Discrete Space-time Models in Biology

Frederic Y.M. Wan

Department of Mathematics, UC, Irvine
Current address: Irvine, CA 92697-3875
E-mail address: fwan@uci.edu
URL: http://www.math.uci.edu/~fwan

2000 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18

Abstract. Course Notes for the MCBU Course Math 113A

Contents

Part 1. Single Population Models 1

Chapter 1. Growth of a Population 3
1. Mathematical Modeling 3
2. Dynamical Systems 6
3. Linearization 8
4. Immigration 12

Chapter 2. Evolution and Equilibria 15
1. The Logistic Growth Model 15
2. Dimensionless Form 16
3. Fixed Points and Their Stability 17
4. Cobweb Graphs 18

Chapter 3. Linear Stability Analysis 21
1. The Mean Value Theorem 21
2. The Basic Stability Theorem 21
3. The Inconclusive Case of $F'(\bar{x}) = 1$ 23
4. When $F'(\bar{x}) = 1$ and $F''(\bar{x}) = 0$ 24
5. Relocation of a Fixed Point to the Origin 25
6. Period Doubling and Cycles 25

Chapter 4. Bifurcation 29
1. Saddle-Node Bifurcation 29
2. Transcritical Bifurcation 31
3. Pitchfork Bifurcation 32
4. Other Types of Bifurcation 33
5. Exercises 33

Chapter 5. Second Order Models 35
1. Fibonacci Rabbits 35
2. Fibonacci Sequence 36
3. Plant Propagation 40
4. Plant Growth without Immigration 41
5. Fixed Points and Stability 44

Chapter 6. External Forcing 49
1. Plant Growth with Immigration 49
2. Variation of Parameters 50
3. Method of Undetermined Coefficients 52

Part 2. Interacting Populations 55

Chapter 7. Linear Systems 57
1. Red Blood Cell Production 57
2. The Matrix Eigenvalue Problem 60
3. The IVP for RBC 61
4. Drug Uptake I 63
5. Drug Uptake II 66

Chapter 8. Markov Chains 71
1. A Forest of Red Oaks and White Pines 71
2. The Steady State for the Red Oak - White Pine Problem 73
3. Fixed Points for General Markov Chains 74
4. Regular Markov Chains 75
5. DNA Mutation 77
6. Absorbing Markov Chains 80
7. Immunity after Recovery 83
8. Expected Transient Stops to an Absorbing State 85
9. Appendix - Proof of Theorem 18 87
10. Exercises 88

Chapter 9. Nonlinear Systems 91
1. Rabbits and Coyotes (Predator-Prey) 91
2. A Resource Limiting Predator-Prey Model 94
3. Viral Dynamics 99
4. Metabolism and Enzyme Kinetics 102

Chapter 10. Bistability 109
1. A Dimerized Reaction 109
2. Two Competing Populations 112
3. Hopf Bifurcation 118

Chapter 11. Mendelian Genetics 123
1. Hardy-Weinberg Stability Theorem 125
2. Selective Breeding 127
3. Gene Frequencies 129
4. Mutation 131

Part 3. Appendices 133

Chapter 12. Linear Equations - Algebra 135
1. Linear Equations 135
2. Matrices 138
3. Gaussian Elimination 142

Chapter 13. Geometry of Linear Equations 147
1. Linear System as a Vector Equation 147
2. Matrix Diagonalization 151
3. Decoupling a Linear System 153

Bibliography 155

Preface

Mathematics is about mathematical structures. geometric, algebraic, topological, etc., independent of specific questions. Applied mathematics is about mathematical issues, evolution, steady states, stability, bifurcation, optimization, etc., independent of specific applications. The course Math 113 on Mathematical Models in Biology is an introduction to mathematical modeling and analysis of phenomena in biological sciences accessible to undergraduates at University of California, Irvine. It is designed specifically to ready undergraduates in mathematics or biomedical sciences to participate in the MCBU summer research program as early as the summer after their sophomore year. In the process, the course introduces the students to the various mathematical issues pertaining to such phenomena and to applied mathematics generally. To be able to analyze the mathematical models we develop and to address the attendent applied mathematical issues, students also need to have a working knowledge of some basic mathematical and computational techniques useful in quantitative studies of biology not already acquired through the prerequisites for Math 113. In general, it is necessary to know about these techniques before we can see how they are used in biology, not to mention doing research in mathematical and computational biology.

Math 113A is about mathematical models of biological phenomena taking place in discrete stages in time or discrete locations in space. To make the course accessible to second year students in Mathematics and in the Biological Sciences with a limited mathematical background of first year calculus, the main mathematical techniques involved are difference equations and matrix algebra with Markov chains being a principal applications of the latter. Through some nonlinear models, the course will introduce students to some techniques for solving nonlinear difference equations as well as expose students to concepts of stability and bifurcation.

Math 113B is about mathematical models of biological phenomena involving ordinary differential equations. To make the course accessible to second year students in the Biological Sciences having taken only first year calculus, the course begins with a short introduction to analytical and numerical methods for first order ODE. Additional mathematical methods for, and numerical simulations of ODE will be introduced as needed by specific mathematical models. Through relevant nonlinear models, the important issues of stability and bifurcation will be revisited in the context of differential equations. Models leading to boundary value problems in ODE will also be discussed enabling us to introduce students to important numerical methods such as the shooting method for these problems. Students interested in a more in depth discussion of these topics are encouraged to take Math 227A and/or Math 290.

Math 113C is designed to introduce students to partial differential equation and stochastic methods through appropriate mathematical models in biological sciences. For example, Turing instability is introduced through the problem of tissue patterning in developmental biology. Much of the useful information that can extracted from PDE models necessarily require numerical simulations; This is less so for ODE and differnece equation models though still extensive and unavoidable at times. As we can see from Math 113A, it takes more than a hundred iterations for a simple first order dynamical systems to show sign of a (two-) cycle. The required computations are best relegated to a computer.code designed for this purpose. For this purpose, students in Math 113 would be well served by a working knowledge

of some useful mathematical software such as Mathematica, MatLab, Maple and others. We will take time out from modeling to introduce Math 113 students to the rudiments of MatLab (or Mathematica).

It is a general objective of the Math 113 sequence to acquaint students in the course with simple and highly idealized mathematical models in the biological sciences and to introduce them to the salient features of the modeling process. While some of the models may seem unrealistic (just as treating planets and projectiles as mass points in physics), we nevertheless gain insight from investigating their implications or use them as stepping stones toward more realistic models. In truth, mathematical modeling is as much an arts as it is a science. There is no single recipe for constructing an appropriate model for any phenomenon to be analyzed. In fact, the same real life phenomenon of interest may be modeled in completely different ways depending on the particular features and issues of the phenomenon of interest of the researcher. For these reasons, the essential ingredients of mathematical modeling are best acquired by exposure to examples and distill from them what general features need to be incorporated in different classes of models. Math 113 is designed to meet this objective.

It is difficult to find a textbook for the intended purpose and curriculum of this course. Two references on mathematical biology (Edelstein-Keshet [**3**] and Voit [**24**]) have been listed mainly to show how difference equations and differential equations enter into the study of different biological phenomena. To provide students with a coherent summary of the material to be covered in the course, this set of course notes was developed as the main reading materials for the prescribed course curriculum for Math 113A (except for sections whose title has an asterisk, "*", indicating that the material is for optional reading). However, much of the learning will be through the exercises in the weekly assignments. It has been documented by research in learning that students learn better by an active learning process.

Frederic Y.M. Wan
Irvine, CA
November 26, 2012

Part 1

Single Population Models

CHAPTER 1

Growth of a Population

1. Mathematical Modeling

These notes are about mathematical models iof biological phenomena and the analysis of these models. If you have had no prior experience with mathematical models or the mathematical modeling process, it is natural to ask what they are about. Instead of giving a general abstract theoretical description, we begin by illustrating mathematical modeling with a simple example from the study of the growth (or decline) of a single population. The example shows some essential features of the modeling process to position ourselves to talk more sophisticatedly about it in subsequent chapters working with more complex models. Since we are interested in learning how to do quantitative research in biological sciences, we work mainly with models for biological phenomena in these notes.

1.1. What is the Question? Mathematical modeling usually starts with a question on a phenomenon of interest. We may want to know why an observed phenomenon takes place, or whether our expectation of what should happen next is justified. In the early days of physics, we have questions such as "Is the Earth's orbit around the Sun circular?" When observational data seemed not to support this expectation, the question was changed to "Is it elliptical?" and "why?" In the biological sciences, a phenomenon of general interest is the growth of a population, which may be the growth of cultured cells, of the virus that is attacking cells in your body, or of the human population on Earth. With the Earth's population reaching 7 billion on October 31, 2011, we are naturally interested in knowing "when will it reach 10 billion?" or "What will it be in year 2050?" Such questions are at the level of those about the Earth's orbit in physics mentioned above and may be investigated by mathematical modeling as Newton did on planetary motion.

1.2. What Do We Know? In order to explain or predict something about population growth, we have to know something about what regulates the growth. In physics, the motion of planets and objects in our daily life is governed by Newton's second law. Unfortunately, we do not have a simple universally applicable law for pupulation growth (or most biological phenomena) analogous to Newton's. For a specific environment, the growth rate may be estimated from available census data or by examining the various birth and death processes. If we want to avoid plunging immediately into such highly technical activities, we may stay at the phenomenological level by extracting from what is observable a reasonable assumption about the relation between current and future population sizes and then see if its consequences fit the facts. Historically, Newton's second laws also started out as a hypothetical assumption; but its predicted consequences fit the corresponding actual observations so perfectly (including the highly sensitive space capsules in outer

space flights today) that it eventually was accepted as the law that governs the dynamics of mass objects (as long as they do not travel near the speed of light).

For population growth, it is an observed fact that people generally beget more people; the greater the current population the larger the future population. It is therefore not unreasonable to hypothesize that the future size of a population should depend on its current size. This simple hypothesis (which may or may not be appropriate) can be translated into a mathematical relation on the evolution of the population size with time, thereby giving us a mathematical model for the growth of the human population (or any other population that behaves similarly).

1.3. A Mathematical Model. To formulate a mathematical model for the growth of a population, let y_n (or equivalently $y(n)$) be the population size (in some biomass unit) at the end of "stage" n. In Table 1 for example, we have from U.S. Census Bureau data the sequence of (roundoff) estimates of the world's population at the end of each of the last six decades. Here we use $n = 1, 2, 3, ..., 6$ to indicate the first decade, the second decade, etc., starting from 1960 (instead of writing out the actual years when the intent is understood). If n ranges over the positive integers, the collection $\{y_n\}$ constitutes an *infinite sequence* of numbers (the kind we encounter in elementrary calculus) that spans all decades starting from 1960.

Table 1

n	1	2	3	4	5	6
$Year$	1960	1970	1980	1990	2000	2010
y_n	3.05	3.72	4.45	5.30	6.10	6.90

*http://upload.wikimedia.org/wikipedia/commons/1/11/World_population_history.svg.

Table 2

$Year$	1927	1959	1974	1987	1999	2011
n	-84	-52	-37	-24	-12	0
$data$	2	3	4	5	6	7

What constitutes a "stage" (through the subscript n) in a temporal sequence is rather arbitary and may be chosen to suit our purpose at hand. For the infinite sequence starting with $n = 1$, the stages may be taken to indicate consecutive years starting from year 0 A.D. (instead of consecutive decades starting from 1960 as in Table 1). The set of data in Table 2 employs another stage designation with consecutive stages corresponding to the number of years after 2011. By letting 2011 correspond to the reference initial year $n = 0$, this stage labeling is particularly appropriate for look aheading in forcasting future year population. Table 3 gives still another choice of stage unit, each stage being a year with 1959 being the reference initial year.

Table 3

$Year$	1804	1927	1959	1974	1987	1999	2011
n	-155	-32	0	15	28	40	52
$data$ ($billions$)	1	2	3	4	5	6	7

To predict the population size for future years, we need to know what drives the growth. The answer to that question can be given at many levels, molecular, phyiological, organismal, etc. For a true understanding of the growth mechanism, we would have to examine the various birth and death proceses and factors influencing them to extract the needed information. For a broad stroke description (at the level of newspaper reporting to lay readers), we may avoid all the technical details

and simply make the phenomenological characterization that *next year's population size is a function of this year's population.* Certainly, if there is no population now, there will no population next year (without immigration). Under normal circumstances, "people beget more people." This phenomenological description of the growth process is summarized in the following mathematical relation

$$y_{n+1} = F_n(y_n) \tag{1.1}$$

with $F_n(0) = 0$ since (without immigration) we must have people to beget offsprings. The relation gives the population in stage $n + 1$ as a function of the population in stage n with the functional relation $F_n(\cdot)$ still to be specified. We expect $F_n(y_n) > 0$ if $y_n > 0$ since more people would beget moe offsrpings; an assumption that holds for human and many other known populations.

By the subscript n in $F_n(\cdot)$, we are allowing this functional depenedence to vary with stage (and therefore time). For example, a severe epidemic of deadly disease may deplete a sizable portion of the Earth's population at some stage and hence chanage he growth of the population during the period of the epidemic. If the growth process is the same for all (foreseeable) future stages, so that $F_n(\cdot)$ is the same for all n, we would drop the subscript to get

$$y_{n+1} = F(y_n). \tag{1.2}$$

In population dynamics, the index n is an indicator of the stage in time. For other applications, it may specify different locations in space, different storage compartments, etc. In all cases, (1.2) when written as $y_{n+1} - y_n = F(y_n) - y_n \equiv f(y_n)$, is known as a *difference equation.* It describes the relationship between two states of the phenomenon under study at consecutive stages. For temporal stages, time is identified and recorded in discrete units of seconds, minutes, hours, days, months, years, decades, or some other unit appropriate for the stages of the particular problem with one to one correspondence to the integers.

The growth process is said to be *autonomous* when the functional relation F does not vary with n but remains the same from stage to stage. Note that even when F does not change with stage changes, the population size generally does. For example, the function $F(y) = y+1$ does not change with stage changes. However, we have from (1.1) or (1.2)

$$y_{n+1} = F(y_n) = y_n + 1$$

so that y_{n+1} differs from y_n with $y_{n+1} - y_n = 1$. Over a sufficiently short time period, the relevant characteristics of the systems being investigated (such as birth and death rate of the population) do not change significantly. For these systems, we may assume the functional relation F is autonomous.if we limit our investigation to a duration when the system characteristics remains more or less unchanged.

For many population including the human population, the autonomous $F(\cdot)$ is an increasing function of its argument, $F'(y) = dF(y)/dy > 0$, since a larger population generally produces more offsprings. However there is generally a diminishing return tothe rate of increase of a population, caused by limited space and other resources. Altogether, we expect the *growth function* $F(y)$ to have the following three properties:

$$i)\ \ F(0) = 0, \quad ii)\ \ F(y) > 0 \ \text{ for } y > 0, \quad iii)\ \ F'(y) > 0, \quad iv)\ \ F''(y) < 0. \tag{1.3}$$

It is understood that y is non-negative ($y \geq 0$) since we cannot have a negative population. Under these assumptions, the graph of $F(y)$ starts at the origin and

is a monotone increasing concave curve. As we shall see, unlike Newton's laws of motion, these properties may not hold for all populations. But they are sufficiently prevalent to be taken as typical for a first attempt to investigate the general growth of a population.

2. Dynamical Systems

2.1. The Growth Rate. As simple looking an equation as (1.2) is, it is often more meaningful to re-write it in an equivalent form by subtracting y_n from both sides of the equation to get

$$y_{n+1} - y_n = F(y_n) - y_n \equiv f(y_n) \tag{2.1}$$

or more compactly as

$$\Delta y_n = f(y_n) \tag{2.2}$$

where

$$\Delta y_n = y_{n+1} - y_n \tag{2.3}$$

is known as the *finite difference* of y_n. It is the analogue of the derivative of $y(t) = y(n)$ $(= y_n)$ in calculus with $\Delta t = 1$ (and hence not taking any limit):

$$\frac{\Delta y_n}{\Delta t} = \frac{y_{n+1} - y_n}{(n+1) - n} = y_{n+1} - y_n.$$

For the more general case where the unit of time increment is allowed to be changed, we write $t = n\Delta t$ (instead of $t = n\cdot$ stage unit) so that

$$y(t) = y(n\Delta t) = y_n.$$

In that case, we have as $\Delta t \to 0$

$$\begin{aligned} \frac{y(t+\Delta t) - y(t)}{(t+\Delta t) - (t)} &= \frac{y((n+1)\Delta t) - y(nt)}{\Delta t} \\ &= \frac{y_{n+1} - y_n}{\Delta t} \to \frac{d}{dt}\left[y(t)\right] \end{aligned}$$

and the difference equation (2.2) becomes an ordinary differential equation (ODE),

$$\frac{d}{dt}\left[y(t)\right] = f(y(t)).$$

Mathematical models involving ODE permeate throughout the biomedical sciences and will be discussed extensively in Math 113B. Here, we focused on problems involving observations and measurements on evolving phenomena in discrete increments of the time units (allowing us to take $\Delta t = 1$). Such phenomena are adequately and appropriately modeled by difference equations..

2.2. The Initial Value Problem. We begin our investigation of the evolution of the human population on Earth by working with the autonomous relation (1.2) until further notice. Starting with a known initial population y_0 at the end of some reference initial year taken to be stage n = 0, we have from (1.2)

$$y_1 = F(y_0), \qquad y_2 = F(y_1) = F(F(y_0)), \qquad \text{.....} \tag{2.4}$$

For the example $F(y) = y + 1$ used earlier, we have

$$\begin{aligned} y_1 &= F(y_0) = y_0 + 1, \\ y_2 &= F(y_1) = F(y_0 + 1) = (y_0 + 1) + 1 = y_0 + 2, \\ y_3 &= F(y_2) = F(y_0 + 2) = y_0 + 3, \quad ....., \end{aligned} \tag{2.5}$$

and by induction

$$y_n = F(y_{n-1}) = y_0 + n$$

giving a general formula for the population at any future stage.

It should be evident that we can obtain the population size of any future year once we know the function $F(\cdot)$ and the population size of some initial reference year. This is accomplished by repeated applications of (1.2) or its finite difference form (2.2), as we did for the particular $F(\cdot)$ above. For a more general $F(\cdot)$ and n large, the task may be tedious and the expression $F(y_n)$ does not readily give any useful information for y_{n+1} beyond the first few stages. Take $F(y) = e^{\cos(y)}$ for example, we have

$$\begin{aligned} y_1 &= F(y_0) = e^{\cos(y_0)}, \\ y_2 &= F(y_1) = F(e^{\cos(y_0)}) = e^{\cos(e^{\cos(y_0)})}, \\ y_3 &= F(y_2) = F(e^{\cos(e^{\cos(y_0)})}) = ... \quad ....., \end{aligned} \tag{2.6}$$

Evidently, the expression for a general stage is too complicated to write down and in any case does not give any hint of the actual population size for stage n. Nevertheless, the actual population size for any stage can be calculated using a laptop computer with mathematical software such as MatLab, Mathematica or Maple.

For the population growth model we adopted, (1.2) or (1.1), it is clear that specification of a reference year end population y_0 is a significant part of the modeling process. Without the population size known at some stage, nothing can be said about the population size of any year, past, present or future. This single data point so important for an answer to our original question is known as the *initial condition* for the difference equation (2.2) or equivalently, the growth relation (1.2).

THEOREM 1. *If the growth of the population is governed by the growth rule (1.2) and the population size of some reference stage, taken to correspond to $n = 0$, is known to be Y, i.e., $y_0 = Y$, then the population y_k of any future stage k is give by*

$$y_k = F^k(Y)$$

where F^k means applying the function $F(\cdot)$ k times as illustrated in (2.6) and (2.5).

In the form (2.2), the growth relation is a *finite difference equation* for the unknown future population size y_n. The interest in how an evolving system changes with time permeates throughout engineering and science (including the social and management sciences). *Evolution* is therefore a major theme and issue in mathematical modeling and the analysis of mathematical models.

Loosely speaking, evolving systems that are modeled by ordinary difference or differential equations (ODE) are generally called dynamical systems. More technically, the term dynamical systems refers to the study of certain issues pertaining to such evolving systems by a class of mathematical methods and approaches. Much of these methods and approaches will be illustrated through examples in these

notes. Hence, it may be said that this course is concerned with dynamical systems generally.

3. Linearization

3.1. Geometric Growth. As stated previously, we assume $F(0) = 0$ since the population does not increase when there is no one available to produce new offsprings. From the definition of derivative in calculus (or a two-term Taylor's expansion for $F(y_n)$ about the origin [**15**]), we have $[F(y_n) - F(0)]/(y_n - 0) = F'(0) + c(y_n - 0)$ for some constant c, or

$$F(y_n) = F'(0)y_n + cy_n^2. \tag{3.1}$$

For sufficiently small y_n, the quadratic term cy_n^2 is an order of magnitude smaller than the first term and can be neglected for a good approximation. In that case, (2.2) simplifies to

$$y_{n+1} = Ay_n \quad \text{or} \quad \Delta y_n = ay_n \tag{3.2}$$

where $A = a + 1 = F'(0) > 0$ is called the *amplification factor* for the population size while a is the *growth rate constant* with unit of 1/time.

Note that equation (3.2) is linear in the unknown y_n and linear equations are known to have many nice things that happen to them. For example, repeated applications of the first relation of (3.2) leads to

$$y_k = y_0 A^k \tag{3.3}$$

where y_0 is the (known) population at some reference year designated to be $n = 0$. For example, we may take $n = 0$ ro correspond to 2011 in which casse $y_0 = 7$ billion (units of biomass). Other options are also possible and may be more appropriate for a particular application. Once the amplication factor $A = F'(0) > 0$ is known, the formula (3.3) gives the answer to the questions posed at the start of this chapter: But even without a specific value for A, the following information is apparent from the solution (3.3) of the IVP:

(1) If $y_0 = 0$, the population remains at zero.
(2) For $y_0 > 0$ (as a negative population is not meaningful in the context of human population growth), the initial population grows without bound as $n \to \infty$ if $A > 1$.
(3) For $y_0 > 0$, the initial population dimimishes and tends to zero as n $\to \infty$ if $(0 <) A < 1$.

Definition 1. *A solution y_E is said to be a fixed point of the difference equation (1.2) if $F(y_E) = y_E$.*

For the linear equation (3.2), $y_E = 0$ is the only fixed point if $A \neq 1$. (If $A = 1$, then any y_0 is a fixed point.) In terms of population growth, a fixed point is said to be an *equilibrium* population since once at that size the population would not change, a situation similar to the at rest position in Newtonian mechanics.

With the extensive censure data available on the world's population, we have many options in choosing the initial condition for the difference equation of the simple growth model (1.2) (or the more general nonautonomous model (1.1)). One previously mentioned choice was $y_0 = 7$ billion by choosing the reference initial year to be 2011. For the purpose of illustrating the predictive capacity of our model, we instead choose n to correspond to year 1959 with $y_0 = 3$ billion. In that case,

the expression (3.3) answers the questions posed at the start of this investigation in terms of the parameter A. It determines that

- the Earth's population in year 2050 will be $3 \times A^{91}$ billions, and
- the Earth's population will reach 10 billiions by year $1959 + k = 1959 + \ell n(10/3)/\ell n(A)$.

The formula (3.3) is nice and compact and hence deserves to be encapsulated in the following theorem:

THEOREM 2. *Under the assumption that the population size of stage $n+1$ is proportional to the population of stage n, the population y_k of any future stage k is uniquely given by (3.3).*

3.2. The Amplification Factor. With $y_0 = 3$ (billion), we need to specify a numerical value for the amplification factor A in order to be able to use the expression (3.3) for determining a numerical value for the population at any other stage (in units of a year) beside $n = 0$. Any meaningful value for A should be consistent with available censure data. To illustrate, we see from Table 3 the Earth's human population in 1974 (witb $n = 15$) was (almost exactly) 4 billions. For (3.3) to be consistent with this piece of data, we must have

$$4 = 3 \times A^{15} \qquad or \qquad A = \left(\frac{4}{3}\right)^{1/15} == 1.019363899.....$$

and therewith

$$y_k = 3 \times (4/3)^{k/15}. \tag{3.4}$$

We can now use the expression (3.4) to answer more concretely the questions we asked at the start of the modeling discussion: By this formula,

- the Earth's population in year 2050 is predicted to be $3 \times (4/3)^{91} = 17.2....$ billion, and
- the Earth's population will reach 10 billiion by year $1959 + 62.776.....\simeq$ 2022

How good are these predictions? We do not know, since we do not have the actual data to check them. The future is not yet here! However, we do have data from the past which can be compared with the prediction by (3.4) as shown in Table 4 below.

Table 4

Year	1804	1927	1959	1974	1987	1999	2011	2022	2050
n	-155	-32	0	15	28	40	52	62.8	91
y_k (*billion*)	0.15	1.62	3	4	5.13	6.46	8.13	10	17.2
data	1	2	3	4	5	6	7	?	?

The predictions by (3.4) seem reasonable for the next two data points after 1974 (1987 and 1999) with a maximum oercentage error of less than 8% which is quite acceptable for most purposes. However, the accuracy (3.4) is not as good if we go further into the future or backward to the past. For example, (3.4) give a population size of 1.62 billion (units of population) for 1927 while the censure data shows that it was 2 billions for a 19% underestimate. The adequacy of (3.4) deteriorates dramatically as we go deeper into the past, giving a population of 0.15 billion when the record shows that it was 1 billion. Going forward in time beyond 1999 is not much better with a percentage error of about 16% for 2011. The

serious discrepancies serve as a warning on the adequacy of the linear model for predicting the distant future. We should therefore consider possible improvement of our model. This we will do in the next chapter.

EXERCISE 1. *Take $y_0 = 7$ billion with $n = 0$ corresponding to year 2011, use the data point of $y_{-51} = 3$ billion to determine the amplification factor A and use the result to determine*

- a) the Earth's population in year 2050 , and
- b) the year when Earth's population reaches 10 billiion

3.3. The Least Square Approximations*. Before moving onto an improved mathematical model, it should be pointed out that there are ways to improve on the predictive power of the linear growth model without changing the actual model. One of these is a better use of the data available from the U.S. Census Bureau record of annual population for determining the two parameters y_0 and A in (3.3). Up to now, we only made use of only two data points to fix these two parameters. In the first attempt, the censure data for the population of 1959 was used as the initial condition to determine y_0 and the data for 1974 was used for determining the amplification factor A. It would seem that we would get better accuracy from the same model if we make more use of the available data (even if not all of them). However, the model (3.3) which gave rise to (3.4) contains only two parameters that can be used to fit the available data. As such, only two data point can be fitted exactly. To make use of more data points, we would have to either modify the model to contain more parameters (possibly as many as the number of data points we wish to fit exactly) or consider an alternative way to take into account more data points without change the linear model (3.3). We will do the first of these in the next and later chapters. In this subsection, we illustrate how the second approach may be implemented.

Suppose we want to make use of the four pieces of census data from the U.S. Census Bureau for $1959, 1974, 1987$ and 1999 (in Table 3 (or Table 4) to determine the two parameters y_0 and A: Clearly, we cannot choose y_0 and A to match all four data points exactly. Whatever choice we make, there would generally be a difference $y_n - \bar{y}_n$ between the value y_n given by (3.3) and the Census Bureau data $\bar{y}_n$ for stage n. Since we cannot make all the differences zero, we will do the best we can. To do that, we need to specify what do we mean by "best"? One possible criterion is to minimize the sum of the square of these discrepancies:

$$E_N^2 = (y_{k_0} - \bar{y}_{k_0})^2 + \text{.....} + (y_{k_N} - \bar{y}_{k_N})^2 = \sum_{i=0}^{N} (y_{k_i} - \bar{y}_{k_i})^2 \tag{3.5}$$

with $k_0 = 0$ corresponding to 1959 as before. (We should not minimize just the sum of all discrepancies as it would allow cancellations of large discrepancies with opposite signs.) For illustrative purpose, we use the four data points corresponding to years 1959 ($k_0 = 0$), 1974 ($k_1 = 15$), 1987 ($k_2 = 28$), and 1999 ($k_3 = 40$). In (3.5), y_{k_i} is given by (3.3) to be $y_0 A^{k_i}$. As such, E_N^2 is a function of y_0 and A. Our (least square) minimization criterion is to choose y_0 and A to render E_N^2 (at least) a local minimum. From elementary calculus, we know that this minimum is

attained at a stationary point of E_N^2 determined by

$$\frac{\partial E_N^2}{\partial y_0} = 2\sum_{i=0}^{3}\left(y_{k_i}-\bar{y}_{k_i}\right)A^{k_i} = 2\sum_{i=0}^{3}\left(y_0A^{k_i}-\bar{y}_{k_i}\right)A^{k_i} = 0 \tag{3.6}$$

$$\frac{\partial E_N^2}{\partial A} = 2\sum_{i=0}^{3}\left(y_{k_i}-\bar{y}_{k_i}\right)k_iy_0A^{k_i-1} = 2\sum_{i=0}^{3}\left(y_0A^{k_i}-\bar{y}_{k_i}\right)k_iy_0A^{k_i-1} = 0 \tag{3.7}$$

The first equation can be solved for y_0 to get

$$\begin{aligned} y_0 &= \frac{\sum_{i=0}^{3}\bar{y}_{k_i}A^{k_i}}{\sum_{i=0}^{3}A^{2k_i}} = \frac{\bar{y}_0A^0+\bar{y}_{k_1}A^{k_1}+\bar{y}_{k_2}A^{k_2}+\bar{y}_{k_3}A^{k_3}}{A^0+A^{2k_1}+A^{2k_2}+A^{2k_3}} \\ &= \frac{3+4A^{15}+5A^{28}+6A^{40}}{1+A^{30}+A^{56}+A^{80}} \end{aligned} \tag{3.8}$$

where we have used the first four pieces of data starting with 1959 ($n=0$). The expression (3.8) can be used to eliminate y_0 from the second condition (3.7) to get

$$\frac{\sum_{m=0}^{3}\bar{y}_mA^{k_m}}{\sum_{m=0}^{3}A^{2k_m}}\sum_{i=0}^{3}k_iA^{2k_i-1} = \sum_{i=0}^{3}k_iA^{k_i-1}\bar{y}_{k_i} \tag{3.9}$$

or, with $k_0=0$ corresponding to 1959 and $N=3$,

$$\frac{3+4A^{15}+5A^{28}+6A^{40}}{1+A^{30}+A^{56}+A^{80}} = \frac{15\cdot4\cdot A^{14}+28\cdot5\cdot A^{27}+40\cdot6\cdot A^{39}}{15\cdot A^{29}+28\cdot A^{55}+40\cdot A^{79}}$$

which can be re-writtem as

$$\begin{aligned} &\left(3+4A^{15}+5A^{28}+6A^{40}\right)\left(15\cdot A^{29}+28\cdot A^{55}+40\cdot A^{79}\right) \\ =\ &\left(15\cdot4\cdot A^{14}+28\cdot5\cdot A^{27}+40\cdot6\cdot A^{39}\right)\left(1+A^{30}+A^{56}+A^{80}\right) \end{aligned} \tag{3.10}$$

Equation (3.10) is a polynomial equation $P_j(A)=0$ of 119 degree (with a zero root of multiplicity 14) and can be solved for A. Maple's *fsolve* and Mathematica's *NSolve* both give just one positive real root (two negative real roots and 102 complex roots):

$$A = 1.015795908....$$

Correspondingly, we get from (3.8)

$$y_0 = 3.178378617...$$

so that (3.3) becomes

$$y_k = (3.178378617...)\times(1.015795908..)^k \tag{3.11}$$

We can now use (3.11) to compute the corresponding y_k for different past and future years:

Table 5

$Year$	1959	1974	1987	1999	2011
y_k	3.178...	4.020...	4.929...	5.949...	7.180...
$data$	3	4	5	6	7

The agreement between the actual data and (3.3) based on the *least squares method* using four data points is evidently "better" than the exact fit with two data points of the previous section. The errors in the predictions for 1959 and 1974 are well below 3% while the error for the other years after 1974 are much smaller than

what we got previously using only the data of 1959 and 1974. However, the accuracy of the prediction deteriorates as we move further away from the stages involved in the least square solution. While census data recorded 1 billion of people in 1804, the least square fit solution predicts only slight more than a quarter of a billion which is unacceptable as an approximation of the data. [It is of some interest to note that our linear growth model with the parameters as determined above predicts the world population to reach 10 billion by $1959 + 73 = 2032$ and 13.23 billion by 2050. On the other hand, United Nation's best estimate is that population will march past 9.3 billion by 2050 and exceed 10.1 billion by the end of the century with the qualification that these estimates could be far more, if birthrates do not continue to drop as they have in the last half-century.]

Table 6

$Year$	1804	1927	1959	1974	1987	1999	2011
n	-155	-32	0	15	28	40	52
y_n	0.28	1.93	3.18	4.02	4.93	5.95	7.18
$data$	1	2	3	4	5	6	7

Can we improve on the accuracy of (3.3) by using more data points from the U.S. Census Burean data for the least square fits? It could be a possible term project for those interested in learning more about statistical methodology. Below is an exercise in this direction:

EXERCISE 2. *Use the 5 data points for 1927, 1959, 1974, 1987, and 1999 to obtain a least square solution for the linear growth model and compare the results for 1804 and 2011 from this solution with the ones in Table 6.*

The polynomial (3.10) may be written as

$$140A^{-12} + 240A^{-24} + 60A^{-25} + \cdots\cdots = 200A^{-12} + 168A^{-24} + 160A^{-25} + \cdots.$$

While higher order terms of the polynomial (3.10) in A may seem negligible, keeping only A^{-26} and neglecting higher order small terms does not give any root.

4. Immigration

Suppose we are interested only in the growth of the population of the United States (and not the world) during a period when the linear growth model is applicable. Let z_n be the U.S. population at the end of year n starting from some reference initial stage. In addition to linear growth of the form Az_n, we, as a nation of immigrants, traditionally admit a certain number of new immigrants each year, say q_n in year n.. The mathematical model in this setting is that of (3.2) augmented by the immigration:

$$z_{n+1} = Az_n + q_n \quad \text{or} \quad \Delta z_n = az_n + q_n \tag{4.1}$$

Given the population of some particular year set to be stage 0,

$$z_0 = Z, \tag{4.2}$$

(and the (scalar) amplification factor A), the first order difference equation (4.1) specifies the U.S. population for other years thereafter.

The linear difference equation (4.1) is a special case of the general first order linear inhomogenous difference equation

$$x_{n+1} = \mu_n x_n + q_n \tag{4.3}$$

where μ_n and q_n are known scalar constants and generally vary with n. Even if $q_n = \mathbf{0}$, the usual method (of assuming solution in the form cA^n) does not apply as long as μ_n varies with n. On the other hand, (4.3) is effectively a recurrence relations giving successive x_n in terms of the same quantity at earlier stages, it is straightforward to deduce the following conclusion:

PROPOSITION 1. *For $q_n = \mathbf{0}$, the unique solution of the IVP*

$$x_{n+1} = \mu_n x_n, \qquad x_0 = Z \tag{4.4}$$

is

$$x_n = Z\ \Pi_{k=0}^{(n-1)} \mu_k. \tag{4.5}$$

PROOF. The solution follows upon writing

$$\begin{aligned} x_n &= \frac{x_n}{x_{n-1}}\frac{x_{n-1}}{x_{n-2}} \cdot \cdots \cdot \frac{x_2}{x_1}\frac{x_1}{x_0}x_0 \\ &= \mu_{n-1}\mu_{n-2} \cdot \cdots \cdot \mu_2\mu_1\mu_0 x_0 = Z\ \Pi_{k=0}^{(n-1)}\mu_k. \end{aligned}$$

To prove uniqueness, suppose there were two solutions y_n and z_n with $y_n - z_n = x_n$. In that case, x_k satisfies the same difference equation (4.4) with $x_0 = y_0 - z_0 = 0$. It follows $x_1 = x_0\mu_0 = 0$ and if $x_n = 0$, then $x_{n+1} = x_n\mu_n = 0$ which completes the proof by induction. □

PROPOSITION 2. *The unique solution of the IVP*

$$x_{n+1} = \mu_n x_n + q_n, \qquad x_0 = Z \tag{4.6}$$

is

$$x_n = \ Z\Pi_{k=0}^{n-1}\mu_k + \sum_{i=0}^{n-1} q_i \left[\Pi_{j=i}^{n-1}\mu_{n-1-j}\right].$$

PROOF. We prove the simpler case of a constant $\mu_n = \mu$ (and leave the general case as an exercise). In this simpler case, we have

$$\begin{aligned} x_1 &= \mu x_0 + q_0 = \mu Z + q_0, \\ x_2 &= \mu x_1 + q_1 = \mu\left[\mu Z + q_0\right] + q_1 \\ &= \mu^2 Z + \mu q_0 + q_1, \end{aligned}$$

By induction, we get

$$\begin{aligned} x_n &= \mu^n Z + \mu^{n-1}q_0 + \mu^{n-2}q_1 + . \cdots \ + .\mu q_{n-2}. + \ q_{n-1} \\ &= Z\mu^n + \sum_{k=0}^{n-1}\mu^{n-1-k}q_k = \mu^n\left[Z + \sum_{k=0}^{n-1}\mu^{-(k+1)}q_k\right]. \end{aligned}$$

Uniqueness is proved as in the previous proposition. □

COROLLARY 1. *If both $\mu_n = \mu$ and $q_n = q$ do not vary with n, then*

$$x_n = Z\mu^n + q\frac{1-\mu^n}{1-\mu}$$

PROOF. (exercise) □

In terms of the U.S. population growth, the model with $\mu_n = \mu$ and $q_n = q$ shows that the model is only realistic if the amplication factor μ is less than 1. For that case, we have

$$x_n \to \frac{q}{1-\mu}$$

as $n \to \infty$. For $\mu > 1$, population size generally grows without bound as $n \to \infty$. For $\mu = 2$ for example, we have to a good approximation for large n

$$x_n = Z2^n + q\frac{1-2^n}{1-2} \simeq (Z+q)2^n.$$

From the discussion of the previous section, we know that the linear model (with or without immigration) ceases to be appropriate and a biologically more realistic model sought instead.

CHAPTER 2

Evolution and Equilibria

1. The Logistic Growth Model

Whether better predictive capacity of the linear growth model may be attained by a least square fit using more data points, it is important to examine the adequacy of that model from the perspetive of known reality, the phenomenon of population growth in our case. Recall that the linear model (3.2) of Chapter 1 was obtained from (1.2) or (2.2) under the assumption that the population size (or its change over a time period) y_n is "small" so that in (the Taylor polynomial approximation of) $F(y_n)$, terms proportional to $(y_n)^m$ for $m > 1$ are negligible compared to the term proportional to y_n and can be omitted. But the U.S. Census Bureau data show that population size increased by 100% between 1959 and 1974 and by 250% by 2011. Such *percentage changes* are not small and the use of linear model over a time span of the order of 52 years may not be appropriate without additional documentation.

While it may be argued that large percentage changes do not necessarily imply a violation of the small population requirement, indefinite (geometric) growth is physically and biologically unrealistic. The limited amount of living space, water and other resources can support only a certain maximum population size, known as the Earth's *carrying capacity*. Intuitively, when the current population is not small compared to the carrying capacity, we expect linear growth should cease to be appropriate and we should keep (at least) the quadratic term in (3.1) of Chapter 1. A two-term Taylor series approximation for $F(y_k)$ gives

$$y_{k+1} = Ay_k + By_k^2 + ... \simeq Ay_k - by_k^2 \tag{1.1}$$

where b is taken to be positive to reflect that fact that the effect of the quadratic term is to dampen the growth rate. It is an observed fact that when the population is "large", environmental or policy constraints generally slow down the growth rate. Cell growth in a Petri dish exemplifies the former and China's one child policy is an example of the latter.

Written as an equaton on the growth rate, the quadratic model (1.1) above becomes

$$y_{k+1} - y_k = ay_k - by_k^2 = ay_k \left(1 - \frac{y_k}{y_c}\right) \tag{1.2}$$

where $y_c = a/b$ is customary known as the *carrying capacity* of the environment. The quadratically nonlinear model in the form (1.2) or (1.1) is known as the *logistic growth* model in population dynamics. It has been used as a mathematical model in studies of many population growth phenomena, quite adequately in many cases.

2. Dimensionless Form

The quantity y_c in (1.2) has the same dimension and unit of measurement as the population size y_k of the k^{th} stage. The second term inside the parentheses is therefore dimensionless, i.e., a pure number with no physical or biological units, corresponding to the ratio of two population sizes . The nonlinear difference equation (1.2) itself may be made completely dimensionless by dividing both sides of the equation by y_c and setting $x_k = y_k/y_c$ to get

$$x_{k+1} = Ax_k - ax_k^2, \qquad \text{with} \qquad A = 1 + a. \tag{2.1}$$

In the process, we have reduced the number of parameters in the difference equation model from two to one, simplifying the analysis of the equation considerably.

It is important to note that the logistic growth equation has been nondimensionalized in other ways. One quite popular way is to write (1.1) as

$$y_{k+1} = Ay_k - by_k^2 = Ay_k\left(1 - \frac{y_k}{y_e}\right) \tag{2.2}$$

where $y_e = A/b$ is also a population size measure (with $y_{k+1} = 0$ if $y_k = y_e$). By setting $z_k = y_k/y_e$, we obtain

$$z_{k+1} = Az_k(1 - z_k) \tag{2.3}$$

which is the form used in [**3**] (where r is used instead of A for the amplification factor of linear growth).

Starting with a prescribed initial population $y_0 = Y$, the population size of any future year can be obtained by repeated applications of the nonlinear difference equation (2.1). The process gives

$$\begin{aligned}
y_1 &= y_0\,(A - by_0) = Y\,(A - bY) \\
y_2 &= y_1\,(A - by_1) = Y\,(A - bY)\,[A - bY\,(A - bY)] \\
y_3 &= Y\,(A - bY)\,[A - bY\,(A - bY)]\,\{A - bY\,(A - bY)\,[A - bY\,(A - bY)]\} \\
&\quad \ldots\ldots
\end{aligned}$$

where $b = a/y_c$. However, the results are unlike those for the linear model where the future population is succinctly summarized by the simple formula (3.3) of Chapter 1 and simple interpretations (in terms of the initial population Y and the growth factor A) can be given as we did in the three items following (3.3). Evidently, not much insight on the evolution of the population can be gained from this method of solution for the nonlinear logistic growth model except to calculuate the numerical values of population size at any specific stage if the parameters A and b are known numerically.

In this chapter, we develop another approach to obtain several types of results for the logistic growth model and other nonlinear first order difference equation models. The new approach is principally geometric in nature, in contrast to the algebraic methods of the previous chapters They are applicable mainly to autonomous difference equations of the form (1.1) where F does not vary with n except through y_n. For the logistic growth model, this requires that A and b not to vary with n, and allows us to re-scale the variable y_n to get (2.1) by setting $x_k = y_k/y_c$.

3. Fixed Points and Their Stability

The key to more useful information for a nonlinear difference equation such as (2.1) is to locate its *equilibrium* solutions. As in the physics of motion, equilbrium configuration of an evolving population is a state of the population that does not change with time. If the population is in that particular state, it will remain in that state forever. Mathematically, this means $y_n = \bar{y}$ for all n. Since any population size is subject to the same growth dynamics, the particular equilibrium configuration $y_n = \bar{y}$ must also satisfy the same difference equation so that

$$\bar{y} = F(\bar{y}) \tag{3.1}$$

In other words, $\bar{y}$ is a root of the nonlinear equation $F(\bar{y}) - \bar{y} = 0$.

For the linear equation (3.3) of Chapter 1 with $A \neq 1$, the only fixed point of the equation is at the origin: $\bar{y} = 0$. For the re-scaled (dimensionless) logistic growth model, the difference equation (2.1) requires a fixed point $x_n = \bar{x}$ to satisfy the algebraic equation

$$\bar{x} = A\bar{x} - a\bar{x}^2, \qquad A = 1 + a$$

for which there are two roots:

$$\bar{x}^{(1)} = 0, \qquad \bar{x}^{(2)} = 1.$$

If population size should be at the level of either of these fixed points at any time, the population would remain in that state forever thereafter.

For other growth function $F(y_n)$, not all critical points can be found explicitly. However, many fast root finding computer software (in Mathematica, Maple or MatLab, etc.) are available for accurate numerical solutions of a single equation of the form $F(t) = t$ or equivalently $F_e(t) \equiv F(t) - t = 0$

Approximate location of fixed points can also be obtained by graphical methods. For example, the two critical points for $F_e(t) = 1 - t - e^{-kt}$ with $k \gg 1$ can be found graphically to be

$$\bar{t}^{(1)} = 0, \qquad 0 < \bar{t}^{(2)} \lesssim 1.$$

by plotting $F_e(t)$ vs. t and looking for locations where the graph crosses the abcissa (the horizontal $t-$axis).

For a very complicated $F(t)$, a different use of graphical methods may be more effective. For example, we can break up the function into two pieces $F(t) = g(t) - h(t)$, plot the two simpler functions $g(y)$ and $h(y)$ and locate their intersections since $F(t) = 0$ corresponds to $g(t) = h(t)$. It may not be so easy to visualize the graph of $F_e(t) = 1 - t - e^{-kt}$ or estimate where it crosses the abscissa, if it crosses at all. But we have no problem visualizing $h(t) = e^{-kt}$ and $g(t) = 1 - t$ and estimating the value $\bar{t}^{(c)}$ of their intersection. Note that there are usually alternative ways to split $F_e(t)$. For example, we could have taken $g(t) = 1 - e^{-kt}$ and $h(t) = t$ instead. While there is not much difference in the two ways in splitting the relative simple $F_e(t)$, we generally should be flexible in exploring more than one options to find an effective splitting to get adequate information about the intersections of the two resulting graphs.

When the initial population Y is not perfectly at a fixed point, the population size of subsequent stages would change with n. Even when the population has been at a (time independent) steady state exactly, some environmental (such as a serious epidemics) or social (such as a devastating war) event may cause a perturbation

from that fixed point. As the perturbed population evolves with n, it is of considerable important to know if it would tend toward the unperturbed fixed point or move away from it. In the first case, the fixed point is said to be (asymptotically) *stable* while in the second case, it is *unstable*. (In some special case, there may be no movement in either direction and the fixed point is said to be *neutrally stable*.) For a planned population, returning to the previous equilibrium size is likely to be desirable; but for replenishing of a depleted commercial fish stock, moving away from a low level equilibrium population size would definitely be preferred.

In population growth models, fixed points correspond to possible steady state populations that do not change with time. In particle dynamics, they correspond to equilibrium configurations or equilibrium state such as a rigid ball at the peak of a (concave) hill top or the same ball at the (convex) bottom of a trough. The two particular examples show that equilibria can be significantly different in nature. At the slightest disturbance, the ball on top of the hill would roll down the hill, running away from its equilibrium position. By contrast, the ball at the bottom of a trough would, upon displacement from its critical point, merely oscillate about its original position. We characterize the trough bottom equilibrium configuration as a stable equilibrium or *stable fixed point* and the hill top configuration as unstable equilibrium or *unstable fixed point*. Depending on the surface of the trough, whether it is smooth or rough (causing frictional resistance), the oscillating ball about the trough bottom may continue to oscillate with the same maxiumum amplitude indefinitely or the oscillation may diminish and eventually return to the resting configuration. In the latter case, the fixed point is said to be *asymptotically stable* and is called an *attractor* in the lingo of dynamical systems. In the former case, the fixed point is (*neutrally*) *stable but not asymptotically stable*.

The situation is not different for biological phenomena. For population growth models, the stability of an equilibrium (steady state) population is also very important in policy, design or medical treatment decisions. An asymptotically stable fish population of adequate size would probably be good for a fishing industry but an asymptotically stable high virus population would be disastrous for a sick patient. It is therefore important to have methods for determining the fixed points and their stability. To the extent that evolution of the dynamical system in question will tell us a great deal about the stability of the fixed points, we will exploit this approach to investigate stabilty in the next section while an analytical approach to stability will be described in the next chapter.

4. Cobweb Graphs

To gain more insight to the solution of the logistic growth equation, we employ a graphical method for generating, recording and tracking successive x_n computed using (2.1). For this purpose, we plot the graph of the growth function $F(x) = x(A-ax)$ with $y = F(x)$ being an upside down parabola crossing the $x-$ axis at the point $\bar{x}^{(1)} = 0$ and $x^{(2)} = A/a$. The parabola is above the abscissa ($y = F(x) > 0$) for $0 < x < A/a$ and below otherwise. We also graph the straight line $y = x$ which intersects the graph of $F(x)$ at $x = 0$ and $x = 1$ (why?). For $A > 1$, the graph $y = F(x)$ is above the straight line $y = x$ for $0 < x < 1$.

Given the initial population $x_0 = \bar{x}_0$ ($= Y/y_c$), we get $x_1 = \bar{x}_0(A-a\bar{x}_0) = F(\bar{x}_0)$ which is a point on the graph for $F(x)$ right above $\bar{x}_0$. To get x_2 computationally, we use x_1 as the argument in $F(\cdot)$. To get the same result graphically for $A > 1$,

we need to locate x_1 on the abcissa (the x - axis) so that we can look for the point $F(x_1)$ on a graph of $F(x)$ above it (giving $x_2 = F(x_1)$). The easiest way to do this without any calculation or measurement to locate x_1 on the abcissa is to move from the point $(\bar{x}_0, F(\bar{x}_0)) = (\bar{x}_0, x_1)$ on the graph of $F(x)$ horizontally until we reach the straight line graph of $y = x$ (Note that y here is just the ordinate of the Cartesian axes and has nothing to do with the unnormalized population size y_n). Drop from there down to a point on the abcissa which would be the location of the value x_1 on that axis. We now repeat the process to get x_2 by locating the point on the graph of $F(x_n)$ right above x_1. Having the point x_2 , we head horizontally toward the line $y = x_n$. Once reaching a point on that line, move vertically (up or down) find $x_3 = F(x_2)$, etc. Note that in this last sequence, we omitted the unessential step of dropping from the $y = x$ line to abcissa before go up to the graph of $F(x)$ to locate x_3, etc.

The entire graphical process is shown schematically in Figure 1 below for a typical initial condition of $\bar{x}_0 = 0.2$. The resulting polygonal path from $(\bar{x}_0, F(\bar{x}_0))$ toward $(1, 1)$ is known as the *cobweb graph* or simply the *cobweb* of the difference equation $x_{n+1} = F(x_n)$.

