta t\right)^3+\cdots.$$



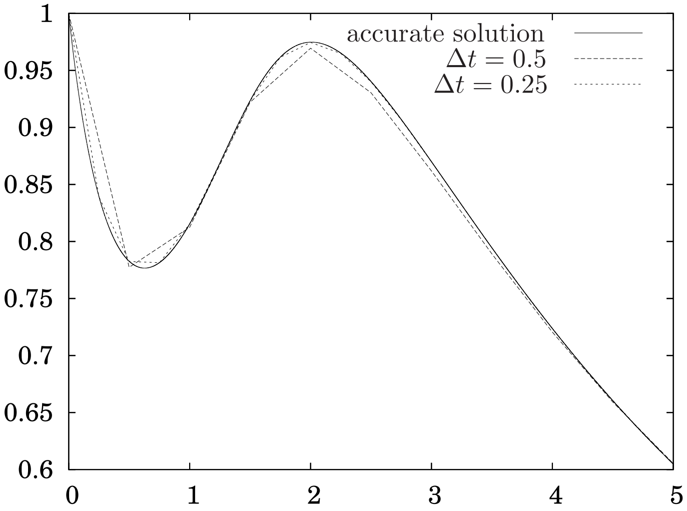


Figure 13.9. Accurate and approximate solutions for example 13.3.4. The fourth order Runge-Kutta has a local truncation error proportional to $(\Delta t)^5$ and an overall accumulated error proportional to $(\Delta t)^4$.

Method	Local Truncation Error	Overall Error
Euler First order Taylor Series First order Runge-Kutta	$\sim (\Delta t)^2$	$\sim (\Delta t)$
Second order Taylor Series	$\sim (\Delta t)^3$	$\sim (\Delta t)^2$
Second order Runge-Kutta Improved Euler	$\sim (\Delta t)^3$	$\sim (\Delta t)^2$
Third order Taylor Series	$\sim (\Delta t)^4$	$\sim (\Delta t)^3$
Third order Runge-Kutta	$\sim (\Delta t)^4$	$\sim (\Delta t)^3$
Fourth order Runge-Kutta	$\sim (\Delta t)^5$	$\sim (\Delta t)^4$

Table 13.1. Local truncation error and overall error for various numerical method schemes for $\Delta t \ll 1$. Equivalent methods are listed in the same row.

The local truncation error is the error introduced at each time step that arises because only a finite number of terms in the Taylor series expansion are used. In contrast, the overall error or global error is the error at a given t. Since the number of time steps required to reach a given t increases as Δt decreases, the overall error is proportional to one order less of Δt than the local truncation error.

13.4.2 Global error

13.4.3 Some subtleties

It is worth emphasizing that the analyses presented in sections 13.4.1 and 13.4.2 are true in general, but that does not preclude the existence of somewhat pathological cases that seemingly defy the rules. Such cases are presented by way of examples.

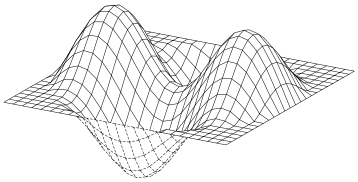
Example 13.4.1 Consider

$$\begin{aligned} \dot{x} + 3x &= 15\left(\cos 3t + \sin 3t\right) \\ x(0) &= 0. \end{aligned}$$

It is straightforward to verify that

$$x(t) = 5 \sin 3t$$

is the solution to this initial value problem. The first three steps of the output of the algorithm for the case where $\Delta t = 0.25$ is as follows. The first column is time, the second is the approximate solution at that time,



the third column is the exact solution and the fourth column is the error:

t	$x(t)$	$5\sin 3t$	$5\sin 3t - x(t)$
0.000000	0.000000	0.000000	0.000000
0.250000	3.750000	3.408194	−0.341806
0.500000	6.237479	4.987475	−1.250004

The first three time steps for the case where the time step has been reduced in half to $\Delta t = 0.125$ is as follows:

t	$x(t)$	$5\sin 3t$	$5\sin 3t - x(t)$
0.000000	0.000000	0.000000	0.000000
0.125000	1.875000	1.831363	−0.043637
0.250000	3.603338	3.408194	−0.195144

Comparing the error after the first time step in each case, since the error has been reduced by approximately a factor of eight, it is tempting to conclude that the method used must be a second order method, *i.e.,* the improved Euler method, 2nd order Runge-Kutta or a second order Taylor series method. However, Figure 13.10 is a plot of the two approximate solutions and the exact solution for the time interval $0 < t \leq 1$. Note that the overall error has been reduced by a factor of two, rather than a factor of four as would be expected by a second order method.

This apparent contradiction is resolved by studying the exact solution. Since

$$x(t) = 5\sin 3t \qquad \blacksquare$$

the Taylor series for $x(t)$ is

$$\begin{aligned} x(t+\Delta t) &= x(t) + \left.\frac{dx}{dt}\right|_t (\Delta t) + \frac{1}{2}\left.\frac{d^2x}{dt^2}\right|_t (\Delta t)^2 + \cdots \\ &= 5\sin 3t + 15\cos 3t\,(\Delta t) - \frac{45}{2}\sin 3t\,(\Delta t)^2 + \cdots \end{aligned}$$

Since every other term is contains $\sin 3t$, at $t = 0$, every other term is zero. Thus when comparing the local truncation error by examining the error after the first time step, a first order method will look like a second order method due to the fact that the coefficient of the $(\Delta t)^2$ term in the Taylor series is zero. Similarly, a third order method will look like a fourth order one, *etc.*. After the first time step, however, the relevant coefficients are nonzero, and hence the global error behaves as expected. A program listing in C for this problem implementing Euler, 2nd order Runge-Kutta, a 2nd order Taylor series expansion and 4th order Runge-Kutta appears in Appendix D.1.4.

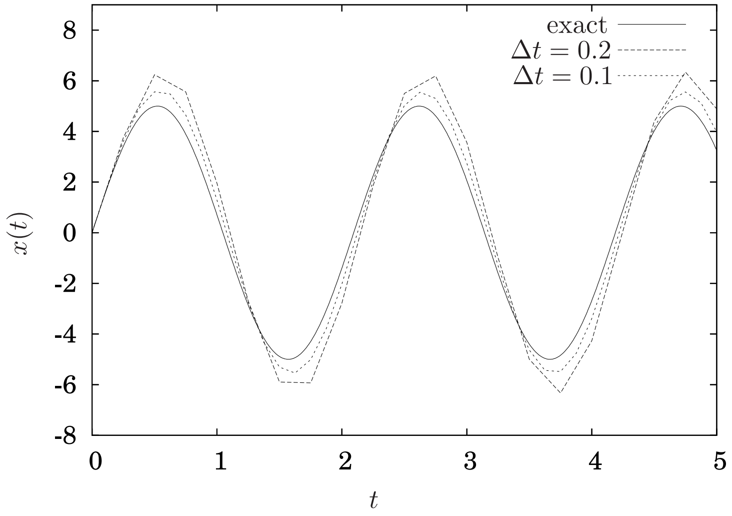


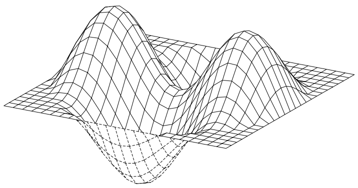
Figure 13.10. Exact and approximate solutions for example 13.4.1 exhibiting an overall accumulated error proportional to Δt.

13.5 Numerical Methods for Higher-Order Systems

All the examples so far have been for first order ordinary differential equations. This section will present the relatively easy extension to systems of first order equations and higher order ordinary differential equations and highlight the one subtlety with respect to computer implementation of the algorithms.

13.5.1 Systems of first order, ordinary differential equations

As a matter of notation, the extension of each of the numerical methods presented in sections 13.1 through 13.3 to systems of differential equations is simply a matter of converting the equations to vector notation. As a matter of substance is is a matter of considering multivariable Taylor series expansions. This section will present the details of Euler's method for systems of equations but simply present the results for the other methods since providing the details would be rather cumbersome with little added pedagogical insight.



13.5.2 Higher order, ordinary differential equations

Consider the system of first order differential equations

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1(t), x_2(t), \ldots, x_n(t), t) \\
\dot{x}_2 &= f_2(x_1(t), x_2(t), \ldots, x_n(t), t) \\
\vdots \quad & \quad \vdots \\
\dot{x}_n &= f_1(x_1(t), x_2(t), \ldots, x_n(t), t).
\end{aligned} \tag{13.23}$$

Expanding each of the x_i in a Taylor series gives

$$\begin{aligned}
x_1(t+\Delta t) &= x_1(t) + \frac{dx_1}{dt}\Delta t + \frac{1}{2}\frac{d^2x_1}{dt^2}(\Delta t)^2 + \cdots \\
x_2(t+\Delta t) &= x_2(t) + \frac{dx_2}{dt}\Delta t + \frac{1}{2}\frac{d^2x_2}{dt^2}(\Delta t)^2 + \cdots \\
\vdots \quad & \quad \vdots \\
x_n(t+\Delta t) &= x_n(t) + \frac{dx_n}{dt}\Delta t + \frac{1}{2}\frac{d^2x_n}{dt^2}(\Delta t)^2 + \cdots ,
\end{aligned}$$

or expressing the derivatives in terms of the functions f_i gives

$$\begin{aligned}
x_1(t+\Delta t) &= x_1(t) + f_1(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t + \frac{1}{2}\frac{df_1}{dt}(\Delta t)^2 + \cdots \\
x_2(t+\Delta t) &= x_2(t) + f_2(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t + \frac{1}{2}\frac{df_2}{dt}(\Delta t)^2 + \cdots \\
\vdots \quad & \quad \vdots \\
x_n(t+\Delta t) &= x_n(t) + f_1(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t + \frac{1}{2}\frac{df_n}{dt}(\Delta t)^2 + \cdots ,
\end{aligned}$$

where the derivatives of each of the f_i are evaluated at $(x_1(t), x_2(t), \ldots, x_n(t), t)$.

13.5.3 Euler's method

Euler's method for a single first order equation was based upon simply keeping the terms in the Taylor series of $x(t)$ up through the Δt term, and the same is easily done in the case of a system of equations. In particular, Euler's method is simply written as

$$\begin{aligned}
x_1(t+\Delta t) &= x_1(t) + f_1(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t \\
x_2(t+\Delta t) &= x_2(t) + f_2(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t \\
\vdots \quad & \quad \vdots \\
x_n(t+\Delta t) &= x_n(t) + f_n(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t
\end{aligned} \tag{13.24}$$

Rewriting all of this in vector notation simplifies the expressions and furthermore makes the relationship between the methods for systems of equations and for a single first order equation transparent. Let

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \tag{13.25}$$

and

$$\mathbf{f}(x_1(t), x_2(t), \ldots, x_n(t), t) = \begin{bmatrix} f_1(x_1(t), x_2(t), \ldots, x_n(t), t) \\ f_2(x_1(t), x_2(t), \ldots, x_n(t), t) \\ \vdots \\ f_n(x_1(t), x_2(t), \ldots, x_n(t), t) \end{bmatrix}$$

or substituting the vector notation for $\mathbf{x}(t)$ from equation 13.25

$$\mathbf{f}(\mathbf{x}(t), t) = \begin{bmatrix} f_1(\mathbf{x}(t), t) \\ f_2(\mathbf{x}(t), t) \\ \vdots \\ f_n(\mathbf{x}(t), t) \end{bmatrix}.$$

Then the original system of equations expressed in equation 13.23 simply becomes

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

which looks remarkably like equation 13.1 with a few of the terms in bold face font. Furthermore, expressing equation 13.24 in this notation reduces the expression to

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \mathbf{f}(\mathbf{x}(t), t)\Delta t, \tag{13.26}$$

which, again, is exactly the same as the equation for Euler's method for a single first order equation with the vector terms in bold face.

Example 13.5.1 Determine an approximate numerical solution to

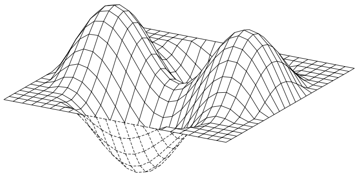
$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= (1 - x^2)y - x \end{aligned}$$

where

$$\begin{aligned} x(0) &= 0.02 \\ y(0) &= 0.0 \end{aligned}$$

using Euler's method. Substituting into equation 13.26 gives the system of equations

$$\begin{aligned} x(t + \Delta t) &= y(t)\Delta t \\ y(t + \Delta t) &= \left(\left(1 - (x(t))^2\right) y(t) - x(t)\right) \Delta t. \end{aligned}$$



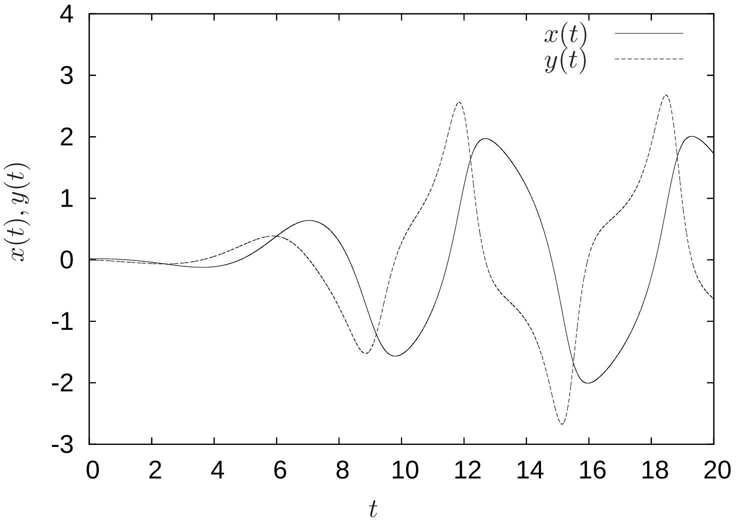


Figure 13.11. Numerical solutions for the system of equations example 13.5.1 using Euler's method.

Figure 13.11 illustrates both components of the solution for $0 < t < 20$. A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4. ■

Observe that the right hand side of equation 13.26 is *evaluated at time t*. It is very easy to write a computer program that does not quite do that, as the following example illustrates.

Example 13.5.2 Consider the system from example 13.5.1 and the following lines of code:

```
x = x + y*dt;
y = y + ((1.0 - x*x)*y - x)*dt;
```

This seemingly incorrectly implements Euler's method because the value for x, which appears on the right hand side of the second, y, equation has already been changed from $x(t)$ to $x(t + \Delta t)$ by the first line.

While this approach deviates from the exact expression for Euler's method (and does indeed result in a different approximate solution since the second equation uses the $x(t+\Delta t)$ values instead of $x(t)$) it is inconsequential with respect to the accuracy of the method. To see this consider the second equation using the “incorrect” method:

$$\begin{aligned} y(t) &= y(t) + ((1 - x(t+\Delta t)x(t+\Delta t))y(t) - x(t+\Delta t))\Delta t \\ &= y(t) + [(1 - (x(t) + f(x,y,t)\Delta t)\,(x(t) + f(x,y,t)\Delta t))\,y(t) - \\ &\quad (x(t) + f(x,y,t)\Delta t)]\,\Delta t \\ &= y(t) + [(1 - x(t)x(t))\,y(t) - x(t)]\,\Delta t + \mathcal{O}\left((\Delta t)^2\right), \end{aligned}$$

where the notation $\mathcal{O}\left((\Delta t)^2\right)$ means a collection of terms that multiply $(\Delta t)^2$.

The bottom line is that while this approach modifies the second equation and “adds” some extra terms to the expression for $y(t+\Delta t)$, these added terms are of a higher order than the accuracy of the method, and hence do not affect the order of accuracy of the approach. ■

13.5.4 Second order Taylor Series

Extending equation 13.24 to the $(\Delta t)^2$ term gives

$$\begin{aligned} x_1(t+\Delta t) &= x_1(t) + f_1(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t + \frac{1}{2}\frac{df_1}{dt}(\Delta t)^2 \quad (13.27) \\ x_2(t+\Delta t) &= x_2(t) + f_2(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t + \frac{1}{2}\frac{df_2}{dt}(\Delta t)^2 \\ &\vdots \qquad \vdots \\ x_n(t+\Delta t) &= x_n(t) + f_n(x_1(t), x_2(t), \ldots, x_n(t), t)\Delta t + \frac{1}{2}\frac{df_n}{dt}(\Delta t)^2. \end{aligned}$$

Since each component of $\mathbf{f}$ possibly depends on *each* x_i as well as t, and each of the x_i depends on t, we have

$$\frac{df_i}{dt} = \frac{\partial f_i}{\partial x_1}f_1 + \frac{\partial f_i}{\partial x_2}f_2 + \cdots + \frac{\partial f_i}{\partial x_n}f_n + \frac{\partial f_i}{\partial t}.$$

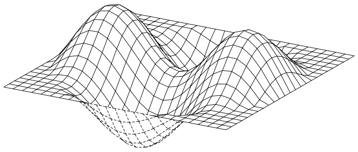
Expanding each derivative term in equation 13.27 would be cumbersome, so to use a more compact notation, recall the definition of the Jacobian

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}.$$

Using this, the the Taylor series to second order may be written in vector form as

$$\mathbf{x}(\mathbf{t}+\boldsymbol{\Delta}\mathbf{t}) = \mathbf{x}(\mathbf{t}) + \mathbf{f}(\mathbf{x},\mathbf{t}) + \frac{\mathbf{1}}{\mathbf{2}}\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\mathbf{f}\,(\boldsymbol{\Delta}\mathbf{t})^{\mathbf{2}}. \qquad (13.28)$$

Obviously, computing all the partial derivatives would be a hassle. There is not much point to doing so since the same accuracy may be obtained by using the Runge-Kutta methods, as is outlined subsequently.



13.5.5 Fourth order Runge-Kutta

Rather than provide the details for every method, this section skips right to the fourth order Runge-Kutta method. From it, the generalizations necessary to implement the other methods for systems of equations should be obvious. Similar to the manner in which Euler's method generalized to the case of a system of equations, fourth order Runge-Kutta may be expressed as the following. For

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

let

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \frac{1}{6}\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\right) \tag{13.29}$$

where

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}\left(\mathbf{x}(t), t\right)\Delta t \qquad (13.30)\\
\mathbf{k}_2 &= \mathbf{f}\left(\mathbf{x}(t) + \frac{1}{2}\mathbf{k}_1, t + \frac{1}{2}\Delta t\right)\Delta t\\
\mathbf{k}_3 &= \mathbf{f}\left(\mathbf{x}(t) + \frac{1}{2}\mathbf{k}_2, t + \frac{1}{2}\Delta t\right)\Delta t\\
\mathbf{k}_4 &= \mathbf{f}\left(\mathbf{x}(t) + \mathbf{k}_3, t + \Delta t\right)\Delta t.
\end{aligned}$$

Note that $\mathbf{k}_1$ through $\mathbf{k}_4$ are vector quantities since $\mathbf{f}(\mathbf{x}(t), t)$ is a vector.

Example 13.5.3 Determine an approximate solution to

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

where

$$\begin{aligned}
x(0) &= 0.02\\
y(0) &= 0.0
\end{aligned}$$

using the fourth order Runge-Kutta method.

Let

$$\mathbf{x}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

and

$$\mathbf{f}(\mathbf{x}(t), t) = \begin{bmatrix} x_2(t) \\ \left(1 - (x_1(t))^2\right) x_2(t) - x_1(t)\sin t \end{bmatrix}.$$

Then

$$
\begin{aligned}
\mathbf{k}_1 &= \begin{bmatrix} k_{11} \\ k_{21} \end{bmatrix} = \begin{bmatrix} x_2 \\ \left(1-(x_1)^2\right) x_2 - x_1 \sin t \end{bmatrix} \Delta t \\
\mathbf{k}_2 &= \begin{bmatrix} k_{12} \\ k_{22} \end{bmatrix} \\
&= \begin{bmatrix} x_2 + \frac{1}{2}k_{21} \\ \left(1-\left(x_1+\frac{1}{2}k_{11}\right)^2\right)\left(x_2+\frac{1}{2}k_{21}\right) - \left(x_1+\frac{1}{2}k_{11}\right)\sin\left(t+\frac{1}{2}\Delta t\right) \end{bmatrix} \Delta t \\
\mathbf{k}_3 &= \begin{bmatrix} k_{13} \\ k_{23} \end{bmatrix} \\
&= \begin{bmatrix} x_2 + \frac{1}{2}k_{22} \\ \left(1-\left(x_1+\frac{1}{2}k_{12}\right)^2\right)\left(x_2+\frac{1}{2}k_{22}\right) - \left(x_1+\frac{1}{2}k_{12}\right)\sin\left(t+\frac{1}{2}\Delta t\right) \end{bmatrix} \Delta t \\
\mathbf{k}_4 &= \begin{bmatrix} k_{14} \\ k_{24} \end{bmatrix} \\
&= \begin{bmatrix} x_2 + k_{23} \\ \left(1-\left(x_1+k_{13}\right)^2\right)\left(x_2+k_{23}\right) - \left(x_1+k_{13}\right)\sin\left(t+\Delta t\right) \end{bmatrix} \Delta t,
\end{aligned}
$$

where all of the x_i terms are evaluated at t. Then finally

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \frac{1}{6}\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\right).$$

A program listing in C for this problem appears in Appendix D.1.4. A program listing in FORTRAN for this problem appears in Appendix D.2.4.
■

Note that when writing a program for the system in example 13.5.3, it may be tempting to compute all the k values for the x term first, followed by all the k values for the y term. However, note that since the equations are coupled, the k_{12} term (the "second" x term), for example, depends upon k_{21} (the "first" y term). Hence, it is necessary to compute all the components of the *vector* $\mathbf{k}_1$ first, followed by all the components of $\mathbf{k}_2$, *etc.* The following example illustrates this fact.

Example 13.5.4 Compute an approximate numerical solution for

$$
\begin{aligned}
\dot{x} &= y \\
\dot{y} &= -x,
\end{aligned}
$$

where

$$
\begin{aligned}
x(0) &= 0 \\
y(0) &= 1
\end{aligned}
$$

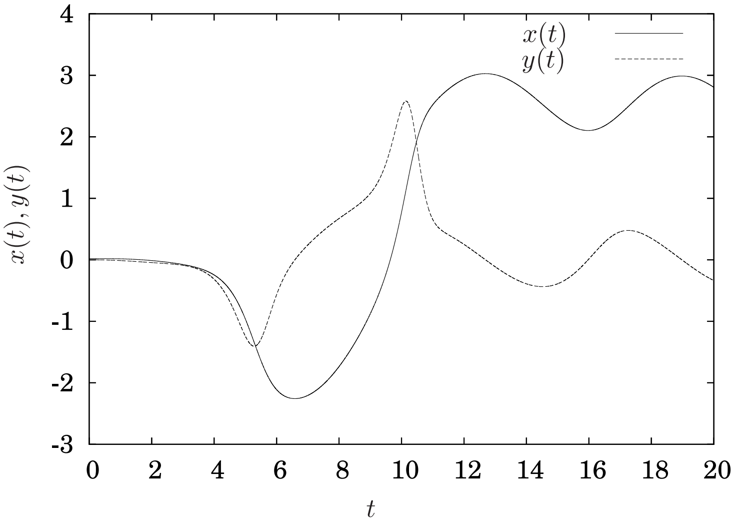


Figure 13.12. Approximate numerical solutions for the system of equations from example 13.5.3 using the fourth order Runge-Kutta method.

using the fourth order Runge-Kutta method. Compare the approximate solution when the terms in the algorithm are computed in the correct and incorrect order.

Using the notation from example 13.5.3, the correct order of computation for the k values is

$$k_{11}, k_{21}, k_{12}, k_{22}, k_{13}, k_{23}, k_{14}, k_{24}.$$

By comparison, the incorrect, but tempting order, is

$$k_{11}, k_{12}, k_{13}, k_{14}, k_{21}, k_{22}, k_{23}, k_{24}.$$

Figure 13.13 illustrates an accurate solution and compares the approximate solutions for $x(t)$ for both cases when $\Delta t = 0.4$. Clearly, the correct approach produces a much more accurate approximate solution. ■

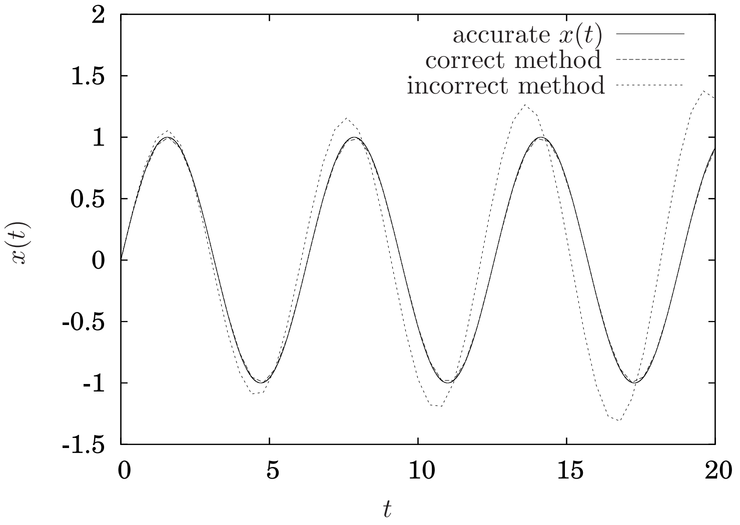


Figure 13.13. Comparison of approximate solutions from example 13.5.4 when making a common error in implementing the fourth order Runge-Kutta algorithm for systems of first order differential equations.

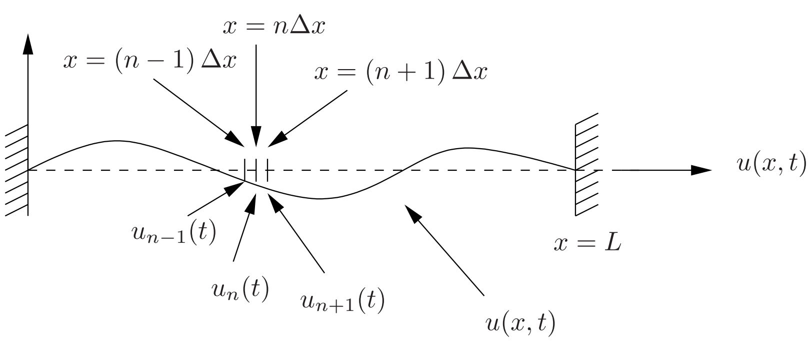


Figure 13.14. Vibrating string.

13.6 Convergence

13.7 Numerical Methods for Partial Differential Equations

The focus of this section will be on the so-called *finite difference method* for partial differential equations. The main idea is, similar to the manner in which time was discritized for determining a numerical approximation to the time derivative for ordinary differential equations, that the spatial dimension(s) must be similarly discritized for partial differential equations with an independent variable corresponding to the spatial direction. Doing so will result in a system of coupled *ordinary* differential equations, which then may be solved using the methods from the previous sections of this chapter.

13.7.1 Finite Difference Approximation

Consider, for example, the the function $u(x,t)$ that describes the solution to the wave equation

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}. \tag{13.31}$$

Obviously, if the solution is known and x is fixed, then the solution is only a function of time. In particular, consider the fixed x values, $n\Delta x$ where $n = 0, 1, 2, \ldots, N$ (so $L = N\Delta x$) and $\Delta x \ll 1$ is a fixed. In that case, define $u_n(t)$ to be

$$u_n(t) = u(n\Delta x, t) \qquad n = 0, 1, 2, \ldots, N$$

which is the motion of the string at the fixed location $x = n\Delta x$ that is only a function of time. This is illustrated in Figure 13.14 for the string example.

Since $u_n(t)$ is only a function of time, it stands to reason that it is governed by an *ordinary* differential equation with independent variable t. The purpose of the finite difference method is to approximate the second order spatial derivative on the left hand side of equation 13.31 with values of $u(x,t)$ at other fixed values.