Figure 1

4.1. Summary of the Graphical Process. To summarize, the graphical process for constructing the cobweb graph consists of the following sequences of steps for a growth function $F(x)$ $(= x(A - ax)$ for our particular application) with x and y denoting the abscissa and ordinate, respectively, of the (two-dimensional) Cartesian coordinate system:

(1) Draw the graph of $y = F(x) = x(A - ax)$. (If it is helpful to be concrete, take $A = 1.5$ and $a = A - 1 = 0, 5)$
(2) Draw also the straight line $y = x$ and mark its two intersection with $y = F(x)$ at the points $(0, 0)$ and $(1, 1)$.
(3) Start from $\bar{x}_0$ along the $x-$ axis, i.e., the point $(\bar{x}_0, 0)$, move vertically toward the graph of $y = F(x)$ to locate the point $(\bar{x}_0, F(\bar{x}_0))$.
(4) Move from $(\bar{x}_0, F(\bar{x}_0))$ horizontally to get to the straignt line (the graph of $y = x$) at the point $(F(\bar{x}_0), F(\bar{x}_0))$ on that line. The x coordinate of that point is x_1 since $x_1 = F(x_0) = F(\bar{x}_0)$.
(5) From the point $(x_1, x_1)) = (F(\bar{x}_0), F(\bar{x}_0))$ on the straight line $y = x$, move up vertically to locate the point $(x_1, F(x_1))$ on the graph $y = F(x)$.
(6) With $x_2 = F(x_1)$, repeat the process by moving from $(x_1, F(x_1)) = (x_1, x_2)$ on the graph $y = F(x)$ horizontally to get to the straightline to locate (x_2, x_2) and then vertically to the graph of $y = F(x)$ to locate $(x_2, F(x_2)) = (x_2, x_3)$, etc.

4.2. Application to the Logistic Growth Equation. We now apply the graphical method above to the logistic growth model with $A > 1$ and find the following results from the cobweb graph constructed:

- For $0 < \bar{x}_0 < 1$, the result is qualitatively the same as the one shown in Figure 1 if the maximum of the downward parabola is attained at $x_{\max} > 1$. The graphical method in this case traces out an ascending stair case moving up and to the right toward the fixed point $(1, 1)$, approaching the fixed point in smaller and smaller steps (but never quite gets there in any finite number of stages). If $x_{\max} < 1$, the cobweb generated by

the graphical porcess is more complicated and will be discussed in more detailed in the next section.

- For $\bar{x}_0 < 0$, the graph $y = F(x)$ lies below the straight line $y = x$. Consequently, the cobweb moves away from the fixed point $(0, 0)$ with $x_n \to -\infty$ decreaseing monotonically toward $-\infty$.

  Together, they suggest (but do not prove) that the fixed point $x^{(1)} = 0$ is unstable.
- For $\bar{x}_0 > 1$, we note that the graph for $F(x)$ is below the straight line $y = x$ for $x > 1$. We again limit discussion here to the case $x_{\max} > 1$ (and postpone discussion of $x_{\max} < 1$ until the next section). Since $F(\bar{x}_0) < \bar{x}_0$ (given the graph of $F(x)$ is below the line $y = x$ in the neighborhood of $x = \bar{x}_0$), the cobweb construction process requires that we move left horizontally from the point $(\bar{x}_0, F(\bar{x}_0))$ to get to a point on that line with coordinates $(F(\bar{x}_0), F(\bar{x}_0)) = (x_1, F(\bar{x}_0))$. At that point we must move vertically downward to get to the graph of $F(x)$ for the value for x_2. As we continue this graphical method, the horizontal and vertical movements are both monotone decreasing, approaching the fixed point $(1, 1)$ as $n \to \infty$.

  We conclude (or more correctly, infer) from these observations that the fixed point $x^{(1)} = 0$ is unstable and $x^{(2)} = 1$ is (asymptotically) stable.

4.3. The Case 0 < A < 1. For this case, the entire graph of $y = F(x)$ lies below the straightline $y = x$. In that case, the same graphical construction gives a polygonal path that lies below y = x and heads toward the origin. We infer from this that the fixed point $x^{(2)} = 1$ is unstable and $x^{(1)} = 0$ is (asymptotically) stable.

4.4. The Linear Growth Case. For a linear growth function $F(x) = Ax$ investigated in Chapter 1, the same graphical technique shows that the only fixed point $x^{(1)} = 0$ is unstable if $A > 1$ and is asymptotically stable if $A < 1$. If $A = 1$, then every initial state $\bar{x}_0$ is a neutrally stable fixed point.

CHAPTER 3

Linear Stability Analysis

1. The Mean Value Theorem

The graphical method for constructing cobwebs of a dynamical system provides us with some indication of the stability of its fixed points for some specific growth function $F(x)$. It does not prove the suggested behavior conclusively. In the next section, we describe an analytical method which proves the stability or instability of a fixed point. The validity of the method hinges on the mean value theorem from elementary calculus. Before we describe the general method of linear stability analysis, it is useful to review this important but somewhat abstract theorem to make it more tangible in the context of growth function.

Suppost $F(x)$ is continuous and continuously differentiable in the range of x of interest and the interval $[\alpha, \beta]$ is inside of this range. The mean value theorem asserts that

$$F(\beta) - F(\alpha) = F'(\gamma)(\beta - \alpha)$$

for some nunber γ inside the interval (α, β), i.e., $\alpha < \gamma < \beta$. The vagueness of the theorem lies in the somewhat unapecified number γ. The vagueness can be removed for any particular application as illustrated here for the example of the logistic growth function.

For the logistic growth function, we have $F(x) = Ax - ax^2$ with $F'(\gamma) = A - 2a\gamma$. Once the interval (α, β) is specified, we can determine γ from the requirement that

$$F'(\gamma) = \frac{F(\beta) - F(\alpha)}{\beta - \alpha}$$

or

$$A - 2a\gamma = \frac{A(\beta - \alpha) - a(\beta^2 - \alpha^2)}{\beta - \alpha} = A - a(\beta + \alpha)$$

which can be solved to get

$$\gamma = \frac{\beta + \alpha}{2}.$$

The explicit solution for γ confirm the assertion that $\alpha < \gamma < \beta$ since the average value of the two end points of the interval lies between them.

2. The Basic Stability Theorem

Before we investigate the stability of a fixed point $\bar{x}$ of the dimensionless dynamical system

$$x_{n+1} = F(x_n) \tag{2.1}$$

analytically, we need to give a mathematical definition of stability and instability.

DEFINITION 2. *A fixed point $\bar{x}$ of (2.1) is said to be **stable** if given any $\varepsilon > 0$, there is a $\delta > 0$ such that if $|\bar{x}_0 - \bar{x}| < \delta$, then $|x_n - \bar{x}_0| < \varepsilon$ for all $n > 0$.*

DEFINITION 3. *A fixed point $\bar{x}$ is said to be **asymptotically stable** or **attracting** if it is stable and there is a number $\delta > 0$ such that the sequence $\{x_n\}$ generated by any initial value $\bar{x}_0$ tends to $\bar{x}$, i.e.,*

$$\lim_{n\to\infty} x_n = \bar{x},$$

whenever $0 < |\bar{x}_0 - \bar{x}| < \delta$.

DEFINITION 4. *A fixed point $\bar{x}$ is said to be **unstable** or **repelling** if it is not stable. In other words, there is a number $\varepsilon > 0$ and some positive integer n such that the n^{th} iterate $x_n = F^n(\bar{x}_0) = F(F(...(F(F(\bar{x}_0)))...))$ of any initial point $\bar{x}_0$ inside the ε-nieghborhood of the fixed point $\bar{x}$, $0 < |\bar{x}_0 - \bar{x}| < \varepsilon$, lies outside the same ε neighborhood, i.e., $|x_n - \bar{x}| > \varepsilon$.*

We are now ready to state and prove the following theorem on local stability or instability of a fixed point:

THEOREM 3. *Suppose the growth function $F(\cdot)$ is continuous and continuously differentiable. A fixed point $\bar{x}$ of (2.1) is attracting/asymptotically stable if $|F'(\bar{x})| < 1$ and is unstable if $|F'(\bar{x})| > 1$. (Note that the theorem is silent on the special case $|F'(\bar{x})| = 1$ which will be discussed later.)*

PROOF. For an initial state $x_0 = \bar{x}_0$, we have from the mean value theorem □

$$x_1 - \bar{x} = F(\bar{x}_0) - \bar{x} = F(\bar{x}_0) - F(\bar{x}) = F'(c)(\bar{x}_0 - \bar{x})$$

for some constant c satisfying $0 < |c - \bar{x}| < |\bar{x}_0 - \bar{x}|$.

(i) For the case $|F'(\bar{x})| < 1$,we have from the continuity of $F'(\cdot)$ that $|F'(x)| < \alpha < 1$ for all x in the interval $|x - \bar{x}| < \beta$. In that case, we have

$$|x_1 - \bar{x}| = |F'(c)|\,|\bar{x}_0 - \bar{x}| < \alpha\,|\bar{x}_0 - \bar{x}| \tag{2.2}$$

for any $\bar{x}_0$ with $|\bar{x}_0 - \bar{x}| < \beta$ (and hence $|c - \bar{x}| < \beta$). Similarly,

$$x_2 - \bar{x} = F(x_1) - \bar{x} = F(x_1) - F(\bar{x}) = F'(d)(x_1 - \bar{x})$$

for some constant d satisfying $0 < |d - \bar{x}| < |x_1 - \bar{x}| < |\bar{x}_0 - \bar{x}|$. It follows that

$$|x_2 - \bar{x}| \le |F'(d)|\,|x_1 - \bar{x}| < \alpha\,|x_1 - \bar{x}| < \alpha^2\,|\bar{x}_0 - \bar{x}| \tag{2.3}$$

for any $\bar{x}_0$ with $|\bar{x}_0 - \bar{x}| < \beta$ (and hence $|d - \bar{x}| < \beta$). By induction, we have

$$|x_n - \bar{x}| < \alpha^n\,|\bar{x}_0 - \bar{x}|\,.$$

for any $\bar{x}_0$ with $|\bar{x}_0 - \bar{x}| < \beta$. It follows from $\alpha < 1$ that

$$\lim_{n\to\infty} x_n = \bar{x},$$

and the fixed point $\bar{x}$ is asymptotically stable.

(ii) For the case $|F'(\bar{x})| > 1$,we have from the continuity of $F'(\cdot)$ that $|F'(x)| > \gamma > 1$ for all x in the interval $|x - \bar{x}| < \varepsilon$. In that case, we have

$$|x_1 - \bar{x}| = |F(\bar{x}_0) - F(\bar{x})| = |F'(p)|\,|\bar{x}_0 - \bar{x}| > \gamma\,|\bar{x}_0 - \bar{x}| > |\bar{x}_0 - \bar{x}| \tag{2.4}$$

for any $\bar{x}_0$ with $|\bar{x}_0 - \bar{x}| < \varepsilon$ (and hence $|p - \bar{x}| < \varepsilon$). The fixed point $\bar{x}$ is unstable (and the proof is complete) if $|x_1 - \bar{x}| > \varepsilon$. If not, i.e., if $|x_1 - \bar{x}| < \varepsilon$, then

$$|x_2 - \bar{x}| = |F(x_1) - F(\bar{x})| = |F'(q)|\,|x_1 - \bar{x}| > \gamma\,|x_1 - \bar{x}| > \gamma^2\,|\bar{x}_0 - \bar{x}|\,.$$

for some constant q satisfying $0 < |q - \bar{x}| < |x_1 - \bar{x}| < \varepsilon$. Again, $\bar{x}$ is unstable if $|x_2 - \bar{x}| > \varepsilon$; otherwise, we repeat the process to get

$$|x_3 - \bar{x}| > \gamma^3 |\bar{x}_0 - \bar{x}| \tag{2.5}$$

etc., until $|x_n - \bar{x}| > \gamma^n |\bar{x}_0 - \bar{x}| > \varepsilon$ for some $n \geq 1$ proving that $\bar{x}$ is an unstable fixed point if $|F'(\bar{x})| > 1$.

EXAMPLE 1. *For the logistic growth model (2.1) of Chapter 2,*

$$x_{n+1} = Ax_n - ax_n^2,$$

we have for the first fixed point $\bar{x}^{(1)} = 0$,

$$F'(\bar{x}^{(1)}) = F'(0) = [A - 2az]_{z=0} = A.$$

where A *is necessarily a positive number. It follows Theorem 3 that* $\bar{x}^{(1)} = 0$ *is an unstable fixed point if* $A > 1$ *and is asymptotically stable if* $0 < A < 1$.

For the other fixed point $\bar{x}^{(2)} = 1$, *we have (with* $A = 1 + a$*)*

$$F'(\bar{x}^{(2)}) = F'(1) = [A - 2az]_{z=1} = 1 - a.$$

If follows from Theorem 3 that $\bar{x}^{(2)} = 1$ *is an unstable fixed point if* $a < 0$ *and is asymptotically stable if* $0 < a \ (< 1)$.

3. The Inconclusive Case of $F'(\bar{x}) = 1$

Theorem 3 has at least the two limitations. The obvious one is that it avoids dealing with the case $|F'(\bar{x})| = 1$. We address this limitation in this section. Consdier first the case $F'(\bar{x}) = 1$. The mean value theorem, which is the same as a two-term Taylor's theorem, is not sufficiently informative for this case. We need to seek additional information by looking at addtional terms in the Taylor polynomial, assuming $F(\cdot)$ is constinuously differentiable to the order needed for our development.

Suppose the initial data $\bar{x}_0$ is not a fixed point but $\bar{x}_0 = \bar{x} + \varepsilon$ for some nonzero constant ε. By a three term Taylor's theorem (including the remainder), we have

$$F(\bar{x}_0) = F(\bar{x} + \varepsilon) = F(\bar{x}) + F'(\bar{x})\varepsilon + \frac{1}{2}F''(c)\varepsilon^2.$$

for some value c with $0 < |c - \bar{x}| < |\bar{x}_0 - \bar{x}|$. .For $F'(\bar{x}) = 1$, we have

$$x_1 - \bar{x} = \varepsilon \left[1 + \frac{1}{2}F''(c)\varepsilon\right] = (\bar{x}_0 - \bar{x}) \left[1 + \frac{1}{2}F''(c)\,(\bar{x}_0 - \bar{x})\right]$$

which, for a sufficiently small ε, shows the stability of the fixed point to depend on the sign of $F''(\bar{x})\,(\bar{x}_0 - \bar{x})$ (with $F''(c)\,(\bar{x}_0 - \bar{x})$ having the same sign by continuity).

Assuming ($F'(\bar{x}) = 1$ and) $F''(\bar{x}) \neq 0$, the nature of stability of the fixed point is characteristically different from the case $F'(\bar{x}) \neq 1$. For the latter, either $\bar{x}$ is (asymptotically) stable (when $F'(\bar{x}) < 1$) or unstable (when $F'(\bar{x}) > 1$) whether or not $\bar{x}_0 - \bar{x} > 0$. This is not the case if $F'(\bar{x}) = 1$. If $F''(\bar{x}) < 0$ and by continuity $F''(c) < 0$, we have

$$|x_1 - \bar{x}| = \eta\,|\bar{x}_0 - \bar{x}|$$

where

$$\eta = \left|1 - \frac{1}{2}\left|F^{''}(c)\right|(\bar{x}_0 - \bar{x})\right| \quad \begin{cases} < 1 & (\bar{x}_0 > \bar{x}) \\ > 1 & (\bar{x}_0 < \bar{x}) \end{cases}.$$

Hence, for $F'(\bar{x}) = 1$ and $F^{''}(\bar{x}) < 0$, the sequence $\{x_n\}$ generated by the difference equation (2.1) approaches the fixed point $\bar{x}$ if $\bar{x}_0 > \bar{x}$ but moves away from the fixed point if $\bar{x}_0 < \bar{x}$. The fixed point is said to be semi-stable from above.

On the other hand, for ($F'(\bar{x}) = 1$ and) $F^{''}(\bar{x}) > 0$ (and by continuity $F^{''}(c) > 0$ for a sufficiently small ε), we have

$$|x_1 - \bar{x}| = \eta\,|\bar{x}_0 - \bar{x}|$$

where

$$\eta = \left|1 + \frac{1}{2}F^{''}(c)\,(\bar{x}_0 - \bar{x})\right| \quad \begin{cases} > 1 & (\bar{x}_0 > \bar{x}) \\ < 1 & (\bar{x}_0 < \bar{x}) \end{cases}.$$

It follows that the fixed point $\bar{x}$ in this case is semi-stable from below. Together we have established the following theorem:

THEOREM 4. *If $F'(\bar{x}) = 1$, the fixed point $\bar{x}$ is semi-stable from above if $F^{''}(\bar{x}) < 0$ and semi-stable from below if $F^{''}(\bar{x}) > 0$.*

4. When $F'(\bar{x}) = 1$ and $F''(\bar{x}) = 0$

The approach to the other inconclusive case, $F'(\bar{x}) = -1$, is qualitatively different from that for $F'(\bar{x}) = 1$. It is most appropriately addressed when we consider second order difference equations. Here we prepare for this discussion (in Section 5 of Chapter 5) by considering the case when $F'(\bar{x}) = 1$ and $F^{''}(\bar{x}) = 0$ which is of interest on its own as it does occur frequently in application. An example is the simple difference equation

$$x_{n+1} = x_n - bx_n^3 \tag{4.1}$$

that has three fixed points at

$$\bar{x}^{(1)} = 0, \qquad \bar{x}^{(2)} = \sqrt{b}, \qquad \bar{x}^{(3)} = -\sqrt{b}$$

For $\bar{x}^{(1)} = 0$, we have

$$F'(0) = 1, \qquad F''(0) = 0, \qquad F'''(0) = -6b.$$

For such dynamical systems, we have the following stability theorem:

THEOREM 5. *Siuppose $\bar{x}$ ist a fixed point of (2.1) with $F'(\bar{x}) = 1$ and $F''(\bar{x}) = 0$. Then the fixed point is (asymptotically) stable if $F'''(\bar{x}) < 0$ and unstable if $F'''(\bar{x}) > 0$.*

PROOF. For $\bar{x}_0 = \bar{x} + \Delta$, we see from the Taylor expansion

$$F(\bar{x}_0) = F(\bar{x}) + \Delta + \frac{1}{3!}F'''(c)\Delta^3 \tag{4.2}$$

for some constant c with $0 < |c - \bar{x}| < |\bar{x}_0 - \bar{x}|$. For sufficiently small $|\Delta| > 0$, $F'''(c)$ has the same sign as $F'''(\bar{x})$. It follows from

$$\begin{aligned} |x_1 - \bar{x}| &= |\bar{x}_0 - \bar{x}| = |F(\bar{x}_0) - F(\bar{x})| \\ &= \left|1 + \frac{1}{3!}F'''(c)\Delta^2\right| |\Delta| \end{aligned}$$

(and analogous relations for $|x_n - \bar{x}|$) that $\bar{x}$ is unstable if $F'''(\bar{x}) > 0$ and asymptotically stable if $F'''(\bar{x}) > 0$. □

Theorem 5 will be needed to determine the stability of a fixed of (2.1) for the inconclusive case of $F'(\bar{x}) = -1$ and $F''(\bar{x}) \neq 0$ (tp be discussed in Section 5 of Chapter 5). Meanwhile, its application to the example (4.1) tells us that the fixed point $\bar{x}^{(1)} = 0$ is asymptotically stable if $b > 0$ and unstable if $b < 0$.

5. Relocation of a Fixed Point to the Origin

For a general growth function $F(\cdot)$, a fixed point $\bar{x}$ of F can be relocated to $x^{(1)} = 0$ by a change of variable $x = z + \bar{x}$ so that $G(z) = F(z + \bar{x})$ has a fixed point at $z = 0$. It is therefore possible to take the fixed point of any growth function to be at $x^{(1)} = 0$ with no loss in generality. In the study of the stability of a fixed point, we will relocate the particular fixed point $x^{(k)}$ being investigated to the origin whenever it simplifies the analysis or development. This will be done by a change of variable $x = z + x^{(k)}$so that we get $z^{(k)} = [z]_{x=x^{(k)}} = 0$ for any fixed k (with the other fixed points of the system relocated elsewhere). For the dynamical system (4.1), we can relocate the fixed point $\bar{x}^{(2)} = \sqrt{b}$ to the origin by the change of variable $x = z + \sqrt{b}$ resulting in

$$z^{(1)} = -\sqrt{b}, \qquad z^{(2)} = 0, \qquad z^{(3)} = -2\sqrt{b}.$$

6. Period Doubling and Cycles

A second important limitation of Theorem 3 is illustrated by the first order dynamical system:

$$x_{n+1} = F(x_n) = 3.2x_n(1 - x_n). \tag{6.1}$$

It has two fixed points at $\bar{x}^{(1)} = 0$ and $\bar{x}^{(2)} = 0.6875.... = 11/16$. As expected, the first fixed point is *unstable* since

$$F'(0) = 3.2 > 1.$$

For the second fixed point, we have

$$F'(11/16) = 3.2 - 6.4\frac{11}{16} = 3.2 - 4.4 = -1.2$$

so that $\bar{x}^{(2)}$ is also unstable by Theorem 3.

We have then the following situation:

- If $x_0 = \bar{x}_0$ is between the two fixed points and sufficiently close to $\bar{x}^{(1)} = 0$, i.e., $\bar{x}_0 = \bar{x}^{(1)} + \varepsilon$ (with $\varepsilon > 0$) then the population would grow (for the next year at least) since

$$x_1 - \bar{x}^{(1)} > x_0 - \bar{x}^{(1)} = \varepsilon > 0.$$

- If $x_0 = \bar{x}_0$ is again between the two fixed points but sufficiently close to $\bar{x}^{(2)} = 11/16$, i.e., $\bar{x}_0 = \bar{x}^{(2)} - \varepsilon$ (with $\varepsilon > 0$), then the population would decline (for the next year at least) since

$$\bar{x}^{(2)} - x_1 > \bar{x}^{(2)} - x_0 = \varepsilon > 0.$$

In the first scenario where the initial population size is slightly greater than $\bar{x}^{(1)}$, the population would move toward the other equilibrium state $\bar{x}^{(2)}$. But if x_k gets sufficiently close to that fixed point from below, we would be in scenario 2 with the population turning around and heading toward $\bar{x}^{(1)}$ (since that fixed point is repelling). On the other hand, if x_k should overshoot $\bar{x}^{(2)}$, i.e., $x_k > \bar{x}^{(2)}$ for some k, then we must have $x_{k+1} < \bar{x}^{(2)}$ since $F'(11/16) = -1.2$ amd we are back to a population below $\bar{x}^{(2)}$ and the previous observation for that case applies. Since $\bar{x}_0 \geq 0$ and there is nowhere else x_k can tend to (as there is no attractor for this problem), this suggests that x_n would (1) oscillate inside the interval between the two fixed points, or (2) bounceback and forth from one side of $\bar{x}^{(2)}$ and the other, or (3) alternating between modes (1) and (2).

For the example (6.1), the evolution of population size starting with an initial population 0f $\bar{x}_0 = 0.025$ can be calculated to get:

n	x_n	n	x_n
0	0.025000	11	0.513337
1	0.078000	12	0.799431
2	0.230131	13	0.513092
3	0.566947	14	0.799452
4	0.785658	15	0.513052
5	0.538878	16	0.799455
6	0.795163	17	0.513046
7	0.521211	18	0.799455
8	0.798560	19	0.513044
9	0.514758	20	0.799455
10	0.799303	21	0.513044

Evidently, the scaled population settles into a cyclic pattern oscillating between two fractions (0.513044... and 0.799455..) of the carrying capacity This steady flip-flopping between two distinct states is called a **two-cycle**.

Cycling among several different number of distinct states is also possible for other first orther dynamical systems. For example, it can be verified that the scaled logistic growth equation $x_{n+1} = 3.5x_n(1 - x_n)$ has a **four - cycle**:

$$(0.87499726..., 0.38281968..., 0.82694071..., 0.50088421...).$$

An N - cycle corresponds to cycling among N distinct states. More precisely, we have the following definition of an N-cycle:

DEFINITION 5. *A first order autonomous dynamical system $y_{n+1} = F(y_n)$ has an N - **cycle** if there exist N distinct states $y_c^{(1)}, ..., y_c^{(N)}$ such that 1) $y_c^{(k)} \neq y_c^{(i)}$ if $i \neq k$; 2) $F(y_c^{(k)}) = y_c^{(k+1)}$, $k = 1, 2, ..., N-1$, and 3) $F(y_c^{(N)}) = y_c^{(1)}$.*

While we are not in a position to state and prove a mathematical result on the existence of a cycle, it is natural to look for a cycle when all the fixed points of a dynamical system are unstable.

For the scaled logistic growth equation $x_{n+1} = F(x_n) = Ax_n(1 - x_n)$ with

$$x_{n+2} = F(x_{n+1}) = F(F(x_n)) \equiv F_2(x_n),$$

it is not difficult to verify that $\bar{x}^{(1)} = 0$ and $\bar{x}^{(2)} = (A-1)/A$ are fixed points of $x_{n+2} = F_2(x_n)$. In addition, there are two others:

$$\begin{pmatrix} \bar{x}^{(3)} \\ \bar{x}^{(4)} \end{pmatrix} = \frac{1}{2}\left[(1 + \frac{1}{A}) \pm \sqrt{\left(\frac{A-1}{A}\right)^2 - \frac{4}{A^2}}\right].$$

However, while $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ are also fixed points of the original equation $x_{n+1} = F(x_n)$, $\bar{x}^{(3)}$ and $\bar{x}^{(4)}$ are not. In this case, it can be verified that

$$F(\bar{x}^{(3)}) = \bar{x}^{(4)}, \qquad F(\bar{x}^{(4)}) = \bar{x}^{(3)}$$

so that the pair constitutes a two-cycle. The example suggests that we look for possible two-cycles among the fixed points of $F_2(x) \equiv F(F(x))$ and, more generally, for possible $k-$cycles among the fixed points of

$$F_k(x) = F^k(x) \equiv F(F(F....F(x))...))),$$

with $F_k(x)$ being the composite function resulting from the defining relation $F_k(x) = F_{k-1}(\ F(x)) = F_{k-2}(F(F(x))) = F_{k-2}(F_2(x))$.

Period doubling involves significant qualitative change in the solution behavior at some critical parameter value. As such it may be considered a kind of *bifurcation* phenomenon (to be discussed in the next chapter). However, the term bifurcation usually pertains to significant qualitative changes of the fixed points whether or not it involves the appearance of cycles. The emerging of cycles are more often discussed in connection with the phenomenon of period doubling,

CHAPTER 4

Bifurcation

The evolution of the population size governed by the logistic growth equation would be different for different sets values for the model parameter A. In general, the changes in the solution are expected to be small for small changes in the value of A. However, this may not be the case for certain special values of the model parameter. Even with a tiny change of the parameter at such a critical value, the solution of the equation is qualitatively and quantitatively different in a very significant way. The phenomenon of very significant (or even catastrophic) changes of the solution behavior associated with a very small change of model parameter values is known as *bifurcation* and the critical (model parameter) values where such phenomena take place are called *bifurcation points*. In the next few sections, we introduce, by way of specific examples, three canonical types of bifurcation for phenomena modeled by a single first order difference equation. These particular type of bifurcation are fundamental because they 1) occur most frequently, 2) also experienced by phenomena modeled by higher order difference equations, 3) other bifurcation phenomena are often either one of these in disguise or minor variations or a combination of some of them, and 4) only one other bifurcation type that is characteristically different from them but does not appear in phenomena modeled by a single first order difference equation.

1. Saddle-Node Bifurcation

Suppose that the re-scaled first order difference equation

$$z_{k+1} = Az_k(1 - z_k) \tag{1.1}$$

with $A > 1$ models the natural growth of a fish population in a fish farm or a regulated natural fishing ground. The model has an unstable equilibrium fish population $\bar{z}^{(1)} = 0$ and an asymptotically stable fish population $\bar{z}^{(2)} = 1$. Suppose also that the commercially valuable fish population is harvested at a constant rate h (in units of biomass per unit stage which is most reasonably taken to be a yearly fishing season). The evolution of the fish population being harvested is then determined by

$$x_{n+1} = Ax_n(1 - x_n) - h \tag{1.2}$$

instead of (1.1). The locations of the two fixed points $x_h^{(k)}$(equilibrium sizes of fish population) now depend on h ; we can determine them by solving the quadratic equation $x - Ax(1 - x) + h = 0$ to get

$$\begin{pmatrix} x_h^{(1)} \\ x_h^{(2)} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \bar{x} - \sqrt{\bar{x}^2 - 4h/A} \\ \bar{x} + \sqrt{\bar{x}^2 - 4h/A} \end{pmatrix}$$

with $\bar{x} = (A-1)/A$ and $0 < x_h^{(1)} < x_h^{(2)} < \bar{x} < 1$ for $0 < h < (A-1)^2/4A$. The two critical points coalesce into one for $h_c = (A-1)^2/4A$ and the difference equation has no equilibrium fish population for $h > h_c = (A-1)^2/4A$.

The same conclusions can be seen readily from the relative position of the graph of the growth curve $F(x_n) = Ax_n(1-x_n)$ and the horizontal line h as shown in Figure 5.

Figure 5: The fixed points of a uniformly harvested fish population.

Figure 5 immediately tells us that there are two critical points $x_h^{(1)}$ and $x_h^{(2)}$ for our fish harvesting model corresponding to the two intersections between the horizontal line and the upside down parabola $Ax(1-x)$. The graph of $Ax(1-x)$ is seen to be above the horizontal line h for y in the interval between the two critical points. We conclude immediately that $x_h^{(1)}$ is unstable while $x_h^{(1)}$ is asymptotically stable.

As h increases from 0 to $h_c = (A-1)^2/4A$, the horizontal line moves up toward to the peak of the growth rate graph $Ax(1-x)$ and the two critical points move toward each other. They coalesce when the line is tangent to the peak at $h = h_c$ with $x_h^{(1)} = x_h^{(2)} = (A-1)/2A$. For larger values of h, the line does not intersect the growth rate curve and the dynamical system does not have any fixed point in this range of the uniform harvest rate.

The change of fixed point count from two to none, transitioning at the *bifurcation point* $h_c = (A-1)^2/4A$ (for which there is only one fixed point), is known as a *saddle-node bifurcation* with respect to the bifurcation parameter h. It is succinctly summarized by the bifurcation diagram in Figure 6.

Figure 6: A saddle-node bifurcation

that plots the the graphs of the two fixed points $x_h^{(1)}$ and $x_h^{(2)}$ as functions of the bifurcation parameter h. It is customary to use thick solid curves to indicate (asymptotically) stable critical points and dashed or dotted curves for unstable critical points.

More generally, for a first order difference equation of the form

$$y_{n+1} = F(y_n, \mu) \tag{1.3}$$

with two fixed points $y^{(1)}(\mu)$ and $y^{(2)}(\mu)$ that depend on the bifurcation parameter μ.

DEFINITION 6. *The dynamical system characterized by the difference equation (1.3) is said to have a saddle-node bifurcation at the bifurcation point $\mu = \mu_c$ if its bifurcation diagram (plotting the graphs of the two fixed points as functions of the bifurcation parameter μ) is qualitatively similar to Figure 6 with the transition from no fixed point to two fixed points occurring at critical value μ_c of the bifurcation parameter.*

In a natural way, the definition of saddle-node bifurcation applies to dynamical systems modeled by single higher order difference equation and systems of first

order difference equations with a transition from two fixed points to no fixed point (or conversely) at some critical value of a bifurcation parameter.

An important observation on the harvested fish population problem should be made at this point. In the notation of (1.3), the tangent to the graph for $F(y;\mu)$ as a function of y is, at bifurcation, horizontal at the fixed point. At a bifurcation point, the slope of $F(y;\mu)$ is $F_{,y}(y;\mu_c)$ with $F_{,y}(y;\mu) = \partial F(y;\mu)/\partial y$. Horizontal tangency at a fixed point y_c at bifurcation corresponds to the two conditions

$$F_{,y}^{(c)} \equiv F_{,y}(y_c;\mu_c) = 0, \qquad F(y_c;\mu_c) = 0. \tag{1.4}$$

We will see that this phenomenon of horizontal tangent at a fixed point at bifurcation persists for the examples in the next two sections. When there are no available cues from various graphical methods to suggest the location of possible bifurcation points, the analytical conditions (1.4) provide a set of criteria for determining them.

2. Transcritical Bifurcation

By working with the difference equation (1.3), we avoid tying bifurcation phenomena to strictly to population growth model, though the latter was used to introduce the saddle-node bifurcation phenomenon. For a different type of bifurcation phenomena, consider

$$y_{n+1} = y_n(\mu + 1 - y_n) \equiv f(y_n, \mu). \tag{2.1}$$

The difference equation also has two fixed points, but now at $y_c^{(1)} = 0$ and $y_c^{(2)} = \mu$. For a positive value of μ in the range $0 < \mu < 2$, the fixed point $y_c^{(2)} = \mu$ is located on the abscissa of the graph of $f(y;\mu)$ vs. y to the right of the origin (the positive y axis). With

$$\frac{\partial f}{\partial y} = \mu + 1 - 2y,$$

we know by Theorem 3 that $y_c^{(1)} = 0$ is unstable and $y_c^{(2)} = \mu$ is asymptotically stable.

As μ decreases but remaining positive, the location of $y_c^{(2)} = \mu$ moves toward the origin while the stability of both fixed points remain unchanged. All the while, the maximum $f(y)$ decreases until $\mu = \mu_c = 0$. For the particular value $\mu_c = 0$, the entire graph of $f(y)$ lies below the y axis but tangent to the abscissa (the $y-$axis) at the origin as shown in Figure 3(a). Further reduction of μ (to negative values) moves the graph further to the left and raises a portion of it above the $y-$axis so that it again crosses the abscissa at two locations. As we can see from the Figure 3(b), one critical point is again at the origin $y_c^{(1)} = 0$ but now the other one, $y_c^{(2)} = \mu\ (< 0)$, is on the negative y axis, illustrating how the location of fixed points may depend on the system parameters.

Figure 3: (a) graph of $f(y;\mu_c)$; (b) graph of $f(y;\mu)$ for $\mu < 0$.

In addition to changing the location of (at least one of) the fixed points, changing μ also affects the stability of the fixed points. For $0 < \mu < 2$, $y_c^{(1)} = 0$ is unstable and $y_c^{(2)} = \mu$ is asymptotically stable. But for $-2 < \mu < 0$, $y_c^{(1)} = 0$ is asymptotically stable and $y_c^{(2)} = \mu$ is unstable. The stability switch takes place at $\mu_c = 0$ where the two fixed points coalesce into one semi-stable fixed point.

Evidently, the dynamical system experiences a bifurcation at $\mu_c = 0$ since its fixed points undergo a change in count, locations, and stability as the bifurcation parameter μ passes through a threshold value. The pivotal value μ_c of the model parameter μ where these changes take place is called a *bifurcation point* of the dynamical system.

For the bifurcation of the dynamical system (2.1), the number of fixed points changes from two to one to two as μ passes through $\mu_c = 0$; the two fixed points also exchange their stability type. This type of bifurcation is called a *transcritical bifurcation*. The behavior of the dynamical system with this type of bifurcation is succinctly summarized by a bifurcatiaon diagram that graphs the two critical points $\{y_c^{(1)}(\mu), y_c^{(2)}(\mu)\}$ as functions of the bifurcation parameter μ as shown in Figure 4. Again, we use thick solid curves to indicate (asymptotically) stable critical points and dashed or dotted curves for unstable critical points.

Figure 4: A transcritical bifurcation.

3. Pitchfork Bifurcation

Transcritical bifurcation is distinct from saddle-node bifurcation. The latter involves a change of the number of fixed points from two to (one to) none. Since all fixed points disappear after bifurcation, nothing can be said about stability type changes after bifurcation. In contrast, the number of fixed points does not change after a transcritical bifurcation, but the stability of all the fixed point change type after bifurcation. The example

$$y' = y(\mu - y^2) \tag{3.1}$$

shows a third kind of bifurcation involving a change in both the count and the stability type of the fixed points involved. For this example, we reverse our previous practice of first deducing the desired results by an analytical method and then confirming them by working with the graphs of the "growth function" $f(y;\mu)$. Here, we first use the graphical method to get the desired results.

For $\mu > 0$, the graph of $f(y;\mu) = y(\mu - y^2)$ is as shown in the Figure 7. There are three isolated fixed points at $y_c^{(1)} = -\sqrt{\mu}$, $y_c^{(2)} = 0$, and $y_c^{(3)} = \sqrt{\mu}$ with the one at the origin being unstable while the other two are asymptotically stable.

Figure 7: Growth rate $f(y;\mu) = y(\mu - y^2)$ vs. y for $\mu > 0$.

For $\mu < 0$, we have $f(y;\mu) = -y(|\mu| + y^2)$ so that its graph crosses the y axis only at the origin as shown in Figure 8. Hence, there is only one critical point $y_c = 0$ and it is asymptotically stable. The transition from three to one critical point occurs at the bifurcation point $\mu_c = 0$ for which the graph of $f(y;\mu_c) = -y^3$ is similar to the one for $\mu < 0$ except that it is tangent to the y axis with $f_{,y}(y_c;\mu_c) = 0$ (and $f_{,yy}(y_c;\mu_c) = 0$ as well).

Figure 8: Growth rate $f(y;\mu) = y(\mu - y^2)$ vs. y for $\mu < 0$.

As in the previous two examples, the growth rate curve at the fixed point again has a horizontal tangent at bifurcation. The various changes associated with bifurcation for this example are different from those of the other two types. There is a change in the number of fixed point and also a change in the stability of the fixed points at the same location before and after bifurcation. These changes are succinctly summarized by the bifurcation diagram in Figure 9.

Figure 9: A pitchfork type bifurcation.

4. Other Types of Bifurcation

The three kinds of bifurcation introduced through the three previous examples in no way exhaust the variety of possible bifurcation types for dynamical systems characterized by a single first order difference equation. However, bifurcation which appears to be different may merely be (i) a basic type in disguise, (ii) a composites several basic bifurcation types, or (iii) their cosmetic variations. Discussion of truly new types of bifurcation including the important Hopf bifurcation and others occurring less frequently in applications can be found in text on dynamical systems (e.g., chapters 3 and 8 of [**17**]).

5. Exercises

EXERCISE 3. *A bacteria population of (initial size) 1000 grows steadily with a 50% increase every hour.*

a) What is the population afer 10 hours?
b) What is the bacteria population after 10 hours if the bacteria population grows from an initial size of 1000 in such a way that the population after $n+1$ hours is $(n+3)/(n+2)$ times the population after n hours instead?

EXERCISE 4. *A person takes a pill containing 200 milligrams of a drug every 4 hours. The drug goes into the bloodstream immediately and the body eliminates 20% of the drug that is in the bloodstream every four hours.*

a) Develop a difference equation model for the amount $x(n)$ (in milligrams) of the drug in the bloodstream after taking the nth pill.
b) Calculate $x(n)$ for $n = 1, 2, 3, 4$, and 5 given the inital bloodstream drug concentration $x(0) = 500$.
c) Determine the equilibrium drug concentration in the body, if it exists.

EXERCISE 5. *When left alone, a fish population in a fish farm doubles (in fish biomass) every six months. But the fishery management harvest h (in units of 1,000 lbs) of fish every six month to be sold for profit.*

a) Develop a difference equation model for the fish biomass $x(n)$ (in units of 1,000 lbs) in the fish farm.
b) For $h = 3$, find the equilibrium fish biomass x_e if it exists.
c) What is the maximum sustainbable yield $h_{\max}$ (maximum value of h that can be sustained indefinitely) for this fish farm?

EXERCISE 6. *Consider again the fish farm above.*

a) Calculate $x(n)$ for $n = 1, 2$, and 3 for $x(0) = x_e + 0.1$
b) Repeat a) for $x(0) = x_e - 0.1$
c) Is x_e stable or unstable?

CHAPTER 5

Second Order Models

1. Fibonacci Rabbits

1.1. The Question. Around year 1202, Leonardo Pisano Fibonacci, considered by some to be the greatest European mathematician of the middle ages, was said to have posed the following problem on the growth of certain rabbit popuplation: "A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on bcomes productive?" We consider here a simpler version of this problem by restricting the production of rabbits to one pair of offsprings per pair for each of the two breeding seasons after the season of the pair's birth (and no further production beyond that). Assuming that all rabbits survive, how many pairs of *new born rabbits* will there be after k generations (seasons)?

1.2. The Model. Since we are asking for a count of new born rabbits only, the problem is relatively easy. Let R_n be the number of **new** born rabbits in year n with $n = 0$ being the year the first pair of rabbits was put in place. Then the new born rabbits in year n should consist of new born rabbits of the year before plus the new borns the year before that (since each of these is to produce a new pair). We can write this as a mathematical relation:

$$R_{n+2} = R_{n+1} + R_n. \tag{1.1}$$

While (1.1) does not tell us how many new borns there will be in year k, we can use it to calculate R_n successively starting from $n = 0$ provided that we know R_0 and R_1. Note that unlike the previous population growth model, the present model gives the population of year n in terms of the populations of two immediate past years, not just one previous year as in (1.2). As such, it is called a *second order difference equation* (while (1.2) is known as a *first order* difference equation).

1.3. A Single Linear Difference Equation. More generally, the solution of a difference equation of order m may be written in as

$$y_{n+m} = F_n(y_{n+m-1}, y_{n+m-2}, ..., y_{n+1}, y_n). \tag{1.2}$$

A dynamical system modeled by an m^{th} order difference equation requires the specification of m initial conditions, usually taken in the form

$$y_k = Y_k \qquad (k = 0, 1, 2, ..., m-1) \tag{1.3}$$

for m (known) prescribed constants $\{Y_k\}$. The m conditions in (1.3) enable us to use (1.2) to calculate y_m which in turn enables us to calculate y_{m+1} and more generally y_{m+k} for $k = 0, 1, 2, 3, ...$, recursively. The following mathematical result is important for both theory and applications of a single linear differemce equations.

THEOREM 6. *The IVP for the (autonomous) m^{th} order linear difference equation*

$$y_{n+m} = a_{m-1}y_{n+m-1} + a_{m-2}y_{n+m-2} + \ ..... \ + a_0 y_n, \tag{1.4}$$

where $a_0, \ ....., a_{m-1}$ are known constants that do not vary with n, and the m initial conditions

$$y_k = Y_k, \qquad k = 0, 1, 2, \cdots, m-1 \tag{1.5}$$

(for prescribed values of $\{Y_k\}$) has one and only one solution/ (In the language of difference equation, the solution of the IVP (1.4)-(1.5) exists and is unique.)

PROOF. The result is a straightforward consequence of the recursion inherent in (1.4). We formalize the argument in the form of a proof by induction. By (1.4), the m prescribed initial conditions determines a unique y_m,

$$y_m = a_{m-1}y_{m-1} + a_{m-2}y_{m-2} + \ ..... \ + a_0 y_0.$$

Suppose $\{y_i\}$ have been determined uniquely for $i = 0, 1, 2, ..., m+k-1$. Then (1.4) gives a unique y_{m+k}

$$y_{m+k} = a_{m-1}y_{m+k-1} + a_{m-2}y_{m+k-2} + \ ..... \ + a_0 y_k.$$

□

THEOREM 7. *The solution of (1.4)-(1.5) where the coefficients $\{a_k\}$ may vary with the index n exists and is unique.*

PROOF. The proof is similar to the previous theorem. □

These observations provide the theoretical basis for computing solutions of linear difference equations.

2. Fibonacci Sequence

2.1. Evolution of the Fibonacci Rabbit Population. Returning to our simplified Fibonacci rabbits problem, we know $R_0 = 1$, since generation (or stage) 0 was the season that the first pair of rabbits was put in place. In season 1, only the pair of rabbits from generation 0 begets a **new** pair; hence, we have $R_1 = 1$ also. We can then apply the formula (1.1) after that with

$$\begin{aligned} R_0 &= 1, \qquad R_1 = 1, \qquad R_2 = R_1 + R_0 = 1 + 1 = 2, \\ R_3 &= R_3 + R_3 = 2 + 1 = 3, \qquad R_4 = R_3 + R_2 = 3 + 2 = 5, \\ R_5 &= R_4 + ... = 8, \qquad R_6 = .R_5 + ... = 13, \ ..... \end{aligned}$$

giving the famous *Fibonacci number* sequence $\{1, 1, 2, 3, 5, 8, 13, 21, ...\}$. With the computing capacity available to us on a laptop computer, it is straightforward to generate the number of new born rabbit pairs in the n^{th} generation for any finite number n using MatLab or any other scientific computing software (such as Maple or Mathematica).

Fibonacci numbers appear in many applications and many in the biological sciences (see the reference by Douady and Couder below). Some examples are:

1) branching in trees,
2) arrangement of leaves on a stem,
3) the fruitlets of a pineapple <http://en.wikipedia.org/wiki/Pineapple>,
4) the flowering of artichoke <http://en.wikipedia.org/wiki/Artichoke>,

5) an uncurling fern,
6) the arrangement of a pine cone <http://en.wikipedia.org/wiki/Pine_cone>.
7) the Yellow Chamomile head showing the arrangement in 21 (blue) and 13 (aqua) spiral
<http://upload.wikimedia.org/wikipedia/commons/5/5a/FibonacciChamomile.PNG>

Applications of Fibonacci numbers in biology are not restricted to plants or the botanical science. For example; they also appear in the description of the family tree of honeybees idealized according to the following rules:

- If an egg is laid by an unmated female, it hatches a male or drone bee.
- If, however, an egg was fertilized by a male, it hatches a female.

Thus, a male bee will always have one parent, and a female bee will have two. If one traces the ancestry of any male bee (1 bee), he has 1 parent (1 bee), 2 grandparents, 3 great-grandparents, 5 great-great-grandparents, and so on. This sequence of numbers of parents is the Fibonacci sequence. The number of ancestors at each level, F_n, is the number of female ancestors, which is F_{n-1}, plus the number of male ancestors, which is F_{n-2}. (All these are under the rather unrealistic assumption that the ancestors at each level are otherwise unrelated.)

Some references on other Fibonacci phenomena are:

- S. Douady and Y. Couder (1996). "Phyllotaxis as a Dynamical Self Organizing Process" (PDF). Journal of Theoretical Biology 178 (178): 255–274.doi:10.1006/jtbi.1996.0026. <http://en.wikipedia.org/wiki/Phyllotaxis>
- Jones, Judy; William Wilson (2006). "Science". An Incomplete Education. Ballantine Books. p. 544. ISBN 978-0-7394-7582-9.
- A. Brousseau (1969). "Fibonacci Statistics in Conifers".Fibonacci Quarterly (7): 525–532.
- "Fibonacci Flim-Flam". <http://www.lhup.edu/~dsimanek/pseudo/fibonacc.htm>
- "Marks for the da Vinci Code: B–". Computer Science For Fun: CS4FN.
- Prusinkiewicz, Przemyslaw; Lindenmayer, Aristid (1990).The Algorithmic Beauty of Plants. Springer-Verlag. pp. 101–107. ISBN 978-0387972978. <http://algorithmicbotany.org/papers/#webdocs>.
- Vogel, H (1979). "A better way to construct the sunflower head". Mathematical Biosciences 44 (44): 179–189.doi:10.1016/0025-5564(79)90080-4.

2.2. Characteristic Values. Evidently, the number of the new born pairs in successive generations in our simplified model of Fibonacci rabbits (or the sequence of Fibonacci numbers) increases rather rapidly as shown by the first 20 numbers in the table below.

R_0	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
1	1	2	3	5	8	13	21	34	55

R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}
89	144	233	377	610	987	1,597	2,584	4,181	6,765

While there are no serious obstacles to obtain R_n by computing recursively using MatLab or other scientific computing software on (1.1), some kind of explicit expression for R_n as a function of n similar to the expression $y_n = y_0 A^n$ for the

geometric growth model (3.2) of Chapter 1. In addition to possible gains in computational efficiency in calculating R_n, such an explicit expression often provides additional insight to the qualitative behavior of the model problem. However, a little reflection or experimentation would show that no single simple expression of the form y_0A^n could generate the Fibonacci sequence. For an expression of the form $R_n = y_0A^n$ to satisfy the two initial condition for the problem, we must have

$$R_0 = 1 = y_0, \qquad R_1 = 1 = y_0A$$

leading to erroneous result

$$R_n = y_0A^n = 1,$$

showing that there would be only one pair of new born rabbits forever. This result is obviously incompatible with the Fibonacci sequence and inconsistent with reality.

Suppose we go ahead and try such a solution for (1.1) anyway by setting $R_k = C\lambda^k$ with two unspecified constants C and λ (changing from y_0 and A to conform with the conventional notations). Since the expression for R_k mist satisfy the difference equation, we substitute the assumed expression for R_k $(= C\lambda^k)$ into (1.1) to get

$$C\lambda^2 = C\lambda + C \qquad or \qquad \lambda^2 - \lambda - 1 = 0$$

given that we want a nontrivial solution (and hence C should not vanish. It follows that there are two possible values of λ, $\lambda = \lambda_i,\ i = 1$ and 2:

$$\begin{pmatrix}\lambda_2\\ \lambda_1\end{pmatrix} = \frac{1}{2}\left[1 \pm \sqrt{5}\right]. \tag{2.1}$$

For each„ the expression $R_k = C_i\lambda_i^k$ is a solution of the difference equation (1.1) for any constant C_i. In other word, both

$$R_k^{(1)} = C_1\lambda_1^k \qquad \text{and} \qquad R_k^{(2)} = C_2\lambda_2^k$$

are solutions of (1.1) for any values of C_1 and C_2. Which solution should we take to provides the correct answer to the question posed earlier: How many pairs of *new born rabbits* will there be after k generations?

Evidently, neither solution is appropriate. If one of them should be chosen, then with $R_0^{(m)} = C_m = 1$, we have $R_k^{(m)} = \lambda_m^k = \left[1 \pm \sqrt{5}\right]^k/2^k$ which is not even an integer for either sign. Instead of discarding both, we note that, for any two constants C_1 and C_2, the linear combination $R_k = R_k^{(1)} + R_k^{(2)} = C_1\lambda_1^k + C_2\lambda_2^k$ is also a solution of (1.1) since

$$\begin{aligned} R_{n+1} + R_n &= (R_{n+1}^{(1)} + R_{n+1}^{(2)}) + (R_n^{(1)} + R_n^{(2)}) \\ &= (R_{n+1}^{(1)} + R_n^{(1)}) + (R_{n+1}^{(2)} + R_n^{(2)}) \\ &= (R_{n+2}^{(1)} + R_{n+2}^{(2)}) = R_{n+2}. \end{aligned}$$

In that case, we have as another possible solution of our Fibonacci rabbits model (1.1):

$$R_k = C_1\lambda_1^k + C_2\lambda_2^k. \tag{2.2}$$

The two initial conditions require

$$R_0 = C_1 + C_2 = 1, \qquad R_1 = C_1\lambda_1 + C_2\lambda_2 = 1.$$

These two linear equations are satisfied by taking

$$C_1 = \frac{1-\lambda_2}{\lambda_1-\lambda_2} = -\frac{\lambda_1}{\sqrt{5}}, \qquad C_2 = -\frac{1-\lambda_1}{\lambda_1-\lambda_2} = \frac{\lambda_2}{\sqrt{5}},$$

giving

$$R_k = \frac{1}{\sqrt{5}}\left[\lambda_2^{k+1} - \lambda_1^{k+2}\right] \tag{2.3}$$

Evidently, the explicit solution (2.3) for the second order difference equation (1.1) is the counterpart of (3.3) for first order differene equation (3.2) of Chapter 1. It can be verified that (2.3) reproduces the Fibonacci sequence as k runs through the nonnegative integers, the fact that the expresion involves $\sqrt{5}$ in a complicate way notwithstanding!

2.3. Superposition Principle I. The observation that linear combinations of two solutions of a (homogeneous) linear difference equation is also a solution of the same equation is sufficiently important to record the general case as the following theorem known as a *superposition principle* for homogeneous difference equations:

THEOREM 8. *(Superpostion Principle I) Let $y_n^{(1)}$ and $y_n^{(2)}$ be two solutions of the (autonomous) m^{th} order linear difference equation*

$$y_{n+m} = a_{m-1}y_{n+m-1} + a_{m-2}y_{n+m-2} + \ ..... \ + a_0 y_n \tag{2.4}$$

where $a_0, \ ....., a_{m-1}$ are known constants that do not vary with n. Then $y_n = c_1 y_n^{(1)} + c_2 y_n^{(2)}$ is also a solution of the same equation for any constants c_1 and c_2.

PROOF. Similar to the proof for the (1.1). □

LEMMA 1. *Let λ_i be a root of the characteristic equation*

$$\lambda^m - \left(a_{m-1}\lambda^{m-1} + a_{m-2}\lambda^{m-2} + \ ..... \ + a_1\lambda + a_0\right) = 0. \tag{2.5}$$

Then $y_n = c\lambda_i^n$ is a solution of the m^{th} order linear difference equation (2.4) for any constant c.

PROOF. (Exercise) □

THEOREM 9. *If $\{\lambda_1, \ \lambda_2, \ ....., \lambda_m\}$ are the m distinct roots of (2.5), then*

$$y_n = c_m\lambda_m^n + c_{m-1}\lambda_{m-1}^n + \ ..... \ + c_1\lambda_1^n \tag{2.6}$$

is the unique solution of the initial value problem (IVP) defined by the difference equation (2.4) and the m initial conditions

$$y_k = Y_k. \quad (k = 0, 1, 2, ..., m-1) \tag{2.7}$$

with $\{c_1, \ c_2, \ ....., c_m\}$ chosen to satisfy the m linear equations

$$c_m\lambda_m^k + c_{m-1}\lambda_{m-1}^k + \ ..... \ + c_1\lambda_1^k = Y_k, \qquad k = 0, 1, 2, ..., m-1.$$

PROOF. Substitute the expression (2.6) for y_k into the diffference equation (2.4) and the initial conditions (2.7) and verify that the these conditions are satisfied. It remains to show that the linear system of equation

$$c_m\lambda_m^k + c_{m-1}\lambda_{m-1}^k + \ ..... \ + c_1\lambda_1^k = Y_k \qquad (k = 0, 1, 2, ..., m-1)$$

has a unique solution for the constants $\{c_1, \ c_2, \ ....., c_m\}$. For this task, we need to know something more about solutions of linear equations (see Appendix). □

3. Plant Propagation

Many biological phenomena may be modeled by a second order linear difference equation with constant coefficients, more general than the equation for the simplified Fibonacci rabbits problem (1.1). One example is the propagation of annual (as opposed to perennial) plants described in [**3**]. Briefly, plants produce seeds in late summer and die at the end of the year. Some of the seeds survive the environment (harsh weather, predators, etc.) and a fraction of these germinates and grows into a new generation of the same plant species. The productive period of surviving seeds is rather short, say two years for the purpose of the present discussion.

3.1. What is the Question? Given what we know about the reproductive process of plant species, will a particular plant specie survive naturally? or is it heading for extinction? This is known as the "survival of the specie" problem.

3.2. What Do We Know? If the plants produce enough seeds, most of the seeds survive, and most of those survived seeds germinate, there would be enough new plants in successive generations for the specie to flourish. But for a given number of seeds produced each year, what survival and germination rates are adequate for the survival of the specie? A more quantitative analysis of the interplay between these contributing factors is needed to determine the ultimate fate of the species. We sketch below a second order difference equation model for the evolution of a particular plant specie.

3.3. The Model. For the phenomenon of annual plant propagation, let p_n be the number of plants in stage (year) n. To simplify our discussion, we consider the case that each plant produces the same amount of seed ρ every late summer and dies, leaving ρp_n amount of seeds to enter the winter months. Of these seeds, a fraction s_1 survives the winter and a fraction g_1 of these $s_1\rho p_n$ seeds germinates into $g_1 r_1 s p_n$ new plants in year $n+1$. The remaining $(1-g_1)s_1\rho p_n$ goes through a second winter with a fraction s_2 surviving by the second spring. Of that $s_2(1-g_1)s_1\rho p_n$ two year old seeds, a fraction g_2 germinates into new plants in year $n+2$. In addition to the these $g_2 s_2(1-g_1)s_1\rho p_n$ plants, there are in period $n+2$ additional plants from the one year old seeds produced by the plants of period $n+1$ equal to $g_1 s_1 \rho p_{n+1}$ (where we have kept the seed production rates and survival and germination fractions for the one year olds unchanged from the previous period). In that case, the total plants p_{n+2} in year $n+2$ is given by

$$p_{n+2} = ap_{n+1} + bp_n \tag{3.1}$$

where

$$a = g_1 s_1 \rho, \qquad b = g_2 s_2 (1-g_1) s_1 \rho. \tag{3.2}$$

We have distinguished the survival and germination fraction of the first year seeds $\{s_1, g_1\}$ from those of the second year seeds $\{s_2, g_2\}$ since the second year seeds tend to be weaker and the corresponding fractions smaller.

In some cases, there may be new plants introduced in any or all periods (through direct addition of new plants or plants resulting from wind blown seeds from neighboring orchards). As in previous population growth models with immigration, we would have an additional prescribed term $f_n \equiv f(n)$ on the right hand side of (3.1):

$$p_{n+2} = ap_{n+1} + bp_n + f_n.$$

The equation may be re-written in the standard form of a general second order equation for a typical unknown y_n $(= p_n)$:

$$y_{n+2} + \alpha y_{n+1} + \beta y_n = f_n \tag{3.3}$$

with $y_n = p_n$, $\alpha = -a$ and $\beta = -b$. The general difference equation (3.3) is said to be *homogeneous* if $f_n = 0$; it is *inhomogeneous* otherwise. Any (non-zero) solution of the homogeneous equation corresponding to (3.3) is called a *complementary solution*. Any solution of the inhomogeneous equation is called a *particular solution*.

Whether it is the inhomogeneous equation (3.3) or its homogeneous counterpart $(f_n = 0)$, we need two initial conditions to determine y_n for all positive integer n. In particular, we have from (3.3) for $n = 0$,

$$y_2 = -\alpha y_1 - \beta y_0 + f_0.$$

Since f_0 is prescribed, we need the values of y_1 and y_0 to determine y_2. Once we have these, we can use (3.3) to determine $y_2, y_3, y_4, \cdots$ recursively. In the problem of propagation of an annual plant species, we need to know the number of plants in (the first) two consecutive periods, p_0 and p_1, in order for the model to determine number of plants in future years The difference equation (3.3) and a set of initial conditions

$$y_0 = P, \qquad y_1 = Q \tag{3.4}$$

define an initial value problem (IVP) which determines the unknown for all $n \geq 2$.

4. Plant Growth without Immigration

4.1. Complementary Solutions. For the plant propagation problem without immigration, we have $f_n = 0$ for all n. In that case, the solution can be expressed in terms of the *characteristic roots* of the difference equation. As in the Fibonacci rabbits problem, we substitute a solution of the form $y_n = C\lambda^n$) into the governing difference equation with $f_n = 0$ to get following *characteristic equation* for the problem

$$\lambda^2 + \alpha\lambda + \beta = 0$$

whose two roots are

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \frac{1}{2}\left(-\alpha \pm \sqrt{\alpha^2 - 4\beta}\right). \tag{4.1}$$

By Superposition Principle I, we have the following complementary solution for the plant propagation problem

$$x_n = C_1\lambda_1^n + C_2\lambda_2^n \tag{4.2}$$

with the two constants C_1 and C_2. The notation x_n is used here to indicate that it is the complementary solution of the difference equation. The two constants of integration are needed (and suffice) for satisfying the two initial conditions (3.4) for the problem.

4.2. Stability of the Fixed Point. For the homogeneous equation (3.1), the only fixed point is at the origin $\bar{x} = 0$. Its stability is seen from (4.2) to be a function of the magnitude of the characteristic roots λ_1 and λ_2 which depend on the five parameters ρ, s_1, g_1, s_2 and g_2 in the model. To be concrete, we take the amount of seeds ρ produced by each plant to be given and fixed each year. The remaining four parameters s_1, g_1, s_2 and g_2 are fractions and therefore in the range $[0, 1]$. Evidently, we have from (4.2)

- $x_n \to \infty$ if $|\lambda_2| > |\lambda_1| > 1$
- $x_n \to 0$ if $|\lambda_1| < |\lambda_2| < 1$

In the first case, the species increase naturally toward the carrying capacity of the environment. In the second case, the species would naturally head toward extinction unless there should be intervention.

For the plant growth problem, we have

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \frac{1}{2}\left(-\alpha \pm \sqrt{\alpha^2 - 4\beta}\right) = \frac{1}{2}\left(a \pm \sqrt{a^2 + 4b}\right).$$

with the non-negative numbers a and b given in terms of the model parameters in (3.2). Since a and b are normally positive, an immediate conclusion for our plant growth model is

$$\lambda_1 < 0 < \lambda_2.$$

On the other hand, the magnitude of $|\lambda_1|$ and $|\lambda_2|$ clearly increases with s_1, s_2 and g_2 (while ρ is fixed), given $a = g_1 s_1 \rho$ and $b = g_2 s_2 (1 - g_1) s_1 \rho$. As functions of g_1, the characteristic roots have no stationary point. In the admissible range $0 \le g_1 \le 1$, we have

$$\max[\lambda_2] = [\lambda_2]_{g_1=1} = s_1\rho, \qquad [\lambda_1]_{g_1=1} = 0, \tag{4.3}$$

and

$$\min[\lambda_2] = [\lambda_2]_{g_1=0} = \sqrt{g_2 s_2 s_1 \rho} = -[\lambda_1]_{g_1=0}. \tag{4.4}$$

keeping in mind $s_2 < s_1$ and $\rho > 1 > g_2$. The following results are immediate consequences of the two expressions for $\max[\lambda_2]$ and $\min[\lambda_2]$:

PROPOSITION 3. *The plant specie modeled by (3.1) heads for extinction if* $s_1\rho < 1$ *and proliferates without bound if* $g_2 s_2 s_1 \rho > 1$.

4.3. Reduction of Order. Before leaving the second order difference equation (3.3), we should address the special case of $\alpha^2 - 4\beta = 0$ for which the two roots are identical so that we have only one characteristic root $\lambda_1 = -\alpha/2$, With $y_n = C_1\lambda_1^n$ containing only one constant, how do satisfy the two initial conditions in this case? Another special case that should be explored further is when $\alpha^2 < 4\beta$ for which the two characteristics are complex numbers. We note that neither applies to the plant model (3.1) for which the two roots are always real and distinct.

For the case of a repeated root so that we have only one solution of the assumed form , $y_n^{(1)} = C_1\lambda_1^n$, the complete solution cannot be in the form (4.2). To find the correct complete solution that would satisfy both initial conditions, we should take advantage of our knowledge of one available solution by writing the complete solution as

$$y_n = y_n^{(1)} u_n \tag{4.5}$$

where u_n is an unknown function of n to be determined. Since u_n is completely arbitrary, there is no loss in generality by taking the complete solution as $y_n^{(1)}u_n$. If the expression on the right side of (4.5) is the solution sought, then it must satisfy the difference equation, i.e.,

$$y_{n+2}^{(1)}u_{n+2} + \alpha y_{n+1}^{(1)}u_{n+1} + \beta y_n^{(1)}u_n = 0 \tag{4.6}$$

with $f_n = 0$ for the annual plant propagation problem. To take advantage of the fact that $y_n^{(1)}$ is a solution of the same difference equation so that

$$y_{n+2}^{(1)} + \alpha y_{n+1}^{(1)} + \beta y_n^{(1)} = 0, \tag{4.7}$$

we add and substract the same terms in the left hand side of (4.6) to get

$$y_{n+2}^{(1)}(u_{n+2} - u_{n+1}) + (y_{n+2}^{(1)} + \alpha y_{n+1}^{(1)} + \beta y_n^{(1)})u_{n+1} - \beta y_n^{(1)}(u_{n+1} - u_n) = 0$$

or, with $v_n = u_{n+1} - u_n$,

$$y_{n+2}^{(1)}v_{n+1} - \beta y_n^{(1)}v_n = 0 \tag{4.8}$$

given (4.7). Now, (4.8) is a first order linear difference equation which can be simplified to

$$v_{n+1} = v_n \tag{4.9}$$

since $y_n^{(1)} = C_1\lambda_1^n = C_1\left(-\frac{1}{2}\alpha\right)^n$ so that

$$\frac{\beta y_n^{(1)}}{y_{n+2}^{(1)}} = \frac{\beta}{\lambda_1^2} = \frac{\beta}{(-\alpha/2)^2} = 1.$$

The solution of (4.9) is known from Chapter 1 to be

$$v_n = v_0 \tag{4.10}$$

for all $n = 0, 1, 2, 3, ...$ where v_0 is an arbitrary constant. Given the defining relation $v_n = u_{n+1} - u_n$, the solution (4.10) is another first order linear difference equation for u_n

$$u_{n+1} - u_n = v_0 \tag{4.11}$$

The solution for (4.11) is also known from Chapter 1 to be

$$u_n = u_0 + v_0 n \qquad (n = 0, 1, 2, ...)$$

Altogether, we have the following theorem:

THEOREM 10. *For the case of $\alpha^2 = 4\beta$ so that there is only one distinct characteristic root $\lambda_1 = -\alpha/2$, the complete solution of the homogeneous second order linear difference equation with constant coefficients (3.1) is*

$$y_n = (u_0 + v_0 n)\, y_n^{(1)} = (c_1 + c_2 n)\, \lambda_1^n = (c_1 + c_2 n)\, (-\alpha/2)^n \tag{4.12}$$

where $c_1 = C_1 u_0$ and $c_2 = C_1 v_0$ are two constants that can be chosed to satisfy the two initial conditions (2.1).

4.4. Complex Roots. When the two characteristic roots are complex (and hence complex conjugate of each other), we can still apply Superposition Principle I to write (4.2). However, we expect the solution for a real life biological phenomenon to be real, we should take the two constants to be complex conjugate of each other as well:

$$x_n = C\lambda_1^n + C^* \left(\lambda_1^*\right)^n \tag{4.13}$$

where $(\ )^*$ denotes the complex conjugate of $(\ )$. Instead of the expression (4.13), it is often preferable to work with actual real expressions for the individual terms. For that purpose, we use the polar representation for a complex nunber $\lambda_1 = \lambda_r + i\lambda_i = \lambda(\cos\theta + i\sin\theta)$ and $C = c_r + ic_i$ to write (4.13) as

$$\begin{aligned} x_n &= 2\lambda^n \left[c_r \cos(n\theta) - c_i \sin(n\theta)\right] \\ &= \lambda^n \left[C_r \cos(n\theta) + C_i \sin(n\theta)\right] \end{aligned} \tag{4.14}$$

for two new real arbitrary constants of integration C_r and C_i.

5. Fixed Points and Stability

5.1. Classification of Fixed Points. Returning to the Fibonacci rabbits problem, we see from the solution (2.3) $R_n \to \infty$ as $n \to \infty$, since $\lambda_2 = \left[1 + \sqrt{5}\right]/2 > 1$. This is not unexpected given that no rabbit ever dies. As we have learned previously from nonlinear growth of the human population that stability and instability are usually associated with some steady state configuration, it is natural to ask what fixed point of the difference equation is repelling and causing the movement toward infinity. For an answer for the Fibonacci rabbits growth, we seek a stage (time) independent solution of our difference equation (1.1) by setting $R_n = \bar{R}$ for all n. In that case the difference equation becomes $\bar{R} = \bar{R} + \bar{R}$ giving $\bar{R} = 0$ as the only fixed point for the equation.

However, not all combinations of initial conditions (3.4) give rise to a solution of the difference equation (1.1) that heads toward infinity. An exception is the combination $R_0 = 1$ and $R_1 = \lambda_1$. (The second of these is not meaningful for the Fibonacci rabbits problem given $\lambda_1 = 1 - \sqrt{5}/2 < 0$. However, (1.1) may be the governing difference equation for another phenomenon for which $\lambda_1 = \left[1 - \sqrt{5}\right]/2$ is meaningful.) This new set of initial conditions is satisfied by $C_1 = 1$ and $C_2 = 0$ giving us

$$R_n = \lambda_1^n.$$

With $0 < |\lambda_1| = \left|1 - \sqrt{5}\right|/2 < 1$, $R_n = \lambda_1^n$ tends to 0 as $n \to \infty$.

While the situation is similar for the plant growth equation (3.1), the fixed point at the origin is asymptotically stable ($x_n \to 0$) for any combination of initial data for the following difference equation

$$x_{n+2} - \frac{1}{2}x_{n+1} + \frac{1}{18}x_n = 0$$

for which $\lambda_1 = 1/6$ and $\lambda_2 = 1/3$. To allow for succinct characterization of the different stability type of fixed points, we introduce the following definitions for a general second or higher order difference equation or a system of several first order equation for many more unknowns:

DEFINITION 7. *A fixed point $\bar{x}$ of an m^{th} order difference equation (or $\bar{\mathbf{x}}$ of m first order linear difference equations) is said to be a* ***sink*** *if $0 < |\lambda_k| < 1$ for all $k = 1, 2, ..., m$. It is said to be a* ***source*** *if $|\lambda_k| > 1$ for all $k = 1, 2, ..., m$.*

DEFINITION 8. *A fixed point $\bar{x}$ of an m^{th} order difference equation (or $\bar{\mathbf{x}}$ of m first order linear difference equations) is said to be a* ***saddle point*** *if (i) $|\lambda_k| \neq 0$ for all $k = 1, 2, ..., m$, (ii) $|\lambda_i| < 1$ for some but some i (so that at least one eigenvalue is less than one in magnitude) and (iii) $|\lambda_j| > 1$ for some but some j (so that at least one eigenvalue is greater than one in magnitude).*

DEFINITION 9. *A fixed point $\bar{x}$ of an m^{th} order difference equation (or $\bar{\mathbf{x}}$ of m first order linear difference equations) is said to be a* ***spiral point*** *if all eigenvalues are complex conjugate pairs with a non-zero real part. It is asymptotically stable if $|\lambda_k| < 1$ for all $k = 1, 2, ..., m$. It is unstable if one or more eigenvalues are of magnitude greater than one, i.e., $|\lambda_p| > 1$ for one or more integer p..*

DEFINITION 10. *A fixed point $\bar{x}$ of an 2^{nd} order difference equation (or $\bar{\mathbf{x}}$ of two first order linear difference equations) is said to be a* ***center*** *if the two eigenvalues are complex conjugate pair with $|\lambda_k| = 1$. The fixed point in this case is stable but not asymptotically stable.*

With the classification above, the fixed point at the origin of the Fibonacci rabbits model (1.1) is a saddle point. On the other hand, the fixed point for the plant growth model (3.1) may be of different type depending on the values of the model parameters and the type of stability may depend on initial data in some cases.

That the stability of a fixed point may depend on the initial condition is not new. We encountered a similar dependence on the initial condition in growth models governed by a single first order nonlinear difference equation

$$x_{n+1} = F(x_n) \tag{5.1}$$

with $F'(\bar{x}) = 1$ and $F''(\bar{x}) \neq 0$ at the fixed point $\bar{x}$. For these models, the $\bar{x}$ was found to be semi-stable according to Theorem 4. At the time, the stability nature of the fixed point for the case of $F'(\bar{x}) = -1$ and $F''(\bar{x}) \neq 0$ was postponed until after a discussion of second order difference equations. Having now some exposure to such difference equations, we are ready to consider this case in the next subsection.

5.2. A Nonlinear 2^{nd} Order Equation. The study of a second or higher order nonlinear difference equation is most conveniently done by converting them to a system of first order nonlinear difference equations. Such systems will be investigated in the next few chapters of these notes. Here, we will consider only issues for a nonlinear second order equation of a special kind relevant to the missing result on the stability of a fixed point of (5.1) with $F'(\bar{x}) = -1$ and $F''(\bar{x}) \neq 0$.

The second order equation of interest here is

$$x_{n+2} = f(x_n) \equiv F(x_{n+1}) = F(F(x_n)) \tag{5.2}$$

where $F(\cdot)$ is as in (5.1), sufficiently continuously differentiable for the validity of the theorems below. For this special second order difference equation, the following basic results are immediate.

THEOREM 11. *If $\bar{x}$ is a fixed point of (5.1), then it is also a fixed point of (5.2). In addition, if $\bar{x}$ is asymptotically stable for (5.1), it is also asymptotically stable for (5.2).*

PROOF. If $F(\bar{x}) = \bar{x}$, then $\bar{x}$ is also a fixed point for (5.2) given

$$F(F(\bar{x})) = F(\bar{x}) = \bar{x}.$$

Suppose $\bar{x}$ is a stable fixed point for (5.1), then for an initial data $\bar{x}_0$ sufficiently close to $\bar{x}$, the sequence $\{x_n\}$ generated by (5.1) starting from $\bar{x}_0$ converges to $\bar{x}$, i.e.,

$$\lim_{n\to\infty} x_n = \bar{x}. \tag{5.3}$$

In that case, the subsequence $\{x_{2n}\}$ of the sequence $\{x_n\}$ also converges to the same fixed point

$$\lim_{n\to\infty} x_{2n} = \bar{x} \tag{5.4}$$

and thereby $\bar{x}$ is also a stable fixed point of (5.2) since $\{x_{2n}\}$ is the sequence generated by (5.2) starting from $\bar{x}_0$.

□

Note that a fixed point of (5.2) needs not be a fixed point for (5.1); a counterexample is a two-cycle of the latter is a fixed point of the former. When $\bar{x}$ is a fixed point of both and is stable fixed point of (5.2), we would like to show that it is also a stable fixed point of (5.1). Before we prove this claim, we need to recall the following fact: The second order difference equation (5.2) generates from the initial data $\bar{x}_0$ only the sequence $\{x_{2k}\}$ which is just the subsequence of $\{x_n\}$ generated by (5.1) with even subscript. In that case, asymptotic stability of a fixed point $\bar{x}$ of (5.2) corresponds to (5.4) as indicated in the proof of Theorem 11 .

5.3. First Order Growth with $F'(\bar{x}) = -1$. We are now ready to state and prove the following theorem needed for our final result:

THEOREM 12. *If $\bar{x}$ is fixed point for both (5.1) and (5.2) and is asymptotically stable for either (5.1) or (5.2), then it is also asymptotically stable for both. If it is unstable for (5.2), it is also unstable for (5.1).*

PROOF. Given Theorem 11, we need only to prove the asymptotic stability oft $\bar{x}$ for (5.1) if $\bar{x}$ is asymptotically stable for (5.2). The latter implies (5.4) for $\bar{x}_0$ inside some small interval I of the fixed point. For a sufficiently small I, $x_1 = F(\bar{x}_0)$ also lies inside the interval and so that be the sequence $\{x_{2k+1}\}$ generated by $f(x) = F(F(x))$ starting from x_1 also converges to the fixed point $\bar{x}$. Combining the two sequences to get (5.3) for any $\bar{x}_0$ inside I (sufficiently small so that $x_1 = F(\bar{x}_0)$ is also inside I). As such, $\bar{x}$ is asymptotically stable for (5.1).
□

THEOREM 13. *If $\bar{x}$ is a fixed point of (5.2) and $F'(x) = -1$, then $\bar{x}$ as a fixed point of (5.2) is (asymptotically) stable if $f'''(\bar{x}) < 0$ and unstable if $f'''(\bar{x}) > 0$.*

PROOF. For the stability of $\bar{x}$ as a fixed point of (5.2), we need to examine

$$f'(x) = \frac{df}{dx} = \left[\frac{dF(y)}{dy}\frac{dy}{dx}\right],$$

where $y = F(x)$. With $[dy/dx]_{x=\bar{x}} = F'(\bar{x}) = -1$, we get

$$f'(\bar{x}) = \left[\frac{dF(y)}{dy}\right]_{y=F(\bar{x})=\bar{x}} (-1) = (-1)(-1) = 1.$$

By Theorem 4, we need to examine $f^{''}(\bar{x})$. From

$$f^{''}(x) = \frac{d}{dx}\left[\frac{dF(y)}{dy}\frac{dy}{dx}\right] = \frac{d^2F(y)}{dy^2}\left(\frac{dy}{dx}\right)^2 + \frac{dF(y)}{dy}\frac{d^2y}{dx^2}$$

we get

$$\begin{aligned} f''(\bar{x}) &= \left[\frac{d^2F(y)}{dy^2}\right]_{y=F(\bar{x})=\bar{x}}(-1)^2 + \left[\frac{dF(y)}{dy}\right]_{y=F(\bar{x})=\bar{x}}F''(\bar{x}) \\ &= F''(\bar{x}) + (-1)F''(\bar{x}) = 0. \end{aligned}$$

In that case, Theorem 5 applies to $f(x)$ requiring information about the sign of $f^{'''}(\bar{x})$. □

COROLLARY 2. *If $\bar{x}$ is a fixed point of (5.1), hence also of (5.2), with $F'(x) = -1$, then $\bar{x}$ is asymptotically stable if $f'''(\bar{x}) < 0$ and unstable if $f'''(\bar{x}) > 0$.*

PROOF. The corollary is an immediate consequence of Theorems 12 and 13. □

Since $\bar{x}$ is a fixed point for (5.1), it seems appropriate to express the condition for stability in terms of $F(\cdot)$ instead of $f(\cdot)$. From

$$f'''(x) = \frac{d^3F(y)}{dy^3}\left(\frac{dy}{dx}\right)^3 + 3\frac{d^2F(y)}{dy^2}\frac{dy}{dx}\frac{d^2y}{dx^2} + \frac{dF(y)}{dy}\frac{d^3y}{dx^3}$$

follows

$$\begin{aligned} f'''(\bar{x}) &= \left[\frac{d^3F(y)}{dy^3}\right]_{y=F(\bar{x})=\bar{x}}(-1)^3 + 3(-1)F''(\bar{x})\left[\frac{d^2F(y)}{dy^2}\right]_{y=F(\bar{x})=\bar{x}} + (-1)F'''(\bar{x}) \\ &= -2F'''(\bar{x}) - 3\left[F''(\bar{x})\right]^2. \end{aligned}$$

The following corollary of the theorem above is an immediate consequence of Theorem 5:

COROLLARY 3. *The fixed point $\bar{x}$ of (2.1) is (asymptotically) stable if $f'''(\bar{x}) = -2F'''(\bar{x}) - 3\left[F''(\bar{x})\right]^2 < 0$ and is unstable if $f'''(\bar{x}) = -2F'''(\bar{x}) - 3\left[F''(\bar{x})\right]^2 > 0$.*

The following examples illustrate applications of this result to problems for which cobweb graphs are not particularly helpful.

EXAMPLE 2. $F(x) = -x + 2x^2$

For this problem, we have one fixed point at $\bar{x} = 0$ with $F'(0) = -1$, $F''(0) = 4$ and $F'''(0) = 0$ so that $f'''(0) = -2F'''(0) - 3\left[F''(0)\right]^2 = -48 < 0$. By Corollary 3 above, the fixed point $\bar{x} - 0$ is asymptotically stable.

REMARK 1. *It should be evident from the example above that if $F(x)$ is any genuine quadratic function of x (and hence $F''(x) \neq 0$), any fixed point $\bar{x}$ of $F(x)$ is asymptotically stable since $F'''(x) = 0$ for all x so that*

$$f'''(\bar{x}) = -2F'''(\bar{x}) - 3\left[F''(\bar{x})\right]^2 = -3\left[F''(\bar{x})\right]^2 < 0.$$

EXAMPLE 3. $F(x) = -x - x^3$

For this problem, we have a fixed point at $\bar{x} = 0$ with $F'(0) = -1$, $F''(0) = 0$ and $F'''(0) = -6$ so that $f'''(0) = -2F'''(0) - 3\left[F''(0)\right]^2 = 12 > 0$. By Corollary 3 above, the fixed point $\bar{x} = 0$ is unstable.

CHAPTER 6

External Forcing

1. Plant Growth with Immigration

1.1. Superposition Principle II. Similar to the Fibonacci rabbits problem, we would like a simple formula for the unknown y_n as a function on n for the plant growth with immigration and more generally for the inhomogeneous difference equation (3.3). For this purpose, we make use of the following superposition principle for inhomogeneous linear difference equations:

THEOREM 14. *(Superposition Principle II): Suppose $y_n^{(1)}$ is a complementary solution and $y_n^{(p)}$ is any solution of the inhomgeneous difference equation (**??**). Then for any constant c_1, the linear combination $c_1 y_n^{(1)} + y_n^{(p)}$ is also a solution of the inhomogeneous equation (**??**).*

PROOF. (exercise) □

Theorem 14 enables us to break the problem down to two simpler tasks. One is to obtain all the complementary solutions and the othre is to obtain any particular solution. These two tasks will be discussed separately in the next few sections.

1.2. A Particular Solution. For $f_n \neq 0$, we need also a particular solution for the inhomogeneous difference equation (3.3) to complete the solution of the IVP. To illustrate, consider the case that ten new plants are added at each stage so that $f_n = 10$. Superposition Principle II allows us to obtain any solution of the inhomogeneous equation

$$y_{n+2} + \alpha y_{n+1} + \beta y_n = 10 \tag{1.1}$$

Given that $f_n = 10$ (or any constant p) does not change from stage to stage, we may seek a solution z_n that also does not depend on stage: $z_n = z$. For such a solution, the difference equation becomes

$$(1 + \alpha + \beta)z = 10$$

or

$$z = \frac{10}{1 + \alpha + \beta}$$

as long as the denominator does not vanish. Note that the method of solution works for any forcing term that is independent of stage, i.e., $f_n = f$ with a particular solution being

$$z = \frac{f}{1 + \alpha + \beta}$$

again assuming $1 + \alpha + \beta \neq 0$. For the plant growth problem, we have $\alpha = -a = -g_1 r_1 s < 0$ and $\beta = -b = -g_2 r_2 (1 - g_1) r_1 s. < 0$ so that the denominator could vanish. However, since the parameters $\{g_1, r_1, g_2, r_2\}$ are all fractions, $1 + \alpha +$

β would vanish only for a special value of s (that depends on these fractions). This exceptional case is treated after we complete the solution for the normal case when $\lambda_1 \neq \lambda_2$ and $1 + \alpha + \beta \neq 0$.

1.3. Solution of the IVP. By Superposition Principle II, we know that for the normal case

$$y_n = x_n + z = C_1\lambda_1^n + C_2\lambda_2^n + \frac{f}{1+\alpha+\beta}$$

is a solution of the inhomogeneous difference equation (1.1). This solution is required to satisfy the two initial conditions (3.4) so that we must have

$$\begin{aligned} C_1 + C_2 + \frac{f}{1+\alpha+\beta} &= P \\ C_1\lambda_1 + C_2\lambda_2 + \frac{f}{1+\alpha+\beta} &= Q \end{aligned}$$

These two linear equations can be solved for the two yet unspecified constants C_1 and C_2 since the determinant of the coefficient matrix $\lambda_2 - \lambda_1 \neq 0$. This complete the solution of the original IVP formulated in subsection 3.4 for $f_n = f$.

It is important to remember that the complementary solution x_n is only a piece of the solution of the solution of the problem (see Superposition Principle II above), not the actual solution for the inhomogeneous equation. The use of the notation x_n was to remind us of that fact. We do not choose C_1 and C_2 to satisfy the initial conditions until we have the complete solution of the difference equation.

2. Variation of Parameters

2.1. The General Method. For the exceptional case of $1 + \alpha + \beta = 0$ (and for more complex forcing function f_n), the assumption of a particular solution in the simple form of $z_n = z$ fails. If we still wish to obtain an explicit solution in terms of elementary functions, the method of reduction of order also applies to inhomogeneous equation (and always so) since there is always at least one complementary solution for the difference equation. However, when 1 is not a double root of the characteristic equation, we should take advantage of the fact that there are two complementary solutions (for our second order equation) for possible simplification of the solution. This is accomplished by the method of variation of parameters. Instead of taking the solution sought in the form (4.5), we make use of both complementary solution by setting

$$y_n = y_n^{(1)}u_n + y_n^{(2)}v_n = \lambda_1^n u_n + \lambda_2^n v_n$$

For such a solution, we have

$$\begin{aligned} y_{n+1} &= y_{n+1}^{(1)}u_{n+1} + y_{n+1}^{(2)}v_{n+1} \\ &= y_{n+1}^{(1)}(u_{n+1} - u_n) + y_{n+1}^{(2)}(v_{n+1} - v_n) + y_{n+1}^{(1)}u_n + y_{n+1}^{(2)}v_n \end{aligned}$$

Since the assume solution has two unknowns u_n and v_n and there is only one difference equation for their determination, we may introduce another condition to specify (and thereby eliminate) one of the unknowns. To simplify the solution process, we take this condition to be

$$y_{n+1}^{(1)}(u_{n+1} - u_n) + y_{n+1}^{(2)}(v_{n+1} - v_n) \equiv y_{n+1}^{(1)}\Delta u_n + y_{n+1}^{(2)}\Delta v_n = 0 \tag{2.1}$$

with $\Delta w_n = w_{n+1} - w_n$, leaving use with the simpler expression for y_{n+1}:

$$y_{n+1} = y^{(1)}_{n+1} u_n + y^{(2)}_{n+1} v_n \tag{2.2}$$

and correspondingly

$$\begin{aligned} y_{n+2} &= y^{(1)}_{n+2} u_{n+1} + y^{(2)}_{n+2} v_{n+1} \\ &= y^{(1)}_{n+2} \Delta u_n + y^{(2)}_{n+2} \Delta v_n + y^{(1)}_{n+2} u_n + y^{(2)}_{n+2} v_n \end{aligned} \tag{2.3}$$

We now make use of the expressions (2.2) and (2.3) in the difference equation to get

$$y^{(1)}_{n+2} \Delta u_n + y^{(2)}_{n+2} \Delta v_n = f_n \tag{2.4}$$

where we have made use of $y^{(k)}_{n+2} + \alpha y^{(k)}_{n+1} + \beta y^{(k)}_n = 0$ for $k = 1$ and 2 to eliminate the other terms. The two conditions (2.1) and (2.4) can be solved for Δu_n and Δv_n to get

$$u_{n+1} - u_n = -\frac{y^{(2)}_{n+1}}{W} f_n, \qquad u_{n+1} - u_n = \frac{y^{(1)}_{n+1}}{W} f_n. \tag{2.5}$$

where

$$\begin{aligned} W &= y^{(1)}_{n+1} y^{(2)}_{n+2} - y^{(1)}_{n+2} y^{(2)}_{n+1} = \lambda_1^{n+1} \lambda_2^{n+2} - \lambda_1^{n+2} \lambda_2^{n+1} \\ &= \lambda_1^{n+1} \lambda_2^{n+1} (\lambda_2 - \lambda_1) = \beta^{n+1} \sqrt{\alpha^2 - 4\beta} \end{aligned}$$

The two relations in (2.5) are two uncoupled first order linear difference equations which we know how to solve for u_n and v_n, respectively.

2.2. Stage Invariant Forcing $f_n = f$.

2.2.1. *The Normal Case:* For the special case $f_n = f$, the two relations (2.5) simplify to

$$u_{n+1} - u_n = -\frac{f \lambda_1^{-n}}{\lambda_1 (\lambda_2 - \lambda_1)}, \qquad v_{n+1} - v_n = \frac{f \lambda_2^{-n}}{\lambda_2 (\lambda_2 - \lambda_1)}$$

The right hand side for these first order difference equaitons being of the form $c_0 \gamma^{-n}$, the solution for u_n and v_n normally may be taken in the form

$$u_n = c_1 + u_0 \lambda_1^{-n}, \qquad v_n = c_2 + v_0 \lambda_2^{-n}, \tag{2.6}$$

respectively to obtain

$$u_0 = c_1 - \frac{f}{(1 - \lambda_1)(\lambda_2 - \lambda_1)}, \qquad v_0 = c_2 + \frac{f}{(1 - \lambda_2)(\lambda_2 - \lambda_1)},$$

Altogether we have

$$y_n = c_1 \lambda_1^n + c_2 \lambda_2^n + \frac{f}{(1 - \lambda_1)(1 - \lambda_2)}$$

with the two constants c_1 and c_2 available to match the two initial conditions (3.4).

2.2.2. *The Exceptional Case:* For the exceptional case $1 + \alpha + \beta = 0$ or $\beta = -(1 + \alpha)$, we have $\alpha^2 - 4\beta = (\alpha + 2)^2$ so that the two characteristic roots are

$$\lambda_1 = \frac{1}{2}\left[-\alpha + (\alpha + 2)\right] = 1, \qquad \lambda_2 = \frac{1}{2}\left[-\alpha - (\alpha + 2)\right] = -(\alpha + 1).$$

In that case, we may proceed to calculate v_n as before but the equation for u_n now becomes

$$u_{n+1} - u_n = \frac{f}{\alpha + 2} \equiv f_0$$

with the right hand side independent of n and a solution of the form $u_n = u_0\lambda_1^{-n}$ $= u_0$ no longer appropriate. Instead, the correct form of the solution for this case was found in (4.11) to be proportional to n. We can obtain it directly by calculating

$$u_1 = u_0 + f_0, \qquad u_2 = u_1 + f_0 = (u_0 + f_0) + f_0 = u_0 + 2f_0,$$

and therewith

$$u_n = u_{n-1} + f_0 = [u_0 + (n - 1)f_0] + f_0 = u_0 + nf_0.$$

where u_0 is a constant still to be specified. Altogether, we have then

$$y_n = c_1\lambda_1^n + c_2\lambda_2^n + \frac{f}{(2 + \alpha)^2}\left[n(2 + \alpha) - 1\right],$$

again with two constants of integration to match the two initial conditions.

3. Method of Undetermined Coefficients

The examples on obtaining a particular solution in the last few sections suggest a general approach to particular solutions for a linear difference equation with constant coefficients known as the *method of undetermined coefficients*. For certain simple forms of f_n, we general expect a particular solution of the difference equation exists in a form proportional to, or a function of f_n For these cases, we can assume such solution form and let the equation determine the unknown constants or whatever remaining freedom there may be. We describe below the method for some simple classes of f_n to illustrate this general approach.

3.1. Power Forcing - $f_n = f_0\gamma^n$.

3.1.1. *The Normal Case:* Assume a particular solution of the form $z_n = c\gamma^n$ so that the difference equation (3.3) becomes

$$(\gamma^2 + \alpha\gamma + \beta)c\gamma^n = f_0\gamma^n.$$

This relation requires

$$c = \frac{f_0}{\gamma^2 + \alpha\gamma + \beta},$$

assuming the denominator does not vanish. If it does for the particular combination of α, β, and γ, we can always use the reduction of order or variation of parameters for an appropriate particular solution. However, our experience with the exceptional case of the plant proliferation problem with immigration suggests a short cut to such a solution to be described in the next subsection.

3.1.2. *The Exceptional Case:* Recall that for the equation $y_{n+2} - y_n = f_0$, we have $\gamma = 1$ and $\gamma^2 + \alpha\gamma + \beta = 1^2 - 1 = 0$. For this equation, a particular solution of the form $z_n = c \cdot 1^n = c$ is inappropriate (since $c \cdot 1^{n+2} - c \cdot 1^n = 0 \neq f_0$). The method of variation of parameters gives one particular solution to be

$$z_n = f_0 \frac{n}{2}$$

Direct verification shows that z_n is a solution of the inhomogenous difference equation (3.3):

$$z_{n+2} - z_n = \frac{f_0}{2}(n + 2 - n) = f_0.$$

The results above suggest the following more general result for a particular solution: method for

PROPOSITION 4. *For the linear second order difference equation (3.3) with $f_n = f_0\gamma^n$, a particular solution exists in the form $z_n = P_k(n)\gamma^n$ for some k^{th} degree polynomial of n, $P_k(n) = c_k n^k + \cdots + c_0$, with k being the multiplicity of γ as a root of the characteristic equation for (3.3): $\lambda^2 + \alpha\lambda + \beta = 0$. In all cases, the unspecified coefficients of $P_k(n)$ are to be determined by matching the same power of n on both sides of the difference equation.*

PROOF. (exercise) □

3.2. Polynomial Forcing - $f_n = f_0 n^k$.

3.2.1. *The Linear Case ($k = 1$):* For $f_n = f_0 n$, it is tempting to consider $z_n = c_1 n$. For such an assumed solution, the left hand side of the difference equation, $z_{n+2} + \alpha z_{n+1} + \beta z_n$, becomes

$$z_{n+2} + \alpha z_{n+1} + \beta z_n = c_1 \left[(1 + \alpha + \beta)n + (2 + \alpha)\right] \tag{3.1}$$

It cannot be equal to $f_0 n$ for any choice of c_1 unless $\alpha = -2$. If $\alpha = -2$, the difference equation is satisfied by taking

$$c_1 = \frac{f_0}{1 + \alpha + \beta} = \frac{f_0}{\beta - 1} \tag{3.2}$$

provided that the denominator does not vanish.

For $2 + \alpha \neq 0$, the relation (3.1) suggests that we consider a particular solution in the form $z_n = c_1 n + c_0$ instead. In that case, the difference equation becomes

$$c_1(1 + \alpha + \beta)n + [c_1(2 + \alpha) + c_0(1 + \alpha + \beta)] = f_0 n.$$

This equation can be satisfied by (3.2) and

$$c_0 = -\frac{2 + \alpha}{1 + \alpha + \beta} c_1$$

again provided $1 + \alpha + \beta \neq 0$.

If $1 + \alpha + \beta = 0$ (so that $\lambda = 1$ is a root of the characeteristic equation of the difference equation (3.3)), our experience with the power forcing case suggest that we consider a polynomial solution one order higher than f_n. We record this observation in the following proposition:

PROPOSITION 5. *For the linear second order difference equation (3.3) with $f_n = f_0 n^k$, a particular solution exists in the form of a polynomial of*

i) degree k with $k+1$ coefficents to be specified if $1+\alpha+\beta \neq 0$;

ii) degree $k+m$ with $k+m+1$ coefficents to be specified if $1+\alpha+\beta = 0$, with m being the multiplicity of the characteristic root $\lambda = 1$.

In all cases, the unspecified coefficients are to be determined by matching the same power of n on both sides of the difference equation.

PROOF. (exercise) □

3.3. Oscillatory Forcing - $f_n = f_0\{\cos(n\pi\gamma), \sin(n\pi\gamma)\}$. For $f_n = f_0 \sin(n\pi\gamma)$, we expect a particular solution to be the linear combination $z_n = C_1 \sin(n\pi\gamma) + C_2 \cos(n\pi\gamma)$, seeing that a solution proportional to either function alone would not do (why?). For the assumed solution, the difference equation becomes

$$s(C_1, C_2)\sin(n\pi\gamma) + c(C_1, C_2)\cos(n\pi\gamma) = f_0 \sin(n\pi\gamma) \tag{3.3}$$

where

$$\begin{aligned} s(C_1, C_2) &= C_1\left[\cos(2\pi\gamma) + \alpha\cos(\pi\gamma) + \beta\right] - C_2\left[\sin(2\pi\gamma) + \alpha\sin(\pi\gamma)\right] \\ c(C_1, C_2) &= C_2\left[\cos(2\pi\gamma) + \alpha\cos(\pi\gamma) + \beta\right] + C_1\left[\sin(2\pi\gamma) + \alpha\sin(\pi\gamma)\right] \end{aligned}$$

Match of coefficients of $\cos(n\pi\gamma)$ and $\sin(n\pi\gamma)$ on both sides of (3.3) requires

$$s(C_1, C_2) = f_0, \qquad c(C_1, C_2) = 0 \tag{3.4}$$

The relations (3.4) are two linear equations for C_1 and C_2 and can be solved for these two constants.

The results above may be extended to give the following proposition:

PROPOSITION 6. *For the linear second order difference equation (3.3) with $f_n = f_s \sin(n\pi\gamma) + f_c \cos(n\pi\gamma)$, a particular solution exists in the form of $z_n = C_1 \sin(n\pi\gamma) + C_2 \cos(n\pi\gamma)$ with the two constants determined by matching $\sin(n\pi\gamma)$ and $\cos(n\pi\gamma)$,respectively, on both sides of the difference equaton.*

3.4. Combinations of Elementary Forcing Functions. (to be written)

Part 2

Interacting Populations

CHAPTER 7

Linear Systems

1. Red Blood Cell Production

1.1. The Problem. Humans and other vertebrate organisms live on oxygens and red blood cells (RBC) are their principal means of delivering oxygen to the body tissues via blood flow. Proper supply and functioning of RBC are therefore critical to their survival. Abnormalities of RBC occur in many forms, including various types of anemias (e.g., the Sickle-cell disease), hemolysis (a term for various form of RBC breakdown), and polycythemias (a general expression for excessive supply of RBC); most of them are life threatening. To deal with these abnormalities by drug treatment or other forms of clinical intervention, we need knowledge and insight to the supply of normal RBC.

1.2. Known Facts. RBC are produced by bone marrow at the rate of about 200 billion (2×10^{11}) per day or about $2,4$ million/sec. Each cell has a half life of about 55-60 days and is lost after 120 days on the average. At any one time, a human adult has about 30 trillions (3×10^{13}) RBC. An average adult has about $8 - 10$ lbs of bone marrow (about 5% of body weight). Bone marrow of mature adult degrades very slowly (rapidly replenished through some feedback mechanism) until old age when the rate becomes high. However, significant loss of RBC (due to disease or abnormal environmental changes) induces production of additonal bone marrow to replenish the excessive loss of RBC at a faster rate.

1.3. A Model on RBC Replenishment. Let R_n and M_n be the amount of RBC and bone marrow, respectively, at stage n. We take here the stage unit to be a day, given the available data on replacement and production of RBC mentioned above. The large size of the RBC population allows us to think of R_n as a real-valued (as opposed to integer-valued) quantity. (Alternatively, we can measure RBC population in units of biomass.) The information on degradation and replenishment activities of the RBC population are summarized by the following pair of relations

$$R_{n+1} = (1-\alpha)R_n + \beta M_n, \quad M_{n+1} = \gamma(R_c - R_n) + (1-\varepsilon)M_n \tag{1.1}$$

where $0 < \alpha, \varepsilon < 1, \beta > 0, \gamma > 0$ and $R_c > 0$ is a maintenance level of RBC.

- $\alpha = 1/120$ given a RBC half life of 60 days
- $\beta = 1.73 \times 10^8 cells/lbs/day$ in order for 10 lbs of bone marrow to replenish the daily loss of about !.73 billion cells (but should be considerably less since bone marrow is also responsible for producing many other cells as well)

- $\gamma = 5 \times 10^{-14} lbs//cell/day$ estimated on the basis of a production of 0.1 lbs of additional bone marrow needed to replenish the loss of 2×10^{11} RBC (= one day's normal production) in about 10 days.
- $\varepsilon = 0$ for mature adults.

The linear system of difference equations is supplemented by two initial conditions:

$$R_0 = \bar{R}, \qquad M_0 = \bar{M} \tag{1.2}$$

EXERCISE 7. *Solve the IVP (1.1) - (1.2) by the method of elinination used for the plant growth problem.*

For more complex phenomena with more evolving unknown quantities, the method of elimination employed for the Fibonacci rabbits becomes tedious and often impractical. For large linear systems of difference equations, the only feasible approach is to work with the system in vector form. We illustrate this approach using the present two unknown problem. With the vector representation

$$\mathbf{y}_n = \begin{pmatrix} R_n \\ M_n \end{pmatrix}, \quad \mathbf{f}_n = \begin{pmatrix} 0 \\ \gamma R_c \end{pmatrix}, . \tag{1.3}$$

we can write the linear system (1.1) as

$$\mathbf{y}_{n+1} = A\mathbf{y}_n + \mathbf{f}_n \tag{1.4}$$

where

$$A = \begin{bmatrix} 1-\alpha & \beta \\ -\gamma & 1-\varepsilon \end{bmatrix} \tag{1.5}$$

is the *coefficient matrix* of the linear system. (A summary of basic matrix algebra and its application can be found in Appendix 1.) Correspondingly, the initial conditions can be written as

$$\mathbf{y}_0 = \mathbf{Y} \equiv \begin{pmatrix} \bar{R} \\ \bar{M} \end{pmatrix}. \tag{1.6}$$

1.4. Superposition and Undetermined Coefficients; The following two theorems are the vector counterparts of the superposition principles for a single second order equation:

THEOREM 15. *(Superposition Principle I) If $\mathbf{y}_n^{(i)}$ and $\mathbf{y}_n^{(j)}$ are two (vector) complementary solutions of the linear system (1.4), so is $c_i\mathbf{y}_n^{(i)} + c_j\mathbf{y}_n^{(j)}$ for any constants c_i and c_j.*

PROOF. With $\mathbf{x}_n = c_i\mathbf{y}_n^{(i)} + c_j\mathbf{y}_n^{(j)}$ for $\mathbf{y}_n$ in the homogeneous equation associated with (1.4), we get

$$\mathbf{x}_{n+1} = c_i\mathbf{y}_{n+1}^{(i)} + c_j\mathbf{y}_{n+1}^{(j)} = c_iA\mathbf{y}_n^{(i)} + c_jA\mathbf{y}_n^{(j)}.$$

Since $\mathbf{y}_n^{(i)}$ and $\mathbf{y}_n^{(j)}$ are complementary solutions, they satisfy (1.4) with $\mathbf{f}_n = \mathbf{0}$ so that

$$\mathbf{x}_{n+1} = A[c_i\mathbf{y}_n^{(i)} + c_j\mathbf{y}_n^{(j)}] = A\mathbf{x}_n.$$

□

THEOREM 16. *(Superposition Principle II) If $\mathbf{y}_n^{(i)}$ is a complementary solution of the linear system (1.4) and $\mathbf{z}_n$ is any particular solution of the inhomogeneous equation (1.4), then $c_i\mathbf{y}_n^{(i)} + \mathbf{z}_n$ is the solution of (1.4) for any constant c_i.*

PROOF. (exercise) □

For the RBC problem, n of (1.4), the vector forcing term $\mathbf{f}_n$ in (1.4) does not vary with n. The results for second order linear equations with stage independent forcing suggest that we consider a time-invariant particular solution: $\mathbf{z}_n = \mathbf{z}$ for which (1.4) becomes $(A - I)\mathbf{z} = -\mathbf{f}_n$ or

$$\mathbf{z} = (I - A)^{-1} \begin{pmatrix} 0 \\ \gamma R_c \end{pmatrix} \tag{1.7}$$

as long as the determinant of $(I - A)$ does not vanish so that $I - A$ is invertible.