In particular, consider approximating the first partial derivative with respect to x at $x = n\Delta x$. This is simple enough using the definition of the derivative

$$\frac{\partial u_n(t)}{\partial x} \approx \frac{u_n(t) - u_{n-1}(t)}{\Delta x}$$

if $\Delta x \ll 1$. Now, since the second derivative is the derivative of the derivative, then

$$\begin{aligned}
\frac{\partial^2 u_n(t)}{\partial x^2} &\approx \frac{\frac{\partial u_{n+1}(t)}{\partial x} - \frac{\partial u_n(t)}{\partial x}}{\Delta x} \\
&\approx \frac{\left(\frac{u_{n+1}(t)-u_n(t)}{\Delta x}\right) - \left(\frac{u_n(t)-u_{n-1}(t)}{\Delta x}\right)}{\Delta x} \\
&= \frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{(\Delta x)^2}.
\end{aligned}$$

Finally, substituting this approximation back into equation 13.31 gives

$$\alpha^2 \frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{(\Delta x)^2} = \frac{d^2 u_n(t)}{dt^2}. \tag{13.32}$$

Since the boundary conditions are specified, $u_0(t)$ and $u_N(t)$ are specified, which means equation 13.32 is $N-1$ ordinary, linear, constant coefficient, homogeneous, second order, coupled differential equations. In a more expanded form (and dropping the explicit dependence of u_n on t)

$$\begin{aligned}
\ddot{u}_1 &= (u_2 - 2u_1 + u_0)\left(\frac{\alpha^2}{\Delta x}\right)^2 \\
\ddot{u}_2 &= (u_3 - 2u_2 + u_1)\left(\frac{\alpha^2}{\Delta x}\right)^2 \\
\ddot{u}_3 &= (u_4 - 2u_3 + u_2)\left(\frac{\alpha^2}{\Delta x}\right)^2 \\
\vdots &= \vdots \\
\ddot{u}_{N-1} &= (u_N - 2u_{N-1} + u_{N-2})\left(\frac{\alpha^2}{\Delta x}\right)^2.
\end{aligned}$$

Thus, determining a numerical approximation to the wave equation amounts to determining the numerical approximations to these $N-1$ ordinary differential equations. Since this is a system of ordinary differential equations, an approximation method will be necessary for the time derivative as well, and may be, for example, implemented using Euler's method, Runge-Kutta, *etc.*

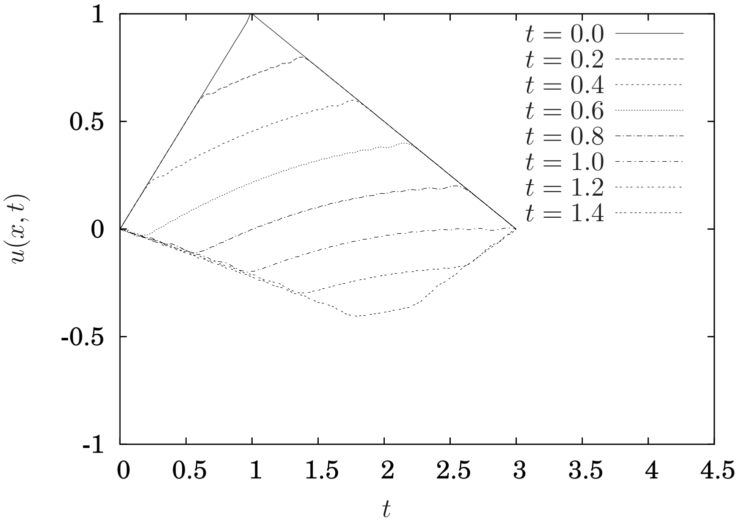


Figure 13.15. Numerical approximate solution for wave equation in example 13.7.1 with $\Delta x = 0.03$ and $\Delta t = 0.000225$. ■

Example 13.7.1 Use the finite difference method and Euler's method to determine an approximate solution to

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}$$

where $L = 3$ and $\alpha = 2$, subjected to the boundary conditions

$$\begin{aligned} u(0,t) &= 0 \\ u(L,t) &= 0 \end{aligned}$$

and initial conditions

$$\begin{aligned} u(x,0) &= \begin{cases} x & x \leq 1 \\ \frac{3-x}{2} & 1 < x \leq 3 \end{cases} \\ \left.\frac{\partial u}{\partial t}\right|_{t=0} &= 0. \end{aligned}$$

This is the same system that was solved analytically in example 12.1.2.

The solution for $\Delta x = 0.03$ and $\Delta t = 0.000225$ for various t values is illustrated in Figure 13.15. Close examination indicates that the solution is not sufficiently accurate. There appears to be dissipation that is not present in the exact solution from example 12.1.2.

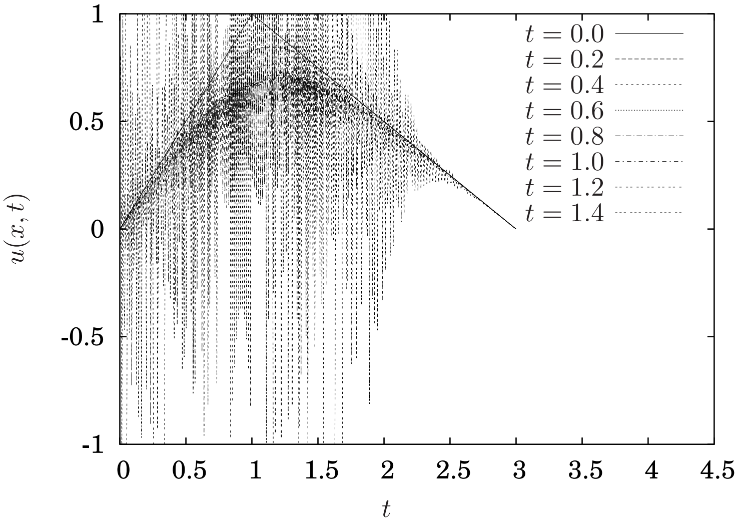


Figure 13.16. Numerical approximate solution for wave equation in example 13.7.1 with $\Delta x = 0.03$ and $\Delta t = 0.000225$.

At this point it is not necessarily clear whether to reduce Δx, Δt or both. Without further information, let us simply reduce Δx by a factor of four to see if there is improvement.

Example 13.7.2 This example considers the same system with a smaller $\Delta x = 0.0075$.

The solution for the same system as in example 12.1.2 with $\Delta x = 0.0075$ and $\Delta t = 0.000225$ for various t values is illustrated in Figure 13.16. This solution, unfortunately, is not even close to the solution from example 12.1.2. Clearly, reducing Δx did not have the intended consequence. ■

The result from example 13.7.2 should be surprising even to the most casual reader. In an attempt to increase the accuracy of the solution, the size of Δx was decreased by a factor of four and the result was that the approximate solution became unstable. Recall that this system is exactly the same as from example 12.1.2, so we know that the exact solution is stable, so the approximate solution became *worse* instead of better when Δx was decreased. The reason this happened is that there is a relationship between Δx and Δt that is necessary for the stability of the numerical solution. Investigating this will be the subject of the next section.

13.7.2 Numerical stability

Heat equation

Let us consider the heat equation with homogeneous boundary conditions

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \alpha^2 \frac{\partial^2 u}{\partial x^2} \\
u(0,t) &= 0 \\
u(L,t) &= 0 \\
u(x,0) &= f(x).
\end{aligned}$$

From section 12.3, the exact solution is given by

$$\begin{aligned}
u(x,t) &= \sum_{n=1}^{\infty} c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \sin \frac{n\pi x}{L} \\
c_n &= \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx.
\end{aligned}$$

Since the solution for $u(x,t)$ is the superposition of an infinite number of modes, let us consider what happens to *one* mode when it is used in the finite difference method. In particular, consider

$$u(x,t) = c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \sin \frac{n\pi x}{L}$$

for some integer n. Using Euler's method for the time steps and the finite difference method for the second derivative,

$$\begin{aligned}
u(x,t+\Delta t) &= u(x,t) + \left(\frac{\alpha}{\Delta x}\right)^2 \left(u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)\right) \Delta t \\
&= c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \sin \frac{n\pi x}{L} + \left(\frac{\alpha}{\Delta x}\right)^2 c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \left(\sin \frac{n\pi\,(x+\Delta x)}{L} - \right. \\
&\quad \left. 2 \sin \frac{n\pi x}{L} + \sin \frac{n\pi\,(x-\Delta x)}{L}\right) \Delta t \\
&= c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \sin \frac{n\pi x}{L} + \left(\frac{\alpha}{\Delta x}\right)^2 c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \cdot \\
&\quad \left(\left(\sin \frac{n\pi x}{L} \cos \frac{n\pi\Delta x}{L} + \sin \frac{n\pi\Delta x}{L} \cos \frac{n\pi x}{L}\right) - 2 \sin \frac{n\pi x}{L}\right. \\
&\quad \left.\left(\sin \frac{n\pi x}{L} \cos \frac{n\pi\Delta x}{L} - \sin \frac{n\pi\Delta x}{L} \cos \frac{n\pi x}{L}\right)\right) \Delta t \\
&= c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \sin \frac{n\pi x}{L} + \left(\frac{\alpha}{\Delta x}\right)^2 c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \cdot \\
&\quad \left(2 \sin \frac{n\pi x}{L} \cos \frac{n\pi\Delta x}{L} - 2 \sin \frac{n\pi x}{L}\right) \Delta t \\
&= c_n e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} \sin \frac{n\pi x}{L} \left[1 + 2\Delta t \left(\frac{\alpha}{\Delta x}\right)^2 \left(\cos \frac{n\pi\Delta x}{L} - 1\right)\right] \qquad (13.33)
\end{aligned}$$

At this point a little interpretation is required.

1. The value for the solution of the nth mode at time $t+\Delta t$ is the value at time t scaled by the term in the parentheses. If the term in parentheses has a magnitude greater than one, the mode will grow; conversely, if the magnitude is less than one, it will decay.

2. Note that
$$-2 \leq \cos\frac{n\pi\Delta x}{L} - 1 \leq 0.$$
Hence, when $n = N$, this term will have the largest magnitude, *i.e.*,
$$\left|\cos\frac{N\pi\Delta x}{L} - 1\right| = 2.$$

3. Do not be tempted to assume that $\cos\frac{n\pi\Delta x}{L} \approx 1$ because Δx is small, because it may be the case that n is large.

4. The exact solution for the left hand side of the equation is
$$\begin{aligned} u(x, t+\Delta t) &= c_n e^{-\frac{\alpha^2 n^2 \pi^2 (t+\Delta t)}{L^2}} \sin\frac{n\pi x}{L} \\ &= c_n \sin\frac{n\pi x}{L} e^{-\frac{\alpha^2 n^2 \pi^2 t}{L^2}} e^{-\frac{\alpha^2 n^2 \pi^2 \Delta t}{L^2}}. \end{aligned}$$
Comparing this with the left hand side of equation 13.33 we can see that the solution will be exactly correct if Δt is such that
$$e^{-\frac{\alpha^2 n^2 \pi^2 \Delta t}{L^2}} = 1 + 2\Delta t \left(\frac{\alpha}{\Delta x}\right)^2 \left(\cos\frac{n\pi\Delta x}{L} - 1\right).$$
Unfortunately, this is a transcendental equation which will be generally difficult to solve; furthermore, different values for n will require different Δt.

5. Pursuing the notion developed above that $n = N$ will be the most unstable mode, let $n = N$ and furthermore, pick Δt so that the right hand side of the equation is zero, *i.e.*,
$$\Delta t = \frac{1}{2}\left(\frac{\Delta x}{\alpha}\right)^2.$$
Substituting this into the right hand side gives
$$\begin{aligned} e^{-\frac{\alpha^2 n^2 \pi^2 \Delta t}{L^2}} &= e^{-\frac{\alpha^2 n^2 \pi^2 \frac{1}{2}\left(\frac{\Delta x}{\alpha}\right)^2.}{L^2}} \\ &= e^{-\frac{\pi^2}{2}}. \end{aligned}$$

Wave equation (incomplete and wrong!)

In order to establish some insight into the reason that reducing Δx made the situation worse in examples 13.7.1 and 13.7.2, consider the exact solution

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} \cos \frac{\alpha n \pi t}{L}.$$

In example 13.7.2 it appears that the higher order modes were unstable. To understand the mechanism for this, substitute one of the modes from the exact solution into equation 13.32. Let $u(x,t)$ be only the ith mode

$$u(x,t) = b_i \sin \frac{i\pi x}{L} \cos \frac{\alpha i \pi t}{L} \qquad 1 < i < N.$$

Then, Euler's method should be of the form

$$\begin{aligned} u(x, t+\Delta t) &= u(x,t) + \dot{u}(x,t)\Delta t \\ \dot{u}(x, t+\Delta t) &= \dot{u}(x,t) + \ddot{u}(x,t)\Delta t, \end{aligned}$$

where $\dot{u}(x,t)$ and $\ddot{u}(x,t)$ are the first and second partial derivatives respectively of $u(x,t)$ with respect to time. Substituting the assumed form for $u(x,t)$ into these equations gives

$$\begin{aligned} u(x, t+\Delta t) &= u(x,t) - b_i \frac{\alpha i \pi}{L} \sin \frac{i\pi x}{L} \sin \frac{\alpha i \pi t}{L} \Delta t \\ \dot{u}(x, t+\Delta t) &= \dot{u}(x,t) - b_i \left(\frac{\alpha i \pi}{L}\right)^2 \sin \frac{i\pi x}{L} \cos \frac{\alpha i \pi t}{L} \Delta t. \qquad (13.34) \end{aligned}$$

Note that these represent the manner in which one mode in the exact solution should appear in Euler's formula. If the terms multiplying Δt in either equation are larger or smaller, then the contribution of the ith mode to the exact solution will either decay or grow, when, in fact, the magnitude should remain constant. Since the first equation is simply the integral of the second, let us consider equation 13.34 only.

Using the approximation for the second derivative from equation 13.32,, Euler's method for $\dot{u}_i(t)$ is

$$\begin{aligned} \dot{u}(x, t+\Delta t) &= \dot{u}(x,t) + \left(\frac{\alpha}{\Delta x}\right)^2 \left(u(x+\Delta x, t) - 2u(x,t) + u(x-\Delta x t)\right)\Delta t \\ &= \dot{u}(x,t) + \left(\frac{\alpha}{\Delta x}\right)^2 b_i \cos \frac{\alpha i \pi t}{L} \left(\sin \frac{i\pi\,(x+\Delta x)}{L} - 2\sin\frac{i\pi x}{L} + \sin \frac{i\pi\,(x-\Delta x)}{L}\right)\Delta t \\ &= \dot{u}(x,t) + \left(\frac{\alpha}{\Delta x}\right)^2 b_i \cos \frac{\alpha i \pi t}{L} \left(2\sin\frac{i\pi x}{L}\cos\frac{i\pi\Delta x}{L} - 2\sin\frac{i\pi x}{L}\right)\Delta t \\ &= \dot{u}(x,t) + 2\left(\frac{\alpha}{\Delta x}\right)^2 b_i \cos \frac{\alpha i \pi t}{L} \sin\frac{i\pi x}{L}\left(\cos\frac{i\pi\Delta x}{L} - 1\right)\Delta t \qquad (13.35) \\ &= \dot{u}(x,t) - 4\left(\frac{\alpha}{\Delta x}\right)^2 b_i \cos \frac{\alpha i \pi t}{L} \sin\frac{i\pi x}{L} \sin^2\frac{i\pi\Delta x}{2L}\Delta t. \qquad (13.36) \end{aligned}$$

Note that the largest the sin^2 term can be is 1 and this will occur when $\frac{i\Delta x}{2L} = 1$, or when $i = \frac{2L}{\Delta x} = 2N$. For this mode, the *magnitude* of $u(x,t)$ is b_i and the magnitude of $\dot{u}(x,t)$ is $\frac{\alpha i\pi}{L} b_i$. Referring to equation 13.36, the magnitude of the right hand side will be

$$\frac{2\alpha N\pi}{L} = \frac{2\alpha N\pi}{L} - 4\left(\frac{\alpha}{\Delta x}\right)^2\left(\sin^2\frac{2N\pi\Delta x}{L}\right)\Delta t,$$

or

$$\frac{2\alpha\pi}{\Delta x} = \frac{2\alpha\pi}{\Delta x}$$

Finish this mess!!!

13.8 Exercises

Problem 13.1 Consider

$$\begin{aligned} \dot{x} + x &= \sin t \\ x(0) &= -1. \end{aligned}$$

1. Write and submit a listing of a computer program to compute an approximate numerical solution for this differential equation using

   (a) Euler's method;

   (b) the second order Taylor series method; and,

   (c) the fourth order Runge-Kutta method.

   You may decided to write three separate programs or include all three methods in one program.

2. For each of the following time steps

   (a) $\Delta t = 0.5$

   (b) $\Delta t = 0.25$

   (c) $\Delta t = 0.125$

   (d) $\Delta t = 0.01$

   submit a plot of the exact solution and the approximate solution using each of the three methods for the time interval $t = 0$ to $t = 10$. Thus, there should be four plots and each plot should have four curves.

3. Submit a plot illustrating the difference between the exact solution and the numerically computed solutions for the same time steps and time interval as in part 2 above. In each case indicate the factor by which the global error changes as the time step changes and indicate whether such a factor would be expected for the global truncation error for the corresponding method.

4. What is the difference between the exact solution and the numerically computed solutions after the first time step for each method and time step size above? Determine the factor by which this error (the local truncation error) changes as the time step changes and indicate whether such a factor corresponds to what is theoretically expected.

Chapter 14

Introduction to Nonlinear Systems

A quick review of the subject matter of this book up to this point will confirm the fact that, with the exception of some specific first order, ordinary differential equations which happen to be exact or separable, all the solution methods covered so far have only been applicable to *linear* differential equations. While the study of nonlinear differential equations is extremely interesting, it is also substantively difficult and rather advanced. This chapter will only introduce some of the reasons why nonlinear systems are important, why they are interesting, and the main tool used to deal with them, which is to simply determine the *linear* differential equation that best approximates the nonlinear equation.

14.1 Motivation: Complexity of Nonlinear Systems

By way of one example, this section illustrates a couple aspects of the complexity of nonlinear systems and introduces the the *phase plane*, which is one specific tool that is useful for two dimensional systems. The example is the famous *Duffing equation* and both the forced and unforced case will be considered.

Example 14.1.1 Consider

$$\ddot{x} + b\dot{x} - x + x^3 = 0. \tag{14.1}$$

Since this equation is nonlinear in x, none of the methods considered previously in this text are applicable to determine a solution, so this example will solve it numerically. Using the fourth order Runge-Kutta method to determine an approximate solution to this equation with

$$\begin{aligned} b &= 0.2 \\ x(0) &= -1 \end{aligned}$$

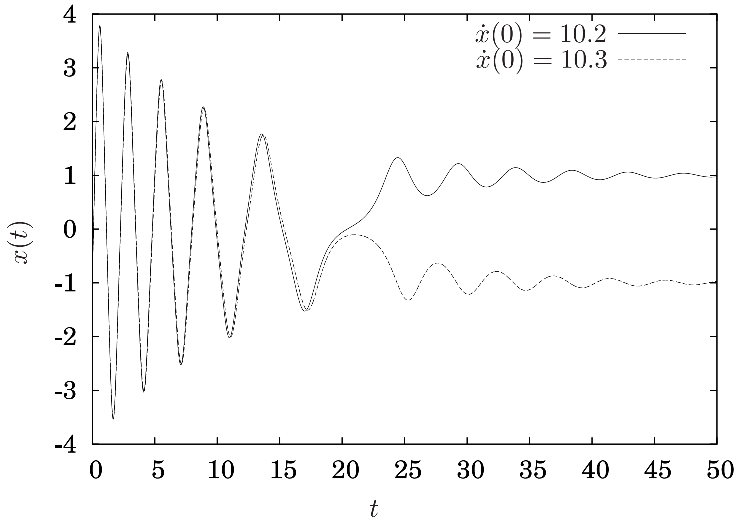


Figure 14.1. Solutions for equation 14.1 from example 14.1.1 for two slightly different initial conditions.

and $\dot{x}(0)$ equal to 10.2 and 10.3, the solutions are illustrated in Figure 14.1.

The obvious feature of these two solutions is that while the initial condition was changed very slightly, and indeed the two solutions were nearly indistinguishable up until approximately $t = 20$, near that time the solutions rather radically diverged and ultimately appeared to approach different steady state values. While it is the case that such features are very sensitive to numerical errors, it will hopefully be clear subsequently that such a feature is inherent in this system and is actually fundamental feature of it.

This example illustrates the fact that solutions to nonlinear differential equations may be sensitive to initial conditions and furthermore may have multiple equilibria. In this case, the equilibria illustrated are the two steady state solutions in Figure 14.1; namely, $\lim_{t\to\infty}(x(t), \dot{x}(t)) = (1, 0)$ and $\lim_{t\to\infty}(x(t), \dot{x}(t)) = (-1, 0)$.

Remark 14.1.2 The approximate numerical solutions used in this example were computed with a very small time step. As of the time of this writing other numerical packages may give a slightly different answers. In particular, since the solutions are so sensitive to initial conditions, slightly different initial conditions may be necessary to produce a similar result. $\diamond$

The following example illustrates the fact that an additional complexity, namely, a time-varying inhomogeneous term, may result in a chaotic solution.

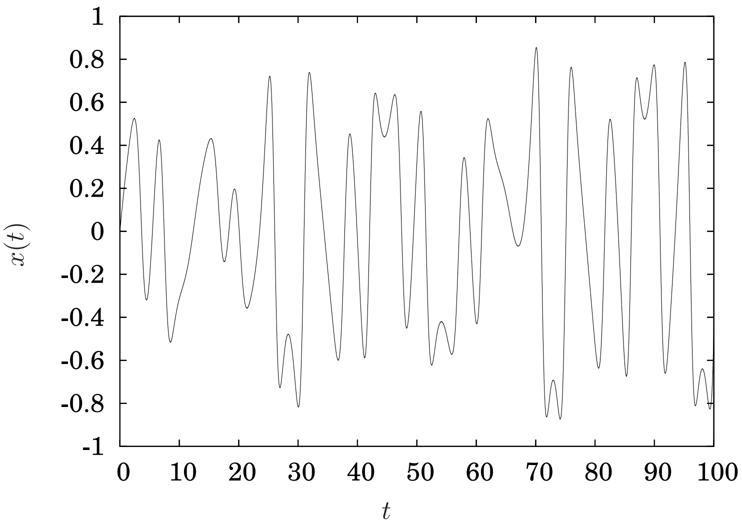


Figure 14.2. Chaotic solution to equation 14.2 from example 14.1.3.

Example 14.1.3 Consider

$$\ddot{x} + b\dot{x} - x + x^3 = \gamma \cos \omega t, \tag{14.2}$$

where

$$\begin{aligned} b &= 0.2 \\ \gamma &= 0.3 \\ \omega &= 1.0. \end{aligned}$$

A plot of a numerical solution to this equation with

$$\begin{aligned} x(0) &= 0 \\ \dot{x}(0) &= 0 \end{aligned}$$

is illustrated in Figure 14.2.

While any precise definition of the term chaos is beyond the scope of this book, it is clear from the solution illustrated in Figure 14.2 that the numerical solution is "chaotic" at least in the sense of the common use of the term. At least for the time interval plotted, the solution does not appear to repeat, *i.e.,* it is non-periodic, and seems to evolve in a rather unpredictable way. ■

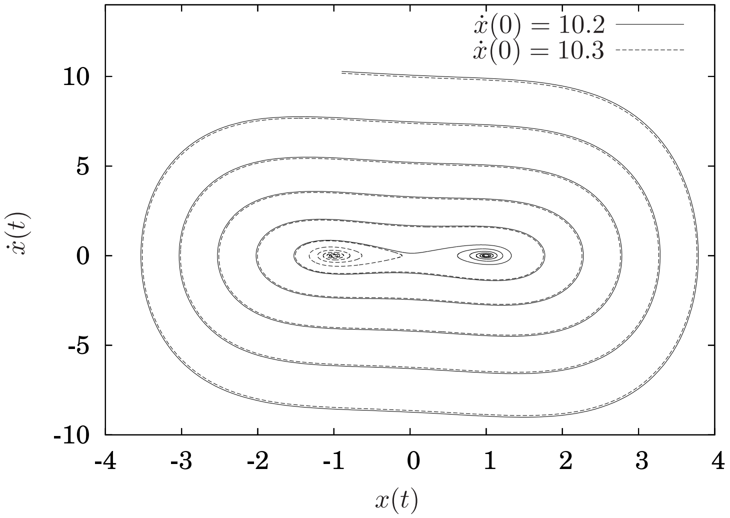


Figure 14.3. Phase portraits for solutions to equation 14.1 from example 14.1.1.

14.1.1 The phase plane

The *phase portrait* of a solution to a differential equation is a parametric plot of the solution *versus* its derivatives up to the derivative that is one less than the order of the differential equation. The parameter that varies in the plot is the independent variable, *i.e.,* usually t. For a second order system, a phase portrait is two dimensional and the domain of the plot is often referred to as the *phase plane*.

Example 14.1.4 The phase portraits for the solutions from Examples 14.1.1 and 14.1.3 are illustrated in Figures 14.3 and 14.4 respectively.

While arguably there is not much to be gained from the second figure, Figure 14.4, the first figure, Figure 14.3 is somewhat enlightening. Judging from the point where the two solutions diverge after closely tracking each other for quite a long time, it is reasonable to infer that the geometric structure of the origin in the phase plane may be significant. In fact, as will be developed subsequently, this is indeed the case. ■

14.1.2 Poincare sections

As indicated in example 14.1.4, other than the chaotic nature of the solution, there is not much to observe from the solution to equation 14.2 illustrated in Figure 14.4. However, one slight modification of the manner in which the data

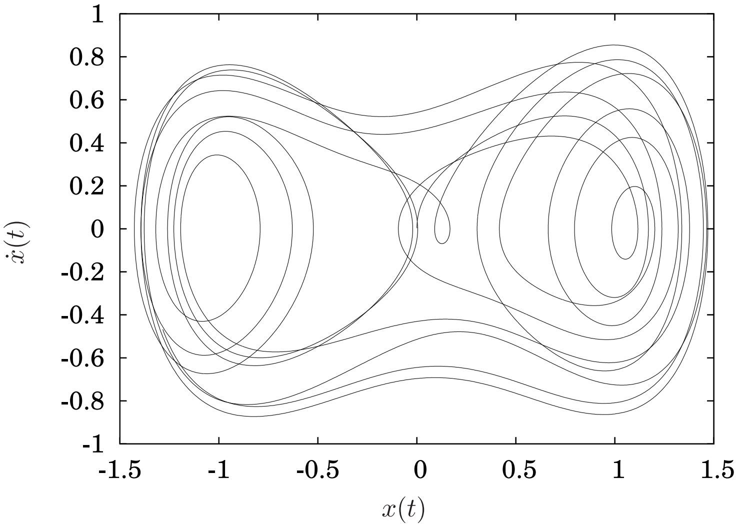


Figure 14.4. Phase portrait for solutions to equation 14.2.

is illustrated reveals some very interesting structure to the solution. In particular, instead of plotting the complete solution curves $x(t)$ *vs.* $\dot{x}(t)$, Figure 14.5 illustrates the discrete values of $x(t)$ *vs.* $\dot{x}(t)$ for $t = 0, 2\pi, 3\pi, 4\pi \ldots$.

Example 14.1.5 Considering again the system in equation 14.2 and computing an approximate numerical solution for the same parameter values and initial conditions, but for a much larger time range $0 \leq t < 3000\pi$, a plot of the discrete values of $x(t)$ *versus* $\dot{x}(t)$ for the $t = 2m\pi$ is illustrated in Figure 14.5.

Note that a rather coherent structure becomes apparent when the data is presented in this manner. This topic will not be pursued further in this text, but the reader should be at least made aware of its existence and its name: a *strange attractor*. ■

14.2 Linearization

One obvious approach to attempt to determine at least the basic features of a nonlinear differential equation is to determine a differential equation that we can solve that is a good approximation to the nonlinear differential equation. In the case where the nonlinear equation is homogeneous and has constant coefficients, if a good *linear* approximation can be determined then, since it can be solved, at least some of the features of the solution of the nonlinear equation may be determined from the solution to the linear one.

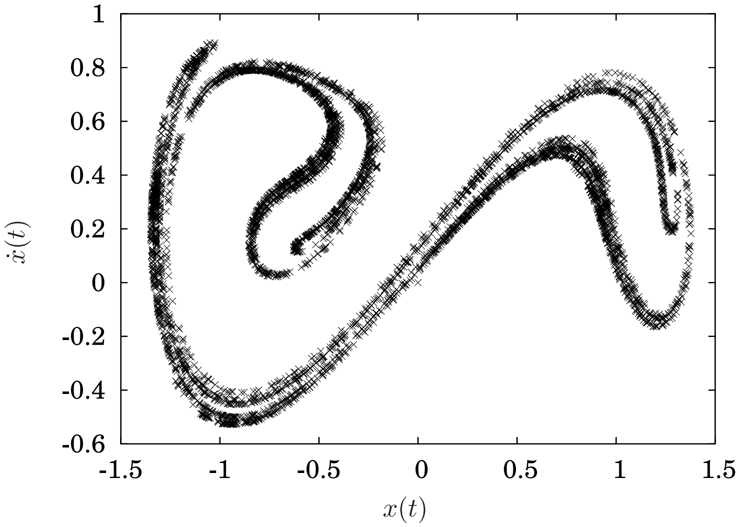


Figure 14.5. Poincare section for the forced Duffing equation in example 14.1.3.

The initial approach presented will be to simply compute a Taylor series for all of the nonlinear terms about some point and keep only the first two terms from the Taylor series, which will result in a linear differential equation. The following example illustrates this approach.

Example 14.2.1 Consider

$$\ddot{x} + \dot{x} - 2x + x^3 = 0. \tag{14.3}$$

Determine a linear approximation for this equation by substituting a Taylor series for the x^3 term, where the Taylor series is computed about $x = \sqrt{2}$. Since the Taylor series for x^3 about $x = x_0$ is

$$x^3 = x_0^3 + 3x_0^2(x - x_0) + 3x_0(x - x_0)^2 + \cdots$$

keeping only the first two terms gives

$$\begin{aligned} \ddot{x} + \dot{x} - 2x + \left(x_0^3 + 3x_0^2\,(x - x_0)\right) &= 0 \\ \ddot{x} + \dot{x} + \left(3x_0^2 - 2\right)x &= 2x_0^3 \end{aligned} \tag{14.4}$$

which is a constant coefficient, linear, inhomogeneous differential equation which we know how to solve.

Substituting $x_0 = \sqrt{2}$ equation 14.4 gives

$$\ddot{x} + \dot{x} + 4x = 4\sqrt{2}, \tag{14.5}$$

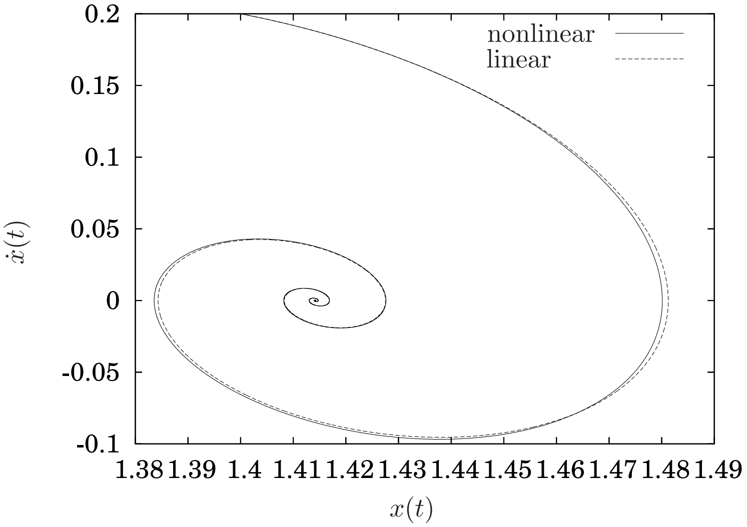


Figure 14.6. Comparison of solutions of equation 14.3 (nonlinear) with equation 14.5 (linear approximation) with initial conditions near $x = \sqrt{2}$.

which is easily solved. The particular solution is $x_p = \sqrt{2}$ and substituting $e^{\lambda t}$ into the homogeneous equation gives

$$\lambda^2 + \lambda + 4 = 0 \quad \Longleftrightarrow \quad \lambda = \frac{-1}{2} \pm i\frac{\sqrt{15}}{2},$$

so the solution to the linear approximation is

$$x(t) = c_1 e^{-\frac{1}{2}t} \cos \frac{\sqrt{15}}{2}t + c_2 e^{-\frac{1}{2}t} \sin \frac{\sqrt{15}}{2}t + \sqrt{2}. \tag{14.6}$$

■

Intuitively, in example 14.2.1, the solutions to the linear approximation, equation 14.5 will be approximately the same as the solutions to equation 14.3 *as long as* $x \approx \sqrt{2}$. Since only the first two terms of the Taylor series were used, then the neglected terms, which were the higher powers of $(x - x_0)$ will only be small if x stays near $\sqrt{2}$. The following example illustrates this fact.

Example 14.2.2 Figure 14.6 illustrates the solutions to equation 14.3 and 14.5 for $x(0) = 1.4$ and $\dot{x}(0) = 0.2$. Note that the approximate solution closely tracks the solution to the nonlinear equation.

If the initial conditions are moved farther away from the point of linearization, say $x(0) = 1.0$ and $\dot{x}(0) = 0.2$, as is illustrated in Figure 14.7 the linear solution is not as good of an approximation to the nonlinear solution as was the case illustrated in Figure 14.6.

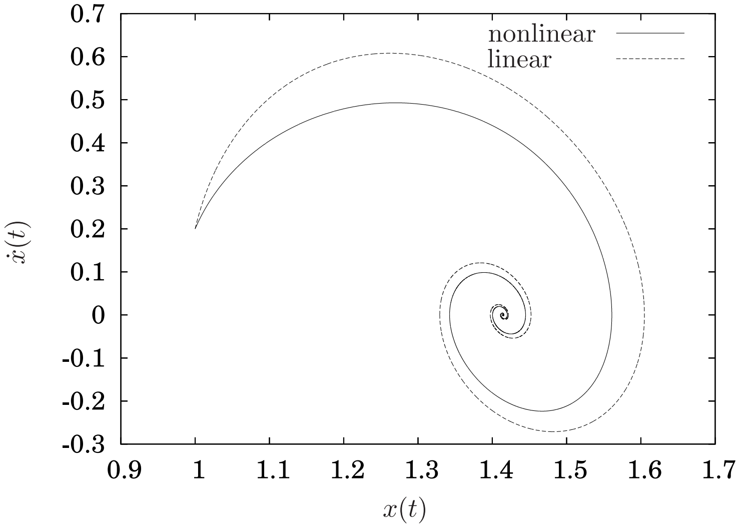


Figure 14.7. Comparison of solutions of equation 14.3 (nonlinear) with equation 14.5 (linear approximation) with initial conditions slightly farther from $x = \sqrt{2}$ than in Figure 14.6.

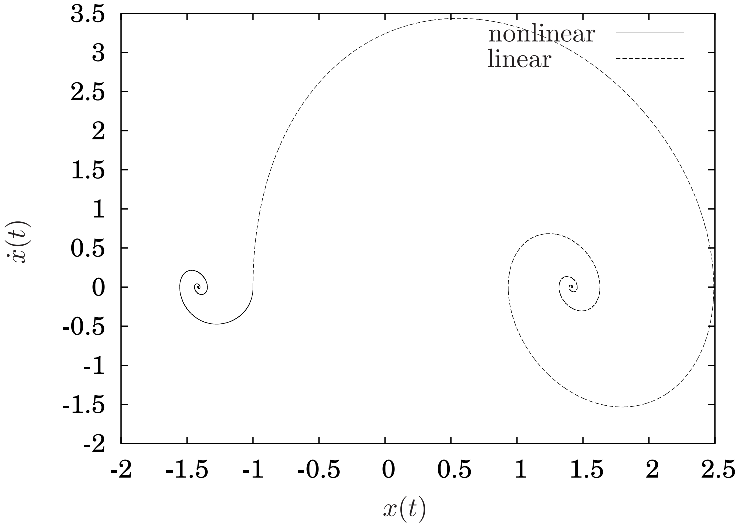


Figure 14.8. Comparison of solutions of equation 14.3 (nonlinear) with equation 14.5 (linear approximation) with initial conditions far from $x = \sqrt{2}$.

However, if the initial conditions are farther away from $x = \sqrt{2}$, say $x(0) = -1$, $\dot{x}(0) = 0.0$, then the two solutions are as illustrated in Figure 14.8. Since the solution is not near the point of linearization, the solution to the linearized differential equation is not even remotely a good approximation to the solution to the nonlinear equation. ■

Now, let us investigate what is happening near the origin.

Example 14.2.3 Determine the best linear approximation to

$$\ddot{x} + \dot{x} - 2x + x^3 = 0$$

for values of x near 0. Substituting $x_0 = 0$ into equation 14.4 gives

$$\ddot{x} + \dot{x} - 2x = 0,$$

which is linear, constant coefficient, homogeneous and ordinary, so solutions are of the form $x = e^{\lambda t}$. Substituting gives the characteristic equation

$$\lambda^2 + \lambda - 2 = 0 \quad \Longleftrightarrow \quad \lambda = 1, -2.$$

Hence, the general solution is of the form

$$x(t) = c_1 e^t + c_2 e^{-2t}.$$

Note that this solution is *unstable* unless $c_1 = 0$. Computing $\dot{x}(t)$ gives

$$\begin{aligned} x(t) &= c_1 e^t + c_2 e^{-2t} \\ \dot{x}(t) &= c_1 e^t - 2c_2 e^{-2t}. \end{aligned}$$

Expressing this in vector form gives

$$\begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^t + c_2 \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-2t}. \tag{14.7}$$

Considering the solution in this manner indicates that in the phase plane, any initial condition which is exactly a multiple of the vector

$$\begin{bmatrix} x(0) \\ \dot{x}(0) \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

will result in $c_1 = 0$ and hence will be stable, *i.e.,*

$$\lim_{t \to \infty} \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Any other initial condition will have a nonzero c_1 and hence will be unstable. If the initial condition is very close to the

$$\begin{bmatrix} x(0) \\ \dot{x}(0) \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

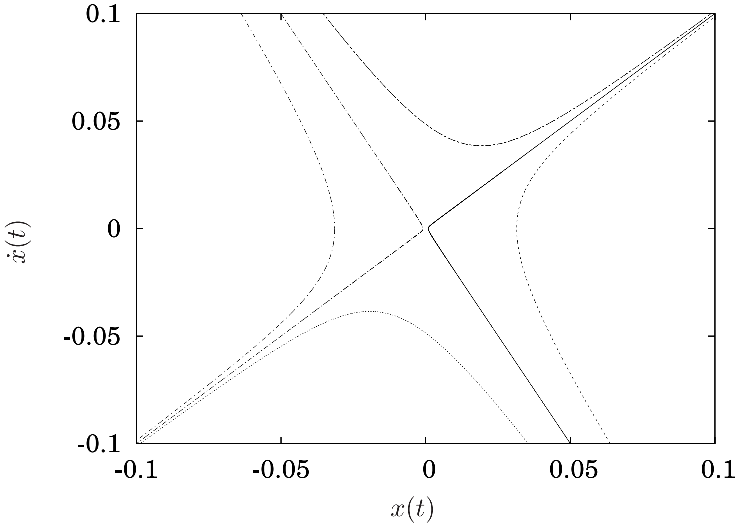


Figure 14.9. Solution of the linear approximation to equation 14.3 near the origin.

vector, then c_1 may be very small and the stable solution may initially dominate; however, due to the exponential term the solution will ultimately be unstable.

In Figure 14.9 of the six solutions plotted, three start in the upper left portion of the plot and initially move toward the origin. Similarly, the other three solutions start in the lower right portion of the graph and initially head to the origin as well. However, ultimately the e^t term dominates and all six solutions ultimately move away from the origin and grow unbounded.

Figure 14.10 illustrates the solutions to the linear approximation near the origin (equation 14.4 with $x_0 = 0$) and the solution to the nonlinear equation (equation 14.3) with the same initial conditions near the origin. While initially the solutions are similar in nature, due to the instability of both solutions they both ultimately leave the domain in which the linear equation is a good approximation to the nonlinear equation. ■

In principle, it is appropriate to compute a Taylor series approximation for any nonlinear terms in a differential equation and keep only the linear terms to determine some of the features of the solution of the nonlinear equation near the point about which the linearization was computed. However, the main utility of linearization is to determine a linear approximation to a differential equation *near an equilibrium point* since that will provide information about the stability of the equilibrium point. Furthermore, if the equilibrium point is indeed stable, then the solutions to the linear approximation will be close to

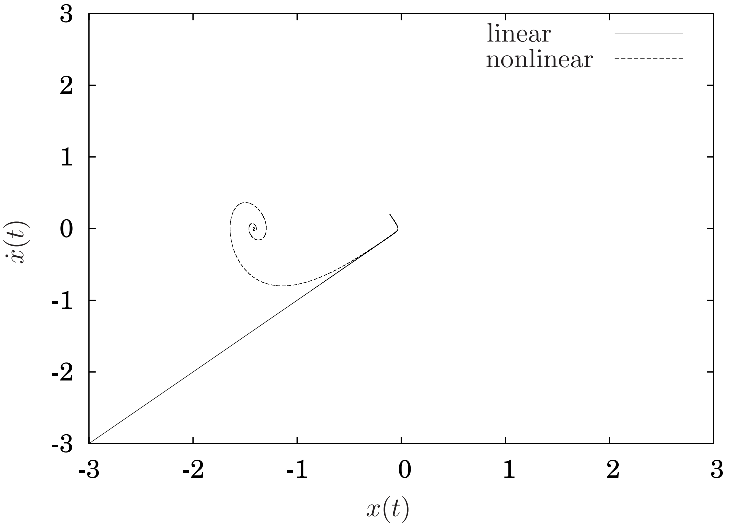


Figure 14.10. Comparison of solutions of linear approximation and nonlinear solution for initial conditions near the origin.

the solutions to the nonlinear equation. The following example illustrates the fact that while a linear approximation computed near a non-equilibrium point gives some information about the nonlinear solution, it is only transiently valid and furthermore, information regarding an equilibrium point of the linearized equation has nothing to do with the nonlinear system.

Example 14.2.4 Determine a linear approximation to

$$\ddot{x} + \dot{x} - 2x + x^3 = 0 \tag{14.8}$$

near $x_0 = 4$. Substituting $x_0 = 4$ into equation 14.4 gives

$$\begin{aligned} \ddot{x} + \dot{x} + -2x\,(64 + 48\,(x-4)) &= 0 \\ \ddot{x} + \dot{x} + 46x &= 128 \end{aligned} \tag{14.9}$$

which has the general solution

$$x(t) = c_1 e^{-\frac{1}{2}t} \cos \frac{\sqrt{183}}{2} t + c_2 e^{-\frac{1}{2}t} \sin \frac{\sqrt{183}}{2} t + \frac{64}{23}.$$

A plot of the the solution to the nonlinear equation and the linear approximation is illustrated in Figure 14.11.

Note that while the solutions stay near $x = 4$, they are nearly identical. However, as expected, as they diverge from $x = 4$ the linear approximation

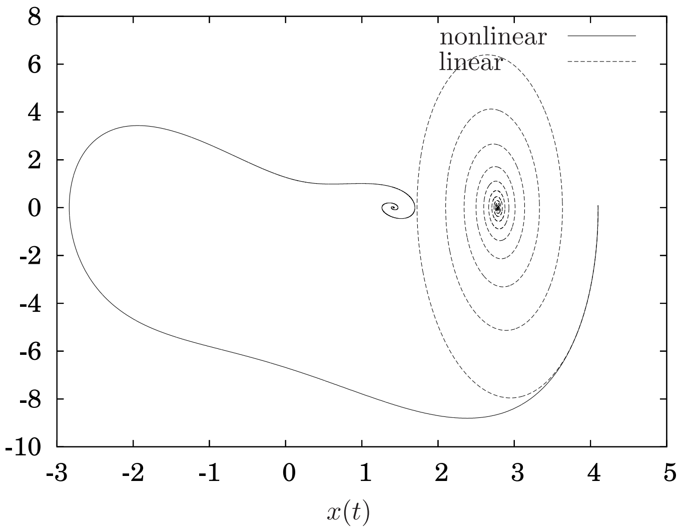


Figure 14.11. Solutions to equation 14.8 and 14.9.

is increasingly less valid. Also note that the linearized equation has an equilibrium at $x = \frac{64}{23}$ which is *not* an equilibrium for the nonlinear equation. The stability of the equilibrium point for the linearized equation at $x = \frac{64}{23}$ has nothing to do with the stability or instability of the nonlinear equation near that point. ■

Typically, linear approximations to nonlinear differential equations are only computed about equilibrium points. This is due to the fact that the linear approximation about a non-equilibrium point in general will have an equilibrium that is not the same as the nonlinear equation. Furthermore, in applications such as feedback control, stabilizing a system to an equilibrium is typically the desired goal and hence it is desirable for the linearized approximation of the nonlinear equation to have an equilibrium in common.

The next sections considers the more standard and systematic approach to doing this; namely, if necessary, converting a system of higher order differential equations to s system of first order equations and computing the Jacobian.

14.3 Jacobian Linearization

This will initially be developed by mirroring the example from the previous section, example 14.2.1.

Example 14.3.1 Convert

$$\ddot{x} + \dot{x} - 2x + x^3 = 0 \tag{14.10}$$

into a system of two first order equations. Using the standard approach of letting

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \end{aligned}$$

the following is equivalent to equation 14.10

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_2 + 2x_1 - x_1^3 \end{bmatrix}. \tag{14.11}$$

Adopting a notation that mirrors that of chapter 6, let

$$\xi = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and

$$f(\xi) = f(x_1, x_2) = \begin{bmatrix} x_2 \\ -x_2 + 2x_1 - x_1^3 \end{bmatrix}.$$

■

Observe carefully that in example 14.3.1, both ξ and $f(\xi)$ are *vectors* and that the whole system may be represented abstractly as

$$\dot{\xi} = f(\xi),$$

as long as the original equation is homogeneous. Since it may be represented so compactly, even though it is, in general, a system of differential equations, it may simply be referred to a *a* or *the* differential equation.

Definition 14.3.2 A point, ξ_0 is an *equilibrium point* of

$$\dot{\xi} = f(\xi)$$

if

$$f(\xi_0) = 0,$$

where the 0 on the right hand side of the equation is a vector of zeros that is the same dimension as ξ and $f(\xi)$. ⋄

Note, that if ξ_0 is an equilibrium point, then

$$\dot{\xi} = f(\xi_0) = 0$$

so

$$\xi(t) = \xi_0$$

is a solution to

$$\dot{\xi} = f(\xi)$$

if

$$\xi(0) = \xi_0.$$

Example 14.3.3 Determine the equilibrium points for equation 14.11. Clearly, $x_2 = 0$ is necessary to make the first component vanish. For the second component any of the three $x_1 = 0$ or $x_1 = \pm\sqrt{2}$. So, any of the three vectors

$$\xi_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} -\sqrt{2} \\ 0 \end{bmatrix}$$

when substituted into equation 14.11 will result in

$$f(\xi_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

■

First we will define the Jacobian and by referring to the previous examples will show that it may be used to determine an equivalent linear approximation to a nonlinear equation near an equilibrium point.

Definition 14.3.4 For a vector valued function, f of a vector

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix},$$

denoted by

$$f(\xi) = \begin{bmatrix} f_1(\xi_1, \xi_2, \ldots, \xi_n) \\ f_2(\xi_1, \xi_2, \ldots, \xi_n) \\ \vdots \\ f_m(\xi_1, \xi_2, \ldots, \xi_n) \end{bmatrix},$$

the *Jacobian matrix* for $f(\xi)$ is given by

$$\frac{\partial f}{\partial \xi} = \begin{bmatrix} \frac{\partial f_1}{\partial \xi_1} & \frac{\partial f_1}{\partial \xi_2} & \cdots & \frac{\partial f_1}{\partial \xi_n} \\ \frac{\partial f_2}{\partial \xi_1} & \frac{\partial f_2}{\partial \xi_2} & \cdots & \frac{\partial f_2}{\partial \xi_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial \xi_1} & \frac{\partial f_m}{\partial \xi_2} & \cdots & \frac{\partial f_m}{\partial \xi_n} \end{bmatrix}$$

where $f(\xi) \in \mathbb{R}^m$, *i.e.,* f is m elements "tall" and $\xi \in \mathbb{R}^n$, *i.e.,* ξ is n elements tall.

⋄

For a system of n first order homogeneous differential equations, the equations themselves will only depend on the n state variables; hence, for systems of first order equations, the Jacobian matrix will always be square. Now, we can define a linearization that will be equivalent to the Taylor series method outlined previously.

Definition 14.3.5 For the system of n first order, homogeneous differential equations

$$\dot{\xi} = f(\xi) \tag{14.12}$$

where $f(\xi_0) = 0$, the *linear approximation to equation 14.12 about* ξ_0 is given by

$$\dot{\xi} = \left.\frac{\partial f}{\partial \xi}\right|_{\xi_0} (\xi - \xi_0)\,. \tag{14.13}$$

$\diamond$

Let us compare the results of using this linearization method with the linearization approximations determined using Taylor series in the previous examples.

Example 14.3.6 Consider

$$\ddot{x} + \dot{x} - 2x + x^3 = 0,$$

which, when converted to two first order equations is given by

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_2 + 2x_1 - x_1^3 \end{bmatrix},$$

where

$$\xi = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix}.$$

The Jacobian for this system of equations is

$$\frac{df}{d\xi} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 2 - 3x_1^2 & -1 \end{bmatrix}.$$

Evaluated at the equilibrium point

$$\xi_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and substituting into equation 14.13 gives

$$\dot{\xi} = \frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

From example 14.2.3, the Taylor series linearization about $x_0 = 0$ was

$$\ddot{x} + \dot{x} - 2x = 0,$$

which, when converted to two first order equations, gives the same result.

Similarly, at

$$\xi_0 = \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$$

the linearization is

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -4 & -1 \end{bmatrix}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}\right).$$

Referring back to example 14.2.1, the Taylor series linearization resulted in

$$\ddot{x} + \dot{x} + 4x = 4\sqrt{2},$$

which is the same. ■

It is worth remarking that the general equation for the Taylor series of a vector valued function of several variables about the point ξ_0 is of the form

$$f(\xi) = f(\xi_0) + \left.\frac{\partial f}{\partial \xi}\right|_{\xi_0} (\xi - \xi_0) + \cdots .$$

Since ξ_0 was assumed to be an equilibrium point in the above development, what is obviously happening is that $f(\xi)$ is replaced by the first two terms in its Taylor series, but the first term $f(\xi_0)$ happens to be zero.

Finally, while there is not anything wrong with equation 14.13, the inhomogeneous term resulting from the "$-\xi_0$" term in $(\xi - \xi_0)$ adds a bit of extra work that is easily avoided. By letting

$$\eta = \xi - \xi_0$$

then, since ξ_0 is a constant, $\dot{\eta} = \dot{\xi}$ and hence the linear approximation can be expressed simply as

$$\dot{\eta} = \left.\frac{\partial f}{\partial \xi}\right|_{\xi_0} \eta. \tag{14.14}$$

Clearly, the origin for η is the fixed point ξ_0, and the constant inhomogeneous term is eliminated by the simply coordinate transformation.

Example 14.3.7 Referring back to example 14.3.6, determine the *homogeneous* linear approximation to

$$\ddot{x} + \dot{x} - 2x + x^3 = 0$$

about the fixed point

$$\xi_0 = \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}.$$

Letting

$$\eta = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$$

and using equation 14.14, then

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -4 & -1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

■

14.4 Geometry and Stability of Equilibrium Points in the Phase Plane

This section will first outline the procedure to solve systems of two first order, linear, homogeneous differential equation. This material will be a bit of a review of the methods from Chapter 6, but will be developed with the ultimate goal of gaining some insight of the relationship between the linear algebra, *i.e.*, eigenvalue and eigenvectors, and the nature and geometry of solutions of systems near equilibrium points.

Examples 14.2.1 and 14.2.3 from Section 14.2 determined linear approximations to the nonlinear Duffing equation and solved them. In the case of example 14.2.1, the solution to the linear approximation near the point $x_0 = \sqrt{2}$ is

$$x(t) = c_1 e^{-\frac{1}{2}t}\cos\frac{\sqrt{15}}{2}t + c_2 e^{-\frac{1}{2}t}\sin\frac{\sqrt{15}}{2}t + \sqrt{2},$$

and in the case of example 14.2.3, the solution to the linear approximation near the point $x_0 = 0$ is

$$x(t) = c_1 e^{t} + c_2 e^{-2t}. \tag{14.15}$$

Note that both of these solutions are easily differentiated. In particular, for the solution of the linearization about the origin, we can write

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{t} + c_2 \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-2t}, \tag{14.16}$$

where $x_1(t) = x(t)$ and $x_2(t) = \dot{x}(t) = \dot{x}_1(t)$ as usual. The second line is computed by simply differentiating the solution from equation 14.15.

Observing the form of equation 14.16, rather than computing it from the solution of the original scalar equation, it seems reasonable that an alternative solution method designed to determine the solution directly from the vector form of the equation would be reasonably useful.

In particular, note that equation 14.14 is a system of first order, homogeneous differential equations, which may be expressed in the form

$$\dot{\eta} = A\eta,$$

where $A \in \mathbb{R}^{2\times 2}$ and

$$A = \left.\frac{\partial f}{\partial \xi}\right|_{\xi_0}.$$

Referring to equation 14.16, it seems reasonable to simply assume a solution of the form

$$\eta(t) = \eta e^{\lambda t}$$

where $\eta \in \mathbb{R}^2$, *i.e.,* it is a vector. Note that $\dot{\eta} = \lambda \eta e^{\lambda t}$. Substituting this into the differential equation gives

$$\lambda \eta e^{\lambda t} = A\eta.$$

Rearranging gives

$$\begin{aligned} A\eta e^{\lambda t} - \lambda \eta e^{\lambda t} &= 0 \\ A\eta e^{\lambda t} - \lambda I \eta e^{\lambda t} &= 0 \\ (A - \lambda I)\,\eta e^{\lambda t} &= 0 \\ (A - \lambda I)\,\eta &= 0, \end{aligned} \tag{14.17}$$

where I is the 2×2 identity matrix. Canceling the $e^{\lambda t}$ terms is justified since it may never be zero.

Hence, solutions of

$$\dot{\eta} = A\eta$$

are of the form

$$\eta(t) = \eta e^{\lambda t}$$

where η and λ satisfy equation 14.17. It is not a coincidence that equation 14.17 is the equation for the eigenvalues and eigenvectors of the matrix A; in fact, one of the primary uses of eigenvalue and eigenvector computations is to solve systems of first order, linear, constant coefficient differential equations.

Recall, that the procedure is to compute the λ values that satisfy equation 14.17 by observing that the equation only has solutions for nonzero η if

$$\det\left(A - \lambda I\right) = 0.$$

Once the values for λ are determined, each value is substituted into equation 14.17 and the corresponding eigenvector, η is computed. Various procedures are necessary depending upon whether the eigenvalues are real or complex and whether or not they are repeated. A compete consideration of all these cases appears in Chapter 6, a summary of which is as follows.

Theorem 14.4.1 *For the linear, homogeneous, constant coefficient system of n first order ordinary differential equations*

$$\dot{\xi} = A\xi,$$

if λ_i are the eigenvalues of A and $\hat{\xi}^i$ are the corresponding eigenvectors, then the general solution $\xi(t)$ depends upon the nature of the eigenvalues as follows.

1. *If the eigenvalues are distinct, then*

$$\xi(t) = c_1\hat{\xi}^1 e^{\lambda_1 t} + c_1\hat{\xi}^2 e^{\lambda_2 t} + \cdots c_n\hat{\xi}^n e^{\lambda_n t}.$$

2. *If there are any complex eigenvalues, say λ_i and $\lambda_{i+1} = \overline{\lambda_i}$ with complex conjugate eigenvectors $\hat{\xi}^1$ and $\hat{\xi}^{i+1} = \overline{\hat{\xi}^i}$ respectively, then the two terms in the general solution corresponding to λ_i and λ_{i+1} satisfy*

$$\begin{aligned} & c_i\hat{\xi}^i e^{\lambda_i t} + c_{i+1}\hat{\xi}^{i+1} e^{\lambda_{i+1} t} = \\ & \quad k_1 e^{\mu t}\left(\mathbf{a}\cos\omega t - \mathbf{b}\sin\omega t\right) + k_2 e^{\mu t}\left(\mathbf{a}\sin\omega t + \mathbf{b}\cos\omega t\right) \end{aligned} \tag{14.18}$$

where $\lambda_i = \mu + i\omega$ and $\hat{\xi}^i = \mathbf{a} + i\mathbf{b}$. It is usually more convenient to have terms in the general solution to be in terms of the trigonometric functions instead of the complex exponentials; hence, it is preferable to replace the left hand side of equation 14.18 with the right hand side for the corresponding terms in the general solution.