For our RBC problem, a particular solution for the normal case (for which the determinant of $I - A$ is not zero) is the solution of

$$(I - A)\mathbf{z} = \begin{bmatrix} \alpha & -\beta \\ \gamma & \varepsilon \end{bmatrix} \mathbf{z} = \begin{pmatrix} 0 \\ \gamma R_c \end{pmatrix}$$

or

$$\mathbf{z}_n = \mathbf{z} = \frac{R_c}{\beta\gamma + \alpha\varepsilon} \begin{pmatrix} \beta\gamma \\ \alpha\gamma \end{pmatrix}. \tag{1.8}$$

We summarize this result for the RBC problem in the following proposition:

PROPOSITION 7. *For the RBC problem, the matrix $A - I$ is not singular (or, equivalently, 1 is not an eigenvalue of A as explained in the next section) and a particular solution of the governing difference equation (1.4) is stage invariant. This particular solution given by (1.7) is also a fixed point of the linear difference equation (1.4).*

For the exceptional case when the determinant of $(I - A)$ does vanish (which may happen for other coefficient matrix but not the one for the RBC problem), we may apply a vector version of reduction of order or an analogue of the method of undetermined coefficients for the exceptional case of a second order equation. For the latter approach, we take

$$\mathbf{z}_n = n\mathbf{c}_1 + \mathbf{c}_0 \tag{1.9}$$

so that (1.4) becomes

$$(n + 1)\,\mathbf{c}_1 + \mathbf{c}_0 = nA\mathbf{c}_1 + A\mathbf{c}_0 + \begin{pmatrix} 0 \\ \gamma R_c \end{pmatrix}$$

resulting in the two requirements

$$(I - A)\mathbf{c}_1 = 0, \qquad (I - A)\mathbf{c}_0 + \mathbf{c}_1 = \begin{pmatrix} 0 \\ \gamma R_c \end{pmatrix}$$

With the vanishing of the determinant of $(I - A)$, we have a nontrivial solution $\mathbf{c}_1 = C_1\mathbf{u}$ for the homogeneous equation for $\mathbf{c}_1$ above, determined up to (at least) one multiplicative constant C_1 (see Appendix 1). The second condition then determines constant C_1 and the vector unknown $\mathbf{c}_0 = C_2\mathbf{w} + \mathbf{t}$. the latter up to the multiplicative constant C_2. This solution process (to be illustrated by the exercise below) can always be executed unless $\lambda = 1$ is a multiple eigenvalue of A. *Eigenvalues* of matrices will be discussed in the next section. For the RBC problem, a

multiple eigenvalue of A is the same as a multiple root of characteristic equation for the second order difference equation for the same problem upon application of the method of elimination.

EXERCISE 8. *Find a particular solution for (1.4) with one free constant available in*

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{f}_n = \begin{pmatrix} \overline{f} \\ \overline{g} \end{pmatrix}$$

using the assummed form (1.9) for a particular solution.

2. The Matrix Eigenvalue Problem

Having obtained a particular solution for the problem, we know from Superposition Principle II that we need only to obtain two (linearly) independent complementary solutions to complete the solution process. Two distinct complementary solutions are needed so that the solution $\mathbf{y}_n$ can satify the two initial conditions. As in the scalar growth model discussed in Chapter 1, the recursive relation (1.4) can again be used to generate the solution $\mathbf{y}_n$ for any $n > 0$. But as seen from the Fibonacci rabbits model and the plant growth model, much more insight can be gained when the complementary solutions of the problem are given as powers of some amplification factor(s).

2.1. Characteristic Values. From the solution obtained by the method of elimination for the Fibonacci rabbits problem and the plant growth problem, we expect a complementary solution for Rn and Mn to be proportional to λ^n for some amplication factor λ. Therefore, we take the complementary solution of (1.4) to be in the form

$$\mathbf{x}_n = \lambda^n \mathbf{v}$$

for some constant λ where $\mathbf{x}_n$ and $\mathbf{v}$ are two component vectors for the RBC problem. For such a solution, (1.4) becomes

$$\mathbf{v}\boldsymbol{\lambda}^{n+1} = A\mathbf{v}\boldsymbol{\lambda}^n \qquad \text{or} \qquad [A - \lambda I]\mathbf{v} = \mathbf{0}. \tag{2.1}$$

which is a homogeneous linear system of equations for the components of the unknown vector $\mathbf{v}$. Such a homogeneous system has a nontrivial solution only if the determinant of the corresponding coefficient matrix vanishes (see Appendix 2).

For a general $m \times m$ coefficient matrix A, this condition requires

$$P_m(\lambda) = (-)^m \det |A - \lambda I| = 0. \tag{2.2}$$

where the degree of the characteristic polynomial m is equal to the number of unknowns in the vector $\mathbf{x}_n$. In other words, all the equations constituent equation of the linear system are not completely distinct, i.e., not linearly independent. Equation (2.2) is known as the *characteristic equation* of the matrix A and the roots of (2.2) are known as the *eigenvalues* of A (corresponding to the characteristic roots in previous chapters). For each distinct eigevnalue λ_k, the homogeneous linear system $[A - \lambda_k I]\mathbf{v}^{(k)} = \mathbf{0}$ has a solution $\mathbf{v}^{(k)}$ determined up to (at least) one multiplicative constant. The solution $\mathbf{v}^{(k)}$ is known as the eigenvector of A associated with that eigenvalue λ_k.

Correspondingly, the homogeneous system $\mathbf{x}_{n+1} = A\mathbf{x}_n$ associated with (1.4) has a solution proportional to $\mathbf{v}^{(k)}\lambda_k^n$. If all m eigenvalues are distinct, we have

then a complete set of m eigenvectors $\{\mathbf{v}^{(k)}\}$ and hence m distinct complementary solution for (1.4). By Superposition Principle I, the linear combination

$$\mathbf{x}_n = C_1\mathbf{v}^{(1)}\boldsymbol{\lambda}_1^n + C_2\mathbf{v}^{(2)}\boldsymbol{\lambda}_2^n + \cdots + C_m\mathbf{v}^{(m)}\boldsymbol{\lambda}_m^n \tag{2.3}$$

for an arbitary set of constants$\{C_1, ..., C_m\}$ is also a complementary solution of the system (1.4).

For the RBC problem, the characteristic equation is

$$\begin{vmatrix} 1-\alpha-\lambda & \beta \\ -\gamma & 1-\varepsilon-\lambda \end{vmatrix} = 0$$

giving the following two eigenvalues:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \frac{1}{2}\left[(2-\alpha-\varepsilon) \pm \sqrt{(\alpha-\varepsilon)^2 - 4\gamma\beta}\right]. \tag{2.4}$$

Up to a multiplicative constant, the corresponding eigenvectors are

$$\mathbf{v}^{(k)} = \begin{pmatrix} 1 \\ \frac{\gamma}{1-\varepsilon-\lambda_k} \end{pmatrix}, \qquad (k=1,2). \tag{2.5}$$

The complementary solution with two constants for the initial conditions is

$$\mathbf{x}_n = C_1\begin{pmatrix} 1 \\ \frac{\gamma}{1-\varepsilon-\lambda_1} \end{pmatrix}\boldsymbol{\lambda}_1^n + C_2\begin{pmatrix} 1 \\ \frac{\gamma}{1-\varepsilon-\lambda_2} \end{pmatrix}\boldsymbol{\lambda}_2^n \tag{2.6}$$

We note in passing that a second order difference equation such as (3.1) for the plant growth problem may be written as a linear syste by setting

$$\mathbf{x}_n = \begin{pmatrix} p_n \\ p_{n+1} \end{pmatrix}$$

and write the single second equation as

$$\mathbf{x}_{n+1} = \begin{bmatrix} 0 & 1 \\ a & b \end{bmatrix}\mathbf{x}_n. \tag{2.7}$$

Readers should verify that the characteristic equation for the vector difference equation (2.7) is the same as the one for (3.1) in the last chapter.

3. The IVP for RBC

3.1. Solution for the IVP. With $\mathbf{x}_n$ given by (2.6) and $\mathbf{z}_n = \mathbf{z}$ by (1.8), the linear combination $\mathbf{y}_n = \mathbf{x}_n + \mathbf{z}_n$ is also a solution of the difference equation (1.4) for the RBC problem by Superposition Principle II:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{x}_n + \mathbf{z}_n \\ &= C_1\mathbf{v}^{(1)}\boldsymbol{\lambda}_1^n + C_2\mathbf{v}^{(2)}\boldsymbol{\lambda}_2^n + \frac{R_c}{\beta\gamma+\alpha\varepsilon}\begin{pmatrix} \beta\gamma \\ \alpha\gamma \end{pmatrix}. \end{aligned}$$

The eigenvalues $\{\boldsymbol{\lambda}_k\}$ and eigenvector $\left\{\mathbf{v}^{(k)}\right\}$ are given in (2.4) and (2.5), respectively. For the solution of the corresponding IVP, the two unknown constants $\{C_k\}$ are chosen to satisfy the two initial conditions

$$\mathbf{y}_0 = \begin{pmatrix} R_0 \\ M_0 \end{pmatrix} = C_1\mathbf{v}^{(1)} + C_2\mathbf{v}^{(2)} + \mathbf{z} = \begin{pmatrix} \bar{R} \\ \bar{M} \end{pmatrix},$$

or

$$\left[\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right]\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \bar{R} \\ \bar{M} \end{pmatrix} - \mathbf{z} \equiv \boldsymbol{\upsilon}.$$

With $P = \left[\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right]$,

$$P\mathbf{C} = \boldsymbol{v} = \begin{pmatrix} \bar{R} \\ \bar{M} \end{pmatrix} - \frac{R_c}{\beta\gamma + \alpha\varepsilon} \begin{pmatrix} \beta\gamma \\ \alpha\gamma \end{pmatrix}$$

where the components of $\boldsymbol{v}$ are known quantities. Since the eigenvectors are distinct in the sense that none of them is a linear combination of the others (i.e., the eigenvectors $\{\mathbf{v}^{(k)}\}$ are linearly independent), the *modal matrix* P (with the eigenvectors as its columns) is invertible. Hence, the linear system of simultaneous equations has a unique solution for the unknown components $\{C_k\}$ of $\mathbf{C}$:

$$\mathbf{C} = P^{-1}\boldsymbol{v}.$$

3.2. The Mature Adult Case. Even without an explicit solution for the IVP, a number of observations can be made of the evolving RBC population and the bone marrow mass. For mature adult, bone marrow remains more or less unchanged with time so that we can take $\varepsilon = 0$. This simplifies the eigenvalues to

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \frac{1}{2}\left[(2-\alpha) \pm \sqrt{\alpha^2 - 4\gamma\beta}\right], \qquad \mathbf{z}_n = R_c \begin{pmatrix} 1 \\ \alpha/\beta \end{pmatrix}$$

With the half life of a RBC being 60 days, 1/120 of the total RBC in the body is lost each day on the average. Under normal conditions, we have $\alpha \simeq 1/120$ (and hence $\alpha^2 = 6.94 \times 10^{-5}$) and $4\gamma\beta = 3.46 \times 10^{-5}$ so that $\alpha^2 > 4\gamma\beta$. In that case, we have $0 < \lambda_2 < \lambda_1 < 1$ with

$$\mathbf{y}_n = \begin{pmatrix} R_n \\ M_n \end{pmatrix} \to \frac{R_c}{\beta\gamma + \alpha\varepsilon} \begin{pmatrix} \beta\gamma \\ \alpha\gamma \end{pmatrix} = R_c \begin{pmatrix} 1 \\ \alpha/\beta \end{pmatrix}$$

as $n \to \infty$. Thus, the RBC count R_n tends to its maintenance count R_c as we would like to see. However, the steady state bone marrow mass is much too high; it probably means that we have underestimated β. Afterall, not all available bone marrow is dedicated to replenishing daily RBC losses. For example, the same bone marrow tissues are also responsible for producing white blood cells and platelets that help with blood clotting. Refinement of the mathematical model to address this and other deficiencies will be left as an exercise.

3.3. The Aging Adult Case. Aging adults are known to suffer a high rate of bone marrow degradation corresponding to a non-negligible loss fraction ε. As a consequence, less RBC are produced to replenish the degraded RBC. In most case, the usual mechanism for replenishing the lost bone marrow through the $(R_c - R_n)$ differential is less effective for the older folks which means a smaller value of γ than normal.

For γ sufficiently small (whether it is due to aging or some illness such as leukemia, lymphoma, etc.) so that $(\alpha - \varepsilon)^2 > 4\gamma\beta$, the two eigenvalues

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \frac{1}{2}\left[(2-\alpha-\varepsilon) \pm \sqrt{(\alpha-\varepsilon)^2 - 4\gamma\beta}\right]. \tag{3.1}$$

continues to be (positive and) less than unity with $0 < \lambda_2 < \lambda_1 < 1$. In that case, the fixed point of the model system remains asymptotically stable with

$$\mathbf{y}_n = \begin{pmatrix} R_n \\ M_n \end{pmatrix} \to \frac{R_c}{\beta\gamma + \alpha\varepsilon} \begin{pmatrix} \beta\gamma \\ \alpha\gamma \end{pmatrix}$$

For γ sufficiently small, the common factor $R_c/(\beta\gamma + \alpha\varepsilon)$ may be approximated by $R_c/\alpha\varepsilon$ to give

$$\lim_{n\to\infty} [\mathbf{y}_n] \simeq \frac{R_c}{\alpha\varepsilon} \begin{pmatrix} \beta\gamma \\ \alpha\gamma \end{pmatrix}$$

In that case, both RBC (R_n) and bone marrow (M_n) tend to zero with γ and some clinical intervention such as drug treatent, blood infusion and/or bone marrow transplant is needed to arrest further deterioration of the individual's health.

EXERCISE 9. *Formulate a linear difference equation model that allows for daily blood infusion to replenish lost RBC.*

4. Drug Uptake I

4.1. The Question. Clinical intervention often involves drug treatment. The rate of uptale of a drug by different body parts is important to the patient, the physician and the pharmaceutical company producing the drug. Knowledge of the distribution of natural biochemcal substances such as hormones and lipoproteins in the body is also important to health maintenance Tracking such uptakes and distributions is often done with radioactive tracers with the substance of interest tagged radioactively and inserted into the body, usually by injection into the blood stream. Blood is then withdrawn at regular time intervals to check the level of radioactivity and thereby providing a way to obtain information of interest, whether it is the absorption or degradation by organs (such as kidney and liver) or excretion along with the usual wastes.

For a first exposure to this area of investigation, we consider here the following very simple problem. A drug is administered daily (say by injection into the blood stream) at the rate of D_n units/day. It is distributed within the body and partially absorbed by the body tissue through a blood tissue exchange. The amount of drug in the blood is excreted through urination at the rate of u per day and that in the tissue is absorbed by the tissue cells and excreted through sweat at a combined rate s per day. Of interest is the amount of the drug retained in (and consumed biochemically by) the body.

4.2. The Mathematical Model. Let B_n and T_n be the amount of the drug in the blood and tissue of the body on day n, respectively. From the information above, the rate of change of the drug in the blood system depends on the ingestion rate D_n, the excretion rates u and s and the exchange with the body tissue system. Without additional information on the exchange process, we take the simplest approach by assuming that it is a linear process so that we can write

$$B_{n+1} - B_n = D_n - uB_n - k_{21}B_n + k_{12}T_n.$$

where k_{21} is the *rate constant* for the transfer of drug from blood to tissue per unit blood volume and k_{12} is the rate constant for the corresponding transfer from tissue to blood per unit tissue volume. The exchange rate constants are normally non-negative. Since we are interested in the case of drug absorption by tissues (possibly of a particular organ), we would have $k_{21} > k_{12}$.

Similarly, the rate of change of the drug in the tissue system is given by

$$T_{n+1} - T_n = -sT_n + k_{21}B_n - k_{12}T_n.$$

The term $-sT_n$ on the right corresponds to a loss of some of the drug entering the tissue cells through biochemical reactions (to accomplish the purpose of the clinical

intervention) and possibly through sweat (which does not contribute to the clinical purpose of the drug intake). In calculating the daily change of B_n and T_n, it has been assumed that there is no loss in the exchange process.

We can simplify the two equations for the dialy change of drug level by combining all terms proportional to B_n and those to T_n to get

$$B_{n+1} = D_n + (1 - u - k_{21})B_n + k_{12}T_n \tag{4.1}$$

$$T_{n+1} = (1 - s - k_{12})T_n + k_{21}B_n \tag{4.2}$$

Regarding the biochemical kinetics characterized by the rate constants k_{ij}, we emphasize that the kinetic exchange is not necessarily symmetric and k_{12} is generally not equal to (and usually less than) k_{21}.

The two first order difference equations are supplemented by two initial conditions. Prior to the administration of the drug, there is no drug concentration in the blood stream or the body tissue. Hence, we have

$$B_0 = T_0 = 0. \tag{4.3}$$

While both difference equations are first order, each equation involves both unknowns. In general, they cannot solved separately for one unknown without simultaneiously considering the other equation. The only exception is when $k_{12}k_{21} = 0$. It is instructive to consider first the special case $k_{21} = 0$ before treating the general case of $k_{21} \neq 0$.

4.3. The Uncoupled System.

4.3.1. $k_{21} = 0$. For the special case $k_{21} = 0$, the relation (4.2) becomes a first order linear difference equation for T_n alone:

$$T_{n+1} = (1 - s - k_{12})T_n$$

Together with the initial condition $T_0 = 0$, it requires

$$T_n = 0 \qquad (n = 0, 1, 2, ...) \tag{4.4}$$

If there is no transfer of drug from blood to body tissues and drug entering the body only through injection into blood stream, there cannot be drug in the body tissue if there is none initially.

In that case, the equation for B_n simplifies to

$$B_{n+1} = D_n + (1 - u)B_n$$

independent of the value of k_{21}. The solution of this equation with the initial condition $B_0 = 0$ is (by the method of solution for single first order linear inhomogeneous difference equation of Proposition 2)

$$B_{n+1} = \sum_{k=0}^{n} (1 - u)^{n-k} D_k, \qquad B_0 = 0. \tag{4.5}$$

where $0 < 1 - u < 1$. Note that at a later stage $N + 1$, the drug concentration in the blood stream from an earlier drug dosage injected at stage k is reduced to $(1 - u)^{N-k} D_k$.

4.3.2. $k_{12} = 0$. For the special case $k_{12} = 0$, the relation (4.1) becomes a first order linear difference equation for B_n alone:

$$B_{n+1} = D_n + (1 - u - k_{21})B_n$$

Together with the initial condition $B_0 = 0$, the solution of the IVP for B_n is

$$B_{n+1} = \sum_{k=0}^{n} (1 - u - k_{21})^{n-k} D_k, \qquad B_0 = 0.$$

If k_{12} also vanishes, the solution reduces to (4.5) and (4.4). On the other hand, for $k_{12} \neq 0$, we have from (4.2) in the form

$$T_{n+1} - (1 - s)T_n = k_{21}B_n \tag{4.6}$$

and $T_0 = 0$, we have a non-zero solution for T_n.

EXERCISE 10. *Obtain the solution of the IVP defined by (4.6) and* $T_0 = 0$.

With a non-zero transfer of drug from blood stream, there should be a noticeable amount of drug in the body tissue after a few days.

4.4. The Coupled System. When $k_{12}k_{21} \neq 0$, the two difference equations must be solved simultaneously. Since there are only two unknowns for this problem, the method of elimination employed for the plant growth problem in the previous chapter is feasible. Since $k_{21} \neq 0$, we can use the second (4.2) to express B_n in terms of T_n:

$$B_n = \frac{1}{k_{21}} \left[T_{n+1} - (1 - s - k_{12})T_n \right]. \tag{4.7}$$

We then use (4.7) to eliminate B_n from (4.1) to get

$$T_{n+2} = \alpha T_{n+1} - \beta T_n + k_{21}D_n \tag{4.8}$$

where

$$\alpha = (2 - s - u - k_{21} - k_{12}) \tag{4.9}$$

$$\beta = (1 - u)(1 - s) - k_{21}(1 - s) - k_{12}(1 - u) \tag{4.10}$$

For our original two simultaneous equation formulation, we have two natural initial conditions given in (4.3). But if we wish to work with (4.8), we would need the initial conditions on T_0 and T_1. We have T_0 from (4.3). For T_1, we make use of the (4.2) which can be re-arranged to read

$$T_{n+1} = (1 - s - k_{12})T_n + k_{21}B_n. \tag{4.11}$$

With $B_0 = T_0 = 0$, we get $T_1 = (1 - s - k_{12})T_0$ and therewith

$$T_0 = 0, \qquad T_1 = 0. \tag{4.12}$$

Equation (4.8) is structurally the same as the governing difference equation for the Fibonacci rabbits with immigration. When the daily dosage of drug intake is the same from day to day, the use of Superposition Principle II as in the solution process for the plant growth problem is also effective for an elementary function solution in terms for the present problem. However, if D_n varies with n and its dependence on n is not given by some elementary functions, the determination of a particular solution is not so straightforward; the method of undetermined

coefficients is not readily applicable. The most direct method of solution would be to use the governing difference equation (4.8) recursively, as in the problem of linear growth with immigration in Chapter 1. We use this approach here to show the kind of solution we get by such an approach, a kind that is not particularly informative about the solution behavior especially for large n.

With the two initial conditions (4.12), we can use (4.8) to calculate T_k sequentially:

$$\begin{aligned}
T_2 &= k_{21}D_0, \qquad T_3 = \alpha T_2 + k_{21}D_1 = k_{21}(D_1 + \alpha D_0),\\
T_4 &= \alpha T_3 - \beta T_2 + k_{21}D_2 = k_{21}(D_2 + \alpha D_1 + \alpha^2 D_0) - \beta k_{21}D_0\\
T_5 &= \alpha T_4 - \beta T_3 + k_{21}D_3 = k_{21}\sum_{i=0}^{3}\alpha^i D_{3-i} - k_{21}\beta(D_1 + 2\alpha D_0)\\
T_6 &= \alpha T_5 - \beta T_4 + k_{21}D_4 = k_{21}\left\{\sum_{i=0}^{4}\alpha^i D_{3-i} - \beta\sum_{i=0}^{2}(i+1)\alpha^i D_{2-i} + \beta^2 D_0)\right\}\\
&\cdots\cdots\cdots .
\end{aligned}$$

The corresponding solutions for B_n are

$$\begin{aligned}
B_0 &= 0, \qquad B_1 = D_0, \qquad B_2 = D_1 + (1 - u - k_{21})D_0\\
B_3 &= D_2 + (1 - u - k_{21})D_1 + (1 - u - k_{21})^2 D_0 + k_{21}k_{12}D_0\\
&\cdots\cdots\cdots
\end{aligned}$$

A general expression for T_{n+2} and B_{n+2} can be obtained by induction but will not be given here since actual solution for a particular case will likely have to be calculated using MatLab or other mathematical software.

For the special case $D_n = D$, the solution (**??**) simplifies considerably with

$$\begin{aligned}
T_0 &= T_1 = 0, \qquad T_2 = k_{21}D, \qquad T_3 = k_{21}D(1 + \alpha),\\
T_4 &= k_{21}D\left\{(1 + \alpha + \alpha^2) - \beta\right\}, \qquad \cdots\cdots\\
B_0 &= 0, \qquad B_1 = D, \qquad B_2 = (2 - u - k_{21})D,\\
&\cdots\cdots
\end{aligned}$$

etc. With some oversimplifications, a general expression for T_n or B_n is in most cases merely more compact without much insight to the nature of the solution (other than the possibility of convergence and boundedness of the solution in the limit as $n \to \infty$). With the computing power available today, it is still significant that we can always get the solution of a problem involving difference equations by calculating the solution for each stage recursivelyusing the difference equations of the mathematical model.

5. Drug Uptake II

5.1. The Problem. Suppose in addition to the blood tissue exchange, there is also an exchange between body tissue and bone (but not between bone and blood directly). Unlike the case of blood and tissue, there is no excretion of the drug in bone parts Otherwise, the biochemical set up is as in the previous section. Of interest is again the amount of the drug retained in (and consumed biochemically by) the body?

5.2. The Mathematical Model. As in the previous section, let B_n and T_n be the amount of the drug in the blood and tissue of the body on day n, respectively. In addition, let B_n^* be the amount of the drug in the body bone on day n. By the statement of the problem above, the rates of change of the drug in the blood, tissue and bone depend on the daily drug intake rate, the excretion/absorption rates and the various assumed exchanges in the body. With the additional information on the exchange process stated above, the equation governing the change of drug concentration in blood stream remains as before:

$$B_{n+1} = D_n + (1 - u - k_{21})\, B_n + k_{12} T_n. \tag{5.1}$$

However, the change of the drug in the tissue system must now be modified to include the exchange between tissue and bone to get

$$T_{n+1} = (1 - s - k_{12} - k_{32}) T_n + k_{21} B_n + k_{23} B_n^*. \tag{5.2}$$

In addition, we now have a third equation for the change in the body bone:

$$B_{nn+1}^* = (1 - k_{23}) B_n^* + k_{32} T_n. \tag{5.3}$$

Here, we assumed that bone does not absorb or excrete the drug intake through the tissue-bone exchange.

The three equations are supplemented by initial condition that there is no drug anywhere in the body before the ingestion of the drug so that

$$B_0 = T_0 = B_0^* = 0. \tag{5.4}$$

Again the exchanges need not be symmetric so that the coefficents k_{ij} may not be equal to k_{ji}.

5.3. Vector Form. The solution of the initial value problem defined by (5.1) - (5.4) can again be obtained by the method of elimination used for the previous problem but is more tedious. For other problems with even more unknowns, the elimination process is impractical and the use of matrix algebra becomes in dispensible. To apply matrix methods, we write the equations for the three-component drug absorption problem as the vector difference equation (1.4) now with

$$A = \begin{bmatrix} k_{11} & k_{12} & 0 \\ k_{21} & k_{22} & k_{23} \\ 0 & k_{32} & k_{33} \end{bmatrix}, \qquad \mathbf{y}_n = \begin{pmatrix} B_n \\ T_n \\ B_n^* \end{pmatrix}, \qquad \mathbf{f}_n = \begin{pmatrix} D_n \\ 0 \\ 0 \end{pmatrix}. \tag{5.5}$$

where

$$k_{11} = 1 - u - k_{21}, \qquad k_{22} = 1 - s - k_{12} - k_{32}, \qquad k_{33} = 1 - k_{23}. \tag{5.6}$$

We are particularly interested in cases where $\mathbf{f}_n$ depends on n, i.e., $\mathbf{f}_n$ is *not* stage invariant, with a more informative solution than the one obtained in the previous section for the two unknown case. Our approach here is again to make use of the two superposition principles.

For an adequate set of complementary solutions of the three unknown drug uptake problem, we need the eigenvalues and eigenvectors, or more briefly, the eigen-pairs, of the matrix A in (5.5). The characteristic equation in this case is

$$(-)^3 \det |A - \lambda I| = \lambda^3 - \alpha\lambda^2 + \beta\lambda - \gamma = 0$$

with

$$\begin{aligned}\alpha &= k_{11}+k_{22}+k_{33}=tr(A),\\ \beta &= (k_{11}k_{22}+k_{33}k_{11}+k_{22}k_{33})-(k_{12}k_{21}+k_{32}k_{23}),\\ \gamma &= \det(A)=k_{11}k_{22}k_{33}-k_{11}k_{32}k_{23}-k_{33}k_{12}k_{21}.\end{aligned}$$

The two superposition principles enable us to determine separately an adequate number of linearly independent complementary solutions and one particular solutions of (1.4) and combine them to get (for the case of three distinct eigenvalues)

$$\mathbf{y}_n{=}C_1\mathbf{v}^{(1)}\boldsymbol{\lambda}_1^n+C_2\mathbf{v}^{(2)}\boldsymbol{\lambda}_2^n+C_3\mathbf{v}^{(3)}\boldsymbol{\lambda}_3^n+\mathbf{z}_n$$

where $\boldsymbol{\lambda}_k$ and $\mathbf{v}^{(k)}$, $k=1,2,3$, are the three distinct eigen-pairs for our three unknown drug uptake problem. By setting the three constants to the values given by

$$\begin{pmatrix} C_1\\ C_2\\ C_3\end{pmatrix}=-\left[\mathbf{v}^{(1)},\mathbf{v}^{(2)},\mathbf{v}^{(3)}\right]^{-1}\mathbf{z}_0,$$

$\mathbf{y}_n$ satisfies the three scalar initial conditions (5.4) and constitutes the unique solution of the IVP.

The following proposition and its proof are analogous to the corresponding existence and uniqueness theorem for a single linear difference equation in chapter 1:

PROPOSITION 8. *The IVP defined by (1.4), (5.5), (5.6) and (5.4) has a unique solution.*

5.4. The Blood-Tissue Exchange Problem. To illustrate the advantage of the use of eigen-pairs for linear difference equations, we return to the two-compartment model for the drug exchange problem in Section 4 written in the vector form (1.4),

$$\mathbf{y}_{n+1}=A\mathbf{y}_n+\mathbf{f}_n,$$

where

$$\mathbf{y}_n=\begin{pmatrix}B_n\\ T_n\end{pmatrix},\quad \mathbf{f}_n=\begin{pmatrix}D_n\\ 0\end{pmatrix},\quad A=\begin{bmatrix}k_{11}&k_{12}\\ k_{21}&k_{22}\end{bmatrix},\tag{5.7}$$

with

$$k_{11}=1-u-k_{21},\qquad k_{22}=1-s-k_{12}.\tag{5.8}$$

The eigenvalues for the matrix A is determined by

$$\det|A-\lambda I|=\lambda^2-(k_{11}+k_{22})\lambda+(k_{11}k_{22}-k_{12}k_{21})=0$$

to be

$$\begin{aligned}2\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} &= tr[A]\pm\sqrt{(tr[A])^2-4\det[A]}\\ &= (k_{11}+k_{22})\pm\sqrt{(k_{11}-k_{22})^2+4k_{12}k_{21}}\end{aligned}\tag{5.9}$$

The solution of the vector difference equation is therefore

$$\mathbf{y}_n=C_1\mathbf{v}^{(1)}\boldsymbol{\lambda}_1^n+C_2\mathbf{v}^{(2)}\boldsymbol{\lambda}_2^n+\mathbf{z}_n,\tag{5.10}$$

where $\left\{\mathbf{v}^{(k)}\right\}$ are the eigenvectors and $\mathbf{z}_n$ is any particular solution. For the solution of the corresponding IVP, the two unknown constants $\{C_k\}$ are chosen to satisfy the two initial conditions $B_0 = T_0 = 0$:

$$\mathbf{y}_0 = \begin{pmatrix} B_0 \\ T_0 \end{pmatrix} = C_1\mathbf{v}^{(1)} + C_2\mathbf{v}^{(2)} + \mathbf{z}_0 = \mathbf{0},$$

or

$$\left[\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right] \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = -\,\mathbf{z}_0.$$

With $P = \left[\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right]$, we get

$$\mathbf{C} = -P^{-1}\mathbf{z}_0\ .$$

Even without explicit expressions for the eigenvectors and the constants $\{C_k\}$ in the solution (5.10), a great deal of information about the solution behavior can be learned from the eigenvalues given that (i) $k_{12} > 0$ and $k_{21} > 0$, and (2) $k_{11} < 1$ and $k_{22} < 1$ (since u and s are both positive):

- From (5.9), we see that both eigenvalues are real whatever the relative magnitude of the four quantities k_{ij} may be.
- If $k_{11} > 0$ and $k_{22} > 0$, then both eigenvalues are positive if $k_{11}k_{22} - k_{12}k_{21} > 0$.
- If $0 < k_{11}\ (< 1)$, $0 < k_{22}\ (< 1)$, and $k_{11}k_{22} - k_{12}k_{21} > 0$, then $0 < \lambda_1 < 1$ and $0 < \lambda_2 < 1$.

With the observations above, we have the following implication on the long term behavior of drug intake:

PROPOSITION 9. *If the various rate constants result in $|\lambda_k| < 1$ for $k = 1, 2$, then $\mathbf{y}_n \simeq \mathbf{z}_n$ for sufficiently large n. In addition, if $D_n = D$ is independent of n, then as $n \to \infty$*

$$\mathbf{y}_n \to (I - A)^{-1} \begin{pmatrix} D \\ 0 \end{pmatrix} = \frac{D}{\Delta} \begin{pmatrix} s + k_{12} \\ u + k_{21} \end{pmatrix}$$

where

$$\begin{aligned} \Delta &= \begin{vmatrix} 1 - k_{11} & -k_{12} \\ -k_{21} & 1 - k_{22} \end{vmatrix} = (u + k_{21})(s + k_{12}) - k_{21}k_{12} \\ &= us + k_{21}s + k_{12}u. \end{aligned}$$

It is significant that, as $n \to \infty$, the limiting values of B_n and T_n are finite and positive. In fact, we have for $k_{12}k_{21} > 0$

$$\frac{D}{\Delta} \begin{pmatrix} s + k_{12} \\ u + k_{21} \end{pmatrix} = \frac{D}{us + k_{21}s + k_{12}u} \begin{pmatrix} s + k_{12} \\ u + k_{21} \end{pmatrix}.$$

With the four parameters k_{12}, k_{21}, u and s being fractions and hence all less than 1, the ratios $(s + k_{12})/\Delta$ and $(u + k_{21})/\Delta$ are generally greater than 1. It follows that the steady state drug concentrations in the blood stream and body tissue are both higher than the amount ingested daily, with

$$B_\infty > D, \qquad T_\infty > D$$

CHAPTER 8

Markov Chains

1. A Forest of Red Oaks and White Pines

Systems of first order linear difference equations arise naturally in the studies of many biological phenomena (other than population growth) for which measurements or observations are available in discrete time units (in generations, years, months, days, hours, minutes, seconds or down to pico (10^{-12}) seconds). We have already encountered such systems in red blood cells production and drug uptake problems. In this chapter, we focus on a special kind of first order linear systems of difference equations known as (first order) *Markov chains.* They are characterized by the dependence of the current state of the biological system only on that of the immediate past, i.e., the state of the previous time unit. We begin with the following simple example to illustrate the nature of such evolving processes:

EXAMPLE 4. *Consider a huge forest comprising of red oaks and white pines, with one or the other tree type found at any particular tree location inside this forest. Both types of trees have (roughly) the same life span. When a tree dies, the replacement tree grown in that location may be of the same type or the other type. Available data show that when a red oak (R) die, it is equally likely replaced by another red oak or by a white pine (W) but the replacement for a white pine is more likely (by a ratio of 3 to 1) to be a red oak. Of interest is how does the forest evolve and what is the distribution of red oaks and white pines after many generations?*

Here we treat the replacement process as a Markov chain, with each stage corresponding to one tree generation. There are only two possible (tree) states at a particular location in the forest at any particular stage, *red oak* and *white pine.* We denote by $x_1(n)$ and $x_2(n)$ the two components of the state vector $\mathbf{x}(n) = (x_1(n), x_2(n))^T$ (with the superscript T indicating the *transpose* of a matrix). To avoid the confusion from multiple subscripts, we use $\mathbf{y}(n)$ for $\mathbf{y}_n$ whenever the components of the vector state $\mathbf{y}$ are subcripted: $\mathbf{y}(n) = (y_1(n), y_2(n), ..., y_m(n))^T$.

For our purpose, $x_1(n)$ and $x_2(n)$ correspond to the fraction of time the tree at a particular location would be red oak and white pine, respectively (or alternatively, the fraction of the forest being red oak and white pine, respectively). The change of $\mathbf{x}(n)$ from one tree generation to the next is indicated by

$$\begin{array}{c} \\ x_1(n+1) \\ x_2(n+1) \end{array} \begin{array}{c} \begin{array}{cc} x_1(n) & x_2(n) \end{array} \\ \begin{bmatrix} 0.5 & 0.75 \\ 0.5 & 0.25 \end{bmatrix} \end{array}$$

and summarized by the transition matrix

$$M = \begin{bmatrix} 0.5 & 0.75 \\ 0.5 & 0.25 \end{bmatrix} \tag{1.1}$$

.

Note that column entries for both columns sum up to 1 as they correspond to the fractions (probabilities) of the forest in the different tree states. The evolution of the two tree fractions (corresponding to the probability distribution for the tree states) from one generation to the next is governed by

$$\mathbf{x}(n+1) = M\mathbf{x}(n), \qquad (n = 0, 1, 2, \ ....), \tag{1.2}$$

Suppose we have as the initial vector $\mathbf{x}(0) = (1, 0)^T$ corresponding to the tree being all red oak at the start. At the next instance, $n = 1$, we have

$$\mathbf{x}(1) = M\mathbf{x}(0) = M \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \tag{1.3}$$

with the components of the vector for $\mathbf{x}(1)$ giving the fraction of the tree being red oak and white pine for the next generation, respectively. At the next stage $(n = 2)$, we have

$$\mathbf{x}(2) = M\mathbf{x}(1) = \begin{bmatrix} 0.5 & 0.75 \\ 0.5 & 0.25 \end{bmatrix} \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix},$$

giving the fraction of trees being red oaks and white pines in the next generation, respectively. If we continue the process and determine $\mathbf{x}(3),\ \mathbf{x}(4),\ \cdots$, we can get the distribution of $\mathbf{x}(n)$ at any future generation n for the different states of the evolving phenomenon, known as the (*fractional* or) *probability distribution* of the phenomenon for brevity.

Before we investigate further the properties of $\mathbf{x}(n)$, it is important to observe that

- all the elements of the column vectors $\mathbf{x}(0), \mathbf{x}(1), \mathbf{x}(2),\ \cdots$ and the elements of the transition matrix M are all *nonnegative*, consistent with the fact that they are fractions or probabilities, and
- the elements of each column of $\mathbf{x}(n)$ and M sum up to 1, consistent with requirement that all the fractions have to add up to the whole.

To facilitate further mathematical analysis and discussion of problems similar to the Red Oak - White Pine Problem, we introduce the following definitions to be used in these notes:

DEFINITION 11. *A* ***probability vector*** *is a column vector* $\mathbf{p} = (p_1, p_2, ..., p_m)^T$ *with* $0 \leq p_k \leq 1$ *and* $p_1 + p_2 + ... + \ p_m = 1$.

DEFINITION 12. *An* $m \times m$ *matrix* M *is a* ***probability matrix*** *if each of its* m *columns is a probability vector.*

A general *Markov chain* (occasionally abbreviated as MC) is characterized by the IVP for a first order vector difference equation

$$\mathbf{x}(n+1) = \mathbf{x}_\infty, \qquad \mathbf{x}(0) = \mathbf{p} \tag{1.4}$$

where the $m \times m$ *Markov transition matrix* M is a probability matrix and the *initial distribution* $\mathbf{p}$ is a probability vector.

It should be noted that some writers prefer to work with row probability vectors $\mathbf{z}(n) = (z_1(n),\ ...., z_m(n))$. The transition matrix in that case corresponds to the transpose of the transition matrix in (1.2) and the state of the evolving phenomenon is then governed by the relation

$$\mathbf{z}(n+1) = \mathbf{z}(n)M^T, \qquad (n = 0, 1, 2,\ ....)$$

instead of (1.2). We will stay with the column probability vector notation throughout these notes.

LEMMA 2. *Product of two probability matrices is a probability matrix. In particular, any power of a probability matrix is a probability matrix.*

PROOF. (Exercise) □

As a special case (application) of this lemma, the product $\mathbf{q} = M\mathbf{p}$ of a Markov transition matrix M and a probability vector $\mathbf{p}$ is a probability vector.

2. The Steady State for the Red Oak - White Pine Problem

As n increases, the expression for $\mathbf{x}(n)$ as given by

$$\mathbf{x}^{(1)}(n) = M\mathbf{x}(n-1) = \cdot\cdot\cdot\cdot\cdot = M^n\mathbf{p}$$

becomes unwieldy if we have general entries for the transition matrix such as $[m_{ij}]$ instead of numerical entries as in (1.1), even for the simple case of $\mathbf{p} = (1, 0)^T$. It is desirable to have some simple expression for $\mathbf{x}(n)$ for general n. For the simple example of the red oak-white pine forest in the last section, we may use either the method of elimination or the matrix eigen-pair approach. Either method leads to the characteristic equation

$$\lambda^2 - 0.75\lambda - 0.25 = 0,$$

which has the two roots

$$\lambda_1 = -0.25, \qquad \lambda_2 = 1$$

and the corresponding eigenvectors are

$$\mathbf{v}^{(1)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \mathbf{v}^{(2)} = \begin{pmatrix} 0.75 \\ 0.5 \end{pmatrix}.$$

As such, $\mathbf{x}^{(1)}(n) = \mathbf{v}^{(1)}\lambda_1^n$ and $\mathbf{x}^{(2)}(n) = \mathbf{v}^{(2)}\lambda_2^n$ are both complementary solutions of the difference equation (1.2) and so is a linear combination of these two solutions, $c_1\mathbf{x}^{(1)}(n) + c_2\mathbf{x}^{(2)}(n)$. The observation gives rise to the following *superposition principle* (previously discussed in conjunction with the Fibonacci rabbits problems and its extension to a general second order difference equation):

DEFINITION 13.

CONCLUSION 1. *Together, we have for a general solution for* $\mathbf{x}(n)$

$$\mathbf{x}(n) = \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} = c_1(-0.25)^n \begin{pmatrix} 1 \\ -1 \end{pmatrix} + c_2(1)^n \begin{pmatrix} 0.75 \\ 0.5 \end{pmatrix}.$$

The two constants of integration c_1 *and* c_2 *are to be determined by the initial probability distribution* $\mathbf{x}(0)$. *For* $\mathbf{x}(0) = (1, 0)^T$, *we have*

$$c_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} + c_2 \begin{pmatrix} 0.75 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

which can be solved to give

$$c_2 = \frac{4}{5}, \qquad c_1 = \frac{2}{5}$$

so that

$$\begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} = \begin{pmatrix} 2/5 \\ -2/5 \end{pmatrix} (-0.25)^n + \begin{pmatrix} 3/5 \\ 2/5 \end{pmatrix}. \tag{2.1}$$

It follows that

$$\lim_{n\to\infty} [\mathbf{x}(n)] = \begin{pmatrix} 3/5 \\ 2/5 \end{pmatrix} \equiv \mathbf{x}_\infty. \tag{2.2}$$

Conclusion 2. *An important result from the mathematical modeling is the somewhat surprising conclusion that white pine continues to have a significant presence in the forest, notwithstanding red oak being highly favored as its replacement, with its fraction of the whole forest converging to a rather substantial 2/5.*

For Markov chains with n state where n is a large integer, it would be impractical to reduce the linear system to a single n^{th} order linear difference equation for one unknown. It is more efficient to obtain the solution corresponding to (2.1) by way of the eigen-pairs of the transition matrix as done for the simple example above.

3. Fixed Points for General Markov Chains

While we have not encountered problems involving a homogeneous linear system of difference equations that has a non-trivial (non-zero) fixed point, the phenomenon is not so rare in both theory and applications of such linear systems. If $\mathbf{x}_\infty$ is a fixed point of a Markov chain, then

$$\mathbf{x}_\infty = M\mathbf{x}_\infty \qquad \text{or} \qquad (M-I)\mathbf{x}_\infty = \mathbf{0} \tag{3.1}$$

This homogeneous linear system has a nontrivial solution since $det[M-I] = 0$ (given the rows adding up to zero). We have then the following result on the existence of a fixed points for a MC:

Lemma 3. *$\lambda = 1$ is always an eigenvalue of the transition matrix M of a MC with an associated eigenvector $\mathbf{x}_\infty$ which may be made a probability vector by scaling its elements to sum to unity.*

Proof. Let $A = M - I$. Then all the rows of A sum to give a zero row. Hence zero is an eigenvalue of A or $\lambda = 1$ is an eigenvalue of M with an associated eigenvector $c\mathbf{v}$. Choose $c = \mathbf{1}/\sum_{i=1}^{m} v_i$ and take $\mathbf{x}_\infty = c\mathbf{v}$. □

Since omponents of $\mathbf{x}(n)$ are fractions of the same whole, they should not tend to infinity as $n \to \infty$. In other words, the fixed point $\mathbf{x}_\infty$ should not be unstable. In fact, the components of the state vector should not exceed unity. This is reflected in the following lemma:

Lemma 4. *All eigenvalues of a Markov transition matrix M must be ≤ 1 in magnitude.*

PROOF. Suppose λ is an eigenvalue of M and $|\lambda| > 1$. Let $\{\lambda, \mathbf{y}\}$ be an eigenpair of M^T (since λ is also an eigenvalue of M^T). Let $\max_{i=1}^{m}[|y_i|] = Y_j > 0$ Then it follows from $M^T\mathbf{y} = \boldsymbol{\lambda}\mathbf{y}$ that

$$|\lambda y_j| = |\lambda|\, Y_j = \left|\sum_{k=1}^{m} M_{kj} y_k\right| \leq Y_j \sum_{k=1}^{m} M_{kj} = Y_j$$

or $|\lambda| \leq 1$, contradicting the supposition $|\lambda| > 1$. □

To have the MC converging to $\mathbf{x}_\infty$ as $n \to \infty$, we must have two other results: i) $\lambda = 1$ is a simple eigenvalue so that there is only one fixed point, and ii) there are no complex eigenvalues of unit magnitude. The former result would eliminate the uncertainty to which steady state does the Markov chain converge. The latter would eliminate the possibility of the chain evolving cyclically. Unfortunately, neither of these is true for MC in general. Both of the following two matrices,

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

qualify for a Markov transition matrix. For the matrix I, $\lambda = 1$ is a double eigenvalue with two associated distinct eigenvectors $(1,0)^T$ and $(0,1)^T$. On the other hand, the two eigenvalues of the matrix J are $\pm i$. There is however one class of Markov chains, known as *regular Markov chains*, that precludes both possibilities; it includes the Red Oak - White chain as a member.

4. Regular Markov Chains

To introduce the notion of a regular MC, we need the concept of a power-positive matrix.

DEFINITION 14. *A matrix M is a **power-positive** if all elements of M^k are positive, i.e., $M^k = C = [C_{ij}]$ with $C_{ij} > 0$ for all $i, j = 1, ..., m$ and $k \geq k_p > 0$ for some integer $k_p \geq 1$.*

DEFINITION 15. *A Markov chain is **regular** if its $m \times m$ transition matrix M is a power-positive probability matrix.*

It is easily seen that the definition of a regular MC excludes both the identity matrix as well as matrices that include J as a submatrix. On the other hand, the transition matrix (1.1) for the Red Oak - White Pine problem of Example 4 is a power-positive probability matrix for $k_p = 1$. We saw from (2.1) that for the initial data $\mathbf{x}(0) = (1,0)^T$, the forest tends to the steady state distribution $\mathbf{x}_\infty$ as given by (2.2). Should we take $\mathbf{x}(0) = (0,1)^T$, we would get instead of (2.1),

$$c_1 = -\frac{3}{5}, \qquad c_2 = \frac{4}{5}$$

and

$$\begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} = \begin{pmatrix} -2/5 \\ 2/5 \end{pmatrix} (-0.25)^n + \begin{pmatrix} 3/5 \\ 2/5 \end{pmatrix}. \tag{4.1}$$

which has the same limiting behavior (2.2) as $n \to \infty$. In fact, it is easy to show for this example that the limiting value of $\mathbf{x}(n)$ is the same for any initial distribution $\mathbf{x}(0)$. We will see that this very special property and related ones are shared more generally by *regular* MC of which the Red Oak - White Pine chain is a member. To

show this, we begin with the following lemma on the successive states of a *regular* MC once it starts with a probability vector:

LEMMA 5. *If M is the transition matrix of a regular MC and $\mathbf{x}(0) = \mathbf{p}$ is a probability vector, then $\mathbf{x}(n)$ is a positive probability vector for sufficiently large n.*

PROOF. (exercise) □

We are now ready to show in the following sequence of results that a regular MC approaches the same limiting behavior as $n \to \infty$, independent of the initial distribution $\mathbf{p}$. First, we show that the limiting behavior, i.e., the fixed point in this case, is unique.

THEOREM 17. *Suppose that M is the transition matrix of a regular MC has one and only one fixed point $\mathbf{x}_\infty$.*

PROOF. For simplicity, we prove the theorem for $M > O$ (and leave the more general case as an exercise). Let probability vectors $\mathbf{x}_\infty$ and $\mathbf{y}_\infty$ be two fixed points and we know there is at least one by a previous lemma). Let $\mathbf{z}_\infty = \mathbf{x}_\infty - \alpha\mathbf{y}_\infty$ with α chosen so that $\mathbf{z}_\infty$ has at least one zero component with all the others positive. Since $\mathbf{x}_\infty$ and $\mathbf{y}_\infty$ are both fixed points of M, we have

$$M\mathbf{z}_\infty = M\mathbf{x}_\infty - \alpha M\mathbf{y}_\infty = \mathbf{x}_\infty - \alpha\mathbf{y}_\infty = \mathbf{z}_\infty$$

But by a previous exercise, we have (with $M > O$) $M\mathbf{z}_\infty > 0$, contradicting the fact that $\mathbf{z}_\infty$ on the right hand side has a zero component unless $\mathbf{z}_\infty = \mathbf{x}_\infty - \alpha\mathbf{y}_\infty = \mathbf{0}$ or $\mathbf{x}_\infty = \alpha\mathbf{y}_\infty$. In that case, we must have $\alpha = 1$ as $\mathbf{x}_\infty$ and $\mathbf{y}_\infty$ are both probability vectors. It follows that there can only be a unique $\mathbf{x}_\infty$ for any two initial distributions (different or not). □

For the fixed point $\mathbf{x}_\infty$ of a *regular* MC to be asymptotically stable, we still have to eliminate the possibility of the transition matrix having complex eigenvalues of unit magnitude. Here, we do so indirectly by proving the following theorem which constitutes the principal result for *regular* MC. Its somewhat intricate proof relegated ton an appendix of this section.

THEOREM 18. *As $n \to \infty$, the vector sequence $\{\mathbf{x}(n) \equiv \mathbf{x}_n\}$ of a regular MC converges to a limiting vector distribution $\mathbf{x}_\infty$ independent of the initial condition.*

PROOF. (see Appendix of this section). □

Below is an application of the theorem above to eliminate the possibility of the tranisition matrix M having complex eigenvalues of unit magnitude.