3. *For any repeated eigenvalues, if λ_i is repeated m times, then there will be m solutions to*
$$\left(A-\lambda_i I\right)^m \hat{\xi} = 0. \tag{14.19}$$
Then for each of the m solutions to equation 14.19, a term of the form
$$\xi(t) = \left(\hat{\xi} + t\left(A-\lambda_i I\right)\hat{\xi} + \frac{t^2}{2!}\left(A-\lambda_i I\right)^2\hat{\xi} + \cdots + \frac{t^{m-1}}{(m-1)!}\left(A-\lambda_i I\right)^{m-1}\hat{\xi}\right)e^{\lambda_i t}$$
will appear in the general solution.

- real versus imaginary eigenvalues/vectors

Example 14.4.2 Consider the second order, nonlinear differential equation
$$\ddot{x} + \dot{x} - 2x + x^2 = 0. \tag{14.20}$$
Determine and solve the linear differential equations that approximate this nonlinear equation near all the equilibria and graphically compare the solutions to the linear approximation to the nonlinear equation.

1. To convert to a system of first order differential equations, let
$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \end{aligned}$$
which gives
$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_2 + 2x_1 - x_1^2 \end{bmatrix}. \tag{14.21}$$
2. The equilibrium points are where the right hand side of equation 14.21 is zero. In particular
$$\begin{bmatrix} x_2 \\ -x_2 + 2x_1 - x_1^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Longleftrightarrow \qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix}. \tag{14.22}$$
3. A Taylor series approximation of the nonlinear term in equation 14.21 about $x = x_0$ gives
$$\ddot{x} + \dot{x} - 2x + \left(x_0^2 + 2x_0\left(x - x_0\right)\right) = 0.$$
   (a) Near $x = 0$, the linear approximation is
$$\ddot{x} + \dot{x} - 2x = 0,$$
   which has a general solution
$$x(t) = c_1 e^t + c_2 e^{-2t}.$$

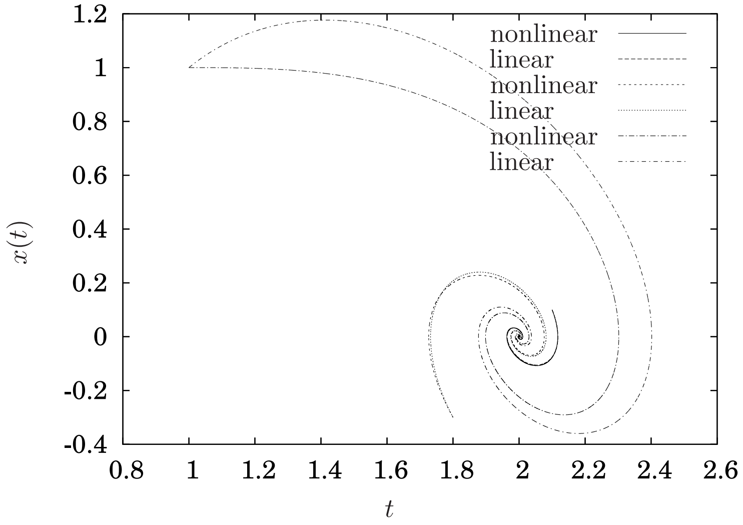


Figure 14.12. Comparison of solutions of nonlinear equation with solutions of the linear approximation near $x_0 = 2$ for equation 14.21.

(b) Near $x = 2$, the linear approximation is

$$\ddot{x} + \dot{x} + 2x = 4,$$

which has a general solution

$$x(t) = c_1 e^{-\frac{1}{2}t} \cos \frac{\sqrt{7}}{2} t + c_2 e^{-\frac{1}{2}t} \sin \frac{\sqrt{7}}{2} t + 2.$$

4. The best way to compare the validity of the solution to the linear approximation is to graphically compare the two solutions.

(a) A plot for three different sets of initial conditions of the solutions of the linear approximation about $x = 2$ and nonlinear equation is illustrated in Figure 14.12. Clearly, the farther the initial conditions are from the equilibrium point the less accurately the solution to the linear approximation approximates the solution to the nonlinear equation.

By inspecting Figure 14.12, it appears that the solution to the linear approximation is a good approximation to the nonlinear equation as long as the solution stays within approximately ±0.25 of the the point about which the system is linearized ($x = 2$).

(b) A plot for three different sets of initial conditions of the solutions of the linear approximation about $x =$ and nonlinear equation is

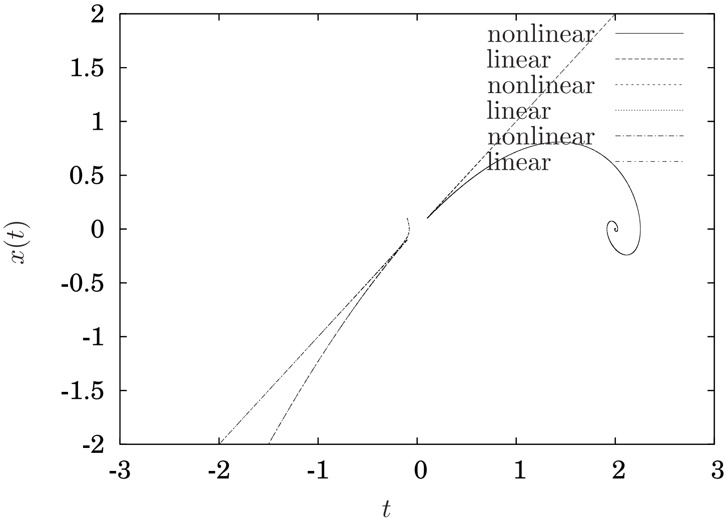


Figure 14.13. Comparison of solutions of nonlinear equation with solutions of the linear approximation near $x_0 = 0$ for equation 14.21.

illustrated in Figure 14.13. Clearly, the farther the initial conditions are from the equilibrium point the less accurately the solution to the linear approximation approximates the solution to the nonlinear equation.

By inspecting Figure 14.13, it appears that the solution to the linear approximation is a good approximation to the nonlinear equation as long as the solution stays within approximately ±0.25 of the the point about which the system is linearized ($x = 0$).

5. The Jacobian for equation 14.22 is

$$\frac{\partial f}{\partial \xi} = \begin{bmatrix} 0 & 1 \\ 2 - 2x_1 & -1 \end{bmatrix}.$$

(a) Near

$$\xi_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

the linear approximation is

$$\dot{\eta} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix} \eta.$$

(b) Near

$$\xi_0 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

the linear approximation is

$$\dot{\eta} = \begin{bmatrix} 0 & 1 \\ -2 & -1 \end{bmatrix} \eta.$$

6. By computing the eigenvalues and eigenvectors, the solutions to the linear approximations are computed as follows.

(a) For

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

the eigenvalues are $\lambda_{1,2} = -2, 1$ and the corresponding eigenvectors are

$$\hat{\xi}_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

and

$$\hat{\xi}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Hence, the general solution is

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c_1 \begin{bmatrix} -2 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{t}.$$

(b) For

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

the eigenvalues are

$$\lambda_{1,2} = -\frac{1}{2} \pm i\frac{\sqrt{7}}{2}$$

and the corresponding eigenvectors are

$$\hat{\xi}_1 = \begin{bmatrix} -\frac{1}{4} - i\frac{\sqrt{7}}{4} \\ 1 \end{bmatrix}$$

and

$$\hat{\xi}_1 = \begin{bmatrix} -\frac{1}{4} + i\frac{\sqrt{7}}{4} \\ 1 \end{bmatrix}$$

As is necessary, the eigenvectors are complex conjugates.

Let

$$\begin{aligned} u(t) &= e^{-\frac{1}{2}t}\left(\begin{bmatrix} -\frac{1}{4} \\ 1 \end{bmatrix}\cos\frac{\sqrt{7}}{2}t + \begin{bmatrix} \frac{\sqrt{7}}{4} \\ 0 \end{bmatrix}\sin\frac{\sqrt{7}}{2}t\right) \\ v(t) &= e^{-\frac{1}{2}t}\left(\begin{bmatrix} -\frac{1}{4} \\ 1 \end{bmatrix}\sin\frac{\sqrt{7}}{2}t - \begin{bmatrix} \frac{\sqrt{7}}{4} \\ 0 \end{bmatrix}\cos\frac{\sqrt{7}}{2}t\right) \end{aligned}$$

and the general solution may be written as

$$\eta(t) = c_1 u(t) + c_2 v(t).$$

Note: this form of the solution is *not* unique. Since eigenvectors may be arbitrarily scaled, it is possible to have an equivalent solution that looks very different. ■

14.5 Introduction to Bifurcation Analysis

This section will present a catalog of typical bifurcations by way of examples. It is not intended to be a complete exposition on the subject; for that, interested readers are referred to [21, 9]. The examples are of the simplest type; namely, first order and generally solvable.

14.5.1 Saddle-node bifurcations

Consider the first order differential equation

$$\dot{x} + x^2 - \mu = 0. \tag{14.23}$$

The equilibria for this equation are $x_0 = \sqrt{\mu}$ if $\mu \geq 0$. If $\mu < 0$ there are no equilibria. These equilibrium values are plotted as a function of μ in Figure 14.14.

About x_0 the linear approximation is

$$\dot{x} + \left(x_0^2 + 2x_0\left(x - x_0\right)\right) - \mu = 0.$$

Substituting $x_0 = \pm\sqrt{\mu}$ gives

$$\dot{x} \pm 2\sqrt{\mu}x = 2\mu,$$

which has the general solution

$$x(t) = ce^{\mp 2\sqrt{\mu}t} + \sqrt{\mu},$$

which is stable for $+\sqrt{\mu}$ and unstable for $-\sqrt{\mu}$. These are indicated in Figure 14.14 with a solid line for the stable upper branch and a dashed line for the unstable lower branch.

Solutions for various initial conditions and $\mu = -0.1$ are illustrated in Figure 14.15. Solutions for various initial conditions and $\mu = 0.5$ are illustrated in Figure 14.16. Note that solutions that start with $x(0) > -\sqrt{\mu}$ are attracted to the upper stable equilibrium. Solutions with $x(0) < -\sqrt{\mu}$ are unstable.

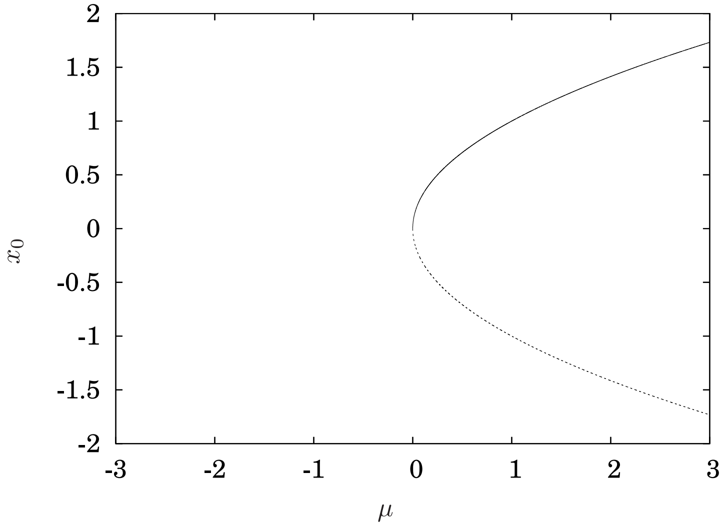


Figure 14.14. Equilibrium values for equation 14.23.

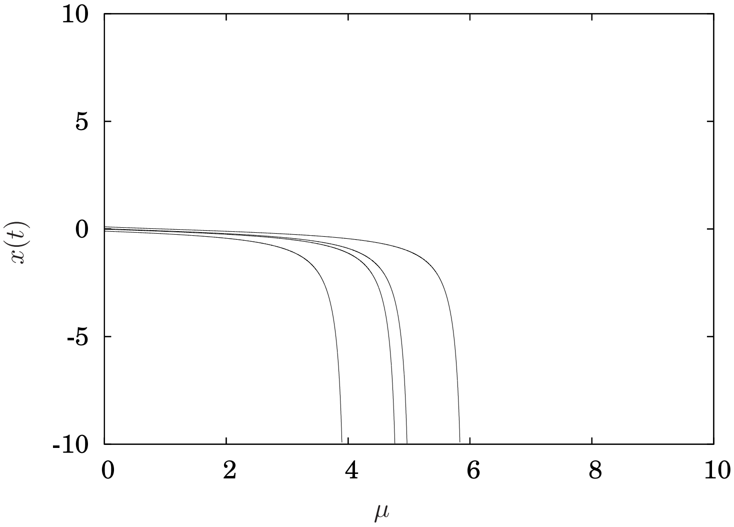


Figure 14.15. Solutions to equation 14.23 for $\mu = -0.1$ and for various initial conditions.

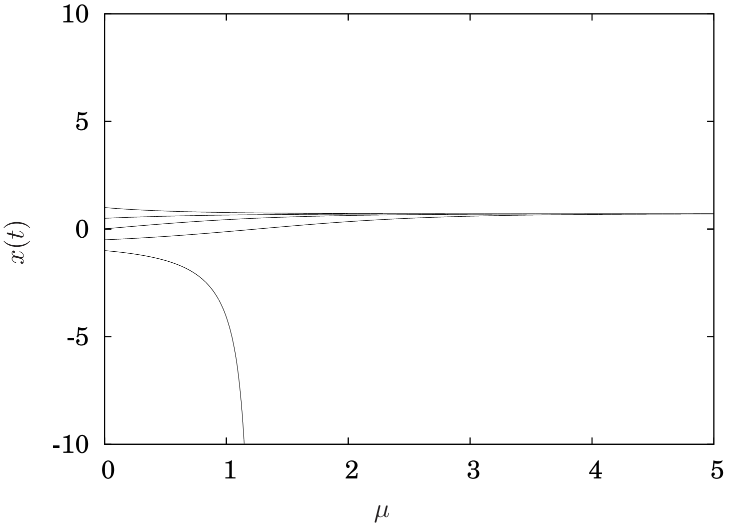


Figure 14.16. Solutions to equation 14.23 for $\mu = 0.5$ and for various initial conditions.

14.5.2 Pitchfork bifurcations

Example 14.5.1 Consider the first order differential equation

$$\dot{x} + x^3 - \mu x = 0. \tag{14.24}$$

This equation is actually separable, but the solution is not in a convenient form for analysis. Also, the solution is not even needed for present purposes.

The way we will proceed is to determine the equilibrium point(s) and compute a linear approximation about each one. For equation 14.24, the equilibrium points satisfy

$$x^3 - \mu x = 0 \qquad \Longleftrightarrow \qquad x = 0, \pm\sqrt{\mu}.$$

So, if $\mu \leq 0$ there is one equilibrium at $x = 0$, and if $\mu > 0$ there are three equilibria: $x = 0$, $x = \sqrt{\mu}$ and $x = -\sqrt{\mu}$. A plot of these equilibrium values *versus* μ is illustrated in Figure 14.17.

This type of bifurcation, for an obvious reason, is called a *pitchfork* bifurcation. The bifurcation aspect arises from the fact that as μ changes from negative to positive values, the number of equilibria changes from one to three.

Now consider the stability of these equilibria. For any value of μ the linear approximation about the $x_0 = 0$ equilibrium is

$$\dot{x} - \mu x = 0,$$

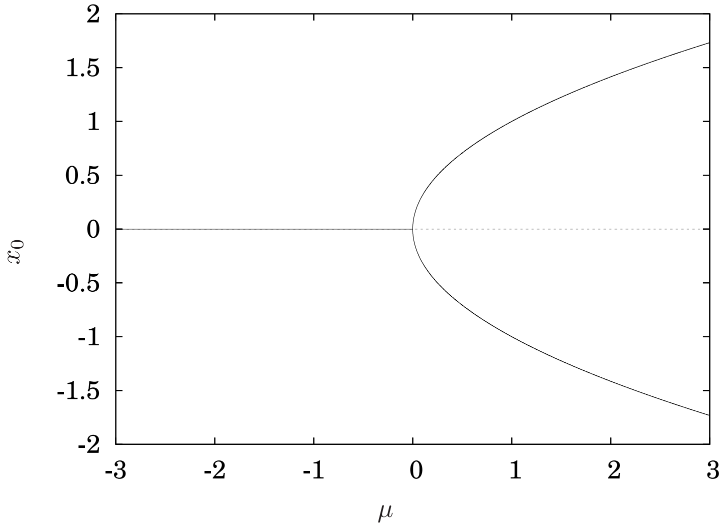


Figure 14.17. Equilibrium values for equation 14.24 *versus* μ.

which has solutions of the form

$$x(t) = ce^{\mu t}.$$

Clearly, for $\mu < 0$ these solutions are stable and for $\mu > 0$ these solutions are unstable. About $x_0 = \pm\sqrt{\mu}$ the linearization is

$$\begin{aligned} \dot{x} + \left(\pm\mu^{\frac{3}{2}} + 3\mu\left(x \mp \sqrt{\mu}\right)\right) - \mu x &= 0 \\ \dot{x} + 2\mu x &= \pm 2\mu^{\frac{3}{2}}. \end{aligned} \tag{14.25}$$

Note the solutions to equation 14.25 is

$$x(t) = ce^{-2\mu} \pm \frac{1}{2}\sqrt{\mu}$$

which is stable regardless of the sign of $\pm\sqrt{\mu}$. Hence, referring back to Figure 14.17, the branches of the equilibrium solutions which are stable are indicated by solid lines and the unstable branch is indicated by dashed lines. Observe that the stability of the $x_0 = 0$ equilibrium switches form stable to unstable as μ switches from negative to positive. The two outer branches for positive μ are stable.

Plots of solutions of equation 14.24 for $\mu = -0.5$ and for various initial conditions are illustrated in Figure 14.18. Note that, the $x_0 = 0$ equilibrium point is stable. Solutions for $\mu = 0.5$ are illustrated in Figure 14.19. Note that the $x_0 = 0$ equilibrium point is unstable; whereas, the $x_0 = \pm\frac{1}{\sqrt{2}}$ equilibria are stable. ■

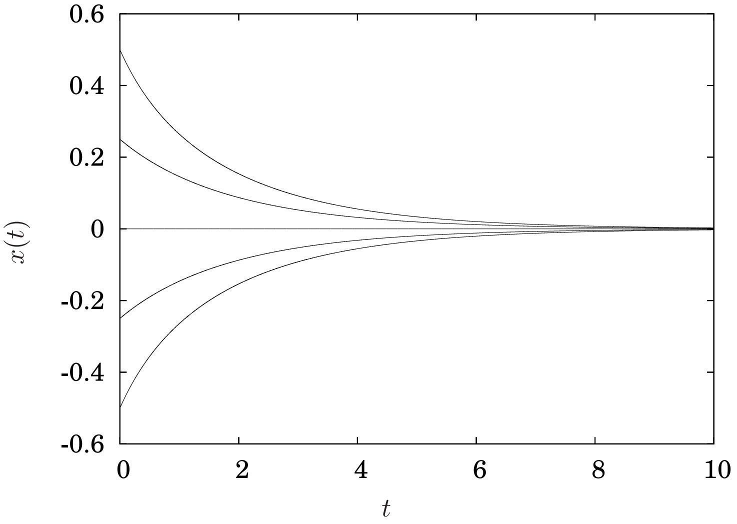


Figure 14.18. Solutions to equation 14.24 for $\mu = -0.5$ and for various initial conditions.

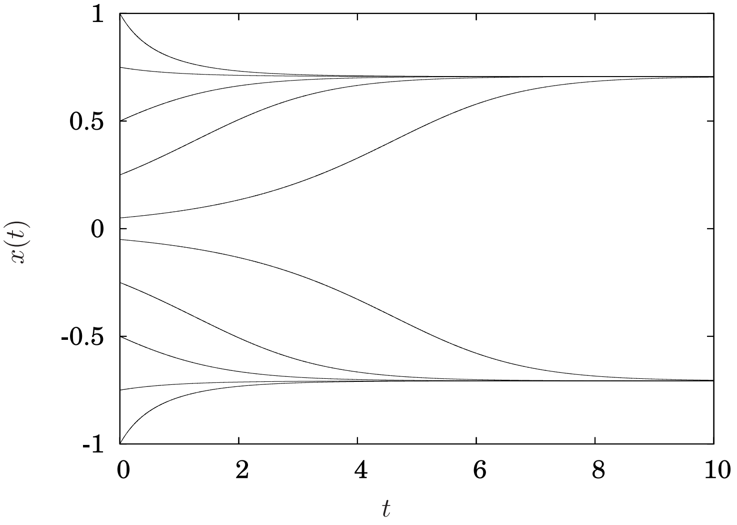


Figure 14.19. Solutions to equation 14.24 for $\mu = 0.5$ and for various initial conditions.

14.6 Exercises

Problem 14.1 Consider

$$\ddot{x} + x - x^3 = 0.$$

1. Write this as two first order ordinary differential equations.

2. Determine all the equilibrium points.

3. Using the Jacobian, determine the differential equation that is the best linear approximation about the equilibrium that is farthest to the right, *i.e.,* about the equilibrium point that has the largest value.

4. Determine the general solution to the linear approximation.

5. Sketch the phase portrait near the equilibrium point. Include the eigenvectors of the Jacobain matrix evaluated at the equilibrium point in the sketch.

Problem 14.2 Consider

$$\ddot{x} + 0.2\dot{x} - x + x^2 = 0.$$

1. Determine a linear differential equation that is the best linear approximation to Equation 14.20 near an arbitrary point $x = x_0$.

2. For $x_0 = 1$ determine the analytical solution to the linear approximation. Make a plot in the phase plane for comparing the solution to the linear approximation and the nonlinear equation for various initial conditions. Include plots of solutions with initial conditions for which the solution to the linear equation is near the solution to the nonlinear equation as well as plots where the solution to the linear equation is not near the solution to the nonlinear equation.

3. Near $x_0 = 0$, determine the solution to the linear approximation. Write it in the form of

$$\frac{d}{dt}\left[\begin{array}{c} x_1(t) \\ x_2(t) \end{array}\right] = c_1\hat{\xi}_1 e^{\lambda_1 t} + c_2\hat{\xi}_2 e^{\lambda_2 t}$$

where $\hat{\xi}_1$ and $\hat{\xi}_2$ are vectors. Plot the two vectors on the phase plane and also plot $-\hat{\xi}_1$ and $-\hat{\xi}_2$. Compare the solution to the linear approximation and the nonlinear approximation for initial conditions on either side of these vectors, *i.e.,* for initial conditions that are like the ×s illustrated in Figure 14.20.

   (a) Plot all the solutions where the range for the axes are $x \in [-.5, .5]$ and $\dot{x} \in [-.5, .5]$.

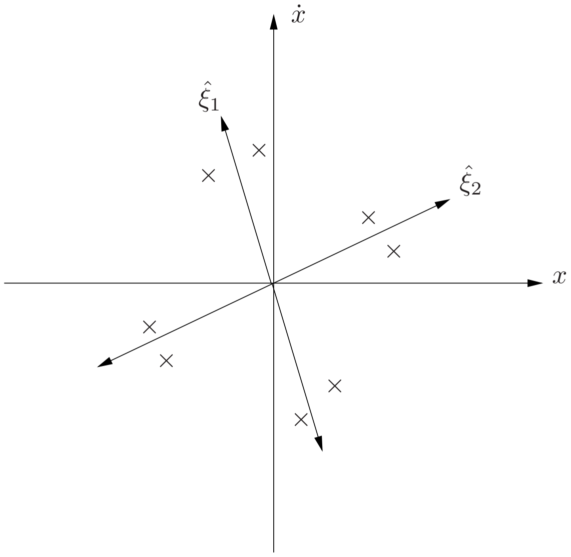


Figure 14.20. Initial conditions for Problem 14.2.

(b) Plot all the solutions where the range for the axes are $x \in [-2, 2]$ and $\dot{x} \in [2, 2]$.

Note: the solutions in this problem are unstable. You may need to adjust the total time for the numerical solution for the nonlinear equation so that the solution does not exceed the maximum value allowable for the simulation and plotting program.

Chapter 15

Perturbation Methods

15.1 Introduction

Based upon the results in Chapter 14, it is possible to obtain an approximation to the solution to a nonlinear differential equation provided that the solution remains sufficiently close to an equilibrium point. This chapter presents a method to determine an approximate solution to a nonlinear equation away from an equilibrium point. However, what is necessary in this chapter is that the nonlinear term in the differential equation must be "small." The approach will be motivated by an example.

Example 15.1.1 Consider

$$\begin{aligned} \ddot{x} + \dot{x} + x - \epsilon x^3 &= 0 \qquad (15.1) \\ x(0) &= 1 \\ \dot{x}(t) &= 0 \end{aligned}$$

where $\epsilon \ll 1$ is a *constant.*

The idea is that if ϵ is small, the it is reasonbale to attempt to compute a solution that is a series expansion in ϵ, *i.e.,*

$$x(t) = x_0(t) + \epsilon x_1(t) + \epsilon^2 x_2(t) + \epsilon^3 x_3(t) + \cdots \qquad \blacksquare$$

where each of the functions, $x_0(t)$, $x_1(t)$, $x_2(t)$, *etc.,* must be computed. To determine these, substitute the assumed form of the solution into Equation 15.1 noting that

$$\begin{aligned} \dot{x}(t) &= \dot{x}_0(t) + \epsilon\dot{x}_1(t) + \epsilon^2\dot{x}_2(t) + \epsilon^3\dot{x}_3(t) + \cdots \\ \ddot{x}(t) &= \ddot{x}_0(t) + \epsilon\ddot{x}_1(t) + \epsilon^2\ddot{x}_2(t) + \epsilon^3\ddot{x}_3(t) + \cdots \end{aligned}$$

which gives

$$\begin{aligned} &\left(\ddot{x}_0(t) + \epsilon\ddot{x}_1(t) + \epsilon^2\ddot{x}_2(t) + \cdots\right) + \left(\dot{x}_0(t) + \epsilon\dot{x}_1(t) + \epsilon^2\dot{x}_2(t) + \cdots\right) \\ + \quad &\left(x_0(t) + \epsilon x_1(t) + \epsilon^2 x_2(t) + \cdots\right) - \epsilon\left(x_0(t) + \epsilon x_1(t) + \epsilon^2 x_2(t) + \cdots\right)^3 = 0. \end{aligned}$$

Chapter 16

Lagrange's Equations

Example 16.0.2 Consider a particle constrained to move along a frictionless wire where the shape of the wire is given by the function $y = f(x)$, illustrated in Figure 16.1. This was considered previously in Example 1.9.9.

The position of the particle is given by

$$\mathbf{x} = \left[\begin{array}{c} x \\ f(x) \end{array} \right]$$

and the velocity is give by

$$\dot{\mathbf{x}} = \left[\begin{array}{c} \dot{x} \\ \frac{df}{dx}\dot{x} \end{array} \right].$$

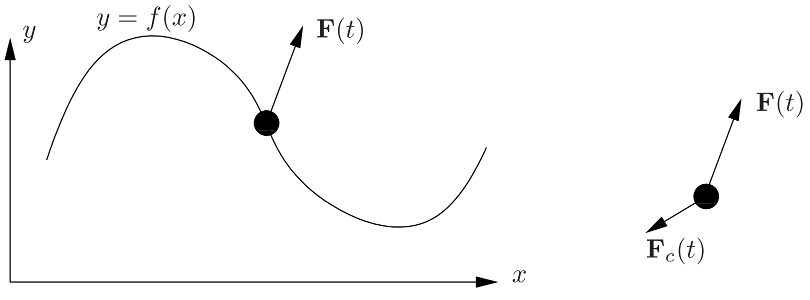


Figure 16.1. System for Example 16.0.2.