LEMMA 6. *The transition matrix M of a regular MC, has no complex eigenvalues with $|\lambda| = 1$.*

PROOF. We take M to be positive to reduce the details of the proof. Suppose μ should be a complex eigenvalue of unit magnitude and $u = v + iw$ is an associated eigenvector with both v and w real. Let $\mathbf{x}_\infty$ be a limiting probability vector of M (assured by Theorem 18) and c be sufficiently large so that both $v + c\mathbf{x}_\infty$ and $w + c\mathbf{x}_\infty$ are both positive vectors. It follows that

$$M(v + iw + c(1+i)\mathbf{x}_\infty) = \mu(v + iw) + c(1+i)\mathbf{x}_\infty$$

with

$$M^n(v + iw + c(1+i)\mathbf{x}_\infty) = \mu^n(v + iw) + c(1+i)\mathbf{x}_\infty. \tag{4.2}$$

As $n \to \infty$, we have from Theorem 18

$$\begin{aligned} M^n(v + iw + c(1+i)\mathbf{x}_\infty) &= M^n((v + c\mathbf{x}_\infty) + i(w + c\mathbf{x}_\infty)) \\ &= M^n(v + c\mathbf{x}_\infty) + iM^n(w + c\mathbf{x}_\infty) \\ &\to \alpha\mathbf{x}_\infty + i\beta\mathbf{x}_\infty \end{aligned}$$

For $|\mu| = 1$, $M^n(v + iw) = \mu^n(v + iw)$ converges as $n \to \infty$ only if $\mu = 1$. □

SUMMARY 1. *A Markov Chain with an $m \times m$ transition matrix M is **regular** if $M^k > O$ for all $k \geq k_p$ for some $k_p \geq 1$. For a regular MC, the following properties hold:*

- If $\mathbf{x}(0) = \mathbf{p}$ is a probability vector, then $\mathbf{x}(n) = M^n\mathbf{p}$ is a **positive** probability when n is sufficiently large ($n \geq k_p$)..
- As $n \to \infty$, $\mathbf{x}(n) \to$ the unique limiting (steady state) distribution $\mathbf{x}_\infty$ which is **independent** of the initial distribution $\mathbf{x}(0) = \mathbf{p}$. Hence $\mathbf{x}_\infty$ is asymptotically stable.
- With $M\mathbf{x}_\infty = \mathbf{x}_\infty$, the limit distribution $\mathbf{x}_\infty$ is a fixed point of M and can be determined by the eigenvector $\mathbf{v}^{(1)}$ of M for the eigenvalue $\lambda_1 = 1$ with $\mathbf{x}_\infty = c\mathbf{v}^{(1)}$ where c is chosen so that $\mathbf{x}_\infty$ is a probability vector.
- The transient distribution $\mathbf{x}(n)$ can be found by solving the linear first order difference equation system.
- Except for $\lambda_1 = 1$, all other eigenvalues of M have less than unit modulus, i.e., $|\lambda_k| < 1$, $1 < k \leq m$. (In particular, there are no complex eigenvalues with a unit modulus.)

5. DNA Mutation

5.1. The Double Helix. The central dogma of biology is "DNA to RNA to proteins," with protein constituting the fundamental units for all parts and functions of living organisms DNA are therefore the basic building blocks for these organisms. DNA (Deoxyribonucleic acid) molecules encode genetic information and these molecules (with their genetic information) are copied and passed on from parent(s) to an offspring. Though highly accurate, the copying process is not immuned from errors that lead to genetic mutation and consequently the evolution of living organisms. To study biological evolution, we need to have an understanding how errors are incurred in the DNA copying process, especially copying errors that result in genetic mutation.

The 1962 Nobel Prize in Physiology or Medicine was awarded to James Watson, Francis Crick and Maurice Wilkins for their discovery of the double helix structure of DNA molecules. In 1953, Watson and Crick saw an x-ray pattern of a crystal of the DNA molecule made by Rosalind Franklin and Maurice Wilkin; it gave Watson and Crick enough information to make an accurate model of the DNA molecule. Their model showed a twisted double helix with little rungs connecting the two helical strands. At each end of a rung of the double helix ladder is one of

four possible molecular subunits (called nucleobases or simply bases): *adenine* (A), *guanine* (G), *cystosine* (C), and t*hymine* (T).

The shape of *adenine* is complementary to *thymine*; they are bound together consistently at opposite ends of a rung of the DNA ladder through a hydrogen bond to form a nucleotide. The nucleobase *guanine* is structurally similar to *adenine* and complementary to *cytosine*. The *guanine-cytosine* pair are also bound together consistently at the opposite ends of a DNA ladder rung. In other words, we always find either A paired with T or G paired with C (but neither A nor T with C or G). Thus, once we know the base at one side of a rung, we can deduce the base at the other end of the same rung. For example, if along one strand of the DNA ladder, we have a base

$$..... \; ATTAGAGCGCGT \; .....$$

then the corresponding sequence along the other strand opposite to the same stretch must be

$$..... \; TAATCTCGCGCA \; .....$$

(Because of their structural similarity, cytosine and thymine are called *pyrimidines* while adenine and guanine are called *purines*.) As such, a DNA molecule (or a segment of it) is specified by a sequence of the four letters A, T, C and G.

Once the model was established, its structure hinted that DNA was indeed the carrier of the genetic code and thus the key molecule of heredity, developmental biology and evolution. Heredity requires genetic information be passed on to offspring. This is accomplished during cell division when the twisted double-helix DNA molecule ladder unzips into two separate strands. One new molecule is formed from each half-ladder and, due to the required pairings, this gives rise to two identical daughter copies from each parent molecule. Though elaborate safeguards are in place to ensure fidelity in the replication, errors still occur though infrequently.

5.2. Mutation due to Base Substitutions. The most common type of errors during the replication process is a replacement of one base by another at a certain site of the strand sequence. For instance, if the sequence in the parent DNA molecule $ATTAGAGC$ should become $ATTACAGC$ in the offspring DNA, then there is a *base substitution* $G \to C$ at the fifth site of the sequence. The base subsitution replaces a pyrimidine by another pyrimidine (or one replaces a purine by another purine) is known as *transition*. A base substitution of a purine by a pyrimidine (or a pyrimidine by a purine) is called a *transversion*. Error types other than base substitution are also possible. These include deletion, insertion and inversion of a section of the sequence. Their occurrence are much less frequent than base substitutions and will be ignore in the discussion herein to focus only on mutations due to base substitutions.

For the restricted problem of mutations by base substitutions only (and for more general version of mutations), one issue of interest is the amount of mutation that has occurred after a number of generations of offspring from a DNA sequence, which may be quite long especially when we are talking about an evolutionary time scale. If data for all generations involved are not available or not used, then the difference between the original sequence of generation n and the sequence of generation $n+m$ does not necessarily give an accurate estimate of the amount of mutation that has occurred since one or more back mutation might have taken place. If the

sequence in the parent DNA molecule $ATTAGAGC$ should become $ATTACAGC$ four generations later, then one base substitution of $G \to C$ at the fifth site of the sequence is only one possible kind of mutation. It may also be a sequence of base substitutions such as $G \to C \to G \to G \to C$. or $.G \to C \to T \to G \to C$, with the former having one back mutation while the latter involves no back mutation at all. A more sophisticated approach is needed to determine the amount of mutation involved between an ancestral DNA site and the same site of its generation n descendent DNA molecule.

5.3. Markov Chain Model. . Given a particular nucleobase at a particular site of an initial (ancestral) DNA sequence, knowing the odds that it would mutate to another nucleobase after replication would allow us to estimate its evolution. For the nucleobase A for example, it would be helpful to know the probability (or the fraction of times) m_{AG} that it would be replace by the nucleobase G, the probability m_{AC} of being replaced by C and the probability m_{AT} of being replaced by T. Since deletion is not allowed, the fraction of times that A remains unchanged would have to be $1 - m_{AG} - m_{AC} - m_{AT} \equiv 1 - \Sigma_A$. Similar odds would be needed for the nucleobases G, C and T. We arrange all these quantities as a matrix M:

$$\begin{array}{c} \\ A_{n+1} \\ G_{n+1} \\ C_{n+1} \\ T_{n+1} \end{array} \begin{array}{c} \begin{array}{cccc} A_n & G_n & C_n & T_n \end{array} \\ \left[\begin{array}{cccc} 1-\Sigma_A & m_{AG} & m_{AC} & m_{AT} \\ m_{GA} & 1-\Sigma_G & m_{GC} & m_{GT} \\ m_{CA} & m_{CG} & 1-\Sigma_C & m_{CT} \\ m_{TA} & m_{TG} & m_{TC} & 1-\Sigma_T \end{array} \right] \end{array},$$

to get the following evolutionary relation for the different bases at a particular site of the DNA sequence:

$$\mathbf{x}_{n+1} = M\mathbf{x}_n$$

where

$$\mathbf{x}_n = (A, G, C, T)_n^T \ ,$$

with a component of $\mathbf{x}_n$ being the fraction of time (or the probability) that the particular nucleobase should occupy the same site by the n^{th} generation. The numerical values of the transition matrix entries $\{m_{ij}\}$ are key to the evolution of the organism; they are normally estimated from available replication data over many generations.

5.4. Equal Opportunity Substituion. To illustrate the use of the Markov chains to learn more about the genetic mutation, we consider the consequences of a highly speculative, one-parameter *equal opportunity* model for the transition matrix M . In this version of the Jukes-Cantor (1969) model, we stipulate that substitution of a nucleobase by any of the other three bases are equally likely with probability $\alpha/3$ where α is a number to be specified (or estimated from available data). It ranges from 10^{-8} mutations per site per year for mitochrondrial DNA to 0.01 mutations per site per year for the influenza A virus DNA. The corresponding transition matrix M take the form

$$M = \left[\begin{array}{cccc} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{array} \right]. \tag{5.1}$$

For such a transition matrix, an ancestral sequence with A at a particular site (so that $\mathbf{x}_0 = (1,0,0,0,0)^T$ would have any one of the four nucleobases at the same site after one replication with odds given by

$$\mathbf{x}_1 = M\mathbf{x}_0 = \begin{pmatrix} 1-\alpha \\ \alpha/3 \\ \alpha/3 \\ \alpha/3 \end{pmatrix}.$$

In other words, a substitution by anyone of the three nucleobase after one replication is equally likely at $\alpha/3$ fractions of the time while the site remains occupied by A at $1-\alpha$ fractions of the time.

The question of interest is what happens after many generations? Does $\mathbf{x}_n$ tend to some limiting $\mathbf{x}_\infty$? Since the transition matrix M is a positive probability matrix, the evolution of the genetic mutation process is a regular Markov chain. In that case, we know that $\mathbf{x}_n$ tends to a unique limiting probability vector $\mathbf{x}_\infty$.

EXERCISE 11. *Determine the limiting probability distribution $\mathbf{x}_\infty$ for the Equal Opportunity Substitution model*

6. Absorbing Markov Chains

Regular Markov chains constitute an important class of Markov chains processes in applications. There are however other Markov chains that are also prevalent in science and engineering. In this section, we examine a class of Markov chains that are characteristically different from *regular* Markov chains.

DEFINITION 16. *A state in a Markov chain is an **absorbing state** if it is impossible to leave it.*

DEFINITION 17. *A Markov chain is said to be an **absorbing MC** if (i) it has at least one absorbing state, and ii) from every non-absorbing (or **transient**) state it is possible to transition to an absorbing state (not necessarily in one step).*

THEOREM 19. *In an absorbing Markov chain, it is a certainty that the process will end up in one of the absorbing states.*

PROOF. (sketched) From a transient state S_j, let n_j be the minimum number of steps required to reach an absorbing state and $p_j < 1$ (since the state S_j is not absorbing) be the probability that, starting from S_j, the process does not reach an absorbing state in n_j steps. Let $n = \max[n_j]$ and $p = \max[p_j]$ where j ranges over all transient states. The probability of not reaching an absorbing state in n steps is less than p, in $2n$ steps is less than p^2, etc. In general, the probability of not reaching an absorbing state in $k \cdot n$ steps is less than p^k. Since $p < 1$, the probability of not reaching an absorbing state tends to zero as $k \to \infty$. □

For an absorbing Markov chain, there are at least three interesting problems:

(1) What is the probability of the process would end up in a particular absorbing state?
(2) On the average, how "long" would it take for the process to reach an absorbing state starting from a non-absorbing state (also known as a *transient state*)?

(3) On the average, how many times does the process pass through each non-absorbing state before ending in an absorbing state?

In the paragraphs below, we provide some insight to these questions.

6.1. A Simple Problem in Infectious Disease. To learn more about absorbing Markov chains, we consider here the simplest problem in infectious diseases. Available clinical data on the inhabitants of a small town show that a fraction p_I of the healthy individuals in month n becomes infected and the remaining fraction $1 - p_I$ remains healthy. Of those infected, a fraction p_R recovers by the following month; another fraction p_D will be dead while the remaining fraction $1 - p_R - p_D$ remaining sick. Of course, those who are dead at any stage will remain dead thereafter. (To simplify the discussion, we assume that all new borns are removed to a safe place in another town without their parents and not counted in the population distribution.) The phenomenon may be treated as a Markov chain of three states H (healthy), S (sick) and D (dead) with the transition matrix M,

$$\begin{array}{c} \\ H_{n+1} \\ S_{n+1} \\ D_{n+1} \end{array} \begin{array}{c} \begin{array}{ccc} H_n & S_n & D_n \end{array} \\ \left[\begin{array}{ccc} 1-p_I & p_R & 0 \\ p_I & 1-p_R-p_D & 0 \\ 0 & p_D & 1 \end{array}\right] \end{array} = M.$$

By setting $\mathbf{x}(n) = (x_1(n),\ x_2(n), x_3(n))^T = (H_n, S_n, D_n)^T$, the distribution of health, infectious and dead fractions (or the distribution vector of the probability of an individual being in the different states) is governed by the first order vector difference equation

$$\mathbf{x}(n+1) = M\mathbf{x}(n), \qquad n = 0, 1, 2, 3, ... \tag{6.1}$$

Evidently, M a probability matrix since $p_{ij} \geq 0$ and $\sum_{i=1}^{m} p_{ij} = 1$ where $m = 3$ in our particular example. The evolution of the distribution of the town inhabitants among the three possible states therefore qualifies for a MC. The chain has an absorbing state $(0, 0, 1)^T$ that can be reached from both transient states - after one or more stages from the infectious state and after two or more stages from the healthy state. As such, our model for the distribution of inhibitants is an absorbing MC.

The transition relation between states in consecutive periods is again a system of linear difference equations. The solution of the initial value problem (IVP) defined by the system (6.1) subject to the (vector) initial condition

$$\mathbf{x}(0) = \mathbf{p} = (p_1, p_2, p_3)^T,$$

with $p_k \geq 0$ $(k = 1, 2, 3)$ and $p_1 + p_2 + p_3 = 1$, provides complete information about the evolution of the distribution of inhabitants with time. However, some useful observations can be made even before solving the IVP.

(1) $\lambda = 1$ is again seen to be an eigenvalue of the transition matrix M above since the rows of the matrix $M - I$ sum up to a zero row.
(2) The single equilibrium state is $\mathbf{x}^{(1)} = (0, 0, 1)^T$ is a fixed points of (6.1).
(3) M is **not** a power positive matrix (see exercise).

With the third property, the Markov chain with transition matrix M does not behave like a *regular* Markov chain. Hence, we expect the behavior of an absorbing Markov chain to be different than what has been observed in the previous section.

6.2. Solution of IVP. To motivate some further developments that will uncover these differences, we consider first the specific example with $p_I = 1/4,\ p_R = 7/16$, and $p_D = 9/16$ so that

$$M = \begin{bmatrix} 3/4 & 7/16 & 0 \\ 1/4 & 0 & 0 \\ 0 & 9/16 & 1 \end{bmatrix}. \tag{6.2}$$

As before, we seek a solution of the difference equation (6.1) in the form $\mathbf{x}(n) = \mathbf{c}\lambda^n$ for some non-zero scalar constant λ and some non-zero vector constant $\mathbf{c}$. For this assumed solution, the vector equation (6.1) becomes

$$[M - \lambda I]\,\mathbf{c} = \mathbf{0}$$

For a nontrivial $\mathbf{c}$, we need $\det[M - \lambda I] = 0$, leading to the characteristic equation

$$(-)^3\,|M - \lambda I| = (\lambda - 1)(\lambda^2 - \frac{3}{4}\lambda - \frac{7}{64}) = 0.$$

Again, $\lambda = 1$ is root of this characteristic equation. The complete set of eigen-pairs are

$$\{1, (0,0,1)^T\},\quad \left\{\frac{7}{8}, (7,2,-9)^T\right\}, \left\{-\frac{1}{8}, (1,-2,1)^T\right\}$$

with the eigenvectors determined up to a multiplicative constant. The general solution of the system $\mathbf{x}(n+1) = M\mathbf{x}(n)$ with the transition matrix (6.2) may be taken to be

$$\mathbf{x}(n) = c_1 \begin{pmatrix} 7 \\ 2 \\ -9 \end{pmatrix} \left(\frac{7}{8}\right)^n + c_2 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \left(-\frac{1}{8}\right)^n + c_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} (1)^n \tag{6.3}$$

where the constants $\{c_k\}$ are determined by the initial probability distribution $\mathbf{x}(0) = \mathbf{p} = (p_1, p_2, p_3)^T$:

$$\mathbf{x}(0) = \begin{bmatrix} 7 & 1 & 0 \\ 2 & -2 & 0 \\ -9 & 1 & 1 \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

or

$$\mathbf{c} = \begin{bmatrix} 1/8 & 1/16 & 0 \\ 1/8 & -7/16 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}.$$

In the limit as $n \to \infty$, we get

$$\lim_{n\to\infty} \mathbf{x}(n) = \begin{pmatrix} 0 \\ 0 \\ p_1 + p_2 + p_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \mathbf{x}_\infty. \tag{6.4}$$

which may be written as

$$\mathbf{x}_\infty = M_\infty \mathbf{p}. \tag{6.5}$$

The solution (6.3) of the IVP and its limiting behavior (6.4) for large n provides the answer to the first question posed in the introductory paragraph of this section on absorbing Markov chains. It gives in a trivial way the probability for the process to end up in the only absorbing state. The process must end up in that absorbing state eventually (with probability 1) whatever the initial distribution of the inhabitants may be. However, the process leading to this result applies to the more general case of multiple absorbing states as to be shown in the next section.

As a model for the spread of an evolving infectious disease, the conclusion that all inhabitants eventually die from the disease is not consistent with reality. Throughout history, all recorded spread of deadly epidemics eventually dissipated and the human population survived, Some were arrested by geographical barrier; others probably by the built-up of the immune system among the fraction that recovered after infection. Whatever the factor or arresting mechanism may be, it should be incorporated into the model in order for it to be more realistic. In the next section we develop an improved model that allow those who recover from the disease to be immuned from future infection.

7. Immunity after Recovery

To improve on the three states infectious disease model, we allow those who recover from the disease to be immuned from contracting it again thereafter at least for the duration of the period of interest (but otherwise the same as the three states model). In that case, we have a Markov chain model of four states, namely *Healthy* (H_n), *Sick* (S_n), *Recovered* (R_n) and *Dead* (D_n), with a transition matrix M given by

$$M = \begin{array}{c} \begin{matrix} H_n & S_n & R_n & D_n \end{matrix} \\ \begin{bmatrix} 3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 0 & 7/16 & 1 & 0 \\ 0 & 9/16 & 0 & 1 \end{bmatrix} \end{array}. \tag{7.1}$$

To the extent that we really do not need (6.3) to tell us that the chain remains in an absorbing state once it is reached, we explore here the possibility of getting more useful information (such as the consequences of not starting from an absorbing state) without expending the effort to solve the IVP to obtain imore nformation than needed in practice. For this purpose, we partition M into four submatrices:

$$M = \begin{bmatrix} R & O \\ Q & I \end{bmatrix} \tag{7.2}$$

where I is the 2×2 identity matrix, O is the 2×2 zero matrix and

$$R = \begin{bmatrix} 3/4 & 0 \\ 1/4 & 0 \end{bmatrix}, \qquad Q = \begin{bmatrix} 0 & 7/16 \\ 0 & 9/16 \end{bmatrix}. \tag{7.3}$$

Correspondingly, we also partitionthe four distinct states of the problem into two groups, the *absorting states* $\mathbf{a}_n$ and the non-absorbing or *transient states* $\mathbf{t}_n$:

$$\mathbf{t}_n = \begin{pmatrix} H_n & S_n \end{pmatrix}^T, \qquad \mathbf{a}_n = \begin{pmatrix} R_n & D_n \end{pmatrix}^T, \tag{7.4}$$

with

$$\mathbf{y}_n = \begin{pmatrix} H_n & S_n & R_n & D_n \end{pmatrix}^T = \begin{pmatrix} \mathbf{t}_n \\ \mathbf{a}_n \end{pmatrix}. \tag{7.5}$$

In that case, the vector difference equation for the four states $\mathbf{y}_{n+1} = M\mathbf{y}_n$ can be written as

$$\begin{pmatrix} \mathbf{t}_{n+1} \\ \mathbf{a}_{n+1} \end{pmatrix} = \begin{bmatrix} R & O \\ Q & I \end{bmatrix} \begin{pmatrix} \mathbf{t}_n \\ \mathbf{a}_n \end{pmatrix} = \begin{pmatrix} R\mathbf{t}_n \\ Q\mathbf{t}_n + \mathbf{a}_n \end{pmatrix}$$

with

$$\begin{pmatrix} \mathbf{t}_1 \\ \mathbf{a}_1 \end{pmatrix} = \begin{bmatrix} R & O \\ Q & I \end{bmatrix} \begin{pmatrix} \mathbf{t}_0 \\ \mathbf{a}_0 \end{pmatrix} = \begin{pmatrix} R\mathbf{t}_0 \\ Q\mathbf{t}_0 + \mathbf{a}_0 \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{t}_2 \\ \mathbf{a}_2 \end{pmatrix} = \begin{bmatrix} R & O \\ Q & I \end{bmatrix} \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{a}_1 \end{pmatrix}$$
$$= \begin{bmatrix} R & O \\ Q & I \end{bmatrix}^2 \begin{pmatrix} \mathbf{t}_0 \\ \mathbf{a}_0 \end{pmatrix} = \begin{pmatrix} R^2\mathbf{t}_0 \\ (QR+Q)\mathbf{t}_0 + \mathbf{a}_0 \end{pmatrix}$$

and by induction

$$\begin{pmatrix} \mathbf{t}_n \\ \mathbf{a}_n \end{pmatrix} = \begin{bmatrix} R & O \\ Q & I \end{bmatrix}^n \begin{pmatrix} \mathbf{t}_0 \\ \mathbf{a}_0 \end{pmatrix} = \begin{pmatrix} R^n\mathbf{t}_0 \\ Q(R^{n-1} + R^{2-2} + \cdots + I)\mathbf{t}_0 + \mathbf{a}_0 \end{pmatrix}$$

or

$$\begin{pmatrix} \mathbf{t}_n \\ \mathbf{a}_n \end{pmatrix} = \begin{bmatrix} R^n & O \\ A_n & I \end{bmatrix} \begin{pmatrix} \mathbf{t}_0 \\ \mathbf{a}_0 \end{pmatrix} = \begin{pmatrix} R^n\mathbf{t}_0 \\ A_n\mathbf{t}_0 + \mathbf{a}_0 \end{pmatrix}$$

with

$$A_n = Q(I-R)^{-1}(I-R^n).$$

This result obtained with no calculations at all is considerably more informative (than the analogue of (6.3)). One immediate observation is the fate of the transient states after n period: For a given starting group distribution, the transient group $\mathbf{t}_n$ after n periods is reduced by a factor R^n.

The elements of the matrix R and Q are typically in the range $[0, 1)$:

$$R = [0 \le R_{ij} < 1], \qquad Q = [0 \le Q_{ij} < 1]$$

since they involve the evolution of the transient state (and therefore cannot be 1). For our example, we have

$$R^n = \begin{bmatrix} 3/4 & 0 \\ 1/4 & 0 \end{bmatrix}^n = \begin{bmatrix} (3/4)^n & 0 \\ (1/4)(3/4)^{n-1} & 0 \end{bmatrix}$$

with

$$\mathbf{t}_n = \begin{pmatrix} H_n \\ S_n \end{pmatrix} = \begin{pmatrix} (3/4)^n H_0 \\ (1/4)(3/4)^{n-1} H_0 \end{pmatrix} \to \mathbf{0} \quad (\text{as } n \to \infty).$$

Note that

$$\lim_{n\to\infty} R^n = \lim_{n\to\infty} \begin{bmatrix} 3/4 & 0 \\ 1/4 & 0 \end{bmatrix}^n = \lim_{n\to\infty} \begin{bmatrix} (3/4)^n & 0 \\ (1/4)(3/4)^{n-1} & 0 \end{bmatrix} = O$$

and

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} \left[Q(I-R^n)(I-R)^{-1}\right] = A$$

with

$$A = Q(I-R)^{-1}. \tag{7.6}$$

It follows that

$$\lim_{n\to\infty} \mathbf{a}_n = \lim_{n\to\infty} \begin{pmatrix} R_n \\ D_n \end{pmatrix} = A\mathbf{t}_0 + \mathbf{a}_0.$$

For our four states model, we have

$$(I-R)^{-1} = \begin{bmatrix} 4 & 0 \\ 3 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 7/16 & 7/16 \\ 9/16 & 9/16 \end{bmatrix}$$

so that

$$\lim_{n\to\infty} \mathbf{a}_n = \lim_{n\to\infty} \begin{pmatrix} R_n \\ D_n \end{pmatrix} = \begin{bmatrix} (7/16)(H_0+S_0) + R_0 \\ (9/16)(H_0+S_0) + D_0 \end{bmatrix}. \tag{7.7}$$

In other words, we have

$$R_n \to \frac{7}{16}(H_0 + S_0) + R_0, \qquad D_n \to \frac{9}{16}(H_0 + S_0) + D_0 \tag{7.8}$$

with

$$\begin{aligned} R_n + D_n &\to \left(\frac{7}{16} + \frac{9}{16}\right)(H_0 + S_0) + R_0 + D_0 \\ &= H_0 + S_0 + R_0 + D_0 = 1 \end{aligned}$$

Note that while it is certain that the population tends to an absorbing state eventually, the final destination depends on the initial distribution of the population among the four states (if there are more than one absorbing states as in the four states model). For a given initial distribution, a fraction of the population recovers (and becomes immune to the disease) while the remaining fraction dies. (Alternatively, the results may be interpreted as a probability $= (7/16)(H_0 + S_0) + R_0)$ for the population to recover eventually and a probability $= (9/16)(H_0 + S_0) + D_0)$ for the population to die out due to the infectious disease.) What these fractions are depends on both the resilience of the population against the disease as characterized by the transition matrix and the initial distribution of the population among the four states. That former is summarized by the matrix A, which is known as the *absorption matrix* for the population for the obvious reason.

In deducing the absorption matrix, it is important to have the possible states of the system ordered in such a way that absorbing state are all adjacent to each other before embarking on process of obtaining the absorption matrix. We emphasize that the absorption matrix contains all the essential information sought about the evolution of the infected population. It is a condensed version of the complete solution (6.4) with all the unessentials omitted. Moreover, we now obtain the needed information by performing simple algebraic operations on the smaller matrices R and Q and not having to solve any matrix eigenvalue problem for the original (larger) transition matrix.

8. Expected Transient Stops to an Absorbing State

There is more than improved computational efficiency and reduction of unessentials to the alternate form of the limiting distribution given in (7.7) and (7.8). The absorption matrix actually provides answers to the two remaining questions posed at the end of the paragraph after Theorem 19 (the first already answered by the limit distribution (6.4) through the solution of the IVP or (7.8) with the help of the absorption matrix (7.6)). We show below how the absorption matrix also provides the answer to the question: *Starting from one of its transient states, how many transient stops does the absorbing chain make on the average before reaching an absorbing state*? (Or how long does the evolving chain survives on the average before being trapped in an absorbing state?) The answer to the other question will also be obtained in the process.

Suppose the given absorbing Markov chain with an $m \times m$ transition matrix $M = [m_{ij}]$ has m_t transient states (and $1 - m_t$ absorbing states). Suppose the MC is in a transient state E_j initially. Let s_{kj} be the *number of stops on the average* (known as the expected number of stops) the absorbing chain makes at a particular transient state E_k before it reaches an absorbing state. If $k \neq j$, the chain can reach the state E_k on the first trial with probability m_{kj}. It may reach

E_k in the second trial with probability $\sum_{\alpha=1}^{m_t} m_{k\alpha}m_{\alpha j}$ by passing through any one intermediate transient state E_α, for $\alpha = 1, 2, \ ..., m_t$ on the first trial. It may also reach E_k in the third trial with probability $\sum_{\alpha=1}^{m_t}\sum_{\beta=1}^{m_t} m_{k\alpha}m_{\alpha\beta}m_{\beta j}$ and so on. If $k = j$, the chain is already in E_j with probability 1. Altogether, starting from in E_i, the expected number of stops the chain makes in E_j is therefore

(8.1)

$$s_{kj} = \delta_{kj}+m_{kj}+\sum_{\alpha=1}^{m_t} m_{k\alpha}m_{\alpha j}+\sum_{\alpha=1}^{m_t}\sum_{\beta=1}^{m_t} m_{k\alpha}m_{\alpha\beta}m_{\beta j}+\sum_{\alpha=1}^{m_t}\sum_{\beta=1}^{m_t}\sum_{\gamma=1}^{m_t} m_{k\alpha}m_{\alpha\beta}m_{\beta\gamma}m_{\gamma j}+ \cdots.$$

where δ_{kj} is Kronecker delta (= 1 if $k = j$ and 0 otherwise). (What we calculated was the sum of different expected numbers of stops to get from transient state E_j to transient state E_k but can be so re-interpreted.) It gives the number of times, on the average, the absorbing chain dwells in the particular non-absorbing state E_k when the chain starts from E_j.

As we allow k and j to range over all the transient states, the relation (8.1) may be written in terms of the two $m_t \times m_t$ matrices $S = [s_{ij}]$ and $R = [m_{k\ell}]$ as

$$S = I + R + R^2 + R^3 \div \ \cdots\cdots$$

To simplify the expression for S, we form $RS = R + R^2 + R^3 + \cdots\cdot = S - I$ to get $I = (I - R)S$ or

$$S = (I - R)^{-1}. \tag{8.2}$$

We have then the following result:

PROPOSITION 10. *The expected number of stops (for all initial transient states) the chain makes in E_k before reaching an absorbing state is sum of the elements of the k^{th} row of the matrix $S = (I - R)^{-1}$.*

PROOF. The expected number of stops the MC makes at E_k is the sum of s_{ij} over all transient states:

$$S_k = \sum_{j=1}^{m_t} s_{kj}.$$

□

The (expected) *transient stop matrix* S also provides the answer to our original question. Let

$$\bar{S}_j = \sum_{k=1}^{m_t} s_{kj} = s_{1j} + s_{2j} + s_{3j} + \cdots\cdot + \ s_{m_t j}.$$

Evidently, starting at the transient state E_j, the sum S_i is the expected number of transient stops incurred by the absorbing chain prior to reaching an absorbing state, leading to the following proposition:

PROPOSITION 11. *The expected number of transient stops made by an absorbing MC starting from E_j is the sum of the elements of the j^{th} column of the matrix $S = (I - R)^{-1}$.*

9. Appendix - Proof of Theorem 18

Below is a proof of Theorem 18: For a *regular* Markov chain with an $m \times m$ transition matrix $M = [m_{ij}]$ and any initial probability distribution $\mathbf{x}(0) = \mathbf{p}$, there exists a probability vector $\mathbf{x}_\infty$ to which the sequence of probability vectors $\{\mathbf{x}(n) = M^n\mathbf{p}\}$ converges as $n \to \infty$. With no loss in generality, we give a proof for the case $M > 0$.

Form $\mathbf{q}^T M^n \mathbf{p} = \mathbf{p}^T (M^T)^n \mathbf{q} = \mathbf{p}^T \mathbf{c}$ for an arbitrary (probability) vector $\mathbf{q} \neq \mathbf{0}$ and set $\mathbf{w}(n) = (M^T)^n \mathbf{q}$. It suffices to show that $\mathbf{w}(n) = (M^T)^n \mathbf{q}$ converges as $n \to \infty$ for any vector $\mathbf{q}$.

Now $\mathbf{w}(n)$ satisfies the difference equation

$$\mathbf{w}(n+1) = M^T \mathbf{w}(n), \qquad \mathbf{w}(0) = \mathbf{q}\ .$$

For each n, let $u(n)$ be the largest component of $\mathbf{w}(n)$ and $v(u)$ the smallest. Since

$$w_i(n+1) = \sum_{j=1}^{m} m_{ji} w_j(n)$$

and since $m_{ij} \geq 0$ and $\sum_{j=1}^{m} m_{ji} = 1$ for $i = 1, 2, ..., m$, it follows that

$$u(n+1) \leq u(n), \qquad v(n+1) \geq v(n).$$

Therefore $\{u(n)\}$ is a monotone decreasing sequence bounded from below by zero; and $\{v(n)\}$ is a monotone increasing sequence bounded from above by 1, respectively. Hence, both converge to a limit, denoted by u_∞ and v_∞, respectively. Our goal is to show $u_\infty = v_\infty$ as we will do below.

With no loss in generality, let $u(n)$ be the first component of $\mathbf{w}(n)$, i.e., $u(n) = w_1(n)$ (and hence $i \neq 1$). In that case, we have

$$\begin{aligned} v(n+1) &= w_i(n+1) = \sum_{j=1}^{m} m_{ji} w_j(n) = \sum_{j=2}^{m} m_{ji} w_j(n) + (m_{1i} - d) w_1(n) + d u(n) \\ &\geq \sum_{j=2}^{m} m_{ji} v(n) + (m_{1i} - d) v(n) + d u(n) = (1-d) v(n) + d u(n). \end{aligned}$$

where $d = \min[m_{pq}] \leq 1/2$ for $m \geq 2$ (and $d > 0$ since M is a positive matrix). Similarly, we have also

$$u(n+1) \leq (1-d) u(n) + d v(n).$$

Combining these two inequalities gives

$$u(n+1) - v(n+1) \leq (1-2d)\,[u(n) - v(n)] \leq (1-2d)^n\,[u(0) - v(0)]\,.$$

Keeping in mind that $0 < d \leq 1/2$ so that $0 \leq 1 - 2d < 1$, the difference $[u(n) - v(n)] \to 0$ as $n \to \infty$ so that $u_\infty = v_\infty$ in the limit.

Thus, not only the two sequences $\{u(n)\}$and $\{v(n)\}$ converge, they both converge to the same limit resulting in

$$\lim_{n \to \infty} \mathbf{w}(n) = (M^T)^n \mathbf{q} = w_\infty (1, 1, ..., 1)^T \tag{9.1}$$

where we have denote by w_∞ the two equal limits u_∞ and v_∞. Since $\mathbf{q}$ is finite and arbitrary, the power matrix $(M^T)^n = (M^n)^T$ converges and $M^n \to$ a well-defined $\bar{M}$. (The k^{th} column of the limiting matrix $\bar{M}$ corresponds to the limiting vector (9.1) for $\mathbf{q} = (\boldsymbol{\delta}_{1k}, \boldsymbol{\delta}_{2k}, ..., \boldsymbol{\delta}_{mk})^T$).

10. Exercises

EXERCISE 12. *Consider the following coin-tossing process involving (the same) two coins, a dime and a quarter. The rules of the game require that the dime is tossed next if a head turns up and the quarter next for a tail. Due to the different engraved patterns on the two sides, the coins are* ***not*** *fair coins. Since the patterns on the coin faces are different, the probability of getting a head for the dime is p_d (obtained as the limit of repeated tossing of the same coin with each outcome independent of the outcomes of past tosses). Similarly, the probability of getting a head for the quarter is p_q. Correspondingly, the probability of getting a tail is $1 - p_q$ for the dime and $1 - p_q$ for the quarter, respectively). We are interested here in whether the dime or quarter would be tossed next after Nth time units. Mathematically, we would like to know the probabilities of getting to toss each of the two coins at the n^{th} toss.*

EXERCISE 13. *Show that the eigenvalues of M^T is the same as that of a transition matrix M and the eigenvector of the M^T is the transpose of the eigenvector of M.*

EXERCISE 14. *If $\mathbf{x}(0) = \mathbf{p}$ is a probability vector, then so is $\mathbf{x}(n) = M^n\mathbf{p}$. Moreover, If M is the transition matrix of a regular MC, then $\mathbf{x}(n)$ is a positive probability vector for sufficiently large n.*

EXERCISE 15. *If $M > O$, show $\mathbf{y} = M\mathbf{x} > 0$ for any probability vector $\mathbf{x}$. (A matrix $M > O$ means that all elements of M are positive. A vector $\mathbf{p} > \mathbf{0}$ means all components of the vector are positive.)*

EXERCISE 16. *If M is the transition matrix of a regular MC and $\mathbf{x}(0) = \mathbf{p}$ is a probability vector, then $\mathbf{x}(n)$ is a positive probability vector for sufficiently large n.*

EXERCISE 17. *Product of two probability matrices is a probability matrix. In particular, any power of a probability matrix is a probability matrix.*

EXERCISE 18. *Prove Theorem* **??** *(corresponding to Superposition Principle I for* complementary solutions *of linear difference equations).*

EXERCISE 19. *(Social Mobility) From data compiled by government census, it is known that a fraction of the offsprings of families in a particular income group becomes significantly more wealthy and another fraction becomes significantly less well off with the rest not doing any better or worse. Divide up families into high (A_1), middle (A_2) and low (A_3) low income groups to get the transition matrix M below, providing a highly simplified summary of the census data. Assuming M does not change with n (which may not be unrealistic for a few generations), a) determine the eigen-pairs of this transition matrix and use it to solve the initial value problem with $\mathbf{x}(0) = \mathbf{p} = (p_1, p_2, p_3)^T$, and b) shows that $\mathbf{x}(n) \to \mathbf{x}_\infty = (0.2, 0.55, 0.25)^T$ as $n \to \infty$, independent of the initial distribution $\mathbf{p}$.*

$$
\begin{array}{c} \text{generation} \\ (n+1)^{th} \backslash n^{th} \end{array}
\begin{array}{c} \\ \text{high} \\ \text{middle} \\ \text{low} \end{array}
\begin{array}{c}
\begin{array}{ccc} \text{high} & \text{middle} & \text{low} \end{array} \\
\left[\begin{array}{ccc} 0.6 & 0.1 & 0.1 \\ 0.3 & 0.8 & 0.2 \\ 0.1 & 0.1 & 0.7 \end{array}\right]
\end{array} = M \tag{10.1}
$$

EXERCISE 20. *Formulate and analyze an improved model for the infectious disease problem of Section 6 that allows those who recover from the disease to be immuned from future infection.*

CHAPTER 9

Nonlinear Systems

1. Rabbits and Coyotes (Predator-Prey)

1.1. The model. Left alone, a rabbit population grows naturally according to its natural growth rate (and not by the rather artificial rules of Fibonacci). When the population is relatvely small, a linear growth model applies leading to geometrical growth discussed briefly in Chapter 1 of these notes. Such growth would be modified when the population is sufficiently large to be affected by resource and space constraints. It may also be modified by the presence of natural predators such as coyotes, foxes and other animals that feed on rabbits without which their population would diminish for a lack of food. Let R_n be a rabbit population at stage n in some biomass unit, and F_n be the fox population at the same stage. The growth or decline of either population clearly depends on the current size of both populations indicated by the mathematical relations

$$R_{n+1} = P(R_n, F_n), \qquad F_{n+1} = Q(R_n, F_n) \tag{1.1}$$

where $P(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ are two prescribed smooth functions, at least twice differentiable in their two arguments R_n and F_n. To the extent that $P(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ do not vary with n, we are restricting to the autonomous case when the growth rates do not change with time (at least not noticeably for the period considered).

To illustrate one of the basic methods to learn about predator-prey type interaction of two populations, we consider the following simple difference equation model of the growth of a rabbit population in the presence of a predator (say fox) population:

$$R_{n+1} = (1+a)R_n - bR_nF_n, \qquad F_{n+1} = (1-c)F_n + dR_nF_n \tag{1.2}$$

with an appropriate choice of stage unit (e.g., a reproductive season). The parameters $(1+a)$, b, $(1-c)$, and d are the relevant gain and loss rate constants. For biologically realistic cases, we take

- a, b, c and d to be positive
- $1-c > 0$ since it is unrealistic to lose the entire fox population is one stage
- $a < 1$ to grow at a rate less than doubling the rabbit population perunit stage (certainly $a < 2$ for less than tripling per stage even in extreme cases)

With $b = d = 0$, the two populations do not interact and the rabbit population would grow geometrically (until resource constraints become important and the linear model ceases to be appropriate) while the fox population would die out for a lack of food.

Starting with some initial populations of rabbits and foxes,

$$R_0 = \bar{R}, \qquad F_0 = \bar{F}, \tag{1.3}$$

we would like to know how the two populations evolve with time acording to the growth dynamics of the Lotke-Volterra model above (with $bd \neq 0$), first formulated in $1925 - 1926$ for predator-prey relations. The information sought is of course provided by the solution of the IVP defined by (1.2) and (1.3). But as we learn in previous chapters of these notes, a simple expression is rarely available for the exact solution of the IVP even for a single nonlinear difference equation, not to mention two simultaneous nonlinear difference.equations. As in the case of a single equation, a great deal of useful information can be obtained from the fixed points of the equations and their stability.

1.2. Fixed Points. To obtain the fixed points of the system (1.2), we consider the possibility of $R_n = R$ and $F_n = F$. In that case, the two difference equations (1.2) become

$$(a - bF)\, R = 0, \qquad (-c + dR)\, F = 0,$$

giving two possible fixed points:

$$\left(R^{(1)}, F^{(1)}\right) = (0, 0), \qquad \left(R^{(2)}, F^{(2)}\right) = (c/d, a/b).$$

Thus, if the initial rabbit and fox population should happen to coincide with one of these two fixed points, the two populatons would remain unchanged for all time. For example, if $\bar{R} = \bar{F} = 0$, then $R_n = F_n = 0$ for all n; if there should be neither rabbit nor fox initially, there would be no rabbits or foxes there after. Unlike linear difference equations, simultaneous nonlinear difference equations often have more than one fixed points. Of interest is the stability of these fixed points.

1.3. The Fixed Point $(0, 0)$ is Unstable. Suppose the initial population of rabbits and foxes are very small, very close to the $(0, 0)$, with

$$R_0 = \varepsilon \bar{r}, \qquad F_0 = \varepsilon \bar{f},$$

where $0 < \varepsilon \ll 1$. In that case, we expect the solution of the IVP would be small at least for a few seasons so that we may write

$$R_n = \varepsilon r_n^{(1)}, \qquad F_n = \varepsilon f_n^{(1)} \tag{1.4}$$

with

$$r_0^{(1)} = \bar{r}, \qquad f_0^{(1)} = \bar{f}.$$

Upon substituting (1.4) into (1.2), we obtain

$$r_{n+1}^{(1)} = (1 + a) r_n^{(1)} - \varepsilon b r_n^{(1)} f_n^{(1)}, \qquad f_{n+1}^{(1)} = (1 - c) f_n^{(1)} + \varepsilon d r_n^{(1)} f_n^{(1)}$$

where we have divided both sides of both equations by $\varepsilon (\neq 0)$. Since $0 < \varepsilon \ll 1$, the last term in each equation is negligibly small compared to the other terms in the same equation, at least for a while. For that duration, we may approximate these governing equations for r_n and f_n by dropping terms involving ε to get

$$r_{n+1}^{(1)} \simeq (1 + a) r_n^{(1)}, \qquad f_{n+1}^{(1)} \simeq (1 - c) f_n^{(1)} \tag{1.5}$$

as adequate approximations of the original nonlinear DE (at least for a few stages). The two DE in (1.5) are linear with explicit solutions

$$r_n^{(1)} = \bar{r}(1 + a)^n, \qquad f_n^{(1)} = \bar{f}(1 - c)^n.$$

With $1 + a > 1$, the fixed point is generally unstable. It is a *saddle point* since $0 < 1 - c < 1$ with the fixed point being asymptotically stable for the special initial condition of $\bar{r} = 0$).

With the initial populations of rabbits and foxes very small, the chances of a fox encountering a rabbits is small. With considerable difficulty in finding food, the fox population dwindles. With few predators around, the rabbits population would geometrically for a while until it becomes sufficiently large to be found by the foxes still around. For our model, the extinction of the fox population is independent of the various growth rate. While this appears unrealistic, keep in mind that the linearization of the original equation is no longer appropriate at that point and we need to return to the nonlinear IVP for a more accurate picture of the growth of the two interacting populations.

1.4. The Stability of $\left(R^{(2)}, F^{(2)}\right) = (c/d, a/b)$. Suppose the initial population of rabbits and foxes are close to, but not exactly the same as, the combination $\left(R^{(2)}, F^{(2)}\right) = (c/d, a/b)$ with

$$R_0 = \frac{c}{d} + \varepsilon \bar{r}, \qquad F_0 = \frac{a}{b} + \varepsilon \bar{f},$$

where $0 < \varepsilon \ll 1$. In that case, we expect the difference between the evolving predator and prey populations to remain close to the fixed point $(c/d, a/b)$ so that we can write the solution of the IVP as

$$R_n = \frac{c}{d} + \varepsilon r_n^{(2)}, \qquad F_n = \frac{a}{b} + \varepsilon f_n^{(2)} \tag{1.6}$$

with

$$r_0^{(2)} = \bar{r}, \qquad f_0^{(2)} = \bar{f}.$$

Upon substituting (1.4) into (1.2) and making use of the fixed point relations

$$a - bF^{(2)} = 0, \qquad -c + dR^{(2)} = 0,$$

we obtain

$$\begin{aligned} r_{n+1}^{(2)} &= (1+a) r_n^{(2)} - b(F^{(2)} r_n^{(2)} + R^{(2)} f_n^{(2)} + \varepsilon r_n^{(2)} f_n^{(2)}), \\ f_{n+1}^{(2)} &= d(F^{(2)} r_n^{(2)} + R^{(2)} f_n^{(2)} + \varepsilon r_n^{(2)} f_n^{(2)}) + (1-c) f_n^{(2)} \end{aligned}$$

where we have divided both sides of both equations by the common factor ε (> 0). Since $0 < \varepsilon \ll 1$, the term multiplied by ε in each equation is negligibly small compared to the other terms in the same equation, at least for a while. For that duration, we may approximate these governing equations for $r_n^{(2)}$ and $f_n^{(2)}$ by dropping terms involving ε to get

$$\begin{pmatrix} r_{n+1}^{(2)} \\ f_{n+1}^{(2)} \end{pmatrix} = J \begin{pmatrix} r_n^{(2)} \\ f_n^{(2)} \end{pmatrix} \tag{1.7}$$

where

$$J = \begin{bmatrix} 1 + a - bF^{(2)} & -bR^{(2)} \\ dF^{(2)} & 1 - c + dR^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & -bc/d \\ da/b & 1 \end{bmatrix} \tag{1.8}$$

as adequate approximations of the original nonlinear DE (at least for a few stages). The two DE in (1.7) are linear with explicit solutions

$$\begin{pmatrix} r_n^{(2)} \\ f_n^{(2)} \end{pmatrix} = c_1 \mathbf{v}^{(1)} \lambda_1^n + c_2 \mathbf{v}^{(2)} \lambda_2^n \ ,$$

where $\left\{\lambda_1, \mathbf{v}^{(1)}\right\}$ and $\left\{\lambda_2, \mathbf{v}^{(2)}\right\}$ are the two eigen-pairs of the coefficient matrix J. The two complex conjugate eigenvalues of $A^{(2)}$ are

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = 1 \pm i\sqrt{ac} = \rho e^{i\theta} = \rho\left(\cos\theta + i\sin\theta\right)$$

with

$$\rho^2 = 1 + ac, \qquad \theta = \tan^{-1}\left(\sqrt{ac}\right).$$

In terms of ρ and θ, we may rewrite the solution for the two homogeneous linear difference equations as

$$\begin{aligned} R_n &= \frac{c}{d} + \varepsilon r_n^{(2)} = \frac{c}{d} + \varepsilon\rho^n\left\{\bar{r}\cos(n\theta) + u\sin(n\theta)\right\} \\ F_n &= \frac{a}{b} + \varepsilon f_n^{(2)} = \frac{a}{b} + \varepsilon\rho^n\left\{\bar{f}\cos(n\theta) + v\sin(n\theta)\right\} \end{aligned}$$

where u and v are known constants found in the solution process (but their explicit expressions will not be needed here).

With $\rho = \sqrt{1+ac} > 1$, both $|R_n|$ and $|F_n|$ tend to ∞ as $n \to \infty$. Hence the fixed point $\left(R^{(2)}, F^{(2)}\right) = (c/d, a/b)$ is also unstable. With $\{\cos(n\theta), \sin(n\theta)\}$ changing sign cyclically as n increases, the polygonal path traced out by the point (R_n, F_n) changing with n spirals outward around the fixed point $\left(R^{(2)}, F^{(2)}\right)$. For n sufficiently large, the polygonal path becomes closer and closer to the positive F and R axes, requiring a separatrix, a boundary between solutions with two different qualitative behavior, to separate the polygonal hyperbola in the neighborhood of $\left(R^{(1)}, F^{(1)}\right) = (0,0)$ which tends to $(\infty, 0)$ as $n \to \infty$ from the polygonal paths that spiral around $\left(R^{(2)}, F^{(2)}\right) = (c/d, a/b)$ with F_n increasing at the lower far end of the first quadrant of the R, F - plane.

We refrain from seeking such a separatrix since the linearized model ceases to be adequate this point. We need to return to the original nonlinear model of a modified form of it to incorporate the effects of resource and space constraints. Nevertheless, the stability study above for the present Lotka-Volterra predator-prey model sufficed for Volterra to explain the observed fluctuations of the shark population and other commercially valuable fish populations (which sharks prey on) in the Adriatic Sea during and after World War I.

2. A Resource Limiting Predator-Prey Model

2.1. Limiting Capacity for Prey Population . We know from our earlier studies of the growth of a single population that such growths are usually constrained by the limited availability of space and other resources. Thus even at a low level of predator population, the prey population does not grow without bound. For that reason, a more realistic model would have the geometric growth rate of the prey be replaced by a logistic type growth rate. In this section, we investigate the consequences of the following modified Lotka-Volterra model:

$$\begin{aligned} R_{n+1} &= (1+a)R_n - \beta R_n^2 - bR_nF_n \equiv g(R_n, F_n). \\ F_{n+1} &= (1-c)F_n + dR_nF_n \equiv h(R_n, F_n) \end{aligned} \tag{2.1}$$

with $R_c = a/\beta$ being the carry capacity of the environment for the rabbit population and all other parameters in the equations are as previously defined gain and loss rate constants with a, b, c and d being positive, $1-c$ also positive and $a < 1$ generally (and may allow $a < 2$ in extreme cases). The two simultaneous nonlinear difference

equations (2.1) are supplemented by the two initial conditions (1.3) to form an IVP for the determination of the predator and prey populations at any subsequenct stage.

We note in passing that we could in principle also modify the second equation of the Lotka-Volterra model to allow for a limiting size for the predator population. However, we refrain from such a modification as there is already a natural mechanism for limiting the predator growth. A large population of predators invariably reduces the size of the prey population. When R_n is sufficiently low so that the unit growth rate dR_n becomes less than the unit loss rate c, the predator population would decline without a limit capacity imposed on the predator population.