Hence, the kinetic energy is

$$\begin{aligned} T &= \frac{1}{2}m\dot{\mathbf{x}}\cdot\dot{\mathbf{x}} \\ &= \frac{1}{2}m\left(\dot{x}^2 + \left(\frac{df}{dx}\right)^2\dot{x}^2\right) \\ &= \frac{1}{2}m\dot{x}^2\left(1 + \left(\frac{df}{dx}\right)^2\right). \end{aligned}$$

Since there is no potential energy, $L = T$. The virtual work done by the applied force is

$$\begin{aligned} Q &= \mathbf{F}\cdot\frac{d\mathbf{x}}{dx} \\ &= F_x + \frac{df}{dx}F_y. \end{aligned}$$

Considering each term in Lagrange's equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = Q$$

we have:

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) &= \frac{d}{dt}\left(m\dot{x}\left(1 + \left(\frac{df}{dx}\right)^2\right)\right) \\ &= m\ddot{x}\left(1 + \left(\frac{df}{dx}\right)^2\right) + m\dot{x}\left(2\frac{df}{dx}\frac{d^2f}{dx^2}\dot{x}\right) \\ &= m\ddot{x}\left(1 + \left(\frac{df}{dx}\right)^2\right) + 2m\dot{x}^2\frac{df}{dx}\frac{d^2f}{dx^2} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial L}{\partial x} &= \frac{1}{2}m\dot{x}^2\left(\frac{d}{dx}\left(\frac{df}{dx}\right)^2\right) \\ &= \frac{1}{2}m\dot{x}^2\left(2\frac{df}{dx}\frac{d^2f}{dx^2}\right) \end{aligned}$$

$$m\dot{x}^2\frac{df}{dx}\frac{d^2f}{dx^2}.$$

So, substituting into Lagrange's equation gives

$$m\ddot{x}\left(1 + \left(\frac{df}{dx}\right)^2\right) + 2m\dot{x}^2\frac{df}{dx}\frac{d^2f}{dx^2} - m\dot{x}^2\frac{df}{dx}\frac{d^2f}{dx^2} = F_x + \frac{df}{dx}F_y,$$

or

$$m\ddot{x}\left(1+\left(\frac{df}{dx}\right)^2\right)+m\dot{x}^2\frac{df}{dx}\frac{d^2f}{dx^2}=F_x+\frac{df}{dx}F_y,$$

which is the same as Equation 1.19. ■

Chapter 17

System Identification

This chapter considers the problem of *system identification* which is the problem of determining the differential equation(s) governing a system based upon experimental data rather than first principles. In principle, it should always be possible to use first principles to determine the governing equations for a given system. However, in practice this is not always the case. First, many engineering systems may simply be too complicated to reduce to a collection of interconnected systems that can be individually modeled. Second, even if the components may be individually modeled, the interaction among them may not be. Finally, even if both of the above are possible, the approximations involved in modeling each individual component may combine in a manner that make the overall model a poor representation of the actual system. Hence, if it is the case that some data is available regarding how the system behaves, it makes sense to use that data t either validate the given model or as a basis for modeling the system.

17.1 The Damped Natural Frequency and Logarithmic Decrement

Consider the problem of modeling the system illustrated in Figure 17.1 where it is the case that the parameters for the model, m, k and b are not known, but what is known is that the system responds in a particular manner as illustrated in Figure 17.2.

Since the system is governed by the differential equation

$$m\ddot{x} + b\dot{x} + kx = 0, \tag{17.1}$$

at first it may seem like a simple matter to find m, b and k to give a response that looks like what is in the figure. In fact, attempting to do so by trial and error is not too difficult. However, since the system is simple enough, we may as well make the effort to at least be a bit more sophisticated about it in order

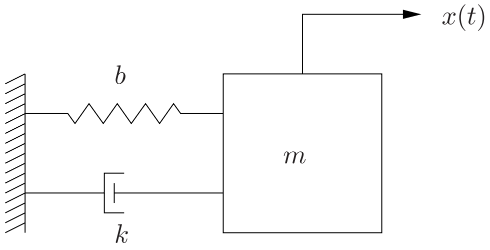


Figure 17.1. Mass–spring–damper system.

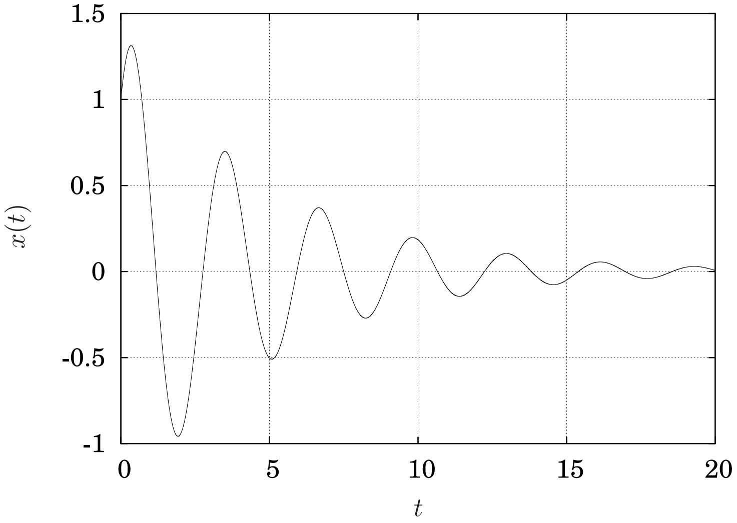


Figure 17.2. Response of a second order system.

to save the time involved in a trial and error method and to gain some insight into the problem at hand.

First, note that there will actually be an infinite number of sets of values for m, b and k that give the same response. This is because of the fact that if $x(t)$ satisfied equation 17.1 it will also satisfy a scaled version of the equation such as

$$\alpha m\ddot{x} + \alpha b\dot{x} + \alpha kx = 0,$$

or, in particular, it will also satisfy

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 0.$$

Since we may arbitrarily scale the equation without changing the solution, it seems reasonable to conclude that we may only find at most *two* of the three parameters. In fact this is the case, as will be outlined subsequently, and hence it makes sense to attempt to find the natural frequency, ω_n and the damping ratio, ζ which are the parameters in the canonical form of the second order linear oscillation equation. Of course, once these two parameters are determined, it will be possible to use their definitions to find all the possible combinations of m, b and k that are equivalent. Furthermore, if one of the parameters can be determined using an independent method, then the unique set of three parameters may be determined.

Recall that the solution to

$$\begin{aligned} \ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x &= 0 \\ x(0) &= x_0 \\ \dot{x}(0) &= \dot{x}_0 \end{aligned}$$

is given by equation 4.16 which is

$$\begin{aligned} x(t) &= e^{-\zeta\omega_n t}\left(c_1\cos\omega_n\sqrt{1-\zeta^2}t + c_2\sin\omega_n\sqrt{1-\zeta^2}t\right) \\ &= e^{-\zeta\omega_n t}\left(c_1\cos\omega_d t + c_2\sin\omega_d t\right). \end{aligned} \tag{17.2}$$

Inspecting equation 17.2 indicates that it should be straightforward to determine ω_d from simply inspecting the period of oscillation in the figure, as is illustrated in Figure 17.3. If the period is given by T, then the relationship between the frequency and period is simply $\omega_d T = 2\pi$, which gives

$$\omega_d = \frac{2\pi}{T}. \tag{17.3}$$

Another quantity that is easy to determine from the response of the system is the ratio of the magnitudes of two successive peaks in the response. Using equation 17.2 we have

$$\frac{x(t+T)}{x(t)} = \frac{e^{-\zeta\omega_n(t+T)}\left(c_1\cos\omega_d\left(t+T\right) + c_2\sin\omega_d\left(t+T\right)\right)}{e^{-\zeta\omega_n t}\left(c_1\cos\omega_d t + c_2\sin\omega_d t\right)}.$$

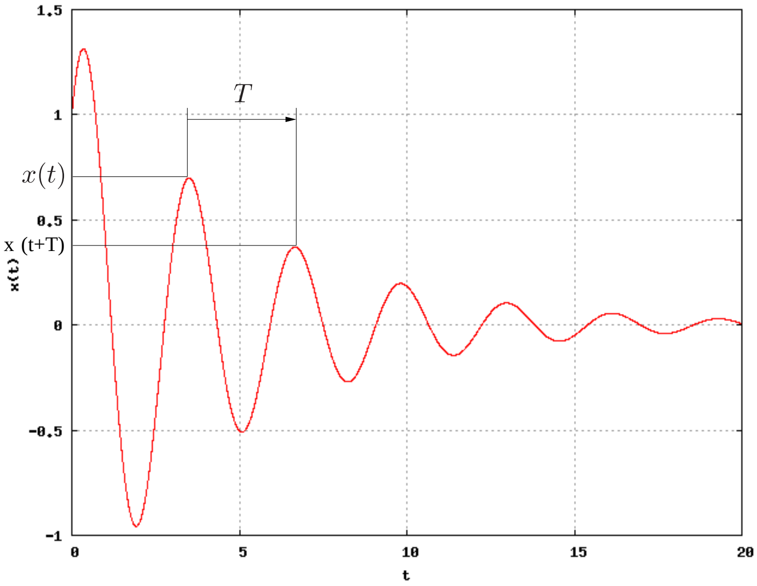


Figure 17.3. Response of a second order system.

However, since the period of oscillation is T, then the sine and cosine terms in the ratio are the same, so

$$\frac{x(t+T)}{x(t)} = \frac{e^{-\zeta\omega_n(t+T)}}{e^{-\zeta\omega_n t}} = e^{-\zeta\omega_n T} = e^{\frac{-\zeta\omega_d T}{\sqrt{1-\zeta^2}}} = e^{\frac{-2\pi\zeta}{\sqrt{1-\zeta^2}}}. \tag{17.4}$$

Hence, the ratio of the magnitude of two successive peaks is a function of the damping ratio only. Simply reading the values of two successive peaks, computing their ratio and then solving equation 17.4 for the damping ratio is all that is necessary. Observe that in the previous computations t was not specified; hence, it does not patter which peaks are used as long as they are successive peaks.

Since the study of linear oscillations is a classical subject, we will take it one step further to make the presentation consistent with the usual treatment. Taking the natural logarithm of both sides of equation 17.4 gives

$$\delta = \ln\left(\frac{x(t+T)}{x(t)}\right) = \ln x(t+T) - \ln x(t) = \frac{-2\pi\zeta}{\sqrt{1-\zeta^2}}.$$

This quantity, δ is called the *logarithmic decrement* and Figure 17.4 is a plot of the logarithmic decrement *versus* damping ratio.

Note for small ζ the logarithmic decrement is approximately linearly related to ζ and is given by

$$\delta \approx -2\pi z.$$

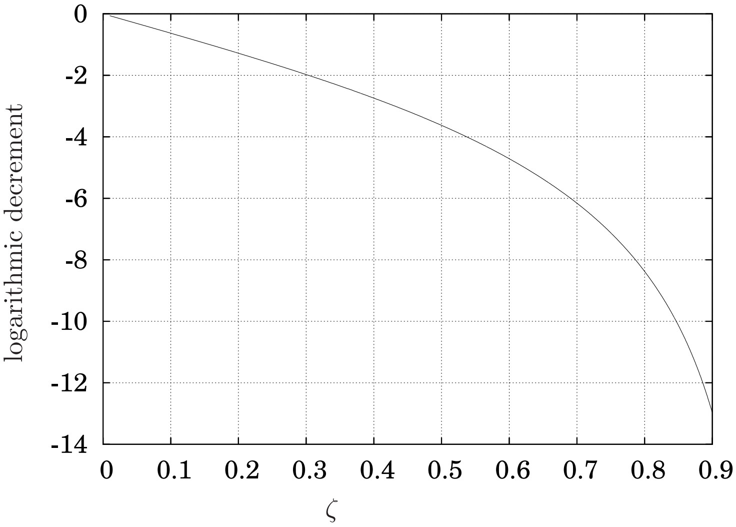


Figure 17.4. Plot of the logarithmic decrement *versus* damping ratio.

Example 17.1.1 Find the damping ratio and natural frequency for the response illustrated in Figure 17.2. Referring to the figure, $T \approx 3$. Hence, $\omega_d \approx \frac{2\pi}{3}$. Also the value of $x(t)$ at the second peak is approximately 0.7 and at the third peak it is approximately 0.4. Hence

$$\delta \approx \ln\left(\frac{0.4}{0.7}\right) = -.56. Since that$$

is a rather small value to use in Figure 17.4 (corresponding to a small value of ζ) so we will use the formula for the approximation for small ζ, which gives

$$\zeta \approx -\frac{\delta}{2\pi} = -\frac{-.56}{2\pi} = 0.089.$$

Using this value gives

$$\omega_n = \frac{\omega_d}{\sqrt{1-\zeta^2}} = 2.1.$$

In fact, the plot was generated using $\zeta = 0.1$ and $\omega_n = 2$, so the approximations involved in reading the values from the graphs and for the linear approximation for the relationship between the damping ratio and logarithmic decrement were really quite good. ■

17.2 Flexibility Influence Coefficients

Palm, page 494.

17.3 Exercises

Problem 17.1

The free response of a second order system is illustrated in Figure 17.5. Determine the natural frequency and the damping ratio.

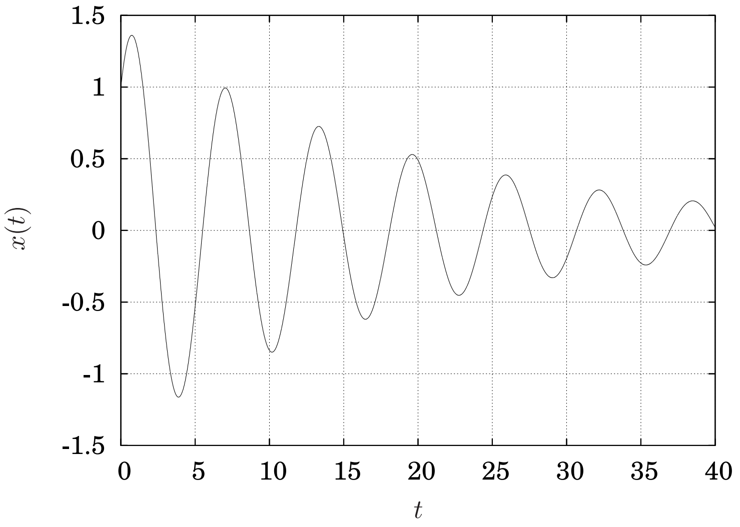


Figure 17.5. System response for Exercise 17.1.

Chapter 18

Symmetry and Transformation Methods

All the material from Chapter 6 related to diagonalization and Jordan canonical form is a special case of coordinate transformation methods. The basic idea is simple: find an alternative set of coordinates in which an equation is particularly simple (and hence, easy to solve). This chapter deals with the generalization of that idea and the manner in which differential equations act under coordinate transformations.

18.1 Coordinate Transformations

While the most general case is rather straightforward to write, it is sufficiently abstract that it is probably easiest to consider a few specific and simple cases before presenting the most general formulation.

18.1.1 Time and Length Scaling

Often it may prove convenient to simply re-scale length or time scales in a problem, which, in fact, are just a very simple subset of the general transformations considered previously.

Consider first, an ordinary, nth order, linear differential equation of the form

$$f_n(t)\frac{d^n x}{dt^n} + f_{n-1}(t)\frac{d^{n-1}x}{dt^{n-1}} + \cdots + f_1(t)\frac{dx}{dt} + f_0(t)x = g(t), \tag{18.1}$$

and consider the new variables τ and y given by

$$\begin{aligned}\tau &= st \\ y &= rx,\end{aligned}$$

or

$$\begin{aligned} t &= \frac{\tau}{s} \\ x &= \frac{y}{r}. \end{aligned}$$

The new coordinates are (y, τ) and the old coordinates are (x, t). Now consider how to express each of the terms in equation 18.1 in terms of y and τ.

Note that

$$y(\tau) = rx(\tau) = rx(st)$$

and

$$x(t) = \frac{y(t)}{r} = \frac{y\left(\frac{\tau}{s}\right)}{r},$$

so that if we know either $x(t)$ or $y(\tau)$ it is easy to compute the other.

Example 18.1.1 Consider the function

$$x(t) = \sin t$$

and consider the time and length scalings

$$\begin{aligned} y &= 5x \\ \tau &= 3t. \end{aligned}$$

Then

$$y(\tau) = rx(st) = 5\sin(3t)$$

The function is illustrated in Figure 18.1. Note that both $x(t)$ and $y(\tau)$ are the same shape, and, in fact, the difference between the two are simple scales of the two axes. ■

Since the differential equation involves derivatives of x with respect to t, converting equation 18.1 to the new coordinates also requires determining how derivatives transform.

Calculus gives

$$\begin{aligned} \frac{dx}{dt} &= \frac{1}{r}\frac{dy}{dt} = \frac{1}{r}\frac{dy}{d\tau}\frac{d\tau}{dt} = \frac{s}{r}\frac{dy}{d\tau} \qquad (18.2) \\ \frac{d^2x}{dt^2} &= \frac{1}{r}\frac{d^2y}{dt^2} = \frac{s}{r}\frac{d}{dt}\frac{dy}{dtau} = \frac{s}{r}\frac{d}{d\tau}\frac{dy}{d\tau}\frac{d\tau}{dt} = \frac{s^2}{r}\frac{d^2y}{d\tau^2} \\ &\vdots \\ \frac{d^nx}{dt^n} &= \frac{s^n}{r}\frac{d^ny}{d\tau^n}. \end{aligned}$$

The approach is first illustrated by means of an example.

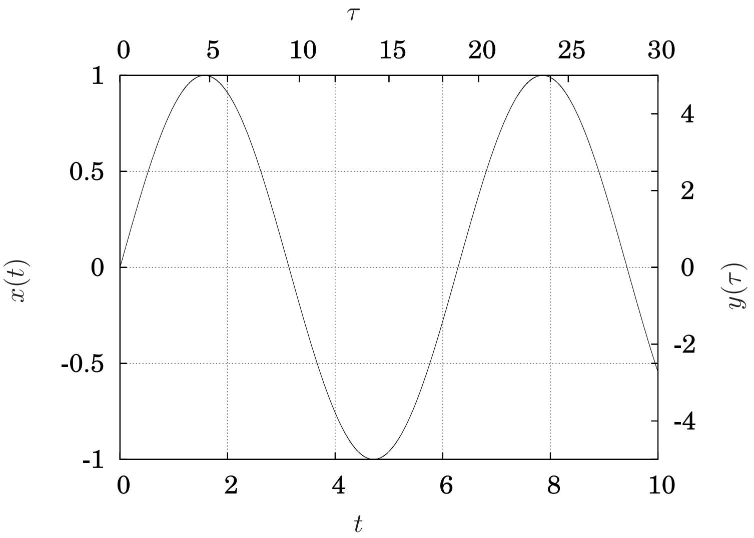


Figure 18.1. A function scaled by 3 in t and by 5 in x.

Example 18.1.2 Consider the initial value problem with an ordinary, second order, constant coefficient, linear, inhomogeneous differential equation of the form

$$\begin{aligned} \ddot{x} + 2\dot{x} + 4x &= 1 \\ x(0) &= 1 \\ \dot{x}(0) &= 2 \end{aligned} \tag{18.3}$$

If

$$\begin{aligned} \tau &= st \\ y &= rx \end{aligned} \tag{18.4}$$

then equation 18.3 expressed in the y and τ coordinates is simply determined by substitution,

$$\begin{aligned} s^2\frac{d^2y}{d\tau^2} + 2s\frac{dy}{d\tau} + 4y &= r \\ y(0) &= r \\ \left.\frac{dy}{d\tau}\right|_{\tau=0} &= 2\frac{r}{s}. \end{aligned} \tag{18.5}$$

Now, for equation 18.5 to represent a scaled version of equation 18.3, the two solutions must be related by *via* equation 18.4. Using, for example, the method of undetermined coefficients for ordinary, second order, linear, constant coefficient, inhomogeneous differential equations, the solution to the initial value problem in equation 18.3 is

$$x(t) = \frac{1}{4} + e^{-t}\left(\frac{3}{4}\cos\sqrt{3}t + \frac{11\sqrt{3}}{12}e^{-t}\sin\sqrt{3}t\right),$$

and the solution to the initial value problem in equation 18.5 is

$$y(\tau) = \frac{r}{4} + e^{-\frac{\tau}{s}}\left(\frac{3r}{4}\cos\frac{\sqrt{3}}{s}\tau + \frac{11\sqrt{3}r}{12}\sin\frac{\sqrt{3}r}{s}\tau\right).$$

Clearly, the relationship between the $y(\tau)$ solution and the $x(t)$ solution is exactly the scales given in equation 18.4. Hence, the differential equation and initial conditions transformed according to the scale rules in equation 18.2 has a solution that is the transform of the original solution. ■

18.1.2 Transformations of nth Order Scalar Equations

18.1.3 General Coordinate Transformations

First, consider the ordinary differential equations with independent variable t. Since any system of differential equations of order higher than one may be converted to a system of first order equations, without loss of generality, we can consider

$$\dot{\xi}(t) = f(\xi(t), t) \tag{18.6}$$

where $\xi \in \mathbb{R}^n$. Of course, "solving" this equation amounts to determining $\xi(t)$ and satisfying any given initial conditions.

Now, consider a coordinate transformation $\psi = \Psi(\xi) \in \mathbb{R}^n$, and $\tau = \phi(t) \in \mathbb{R}$, *i.e.,* ψ and τ are the new dependent and independent variables, respectively. Also, assume that these coordinate transformations are invertible. Particular examples will be presented shortly, but assume that these transformations are given and consider the problem of transforming the differential equation given in equation 18.6 to the new coordinates. By a simple application of the chain rule

$$\begin{aligned} \frac{d\psi}{d\tau} &= \frac{d\psi}{dt}\frac{dt}{d\tau} \\ &= \frac{d\Psi}{d\xi}\frac{d\xi}{dt}\frac{dt}{d\tau} \end{aligned}$$

Appendix A

Some Complex Variable Theory

This appendix presents a very short overview of complex variable theory. An interested reader is referred to [3] for a complete exposition.

A.1 Complex Numbers

Historically, of course, imaginary numbers have a natural association with the square root of negative numbers. We will develop the definitions of complex numbers in a more deductive manner and then show that the approach is consistent with the more historical view.

All readers should be familiar with the usual notion that a complex numer has a real and imaginary component, where we may write

$$s = \sigma + i\omega$$

where s is the complex number, σ is its real part and ω is the imaginary part . Since a complex number has two components, it may naturally be considered an ordered pair of numbers. The only twist is to ensure that we define multiplication correctly.

Definition A.1.1 A complex number is an ordered pair of real numbers,

$$s = (a, b)$$

where for

$$s_1 = (a_1, b_1)$$

and

$$s_2 = (a_2, b_2)$$

addition is defined by

$$s_1 + s_2 = (a_1 + a_2, b_1 + b_2)$$

and multiplication is defined by

$$s_1 s_2 = (a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1)\,. \qquad \diamond$$

This definition is consistent with the idea of using i because if we write

$$s_1 = a_1 + ib_1$$

and

$$s_2 = a_2 + ib_2$$

then

$$\begin{aligned} s_1 + s_2 &= a_1 + ib_1 + a_2 + ib_2 \\ &= (a_1 + a_2) + i\,(b_1 + b_2) \end{aligned}$$

and

$$\begin{aligned} s_1 s_2 &= (a_1 + ib_1)(a_2 + ib_2) \\ &= a_1 a_2 + ib_1 a_2 + a_1 ib_2 + ib_1 ib_2 \\ &= (a_1 a_2 - b_1 b_2) + i\,(a_1 b_2 + b_1 a_2)\,. \end{aligned}$$

It follows from the definition of addition and multiplication that the additive inverse of

$$s = a + ib$$

is

$$-s = -a - ib$$

and the multiplicative inverse is

$$s^{-1} = \left(\frac{a}{a^2 + b^2}, -\frac{b}{a^2 + b^2} \right).$$

Using the multiplicative inverse, division may be defined as

$$\frac{s_1}{s_2} = s_1 s_2^{-1}.$$

An alternative representation is in polar coordinates where s is represented by a magnitude and phase which are the usual Euclidean norm and angle if the number is plotted in its Cartesian coordinates. Referring to Figure A.1, clearly if $s = a + ib$, then

$$\begin{aligned} r &= \sqrt{a^2 + b^2} \\ &= |s| \end{aligned}$$

and

$$\begin{aligned} \theta &= \tan^{-1}\left(\frac{b}{a} \right) \\ &= \angle s. \end{aligned}$$

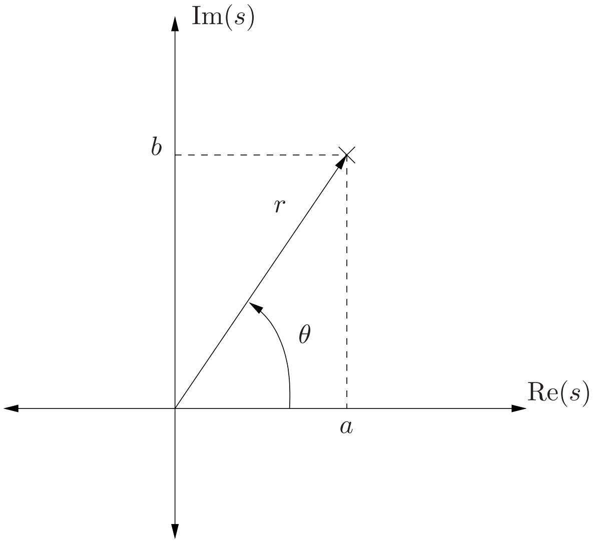


Figure A.1. Cartesian, $s = a + ib$ and polar, $s = (r, \theta)$ forms of a complex number, s.

The number (angle) θ is called an *argument* of the complex number s. There are an infinite number of arguments which differ by a multiple of 2π. The *principal value* of the arguments in the unique value $\theta \in (-\pi, \pi]$.

The previous two equations relate the Cartesian for to polar form. Going from polar for to Cartesian form is simple geometry and is given by

$$s = r(\cos\theta + i\sin\theta).$$

The Cartesian form is easy to use for addition and subtraction since if $s_1 = a_1 + ib_1$ and $s_2 = a_2 + ib_2$, then

$$s_1 + s_2 = (a_1 + a_2) + i(b_1 + b_2).$$

However, multiplication is easier in polar form. In particular, if $s_1 = (r_1, \theta_1)$ and $s_2 = (r_2, \theta_2)$, then the product is

$$s_1 s_2 = (r_1 r_2, \theta_1 + \theta_2)$$

and the quotient is

$$\frac{s_1}{s_2} = \left(\frac{r_1}{r_2}, \theta_1 - \theta_2\right).$$

This multiplication rule is easily seen by writing

$$s_1 = r_1(\cos\theta_1 + i\sin\theta_1)$$

and

$$s_2 = r_2 \left(\cos\theta_2 + i\sin\theta_2\right).$$

Taking the product

$$\begin{aligned} s_1 s_2 &= r_1 \left(\cos\theta_1 + i\sin\theta_1\right) r_2 \left(\cos\theta_2 + i\sin\theta_2\right) \\ &= r_1 r_2 \left[\cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2 + i\left(\sin\theta_1\cos\theta_2 + \sin\theta_2\cos\theta_1\right)\right] \\ &= r_1 r_2 \left[\cos\left(\theta_1+\theta_2\right) + i\sin\left(\theta_1+\theta_2\right)\right] \end{aligned}$$

so

$$s_1 s_2 = \left(r_1 r_2, \theta_1 + \theta_2\right).$$

A.2 Functions of a Complex Variable

The most important function of a complex variable is the exponential function due to the fact that exponentials are solutions to homogeneous, linear, constant coefficient, ordinary differential equations.

Definition A.2.1 If $s = a + ib$, define

$$e^s = e^a \left(\cos b + i\sin b\right). \qquad \diamond$$

Remark A.2.2 Note that this is a *definition*, which we choose to adopt. It remains to determine whether or not it is a useful definition or whether it reduces to the usual form when the imaginary part of s is zero. $\diamond$

Developing calculus for functions of a complex variable is beyond the scope of this book. However, if we have a complex-valued function of a real variable,

$$f(t) = f_r(t) + i f_i(t)$$

then is makes sense to define

$$\frac{df}{dt}(t) = \frac{df_r}{dt}(t) + i\frac{df_i}{dt}(t).$$

The property that we must verify for exponentials is

$$\begin{aligned} \frac{d}{dt}e^{st} &= \frac{d}{dt}e^{(a+ib)t} \\ &= \frac{d}{dt}\left(e^{at}\left(\cos bt + i\sin bt\right)\right) \\ &= ae^{at}\left(\cos bt + i\sin bt\right) + be^{at}\left(-\sin bt + i\cos bt\right) \\ &= ae^{at}\left(\cos bt + i\sin bt\right) + ibe^{at}\left(\cos bt + i\sin bt\right) \\ &= (a+ib)\left(e^{at}\left(\cos bt + i\sin bt\right)\right), \end{aligned}$$

so the usual rule for differentiating an exponential function holds.

A.3 Partial Fraction Decomposition

This subject is not limited to the field of complex variables, but since it appears in the process of solving for inverse Laplace transforms, it is included with the supplemental material, which is primarily complex-variable in nature.

The use of the partial fraction decomposition in this text is exclusively for the means of decomposing a rational function[1] into a linear combination of terms that appear in a Laplace transfor table (Table 8.1). If it is possible to do this, then it completely avoids the rather arduous exercise of evaluting the inverse Laplace transform, which is given by Definition 8.3.2.

We will use a partial fraction decomposition to reduce the degree of the polynomial appearing in the denominator of a rational function, which we will always assume to be proper.[2] Reducing the degree of the denominator is useful because, referring to Table 8.1, the distinguishing features of different elements of the table are the denominators of the functions.

The approach is to express a rational function in the form

$$\frac{N(s)}{D(s)} = \frac{N_1(s)}{D_1(s)} + \frac{N_2(s)}{D_2(s)} + \cdots + \frac{N_n(s)}{D_n(s)},$$

where the $D_i(s)$ satisfy

$$D(s) = D_1(s)D_2(s)\cdots D_n(s)$$

and are of a desired form, *i.e.,* for our purposes of the form of the denominator of elements in Table 8.1. The $N_i(s)$, then are simply polynomials in s that make the equality hold. First we will state a proposition that is helpful for computing the $N_i(s)$.

Proposition A.3.1 *If the function is proper, then the order each $N_i(s)$ will be less than the order of the corresponding $D_i(s)$.*

Proof If $P(s)$ is a polynomial in s, let $\mathcal{O}(P(s))$ denote the order of $P(s)$. Writing the decomposition and then putting it over a common denominator gives

$$\begin{aligned}\frac{N(s)}{D(s)} &= \frac{N_1(s)}{D_1(s)} + \frac{N_2(s)}{D_2(s)} + \cdots + \frac{N_n(s)}{D_n(s)} \\ &= \frac{(N_1(s)D_2(s)D_3(s)\cdots D_n(s)) + (N_2(s)D_1(s)D_3(s)\cdots D_n(s)) + \cdots}{D_1(s)D_2(s)\cdots D_n(s)} \\ &= \frac{\sum_{i=1}^{n}\prod_{j=1,j\neq i}^{n} N_i(s)D_j(s)}{\prod_{i=1}^{n} D_i(s)}. \qquad \text{(A.1)}\end{aligned}$$

At least one term in the sum in the numerator on the right in Equation A.1 must have the same order as $N(s)$. Since $\mathcal{O}(N(s)) < \mathcal{O}(D(s))$, then each term

[1] A *rational function*, is a function that may be written as a ratio of polynomials.

[2] A rational function is proper if the degree of the numerator is less than the degree of the denominator.

in the sum in the numerator on the left hand side of Equation A.1 has lower order than $D(s)$. Since

$$\mathcal{O}\left(P_1(s)P_2(s)\right) = \mathcal{O}\left(P_1(s)\right) + \mathcal{O}\left(P_2(s)\right)$$

then

$$\mathcal{O}\left(D(s)\right) = \sum_{j=1}^{n} \mathcal{O}\left(D_j(s)\right).$$

Then, for any $i \in \{1, \ldots n\}$,

$$\begin{aligned}
\mathcal{O}\left(D(s)\right) &> \mathcal{O}\left(N_i(s)D_1(s)D_2(s)\cdots D_{i-1}(s)D_{i+1}(s)\cdots D_n(s)\right) \\
&= \mathcal{O}\left(N_i(s)\right) + \mathcal{O}\left(D_1(s)\right) + \cdots \mathcal{O}\left(D_{i-1}(s)\right) + \\
&\quad \mathcal{O}\left(D_{i+1}(s)\right) + \cdots + \mathcal{O}\left(D_n(s)\right) \\
&= \mathcal{O}\left(N_i(s)\right) + \mathcal{O}\left(D(s)\right) - \mathcal{O}\left(D_i(s)\right).
\end{aligned}$$

Hence

$$\mathcal{O}\left(D_i(s)\right) > \mathcal{O}\left(N_i(s)\right).$$

□

The proof was a bit detailed, but what it tells us is that if we need to assume a form for the numerator of one of the fractions, $N_i(s)$ the largest its order can be is one less than the order of the denominator, $D_i(s)$.

Example A.3.2 Compute the partial fraction decomposition for

$$G(s) = \frac{s+1}{(s+2)(s+3)}$$

so the result is a linear combination of terms appearing in Table 8.1.

Since they correspond to an entry in Table 8.1, we will pick the two denominators to be

$$\begin{aligned}
D_1(s) &= s+2 \\
D_2(s) &= s+3.
\end{aligned}$$

Since both of these have order 1 in s, then each numerator must be of order 0, *i.e.,* a constant. Hence

$$\frac{s+1}{(s+2)(s+3)} = \frac{c_1}{s+2} + \frac{c_2}{s+3}.$$

The task now is to determine c_1 and c_2 so that the equality holds. We will present two ways to do this.

1. One way would be to put the right hand side over the common denominator and equate the resulting numerators, *i.e.,*

$$\begin{aligned}
\frac{s+1}{(s+2)(s+3)} &= \frac{c_1}{s+2} + \frac{c_2}{s+3} \\
&= \frac{c_1(s+3) + c_2(s+2)}{(s+2)(s+3)}.
\end{aligned}$$

Since the equality must hold, the numerators must be equal. So

$$s+1=c_1\left(s+3\right)+c_2\left(s+2\right).$$

Also, for this to hold for any s, the coefficient of each power of s must be equal so

$$s+1=\left(c_1+c_2\right)s+\left(3c_1+2c_2\right)$$

requires

$$\begin{aligned} c_1+c_2 &= 1 \\ 3c_1+2c_2 &= 1 \end{aligned}$$

which gives

$$\begin{aligned} c_1 &= -1 \\ c_2 &= 2. \end{aligned}$$

Hence,

$$\frac{s+1}{\left(s+2\right)\left(s+3\right)}=\frac{-1}{s+2}+\frac{2}{s+3}.$$

2. Another way to determine an equation to compute the numerators is to multiple each side of the expression by the denominator corresponding to the numerator we want to compute and then take the limit as s approches the value of the pole location for that term. So for

$$\frac{s+1}{\left(s+2\right)\left(s+3\right)}=\frac{c_1}{s+2}+\frac{c_2}{s+3}$$

to determine c_1, multiply both sides by $(s+2)$ and let $s=-2$, *i.e.*,

$$\frac{s+1}{\left(s+2\right)\left(s+3\right)}\left(s+2\right)=\frac{c_1}{s+2}\left(s+2\right)+\frac{c_2}{s+3}\left(s+2\right)$$

or

$$\frac{s+1}{s+3}=c_1+\frac{c_2}{s+3}\left(s+2\right).$$

Evaluating this as $s\to-2$, then

$$\frac{-2+1}{-2+3}=c_1+\frac{c_2}{-2+3}\left(-2+2\right).$$

Since the last term is zero,

$$c_1=-1.$$

Similarly,

$$\frac{s+1}{\left(s+2\right)\left(s+3\right)}\left(s+3\right)=\frac{c_1}{s+2}\left(s+3\right)+\frac{c_2}{s+3}\left(s+3\right)$$

or

$$\frac{s+1}{(s+2)} = \frac{c_1}{s+2}(s+3) + c_2.$$

Evaluating this as $s \to -3$ gives

$$c_2 = 2.$$

Hence

$$\frac{s+1}{(s+2)(s+3)} = \frac{-1}{s+2} + \frac{2}{s+3}.$$

■

Either approch works for complex conjugate poles as well.

Example A.3.3 Compute the partial fraction decomposition for

$$G(s) = \frac{1}{(s+2)(s^2+2s+2)}$$

so the result is a linear combination of terms appearing in Table 8.1. The roots for the second term in the denominator are $s = -1 \pm i$. We could factor it, but the form $(s+1)^2+1$ is what appears in Table 8.1. Hence we wish to determine c_1, c_2 and c_3 such that

$$\frac{1}{(s+2)(s^2+2s+2)} = \frac{c_1}{s+2} + \frac{c_2 s + c_3}{(s+1)^2+1}. \tag{A.2}$$

1. Combining the terms on the right hand side gives

$$\frac{1}{(s+2)(s^2+2s+2)} = \frac{c_1\left[(s+1)^2+1\right] + (c_2 s + c_3)(s+2)}{(s+2)(s^2+2s+2)},$$

and equating the numerators gives

$$\begin{aligned} 1 &= c_1\left(s^2+2s+2\right) + (c_2 s + c_3)(s+2) \\ &= (c_1+c_2)s^2 + (2c_1+2c_2+c_3)s + (2c_1+2c_3). \end{aligned}$$

Since this equality must hold for all s, the coefficients of each power of s must be equal so

$$\begin{aligned} c_1 + c_2 &= 0 \\ 2c_1 + 2c_2 + c_3 &= 0 \\ 2c_1 + 2c_3 &= 1, \end{aligned}$$

which gives

$$\begin{aligned} c_1 &= \frac{1}{2} \\ c_2 &= -\frac{1}{2} \\ c_3 &= 0. \end{aligned}$$

Hence,

$$\frac{1}{(s+2)\left(s^2+2s+2\right)} = \frac{1}{2\,(s+2)} - \frac{s}{2\left[(s+1)^2+1\right]}.$$

2. Alternatively, multiplying both sides of Equation A.2 by $s+2$ and computing the limit as $s \to -2$ gives

$$\begin{aligned} c_1 &= \lim_{s\to -2} \frac{1}{s^2+2s+2} \\ &= \frac{1}{2}. \end{aligned}$$

Similarly multiplying both sides of Equation A.2 by s^2+2s+2 and computing the limit as $s \to -1+i$ gives

$$\lim_{s\to -1+i} (c_2 s + c_3) = \lim_{s\to -1+i} \frac{1}{s+2}$$

which gives

$$\begin{aligned} c_2\,(-1+i) + c_3 &= \frac{1}{-1+i+2} \\ &= \frac{1}{1+i}\frac{1-i}{1-i} \\ &= \frac{1-i}{2}. \end{aligned}$$

Equating the real and imaginary parts gives

$$\begin{aligned} -c_2 + c_3 &= \frac{1}{2} \\ c_2 &= -\frac{1}{2}, \end{aligned}$$

and hence

$$c_3 = 0,$$

which is the same answer as before. ■

Appendix B

Linear Algebra Review

This appendix reviews some basic concepts from linear algebra. In particular, the definition of a linear vector space and transformations between them are considered.

B.1 Linear Vector Spaces

The most fundamental object in linear algebra is a *vector space.* A vector space is a generalization of the usual notion of a collection of vectors in Euclidean space and is useful to because such a generalized space will have all the properties that the set of vectors has. Instead of simply defining a vector space, let us present a list of those so-called useful properties and give examples of sets of objects other than vectors that also exhibit them or examples of objects that do not satisfy them.

B.1.1 Properties of vector operations in the Euclidean plane

While the notation will be abandoned subsequently, for this introductory section the common practice of denoting vectors with bold letters will be used. Also, to distinguish it from addition of real numbers, vector addition will be denoted bold plus sign. Also, while a vector space is fundamentally a set, the important properties that define it as a vector space are related to operations on these vectors, particularly, how they add and how they are scaled.

Specifically, define vector addition in the usual "head to tail" manner (illustrated in Figure B.1) and define

Property B.1.1 Vector addition is *commutative, i.e.,* for vectors $\mathbf{x_1}$ and $\mathbf{x_2}$,

$$\mathbf{x_1} \boldsymbol{+} \mathbf{x_2} = \mathbf{x_2} \boldsymbol{+} \mathbf{x_1}.$$

This property is illustrated in the usual way in Figure B.1. ⋄

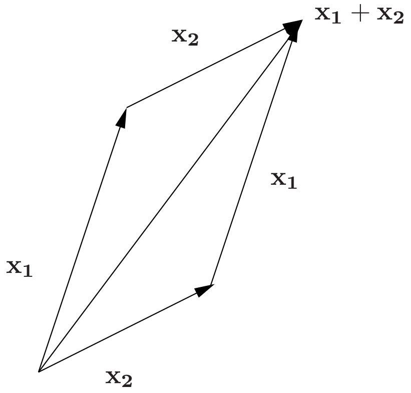


Figure B.1. Vector addition is commutative.

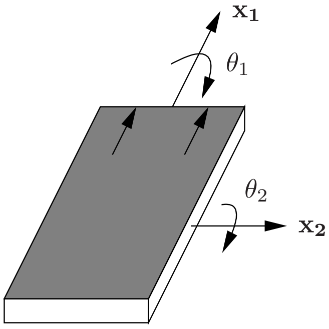


Figure B.2. Rotating a rigid body.

Example B.1.2 An example of an operation that is not commutative is rigid body rotations. Consider the book illustrated in Figure B.2 where the front cover is shaded and the top is indicated by arrows. If the book is first rotated about axis x_1 by an angle of 90° and then about axis x_2 by an amount 90°, the final orientation is illustrated in Figure B.3. Positive directions of rotation are given by the right hand rule. In contrast, if the body is rotated about axis x_2 by an amount 90° followed by a rotation about x_1 by an amount 90°, the final orientation is illustrated in Figure B.4.

If we use some sort of mathematical operation to represent these rotations, it may **not** commute, because one rotation followed by another rotation is, in general, not equal to the reverse order of rotations. While it is outside the scope of this text, it should not come as a surprise that rigid body rotations are represented by matrices, and one rotation follwed by an-

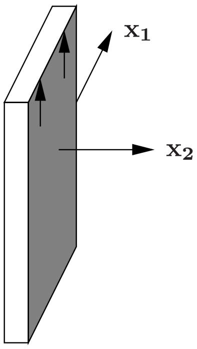


Figure B.3. Final book position after a rotation about $\mathbf{x_1}$ of 90° followed by a rotation about $\mathbf{x_2}$ by an amount 90°.

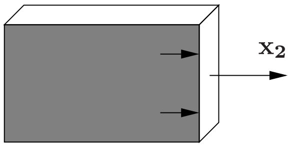


Figure B.4. Final book position after a rotation about $\mathbf{x_2}$ of 90° followed by a rotation about $\mathbf{x_1}$ by an amount 90°.

other is represented by matrix multiplication, which does not commute. An interested reader is referred to [5] for a complete exposition on rigid body kinematics and [16] for a more advanced treatment. ■

Property B.1.3 Vector addition is *associative*, *i.e.,* if $\mathbf{x_1}$, $\mathbf{x_2}$ and $\mathbf{x_3}$ are vectors, then

$$\mathbf{x_1} + (\mathbf{x_2} + \mathbf{x_3}) = (\mathbf{x_1} + \mathbf{x_2}) + \mathbf{x_3}. \qquad \diamond$$

An example of a nonassiciative operation is the cross product in $\mathbb{R}^3$.

Example B.1.4 Let $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ deonte the usual coordinate axes in $\mathbb{R}^3$. Then observing that

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}$$

and

$$\mathbf{i} \times \mathbf{i} = \mathbf{0}, \qquad ■$$

then

$$\begin{aligned} \mathbf{i} \times (\mathbf{i} \times \mathbf{j}) &= \mathbf{i} \times \mathbf{k} \\ &= -\mathbf{j}. \end{aligned}$$

However

$$\begin{aligned} (\mathbf{i} \times \mathbf{i}) \times \mathbf{j} &= \mathbf{0} \times \mathbf{k} \\ &= \mathbf{0}. \end{aligned}$$

Property B.1.5 Vector addition has an *identity element, i.e.,* there is a zero vector, $\mathbf{0}$ such that for any vector $\mathbf{x}$,

$$\mathbf{x} + \mathbf{0} = \mathbf{x}. \qquad \diamond$$

In the case of vectors in Euclidean space, the zero vector has no length.

Property B.1.6 Vector addition has an *inverse*, *i.e.,* for each vector $\mathbf{x}$, there exists another vector denoted by $-\mathbf{x}$ such that

$$\mathbf{x} + (-\mathbf{x}) = \mathbf{0}. \qquad \diamond$$

In the case of vectors in Euclidean space, the additive inverse of a vector is a vector with the same lenght, but with the opposite orientation.

Property B.1.7 Scalar multiplication distributes over vector addition. So, for vectors $\mathbf{x_1}$ and $\mathbf{x_2}$, and a real number α,

$$\alpha (\mathbf{x_1} + \mathbf{x_2}) = \alpha \mathbf{x_1} + \alpha \mathbf{x_2}. \qquad \diamond$$

Example B.1.8 Considering the two vectors in Figure B.1 again, if we double the sum, it is equal to doubling the length of each vector first, and then adding them. This is illustrated in Figure B.5. ■

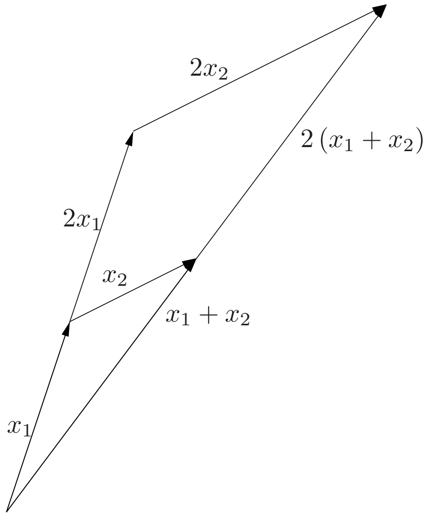


Figure B.5. Scalar multiplication distributes over vector addition.

Property B.1.9 Addition of two real numbers, a and b, distributes, *i.e.,*

$$(a + b)\,\mathbf{x} = a\mathbf{x} + b\mathbf{x}.$$

In other words, it does not matter if you add a and b first and then scale the vector, or if you multiply the vector individually by a and b and take the sum.$\diamond$

Example B.1.10 Considering the vector $\mathbf{x}$ illustrated in Figure B.6, and the scalars $a = 1.5$ and $b = 1.5$, it is the case that

$$(1.5 + 1.5)\,\mathbf{x} = 3\mathbf{x},$$

as is illustrated in Figure B.6. ■

Property B.1.11 Scalar multiplication of a vector is compatable with multiplication of real numbers, *i.e.,* for real numbers a and b

$$(ab)\,\mathbf{x} = a\,(b\mathbf{x})\,.$$

In other words, it does not matter if you multiply a and b together first and then scale the vector $\mathbf{x}$ or if you scale the vector by one of them followed by scaling by the other. $\diamond$

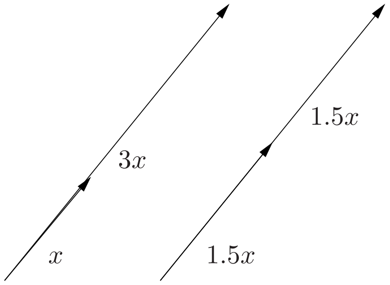


Figure B.6. Scalar addition distributes.

B.1.2 Definition and examples of vector spaces

A vector space is simply any set where you can add elements and scale them. To add some degree of generality, we will allow the vectors to be scaled by either real or complex numbers.

Definition B.1.12 Let the set[1] $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{C}$ and let V be a set with

1. a mapping $V \times V \to V$ called *vector addition* and denoted by $\mathbf{x_1} + \mathbf{x_2}$ for $\mathbf{x_1}$ and $\mathbf{x_2} \in V$; and,
2. a mapping $\mathbb{F} \times V \to V$ called *scalar multiplication* and denoted by $a\mathbf{x}$ for $a \in \mathbb{F}$ and $\mathbf{x} \in V$

where the mappings satisfy the following:

1. $\mathbf{x_1} + \mathbf{x_2} = \mathbf{x_2} + \mathbf{x_1}$;
2. $(\mathbf{x_1} + \mathbf{x_2}) + \mathbf{x_3} = \mathbf{x_1} + (\mathbf{x_2} + \mathbf{x_3})$;
3. $\exists \mathbf{0} \in V$ such that $\mathbf{0} + \mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in V$;
4. for each $\mathbf{x} \in V, \exists {-\mathbf{x}}$ such that $\mathbf{x}+(-\mathbf{x}) = \mathbf{0}$;
5. $(ab)\,\mathbf{x} = a\,(b\mathbf{x})$ for all $a, b \in \mathbb{F}$ and for all $\mathbf{x} \in V$;
6. $1\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in V$;
7. $0\mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in V$;
8. $a\,(\mathbf{x_1} + \mathbf{x_2}) = a\mathbf{x_1}+b\mathbf{x_2}$ for all $a \in \mathbb{F}$ and for all $\mathbf{x_1}, \mathbf{x_2} \in V$; and,
9. $(a + b)\,\mathbf{x} = a\mathbf{x}+b\mathbf{x}$ for all $a, b \in \mathbb{F}$ and for all $\mathbf{x} \in V$. ⋄

[1] The set $\mathbb{F}$ is generally a *field*, but for the purposes of this book it will always be the real or complex numbers.

Example B.1.13 Consider the set of polynomials of the independent variable t with real coefficients and degree less than or equal to n. Denote this set by $P(t,n)$. Any element of $P(t,n)$ my be expressed as

$$\alpha_n t^n + \alpha_{n-1} t^{n-1} + \cdots + \alpha_1 t + \alpha_0 \in P(t,n).$$

If addition and scalar multiplication are defined in the usual manner, *i.e.,*

$$\begin{aligned} &\left(\alpha_n t^n + \alpha_{n-1} t^{n-1} + \cdots + \alpha_1 t + \alpha_0\right) + \\ &\qquad \left(\beta_n t^n + \beta_{n-1} t^{n-1} + \cdots + \beta_1 t + \beta_0\right) \\ &= \quad \left(\alpha_n + \beta_n\right) t^n + \cdots + \left(\alpha_1 + \beta_1\right) t + \left(\alpha_0 + \beta_0\right) \end{aligned} \tag{B.1}$$

and

$$\beta \left(\alpha_n t^n + \alpha_{n-1} t^{n-1} + \cdots + \alpha_1 t + \alpha_0\right) = \left(\beta\alpha_n\right) t^n + \cdots + \left(\beta\alpha_1\right) t + \left(\beta\alpha_0\right) \tag{B.2}$$

then $P(t,n)$ is a vector space.

To actually *prove* this, we must verify each of the properties. This is generally a somewhat arduous exercise, but it is worth doing at least a few times.

1. For

   $$p_1 = \alpha_n t^n + \alpha_{n-1} t^{n-1} + \cdots + \alpha_1 t + \alpha_0$$

   and

   $$p_2 = \beta_n t^n + \beta_{n-1} t^{n-1} + \cdots + \beta_1 t + \beta_0$$

   we may write

   $$\begin{aligned} p_1 + p_2 &= \left(\alpha_n t^n + \cdots + \alpha_0\right) + \left(\beta_n t^n + \cdots + \beta_0\right) &\text{(B.3)} \\ &= \left(\alpha_n + \beta_n\right) t^n + \cdots + \left(\alpha_1 + \beta_1\right) t + \left(\alpha_0 + \beta_0\right) &\text{(B.4)} \\ &= \left(\beta_n + \alpha_n\right) t^n + \cdots + \left(\beta_1 + \alpha_1\right) t + \left(\beta_0 + \alpha_0\right) &\text{(B.5)} \\ &= \left(\beta_n t^n + \cdots + \beta_0\right) + \left(\alpha_n t^n + \cdots + \alpha_0\right) &\text{(B.6)} \\ &= p_2 + p_1. &\text{(B.7)} \end{aligned}$$

   These steps are justified as follows.

   (a) The step from Equation B.3 to B.4 is by the definition of vector addition in $P(t,n)$ given by Equation B.1.

   (b) The step from Equation B.4 to B.5 is justified because the coefficients in the polynomial are real and real addition commutes.

   (c) The step from Equation B.5 to B.6 is by the definition of addition in $P(t,n)$.

   Observe that, basically, addition of elements in $P(t,n)$ is defined in such a manner that the commutative property of addition of real numbers gave rise to the property that addition of two polynomials was also commutative.

2. For $p_1, p_2, p_3 \in P(t,n)$, where

$$\begin{aligned} p_1 &= \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \\ p_2 &= \beta_n t^n + \cdots + \beta_1 t + \beta_0 \\ p_3 &= \gamma_n t^n + \cdots + \gamma_1 t + \gamma_0 \end{aligned}$$

we have already that

$$(p_1 + p_2) = (\alpha_n + \beta_n)\, t^n + \cdots + (\alpha_1 + \beta_1)\, t + (\alpha_0 + \beta_0)$$

and

$$(p_2 + p_3) = (\beta_n + \gamma_n)\, t^n + \cdots + (\beta_1 + \gamma_1)\, t + (\beta_0 + \gamma_0)\, .$$

Hence,

$$\begin{aligned} (p_1 + p_2) + p_3 &= ((\alpha_n + \beta_n) + \gamma_n)\, t^n + \cdots + ((\alpha_0 + \beta_0) + \gamma_0) \\ &= (\alpha_n + (\beta_n + \gamma_n))\, t^n + \cdots + (\alpha_0 + (\beta_0 + \gamma_0)) \\ &= p_1 + (p_2 + p_3)\, . \end{aligned}$$

Again, the associative property of vector addition basically follows from the definition of addition and the fact that real number addition is associative.

3. Define the zero polynomial as

$$p_0 = 0t^n + \cdots + 0t + 0.$$

Then for any other

$$p = \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0$$

we have

$$\begin{aligned} p_0 + p &= (0 + \alpha_n)\, t^n + \cdots + (0 + \alpha_1)\, t + (0 + \alpha_0) \\ &= \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \\ &= p. \end{aligned}$$

4. Since for any $\alpha \in \mathbb{R}$ there exists $-\alpha \in \mathbb{R}$, then for any

$$p = \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \in P(t,n)$$

there exists a

$$(-p) = (-\alpha_n)\, t^n + \cdots + (-\alpha_1)\, t + (-\alpha_0) \in P(t,n)$$

such that

$$\begin{aligned} p+(-p) &= (\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0) + \\ &\quad ((-\alpha_n)\, t^n + \cdots + (-\alpha_1)\, t + (-\alpha_0)) \\ &= (\alpha_n - \alpha_n)\, t^n + \cdots + (\alpha_1 - \alpha_1)\, t + (\alpha_0 - \alpha_0) \\ &= 0t^n + \cdots + 0t + 0 \\ &= p_0. \end{aligned}$$

5. For

$$p = \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \in P(t,n)$$

we have

$$\begin{aligned} (ab)\, p &= (ab)\,(\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0) && \text{(B.8)} \\ &= ((ab)\,\alpha_n)\, t^n + \cdots + ((ab)\,\alpha_1)\, t + ((ab)\,\alpha_0) && \text{(B.9)} \\ &= (a\,(b\alpha_n))\, t^n + \cdots + (a\,(b\alpha_1))\, t + (a\,(b\alpha_0)) && \text{(B.10)} \\ &= a\,((b\alpha_n)\, t^n + \cdots + (b\alpha_1)\, t + (b\alpha_0)) && \text{(B.11)} \\ &= (a)\,(bp)\,. && \text{(B.12)} \end{aligned}$$

The justification for each step is as follows.

(a) The step from Equation B.8 to B.9 is the definition of scalar multiplication in $P(t,n)$ given by Equation B.2.

(b) The step from Equation B.9 to B.10 is justified because multiplication of real numbers is associative.

(c) The step from Equation B.10 to B.11 is the definition of scalar multiplication in $P(t,n)$.