Again, insight to the evulation of the two populations may be obtained by investigating the fixed points of the model system and their stability. For a fixed point, we have $R_n = R$ and $F_n = F$ so that two equations in (2.1) become

$$[a - \beta R - bF]\, R = 0, \qquad (-c + dR)\, F = 0. \tag{2.2}$$

The two steady state conditions (2.2) now give rise to three possible fixed points:

$$\begin{aligned} \left(R^{(1)}, F^{(1)}\right) &= (0,0), \qquad \left(R^{(2)}, F^{(2)}\right) = (a/\beta, 0), \\ \left(R^{(3)}, F^{(3)}\right) &= \left(\frac{c}{d}, \frac{a}{b}\left(1 - \frac{c/d}{a/\beta}\right)\right). \end{aligned} \tag{2.3}$$

Thus, if the initial rabbit and fox population should happen to coincide with anyone of these three fixed points, the two populatons would remain unchanged for all stages. For example, if $\bar{R} = a/\beta$ and $\bar{F} = 0$, then $R_n = a/\beta$ and $F_n = 0$ for all n. If there should be no fox population initially, there would be no fox offsprings and the rabbits are free to grow to its carrying capacity.

The two nonlinear difference equation has three fixed points. In general, there is no correlation between the number or order of the difference equations and the number of fixed points. Some nonlinear systems may even have an infinite number of fixed and others have none.

2.2. Linearization Near a Fixed Point. Next, we want to learn about the stability of each fixed point. Suppose the initial population of rabbits and foxes are close to, but not exactly the same as, one of the fixed points, say $\left(R^{(k)}, F^{(k)}\right)$ with

$$R_0 = R^{(k)} + \varepsilon \bar{r}, \qquad F_0 = F^{(k)} + \varepsilon \bar{f},$$

where $0 < \varepsilon \ll 1$. In that case, we expect the difference between the evolving predator and prey populations to remain close to the fixed point $\left(R^{(k)}, F^{(k)}\right)$ so that we can write the solution of the IVP as

$$R_n = R^{(k)} + \varepsilon r_n^{(k)}, \qquad F_n = F^{(k)} + \varepsilon f_n^{(k)} \tag{2.4}$$

(at least for a few stages) with

$$r_0^{(k)} = \bar{r}, \qquad f_0^{(k)} = \bar{f}.$$

Upon substituting (2.4) into (2.1) to get

$$R_{n+1} = g(R^{(k)} + \varepsilon r_n^{(k)}, F^{(k)} + \varepsilon f_n^{(k)}), \qquad F_{n+1} = h(R^{(k)} + \varepsilon r_n^{(k)}, F^{(k)} + \varepsilon f_n^{(k)})$$

Following the approach used for the Lotka-Volterra model, we wish to linearize the nonlinear difference equations. We do this for a general pair of growth functions

$g(x,y)$ and $h(x,y)$ by a Taylor series expansion in the parameter ε about $\varepsilon = 0$ for each to get

$$R^{(k)} + r_{n+1}^{(k)} = g(R^{(k)}, F^{(k)}) + \varepsilon r_n^{(k)} g_{,x}(R^{(k)}, F^{(k)}) + \varepsilon f_n^{(k)} g_{,y}(R^{(k)}, F^{(k)}) + O(\varepsilon^2) \tag{2.5}$$

$$F^{(k)} + f_{n+1}^{(k)} = h(R^{(k)}, F^{(k)}) + \varepsilon r_n^{(k)} h_{,x}(R^{(k)}, F^{(k)}) + \varepsilon f_n^{(k)} h_{,y}(R^{(k)}, F^{(k)}) + O(\varepsilon^2) \tag{2.6}$$

where $g_{,x}(R,F)$ is the partial derivative of $g(x,y)$ with respect to x evaluated at the point (R,F), etc. The term $O(\varepsilon^2)$ in each equation is an abbreviation for the sum of all the remaining terms in the Taylor expansion each proportional to ε^m for $m \geq 2$. Given the fixed point relations (2.2) or more generally

$$R^{(k)} = g(R^{(k)}, F^{(k)}), \qquad F^{(k)} = h(R^{(k)}, F^{(k)}), \tag{2.7}$$

the two equations(2.5) and (2.6) simplify to

$$\begin{aligned} r_{n+1}^{(k)} &= g_{,x}(R^{(k)}, F^{(k)}) r_n^{(k)} + g_{,y}(R^{(k)}, F^{(k)}) f_n^{(k)} + O(\varepsilon) \\ f_{n+1}^{(k)} &= h_{,x}(R^{(k)}, F^{(k)}) r_n^{(k)} + h_{,y}(R^{(k)}, F^{(k)}) f_n^{(k)} + O(\varepsilon) \end{aligned}$$

where we have divided both sides of both equations by the common factor ε (> 0). Since $0 < \varepsilon \ll 1$, the $O(\varepsilon)$ term in each equation is negligibly small compared to the other terms in the same equation, at least for a while. For that duration, we may approximate these governing nonlinear difference equations for $r_n^{(k)}$ and $f_n^{(k)}$ by dropping terms involving ε to get following linear difference equation system for the vector state variable $\mathbf{x}_n = (r_n^{(k)}, f_n^{(k)})^T$:

$$\mathbf{x}_{n+1} = J(R^{(k)}, F^{(k)}) \mathbf{x}_n \tag{2.8}$$

where

$$J(R^{(k)}, F^{(k)}) = \begin{bmatrix} g_{,x}(R^{(k)}, F^{(k)}) & g_{,y}(R^{(k)}, F^{(k)}) \\ h_{,x}(R^{(k)}, F^{(k)}) & h_{,y}(R^{(k)}, F^{(k)}) \end{bmatrix} \tag{2.9}$$

is known as the *Jacobian matrix* for the two functions $g(x,y)$ and $h(x,y)$ evaluated at $(x,y) = (R^{(k)}, F^{(k)})$ (with $g_{,x} = \partial g / \partial x$, etc.). The Jacobian matrix $J(R^{(2)}, F^{(2)})$ for the nontrivial fixed point of the Lotka-Volterra model of the previous section is given by (1.8). We now use (2.9) to study the stability of the three fixed points of the modified Lotka-Volterra model (2.1) of this section.

2.3. Stability of $(R^{(1)}, F^{(1)}) = (0,0)$. For our modified Lotka-Volterra model (2.1), the Jacobian matrix at $(R^{(1)}, F^{(1)}) = (0,0)$ is

$$J(0,0) = \begin{bmatrix} (1+a) - 2\beta x - by & -bx \\ dy & 1 - c + dx \end{bmatrix}_{(x,y)=(0,0)} = \begin{bmatrix} 1+a & 0 \\ 0 & 1-c \end{bmatrix}$$

which is the same as that for the Lotka-Volterra model. Therefore, we conclude again

PROPOSITION 12. *Given $a > 0$, the fixed point $(R^{(1)}, F^{(1)}) = (0,0)$ is unstable since $\lambda_1^{(1)} = 1 + a > 1$.*

The fixed point at the origin of the (R,F) - plane is a *saddle point* since $0 < 1 - c < 1$ would be asymptotically stable for the special initial condition $\bar{r} = 0$.

2.4. Stability of $(R^{(2)}, F^{(2)}) = (a/\beta, 0)$. For the fixed point $(R^{(2)}, F^{(2)}) = (a/\beta, 0)$, the Jacobian for our modified Lotka-Volterra model (2.1) is

$$\begin{aligned} J(a/\beta, 0) &= \left[\begin{array}{cc} (1+a) - 2\beta x - by & -bx \\ dy & 1 - c + dx \end{array} \right]_{(x,y)=(a/\beta,0)} \\ &= \left[\begin{array}{cc} 1-a & -ba/\beta \\ 0 & 1 - c + ad/\beta \end{array} \right] \end{aligned}$$

The general solution of the linear system (2.8) for this fixed point is

$$\begin{aligned} \mathbf{x}_n^{(2)} &= \left(\begin{array}{c} r_n^{(2)} \\ f_n^{(2)} \end{array} \right) = \bar{r} \left(\begin{array}{c} 1 - \frac{\bar{f}}{\bar{r}} \frac{abc}{\beta} \left(1 - \varrho - \frac{a}{c} \right) \\ 0 \end{array} \right) \lambda_1^n \\ &\quad + \bar{f} \left(\begin{array}{c} \frac{abc}{\beta} \left(1 - \varrho - \frac{a}{c} \right) \\ 1 \end{array} \right) \lambda_2^n \end{aligned}$$

where

$$\varrho = \frac{a/\beta}{c/d}$$

is the *reproductive ratio* of the rabbit and fox. The two eigenvalues of $J(a/\beta, 0)$ are

$$\lambda_1 = 1 - a, \qquad \lambda_2 = 1 + c\,(\varrho - 1)\,.$$

The stability of the fixed point $(a/\beta, 0)$ depends on the magnitude of $\lambda_1$and λ_2.

PROPOSITION 13. *The fixed point $(R^{(2)}, F^{(2)}) = (a/\beta, 0)$ is unstable if $\varrho > 1$, i.e., $c/d < a/\beta$.*

When gain-loss ratio of the rabbits in large compared to that of the foxes, there are plenty of rabbits to feed the few foxes around. It is not surprising that the fox population will increase and move away from the fixed point $(R^{(2)}, F^{(2)})$ which would have no foxes.

In general, the situation is more complicated if $\varrho < 1$. However, by the restriction we have already imposed on the magnitude of a and c, the following conclusion is also immediate:

PROPOSITION 14. *If $\varrho < 1$, the fixed point $(R^{(2)}, F^{(2)}) = (a/\beta, 0)$ is asymptotically stable.*

PROOF. We have the restriction $0 < a < 2$ so that $|\lambda_1| = |1 - a| < 1$. Since $0 < c < 1$, we have $0 < c\,(1 - \varrho) < 1$ for $0 < 1 - \varrho < 1$ so that $0 < |\lambda_2| = |1 - c\,(1 - \varrho)| < 1$. □

With a small initial fox population and its gain-loss rate is low, the fox population will head for extinction as the reproductive rate of the predator due the presence of a unit prey does not make up for the natural attrition rate of the predator. This leaves the rabbits to grow in an environment of limited resources tending toward the carrying capacity a/β of its logistic growth.

2.5. Stability of $(R^{(3)}, F^{(3)})$. For the fixed point $(R^{(3)}, F^{(3)})$, we have for our modified Lotka-Volterra model (2.1)

$$\begin{aligned} J(R^{(3)}, F^{(3)}) &= \left[\begin{array}{cc} (1+a) - 2\beta x - by & -bx \\ dy & 1 - c + dx \end{array} \right]_{(x,y)=(R^{(3)},F^{(3)})} \\ &= \left[\begin{array}{cc} 1 - \frac{\beta c}{d} & -\frac{bc}{d} \\ \frac{\beta c}{b}\left(\frac{a/\beta}{c/d} - 1\right) & 1 \end{array} \right] \end{aligned}$$

The general solution for the linear system (2.8) is

$$\mathbf{x}_n = \left(\begin{array}{c} r_n^{(3)} \\ f_n^{(3)} \end{array} \right) = c_1 \mathbf{u}^{(3)} \lambda_1^n + c_2 \mathbf{v}^{(3)} \lambda_2^n$$

where $\left\{\lambda_1^{(3)}, \mathbf{u}^{(3)}\right\}$ and $\left\{\lambda_2^{(3)}, \mathbf{v}^{(3)}\right\}$ are the two eigen-pairs for the Jacobian matrix $J(R^{(3)}, F^{(3)})$. The eigenvalues $\left\{\lambda_1^{(3)}, \lambda_2^{(3)}\right\}$ being the roots of the quadratic equation

$$\lambda^2 - \left(2 - \frac{\beta c}{d}\right)\lambda + \left(1 - \frac{\beta c}{d}\right) + ac\left(1 - \frac{c/d}{a/\beta}\right) = 0$$

or

$$\left\{ \begin{array}{c} \lambda_1^{(3)} \\ \lambda_2^{(3)} \end{array} \right\} = \frac{1}{2}\left\{ \left(2 - \frac{\beta c}{d}\right) \pm \sqrt{\left(\frac{\beta c}{d}\right)^2 - 4\frac{\beta c^2}{d}(\varrho - 1)} \right\}$$

PROPOSITION 15. *The fixed point $(R^{(3)}, F^{(3)})$ is unstable if $\varrho < 1$.*

PROOF. If $\varrho < 1$, we have

$$\lambda_1^{(3)} > \frac{1}{2}\left\{ \left(2 - \frac{\beta c}{d}\right) + \sqrt{\left(\frac{\beta c}{d}\right)^2} \right\} = 1.$$

and the fixed point is unstable. □

For if $\varrho > 1$, we have two separate cases:

PROPOSITION 16. *$(R^{(3)}, F^{(3)})$ is asymptotically stable if*

$$\frac{\beta}{d} > 4(\varrho - 1) \qquad \textit{and} \qquad \frac{\beta c}{d} < 2. \tag{2.10}$$

PROOF. When (2.10) holds, we have

$$i) \quad \lambda_1^{(3)} > \frac{1}{2}\left(2 - \frac{\beta c}{d}\right) > \lambda_2^{(3)},$$

$$ii) \quad \lambda_1^{(3)} < \frac{1}{2}\left\{ \left(2 - \frac{\beta c}{d}\right) + \sqrt{\left(\frac{\beta c}{d}\right)^2} \right\} = 1,$$

and

$$iii) \quad \lambda_2^{(3)} > \frac{1}{2}\left\{ \left(2 - \frac{\beta c}{d}\right) - \sqrt{\left(\frac{\beta c}{d}\right)^2} \right\} = 1 - \frac{\beta c}{d}.$$

They can be combined to give

$$0 < 1 - \frac{\beta c}{2d} < \lambda_1^{(3)} < 1$$

and

$$1-\frac{\beta c}{d}<\lambda_2^{(3)}<1-\frac{\beta c}{2d}$$

with

$$\left|\lambda_2^{(3)}\right|<1.$$

□

PROPOSITION 17. *$(R^{(3)},F^{(3)})$ is also asymptotically stable if*

$$\frac{\beta}{4d}<\varrho-1<\frac{1}{c}. \tag{2.11}$$

PROOF. When $\varrho-1>\beta/4d$ holds, the eigenvalues are complex conjugate pair with modulus

$$\begin{aligned}\left|\lambda_k^{(3)}\right| &= \left(1-\frac{\beta c}{2d}\right)^2+\frac{\beta c^2}{d}(\varrho-1)-\left(\frac{\beta c}{2d}\right)^2\\ &= 1-\frac{\beta c}{d}+\frac{\beta c^2}{d}(\varrho-1)\end{aligned}$$

If $0<c(\varrho-1)<1$, then $\left|\lambda_k^{(3)}\right|<1$ rendering $(R^{(3)},F^{(3)})$ asymptotically stable. □

3. Viral Dynamics

3.1. A Simple Model. Human being are subject the intrusion of bacteria and virus. Some are harmless or beneficial to our body. An example is the collection of bacteria in our intestines. Others are harmful to our health and are to be eliminated. For this purpose, the human host routinely produces antibodies to destroy unwanted bacteria, viruses and other pathogens. For a simple model of the interaction between antibodies and pathogens, we let a_n and v_n be the concentration of antibodies and virus at stage n and idealize the actual problem using the following simplifying assumptions:

- Antibodies in a human host are generally lost over a period of time. Some degrade naturally after a finite life span and others are destroyed when performing the task of neutralizing virus particles (virions).
- The human host regular produces antibodies at the rate of an amount s_n per unit stage while losing a fraction μ of the existing antibodies through degradation and lost to the host.
- When left unchecked, the virions would grow and increase by a fraction r in each stage.
- The presence of virions stimulate a higher production rate of antibodies in proportion to the product v_na_n of the current antibody and virus concentration.
- The presence of antibodies enables the human host to eliminate a fraction of the bacteria also in proportion to the product of v_na_n.

With these hypotheses, we have the following pair of nonlinear difference equations governing the evolution of the antibody and virus concentrations:

$$a_{n+1} = s + (1-\mu)a_n + \beta a_n v_n = g(a_n, v_n) \tag{3.1}$$

$$b_{n+1} = (1+r)v_n - \gamma a_n v_n = h(a_n, v_n) \tag{3.2}$$

In (3.1) and (3.2), the parameters μ, β, r and γ are prescribed positive rate constants. While the infusion may vary from stage to stage, it will be assumed to be uniform in time in the following development so that $s_n = s > 0$. Given the initial concentrations of antibodies and virions,

$$a_0 = A, \qquad v_0 = V, \tag{3.3}$$

the evolution of two concentrations with stage is governed by the IVP defined by (3.1) - (3.3).

It is of some interest to point out that unlike the rabbit - fox problem where there is no immigration of either rabbits or foxes, here there is a regular infusion of antibodies. Under normal circumstances, the antibody synthesis rate s per unit stage is generated internally as a part of human physiochemistry. However, it may also be indirectly and artificially stimulated by external means (such as vaccination) should it be needed. The main question of interest is the prognosis of the long term health of the host. In terms of information from our simple model, we want to know the fixed points of the system (3.1) and the stability of each.

3.2. Fixed Points. To seek a possible stage-independent steady state, we look for a fixed point of the two difference equations by taking $a_n = a$ and $v_n = v$ so that (3.1) and (3.2) become

$$\begin{aligned} g(a,v) - a &= s - a(\mu - \beta v) = 0, \\ h(a,v) - v &= (r - \gamma a)\, v = 0. \end{aligned}$$

The second of these requires

$$v^{(1)} = 0 \qquad \text{or} \qquad a^{(2)} = \frac{r}{\gamma}.$$

Corresponding to each of these, the first equation determine a unique solution for the other unknown:

$$a^{(1)} = \frac{s}{\mu}\,, \qquad v^{(2)} = \frac{\mu}{\beta}\left(\frac{r}{\gamma} - \frac{s}{\mu}\right).$$

We have then two fixed points for the system of nonlinear difference equations:

$$(a^{(1)}, v^{(1)}) = \left(\frac{s}{\mu}, 0\right), \qquad (a^{(2)}, v^{(2)}) = \left(\frac{r}{\gamma}, \frac{\mu}{\beta}\left(\frac{r}{\gamma} - \frac{s}{\mu}\right)\right), \tag{3.4}$$

It should be noted that the two fixed points collapse to one so that

$$(a^{(1)}, v^{(1)}) = \left(\frac{s}{\mu}, 0\right) = (a^{(2)}, v^{(2)})$$

if $r/\gamma = s/\mu$ or, what is the same, if the *gain-loss ratio* r/γ of the virus is the same as the *gain-loss ratio* s/μ of the antibodies:

$$R = \frac{r/\gamma}{s/\mu} = 1.$$

The ratio R is called the *reproductive ratio* of the two populations.

3.3. Stability of $(a^{(1)}, v^{(1)})$. The key to the nature of the stability of $(a^{(1)}, v^{(1)})$ is the Jacobian matrix for that fixed point. It is straightforward to calculate the various partial derivatives to get

$$\begin{aligned} J^{(1)} &= J(\frac{s}{\mu}, 0) = \begin{bmatrix} g_{,a}\left(s/\mu, 0\right) & g_{,v}\left(a^{(k)}, v^{(k)}\right) \\ h_{,a}\left(s/\mu, 0\right) & h_{,v}\left(s/\mu, 0\right) \end{bmatrix} \\ &= \begin{bmatrix} 1-\mu & \beta s/\mu \\ 0 & 1+r-\gamma s/\mu \end{bmatrix} \end{aligned}$$

The two eigenvalues of $J^{(1)}$ are

$$\lambda_1^{(1)} = 1-\mu, \qquad \lambda_2^{(1)} = 1+\frac{\gamma s}{\mu}(R-1) \tag{3.5}$$

It follows from (3.5) that the fixed point $(s/\mu, 0)$ is a saddle point and *unstable* if $R > 1$ and is *asymptotically stable* if $R < 1$.

3.4. Stability of $(a^{(2)}, v^{(2)})$. For the fixed point $(a^{(2)}, v^{(2)})$, the Jacobian matrix is

$$\begin{aligned} J^{(2)} &= J(a^{(2)}, v^{(2)}) \\ &= \begin{bmatrix} 1-\gamma s/r & \beta r/\gamma \\ (1-R)s\gamma^2/\beta r & 1 \end{bmatrix} \end{aligned}$$

The two eigenvalues of $J^{(2)}$ are the solution of the quadratic equation

$$\lambda^2 - \left(2 - \frac{s\gamma}{r}\right)\lambda + \left(1 - \frac{s\gamma}{r} + \gamma s(R-1)\right),$$

or

$$\begin{pmatrix} \lambda_1^{(2)} \\ \lambda_2^{(2)} \end{pmatrix} = \left(1 - \frac{s\gamma}{2r}\right) \pm \sqrt{\left(\frac{s\gamma}{2r}\right)^2 - \gamma s(R-1)}. \tag{3.6}$$

It follows from (3.5) that

PROPOSITION 18. *The fixed point $(a^{(2)}, v^{(2)})$ is unstable and the fixed point $(a^{(1)}, v^{(1)})$ is asymptotically stable if $R < 1$.*

If, on the other hand, $R > 1$, then the situation is a little more complicated as indicated by the following conclusion:

PROPOSITION 19. *The fixed point $(a^{(2)}, v^{(2)})$ is asymptotically stable if either*

1. *$0 < R - 1 < s\gamma/4r^2$ with $0 < \lambda_2^{(2)} < \lambda_1^{(2)} < 1,$. or*
2. *$R - 1 > s\gamma/4r^2 > 0$ but $R - 1 < 1/r$ with $\lambda_1^{(2)}$ and $\lambda_2^{(2)}$ being complex conjugates and $0 < \left|\lambda_1^{(2)}\right| = \left|\lambda_2^{(2)}\right| < 1$.*

For the remaining possible but unlikely combination of parameter values, we have

PROPOSITION 20. *The fixed point $(a^{(2)}, v^{(2)})$ is unstable if $R - 1 > 1/r$ (and $R - 1 > s\gamma/4r^2 > 0$)*

For this last case, the polygonal trajectory of (a_n, v_n) spirals away from the fixed point $(a^{(2)}, v^{(2)})$.

3.5. Transcritical Bifurcation. We learned from the expressions (3.4) for the two fixed points that they coalesce for $R = 1$ so that the gain-loss ratio of both antibodies and virions are the same.but are distinct away from this critical value of the reproductive ratio. It appears that $R_c = 1$ is a likely bifurcation point and we should consider constructing a bifurcation diagram with R as the bifurcation parameter for the evolution of the two interacting populations.

For $R < 1$, the fixed point $\left(a^{(2)}, v^{(2)}\right)$ has a negative value for the virus concentration which is biologically not realizable while the only realizable fixed point $\left(a^{(1)}, v^{(1)}\right) = (s/\mu, 0)$ is *asymptotically stable* by Proposition 18 (which also concludes that $\left(a^{(2)}, v^{(2)}\right)$ is unstable for this range of reproductive ratio. When the gain-loss ratio for the antibody population is larger than that of the virus, we do expect the virus to be wiped out eventually.

For $R > 1$ (at the other side of the bifurcation point), the gain-loss ratio of the antibody is smaller than that of the virus, we expect the virus to be resistant to the action of the antibodies. With an "adequate" (daily) infusion rate of antibodies (so that $R - 1 < 1/r$) , Proposition 19 assures us that the struggle tends to a standoff, with (a_n, v_n) converging (either directly or spirally) to a steady state concentration for both population corresponding to $\left(a^{(2)}, v^{(2)}\right)$. Whether the human host can withstand a virus concentration at the level $v^{(2)}$ depends on the individual's constitution. It may or may not require clinical intervention to sustain the host, hopefully putting him or her on the road to recovery.

Together, the bifurcation diagram for the stability of the the two fixed points show a transcritical bifurcation at the bifurcation point $R = 1$.

The remaining possible combination of $R - 1 > s\gamma/4r^2$ and $R - 1 > 1/r$ corresponds to a high gain-loss ratio for virus, i.e., an extremely high virus proliferation rate, so that the virus population increases cyclically without bound. The host constitution is not strong enough to produce enough antibodies to keep the virus population under control. External intervention is necessary in order for the human host to survive the virus attack. It should be kept in mind that there are limitations to the conclusions based on linearization of the model equations for this case. The prediction may be modified when either population gets sufficiently large; the original model equations may provide a different (or at least more accurate) description of the evolution of the two populations.

4. Metabolism and Enzyme Kinetics

4.1. Metabolic Reactions. The partial debunking of the Central Dogma of biology by the discovery of virus notwithstanding, *proteins* (or their subunits of *amino acid* chains) are still central to sustaining life and reproduction. Among the various types of proteins is a group known as *enzymes* that are indispensible catalysts in the conversion of macro-molecules (such as the food we injest) to smaller molecules (such as sugar and lipid) more useful to human body. The conversion is done through one or more special biochemical reactions, known as *metabolic reactions.* In such reactions, some molecule or a group of them, known as the *substrate, S,* in the biochemistry (not to be confused with the substrates in material sciences), is bound to an enzyne E to form an enzyme-substrate complex C. The resulting molecular complex then transforms the substrate into a *product* (a term preferred by writers of enzyme kinetics over the more or less equivalent designation of *metabolite*), usually requiring some energy for the work done in this process.

(We may think of this type of metabolic process as one of an enzyme taking in a (raw material) substrate and making it into one or more finished products.) Upon releasing the product(s), the responsible enzyne is then freed and becomes available to repeat the same process with substrates still available for binding. In this way, an enzyme acts only as a catalyst in making the product or products and remains intact after the releasing the finished products. Such constructive metabolic reactions are known as *anabolisms*.

An example of an anabolic reaction is the (rennet) coagulation (or curd formation) resulting from adding the enzyme *rennin* to *milk*. In this reaction, the substrate is milk protein. Another example is adding the enzyme *lactasse* to the substrate *lactose* to result in the two products *glucose* and *galactose*. Not all metabolic reactions are anabolic. The enzyme *catalasse*, when added to the hydrogen peroxide, cause that chemical to decompose into it constituent parts. These parts are not products since they are not closer in usefulness for sustaining life or reproduction than the substrate. More directly relevant to the human body are the metabolic reactions that break down large protein molecules into amino acids and other simple compounds. Such destructive type of metabolic reactions are known as *protein catabolism*. Breaking down food molecules constitutes an essential part of our digestive process. Only smaller and simpler molecules can be transported through the plasma membrane of cells into the cell interior for production of new proteins. Other catabolic reactions may be for the purpose of releasing energy (and component parts) needed by anabolic reactions.

4.2. Anabolic Enzyme Kinetics. As an introduction to metabolic reactions, we consider the simplest model of anabolic enzyme kinetics involving one enzyme, one substrate and one product formed by way of the enzyme-substrate complex. The mathematical model is essentially a matter of bookkeeping of the concentration of substrate, enzyme and the compound form from them. The compound concentration C_{n+1} in stage $n+1$ must be

(1) what is already there at stage n , C_n,
(2) plus the new compounds formed from the binding of enzyme and substrate during that stage, $k_{se}E_nS_n$ with a binding rate constant k_{se},
(3) minus the dissociation of existing complex $k_{rb}C_n$ with a rate constant k_{rb}
(4) minus those lost from having completed the conversion of the substrate into a product, $k_{cp}C_n$, freeing up the enzyme for new anabolic reaction in the process.

Together, the accounting leads to the following nonlinear difference equation

$$C_{n+1} = C_n + k_{se}E_nS_n - k_{cp}C_n - k_{rb}C_n \tag{4.1}$$

The accounting of the substrate concentration is similar leading to

$$S_{n+1} = S_n - k_{se}E_nS_n + k_{rb}C_n. \tag{4.2}$$

The right hand side is the sum of the substrate concentration already there in stage n, the amount freed up by the dissociation of existing compound, and the loss from binding with substrate.

Finally the transformation of the enzyme-substrate complex into a product gives rise to the accounting equation

$$P_{n+1} = P_n + k_{cp}C_n \tag{4.3}$$

In principle, there should be another accounting equation for the enzyme concentration:

$$E_{n+1} = E_n - k_{se}E_nS_n + (k_{cp} + k_{rb})C_n \tag{4.4}$$

However, by adding (4.4) and (4.1), we get the conservation law

$$E_{n+1} + C_{n+1} = E_n + C_n$$

so that

$$E_n + C_n = E_0 + C_0 = I_0 \qquad \text{or} \qquad E_n = I_0 - C_n \tag{4.5}$$

where I_0 is the *known* sum of the initial enzyme concentration E_0 and enzyme-substrate complex concentration C_0.

With (4.5), we may rewrite the first two accounting equations as

$$C_{n+1} = k_{se}I_0S_n + (1 - k_{cp} - k_{rb})C_n - k_{se}S_nC_n \tag{4.6}$$

The second accounting equation for the substrate concentration similarly leads to

$$S_{n+1} = (1 - k_{se}I_0)S_n + k_{rb}C_n + k_{sc}C_nS_n \tag{4.7}$$

The two difference equations (4.6)-(4.7) together with the prescribed initial conditions

$$C_0 = \bar{c}, \qquad S_0 = \bar{s} \tag{4.8}$$

define an IVP that determines the substrate and complex concentrations for future stages. The single first order difference equation (4.3) and the initial condition

$$P_0 = \bar{p}$$

determine the amount of product for future stages. Future enzyme concentration can also be calculated from (4.5).

The system of equations for our simple metabolic reaction was first formulated by Leonor Michaelis and Maud Menton in 1913 based on Victor Henri's discovery ten years earlier of the binding between substrate to a catalyst enzyme protein to initiate the production of metabolites. It is known as the Michaelis-Menton kinetics.

4.3. Fixed Point Analysis. Experience from the last three subsections would have looking for fixed points and their stability. For our enzyme kinetics problem, a stage independent solution (S, C) reduces (4.6) - (4.8) to

$$\begin{aligned} 0 &= k_{se}I_0S - (k_{cp} + k_{rb})C - k_{se}SC, \\ 0 &= -k_{se}I_0S + k_{rb}C + k_{sc}CS. \end{aligned}$$

Upon adding both sides of these two equations results in

$$-k_{cp}C = 0$$

leading to

$$C = S = 0$$

as the only fixed point of the enzyme kinetics equations.

For its stability, we calculate the corresponding Jacobian matrix

$$J = \begin{bmatrix} 1 - k_{cp} - k_{rb} & k_{se}I_0 \\ k_{rb} & 1 - k_{se}I_0 \end{bmatrix}$$

where biological reality requires $0 < 1 - k_{cp} - k_{rb} < 1$ and $0 < 1 - k_{se}I_0 < 1$. From $|J - \lambda I| = 0$, we get

$$\lambda^2 - (2 - k_{cp} - k_{rb} - k_{se}I_0)\lambda + \{1 - k_{cp} - k_{rb} - k_{se}I_0(1 - k_{cp})\} = 0.$$

with $0 < 1 - k_{cp} < 1$ in addition to the restriction on the parameters previously stipulated above. It follows that

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \alpha \pm \sqrt{\alpha^2 - \beta}$$

where the positive constants α and β

$$\begin{aligned} \alpha &= \frac{1}{2}(1 - k_{cp} - k_{rb}) + \frac{1}{2}(1 - k_{se}I_0) < 1 \\ \beta &= 2\alpha + k_{se}I_0(1 - k_{cp}) \end{aligned}$$

Since $\alpha^2 - \beta < \alpha^2$ and $0 < \alpha < 1$, we have

$$0 < \lambda_1 < 2\alpha < 2, \qquad 0 < \lambda_2 < \alpha$$

and the fixed point is asymptotically stable.

If the simple outcome above is disappointing in light of the more interesting and complex results for the previouse three examples, it should not be. In fact, it should have been anticipated without carrying out the fixed point analysis. Given the fact that enzymes serve only as a catalyst and its concentration is conserved, the metabolic reaction can continue until the substrates are all converted into products. With substrates exhausted, no additional enzyne-substrate complexes can be form after all existing one have been converted into products. It is not surprising then that the only fixed point of the problem should be $(0, 0)$ and it is asymptotically stable. We really do not need to do any fixed point analysis to arrive at that result.

4.4. Constant Substrate Approximation. Instead of looking for the limiting behavior (which is rather intuitively evident), the interest in enzyme-substratc reactions is the reaction process itself, particularly the speed of conversion from substrate to product. Depending on that speed, the metabolic reaction may or may not accomplish its mission in sustaining life. More often, the product of a particular anabolic reaction is only an intermediate step toward the human needs. It has to be produced in time for the next set of reactions that make use of it. There are also anabolic reactions that lead to products poisonous to health. In all case, clinical intervention may be necessary when a needed product is produced too slowly or too much of an unwanted product is produced. For the answer on the speed or other features of a reaction, we need to know more than the steady state behavior, possibly the actual evoluation of the various concentration in extreme cases,

Unfortunately, the two equations for C_n and S_n are nonlinear and no simple explicit solution is readily available. For the evolutionary details of the reaction, we describe the following classical approximate approach to the IVP for the case where there is an abundance of substrates. With the change of substrate concentration being only a tiny fraction of the initial substrate concentration, S_n remains pretty much unchanged as n increases, so that $S_{n+1} \simeq S_n \simeq S_0$ and (4.6) may be approximated adequately by

$$C_{n+1} = k_{se}I_0S_0 + (1 - \alpha)C_n \tag{4.9}$$

where

$$\alpha = k_{cp} + k_{rb} + k_{se}S_0. \tag{4.10}$$

The approximate version of (4.6) is a linear first order difference equaiton with immigration whose solution is

$$C_n = \frac{k_{se}}{\alpha} I_0 \left\{1 - (1-\alpha)^n\right\} + C_0(1-\alpha)^n \tag{4.11}$$

Note that whatever the values for the parameters k_{cp}, k_{rb} and $k_{se}S_0$may be, we must have $0 < 1 - \alpha < 1$ since we cannot take away (by the various molecular processes) more that the whole substrate concentration at any stage. In that case, we have

$$\lim_{n\to\infty} C_n = \frac{k_{se}}{\alpha} I_0 S_0 = \frac{I_0 S_0}{K_m + S_0} \equiv C_\infty$$

where

$$K_m = \frac{k_{cp} + k_{rb}}{k_{se}} \tag{4.12}$$

is known as the *Michaelis constant*. This limiting behavior for C_n simply says that the eventual concentration of enzyme-substrate complexes corresponds to that fraction of initial coneentration of enzymes and complexes not lost to dissociation or conversion to product.

Whatever the limiting value of the complex concentration, the existence of a positive C_∞ is inconsistent with $(0, 0)$ being the only fixed point and the intuition that all complexes will be turned into products. It is not difficult to see that the inconsistency exists only because we apply the constant substrate approximation beyond it range of applicability. Sooner or later, the substrate would be reduced so significantly that the reduction is no longer an insignificant fraction of the initial concentration. By then, we should not continue to apply the result based on the constant substrate concentration.

4.5. Quasi-Steady State Approximation. To handle the range of substrate concentration that deviates substantially from the initial concentration, we make use of the fact that the complex concentration is approaching C_∞ at that point. This means C_n is no longer changing significantly. We may therefore make the reasonable quasi-steady state approximation $C_{n+1} \simeq C_n$ so that (4.6) may be approximated by

$$k_{se}I_0S_n - (k_{cp} + k_{rb})C_n - k_{se}S_nC_n = 0 \tag{4.13}$$

or, with $I_0 = E_0$ (given that there is no enzyme-substrate complexes at the start of the reaction)

$$C_n = \frac{E_0 S_n}{K_m + S_n} \tag{4.14}$$

This expression may be used to eliminate C_n from (4.13) to get a single first order difference equation for S_n:

$$S_{n+1} = S_n - \frac{k_{cp}E_0S_n}{K_m + S_n} \tag{4.15}$$

There is no simple analytic solution for the nonlinear difference equation (4.15). Moreover, there is also no definitive choice for an "initial" condition for S_k at stage k when we begin to apply the quasi-steady state approximation for the substrate

concentration to supplement (4.13) in this range. If we should take S_k to be the substrate concentration corresponding to C_∞ approximately, then the relation (4.13) requires $S_k = S_0$.

Fortunately, for the issue product conversion rate (conversion speed or velocity of reactions), we do not need to solve this IVP (with uncertain "initial" condition). The velocity of reaction (= product conversion rate) is given by

$$\begin{aligned} V &= P_{n+1} - P_n = k_{cp} C_n \\ &= \frac{k_{cp} E_0 S_n}{K_m + S_n} \end{aligned}$$

Since V is a monotone increasing function of S_n (why?), its maximum value $V_{\max}$ is given by

$$V_{\max} = k_{cp} E_0. \tag{4.16}$$

and we can rewrite the expression for V as

$$V = \frac{V_{\max} S_n}{K_m + S_n}. \tag{4.17}$$

With (4.17), it is not difficult see the Michaelis constant is that value of substrate corresponding a conversion rate equal to half of the maximum value:

$$[V]_{S_n = K_m} = \frac{1}{2} V_{\max}$$

4.6. The Lineweaver-Burk Plot. Since V is a monotone increasing function of S_n and S_n is a monotone decreasing function of stage (as more and more substrates have been converted to products), the maximum value of V for the prescribed initial substrate is attained at the initial substrate concentration S_0 and given by

$$V_0 = \frac{V_{\max} S_0}{K_m + S_0}. \tag{4.18}$$

This maximum value V_0 of V for a prescribed set of initial data should not be confused with $V_{\max} (> V_0)$ which is approached only as $S_n \to \infty$. For a particular reaction, this maximum reaction velocity (or product conversion speed) is the principal item of interest for the purpose of the human body working toward some proper outcome. For that reason, V_0 is simply called *the reaction velocity* V in the metabolic reaction literature with the understanding that it is for the initial substrate concentration S_0. To avoid confusion, we continue to call it V_0 in the subsequent development.

Evidently, the reaction velocity of a anabolic reaction is known once with have the values of the two parameters $V_{\max}$ and K_m. These values will have to be obtained from suitable data. To assist with the estimation of these two parameter values, we re-arrange the relation (4.18) to read

$$\frac{1}{V_0} = \frac{1}{V_{\max}} \left(1 + \frac{K_m}{S_0} \right). \tag{4.19}$$

With $v = 1/V_0$ as a function of $s = 1/S_0$, the relation (4.19) is linear. The graph of v vs. s is a straight line with $v_{\max} = 1/V_{\max}$ being the intercept on the v–axis and $\kappa = K_m/V_{\max}$ as the slope of the straightline.

Suppose we have run the reaction for a set of initial substrate concentrations $\{S_0^{(k)}\}$ with $\{V_0^{(k)}\}$ the corresponding initial reaction velocity. With error in measurements and the set up for the repeated reaction runs not perfectly identical (ambient temperature not the same for all runs, etc.), the points $(1/S_0^{(k)}, 1/V_0^{(k)})$ would not fall on a straightline. They can be fitted by a straightline $v = v_{\max} + \kappa s$ by determining $v_{\max}$ and κ by a least square fit (as in Chapter 1 of these notes). With $V_{\max} = 1/v_{\max}$ and $K_m = \kappa V_{\max}$, the reaction velocity (4.17) and (4.18) can now be used to determine the speed of conversion to products for any substrate concentration.

Because the Michaelis constant is not determined exactly but only by a least square fit, it is an approximation of its true value. As a check of its degree of accuracy, we may use the fact that $V = V_{\max}/2$ when $S_n = K_m$ as another way of calculating K_m. Go the point $v = 2/V_{\max}$ and see from the straightline graph the corresponding value of s_2 . The resulting $K_m = 1/s_2$ will not likely be the same as that already determined from the slope κ. However, if the discrepancy should be large, more data point should be used for the least square fit for the straightline.

CHAPTER 10

Bistability

As seen from the first three sections of this chapter and the related exercises, fixed points and their stability have been found useful in the study of interacting populations. In these models involving nonlinear interactions, there exist several fixed points but only one of them, at most, is asymptotically stable though that stable fixed point may be in a different location depending on the values of the various rate constants. One of the important phenomena of interest in biology is the presence of more than one stable fixed points giving rise to an uncertainty in the steady state behavior of the interacting populations. The obvious question, both interesting and important, is to which stable fixed point do the populations converge eventually? If the populations are in one stable fixed point, what does it take to move to the other stable fixed points?

In this section, we discuss a few simple mathematical example of two intereacting populations for which there are two stable fixed points. Some biological motivation will be provided for these examples of bistability.

1. A Dimerized Reaction

1.1. The Model. The first mathematical model of two populations S_n and C_n to be considered is to interact according to the following two difference equations:

$$S_{n+1} = (1-\alpha)S_n + \beta C_n \tag{1.1}$$

$$C_{n+1} = (1-\gamma)C_n + \frac{\mu S_n^2}{1+S_n^2} \tag{1.2}$$

For some biological motivation, we recall from the previous section the two equations for substrate and substrate-enzyme complex concentrations (4.2) and (4.1) but specialized them to the case of abundant enzymes (relative to the substrates) so that we may take E_n to be approximately E_0 at least for a few stages:

$$S_{n+1} = (1-k_{se}E_0)S_n + k_{rb}C_n. \tag{1.3}$$

$$C_{n+1} = C_n(1-k_{cp}-k_{rb}) + k_{se}E_0S_n^* \tag{1.4}$$

In the last term of the equation for C_{n+1}, we introduce notion of a dimer.which is formed by two copies of the substrate molecule. Its contriubtion to the synthesis of the substrate-enzyme complex is taken to be equivalent to $S_n^* = S_n^2/(1+S_n^2)$ instead.

1.2. Fixed Points and Bifurcation. For the fixed points (S, C) of this system, we have

$$C = \frac{\alpha}{\beta}S, \qquad C = \frac{\mu}{\gamma}\frac{S^2}{1+S^2}.$$

The two relations give rise to three fixed points,

$$S^{(1)} = 0, \qquad \begin{pmatrix} S^{(2)} \\ S^{(3)} \end{pmatrix} = \sigma \mp \sqrt{\sigma^2 - 1}, \tag{1.5}$$

$$C^{(k)} = \frac{\alpha}{\beta} S^{(k)}. \tag{1.6}$$

The number of fixed points depends on the magnitude of σ with

$$2\sigma = \frac{\mu\beta}{\alpha\gamma} > 0. \tag{1.7}$$

- There is only one fixed point $(S^{(1)}, C^{(1)}) = (0, 0)$ if $0 < \sigma < 1$.
- There are two fixed points with $(S^{(2)}, C^{(2)}) = (\sigma, \mu/2\gamma)$ if $\sigma = 1$.
- There are three fixed points given by (1.5) and (1.6) if $\sigma > 1$.

Evidently, there is a bifurcation point at $\sigma = 1$. With the number of fixed points changing from 3 to 1 as σ decrease from $\sigma > 1$ through the bifurcation point to $\sigma < 1$, the bifurcation might appear to be the pitchfork type. However, a plot of a bifurcation diagram (even before determining the stability of these fixed points would show that the bifurcation is *saddle-node* with $(S^{(2)}, C^{(2)})$ and $(S^{(3)}, C^{(3)})$ for $\sigma > 1$ coalescing at $\sigma = 1$ and disappearing for smaller values of σ. All the while the fixed point $(S^{(1)}, C^{(1)}) = (0, 0)$ is unaffected by the change in σ, both its location and (as we shall see) its stability.

1.3. Linear Stability Analysis. For the purpose of illustrating bistability, we are interested principally in the range $\sigma > 1$ for which we have three distinct fixed points. For the stability of these fixed points, we need the Jacobian matrix for the problem:

$$J_k = \begin{bmatrix} 1 - \alpha & \beta \\ \frac{2\mu S^{(k)}}{\left(1 + \left[S^{(k)}\right]^2\right)^2} & 1 - \gamma \end{bmatrix}.$$

1.3.1. *Stability of* $(S^{(1)}, C^{(1)})$. For the fixed point $(0, 0)$, the Jacobian becomes

$$J_1 = \begin{bmatrix} 1 - \alpha & \beta \\ 0 & 1 - \gamma \end{bmatrix}.$$

The corresponding eigenvalues are

$$0 < \lambda_1^{(1)} = 1 - \alpha < 1, \qquad 0 < \lambda_2^{(1)} = 1 - \gamma < 1.$$

PROPOSITION 21. *The fixed point* $(S^{(1)}, C^{(1)}) = (0, 0)$ *is asymptotically stable for any positive value of* σ *(and certainly for* $\sigma > 1$*)*

Since there is no change in the stability of $(S^{(1)}, C^{(1)}) = (0, 0)$ for all σ, the fixed point is not involved in any bifurcation that may be in the reaction.

1.3.2. *Stability of* $(S^{(k)}, C^{(k)})$ *for* $\sigma > 1$. For the other two fixed points with $S^{(k)} = \sigma \mp \sqrt{\sigma^2 - 1}$, the Jacobian matrix is

$$J_k = \begin{bmatrix} 1 - \alpha & \beta \\ \frac{\mu/2\sigma^2}{\sigma \mp \sqrt{\sigma^2 - 1}} & 1 - \gamma \end{bmatrix} = \begin{bmatrix} 1 - \alpha & \beta \\ \frac{\mu}{2\sigma^2}\left(\sigma \pm \sqrt{\sigma^2 - 1}\right) & 1 - \gamma \end{bmatrix} \qquad (k = 2, 3). \tag{1.8}$$

For $\sigma > 1$, the two "larger" eigenvalues $\lambda_1^{(k)}$ of the Jacobian matrix (1.8) for the fixed points $(S^{(k)}, C^{(k)})$, $k = 2$ and 3, are

$$\begin{pmatrix} \lambda_1^{(2)} \\ \lambda_1^{(3)} \end{pmatrix} = \frac{1}{2}\left\{(2-\alpha-\gamma) + \sqrt{(\alpha-\gamma)^2 + \frac{2\beta\mu}{\sigma^2}\left(\sigma \pm \sqrt{\sigma^2-1}\right)}\right\}$$

while the two "smaller" eigenvalues $\lambda_2^{(k)}$ are

$$\begin{pmatrix} \lambda_2^{(2)} \\ \lambda_2^{(3)} \end{pmatrix} = \frac{1}{2}\left\{(2-\alpha-\gamma) - \sqrt{(\alpha-\gamma)^2 + \frac{2\beta\mu}{\sigma^2}\left(\sigma \pm \sqrt{\sigma^2-1}\right)}\right\}$$

For the fixed point $(S^{(2)}, C^{(2)}) = (\sigma - \sqrt{\sigma^2-1})(1, \alpha/\beta)$, we have

$$\begin{aligned} \lambda_1^{(2)} &> \frac{1}{2}\left\{(2-\alpha-\gamma) + \sqrt{(\alpha-\gamma)^2 + \frac{2\beta\mu}{\sigma}}\right\} \\ &= \frac{1}{2}\left\{(2-\alpha-\gamma) + \sqrt{(\alpha-\gamma)^2 + 4\alpha\gamma}\right\} = 1 \end{aligned}$$

and consequently the following result:

PROPOSITION 22. *The fixed point $(S^{(2)}, C^{(2)})$ is unstable for $\sigma > 1$.*

For the fixed point $(S^{(3)}, C^{(3)})$, we have

$$\begin{aligned} 0 &< 1 - \frac{1}{2}(\alpha+\gamma) < \lambda_1^{(3)} \\ &< \frac{1}{2}\left\{(2-\alpha-\gamma) + \sqrt{(\alpha-\gamma)^2 + \frac{2\beta\mu}{\sigma}}\right\} = 1. \end{aligned} \tag{1.9}$$

Similarly, we have

$$\begin{aligned} \lambda_2^{(3)} &< \frac{1}{2}\left\{(2-\alpha-\gamma) - \sqrt{(\alpha-\gamma)^2}\right\} \\ &< \frac{1}{2}\begin{cases} 2-2\alpha & (\alpha < \gamma) \\ 2-2\gamma & (\gamma < \alpha) \end{cases} < 1. \end{aligned}$$

To show $\lambda_2^{(3)} > 0$, we note that

$$\begin{aligned} 4(1-\alpha)(1-\gamma) + \frac{2\beta\mu}{\sigma^2}[\sigma - \sqrt{\sigma^2-1}] &> 4(1-\alpha)(1-\gamma) + \frac{2\beta\mu}{\sigma} \\ &= 4(1-\alpha-\gamma), \end{aligned}$$

we have

$$\begin{aligned} \lambda_2^{(3)} &> \frac{1}{2}\left\{(2-\alpha-\gamma) - \sqrt{(\alpha-\gamma)^2 + \frac{2\beta\mu}{\sigma}}\right\} \\ &= \frac{1}{2}\left\{(2-\alpha-\gamma) - \sqrt{(\alpha-\gamma)^2 + 4\alpha\gamma}\right\} \\ &= \frac{1}{2}\left\{(2-\alpha-\gamma) - (\alpha+\gamma)\right\} = 1 - (\alpha+\gamma) > -1 \end{aligned}$$

Altogether, we have $\left|\lambda_2^{(3)}\right| < 1$ and proved the following proposition: (

PROPOSITION 23. *The fixed point $(S^{(3)}, C^{(3)})$ is asymptotically stable for $\sigma > 1$.*

Altogether for the range $\sigma > 1$, we have (for the biologically realistic range of values of α and γ) two *asymptotically stable* fixed points at $(S^{(1)}, C^{(1)})$ and $(S^{(3)}, C^{(3)})$. This establishes the possible existence of bistability for this reaction.

With the two distinct fixed points $(S^{(2)}, C^{(2)})$ and $(S^{(3)}, C^{(3)})$ exhibiting different stability types for $\sigma > 1$, coalescing to same fixed point for $\sigma = 1$ and transitioning to no fixed point for smaller σ, the reaction has a saddle-node bifurcation with $\sigma = 1$ bing the bifurcation point.

2. Two Competing Populations

Many biological processes involve the phenomenon of bistability. While the mathematics involved is similar to that used in the simple mathematical example above, the description of most of these processes would require considerable background information in biology and biochemistry. For example, cytoplasmic calcium concentration,denoted by $\left[Ca^{2+}\right]_i$, is of great importance for the life and dealth of cells and bistability occurs in conjunction with the $\left[Ca^{2+}\right]_i$ concentration in many cells, including bullfrog sympathetic ganglion neurons and pancreatic beta cells (see [**4**] and references therein). Another group of rather complex biological processes exhibiting bistability is the area of morphogenesis in developmental biology. One simpler one is the role of cytoplasmic calcium concentration ($\left[Ca^{2+}\right]_i$ again) in the folding of a sheet of cells in the shaping of tissues and organs. The effects of $\left[Ca^{2+}\right]_i$ on the Newtonian dynamics of the shape changes of a cell sheet lead to the possibility of bistability on the sheet shape in a certain range of system parameter values (see [**14**] and references therein). Instead of embarking on a description of the modeling and analysis of these or other rather complex biological processes at the cellular level, we formulate and analyze in this section a model of two competing populations which is only a natural extension of those we have become totalling familiar and at ease in this and earlier chapters.