6. For

$$p = \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \in P(t,n)$$

we have

$$\begin{aligned} 1p &= 1\,(\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0) \\ &= (1\alpha_n) + \cdots + (1\alpha_1)\, t + (1\alpha_0) \\ &= \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \\ &= p. \end{aligned}$$

7. For

$$p = \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \in P(t,n)$$

we have

$$\begin{aligned} 0p &= 0\,(\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0) \\ &= (0\alpha_n) + \cdots + (0\alpha_1)\, t + (0\alpha_0) \\ &= 0t^n + \cdots + 0t + 0 \\ &= \mathbf{0}. \end{aligned}$$

8. For

$$\begin{aligned} p_1 &= \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \\ p_2 &= \beta_n t^n + \cdots + \beta_1 t + \beta_0 \end{aligned}$$

we have

$$\begin{align} a\left(p_1 + p_2\right) &= a\left[\left(\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0\right) + \right. \nonumber \\ &\quad \left.\left(\beta_n t^n + \cdots + \beta_1 t + \beta_0\right)\right] \tag{B.13} \\ &= a\left[\left(\alpha_n + \beta_n\right) t^n + \cdots + \left(\alpha_0 + \beta_0\right)\right] \tag{B.14} \\ &= \left[a\left(\alpha_n + \beta_n\right) t^n + \cdots + a\left(\alpha_0 + \beta_0\right)\right] \tag{B.15} \\ &= \left[\left(a\alpha_n + a\beta_n\right) t^n + \cdots + \left(a\alpha_0 + a\beta_0\right)\right] \tag{B.16} \\ &= \left(\left(a\alpha_n\right) t^n + \cdots + \left(a\alpha_1\right) t + \left(a\alpha_0\right) + \right. \nonumber \\ &\quad \left.\left(a\beta_n\right) t^n + \cdots + \left(a\beta_1\right) t + \left(a\beta_0\right)\right) \tag{B.17} \\ &= ap_1 + ap_2. \tag{B.18} \end{align}$$

Each step is justified as follows.

(a) The step from Equation B.13 to B.14 is justified by the definition of addition in $P(t, n)$ given by Equation B.1.

(b) The step from Equation B.14 to B.15 is justified by the definition of scalar multiplication in $P(t, n)$ given by Equation B.2.

(c) The step from Equation B.15 to B.16 is justified by the fact that multiplication of real numbers distributes over addition of real numbers.

(d) The step from Equation B.16 to B.17 is justified by the definition of addition in $P(t, n)$.

9. For

$$p = \alpha_n t^n + \cdots + \alpha_1 t + \alpha_0 \in P(t, n)$$

and $a, b \in \mathbb{R}$, we have

$$\begin{align} (a + b)\, p &= (a + b)\left(\alpha_n t^n + \cdots + \alpha_1 t + \alpha_0\right) \tag{B.19} \\ &= \left((a + b)\,\alpha_n\right) t^n + \cdots + \left((a + b)\,\alpha_0\right) \tag{B.20} \\ &= \left(a\alpha_n + b\alpha_n\right) t^n + \cdots + \left(a\alpha_0 + b\alpha_0\right) \tag{B.21} \\ &= \left(a\alpha_n\right) t^n + \cdots + \left(a\alpha_0\right) + \nonumber \\ &\quad \left(b\alpha_n\right) t^n + \cdots + \left(b\alpha_0\right) \tag{B.22} \\ &= ap + bp. \tag{B.23} \end{align}$$

Each step is justified as follows.

(a) The step from Equation B.19 to B.20 is justified by the definition of scalar multiplication in $P(t, n)$ given by Equation B.2.

(b) The step from Equation B.20 to B.21 is justified by the fact that multiplication of real numbers distributes over addition of real numbers.

(c) The step from Equation B.21 to B.22 is justified by the definition of vector addition in $P(t, n)$ given by Equation B.1. ■

Before giving examples, we must discuss the *closure* property. The idea of closure is that you can not add to vectors in a vector space to create a vector not in it. Similarly, you can not scale a vector and leave the vector space.

Property B.1.14 COMPLETE! ◇

The remaining sections in this appendix were moved from Chapter 6 to here.

B.1.3 Linear independence

Consider the set of vectors $\left\{\xi^1, \ldots, \xi^k\right\} \in \mathbb{R}^n$, *i.e.*, k vectors that are n elements "tall" such as

$$\xi^i = \begin{bmatrix} \xi_1^i \\ \xi_2^i \\ \vdots \\ \xi_n^i \end{bmatrix}.$$

Definition B.1.15 (Linear (in)dependence) The set $\left\{\xi^1, \ldots, \xi^k\right\}$ is *linearly ~~in~~dependent* if $\exists$ scalars $\alpha_1, \ldots, \alpha_k$, where at least one $\alpha_i \neq 0$ such that

$$\alpha_1 \xi^1 + \alpha_2 \xi^2 + \cdots + \alpha_k \xi^k = \sum_{i=1}^{k} \alpha_i \xi^i = 0.$$

◇

If the set is ~~non~~ not linearly dependent, then it is *linearly independent.*

A simple example is in order.

Example B.1.16 Let $n = 3$ and

$$\xi^1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \xi^2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \xi^3 = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}.$$

Clearly, determining linear dependence or independence by inspection is not easy. So we try to solve

$$\alpha_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

or, as three scalar equations

$$\begin{aligned}\alpha_1+\alpha_2+5\alpha_3 &= 0\\ 2\alpha_1+\alpha_2+7\alpha_3 &= 0\\ 3\alpha_1+\alpha_2+9\alpha_3 &= 0.\end{aligned}$$

A tedious calculation gives

$$\begin{aligned}\alpha_1 &= 2\\ \alpha_2 &= 3\\ \alpha_3 &= 1,\end{aligned}$$

which determines that the set of vectors $\{\xi^1,\xi^2,\xi^3\}$ is linearly dependent.■

An easier approach is to recall the following basic result from linear algebra [7].

Proposition B.1.17 *If $A\in\mathbb{R}^{n\times n}$ and if $\det(A)=0$ then the set of vectors that are the columns of A are linearly dependent. Also, the set of vectors that are the rows of A are linearly dependent. If $\det(A)\neq 0$ then the columns and rows are linearly independent.*

Example B.1.18 Considering the system in Example B.1.16, an easy computation gives

$$\det\left(\begin{bmatrix}1&1&5\\2&1&7\\3&1&9\end{bmatrix}\right)=0$$

thus confirming the result from Example B.1.16 that the vectors are linearly dependent. ■

The primary utility of the notion of linear independence is that in a n dimensional vector space, a set of n linearly independent vectors, $\{\xi^1,\ldots,\xi^n\}$, form a *basis* for the vector space. Thus any vector in that space can be written as a linear combination, *i.e.*, $\xi=\sum_{i=1}^n\alpha_i\xi^i$.

Remark B.1.19 ~~Relationship with the Wronskian~~

Recall from Chapter 3 that we were concerned with linearly independent functions, and in particular used the notion of the Wronskian in Definition 3.2.6 to determine whether or not a set of functions was linearly independent. Analogous to the definition for vectors, a set of functions, $\{x_1(t),x_2(t),\ldots,x_n(t)\}$ is linearly dependent on an interval, $\mathcal{I}$, if there exists constants, $c_1,c_2,\ldots,c_n$, where not all of the constants are zero, such that

$$c_1x_1(t)+c_2x_2(t)+\cdots+c_nx_n(t)=0. \tag{B.24}$$

Differentiating Equation B.24 $n-1$ times gives the system of equations

$$\begin{array}{rcl} c_1 x_1(t) + c_2 x_2(t) + \cdots + c_n x_n(t) & = & 0 \\ c_1 \frac{dx_1}{dt}(t) + c_2 \frac{dx_2}{dt}(t) + \cdots + c_n \frac{dx_n}{dt}(t) & = & 0 \\ c_1 \frac{d^2 x_1}{dt^2}(t) + c_2 \frac{d^2 x_2}{dt^2}(t) + \cdots + c_n \frac{d^2 x_n}{dt^2}(t) & = & 0 \\ & \vdots & \\ c_1 \frac{d^{n-1} x_1}{dt^{n-1}}(t) + c_2 \frac{d^{n-1} x_2}{dt^{n-1}}(t) + \cdots + c_n \frac{d^{n-1} x_n}{dt^{n-1}}(t) & = & 0 \end{array}$$

which can be written in matrix form as

$$\begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \\ \frac{dx_1}{dt}(t) & \frac{dx_2}{dt}(t) & \cdots & \frac{dx_n}{dt}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{n-1}x_1}{dt^{n-1}}(t) & \frac{d^{n-1}x_2}{dt^{n-1}}(t) & \cdots & \frac{d^{n-1}x_n}{dt^{n-1}}(t) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \qquad \text{(B.25)}$$

A solution to Equation B.25 requires

$$\begin{vmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \\ \frac{dx_1}{dt}(t) & \frac{dx_2}{dt}(t) & \cdots & \frac{dx_n}{dt}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{n-1}x_1}{dt^{n-1}}(t) & \frac{d^{n-1}x_2}{dt^{n-1}}(t) & \cdots & \frac{d^{n-1}x_n}{dt^{n-1}}(t) \end{vmatrix} = 0.$$

◇

B.1.4 Eigenvalues and eigenvectors

Given a matrix $A \in \mathbb{R}^{n \times n}$ and a vector $\xi \in \mathbb{R}^n$, the product $y = A\xi$ is simply another vector in $\mathbb{R}^n$. However, there are two classes of the vectors x that give a special result when multiplied into A. The first special case is then the resulting vector is all zeros and the second special case is when the resulting vector is just a scaled version of x. The following two definitions elaborate upon this.

Definition B.1.20 (Null Space) The *null space* of a matrix $A \in \mathbb{R}^{n \times n}$, denoted by $\mathcal{N}(A)$, is the set of all vectors $\xi \in \mathbb{R}^n$ such that

$$A\xi = 0.$$

In this case 0 is the vector in $\mathbb{R}^n$ full of n zeros. ◇

Definition B.1.21 (Eigenvectors and Eigenvalues) An *eigenvector* of a matrix $A \in \mathbb{R}^{n \times n}$ is a non-zero vector, $\hat{\xi}$, such that

$$A\hat{\xi} = \lambda\hat{\xi}.$$

◇

The number λ, which may be real or complex, is the associated *eigenvalue*.

To compute eigenvalues and eigenvectors, note that

$$A\hat{\xi} = \lambda\hat{\xi} \qquad \Longrightarrow \qquad A\hat{\xi} - \lambda\hat{\xi} = (A - \lambda I)\,\hat{\xi} = 0, \tag{B.26}$$

where I is the $n \times n$ identity matrix. By Cramer's rule, Equation B.26 has a solution if and only if

$$\det(A - \lambda I) = 0. \tag{B.27}$$

Equation B.27 is an nth degree polynomial in λ ~~and hence has~~ with n solutions, and is called the *characteristic equation*. Thus, $A \in \mathbb{R}^{n\times n}$ has n eigenvalues. At this point, all we know is that there are n eigenvalues. Note that the eigenvalues may be all real and distinct, or some of them may be repeated and/or complex conjugate pairs.

To compute the eigenvalue associated with a particular eigenvalue λ, simply substitute the value for λ into Equation B.26 and solve for each component of $\hat{\xi}$. As the following example illustrates, the eigenvector can only be determined up to a unique scaling factor.

Example B.1.22 Compute the eigenvalues and eigenvectors of

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

First, to compute the eigenvalues,

$$\begin{aligned}
\det(A - \lambda I) &= \det\left(\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\
&= \det\left(\begin{bmatrix} 1-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix}\right) \\
&= (1-\lambda)(3-\lambda) - 2 \\
&= \lambda^2 - 4\lambda + 1 \\
&= 0.
\end{aligned}$$

Thus,

$$\lambda = 2 \pm \sqrt{3}.$$

To compute the eigenvectors, substituting the two values for λ into Equation B.27 gives

$$\left(A - \left(2+\sqrt{3}\right) I\right) = \begin{bmatrix} 1-2-\sqrt{3} & 2 \\ 1 & 3-2-\sqrt{3} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$$

which gives

$$\begin{aligned}
\left(-1-\sqrt{3}\right)\xi_1 + 2\xi_2 &= 0 \\
\xi_1 + \left(1-\sqrt{3}\right)\xi_2 &= 0.
\end{aligned}$$

A quick computation will show that if we try to solve for one variable, say ξ_2, from one of the equations and substitute into the other equation, we will end up with the degenerate equation $0 = 0$. This is precisely due to the fact that we are trying to solve a system of linearly dependent equations. Thus there are an infinite number of solutions.

The most straightforward approach may be to simply set one of the variables equal to one and solve for the others. So, in this example, arbitrarily let $\xi_2 = 1$. Both equations then give $\xi_1 = \sqrt{3}-1$, and hence the eigenvector corresponding to the eigenvalue $\lambda = 2 + \sqrt{3}$ is

$$\hat{\xi} = \begin{bmatrix} \sqrt{3}-1 \\ 1 \end{bmatrix}.$$

Note that **any** vector of the form

$$\hat{\xi} = \alpha \begin{bmatrix} \sqrt{3}-1 \\ 1 \end{bmatrix},$$

where α is a real or complex number is also an eigenvector corresponding to the eigenvalue $\lambda = 2 + \sqrt{3}$.

A similar computation (and again arbitrarily setting $\xi_2 = 1$) gives

$$\hat{\xi} = \begin{bmatrix} -\sqrt{3}-1 \\ 1 \end{bmatrix}$$

as an eigenvector corresponding to the eigenvalue $\lambda = 2 - \sqrt{3}$. ■

~~It will sometimes be the case that there is more than one linearly independent solution to the eigenvector problem. In that case it will be useful to have a more systematic approach to determining the linearly independent solutions.~~ In order to be more systematic in the approch to computing eigenvectors recall that ~~Recall~~ to solve a set of linear equations

$$Ax = b,$$

where $A \in \mathbb{R}^{n\times n}$, $b, x \in \mathbb{R}^n$ where A and b are given and x is to be determined, one approach is to construct the augmented matrix

$$\left[\begin{array}{c|c} A & b \end{array}\right]$$

and use row reduction operations to convert the left part of the augmented matrix to a convenient form (typically triangular form). In the case of determining eigenvectors, b will be a column of zeros, so the problem will be somewhat simpler. The details of the approach will be illustrated by the following example.

Example B.1.23 Determine the eigenvalues and eigenvectors of

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

We have

$$\begin{aligned}
\det\left(A-\lambda I\right) &= \left(1-\lambda\right)\begin{vmatrix} 2-\lambda & 0 & 0 \\ 0 & 1-\lambda & 1 \\ 0 & -1 & 3-\lambda \end{vmatrix} \\
&= \left(1-\lambda\right)\left(2-\lambda\right)\begin{vmatrix} 1-\lambda & 1 \\ -1 & 3-\lambda \end{vmatrix} \\
&= \left(1-\lambda\right)\left(2-\lambda\right)\left(\left(1-\lambda\right)\left(3-\lambda\right)+1\right) \\
&= \left(1-\lambda\right)\left(2-\lambda\right)\left(\lambda^2-4\lambda+4\right) \\
&= \left(1-\lambda\right)\left(2-\lambda\right)\left(\lambda^2-4\lambda+4\right) \\
&= \left(1-\lambda\right)\left(2-\lambda\right)^3 .
\end{aligned}$$

~~So, $\lambda = 1$ has an algebraic multiplicity of one and $\lambda = 2$ has an algebraic multiplicity of three.~~ So, $\lambda = 1$ is an eigenvalue and $\lambda = 2$ is an eigenvalue that is repeated three times.

Note that, in general, for matrices larger than two by two we will not be able to do such computations by hand. It was only due to the particular structure of the way the zeros were arranged in A that allowed us to do it in this example. ~~For larger matrices,~~ In general, for matrices larger than 2×2 using a computer program or calculator will be necessary to compute the eigenvalues, which are the roots of the characteristic equation.

Now, to compute the eigenvectors, substituting $\lambda = 1$ into $\left(A-\lambda I\right)\hat{\xi}^1 = 0$ gives

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} \hat{\xi}_1^1 \\ \xi_2^1 \\ \xi_3^1 \\ \xi_4^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

which, in augmented matrix form is

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & -1 & 2 & 0 \end{array}\right].$$

Interchanging the first and fourth rows gives

$$\left[\begin{array}{cccc|c} -1 & 0 & -1 & 2 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

and subtracting the first row from the second and third rows gives

$$\left[\begin{array}{cccc|c} -1 & 0 & -1 & 2 & 0 \\ 0 & 1 & 1 & -2 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

which is in upper triangular form. If we choose $\hat{\xi}_4^1 = 1$, then from the third row we have $\hat{\xi}_3^1 = 1$. Substituting both of these values into the second row gives $\hat{\xi}_2^1 = 1$ and finally the first row gives $\hat{\xi}_1^1 = 1$. Hence

$$\hat{\xi}^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Now, for $\lambda = 2$ we have $\det(A - 2I)\,\hat{\xi} = 0$ as

$$\begin{bmatrix} -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 \\ -1 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\xi}_1^2 \\ \hat{\xi}_2^2 \\ \hat{\xi}_3^2 \\ \hat{\xi}_4^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

or, in augmented matrix form

$$\left[\begin{array}{cccc|c} -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 & 0 \\ -1 & 0 & -1 & 1 & 0 \end{array}\right]$$

and, without elaborating on all the details, the row reductions gives

$$\begin{aligned} \left[\begin{array}{cccc|c} -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 & 0 \\ -1 & 0 & -1 & 1 & 0 \end{array}\right] &\Longleftrightarrow \left[\begin{array}{cccc|c} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \\ &\Longleftrightarrow \left[\begin{array}{cccc|c} -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \\ &\Longleftrightarrow \left[\begin{array}{cccc|c} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]. \end{aligned}$$

The procedure we will adopt is the following.

1. Inspecting each row in the reduced matrix, we will identify the variables as *not* free if they are the first nonzero term in any row. So, in the proceeding matrix, the components $\hat{\xi}_1$ and $\hat{\xi}_3$ are not free.

2. The remaining variables are free. Choose one of the free variables to be equal to one and the rest of the free variables to be equal to zero and compute the remaining components. This will give one eigenvector.

3. To compute another linearly independent eigenvector, choose another of the free variables to be one and the rest to be zero, and compute the remaining components. Continue through the entire list of free variables. This will result in a linearly independent set of eigenvectors.

Returning to the example, $\hat{\xi}_4^2$ and $\hat{\xi}_2^2$ are free. Choosing $\hat{\xi}_4^2 = 1$ and $\hat{\xi}_2^2 = 0$ gives

$$\hat{\xi}^2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

and choosing $\hat{\xi}_4^3 = 0$ and $\hat{\xi}_2^3 = 1$ gives

$$\hat{\xi}^3 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

■

B.2 Matrix Computations

It will be necessary to be able to compute matrix determinants and inverses. This section reviews how to do so. It is the way to do it by hand and is also the way that a computer program does it.

B.2.1 Computing Determinants

Define the determinant of a 2 × 2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

by

$$\det A = a_{11}a_{22} - a_{12}a_{21}.$$

For matrices larger than 2 × 2, we will use the following theorem is from [7].

Theorem B.2.1 *Let $A = (a_{ij})$ be an $n \times n$ matrix, where $n \geq 2$. Let A_{ij} be the $(n-1) \times (n-1)$ matrix formed by deleting row i and column j from A. Defining the cofactor*

$$c_{ij} = (-1)^{i+j} \det A_{ij}$$

we then have the expansion by rot i:

$$\det A = a_{i1}c_{i1} + a_{i2}c_{i2} + \cdots + a_{in}c_{in}$$

and we have the expansion by column j:

$$\det A = a_{1j}c_{1j} + a_{2j}c_{2j} + \cdots + a_{nj}c_{nj}.$$

The 3×3 case should probably be memorized.

Corollary B.2.2 *For*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

applying Theorem B.2.1 gives

$$\det A = a_{11}\left(a_{22}a_{33} - a_{23}a_{32}\right) - a_{12}\left(a_{21}a_{33} - a_{23}a_{31}\right) + a_{13}\left(a_{21}a_{32} - a_{22}a_{31}\right).$$

Example B.2.3 Compute the determinant of

$$A = \begin{bmatrix} -2 & 0 & 1 & -3 \\ -1 & -1 & 1 & -3 \\ 2 & -2 & -3 & -1 \\ 0 & 0 & 0 & -4 \end{bmatrix}.$$

■

Clearly, it is best to expand on a row or column with the most zeros. Expanding across the fourth row gives

$$\begin{aligned} \det A &= 0\left(-1\right)^{4+1}\det A_{41} + 0\left(-1\right)^{4+2}\det A_{42} + \\ &\quad 0\left(-1\right)^{4+3}\det A_{43} + -4\left(-1\right)^{4+4}\det A_{44} \\ &= -4\begin{vmatrix} -2 & 0 & 1 \\ -1 & -1 & 1 \\ 2 & -2 & -3 \end{vmatrix} \\ &= -4\left[-2\left(3+2\right) - 0\left(4-2\right) + 1\left(2+2\right)\right] \\ &= 24. \end{aligned}$$

B.2.2 Computing a Matrix Inverse

Let A^{-1} be the matrix From [7] is the following useful theorem to compute matrix inverses by hand.

Theorem B.2.4 *Let A be an $n \times n$ matrix, with $n > 1$. Let A_{ij} be the $(n-1) \times (n-1)$ matrix formed by deleting row i and column j from A. Define the cofactor matrix*

$$\operatorname{cof} A = C = \left[\left(-1\right)^{i+j}\det A_{ij}\right] \qquad \left(i, j = 1, \ldots, n\right).$$

Let $\Delta = \det A$. Then if $\Delta \neq 0$

$$A^{-1} = \frac{1}{\Delta}C^T.$$

Example B.2.5 Compute the inverse of

$$A = \begin{bmatrix} -2 & 0 & 1 & -3 \\ -1 & -1 & 1 & -3 \\ 2 & -2 & -3 & -1 \\ 0 & 0 & 0 & -4 \end{bmatrix}.$$

From Example B.2.3 we know that $\Delta = 24$. The terms in the cofactor matrix are

$$\begin{aligned} C_{11} &= (-1)^{1+1} \begin{vmatrix} -1 & 1 & -3 \\ -2 & -3 & -1 \\ 0 & 0 & -4 \end{vmatrix} \\ &= (-1)^{1+1} (-1)^{3+3} [-4 (3+2)] \\ &= -20, \end{aligned}$$

$$\begin{aligned} C_{12} &= (-1)^{1+2} \begin{vmatrix} -1 & 1-3 & \\ 2 & -3 & -1 \\ 0 & 0 & -4 \end{vmatrix} \\ &= -1 [-1 (12) - 1 (-8) - 3 (0)] \\ &= 4 \end{aligned}$$

$$\begin{aligned} C_{13} &= (-1)^{1+3} \begin{vmatrix} -1 & -1 & -3 \\ 2 & -2 & -1 \\ 0 & 0 & -4 \end{vmatrix} \\ &= (1) [-1 (8) - (-1) (-8) + 0] \\ &= -16 \end{aligned}$$

and

$$\begin{aligned} C_{14} &= (-1)^{1+4} \begin{vmatrix} -1 & -1 & 1 \\ 2 & -2 & 3 \\ 0 & 0 & 0 \end{vmatrix} \\ &= 0, \end{aligned}$$

and so on. Completing the tedious calculations gives

$$A^{-1} = \frac{1}{24} \begin{bmatrix} -20 & 4 & -16 & 0 \\ 8 & -16 & 16 & 0 \\ -4 & -4 & -8 & 0 \\ 22 & -14 & 26 & -6 \end{bmatrix}$$

■

Appendix C

Detailed Computations

This appendix contains some of the important, but detailed or cumbersome computations, inclusion of which in the main text perhaps would be distracting.

C.1 Computations Related to Fourier Series

Proposition C.1.1 *The integral*

$$\int_0^L \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{m\pi x}{L}\right)dx = \begin{cases} 0 & m \neq n \\ \frac{L}{2} & m = n \end{cases}$$

$m, n \in \mathbb{N}$.

Proof The case for $n \neq m$ is shown simply by integration by parts.

$$\begin{aligned}\int_0^L \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{m\pi x}{L}\right)dx &= -\frac{L}{m\pi}\sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right)\Big|_0^L + \\ &\quad \frac{n}{m}\int_0^L \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right)dx \\ &= \frac{n}{m}\int_0^L \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right)dx,\end{aligned}$$

which is clearly asking us to integrate by parts again. So

$$\begin{aligned}\frac{n}{m}\int_0^L \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right)dx &= \frac{n}{m}\,\frac{L}{n\pi}\cos\left(\frac{m\pi x}{L}\right)\sin\left(\frac{n\pi x}{L}\right)\Big|_0^L + \\ &\quad \left(\frac{n}{m}\right)^2\int_0^L \sin\left(\frac{m\pi x}{L}\right)\sin\left(\frac{n\pi x}{L}\right)dx.\end{aligned}$$

Thus

$$\int_0^L \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{m\pi x}{L}\right)dx = \left(\frac{n}{m}\right)^2\int_0^L \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{m\pi x}{L}\right)dx.$$

Hence, if $n \neq m$ the integral must be zero.

The case where $n = m$ is simply done by a trigonometric substitution. Recall that

$$\begin{aligned}\cos 2\theta &= \cos\theta\cos\theta - \sin\theta\sin\theta \\ &= \cos^2\theta - \sin^2\theta \\ &= \left(1 - \sin^2\theta\right) - \sin^2\theta \\ &= 1 - 2\sin^2\theta,\end{aligned}$$

so

$$\sin^2\theta = \frac{1-\cos 2\theta}{2}.$$

□

Hence

$$\begin{aligned}\int_0^L \sin^2\left(\frac{n\pi x}{L}\right) dx &= \frac{1}{2}\int_0^L 1 - \cos\left(\frac{2n\pi x}{L}\right) dx \\ &= \frac{x}{2}\Big|_0^L - \frac{L}{4n\pi}\sin\left(\frac{2n\pi x}{L}\right)\Big|_0^L \\ &= \frac{L}{2}.\end{aligned}$$

Note that $\mathbb{N}$ is the set of *natural numbers*, $\mathbb{N} = \{1, 2, 3, \ldots\}$.

Proposition C.1.2 *The integral*

$$\int_0^L \cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right) dx = \begin{cases} 0 & m \neq n \\ \frac{L}{2} & m = n \end{cases}$$

Proof This exactly mirrors the proof to Proposition C.1.1. □

Proposition C.1.3 *The integral*

$$\int_0^L \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{m\pi x}{L}\right) dx = 0$$

for all $m, n \in \mathbb{N}$.

C.2 Detailed Runge-Kutta Derivations

C.2.1 Third order Runge-Kutta method

The third order Runge-Kutta formula is derived by equating

$$\begin{aligned}x(t+\Delta t) &= x(t) + f(x(t),t)\Delta t + \frac{1}{2}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right)\Bigg|_{(x(t),t)}(\Delta t)^2 + \\ &\quad \frac{1}{6}\left[\left(\frac{\partial^2 f}{\partial x^2}f + \frac{\partial^2 f}{\partial x\partial t}\right)f + \frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial x}f + \frac{\partial f}{\partial t}\right) + \right. \\ &\quad \left.\frac{\partial^2 f}{\partial x\partial t}f + \frac{\partial^2 f}{\partial t^2}\right]\Bigg|_{(x(t),t)}(\Delta t)^3\end{aligned}$$

with

$$\begin{aligned} x(t + \Delta t) &= c_1 f + \\ &\quad c_2 f(x + c_3 f \Delta t, t + c_4 \Delta t) + \\ &\quad c_5 f(x + c_6 f + c_7 f(x + c_8 f \Delta t, t + c_9 \Delta t) \end{aligned} \tag{C.1}$$

(if no arguments to f are specified, it is evaluated at $(x(t), t)$).

To determine the coefficients, equation C.1 must be expanded to third order. Since

$$f(x + a, t + b) = f + \frac{\partial f}{\partial x} a + \frac{\partial f}{\partial t} b + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} a^2 + \frac{1}{2} \frac{\partial^2 f}{\partial t^2} b^2 + \frac{1}{2} \frac{\partial^2 f}{\partial x \partial t} ab + \cdots$$

where f and all the derivative terms are evaluated at (x, t), to third order equation C.1 is...

Appendix D

Example Programs

D.1 C Programs

D.1.1 Programs from Chapter 1

Program for example 1.10.1

```
/*  Example C program to determine an approximate solution to
 *
 *  x' = sin(2 t)
 *  x(0) = 3
 *
 *  using Euler's method.
 *
 *  To compile on the unix platfom, run "gcc example.c -lm" and then type
 *  "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  int n;
  float x,t,dt,f;
  FILE *fp;

  fp = fopen("eulerexample05.d","w");

  n = 0;
  dt = 0.5;
  x = 3.0;

  for(t=0;t<10;t+=dt) {
    f = sin(2.0*t);
    fprintf(fp,"%f \t %d \t %f \t %f \t ",t,n,x,f);
    x += f*dt;
    fprintf(fp,"%f \t %f\n",x,7.0/2.0 - cos(2.0*(t+dt))/2.0);
    n++;
  }
  fclose(fp);

  return 0;
}
```

Program for example 1.10.2

```
/*  Example C program to determine an approximate solution to
 *
 *  x' = 75 x (1 - x)
 *
 *  using Euler's method.
 *
 *  To compile on the unix platfom, run "gcc example.c -lm" and then type
 *  "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  int n;
  double x,t,dt,f;
  FILE *fp;

  fp = fopen("output.d","w");

  n = 0;
  dt = 0.1;
  x = 1.0/(1.0+exp(75.0));

  for(t=-1;t<1;t+=dt) {
    f = 75.0*x*(1-x);
    fprintf(fp,"%f\t%d\t%f\t%f\t%f\n ",t,n,x,f,x+f*dt);
    x += f*dt;
    n++;
  }
  fclose(fp);

  return 0;
}
```

Program for example 1.10.3

```
/*  Example C program to determine an approximate solution to
 *
 *  x’’ + sin(t) x’ + cos(t) x = exp(-5*t)
 *
 *  using Euler’s method.
 *
 *  To compile on the unix platfom, run "gcc example.c -lm" and then type
 *  "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  int n;
  float x[2],t,dt,f[2];
  FILE *fp;

  fp = fopen("output.d","w");

  n = 0;
  dt = 0.01;
  x[0] = 2;
  x[1] = 5;

  for(t=0;t<30;t+=dt) {
    f[0] = x[1];
    f[1] = exp(-5*t) - sin(t) * x[1] - cos(t)*x[0];
    fprintf(fp,"%f\t%d\t%f\t%f\t%f\t%f\t%f\t%f\n ",t,n,
    x[0],f[0],x[1],f[1],x[0]+f[0]*dt,x[1]+f[1]*dt);
    x[0] += f[0]*dt;
    x[1] += f[1]*dt;
    n++;
  }
  fclose(fp);

  return 0;
}
```

D.1.2 Programs from Chapter 2

D.1.3 Programs from Chapter 3

D.1.4 Programs from Chapter 13

Program for example 13.1.1

```
/*  This is from the file C/eulererroranalysis.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x' = 5 x
 *  x(0) = 1
 *
 *  using Euler's method.  The exact solution is also printed to the
 *  data file for comparsion purposes.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  float x,t,dt;
  FILE *fp;

  fp = fopen("data.d","w");

  dt = 0.1;
  x = 1.0;

  for(t=0;t<=1;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\n",t,x,exp(5.0*t));
    x += 5.0*x*dt;
  }
  fprintf(fp,"%f\t%f\t%f\n",t,x,exp(5.0*t));
  fclose(fp);

  return 0;
}
```

Program for example 13.1.2

```
/*  This is from the file C/eulererroranalysis2.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x' = -sin t
 *  x(0) = 1
 *
 *  using Euler's method.  The exact solution is also printed to the
 *  data file for comparsion purposes.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  float x,t,dt;
  FILE *fp;

  fp = fopen("eulererroranalysis2.d","w");

  dt = 1.0;
  x = 1.0;

  for(t=0;t<30;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\n",t,x,cos(t));
    x += -sin(t)*dt;
  }
  fprintf(fp,"%f\t%f\t%f\n",t,x,cos(t));
  fclose(fp);

  return 0;
}
```

Program for example 13.2.1

```
/*  This is from the file C/secondordertaylor.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x' = 5 x
 *  x(0) = 1
 *
 *  using a second order Taylor series expansion.  The exact solution
 *  is also printed to the data file for comparsion purposes.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  float x,t,dt;
  FILE *fp;

  fp = fopen("secondordertaylor.d","w");

  dt = 0.1;
  x = 1.0;

  for(t=0;t<=1;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\n",t,x,exp(5.0*t));
    x += 5.0*x*dt + 25.0/2.0*x*pow(dt,2);
  }
  fprintf(fp,"%f\t%f\t%f\n",t,x,exp(5.0*t));
  fclose(fp);

  return 0;
}
```

Program for example 13.2.2

```
/*  This is from the file C/secondordertaylorhard.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x' = -x^3 + sin(t x)
 *  x(0) = 1
 *
 *  using a second order Taylor series expansion.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  float x,t,dt;
  FILE *fp;

  fp = fopen("secondordertaylor2a.d","w");

  dt = 0.2;
  x = 1.0;

  for(t=0;t<5;t+=dt) {
    fprintf(fp,"%f\t%f\n",t,x);
    x += (-pow(x,3) + sin(t*x))*dt
      + 1.0/2.0*((-3*pow(x,2) + t*cos(t*x))*(-pow(x,3) + sin(t*x))
 + x*cos(t*x))*pow(dt,2);
  }
  fprintf(fp,"%f\t%f\n",t,x);
  fclose(fp);

  return 0;
}
```

Program for example 13.3.1

```
/*  This is from the file C/rk2.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x’ = -x^3 + sin(t x)
 *  x(0) = 1
 *
 *  using the second order Runge-Kutta (or improved Euler) method. The
 *  exact solution is also printed to the data file for comparsion
 *  purposes
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  float x,t,dt;
  FILE *fp;

  fp = fopen("rk2.d","w");

  dt = 0.1;
  x = 1.0;

  for(t=0;t<=1;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\n",t,x,exp(5.0*t));
    x += 5.0*x*dt;
  }
  fprintf(fp,"%f\t%f\t%f\n",t,x,exp(5.0*t));
  fclose(fp);

  return 0;
}
```

Program for example 13.3.2

```
/*  This is from the file C/rk2hard.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x’ = -x^3 + sin(t x)
 *  x(0) = 1
 *
 *  using the second Runge-Kutta method.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */


#include<stdio.h>
#include<math.h>

double f(double x, double t);

int main() {

  double x,t,dt;
  FILE *fp;

  fp = fopen("secondorderrk2a.d","w");

  dt = 0.2;
  x = 1.0;

  for(t=0;t<5;t+=dt) {
    fprintf(fp,"%f\t%f\n",t,x);
    x += dt/2*(f(x,t) + f(x+f(x,t)*dt,t+dt));
  }
  fprintf(fp,"%f\t%f\n",t,x);
  fclose(fp);

  return 0;
}

double f(double x, double t) {
  return -pow(x,3) + sin(t*x);
}
```

Program for example 13.3.3

```
/*  This is from the file C/rk3.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x’ = -x^3 + sin(t x)
 *  x(0) = 1
 *
 *  using the third order Runge-Kutta method.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */


#include<stdio.h>
#include<math.h>

double f(double x, double t);

int main() {

  double x,t,dt;
  double v1,v2,v3;
  FILE *fp;

  fp = fopen("rk3.d","w");

  dt = 0.5;
  x = 1.0;

  for(t=0;t<5;t+=dt) {
    fprintf(fp,"%f\t%f\n",t,x);
    v1 = f(x,t)*dt;
    v2 = f(x+v1/2.0,t+dt/2.0)*dt;
    v3 = f(x+2.0*v2-v1,t+dt)*dt;
    x += 1.0/6.0*(v1+4.0*v2+v3);
  }
  fprintf(fp,"%f\t%f\n",t,x);
  fclose(fp);

  return 0;
}
```

```
double f(double x, double t) {
  return -pow(x,3) + sin(t*x);
}
```

Program for example 13.3.4

```
/*  This is from the file C/rk4.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x’ = -x^3 + sin(t x)
 *  x(0) = 1
 *
 *  using the fourth order Runge-Kutta method.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */


#include<stdio.h>
#include<math.h>

double f(double x, double t);

int main() {

  double x,t,dt;
  double k1,k2,k3,k4;
  FILE *fp;

  fp = fopen("rk4a.d","w");

  dt = 0.25;
  x = 1.0;

  for(t=0;t<5;t+=dt) {
    fprintf(fp,"%f\t%f\n",t,x);
    k1 = f(x,t)*dt;
    k2 = f(x+k1/2.0,t+dt/2.0)*dt;
    k3 = f(x+k2/2.0,t+dt/2.0)*dt;
    k4 = f(x+k3,t+dt)*dt;

    x += 1.0/6.0*(k1 + 2.0*k2 + 2.0*k3 + k4);
  }
  fprintf(fp,"%f\t%f\n",t,x);
  fclose(fp);

  return 0;
}

double f(double x, double t) {
  return -pow(x,3) + sin(t*x);
}
```

Program for example 13.4.1

```
/*  This is from the file C/subtleerror.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x' + 3 x = 15(cos(3 t) + sin(3 t))
 *  x(0) = 1
 *
 *  using Euler's method, 2nd order RK, a 2nd order Taylor series
 *  expansion and 4th order RK.  The exact solution is also printed to
 *  the data file for comparsion purposes.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

double f(double x, double t);

main() {

  double xe,xie,t,dt=0.125;
  double xts,x4rk;
  double w1,w2,w3,w4;
  double t_final=5;
  double exact;
  FILE *fp;

  fp = fopen("subtledata2.d","w");
  xe = 0.0;                     /* euler's method */
  xie = 0.0;                    /* 2nd order RK */
  xts = 0.0;                    /* 2nd order TS */
  x4rk = 0.0;                   /* 4th order RK */
  exact = 0.0;

  for(t=0;t<=t_final;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\t%f\t%f\t%f\t%f\t%f\t%f\t%f\n",
    t,xe,xie,xts,x4rk,exact,exact-xe,exact-xie,exact-xts,exact-x4rk);

    xe += f(xe,t)*dt;
    xie += (f(xie,t)+f(xie+f(xie,t)*dt,t+dt))*dt/2.0;
    xts += f(xts,t)*dt + pow(dt,2)/2.0*
```

```
      (-3.0*f(xts,t)+45.0*(-sin(3.0*t) + cos(3.0*t)));
    w1 = f(x4rk,t)*dt;
    w2 = f(x4rk+w1/2.0,t+dt/2.0)*dt;
    w3 = f(x4rk+w2/2.0,t+dt/2.0)*dt;
    w4 = f(x4rk+w3,t+dt)*dt;
    x4rk += 1.0/6.0*(w1 + 2.0*w2 + 2.0*w3 + w4);
    exact = 5.0*sin(3.0*(t+dt));
  }
    fclose(fp);
}


double f(double x, double t) {
  return 15.0*(cos(3.0*t)+sin(3.0*t)) - 3.0*x;
}
```

Program for example 13.5.1

```
/*  This is from the file C/systemeuler.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x’ = y
 *  y’ = (1 - x^2)y - x
 *  x(0) = 0.0 2
 *  y(0) = 0.0
 *
 *  using Euler’s method.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

int main() {

  double x[2],t,dt;
  double copy[2];
  int i;
  FILE *fp;

  fp = fopen("system.d","w");

  dt = 0.001;
  x[0] = 0.02;
  x[1] = 0.0;

  for(t=0;t<=20;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\n",t,x[0],x[1]);
    for(i=0;i<2;i++)
      copy[i] = x[i];

    x[0] += copy[1]*dt;
    x[1] += ((1.0-pow(x[0],2))*x[1]-x[0])*dt;
  }
  fclose(fp);

  return 0;
}
```

Program for example 13.5.3

```
/*  This is from the file C/systemrk4.c
 *
 *  Example C program to determine an approximate solution to
 *
 *  x’ = y
 *  y’ = (1 - x^2)y - x sin(t)
 *  x(0) = 0.0 2
 *  y(0) = 0.0
 *
 *  using the fourth order Runge-Kutta method.
 *
 *  To compile on the unix platfom, run "gcc <filename>.c -lm" and
 *  then type "a.out" to execute the program.
 *
 */

#include<stdio.h>
#include<math.h>

double f(double x, double y, double t);
double g(double x, double y, double t);

int main() {

  double x,y,t,dt;
  double v1,v2,v3,v4,w1,w2,w3,w4;
  FILE *fp;

  fp = fopen("systemrk4.d","w");

  dt = 0.001;
  x = 0.02;
  y = 0.0;

  for(t=0;t<=20;t+=dt) {
    fprintf(fp,"%f\t%f\t%f\n",t,x,y);
    v1 = f(x, y, t)*dt;
    w1 = g(x, y, t)*dt;
    v2 = f(x+v1/2.0, y+w1/2.0, t+dt/2.0)*dt;
    w2 = g(x+v1/2.0, y+w1/2.0, t+dt/2.0)*dt;
    v3 = f(x+v2/2.0, y+w2/2.0, t+dt/2.0)*dt;
    w3 = g(x+v2/2.0, y+w2/2.0, t+dt/2.0)*dt;
    v4 = f(x+v3, y+w3, t+dt)*dt;
    w4 = g(x+v3, y+w3, t+dt)*dt;
    x += (v1 + 2.0*v2 + 2.0*v3 + v4)/6.0;
    y += (w1 + 2.0*w2 + 2.0*w3 + w4)/6.0;
  }
  fclose(fp);

  return 0;
}

double f(double x, double y, double t) {
  return y;
}

double g(double x, double y, double t) {
  return (1.0 - pow(x,2))*y - x*sin(t);
}
```

D.2 FORTRAN Programs

D.2.1 Programs from Chapter 1

Program for example 1.10.1

```
      program eulerexample

c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x’ = sin(2*t)
c
c     using Euler’s method.
c
c     To compile this on a unix machine, type ’f77 exulerexample.f’
c     and then type ’a.out’ to execute it.

      real x,t,dt,f
      integer n


      open(unit=13,file="output.d")

      n = 0
      dt = 0.01
      x = 3.0

c 100  format(f3.5,i4,f3.5,f3.5,f3.5,f3.5)
      do 10 t = 0, 10, dt
         f = sin(2.0*t)
         write(13,*) t,n,x,f,x+f*dt, 7.0/2.0 - cos(2.0*(t+dt))/2.0
         x = x + f*dt
         n = n + 1
 10   continue
      stop
      end
```

Program for example 1.10.2

```
      program eulerexample

c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x' = 1/(1 + exp(-10*(t-5)))
c
c     using Euler's method.
c
c     To compile this on a unix machine, type 'f77 exulerexample.f'
c     and then type 'a.out' to execute it.

      double precision x,t,dt,f
      integer n

      open(unit=13,file="output.d")

      n = 0
      dt = 0.00001
      x = 1/(1+exp(75.0))

      do 10 t = -1, 1, dt
         f = 75*x*(1-x)
         write(13,*) t,n,x,f,x+f*dt
         x = x + f*dt
         n = n + 1
 10   continue
      stop
      end
```

Program for example 1.10.3

```
      program eulerexample

c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x’’ + sin(t) x’ + cos(t) x = exp(-5*t)
c     x(0) = 2
c     x’(0) = 5
c
c     using Euler’s method.
c
c     To compile this on a unix machine, type ’f77 exulerexample.f’
c     and then type ’a.out’ to execute it.

      double precision x(2),t,dt,f(2)
      integer n

      open(unit=13,file="output.d")

      n = 0
      dt = 0.02
      x(1) = 2.0
      x(2) = 5.0

      do 10 t = 0, 30, dt
         f(1) = x(2)
         f(2) = exp(-5.0*t) - sin(t)*x(2) - cos(t)*x(1)
         write(13,*) t,x(1),x(2)
         x(1) = x(1) + f(1)*dt
         x(2) = x(2) + f(2)*dt
         n = n + 1
 10   continue
      stop
      end
```

D.2.2 Programs from Chapter 2

D.2.3 Programs from Chapter 3

D.2.4 Programs from Chapter 13

Program for example 13.1.1

```
      program eulererroranalysis

c     This is from the file FORTRAN/eulererroranalysis.f
c
c     This is a sample FORTRAN program that solves the differential
c     equation
c
c     x’ = 5 x
c     x(0) = 1
c
c     using Euler’s method.
c
c     To compile this on a unix machine, type ’f77 <filename>.f’
c     and then type ’a.out’ to execute it.

      double precision x,t,dt

      open(unit=13,file="fortrandata.d")

      dt = 0.1
      x = 1.0

      do 10 t = 0, 1, dt
         write(13,*) t,x,exp(5*t)
         x = x + 5*x*dt
 10   continue
      write(13,*) t,x,exp(5*t)
      stop
      end
```

Program for example 13.1.2

```
      program eulererroranalysis2

c     This is from the file FORTRAN/eulererroranalysis2.f
c
c     This is a sample FORTRAN program that solves the differential
c     equation
c
c     x' = -sin(t)
c     x(0) = 1
c
c     using Euler's method.
c
c     To compile this on a unix machine, type 'f77 <filename>.f'
c     and then type 'a.out' to execute it.

      double precision x,t,dt;

      open(unit=13,file="fortrandata.d")

      dt = 1
      x = 1.0

      do 10 t = 0, 30, dt
         write(13,*) t,x,cos(t)
         x = x - sin(t)*dt
 10   continue
      write(13,*) t,x,cos(t)
      stop
      end
```

Program for example 13.2.1

```
      program secondordertaylor

c     This is from the file FORTRAN/secondordertaylor.f
c
c     This is a sample FORTRAN program that solves the differential
c     equation
c
c     x' = 5 x
c     x(0) = 1
c
c     using a second order Taylor series expansion
c
c     To compile this on a unix machine, type 'f77 <filename>.f'
c     and then type 'a.out' to execute it.

      double precision x,t,dt;

      open(unit=13,file="secondordertaylor.d")

      dt = 0.1
      x = 1.0

      do 10 t = 0, 1, dt
         write(13,*) t,x,exp(5*t)
         x = x + 5*x*dt + 25/2*x*dt**2
 10   continue
      write(13,*) t,x,exp(5*t)
      stop
      end
```

Program for example 13.2.2

```
      program secondordertaylor2

c     This is from the file FORTRAN/secondordertaylorhard.f
c
c     This is a sample FORTRAN program that solves the differential
c     equation
c
c     x' = -x^3 + sin(t x)
c     x(0) = 1
c
c     using a second order Taylor series expansion
c
c     To compile this on a unix machine, type 'f77 <filename>.f'
c     and then type 'a.out' to execute it.

      double precision x,t,dt;

      open(unit=13,file="secondordertaylor2a.d")

      dt = 0.2
      x = 1.0

      do 10 t = 0, 5, dt
         write(13,*) t,x
         x = x + (-x**3 + sin(t*x))*dt
     c   + 1.0/2.0*((-3*x**2 + t*cos(t*x))*(-x**3 + sin(t*x)) +
     c   x*cos(t*x))*dt**2
 10   continue
      write(13,*) t,x
      stop
      end
```

Program for example 13.3.1

```
      program rk2

c     This is from the file FORTRAN/rk2.f
c
c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x' = 5*x
c     x(0) = 1
c
c     using the second order Runge-Kutta method.
c
c     To compile this on a unix machine, type 'f77 <filename>.f'
c     and then type 'a.out' to execute it.

      double precision x,t,dt;

      open(unit=13,file="secondorderrk.d")

      dt = 0.1
      x = 1.0

      do 10 t = 0, 1, dt
         write(13,*) t,x,exp(5.0*t)
         x = x + dt/2.0*(f(x,t) + f(x+f(x,t)*dt,t+dt))
 10   continue
      write(13,*) t,x,exp(5.0*t)
      stop
      end

      double precision function f(x,t)
      double precision x,t

      f = 5*x

      return
      end
```

Program for example 13.3.2

```
      program secondorderrk

c     This is from the file FORTRAN/rk2hard.f
c
c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x’ = -x^3 + sin(t x)
c     x(0) = 1
c
c     using the second order Runge-Kutta method.
c
c     To compile this on a unix machine, type ’f77 <filename>.f’
c     and then type ’a.out’ to execute it.

      double precision x,t,dt;

      open(unit=13,file="secondorderrk2a.d")

      dt = 0.2
      x = 1.0

      do 10 t = 0, 5, dt
         write(13,*) t,x
         x = x + dt/2.0*(f(x,t) + f(x+f(x,t)*dt,t+dt))
 10   continue
      write(13,*) t,x
      stop
      end

      double precision function f(x,t)
      double precision x,t

      f = -x**3 + sin(t*x)

      return
      end
```

Program for example 13.3.3

```
      program rk3

c     This is from the file FORTRAN/rk3.f
c
c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x’ = -x^3 + sin(t x)
c     x(0) = 1
c
c     using the third order Runge-Kutta method.
c
c     To compile this on a unix machine, type ’f77 <filename>.f’
c     and then type ’a.out’ to execute it.

      double precision x,t,dt;
      double precision v1,v2,v3;

      open(unit=13,file="rk3.d")

      dt = 0.25
      x = 1.0

      do 10 t = 0, 5, dt
         write(13,*) t,x
         v1 = f(x,t)*dt
         v2 = f(x+0.5*v1,t+0.5*dt)*dt
         v3 = f(x+2.0*v2-v1,t+dt)*dt
         x = x + (v1 + 4*v2 + v3)/6.0
 10   continue
      write(13,*) t,x
      stop
      end

      double precision function f(x,t)
      double precision x,t

      f = -x**3 + sin(t*x)

      return
      end
```

Program for example 13.3.4

```
      program rk4

c     This is from the file FORTRAN/rk4.f
c
c     This is a sampe FORTRAN program that solves the differential
c     equation
c
c     x' = -x^3 + sin(t x)
c     x(0) = 1
c
c     using the fourth order Runge-Kutta method.
c
c     To compile this on a unix machine, type 'f77 <filename>.f'
c     and then type 'a.out' to execute it.

      double precision x,t,dt;
      double precision k1,k2,k3,k4

      open(unit=13,file="rk4.d")

      dt = 0.25
      x = 1.0

      do 10 t = 0, 5, dt
         write(13,*) t,x
         k1 = f(x,t)*dt
         k2 = f(x+k1/2.0,t+dt/2.0)*dt
         k3 = f(x+k2/2.0,t+dt/2.0)*dt
         k4 = f(x+k3,t+dt)*dt
         x = x + (k1 + 2.0*k2 + 2.0*k3 + k4)/6.0
 10   continue
      write(13,*) t,x
      stop
      end

      double precision function f(x,t)
      double precision x,t

      f = -x**3 + sin(t*x)

      return
      end
```

Program for example 13.4.1

```
still needs to be written!!!!!!!!!!
```

Program for example 13.5.1

```
      program systemeuler

c     This is from the file FORTRAN/systemeuler.f
c
c     This is a sample FORTRAN program that solves the differential
c     equation
c
c     x’ = y
c     y’ = (1 - x^2)y - x
c     x(0) = 0.02
c     y(0) = 0.0
c
c     using Euler’s method.
c
c     To compile this on a unix machine, type ’f77 <filename>.f’
c     and then type ’a.out’ to execute it.

      double precision x(2),t,dt
      double precision copy(2)

      open(unit=13,file="systemfortran.d")

      dt = 0.001
      x(1) = 0.02
      x(2) = 0.0

      do 10 t = 0, 20, dt
         write(13,*) t,x(1),x(2)
         copy(1) = x(1)
         copy(2) = x(2)
         x(1) = x(1) + (copy(2))*dt
         x(2) = x(2) + ((1.0 - copy(1)**2)*x(2) - x(1))*dt
 10   continue
      stop
      end
```

Program for example 13.5.3

```
      program systemrk4

c     This is from the file FORTRAN/systemrk4.f
c
c     This is a sample FORTRAN program that solves the differential
c     equation
c
c     x’ = y
c     y’ = (1 - x^2)y - x sin(t)
c     x(0) = 0.02
c     y(0) = 0.0
c
c     using the fourth order Runge-Kutta method.
c
c     To compile this on a unix machine, type ’f77 <filename>.f’
c     and then type ’a.out’ to execute it.

      double precision x,y,t,dt
      double precision v1,v2,v3,v4,w1,w2,w3,w4

      open(unit=13,file="systemrk4f.d")

      dt = 0.001
      x = 0.02
      y = 0.0

      do 10 t = 0, 20, dt
         write(13,*) t,x,y

         v1 = f(x, y, t)*dt
         w1 = g(x, y, t)*dt

         v2 = f(x+v1/2.0, y+w1/2.0, t+dt/2.0)*dt
         w2 = g(x+v1/2.0, y+w1/2.0, t+dt/2.0)*dt

         v3 = f(x+v2/2.0, y+w2/2.0, t+dt/2.0)*dt
         w3 = g(x+v2/2.0, y+w2/2.0, t+dt/2.0)*dt

         v4 = f(x+v3, y+w3, t+dt)*dt
         w4 = g(x+v3, y+w3, t+dt)*dt

         x = x + (v1 + 2.0*v2 + 2.0*v3 + v4)/6.0
         y = y + (w1 + 2.0*w2 + 2.0*w3 + w4)/6.0
 10   continue
```

```
      stop
      end

      double precision function f(x,y,t)
      double precision x,y,t

      f = y

      return
      end

      double precision function g(x,y,t)
      double precision x,y,t

      g = (1.0-x**2)*y - x*sin(t)

      return
      end
```

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. Foundations of Mechanics. Addison Wesley, 2nd edition, 1978. 21

[2] V.I. Arnold. Mathematical Methods of Classical Mechanics. Springer, 2nd edition, 1997. 21

[3] Ruel V. Churchill, James W. Brown, and Roger F. Verhey. Complex Variables and Applications. McGraw-Hill, third edition, 1974. 555

[4] H.C. Corben and Philip Stehle. Classical Mechanics. Wiley, 1950. 21

[5] John J. Craig. Introduction to Robotics. Addison-Wesley, second edition, 1989. 568

[6] Gene Franklin, J.D. Powell, and Abbas Emami-Naeini. Feedback Control of Dynamic Systems. Prentice Hall, fifth edition, 005. 326

[7] Joel N. Franklin. Matrix Theory. Prentice-Hall, Inc., 1968. 182, 576, 582, 583

[8] Herbert Goldstein. Classical Mechanics. Addison-Wesley Publishing Company, Reading, MA, 2nd edition, 1980. 21

[9] Morris W. Hirsch and Stephen Smale. Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, Inc., 1974. 182, 531

[10] R. A. Howland. Intermediate Dynamics: A Linear Algebraic Approach. Springer, 2005. 21, 27

[11] Thomas W. Hungerford. Algebra. Springer-Verlag, 1974. 4

[12] Frank P. Incropera and David P. DeWitt. Fundamentals of Heat and Mass Transfer. Wiley, second edition, 1895. 55

[13] Benjamin C. Kuo. Automatic Control Systems. Prentice Hall, seventh edition, 1995. 326

[14] Ronald Mancini. Feedback amplifier analysis tools. Technical report, Texas Instruments, 2001. 269

[15] J. L. Meriam and L. G. Kraige. Engineering Mechanics , Dynamics. Wiley, 5th edition, 2001. 21

[16] R. M. Murray, Z. X. Li, and S. S. Sastry. A Mathematical Introduction to Robotic Manipulation. CRC Press, Boca Raton, Florida, 1994. 568

[17] Isaac Netwon. Philosophiae naturalis principia mathematica, 1687. 21

[18] Isaac Newton. The Principia. Prometheus Books, 1995. 21

[19] L. Schwartz. Theorie des Distributions. Hermann, Paris, 1966. 232

[20] Barry N. Taylor, editor. The International System of Units (SI). NIST Special Publication 330. United States Department of Commerce, 2001. xxv, 20, 21

[21] S. Wiggins. Introduction to Applied Nonlinear Dynamical Systems and Chaos. Springer, 1990. 531

[22] L.A. Zadeh. Fuzzy sets. Information and Control, 8:338–353, 1965. 4

Index