2.1. Leopards and Hyenas. Leopards and Hyenas both feed on smaller animals such impalas, gazelles, dikdiks and other members of the antelope family (and other small animals cohabiting in the same habitat). As the only predator for these smaller animals, each (Leopards or Hyenas) population would multiply subject to some size limiting growth rate.which we will take to be the logistic growth rate. When both xo-exist in the same habitat, as in the East Africa game reserves (e.g., Serengeti National Park), they would have to compete with each other for the same preys. While they do not prey on each other (except for occasional fights for the same kills), the presence of each inhibits the growth of the other. The growth of the two competing populations is adequately modeled by the following two first order nonlinear difference equations:

$$\begin{aligned} L_{n+1} &= (1+a)L_n - \alpha L_n^2 - bL_nH_n \equiv g(L_n, H_n). \\ H_{n+1} &= (1+c)H_n - \gamma H_n^2 - dL_nH_n \equiv h(L_n, H_n) \end{aligned} \tag{2.1}$$

As usual, we limit

- α, γ, b and d to be all positive;
- a and c to be generally in the interval $(0, 1)$, i.e., $0 < a, c < 1$ allowing $0 < a, c < 2$ in extreme cases.

Note that without the two interaction terms ($b = d = 0$), the leopards and hyenas simply grow according to their own logistic growth rate. The terms proportional to $-L_nH_n$ constitute the negative effect (inhibition) on the growth of the competiting population. As such, the system (2.1) is a generalization of the model in an assigned exercise where there is no limit imposed on the natural growth of either competing populations and the inhibiting effect on each other is identical.

2.2. Fixed Points. With $L_n = L$ and $H_n = H$, the system (2.1) becomes

$$\begin{aligned} g(L,H) - L &= L(a - \alpha L - bH) = 0, \\ h(L,H) - H &= H(c - \gamma H - dL) = 0. \end{aligned} \tag{2.2}$$

The four solutions of (2.2) are

$$\begin{aligned} (L^{(1)}, H^{(1)}) &= (0,0), \qquad (L^{(2)}, H^{(2)}) = \left(\frac{a}{\alpha}, 0\right), \\ (L^{(3)}, H^{(3)}) &= \left(0, \frac{c}{\gamma}\right), \\ (L^{(4)}, H^{(4)}) &= \left(\frac{bc - a\gamma}{db - \alpha\gamma}, \frac{da - \alpha c}{db - \alpha\gamma}\right). \end{aligned} \tag{2.3}$$

Since $a > 0$ and $c > 0$, the fixed points $(L^{(2)}, H^{(2)})$ and $(L^{(3)}, H^{(3)})$ do *not* degenerate to coincide with $(L^{(1)}, H^{(1)}) = (0,0)$ or with each other. (Otherwise, there may be bifurcation in the model.) The possible coalescing of $(L^{(4)}, H^{(4)})$ and one or more other fixed points will be investigated after we determine the stability of some of the fixed points.

2.3. Linear Stability Analysis. To determine the stability of the fixed points, we need the Jacobian matrix for the system (2.1):

$$J_k = \begin{bmatrix} 1 + \alpha - 2\alpha L^{(k)} - bH^{(k)} & -bL^{(k)} \\ -dH^{(k)} & 1 + \gamma - 2\gamma H^{(k)} - dL^{(k)} \end{bmatrix}$$

2.3.1. *Stability of* $(L^{(1)}, H^{(1)})$. For the fixed point $(L^{(1)}, H^{(1)}) = (0,0)$, the Jacobian becomes

$$J_1 = \begin{bmatrix} 1 + \alpha & 0 \\ 0 & 1 + \gamma \end{bmatrix}.$$

The two eigenvalues are

$$\lambda_1^{(1)} = 1 + \alpha > 1, \qquad \lambda_2^{(1)} = 1 + c > 1.$$

PROPOSITION 24. *The the fixed point* $(L^{(1)}, H^{(1)}) = (0,0)$ *is unstable independent of the values of all the other parameters.*

The fact that $(L^{(1)}, H^{(1)}) = (0,0)$ do not it stability type for all parameter values (given the restrictions $0 < a, c < 1$) shows that this fixed point is not involved in any bifurcation of the growth of the two competing populations.

2.3.2. *Stability of* $(L^{(2)}, H^{(2)})$. For the fixed point $(L^{(2)}, H^{(2)}) = (a/\alpha, 0)$, the Jacobian becomes

$$J_2 = \begin{bmatrix} 1 - \alpha & -ba/\alpha \\ 0 & 1 - (da - \alpha c)/\alpha \end{bmatrix}.$$

The two eigenvalues are

$$\lambda_1^{(2)} = 1 - \alpha < 1, \qquad \lambda_2^{(2)} = 1 + c - \frac{da}{\alpha} = 1 - \frac{da - \alpha c}{\alpha}.$$

PROPOSITION 25. *The fixed point* $(L^{(2)}, H^{(2)}) = (a/\alpha, 0)$ *is*
i) unstable if $a/\alpha < c/d$,
ii) asymptotically stable if $-1 < 1 + c - da/\alpha < 1$ *or*

$$\frac{c}{d} < \frac{a}{\alpha} < \frac{2+c}{d}.$$

In general, we expect $0 < a, c < 1$ so that $(2+c)/d > 2/d$ while $a/\alpha < 1/\alpha$. Together, these inequalities require $c/a < d/\alpha < 2$ which is expected to be met by biologically realistic competing populations so that the fixed point is asymptotically stable if

$$\frac{c}{d} < \frac{a}{\alpha}. \tag{2.4}$$

The fact that $(L^{(2)}, H^{(2)}) = (a/\alpha, 0)$ switches its stability as the parameter $\mu = c\alpha - ad$ changes sign suggests the possibility of a bifurcation of the growth of the two competing populations at $\mu = 0$. This possibility will be investigated after we have determined the stability of the remaining fixed points.

2.3.3. *Stability of* $(L^{(3)}, H^{(3)})$. For the fixed point $(L^{(3)}, H^{(3)}) = (0, c/\gamma)$, the Jacobian becomes

$$J_3 = \begin{bmatrix} 1 - (bc - \gamma a)/\gamma & 0 \\ -dc/\gamma & 1 - c \end{bmatrix}.$$

The two eigenvalues are

$$\lambda_1^{(3)} = 1 - \frac{bc - \gamma a}{\gamma}, \qquad \lambda_2^{(3)} = 1 - c$$

PROPOSITION 26. *The fixed point* $(L^{(3)}, H^{(3)}) = (0, c/\gamma)$ *is*
i) unstable if $c/\gamma < a/b$,
ii) asymptotically stable if $-1 < 1 + a - bc/\gamma < 1$ *or*

$$\frac{a}{b} < \frac{c}{\gamma} < \frac{2+a}{b}.$$

In general, we expect $0 < a, c < 1$ so that $(2+a)/b > 2/b$ while $c/\gamma < 1/\gamma$. Together, these inequalities require $a/c < b/\gamma < 2$ which is expected to be met by biologically realistic competing populations so that the fixed point is asymptotically stable if

$$\frac{a}{b} < \frac{c}{\gamma}. \tag{2.5}$$

Similar to the situation encountered in the stability of $(L^{(2)}, H^{(2)})$, the fact that $(L^{(3)}, H^{(3)}) = (0, c/\gamma)$ switches its stability as the parameter $\zeta = a\gamma - bc$ changes sign suggests the possibility of a bifurcation at $\zeta = 0$. This possibility will be discussed after we have determined the stability of $(L^{(4)}, H^{(4)})$.

2.4. Existence of Bistability. If $c/d < a/\alpha$ and $a/b < c/\gamma$, both $(L^{(2)}, H^{(2)}) = (a/\alpha, 0)$ and $(L^{(3)}, H^{(3)}) = (0, c/\gamma)$ are asymptotically stable. It follows that bistability exists for our competing populations of leopards and hyenas, assuming that the biologically realistic conditions

$$\frac{c}{\gamma} < \frac{2+a}{b}, \qquad \frac{a}{\alpha} < \frac{2+c}{d}$$

are also met. Having established the existence of this bistability, there remains two questions of interest:

1) Is $(L^{(4)}, H^{(4)})$ stable or unstable when $(L^{(2)}, H^{(2)})$ and $(L^{(3)}, H^{(3)})$ are asymptotically stable?

2) Are there other bistability configurations for our model?

We address the first question in the next subsection assuming the conditions (2.4) and (2.5) to ensure the asymptotic stability of $(L^{(2)}, H^{(2)})$ and $(L^{(3)}, H^{(3)})$.

2.5. Stability of $(L^{(4)}, H^{(4)})$. For the fixed point $(L^{(4)}, H^{(4)}) = (bc - a\gamma, da - \alpha c)/(db - \alpha\gamma)$, the Jacobian becomes

$$J_4 = \begin{bmatrix} 1 - \alpha L^{(4)} & -bL^{(4)} \\ -dH^{(4)} & 1 - \gamma H^{(4)} \end{bmatrix}.$$

after using the expressions for $L^{(4)}$ and $H^{(4)}$ to simplify the elements of J_4. The two eigenvalues are

$$\begin{pmatrix} \lambda_1^{(4)} \\ \lambda_2^{(4)} \end{pmatrix} = \left\{ \frac{1}{2}\left(2 - \alpha L^{(4)} - \gamma H^{(4)}\right) \pm \sqrt{(\alpha L^{(4)} - \gamma H^{(4)})^2 + 4bdL^{(4)}H^{(4)}} \right\}.$$

We are interested here in the special case

$$\frac{c}{d} < \frac{a}{\alpha}, \qquad \frac{a}{b} < \frac{c}{\gamma} \tag{2.6}$$

for which both $(L^{(2)}, H^{(2)})$ and $(L^{(3)}, H^{(3)})$ are asymptotically stable. Note that from (2.6) follows the inequality $(c/d)(a/b) < (a/\alpha)(c/\gamma)$ or

$$\alpha\gamma < bd \tag{2.7}$$

given $ac > 0$. Important consequences of (2.7) include

- $L^{(4)} > 0$ and $H^{(4)} > 0$ so that the fixed point $(L^{(4)}, H^{(4)})$ is biologically realizable.
- $4bdL^{(4)}H^{(4)} > 4\alpha\gamma L^{(4)}H^{(4)}$ so that

$$\sqrt{(\alpha L^{(4)} - \gamma H^{(4)})^2 + 4bdL^{(4)}H^{(4)}} > \sqrt{(\alpha L^{(4)} + \gamma H^{(4)})^2}. \tag{2.8}$$

PROPOSITION 27. *If either inequality of (2.6) holds along with (2.7), the fixed point $(L^{(4)}, H^{(4)})$ is unstable.*

PROOF. It follows from (2.8) that

$$\lambda_1^{(4)} > \left\{ \frac{1}{2}\left(2 - \alpha L^{(4)} - \gamma H^{(4)}\right) + \sqrt{(\alpha L^{(4)} + \gamma H^{(4)})^2} \right\} = 1.$$

□

COROLLARY 4. *If both inequalities of(2.6), the fixed point $(L^{(4)}, H^{(4)})$ is unstable.*

PROOF. The two inequalities in (2.6) implies (2.7) so that Proposition 27 applies. □

It is of some interest to investigate the stability of $(L^{(4)}, H^{(4)})$ for the inequalities

$$\frac{c}{d} > \frac{a}{\alpha} \quad \text{and} \quad \frac{a}{b} > \frac{c}{\gamma}, \tag{2.9}$$

instead of (2.6). From (2.9), we have $(c/d)(a/b) > (a/\alpha)(c/\gamma)$ so that

$$\alpha\gamma > bd. \tag{2.10}$$

It follows from (2.10)

$$\sqrt{(\alpha L^{(4)} - \gamma H^{(4)})^2 + 4bdL^{(4)}H^{(4)}} < \sqrt{(\alpha L^{(4)} + \gamma H^{(4)})^2}$$

so that

$$\lambda_1^{(4)} < \left\{ \frac{1}{2}\left(2 - \alpha L^{(4)} - \gamma H^{(4)}\right) + \sqrt{(\alpha L^{(4)} + \gamma H^{(4)})^2} \right\} = 1.$$

At the same time,

$$\begin{aligned} \lambda_2^{(4)} &> \left\{ \frac{1}{2}\left(2 - \alpha L^{(4)} - \gamma H^{(4)}\right) - \sqrt{(\alpha L^{(4)} + \gamma H^{(4)})^2} \right\} \\ &= 1 - \alpha L^{(4)} - \gamma H^{(4)}, \end{aligned}$$

so that

$$1 - \left(\alpha L^{(4)} + \gamma H^{(4)}\right) < \lambda_2^{(4)} < 1 - \frac{1}{2}\left(\alpha L^{(4)} + \gamma H^{(4)}\right) < 1. \tag{2.11}$$

PROPOSITION 28. *If either of the inequality in (2.9) holds along with (2.10), then the fixed point* $(L^{(4)}, H^{(4)})$ *is asymptotically stable.*

PROOF. Given (2.11), it remains to show that $1 - \left(\alpha L^{(4)} + \gamma H^{(4)}\right) > 0$ to complete the proof. This follows from the fact that

$$L_{n+1} = (1+a)L_n - \alpha L_n^2 - bL_nH_n = g(L_n, H_n) > 0.$$

More specifically, we have from $g(L^{(4)}, H^{(4)}) > 0$ and $L^{(4)}H^{(4)} > 0$

$$\begin{aligned} 0 &< 1 - \alpha L^{(4)} + a\left(1 - \frac{b}{a}H^{(4)}\right) \\ &< 1 - \alpha L^{(4)} + 1 - \frac{\gamma}{c}H^{(4)} \\ &< 1 - \alpha L^{(4)} + 1 - \gamma H^{(4)} \end{aligned}$$

for the second inequality in (2.9). If the first inequality in (2.9) should hold instead, a similar argument through $h(L^{(4)}, H^{(4)}) > 0$ leads to the same conclusion. □

COROLLARY 5. *If both inequalities in (2.9) hold, then the fixed point* $(L^{(4)}, H^{(4)})$ *is asymptotically stable.*

PROOF. The two inequalities in (2.9) implies (2.9) so that Proposition 28 applies. □

2.6. Other Bistability Configurations?

2.6.1. *No Bistability with Unstable* $(L^{(2)}, H^{(2)})$ *and* $(L^{(3)}, H^{(3)})$. If instead of (2.6), we have the opposite inequalities then $(L^{(2)}, H^{(2)})$ and $(L^{(3)}, H^{(3)})$ are unstable. Since $(L^{(1)}, H^{(1)})$ is also unstable independent of the relative size of the four products $bc, a\gamma, da$ and αc (not to mention $\alpha\gamma$ and bd), we have immediately the following proposition:

PROPOSITION 29. *Bistability does not exist if both inequalities in (2.9) hold.*

PROOF. The three fixed points $(L^{(1)}, H^{(1)})$, $(L^{(2)}, H^{(2)})$ and $(L^{(3)}, H^{(3)})$ are all unstable. Only $(L^{(4)}, H^{(4)})$ is asymptotically stable for this case by Proposition 28. □

2.6.2. *Stable* $(L^{(2)}, H^{(2)})$ *and Unstable* $(L^{(3)}, H^{(3)})$. Suppose (2.4) holds but (2.5) does not so that $(L^{(2)}, H^{(2)})$ is asymptotically stable but $(L^{(3)}, H^{(3)})$ is unstable. Whether there is bistability now depends on the stability of

$$(L^{(4)}, H^{(4)}) = \left(\frac{bc - a\gamma}{db - \alpha\gamma}, \frac{da - \alpha c}{db - \alpha\gamma} \right) \tag{2.12}$$

Now, the product $(c/d)(a/b)$ may be less than or greater than $(a/\alpha)(c/\gamma)$ (or db may be less or greater than $\alpha\gamma$ given the inequality $a/b > c/\gamma$ may more than compensate for (2.4), i.e.,

$$i)\ db - \alpha\gamma > 0, \qquad \text{or} \qquad ii)\ db - \alpha\gamma < 0.$$

(i) $db - \alpha\gamma > 0$: Together with $c/d < a/\alpha$, we have from Proposition 27 that $(L^{(4)}, H^{(4)})$ is unstable. This leaves $(L^{(2)}, H^{(2)})$ as the only asymptotically stable fixed point. Furthermore and hence no bistability in this case.

In addition, with $a/b > c/\gamma$ so that

$$L^{(4)} = \frac{bc - a\gamma}{db - \alpha\gamma} < 0$$

(while $H^{(4)}$ can be seen to be positive), the corresponding fixed point is not biologically realizable as we cannot have a negative leopard population.

(ii) $db - \alpha\gamma < 0$: Together with $a/b > c/\gamma$, we have from Proposition 28 that $(L^{(4)}, H^{(4)})$ is asymptotically stable. This fixed point and the asymptotically stable $(L^{(2)}, H^{(2)})$ constitute bistability for this case (while the other two remaining fixed points are unstable).

However, with $da > \alpha c$ so that

$$H^{(4)} = \frac{da - \alpha c}{db - \alpha\gamma} < 0$$

(while $L^{(4)}$ can be seen to be positive), $(L^{(4)}, H^{(4)})$ is not biologically realizable as we cannot have a negative hyenas population.

We summarize the results above in the following proposition:

PROPOSITION 30. *If* $c/d < a/\alpha$ *and* $a/v > c/\gamma$ *so that* $(L^{(2)}, H^{(2)})$ *is asymptotically stable and* $(L^{(3)}, H^{(3)})$ *is unstable, bistability occurs only if* $db < \alpha\gamma$ *with a biologically unrealizable* $(L^{(4)}, H^{(4)})$ *being asymptotically stable in that case.*

2.6.3. *Stable* $(L^{(3)}, H^{(3)})$ *and Unstable* $(L^{(2)}, H^{(2)})$. Now suppose (2.5) holds but (2.4) does not so that $(L^{(3)}, H^{(3)})$ is asymptotically stable but $(L^{(2)}, H^{(2)})$ is unstable. Whether there is bistability depends again on the stability of $(L^{(4)}, H^{(4)})$. The situation is a mirror image of the last subsection. Without repeating a similar argument, we summarize below a similar set of results regarding bistability for this case.

(i) $db - \alpha\gamma > 0$: Together with $a/b < c/\gamma$, we have from Proposition 27 that $(L^{(4)}, H^{(4)})$ is unstable. This leaves $(L^{(3)}, H^{(3)})$ as the only asymptotically stable fixed point and the model exhibits no bistability in this case.

In addition, with $c/d > a/\alpha$ so that

$$H^{(4)} = \frac{da - \alpha c}{db - \alpha\gamma} < 0$$

(while $L^{(4)}$ can be seen to be positive), the corresponding fixed point is not biologically realizable as we cannot have a negative leopard population.

(ii) $db - \alpha\gamma < 0$: Together with $c/d > a/\alpha$, we have from Proposition 28 that $(L^{(4)}, H^{(4)})$ is asymptotically stable. This fixed point and the asymptotically stable $(L^{(3)}, H^{(3)})$ constitute bistability for this case (while the other two remaining fixed points are unstable).

However, with $a/b < c/\gamma$ so that

$$L^{(4)} = \frac{bc - a\gamma}{db - \alpha\gamma} < 0$$

(while $H^{(4)}$ can be seen to be positive), $(L^{(4)}, H^{(4)})$ is not biologically realizable as we cannot have a negative leopard population.

We summarize the results above in the following proposition:

PROPOSITION 31. *If $a/b < c/\gamma$ and $c/d > a/\alpha$ so that $(L^{(3)}, H^{(3)})$ is asymptotically stable and $(L^{(2)}, H^{(2)})$ is unstable, bistability occurs only if $db < \alpha\gamma$ with a biologically unrealizable $(L^{(4)}, H^{(4)})$ being asymptotically stable in that case.*

2.7. Multiple Bifurcation Parameters.

2.7.1. *Bifurcation in the Parameter $\mu = c\alpha - ad$.* We found from stability studies above that

- If $\mu < 0$, the fixed point $(L^{(2)}, H^{(2)}) = (a/\alpha, 0)$ is asymptotically stable and $(L^{(4)}, H^{(4)}) = (bc - a\gamma, da - \alpha c)/(db - \alpha\gamma)$ is unstable (assuming for the latter $\Delta = db - \alpha\gamma > 0$ to keep $H^{(4)} > 0$ to be biologically realistic).
- If $\mu > 0$, we have from Proposition **??** that $(L^{(2)}, H^{(2)})$ is unstable and $(L^{(4)}, H^{(4)})$ is asymptotically stable (assuming for the latter $\Delta = db - \alpha\gamma < 0$ to keep $H^{(4)} > 0$ to be biologically realistic).
- If $\mu = 0$, we have $H^{(4)} = 0$ and $L^{(4)} = a/\alpha$ so that $(L^{(4)}, H^{(4)})$ degenerates to coincide with $(L^{(2)}, H^{(2)})$.

Together, they show that the growth of the competing populations has a transcritical bifurcation at $\mu = c\alpha - ad = 0$.

Similarly, the stability of $(L^{(3)}, H^{(3)})$ and $(L^{(4)}, H^{(4)})$ switches as ζ changes sign and $(L^{(4)}, H^{(4)})$ coincides with $(L^{(3)}, H^{(3)}) = (0, c/\gamma)$ at $\zeta = 0$. Hence, The competing population growth has another transcritical bifurcation at $\zeta = a\gamma - bc = 0$.

It is left to the reader to investigate the consequences of $\mu = 0$ and $\zeta = 0$ simultaneously. Note that we have also $\Delta = 0$ in that case.

3. Hopf Bifurcation

In the last few sections, we saw how the three kinds of bifurcation (saddle-node, transcritical and pitchfork) first discussed in Chapter 4 of these notes for a single first order difference equation also occur for second (or higher) order systems of interacting populations. For model involving only a single first order difference equation, these are the three basic types of bifurcation and others are usually distorted version or combination of them. However, for higher order systems, there is a fourth basic bifurcation type, known as the Hopf bifurcation. It is distinctly different from the other three. One of its distinct features is that they involve Jacobina matrices with complex eigenvalues. In this section, we use an example to

show the nature of a Hopf bifurcation and howpolar coordinates play an important role in studying such bifurcation.

3.1. A Mathematical Model. We consider here two interacting populations x_n and y_n at stage n evolving according to the following two nonlinear difference equations:

$$\begin{aligned} x_{n+1} &= (2+\mu-r_n^2)(0.3x_n - 0.4y_n) \\ y_{n+1} &= (2+\mu-r_n^2)(0.4x_n + 0.3y_n) \end{aligned} \tag{3.1}$$

where $r_n^2 = x_n^2 + y_n^2$ (the square of the *radial distance* between the origin of the $x, y-$plane and point (x_n, y_n) in that plane. These equations are supplemented by the initial conditions

$$x_0 = X, \qquad y_0 = Y \tag{3.2}$$

To be biologically realistic, we limit μ to be greater than -2, i.e., $\mu > -2$.

While the model is somewhat contrived, it is intended to provide a simple illustration of the phenomenon of Hopf bifurcation. It should be noted however that the appearance of the combination $r_n^2 = x_n^2 + y_n^2$ in the dynamics of the interacting processes is not so infrequent. Many factors influencing the evolution of the processes depends only on the distance from the source and not on the orientation. Central force fields in physics (is (such as the gravitational field determining planetary motion for example) are prime and ubiquitous examples of such factors. A source radiating heat is another.

3.2. Stability of the Fixed Point at the Origin. It is straightforward to verify that $(0,0)$ is a fixed point of our nonlinear system (3.1). To investigate its (linear) stability, we calculate the Jacobian matrix for that fixed point to get

$$J(0,0) = \begin{bmatrix} 0.3(2+\mu) & -0.4(2+\mu) \\ 0.4(2+\mu) & 0.3(2+\mu) \end{bmatrix}.$$

Its eigenvalues are determined by

$$[\lambda - 0.3(2+\mu)]^2 + [0.4(2+\mu)]^2 = 0$$

so that

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = (2+\mu)(0.3 \pm 0.4i) \equiv \frac{1}{2}(2+\mu)(\cos\theta \pm i\sin\theta. \tag{3.3}$$

with

$$|\lambda_k| = 0.5(2+\mu).$$

PROPOSITION 32. *The fixed point* $(0,0)$ *is unstable if* $\mu > 0$ *and asymptotically stable if* $-2 < \mu < 0$.

For the complex eigenvalues, we may write the general solution of the linearized system about the fixed point,

$$\begin{aligned} u_{n+1} &= (2+\mu)(0.3u_n - 0.4v_n) \\ v_{n+1} &= (2+\mu)(0.4u_n + 0.3v_n), \end{aligned} \tag{3.4}$$

as

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} = \frac{1}{2}(2+\mu)^n \begin{bmatrix} \cos(n\theta) & \sin(n\theta) \\ \sin(n\theta) & -\cos(n\theta) \end{bmatrix} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \tag{3.5}$$

where θ is as previously determined in (3.3) as the phase angle of the eigenvalues

$$\tan\theta = \frac{4}{3}.$$

The solution of the IVP for the linearized system is a polygonal path spiraling into the origin of the $u, v-$plane (which approximates the $x, y-$plane for an initial point (X, Y) close to the origin) for $-2 < \mu < 0$ .and spiraling outward away from the origin for $\mu > 0$.

3.3. Bifurcation at $\mu = 0$. It is possible to show that there are no other fixed points for the system (3.1). Otherwise, we have $2 + \mu - r_n^2 \neq 0$ so that

$$\frac{x}{y} = \frac{0.3x - 0.4y}{0.4x + 0.3y}$$

or

$$0.4x^2 + 0.3yx = 0.3xy - 0.4y^2.$$

It follows that $(0, 0)$ is the only fixed point of (3.1). For the case $\mu > 0$, it would appear that (x_n, y_n) should spiral outward toward infinity as $n \to \infty$ and we have a new kind of bifuration (unlike the three previously encountered). While we do have a new kind of bifurcation, there is more to it than what meets the eyes.

3.4. An Equation for r_n. Since $r_n^2 = x_n^2 + y_n^2$ appears in the system (3.1) and r_n^2 increases without bound according to the linearized system (3.4), we may want to see how this quantity varies according to the original nonlinear system (3.1). To see if it is possible to get an equation for r_n^2 (or r_n) alone, we use the two nonlinear difference equations to form $x_{n+1}^2 + y_{n+1}^2 = r_{n+1}^2$ to get

$$\begin{aligned} r_{n+1}^2 &= (2 + \mu - r_n^2)^2 \left\{(0.3x_n - 0.4y_n)^2 + (0.4x_n + 0.3y_n)^2\right\} \\ &= (2 + \mu - r_n^2)^2 \left(0.25\right)(x_n^2 + y_n^2) = \frac{1}{4}(2 + \mu - r_n^2)^2 r_n^2 \end{aligned}$$

or

$$r_{n+1} = \frac{1}{2} r_n (2 + \mu - r_n^2) \equiv F(r_n) \tag{3.6}$$

which is an equation for r_n alone! (This is not always the case but the result of this effort would still be informative but will not be pursued in these notes.) As such, we can apply to (3.6) the knowledge we have learned in the first few chapters on single first order difference equations

Evidently, (3.6) has three fixed points: 0 and $\pm\sqrt{\mu}$ if $\mu > 0$ and only one fixed point $r_n = 0$ if $\mu < 0$. (We will not be concerned with the fixed point $-\sqrt{\mu}$ even when it is real since it r_n is the radial distance of the point (x_n, y_n) from the origin and cannot be negative. For the linear stability of these fixed points, we differentiate $F(r)$ to get

$$F'(r) = \frac{1}{2}(2 + \mu - 3r^2).$$

With $F'(0) = \frac{1}{2}(2 + \mu)$, the fixed point $r = 0$ is asymptotically stable if $-2 < \mu < 0$ and unstable if $\mu > 0$. The result re-affirms what we have found by a linear stability analysis for the fixed point $(0, 0)$ of the original system. By plotting a cobweb graph for (3.6), we have the following more general result for the same fixed point:

PROPOSITION 33. *All polygonal solution trajectories of (3.1) starting at (X, Y) with $0 < \sqrt{X^2 + Y^2} < \sqrt{\mu}$spiral into the origin as $n \to \infty$ if $-2 < \mu < 0$ and away from the origin if $\mu > 0$*

More interesting is the stability of the fixed point $r = \sqrt{\mu}$ of the equation for r for which

$$F'(\sqrt{\mu}) = \frac{1}{2}(2 + \mu - 3\mu) = 1 - \mu.$$

It follows that the fixed point $r = \sqrt{\mu}$ is asymptotically stable for $0 < \mu < 2$ (and unstable for $\mu < 0$). Again, we can get more from the cobweb graph of (3.6):

PROPOSITION 34. *For $0 < \mu < 2$, any solution trajectory starting inside the circle $r = \sqrt{\mu}$ of the $x, y-$plane would spiral toward this circle and not toward infinity!*

So the bifurcation at $\mu = 0$ not only involves a change the stability of the fixed point $(0, 0)$ but also the occurrence of a circle to which solution trajectories approach for the case $0 < \mu < 2$ (with $\mu > 2$ not likely to be biologically relevant). As an exercise, investigate the behavior or trajectories starting at a point on the circle.

CHAPTER 11

Mendelian Genetics

While much more can be said about first (or higher) order Markov chain, results are more difficult to obtain for nonlinear discrete stochastic processes. This chapter starts with one well-known problem of this type, the simplest type of genetics problem, to illustrate the difference of these processes from Markov chains. Some techniques for solving nonlinear difference equation ensue.

Genetics is a branch of biological science that investigates the mechanism for passing physiological traits from one generation to the next. Genetic analysis predates Gregor Mendel, but Mendel introduced a number of innovations to the science of genetics. They enabled him to formulate laws that provide the theoretical basis of our understanding of the genetics of inheritance. Briefly, Mendel concluded that the hereditary determinants are of a particulate nature. These determinants are called **genes**.

To start, we limit our discussion first to the inheritance of physical traits with two **phenotypes:** Either you are an albino or you are not; either the surface of an organ is wrinkled or it is smooth. (Other physical traits such as the color of your eyes have more than two phenotypes.) Each parent has a **gene pair** in each cell for each trait of interest. Each of the two members of the gene pair is called an **Allele.** For our two phenotype genetics, there are two possible forms of an allele (of the same gene-pair), denoted by D (for dominant) and R (for recessive), though A and a are sometimes used instead by some writer. The two alleles from the same gene-pair may be the same or different allele type. So a gene pair can be one of the three **genotypes**: $\mathbf{D} = (D, D)$, $\mathbf{H} = (D, R) = (R, D)$ and $\mathbf{R} = (R, R)$, known as dominant, hybrid and recessive genotype.

Below are some terminology in Mendelian genetics:

- **Allele** is one member of a gene pair with two possible forms: wrinkled and smooth are the alleles for the surface appearance of an organ. (More than two alleles can exist for any specific gene, but only two of them will be found within any individual.)
- **Allelic pair** is the combination of two alleles which comprise the gene pair
- **Homozygote** is an individual which contains only one allele at the allelic pair; for example DD is homozygous dominant and RR is homozygous recessive; pure lines are homozygous for the gene of interest
- **Heterozygote** is an individual which contains one of each member of the gene pair; for example the DR heterozygote
- **Genotype** is the specific allelic combination for a certain gene or set of genes
- An allele of a gene is **dominant** if an organism of genotype $\mathbf{D} = (D, D)$ is indistinguishable from one of the genotype $\mathbf{H} =(D, R) = (R, D)$

- An allele of a gene is **recessive** if an organism of genotype $\mathbf{R} = (R, R)$ appear to be different from one of the genotype $\mathbf{H}$

For example, for the gene controlling sickle cell anemia, an individual with (R, R) pair would show severe anemia while neither a $\mathbf{D}$ (dominant) genotype nor a $\mathbf{H}$ (hybrid) genotype shows such a trait. The genotypes themselves are called *dominant*, *hybrid* and *recessive*, respectively. While the two different variation (wrinkled and smooth) of a particular phusiological characteristic (smoothness of an organ) are called (the two) phenotypes of the particular characteristic.

In the simplest setting, a gene is drawn for the gene pair of a trait from each of two parents, male and female, to form the genotype of the offspring for this trait. Allele frequencies in a population is to be the same across generations; the static allele frequencies effectively assumes: no mutation (the alleles don't change), no migration or emigration (no exchange of alleles between populations), infinitely large population size, and no selective pressure for or against any genotypes. Genotype frequencies is also to be static while mating is random.

Together, the two alleles comprise the gene pair. With each offspring gene pair containing one member of each parent's gene pair in accordance with Mendel's two laws of genetics:

Mendel's First Law - *The Law of Segregation*: For the pair of alleles an offspring has of some gene (or at some genetic locus), one is a copy of a randomly chosen one in the father, and the other is a copy of a randomly chosen one in the mother.

Mendel's Second Law - *The Law of Independent Assortment*: Each allele of a parent's allelic pair has an equal chance to be the one copied for the offspring, and that the copying of alleles to different offspring or from different parents are independent.

(Today, we know that some genes are in fact "linked" and are inherited together, but for the most part Mendel's laws have proved surprisingly robust.)

Let $(d, 2h, r)$ be the probability of the offspring be of the dominant, hybrid and recessive genotype, respectively. The distribution of the three probabilities evidently depends of the genotypes of the two parents. Below is a table of distributions for the different combinations of parent genotypes with the explanations given in bullets to follow:

$m \backslash f$	(D, D)	(D, R)	(R, R)
(D, D)	$(1.0, 0)$	$(\frac{1}{2}, \frac{1}{2}, 0)$	$(0, 1, 0)$
(D, R)	$(\frac{1}{2}, \frac{1}{2}, 0)$	$(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	$(0, \frac{1}{2}, \frac{1}{2})$
(R, R)	$(0, 1, 0)$	$(0, \frac{1}{2}, \frac{1}{2})$	$(0, 0, 1)$

(??)

- If the genotypes of both parents are dominant, denoted by $\mathbf{D}_m$ and $\mathbf{R}_f$ (with the subscript m and f indicating male and female, respectively), it is certain that the genotype of the offspring will be dominant and hence $(d, 2h, r) = (1, 0, 0)$.
- Similarly the genotype of the offspring will be recessive with $(d, 2h, r) = (0, 0, 1)$, if the genotypes of the parents are $\mathbf{R}_m$ and $\mathbf{R}_f$.
- If the genotypes of the two parents are i) $\mathbf{D}_m$ and $\mathbf{H}_f$, respectively, or ii) $\mathbf{D}_f$ and $\mathbf{H}_m$, respectively, then the offspring genotype would be $\mathbf{D}$ (with

probability $1{\cdot}\frac{1}{2} = \frac{1}{2}$) or **H** (with probability $1{\cdot}\frac{1}{2} = \frac{1}{2}$), corresponding to the entry $(\frac{1}{2}, \frac{1}{2}, 0)$ in the first super- and sub-diagonal position of the matrix (**??**).

- The offspring genotype from parents of genotypes iii) $\mathbf{R}_m$ and $\mathbf{H}_f$, respectively, or ii) $\mathbf{R}_f$ and $\mathbf{H}_m$, respectively, are similarly **R** and **H**, both with probability $1{\cdot}\frac{1}{2} = .\frac{1}{2}$, giving the entry $(0, \frac{1}{2}, \frac{1}{2})$ in the last super- and sub-diagonal position of the table (**??**)..

This leaves the complicated case of both parents being of the hybrid genotype, $\mathbf{H}_m$ and $\mathbf{H}_f$, each containing one allele for the dominant phenotype and one for the recessive phenotype. To get a dominant genotype offspring, we need a dominant allele from each parent. Each of these occurs with a probability of $\frac{1}{2}$ resulting in a probability of $r = \frac{1}{4}$ the event of a dominant offspring. Similarly, the probability of a recessive genotype offspring is also $d = \frac{1}{4}$. For a hybrid offspring, we can get it in two ways: a dominant allele from the "male" parent and a recessive allele from the "female" parent and the mirror image of this construction. Each of these two elementary events is with a probability of $\frac{1}{4}$ totaling to a probability of $2h = \frac{1}{2}$ for a hybrid genotype offspring. Altogether, the probability distribution for the genotype offspring of a pair of hybrid (genotype) parents is $(d, 2h, r)^T = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})^T$).as shown in the center box of the table above.

1. Hardy-Weinberg Stability Theorem

In evolutionary biology, we are interested more than just the next generation offsprings but multi-generational evolution of the genotypes of offsprings. Starting with an initial distribution of $\boldsymbol{\delta}_0 = (d_0, 2h_0, r_0)^T$, the genotype of subsequent generations is **not** determined by a Markov Chain, $\boldsymbol{\delta}_{n+1} \neq M\boldsymbol{\delta}_n$, where we now use a subscripted variable, x_n, instead of the previous form of $x(n)$. For example, a dominant offspring in the $(n+1)^{th}$ generation given the n^{th} generation probability distribution vector $\boldsymbol{\delta}_n = (d_n, 2h_n, r_n)^T$ can come from combinations of dominant or hybrid parents. The probability of a dominant allele from the male parent of the n^{th} generation is $1 \cdot d_n + \frac{1}{2} \cdot (2h_n)$; the same is true for the female parent. Together, they give the probability of $d_{n+1} = (d_n + h_n)^2$ for a *dominant* genotype offspring. Similarly, the probability of a *recessive* genotype offspring is $r_{n+1} = (r_n + h_n)^2$. On the other hand, we can get a hybrid offspring in two ways. One is to get a dominant allele from the male parent with probability $(d_n + h_n)$ and a recessive allele from the female parent with a probability $(r_n + h_n)$ so that the probability of a hybrid genotype offspring from this combination is $(d_n + h_n)(r_n + h_n)$. Now, we can also get a hybrid offspring through a dominant allele from the female parent and a recessive allele from the male parent with the same probability so that.$2h_{n+1} = 2(d_n + h_n)(r_n + h_n)$. These observations are summarized as the following systems of three *difference equations*:

$$\begin{aligned} d_{n+1} &= (d_n + h_n)^2, \\ 2h_{n+1} &= 2(d_n + h_n)(r_n + h_n), \\ r_{n+1} &= (r_n + h_n)^2. \end{aligned} \tag{1.1}$$

Unlike Markov chains, the difference equations that govern the probability distribution $\boldsymbol{\delta}_n = (d_n, 2h_n, r_n)^T$ are not linear and solutions of the form $c\lambda^n$ is generally not applicable. In fact, there are much less general methods for analyzing the

solution of nonlinear difference equations (even less than the corresponding linear ODE).

For the present nonlinear system (1.1), we note that of solution for

$$\begin{aligned} d_1 &= (d_0 + h_0)^2, \\ 2h_1 &= 2(d_0 + h_0)(r_0 + h_0), \\ r_1 &= (r_0 + h_0)^2. \end{aligned} \tag{1.2}$$

and

$$\begin{aligned} d_2 &= (d_1 + h_1)^2 = \left[(d_0 + h_0)^2 + (d_0 + h_0)(r_0 + h_0)\right]^2 \\ &= (d_0 + h_0)^2 \left[(d_0 + h_0) + (r_0 + h_0)\right]^2 = (d_0 + h_0)^2. \end{aligned} \tag{1.3}$$

Similarly, we have

$$\begin{aligned} 2h_2 &= 2(d_1 + h_1)(r_1 + h_1) \\ &= 2\left[(d_0 + h_0)^2 + (d_0 + h_0)(r_0 + h_0)\right]\left[(r_0 + h_0)^2 + (d_0 + h_0)(r_0 + h_0)\right]^2 \\ &= 2(d_0 + h_0)(r_0 + h_0), \end{aligned} \tag{1.4}$$

and

$$\begin{aligned} r_2 &= (r_1 + h_1)^2 = \left[(r_0 + h_0)^2 + (d_0 + h_0)(r_0 + h_0)\right]^2 \\ &= (r_0 + h_0)^2 \left[(d_0 + h_0) + (r_0 + h_0)\right]^2 = (r_0 + h_0)^2. \end{aligned} \tag{1.5}$$

Upon repeating the calculations for $\boldsymbol{\delta}_3$, $\boldsymbol{\delta}_4$, $\cdots$, we have the following celebrated Hardy-Weinberg stability theorem in (two allele -) Mendelian genetics:

THEOREM 20. *Given the initial probability distribution* $\boldsymbol{\delta}_0 = (d_0, 2h_0, r_0)^T$, *the subsequent probability distribution* $\boldsymbol{\delta}_n$ *is invariant after one generation with*

$$d_n = (d_0 + h_0)^2, \quad 2h_n = 2(d_0 + h_0)(r_0 + h_0), \quad r_n = (r_0 + h_0)^2 \quad (n \geq 1).$$

PROOF. (by induction) □

In the language of difference equations, the probability distribution of genotypes evolves into a steady state. In the case of a *regular* Markov chain for which, starting with an initial distribution that is not the steady state, the latter is approached through a converging process and reached only in the limit. For the present simple Mendelian model of genetic evolution, the equilibrium configuration is reached in two generations and does not change thereafter. The development and attainment of an equilibrium genotype distribution in this model is remarkably rapid and its implication is of greatest significance. It is a very much consistent with the physical traits in a population being very stable. However, if it were completely stable, there would be no changes in physical traits, and there would be no evolution.

Fortunately, the Hardy-Weinberg equilibrium distribution is, in the language of dynamical systems, stable but not asymptotically stable. Suppose at some stage N (> 1), there is a perturbation from the equilibrium distribution so that we have $\boldsymbol{\delta}_N^* = (d_N^*, 2h_N^*, r_N^*)^T$ instead of $\boldsymbol{\delta}_N = \left((d_0 + h_0)^2, 2(d_0 + h_0)(r_0 + h_0), (r_0 + h_0)^2\right)^T$.

By the Hardy-Weinberg stability theorem, the genotype probability distributions for all future generations would be

$$\boldsymbol{\delta}^*_{N+k} = \left(\left(d^*_N + h^*_N\right)^2, 2\left(d^*_N + h^*_N\right)\left(r^*_N + h^*_N\right), \left(d^*_N + h^*_N\right)^2 \right)^T \qquad (k \geq 1).$$

In other words, the genotype distribution quickly reaches another steady state configuration. If $\boldsymbol{\delta}^*_N$ is close to $\boldsymbol{\delta}_N$, then the new equilibrium configuration $\boldsymbol{\delta}^*_{n+k}$ would be close to $\boldsymbol{\delta}_{N+k} = \boldsymbol{\delta}_N$. In that sense, the Hardy-Weinberg equilibrium configuration is stable but not asymptotically stable; the perturbed genotype distribution does not evolve and return to the equilibrium configuration before perturbation. Thus, Hardy-Weinberg is compatible with the view that evolution is a process for physical traits to change from some existing state apparently with a high degree of stability.

To see what may be responsible for the observed evolution, it is important to make explicit the assumptions in the simple Mendelian model of genetics that led to the Hardy-Weinberg law. These include

- The controlling genes have only two trait alleles
- The genotypes
- The population is bisexual with the same distribution of genotypes in both
- Generations are discrete
- A pair of male and female parents is selected in random in each generation to produce an offspring
- The offspring genotype is determined by an allele from a randomly selected gene from each parent.

Clearly, Hardy-Weinberg law may not apply when anyone of these assumptions is violated. Additional biological processes that are implicitly excluded from the model that led to the Hardy-Weinberg law include

- mutation
- nonrandom mating (inbreeding, selective breeding, assortative mating, etc.)
- natural selection
- gene flow
- genetic drift

While some of these exclusions are consequences of the assumptions listed earlier, they are mentioned explicitly because of their importance as biological processes that may lead to evolutionary changes. We examine a few of these in some later sections.

2. Selective Breeding

Instead of random mating, suppose only the dominant genotype of one parent is allowed to breed. For example, a plant flower consists of both the male (with pollens corresponding to sperms) and female (center of the stamen) parts. A honey bee usually does the transfer of pollens to complete the cycle. As such, flowers are said to self-fertilize. A flower grower may retain only pollens of genotype for more brilliant flower colors for the bee to transfer. With male parent gene to have both dominant alleles, the sample space for the offspring genotype consists of only two elementary events $\{(D_m, D_f), (D_m, R_f)\}$ with a recessive genotype offspring being an impossibility (and hence $r_n = 0$ for all $n > 0$). The probability d_{n+1} of a

dominant offspring genotype is then $1 \cdot (d_n + h_n)$ and the probability $2h_{n+1}$ of a *hybrid* offspring is $1 \cdot (r_n + h_n)$ with $r_{n+1} = 0 \cdot (r_n + \ h_n) = 0$ for $n > 0$.

For $n = 0$, we have

$$d_1 = (d_0 + \ h_0)\,, \qquad 2h_1 = (r_0 + \ h_0)\,, \qquad r_1 = 0. \tag{2.1}$$

with no restriction on the known (prescribed) initial distribution $\boldsymbol{\delta}_0 = (d_0, 2h_0, r_0)^T$ other than that it be a probability vector. For $n > 1$, we have

$$d_2 = (d_1 + \ h_1) = \left[(d_0 + \ h_0) + \frac{1}{2}\,(r_0 + \ h_0)\right] = 1 - \frac{1}{2}\,(r_0 + \ h_0)\,, \tag{2.2}$$

$$2h_2 = (r_1 + \ h_1) = \frac{1}{2}\,(r_0 + \ h_0)\,, \qquad r_2 = 0, \tag{2.3}$$

and

$$d_3 = (d_2 + \ h_2) = 1 - \frac{1}{4}\,(r_0 + \ h_0)\,, \tag{2.4}$$

$$2h_3 = (r_2 + \ h_2) = \frac{1}{4}\,(r_0 + \ h_0)\,, \qquad r_3 = 0, \tag{2.5}$$

etc. By induction, we get for $n > 0$ the following result:

PROPOSITION 35. *For selective breeding with the gene of one breeding parent to have two dominant alleles, the components of the genotype distribution at the* $(n+1)^{th}$ *stage is given by*

$$d_{n+1} = (d_n + \ h_n) = 1 - \frac{1}{2^n}\,(r_0 + \ h_0)\,, \tag{2.6}$$

$$2h_{n+1} = (r_n + \ h_n) = \frac{1}{2^n}\,(r_0 + \ h_0)\,, \qquad r_{n+1} = 0. \tag{2.7}$$

PROOF. (by induction) □

In the limit as $n \to \infty$, we get

$$\boldsymbol{\delta}_n \to (1, 0, 0)^T$$

as we would expect (when there no other changes in the genetic environment).

It is worth mentioning that the relations (2.6) and (2.7) constitute a set of linear difference equations and hence amenable to an explicit solution of the form $c\lambda^n$with the constant λ to be determined by the method Section 3 of Chapter 1. Equation (2.6) and the first equation of (2.7) with $r_n = 0$ (by the second equation in (2.7)) may be written in matrix form,

$$\boldsymbol{\delta}_{n+1} = \begin{pmatrix} d_{n+1} \\ 2h_{n+1} \end{pmatrix} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{pmatrix} d_n \\ 2h_n \end{pmatrix}, \qquad (n \geq 1) \tag{2.8}$$

with the initial conditions (2.1)

$$\boldsymbol{\delta}_1 = \begin{pmatrix} d_1 \\ 2h_1 \end{pmatrix} = \begin{pmatrix} d_0 + h_0 \\ r_0 + h_0 \end{pmatrix}. \tag{2.9}$$

A solution proportional to λ^n,i.e., $\boldsymbol{\delta}_n = \mathbf{x}\lambda^n$, is possible. The linear system of two difference equations requires the constant λ to be a root of the quadratic equation

$$2\lambda^2 - 3\lambda + 1 = 0,$$

namely $\lambda_1 = 1$ and $\lambda_2 = \frac{1}{2}$. Superposition of the two linearly independent solutions corresponding to the two roots and the auxiliary conditions at $n = 0$ give

$$\begin{aligned} d_n &= d_0 + 2h_0\left(1 - \frac{1}{2^{n+1}}\right) + r_0\left(1 - \frac{1}{2^n}\right) \\ 2h_{n+1} &= \frac{1}{2^n}(r_0 + h_0) \end{aligned}$$

the same as previously obtained.

3. Gene Frequencies

Let

$$p_n = d_n + h_n, \qquad q_n = h_n + r_n. \tag{3.1}$$

Evidently, p_n and q_n are, respectively, the frequency of the dominant gene and recessive gene. For the two allele Mendelian model in the first section of this chapter, the evolution of genotype distribution may be rewritten as

$$d_{n+1} = p_n^2, \qquad 2h_{n+1} = 2p_nq_n, \qquad r_{n+1} = q_n^2.$$

It follows that

$$p_{n+1} = p_n^2 + p_nq_n = p_n, \qquad q_{n+1} = q_n^2 + p_nq_n = q_n,$$

for $n = 0, 1, 2, ...$ with $p_{n+1} + q_{n+1} = p_n + q_n = ... = p_0 + q_0 = 1$. Whatever the initial gene frequency distribution, it remains the same thereafter.

PROPOSITION 36. *In the two allele Mendelian model of population genetics, gene frequency distribution is conserved.*

To illustrate the necessary care needed in the delineation of gene frequencies, we consider the following problem of selective breeding that is the opposite of the one discussed in the previous section. In the new problem, those individuals of recessive genotype do not participate in the reproductive process. For example, they may expire prior to reproductive age or may simply be prohibited from participating in reproduction.

Suppose we start with a genotype distribution of $\boldsymbol{\delta}_0 = (d_0, 2h_0, r_0)$. By the Mendelian model without the participation of the recessive genotype in reproduction, we now have the following genotype distribution for the next generation (instead of (1.1) for $n = 1$):

$$d_1 = (d_0 + h_0)^2, \qquad 2h_1 = 2(d_0 + h_0)h_0, \qquad r_1 = h_0^2. \tag{3.2}$$

Before we proceed to calculate the genotype distribution of the next generation, it is important to note that

$$\begin{aligned} d_1 + 2h_1 + r_1 &= (d_0 + h_0)^2 + 2(d_0 + h_0)h_0 + h_0^2, \\ &= (d_0 + 2h_0)^2 = (1 - r_0)^2. \end{aligned}$$

Evidently, given that not the entire gene pool is allowed to participate in reproduction, the parts of the pool allowed to participate do not add up to the whole. To

focus on the part of the pool allowed to participate in reproduction as the whole pool for the reproduction of the next generation, we set .

$$\boldsymbol{\delta}_0^* = (d_0^*, 2h_0^*, r_0^*) \ = \left(\frac{d_0}{1-r_0}, \frac{2h_0}{1-r_0}, 0\right),$$

with the frequency distribution taken to be that for the genes allowed to participate in reproduction

$$p_0 = d_0^* + h_0^* = \frac{d_0 + h_0}{1-r_0}, \qquad q_0 = r_0^* + h_0^* = \frac{h_0}{1-r_0} \tag{3.3}$$

and

$$p_0 + q_0 = \frac{d_0 + 2h_0}{1-r_0} = 1.$$

With recessive genotype individuals not participating in the reproductive process, the next generation's genotype distribution is appropriately given by

$$\begin{aligned} d_1 &= (d_0^* + h_0^*)^2 = p_0^2, \qquad r_1 = (r_0^* + h_0^*)^2 = q_0^2, \\ 2h_1 &= 2(d_0^* + h_0^*)(r_0^* + h_0^*) = 2p_0q_0, \end{aligned}$$

with

$$\begin{aligned} d_1 + 2h_1 + r_1 &= \left(\frac{d_0 + h_0}{1-r_0}\right)^2 + 2\left(\frac{d_0 + h_0}{1-r_0}\right)\frac{h_0}{1-r_0} + \left(\frac{h_0}{1-r_0}\right)^2 \\ &= .\left(\frac{d_0 + 2h_0}{1-r_0}\right)^2 = 1. \end{aligned} \tag{3.4}$$

For the n^{th} generation, we again take the gene frequency to be that of the genes allowed to participate:

$$p_n = \frac{d_n + h_n}{1-r_n}, \qquad q_n = \frac{h_n}{1-r_n}.$$

with

$$d_{n+1} = p_n^2, \qquad 2h_{n+1} = 2p_nq_n, \qquad r_{n+1} = q_n^2.$$

LEMMA 7. $\qquad d_{n+1} + 2h_{n+1} + r_{n+1} = 1 \qquad (n \geq 0)$

PROOF. The initial distribution is a probability vector so that $d_0 + 2h_0 + r_0 = 1$ and consequently $d_1 + 2h_1 + r_1 = 1$ by (3.4). Suppose $d_k + 2h_k + r_k = 1$ holds for $k = n$; then we have

$$\begin{aligned} d_{n+1} + 2h_{n+1} + r_{n+1} &= p_n^2 + \ 2p_nq_n + q_n^2 = (p_n + \ q_n)^2 \\ &= \left(\frac{d_n + 2h_n}{1-r_n}\right)^2 = \left(\frac{1-r_n}{1-r_n}\right)^2 = 1. \end{aligned}$$

□

PROPOSITION 37. *The evolution of gene frequencies for $n > 0$ is given by*

$$q_{n+1} = \frac{q_n}{1+q_n}, \qquad p_{n+1} = \frac{1}{1+q_n}. \tag{3.5}$$

PROOF.

$$\begin{aligned} p_{n+1} &= \frac{d_{n+1}+h_{n+1}}{1-r_{n+1}} = \frac{p_n^2+p_nq_n}{1-q_n^2} = \frac{p_n}{1-q_n^2} = \frac{1}{1+q_n}, \\ q_{n+1} &= \frac{h_{n+1}}{1-r_{n+1}} = \frac{q_n^2+p_nq_n}{1-q_n^2} = \frac{p_nq_n}{1-q_n^2} = \frac{q_n}{1+q_n}. \end{aligned}$$

□

The first order difference equation for q_n in (3.5) and the initial condition for q_0 in (3.3) define a nonlinear IVP. Its solution is obtained by re-arranging the difference equation into a linear difference equation:

$$q_{n+1} = \frac{q_n}{1+q_n} = \frac{1}{1+q_n^{-1}}$$

or

$$x_{n+1} = 1 + x_n,$$

where $x_n = 1/q_n$ and $x_0 = 1/q_0 = (1-r_0)/h_0$. The solution for IVP for x_n is

$$x_n = n + x_0 = \frac{nh_0 + (1-r_0)}{h_0}.$$

From the calculations above, we have the following result for the evolution of the gene frequency distribution:

PROPOSITION 38. *The frequency of the recessive gene pool decreases to zero as* $n \to \infty$ *with*

$$q_n = \frac{1}{n+x_0} = \frac{h_0}{nh_0+(1-r_0)} \sim \frac{1}{n}.$$

Correspondingly, the frequency of the dominant gene pool increases slowly toward unity:

$$p_n = \frac{n+x_0-1}{n+x_0} = \frac{(n-1)h_0+(1-r_0)}{nh_0+(1-r_0)} \sim 1 + O(\frac{1}{n}).$$

4. Mutation

Under the appropriate idealized conditions, we were led to the Hardy Weinberg law of Section 2 which predicts genetic stability for the population after a generation, The theoretical prediction turns out to be quite consistent with the observed persistency in the heredity of traits. However, changes do occurs naturally, albeit very infrequently and/or very slowly. This happens even under selecting breeding of Section 6 where only the dominant genotype of one parent is allowed to breed; a recessive genotype occurs on rare occasions, when the modeling result predicts that it should not. We simply call this observed process a gene **mutation** without getting into the biological details of how it takes place.

For a simple modeling of the phenomenon of mutation, we consider the situation that whenever a dominant gene is transmitted, there is a small probability α $(0 < \alpha \ll 1)$ that the gene will mutate to a recessive gene. We suppose that the selection occurs after selection of the dominant gene from a parent. Otherwise, we retain all the hypotheses of the Mendelian (*panmixia*) model. In that case, the Mendelian model (1.1) governing the evolution of genotypes is modified by a

reduction of the dominant gene frequency and an increase in the recessive gene frequency. The modification is most simply done by working with dominant and recessive gene frequencies, p_n and q_n, introduced in the last section.

The dominant offspring genotype probability d_{n+1} is the product of the (available) dominant gene frequency from the two parents both now reduced to $(1-\alpha)p_n$ by a loss αp_n due to mutation

$$d_{n+1} = (1-\alpha)p_n^2. \tag{4.1}$$

The recessive gene frequency available for the offsprings is enhanced by mutation from q_n to $\alpha p_n + q_n$ thereby giving

$$2h_{n+1} = 2(1-\alpha)p_n(\alpha p_n + q_n) \tag{4.2}$$

$$r_{n+1} = .(\alpha p_n + q_n)^2 \tag{4.3}$$

Note that

$$d_{n+1} + 2h_{n+1} + r_{n+1} = (p_n + q_n)^2 = 1$$

so that gene frequency is conserved.

Instead of solving for the genotype distribution, we form

$$p_{n+1} = d_{n+1} + h_{n+1} = (1-\alpha)p_n\left[(1-\alpha)p_n + (\alpha p_n + q_n)\right] = (1-\alpha)p_n.$$

It follows immediately that

$$p_n = p_0(1-\alpha)^n$$

where p_0 is the initial dominant gene frequency. Correspondingly, the recessive gene frequency is

$$\begin{aligned} \alpha p_n + q_n &= \alpha p_n + (1-p_n) = 1 - (1-\alpha)p_n \\ &= 1 - p_0(1-\alpha)^{n+1} \equiv 1 - \pi_n. \end{aligned}$$

In terms of $\pi_n = p_0(1-\alpha)^{n+1}$, we have

$$d_{n+1} = \pi_n^2, \qquad 2h_{n+1} = 2\pi_n(1-\pi_n), \qquad r_{n+1} = (1-\pi_n)^2.$$

When there is no mutation so that $\alpha = 0$ and therewith $\pi_n = p_0$, the above results reduce to the Hardy-Weinberg scenario with

$$d_{n+1} = p_0^2, \qquad 2h_{n+1} = 2p_0q_0, \qquad r_{n+1} = q_0^2.$$

For $0 < \alpha < 1$, we have $\pi_n = p_0(1-\alpha)^{n+1} \to 0$ so that

$$d_{n+1} \to 0, \qquad 2h_{n+1} \to 0, \qquad r_{n+1} \to 1.$$

if the particular type of mutation is the only evolutionary process at work (which fortunately is not).

Part 3

Appendices

CHAPTER 12

Linear Equations - Algebra

1. Linear Equations

Systems of linear equations play a very significant role in much of science and engineering. They play an equally important role in difference (and differential) equations. In this section, we discuss briefly the solution of such systems facilitated by the use of matrix notations.

1.1. Cramer's Rule. Suppose we want solve for the unknowns x_1 and x_2 in the linear system

$$a_{11}x_1 + a_{12}x_2 = b_1, \qquad a_{21}x_1 + a_{22}x_2 = b_2, .$$

with known coefficients $\{a_{ij}\}$ and right hand members $\{b_k\}$. This is easily done by using one equation to eliminate an unknown from the other equation to get

$$Dx_j = D_j, \tag{1.1}$$

$j = 1, 2$, where the number

$$D = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}, \tag{1.2}$$

is the *determinant* of the coefficients $\{a_{ij}\}$ of the linear system, and where the other two determinants involving the right hand members are

$$D_1 = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} = b_1a_{22} - b_2a_{12}, \qquad D_2 = \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix} = b_2a_{11} - b_1a_{21}. \tag{1.3}$$

For linear systems with more than two equations, the same method of elimination can also be used to obtain the unknowns in terms of similar determinants in the form (1.1) where now $j = 1, 2, ..., n$ where n is number of equations (which is the same as the number of unknowns). For the three equations

$$\begin{aligned} 3x_1 + 2x_2 + 2x_3 &= 3 \\ x_1 + 4x_2 + x_3 &= 3 \\ 2x_1 + 4x_2 + x_3 &= 3 \end{aligned} \tag{1.4}$$

for example, the relevant determinants in the Cramer's Rule (1.1) are

$$D = \det \begin{bmatrix} 3 & 2 & 2 \\ 1 & 4 & 1 \\ 2 & 4 & 1 \end{bmatrix} = \begin{vmatrix} 3 & 2 & 2 \\ 1 & 4 & 1 \\ 2 & 4 & 1 \end{vmatrix}, \quad D_1 = \det \begin{vmatrix} 1 & 2 & 2 \\ 1 & 4 & 1 \\ 1 & 4 & 1 \end{vmatrix} = \cdots, \tag{1.5}$$

$$D_1 = \cdots = \begin{vmatrix} 3 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 1 \end{vmatrix}, \qquad D_3 = . \det \begin{bmatrix} 3 & 2 & 1 \\ 1 & 4 & 1 \\ 2 & 4 & 1 \end{bmatrix} = \begin{vmatrix} 3 & 2 & 1 \\ 1 & 4 & 1 \\ 2 & 4 & 1 \end{vmatrix}. \tag{1.6}$$

In general, the numerical value of the determinant

$$|\begin{smallmatrix}\cdots\\ \cdots\end{smallmatrix}| = \det \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot \\ a_{21} & \cdot & \cdot & \\ \cdot & \cdot & & \\ \cdot & & & \end{bmatrix}$$

of an $n \times n$ array of numbers $\{a_{ij}\}$ is obtained by the rule of expansion in *cofactors..*

DEFINITION 18. *The* ***cofactor*** A_{ij} *of the entry (or element)* a_{ij} *in the location of the* i^{th} *row and* j^{th} *column of the determinant* $D = |\begin{smallmatrix}\cdots\\ \cdots\end{smallmatrix}|$ *is defined to be*

$$A_{ij} = (-1)^{i+j} M_{ij} \tag{1.7}$$

where .M_{ij} *is (the original determinant)* D *with the* i^{th} *row and* j^{th} *column removed.*

With determinant of a single element $|a_{22}|$ being the element a_{22} itself, the *rule of cofactor* for an $n \times n$ determinant is

$$D = \sum_{k=1}^{n} a_{ik} A_{ik} = \sum_{k=1}^{n} a_{ki} A_{ki} \tag{1.8}$$

For the 2×2 determinant D in (1.2), we have from the rule of cofactors:

$$A_{11} = (-1)^{1+1} |a_{22}| = a_{22}, \qquad A_{12} = (-1)^{1+2} |a_{21}| = -a_{21}$$

so that

$$D = a_{11} a_{22} + a_{12} (-a_{21})$$

For the 3×3 determinant (1.5), we have

$$A_{11} = (-1)^{1+1} \begin{vmatrix} 4 & 1 \\ 4 & 1 \end{vmatrix}, \qquad A_{12} = (-1)^{1+2} \begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix}, \qquad A_{13} = (-1)^{1+3} \begin{vmatrix} 1 & 4 \\ 2 & 4 \end{vmatrix},$$

so that

$$D = 3(0) + 2(1) + 2(-4) = -6$$

We may also use

$$A_{22} = (-1)^{2+2} \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix}, \qquad A_{32} = (-1)^{2+3} \begin{vmatrix} 3 & 2 \\ 1 & 1 \end{vmatrix},$$

so that

$$D = 2(1) + 4(-1) + 4(-1) = -6.$$

Other combinations of cofactors for (1.8) are also possible for the same result.

A proof of Cramer's as a formula for the solution of a linear system can be found in most older text on linear systems or on internet <http://en.wikipedia.org/wiki/Cramer's_rule#Proof>. Most proofs are based on a few simple properties of determinants. A subsequent subsection discusses some relevant properties and their relations to Cramer's formula (1.8). These properties and other aspects of linear systems of equations (such as Gaussian elimination) are more efficiently and effectively described with the help of matrix notations. The next subsection introduces the concept and algebra of matrices with significant impact and applications well beyond linear systems of algebraic equations.

1.2. Determinants (optional). The following properties are intrinsic to determinants as introduced in the Cramer's rule for linear systems of equations. Their validity will be demonstrated for either the 2×2 case or the 3×3 case but is easily extended to the $n \times n$ case.

1.2.1. *Property (1) Linearity With Respect to a Column:*

$$(1.9) \qquad (a) \qquad \begin{vmatrix} b_1+c_1 & a_{12} & a_{13} \\ b_2+c_2 & a_{22} & a_{23} \\ b_3+c_3 & a_{23} & a_{33} \end{vmatrix} = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{23} & a_{33} \end{vmatrix} + \begin{vmatrix} c_1 & a_{12} & a_{13} \\ c_2 & a_{22} & a_{23} \\ c_3 & a_{23} & a_{33} \end{vmatrix}$$

The left hand side of (1.9) is the solution $Dx_1^{(b+c)}$ of the linear system

$$(1.10) \qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= r_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= r_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= r_3 \end{aligned}$$

with right hand members $\{r_1 = b_1 + c_1, r_2 = b_2 + c_2, r_3 = b_3 + c_3\}$ while the right hand side of (1.9) is the sum of the solutions $Dx_1^{(b)}$ and $Dx_1^{(c)}$ of the linear system with right hand member $\{b_1, b_2, b_3\}$ and $\{c_1, c_2, c_3\}$, respectively. The two sides must be equal given

$$\begin{aligned} Dx_1^{(b+c)} &= (b_1+c_1)A_{11} + (b_2+c_2)A_{21} + (b_3+c_3)A_{31} \\ &= (b_1A_{11} + b_2A_{21} + b_3A_{31}) + (c_1A_{11} + c_2A_{21} + c_3A_{31}) \\ &= \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{23} & a_{33} \end{vmatrix} + \begin{vmatrix} c_1 & a_{12} & a_{13} \\ c_2 & a_{22} & a_{23} \\ c_3 & a_{23} & a_{33} \end{vmatrix} = Dx_1^{(b)} + Dx_1^{(c)} \end{aligned}$$

$$(1.11) \qquad (b) \qquad \begin{vmatrix} \alpha b_1 & a_{12} & a_{13} \\ \alpha b_2 & a_{22} & a_{23} \\ \alpha b_3 & a_{23} & a_{33} \end{vmatrix} = \alpha \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{23} & a_{33} \end{vmatrix}$$

The left hand side is the solution $Dx_1^{(\alpha)}$ of (1.11) with $(r_1, r_2, r_3) = (ab_1, ab_2, ab_3)$. By

$$\begin{aligned} Dx_1^{(\alpha)} &= (\alpha b_1)A_{11} + (\alpha b_2)A_{21} + (\alpha b_3)A_{31} \\ &= \alpha\left(b_1A_{11} + b_2A_{21} + b_3A_{31}\right) \\ &= \alpha \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{23} & a_{33} \end{vmatrix} = \alpha Dx_1^{(1)}, \end{aligned}$$

it must be equal to $\alpha\left[Dx_1^{(1)}\right]$ where $x_1^{(1)}$ is the solution of the same system with right hand side (b_1, b_2, b_3).

1.2.2. *Property (2) Adjacent Column Exchange ⇔ Sign Change of the Determinant:* This property holds for the 2×2 case since

$$\begin{vmatrix} a_{12} & a_{11} \\ a_{22} & a_{21} \end{vmatrix} = a_{12}a_{21} - a_{11}a_{22} = -\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}.$$

We illustrate the argument for the general $n \times n$ case by working with a 3×3 system. Suppose we exchange the first and second column of D corresponding to re-writing the system $Ax = b$ as

$$A' \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix} = \begin{bmatrix} a_{12} & a_{11} & a_{13} \\ a_{22} & a_{21} & a_{23} \\ a_{32} & a_{31} & a_{33} \end{bmatrix} \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix} = \mathbf{b}$$

For the re-arranged system, we have

$$\det[A']x_1 = \begin{vmatrix} a_{12} & b_1 & a_{13} \\ a_{22} & b_2 & a_{23} \\ a_{32} & b_3 & a_{33} \end{vmatrix} \equiv D_1'$$

This must be the same as

$$\det[A]x_1 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{23} & a_{33} \end{vmatrix} = D_1$$

requiring either $\det[A'] = -\det[A]$ **and** $D_1' = -D_1$ or no sign change for both. But we must have a sign change for the 2×2 case; a sign change for the 3×3 case follows from expansion in terms of the cofactors of the third column.

1.2.3. *Property (3) If D has two equal columns, then $D = 0$:* By exchange the two equal columns of D to get D', then $D' = -D$.by Property (2). But $D = D'$; hence D must vanish as seen from the 2×2 case:

$$\det \begin{bmatrix} a_{11} & a_{11} \\ a_{21} & a_{21} \end{bmatrix} = a_{11}a_{21} - a_{21}a_{11} = 0.$$

Taking as definition of determinants, the three properties above suffice for the purpose of proving Cramer's rule solves the linear system for the general $n \times n$ case.

DEFINITION 19. *The numerical value of a scalar quantity D is the determinant $det[a_{ij}]$ of an $n \times n$ matrix $[a_{ij}]$ if it satisfies properties (1) - (3) above.*

THEOREM 21. *With determinants as defined above, Cramer's rule solves a system of n linear equations if the determinant of the system's coefficient matrix is not zero.*

PROOF. See <http://en.wikipedia.org/wiki/Cramer's_rule#Proof>. □

1.2.4. *Property (4) The determinant of the identity matrix is 1, $det[I] = 1$.* Other useful properties of determinants can be found in [**16**] and references therein (see also <http://en.wikipedia.org/wiki/Cramer's_rule#Proof>).

2. Matrices

2.1. Matrix Notations. To motivate these notations, we begin with the familiar physical vectors in three dimensions denoted by bold face letters, $\mathbf{v}$, $\mathbf{w}$, etc., or columns of their respective component coordinates,

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix},$$

etc. The *scalar product* (also called *dot product* or *inner product*) of two vectors, denoted by $\mathbf{v}\cdot\mathbf{w}$, is defined as the sum of the products of their respective coordinate in the same direction,

$$\mathbf{v} \cdot \mathbf{w} = v_1w_1 + v_2w_2 + v_3w_3 = \mathbf{w} \cdot \mathbf{v}.$$

For a more productive development, we adopt the convention for writing the scalar product as

$$\begin{aligned}\mathbf{v}\cdot\mathbf{w} &= \mathbf{v}^T\mathbf{w} = (v_1,v_2,v_3)\begin{pmatrix} w_1\\ w_2\\ w_3\end{pmatrix}\\ &= v_1w_1+v_2w_2+v_3w_3 = \sum_{i=1}^{3} v_iw_i\end{aligned}$$

where

$$\mathbf{u}^T = \begin{pmatrix} u_1\\ u_2\\ u_3\end{pmatrix}^T = (u_1,u_2,u_3)$$

is called the *transpose* of the vector $\mathbf{u}$. Evidently, we have the following property of the matrix description of the dot product:

$$(i)\qquad \mathbf{w}^T\mathbf{v} = (w_1,w_2,w_3)\begin{pmatrix} v_1\\ v_2\\ v_3\end{pmatrix} = \mathbf{w}\cdot\mathbf{v} = \mathbf{v}^T\mathbf{w}. \tag{2.1}$$

With (2.1), we can now write the first equation of the linear system (1.4) as

$$\mathbf{a}_1\cdot\mathbf{x} = \begin{pmatrix} 3\\ 2\\ 2\end{pmatrix}\cdot\begin{pmatrix} x_1\\ x_2\\ x_3\end{pmatrix} = (3,2,2)\begin{pmatrix} x_1\\ x_2\\ x_3\end{pmatrix} = \mathbf{a}_1^T\mathbf{x}$$

and the entire system of three equations as

$$\begin{pmatrix} \mathbf{a}_1\cdot\mathbf{x}\\ \mathbf{a}_2\cdot\mathbf{x}\\ \mathbf{a}_3\cdot\mathbf{x}\end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^T\mathbf{x}\\ \mathbf{a}_2^T\mathbf{x}\\ \mathbf{a}_3^T\mathbf{x}\end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^T\\ \mathbf{a}_2^T\\ \mathbf{a}_3^T\end{pmatrix}\mathbf{x} = \begin{pmatrix} 3\\ 3\\ 3\end{pmatrix}, \tag{2.2}$$

where

$$\mathbf{a}_1^T = (3,2,2),\qquad \mathbf{a}_2^T = (1,4,1),\qquad \mathbf{a}_3^T = (2,4,1),\qquad \mathbf{x}^T = (x_1,x_2,x_3). \tag{2.3}$$

The following properties of transposing follow from the definition of the transpose operation:

$$(ii)\qquad \left(\mathbf{u}^T\right)^T = (u_1,u_2,u_3)^T = \begin{pmatrix} u_1\\ u_2\\ u_3\end{pmatrix}, \tag{2.4}$$

$$(iii)\qquad \left(\mathbf{u}^T\mathbf{v}\right)^T = \left[(u_1,u_2,u_3)^T\begin{pmatrix} v_1\\ v_2\\ v_3\end{pmatrix}\right]^T = (v_1,v_2,v_3)^T\begin{pmatrix} u_1\\ u_2\\ u_3\end{pmatrix} = \mathbf{v}^T\mathbf{u}. \tag{2.5}$$

Regarding property (iii), we note that $\left(\mathbf{u}^T\mathbf{v}\right)^T$ cannot be equal to $(\mathbf{u}^T)^T\mathbf{v}^T = \mathbf{u}\mathbf{v}^T$ since the former is a number while

$$\mathbf{u}\mathbf{v}^T = \begin{pmatrix} u_1\\ u_2\\ u_3\end{pmatrix}(v_1,v_2,v_3) = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3\\ u_2v_1 & u_2v_2 & u_2v_3\\ u_3v_1 & u_3v_2 & u_3v_3\end{bmatrix} \tag{2.6}$$

is a collection of products arranged in the order of appearance of the components of $\mathbf{u}$ and $\mathbf{v}$. The ollection ordered elements on the right hand side of (2.6) is called

a *matrix*. More specifically, it is a 3×3 matrix since it has three rows and three columns.

If we line up the three vector transpose $\mathbf{a}_k^T$ in three rows and denote the resulting square array by A with

$$A = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \mathbf{a}_3^T \end{pmatrix} = \begin{bmatrix} 3 & 2 & 2 \\ 1 & 4 & 1 \\ 2 & 4 & 1 \end{bmatrix}, \tag{2.7}$$

then A is also a 3×3 matrix. In terms of the matrix A, (2.2) may be written compactly as

$$A\mathbf{x} = \mathbf{b}, \tag{2.8}$$

where

$$\mathbf{b} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} .. \tag{2.9}$$

Evidently the preceding development can be extended to the general case of n linear equations for n unknowns $\{x_1, x_2, , ..., \ x_{n-1}, x_n\}$. Specifically, we an write the linear system

$$a_{k1}x_1 + a_{k2}x_2 + \ \cdots\cdots \ + a_{k(n-1)}x_{n-1} \ + a_{kn}x_n = b_k. \qquad (k = 1, 2, ..., n). \tag{2.10}$$

as a vector equation of the form (2.8), now with

$$A = \begin{bmatrix} a_{11} & a_{12} & . & . & . & a_{1n} \\ a_{21} & . & . & . & & . \\ . & . & . & & . & . \\ . & & . & . & & . \\ . & & & . & . & . \\ a_{n1} & a_{n2} & . & . & . & a_{nn} \end{bmatrix}, \tag{2.11}$$

and

$$\mathbf{x} = (x_1, x_2, , ..., x_{n-1}, x_n)^T, \qquad \mathbf{b} = (b_1, b_2, ..., b_{n-1}, b_n)^T. \tag{2.12}$$

With algebraic operations such as equality, sum and dot product of two n-tuple vectors and multiplication of an n-tuple vector by a scalar specified to be the same as those for physical vectors,

$$\begin{aligned} \mathbf{v} &= \mathbf{u} \ <=> \ v_k = u_k \quad (k = 1, 2, ..., n) \\ c_1\mathbf{v} + c_2\mathbf{u} &= (c_1v_1 + c_2u_1, ..., c_1v_n + c_2u_n)^T \\ \mathbf{v}^T\mathbf{u} &= \mathbf{u}^T\mathbf{v} = v_1u_1 + ... + v_nu_n, \end{aligned} \tag{2.13}$$

for $n > 3$, n - tuples such as (2.12) have all the characteristics of a physical vector except with more coordinate components (and hence treated as vectors in higher *dimensions*). We continue to call them vectors in developing the (linear) *algebra* (the study of the rules of operations and relations) of a systems of n linear equations. The field of *linear algebra* is also concerned with the more general case of n equations for m unknowns where m may or may not be equal n.

In writing a linear system of equations as an equation for the unknown vector $\mathbf{x}$, the coefficients of the unknowns in these equations are collected as an array of numbers A. The array is square when the number of unknowns is identical to the

number of equations to be satisfied by them. We call such an array a *matrix*. . Evidently, a vector is an $n \times 1$ (or a column) matrix and its transpose is an $1 \times n$ (or a row) matrix. The multiplication of a vector $\mathbf{x}$ by a matrix $\mathbf{A}$ corresponds to dot products of the rows (or more correctly, the vectors corresponding to the transpose of the different rows) of $\mathbf{A}$ with the vector $\mathbf{x}$ as in (2.2) and (2.8).

Multiplication of a vector by a matrix (as in (2.8)) can be generalized to multiplication of matrices. For example, if A is an $n \times m$ matrices and B is an $m \times k$ matrix, then AB is an $n \times k$ matrix F with its element F_{ij} given by

$$AB = F = [F_{ij}] = \left[\mathbf{a}_i^T \mathbf{b}_j\right] = \left[\sum_{\ell=1}^{m} a_{i\ell} b_{\ell j}\right].$$

where $\mathbf{a}_i^T$ is the i^{th}row of A and $\mathbf{b}_j$ is the j^{th} column of B. Unless $n = k$, the matrix product BA is not meaningful (and is therefore not admissible) because scalar products do not apply to two vectors of different dimensions.

Writing a set of linear equations as a vector equation (2.8) has enabled researchers of the past few centuries to explore and exploit the rich properties of such systems to the benefits of all sciences. The following sections report some of the useful basic findings of their effort and should be mastered by all students of science (including the social sciences) and engineering.

2.2. Matrix Inversion. Now that we have conjure up notations that allow us to write linear systems in the form of $A\mathbf{x} = \mathbf{b}$, similar to $ax = b$ for the one equation and one unknown case, it is natural to want to simply divide both sides by A to get $\mathbf{x} = A^{-1}\mathbf{b}$ (analogous to $x = a^{-1}b$ for the one unknown case). However, A is not a number but a square array of numbers, we need to say what we mean by A^{-1}, called the *inverse* of A, for the case of more than one unknowns. The meaning of A^{-1} follows quite naturally from the desired outcome indicated above. A^{-1} should be a matrix which when it (pre-) multiplies both side of the given vector equation (2.8), the left hand side should become just $\mathbf{x}$ so that

$$A^{-1}A\mathbf{x} = \mathbf{x} = A^{-1}\mathbf{b}.$$

This means

$$A^{-1}A = I$$

where I is the *identity matrix* with all diagonal elements equal to 1 and all others equal to zero, i.e.,

$$I = [\delta_{ij}] = \begin{bmatrix} 1 & 0 & . & . & 0 & 0 \\ 0 & 1 & 0 & . & . & 0 \\ . & 0 & . & . & . & . \\ . & . & . & . & 0 & . \\ 0 & . & . & 0 & 1 & 0 \\ 0 & 0 & . & . & 0 & 1 \end{bmatrix}, \tag{2.14}$$

where the *Kronecker delta* δ_{ij} is defined to be 1 when $i = j$ and 0 otherwise. Evidently, we have

$$I\mathbf{v} = \mathbf{v}, \qquad \mathbf{IA} = \mathbf{AI} = \mathbf{A} \tag{2.15}$$

so that I in matrix theory plays the role of 1 for scalars. We leave as an exercise to show

$$AA^{-1} = A^{-1}A = I. \tag{2.16}$$

The relation (2.16) offers a way to compute the inverse of A. Let $\mathbf{x}^{(k)}$ be the k^{th} column of A^{-1}. Then

$$A\mathbf{x}^{(k)} = \begin{pmatrix} \delta_{1k} \\ . \\ . \\ \delta_{nk} \end{pmatrix}$$

We solve this equation for $k = 1, 2, 3, ..., n$ and obtain all the columns of the matrix A^{-1}. The solution of the linear system (2.8) is then given by the product of the matrix $A^{-1}$and the vector $\mathbf{b}$, the right hand side of (2.8). As such, we have another way of solving the linear equations in addition to Cramer's rule (**??**).

Benefited from the matrix notations, the new approach of solving linear equations with A^{-1} not only is more elegant, but also enable us to deduce more information about linear systems and their solutions. For the latter reason, it has and will continue to have a niche in the theory of linear algebra. However, in an age when linear systems to be solved involve a large number of equations and unknowns (with n as large as a million) and repeatedly for a large number of different vectors $\mathbf{b}$, *efficiency* of any method of solution is paramount. While efficiency is technically measured by the number of operations (principally the number of multiplications and function evaluations) required by an algorithm, known as its *operation count*, the end effect of efficiency is felt in speed, i.e., the time it would take to get the solution of the problem.

At the most primitive level, time is money and we are talking big money in large scale and high performance computing (on laptops or supercomputers). More importantly however, many problems need to be solved in real time. For calculations related to in flight correction of the orbit of a manned spaceship or interactive decision on adaptive surgical operations of a life saving (robotic) surgery, efficiency could mean success or failure, or, in extreme cases, the difference between life and death. For these reasons, numerical linear algebra took center stage as computer became more and more sophisticated in the 20th century and has continues unabated since. From the efficiency viewpoint, both Cramer's rule and computing A^{-1} are usually not used in mathematical software for solving linear equations, e.g., Mathematica, Maple and MatLab. For some problems, they can even be disastrous (see a later section for more specific comparisons of efficiency by operation counts).

3. Gaussian Elimination

3.1. An Example. For a system involving a large number of unknowns, Cramer's rule is known to be highly inefficient (by comparing the operation counts, i.e., the number of multiplications and function evaluations involved, with those of other methods. By that criterion, the method of Gaussian elimination is known to be a more efficient method [**16, 23**]. To develop this more efficient method, let us solve the illustrative example (1.4) as we would do naturally without making use of any formal method described previously. What we would normally do would be to use two of the equations to eliminate two of the unknown leaving us with one equation for one unknown which is easily solved. This approach can be repeated for the other two unknowns as indicated at the start of this Appendix. Suppose we carry out this program more slowly and with a general right hand side $\mathbf{b} = (b_1, b_2, b_3)^T$ (instead of $(3, 3, 3)^T$).

Use the first equation to eliminate the unknown x_1 from the last two equations leaving us with

$$\begin{array}{rrrcl} 3x_1 + & 2x_2 + & 2x_3 & = & b_1 \\ & (4-2\cdot 1/3)x_2 + (1-2\cdot 1/3)x_3 & & = & b_1 - 1/3\cdot b_1 \\ & (4-2\cdot 2/3)x_2 + (1-2\cdot 2/3)x_3 & & = & b_1 - 2/3\cdot b_1 \end{array} \tag{3.1}$$

where we have kept the details of the elimination process. More specifically, in order to eliminate x_1 from the second equation of (1.4), we multiplied the first equation of that original system through by $\ell_{21} = 1/3$ and subtract the resulting equation from the second to get the second equation in (3.1). Similarly, we multiplied the first equation of (1.4) by $\ell_{31} = 2/3$ and subtract the result from the third equation to get the third line of (3.1). Having seen the details, we now clean up (3.1) to get

$$\begin{array}{rrrcl} 3x_1 + & 2x_2 + & 2x_3 & = & b_1 \\ & (10/3)x_2 + (1/3)x_3 & & = & b_2 - \ell_{21}b_1 \\ & (8/3)x_2 - (1/3)x_3 & & = & b_3 - \ell_{31}b_1 \end{array} \tag{3.2}$$

where we have used ℓ_{21} and ℓ_{31} on the right hand side instead of their actual values for this example for reasons that will become clear later.

Now, instead of solving the last two equations for x_2 and x_3 and use the results to eliminate these two unknowns from the first, we simply multiply the second equation in the modified system (3.2) by $\ell_{32} = 4/5$ and use the result to eliminate x_2 from the last equation to get

$$\begin{array}{rrrcl} 3x_1 + & 2x_2 + & 2x_3 & = & b_1 \\ & (10/3)x_2 + (1/3)x_3 & & = & b_2 - \ell_{21}b_1 \\ & (-3/5)x_3 & & = & b_3 - \ell_{32}b_2 - (\ell_{31} - \ell_{32}\ell_{21})b_1 \end{array} \tag{3.3}$$

The modified system (3.3) in the form

$$U\mathbf{x} = \mathbf{c}, \tag{3.4}$$

where the coefficient matrix

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} 3 & 2 & 2 \\ 0 & \frac{10}{3} & \frac{1}{3} \\ 0 & 0 & -\frac{3}{5} \end{bmatrix} \tag{3.5}$$

is *upper triangular*. The process of forward elimination leading to (3.4)-(3.5) is known as *row echelon reduction*. The corresponding modified right hand side is

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}. = \begin{pmatrix} b_1 \\ b_2 - \ell_{21}b_1 \\ b_3 - \ell_{32}b_2 - (\ell_{31} - \ell_{21}\ell_{32})b_1 \end{pmatrix} \tag{3.6}$$

Evidently, the reduced system (3.4)-(3.5) (or what is the same, (3.3)) is easily solved for the unknowns. Starting with the last equation which involved only one unknown, we get $x_3 = c_3/(-3/5)$. With x_3 known, the second equation can then be solved for the only unknown x_2. We repeat the process for the first equation to get x_1. The final result of this process of *backward substitution* with $\mathbf{b} = (3, 3, 3)^T$

is

$$\mathbf{x} = \begin{pmatrix} (c_1 - 2(x2 + x3))/3 \\ 3(c_2 - x_3/3)/10 \\ -3c_3//5 \end{pmatrix} = . \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}$$

where x_3 in the first and second component is known from the third and x_2 in the first component is known from the second.

3.2. Efficiency . The method of forward elimination followed by backsubstitution for the solution of linear systems is commonly known as Gaussian elimination in honor of the German mathematician Carl Friedrich Gauss. Because of its efficiency compared to other direct methods such as computing the inverse or Cramer's rule, it has become the backbone of most mathematical software for solving systems of linear equations using a "*direct*"method (as opposed to "*iterative*" methods which will not be discussed here), possibly with refinements (such as pivoting, partial pivoting, pre-conditioning, etc., not discussed herein) needed for matrices with various pathological idiosyncrasies. Given the importance of linear systems in so many areas of science and engineering, mathematical software are now available to take the drudgeries out of solving such systems. An example of mathematical software available for linear systems is the command "LinearSolve[A,b]" in Mathematica which returns the unique solution, if it exists, as a row vector (a $1 \times n$ matrix, in braces, e.g., $\{0, \frac{1}{2}, 1\}$ for the problem (1.4).

Any claim of efficiency should be established by comparing operation counts of the methods. A discussion of the operation counts for both Gaussian elimination and computing A^{-1} can be found in many texts in numerical linear algebra (see [**16**] and [**23**] for examples). By a careful estimate, the operation count for computing A^{-1} is n^3 for an $n \times n$ matrix (which is remarkably low considering the type of calculations involved [**16**]). Up to the late nineteen eighties, the operation count for Gaussian elimination was thought to be $n^3/3$. The count is mathematically correct and its derivation is still given in many elementary texts on linear algebra.. It is is modestly better theoretically and considerably more efficient for large n (than n^3). However, with a clever use of computer arithmetic in carrying out the algorithm, the count was lowered to $n^{\log_2 7}/3 \approx n^{2.8}$ around 1987 and gradually reduced further to below $n^{2.4}$ shortly thereafter (and confirmed to be at most $n^{2.37}$by 1988). The expectation is that the exponent would continue to decrease toward n^2 with further effort of improvement.

With the improvements, the difference in efficiency between Gaussian elimination and computing A^{-1} is now substantial. To illustrate with $n = 1000 = 10^3$, we have $n^3 = 10^9$, $n^2 = 10^6$ and $n^{2.37} = 10^{7.11} = 1.29 \times 10^7$; Gaussian elimination is more than two hundred times faster for the confirmed operation count of $n^{2.37}$ and three thousand times faster for the conjectured lower bound count. Clearly, we should not solve linear systems numerically by finding A^{-1} when n is large. (Of course, there may still be good reasons to work with inverses for theoretical developments.)

It should not be left unreported that evaluating a determinant is known to require $n!$operations. For large n, $n!$ is proportional to $n^{n+\frac{1}{2}}$. Thus, Cramer's rule is definitely a no-no for a large linear systems.

3.3. LU Decomposition. The factors $\{\ell_{ij}\}$ used in row-echelon reduction appear to have no particular relevance beyond the one time usage in the elimination

of one column of elements in the original matrix A. But what if we want to solve the same set of equations again by the same method but now for a different right hand vector $\mathbf{b}$. The row-echelon reduction process would lead again to the same upper triangular matrix U, given that nothing has changed in A. On the other hand, the modified right hand side $\mathbf{c}$ is now different because $\mathbf{b}$ has changed. With $\mathbf{b}$ replaced by $\mathbf{b}'$, we would have to re-calculate the corresponding $\mathbf{c}'$ to get the new solution $\mathbf{x}'$:

$$U\mathbf{x}' = \mathbf{c}'$$

From the right hand side of (3.6), we see that we need the factors $\{\ell_{ij}\}$ in order to get $\mathbf{c}'$. In other words, there is a good reason for saving the factors $\{\ell_{ij}\}$ in case we we should need them again for the solution of a new right hand member $\mathbf{b}'$. To conserve storage space requirements when n is large, a long standing practice is to store $\{\ell_{ij}\}$ in the effectively unused locations in the upper triangular matrix U, namely the 0 entries of that matrix. Instead of U, we have for the linear system (1.4)

$$\bar{U} == \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ \ell_{21} & u_{22} & u_{23} \\ \ell_{31} & \ell_{32} & u_{33} \end{bmatrix} = \begin{bmatrix} 3 & 2 & 2 \\ \frac{1}{3} & \frac{10}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{4}{5} & -\frac{3}{5} \end{bmatrix}$$

after forward elimination. For a general $n \times n$ matrix A, the factors $\{\ell_{ij}\}$ are similarly tugged away in the unused location of the upper triangular matrix U to get a corresponding $\bar{U}$. .

An alternative characterization of Gaussian elimination is to observe that the matrix A may be factored into a product of two triangular matrices L and U,

$$A = LU. \tag{3.7}$$

In (3.7), the matrix U is the *upper* triangular matrix in (3.4) resulting from forward elimination by row-echelon reduction, and L is the *lower* triangular matrix

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 4/5 & 1 \end{bmatrix}, \tag{3.8}$$

whose elements below the main diagonal are the factors $\{\ell_{ij}\}$ found in the forward elimination calculations. This is proved by applying Gaussian elimination to both sides of (3.7). Note that $L + U = \bar{U} + I$ (not $\bar{U}$)

With (3.7), we see that the solution of the original linear system $A\mathbf{x} = (\mathbf{LUx} =)\ \mathbf{b}$ by Gaussian elimination is effectively accomplished in two steps:

$$L\mathbf{y} = \mathbf{b}, \qquad U\mathbf{x} = \mathbf{y}. \tag{3.9}$$

We first forward eliminate by solving $L\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$ and then backward substitute by solving $U\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$. Since both L and U are triangular, both solution require minimal calculations, but each with an operation count proportional to $n^2$for large n nevertheless. As such, the elegant *LU decomposition* (3.7), though theoretically useful, should not be used in the form (3.9) for the actual solution of the linear system $A\mathbf{x} = \mathbf{b}$. There is no need to do the extra calculations for the inverse of L and U in an efficient solution process.

CHAPTER 13

Geometry of Linear Equations

1. Linear System as a Vector Equation

Linear equations do not always have a solution or they may have many solutions. The first of these situations can be seen from Cramer's rule when the determinant $D = |A|$ of the coefficient matrix A vanishes while some or all of the other determinants $\{D_k\}$ do not (see (**??**)). The usual reason for this case is that the equations are inconsistent. The second possibility occurs when all four determinants involved in Cramer's rule vanish simultaneously so that we have to re-examine the original linear system for the correct solution, if there should be one. One possibility is that one or more of the equations are redundant so that we have fewer equations than there are unknowns. Instead of repeating a discussion of these familiar mathematical facts, we examine these two situations from a geometrical perspective.

For this purpose, we write a linear system of three equations as

$$x_1\mathbf{A}_1 + x_2\mathbf{A}_2 + x_3\mathbf{A}_3 = \mathbf{b}. \tag{1.1}$$

where the vectors $\{\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3\}$ are the columns of the coefficient matrix A of the linear system

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}. = [\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]$$

with

$$\mathbf{A}_1 = (3, 1, 2)^T, \qquad \mathbf{A}_2 = (2, 4, 4)^T, \qquad \mathbf{A}_3 = (2, 1, 1)^T$$

for the system (1.4). (Note that $\mathbf{A}_k$ is neither $\mathbf{a}_k$ nor $\mathbf{a}_k^T$ of (2.3).) In the form (1.1), the left hand side is a linear combination of the three columns of the matrix A. As such, solving the linear system is seen to be finding a linear combination of the three vectors $\{\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3\}$ equal to a given vector $\mathbf{b}$. This is known to be possible if the three column vectors are linearly independent (not colinear or coplanar in the case of physical vectors in three dimensions). In that case, there is a unique set of coordinates $\{x_1,x_2,x_3\}$ that would accomplishes the task.

Consider now the following system of three equations

$$\begin{aligned} 3y_1 + 2y_2 - y_3 &= 3 \\ y_1 + 4y_2 + 3y_3 &= 3 \\ 2y_1 + 4y_2 + 2y_3 &= 3 \end{aligned} \tag{1.2}$$

or, in matrix notation,

$$B\mathbf{y} = \begin{bmatrix} 3 & 2 & -1 \\ 1 & 4 & 3 \\ 2 & 4 & 2 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}.$$

Apply row echelon reduction to this new system to get at the end of forward elimination

$$\begin{bmatrix} 3 & 2 & -1 \\ 0 & 10/3 & 10/3 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ -3/5 \end{pmatrix}. \tag{1.3}$$

The last equation of (1.3) cannot be met since it requires 0 equal to $-3/5$. The system (1.2) therefore has no solution and it is easy to see that the three equations involved are not consistent with each other.

Analogous to (1.1), the new system may be taken in the form

$$y_1\mathbf{B}_1 + y_2\mathbf{B}_2 + y_3\mathbf{B}_3 = \mathbf{b}.$$

where $\mathbf{B}_k$ is the k^{th} column of the matrix B. In this form, we are asked to find a set of coordinates $\{y_1,y_2,y_3\}$ to express $\mathbf{b}$ in terms of the three (basis) vectors $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3\}$. To accomplish this task is to solve the linear system (1.2) for the unknowns $\{y_1,y_2,y_3\}$ which we know to be impossible by (1.3). But in the context of vectors, this means the three columns $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3\}$ are not linearly independent; one must be a linear combination of the other two. In the language of linear algebra, there is a set of constants $\{p_1,p_2,p_3\}$, not all zeros, for which

$$p_1\mathbf{B}_1 + p_2\mathbf{B}_2 + p_3\mathbf{B}_3 = \mathbf{0}.$$

Not surprisingly, the application of row-echelon reduction to the corresponding linear system results in

$$\begin{bmatrix} 3 & 2 & -1 \\ 0 & 10/3 & 10/3 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{1.4}$$

giving $\mathbf{p} = c_0(1, -1, 1)^T$ for any constant c_0 so that

$$\mathbf{B}_2 = \mathbf{B}_1 + \mathbf{B}_3.$$

The three vectors $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3\}$ are therefore not linearly independent and a truly three dimensional vector such as $\mathbf{b} = (3, 3, 3)^T$ would have a component perpendicular (orthogonal) to the plane spanned by the three columns of B and cannot be a linear combination of the three vectors $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3\}$.

The situation would be different if the vector on the right hand side of the system should happen to lie in the plane spanned by the columns of B . For example, if $\mathbf{b}$ should be replaced by $\mathbf{b}' = (3, 3, 18/5)^T$, forward elimination would lead to

$$\begin{bmatrix} 3 & 2 & -1 \\ 0 & 10/3 & 10/3 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}$$

or

$$\begin{aligned} 3y_1' + \quad 2y_2' - \quad y_3' &= 3 \\ (10/3)y_2' + (10/3)y_3' &= 2 \\ 0 &= 0 \end{aligned}$$

giving

$$\begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} -3/5 \\ 3/5 \\ 0 \end{pmatrix} + y_3' \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

where y_3' is unspecified.

It should be mentioned that given the zero row in the row-echelon reduced B as found in (1.4), we know from the theory of determinants that $|B|$ vanishes and the linear system $B\mathbf{x} = \mathbf{b}$ generally does not have a solution. That was the case for $\mathbf{b} = (3, 3, 3)^T$ and other truly 3-dimensional vectors. For a class of right hand members such as $\mathbf{b} = (3, 3, 18/5)^T$, the system has multiple solution, the vanishing of the determinant of the coefficient matrix B notwithstanding. By visualizing linear systems as vectors, we now have a succinct geometrical description of these right hand members; they are those that lie in the plane spanned by the only two linearly independent vectors associated with the columns of B. In fact, the three phenomena of vanishing of the determinant $|B|$ of the coefficient matrix, zero rows in the row-echelon reduced matrix U' and linearly independence of the columns of the coefficient matrix B are now seen to be different manifestation of the same characteristic of B. They have been unified by the geometrical visualization of a linear systems as combination of vectors.

There is another related way of thinking about linear systems $A\mathbf{x} = \mathbf{b}$ in terms of vectors. When multiplied by A, the vector $\mathbf{x}$ is transformed into another vector $\mathbf{b}$. In general, the output vector $\mathbf{b}$ is different from the input vector $\mathbf{x}$ in both magnitude and direction. By the *magnitude* of an n-dimensional vector $\mathbf{v}$, we mean

$$|\mathbf{v}| = \sqrt{\mathbf{v}^T\mathbf{v}} = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}.$$

similar to that for physical vectors. By dividing it by its magnitude $|\mathbf{v}|$, an n-dimensional vector $\mathbf{v}$ is *normalized* to unit length:

$$u = \frac{\mathbf{v}}{|\mathbf{v}|}.$$

Two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are *in the same direction* (or have same orientation) if they are equal after normalization:

$$\frac{\mathbf{v}_1}{|\mathbf{v}_1|} = \frac{\mathbf{v}_2}{|\mathbf{v}_2|}.$$

Given an $n \times n$ matrix A, it is of considerable importance in applications to ask whether there is a *non-zero* n - dimensional input vector $\mathbf{v}^{(i)} = (v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, ..., v_n^{(i)})^T$ for which the output from multiplication by A has the same orientation as the input so that

$$A\mathbf{v}^{(i)} = \lambda_i \mathbf{v}^{(i)} \tag{1.5}$$

where λ_i is some scalar constant corresponding to the magnification factor for the input vector. (It is unnecessary to normalize $\mathbf{v}^{(i)}$ before multiplying it by A in this case since the magnitude factors $\left|\mathbf{v}^{(i)}\right|$ on both side of the equation cancel each other.)

If it exists, a vector that satisfies (1.5) for some λ_i is called the *eigenvector* of the matrix A and the corresponding change of magnitude factor λ_i is called the *eigenvalue* of A for that eigenvector. Together, $\{\lambda_i, \mathbf{v}^{(i)}\}$ is an eigen-pair of A. We show in the next section that eigen-pairs always exist and how they can be found computationally. Note that the zero vector $\mathbf{0} = (0, 0, ..., 0)^T$ satisfies (1.5) for

any λ_i but is *not* an eigenvector as it has been ruled out by the definition which requires an eigenvector must be a non-zero vector.

The Matrix Eigenvalue Problem

We illustrate a straightforward (but not necessarily the most efficient) way to find eigen-pairs with the following matrix

$$C = \begin{bmatrix} 3 & 2 & 2 \\ 1 & 4 & 1 \\ -2 & -4 & -1 \end{bmatrix}.$$

Note that C is just A of (2.7) with a change of sign in the last row. With the change, the system $C\mathbf{x} = \mathbf{d}$ wound be the same as $A\mathbf{x} = \mathbf{b}$ only if $\mathbf{d} = (\mathbf{3}, \mathbf{3}, \mathbf{-3})^T$ but not otherwise. Also, $C\mathbf{v} = \lambda\mathbf{v}$ is not the same as $A\mathbf{x} = \lambda\mathbf{x}$. While we could have worked with the matrix A, the solution comes out nicer for the eigenvalue problem with C as the coefficient matrix. Now, the condition $C\mathbf{v} = \lambda\mathbf{v}$ is the same as

$$[C - \lambda I]\,\mathbf{v} = 0. \tag{1.6}$$

The homogeneous system (1.6) has a nontrivial solution only if the determinant of the matrix $C - \lambda I$ vanishes. This requirement leads to the conditions

$$\begin{vmatrix} 3-\lambda & 2 & 2 \\ 1 & 4-\lambda & 1 \\ -2 & -4 & -1-\lambda \end{vmatrix} = 0,$$

or, upon evaluating the determinant,

$$P_3(\lambda) \equiv 6 - 11\lambda + 6\lambda^2 - \lambda^3 = 0.$$

The characteristic polynomial $P_3(\lambda)$ has three distinct roots

$$\lambda_1 = 3, \qquad \lambda_1 = 2, \qquad \lambda_1 = 1.$$

For each eigenvalue λ_k, we solve by forward elimination and backward substitution the redundant homogeneous linear system (1.6) to get the corresponding eigenvector $\mathbf{v}^{(k)}$ up to an arbitrary multiplicative constant similar to (1.4). In this way, we obtain

$$\begin{aligned} \left\{\lambda_1, \mathbf{v}^{(1)}\right\} &= \left\{3, (0, -1, 1)^T\right\}, \\ \left\{\lambda_2, \mathbf{v}^{(2)}\right\} &= \left\{2, (-2, 1, 0)^T\right\}, \\ \left\{\lambda_3, \mathbf{v}^{(3)}\right\} &= \left\{1, (-1, 0, 1)^T\right\}. \end{aligned}$$

Evidently, the process of finding eigen-pairs of matrices above can be extended to general $n \times n$ matrices. The process would be more tedious. Fortunately, mathematical software are available to take care of the drudgeries. For example, the command

"Eigenvalues[C]"

in Mathematica with the input for the matrix C in the form

"C ={{3, 2, 3},{1, 4, 1},{ − 2, −4, −1}}",

returns the three eigenvalues $\{3, 2, 1\}$ as output while the command

"Eigenvectors[C]"

gives as output the three eigenvectors up to a multiplicative constant. There are also commands that give all the eigen-pairs and the characteristic polynomials for the eigenvalues, respectively.

Occasionally, the characteristic polynomial $P_n(\lambda)$ has *complex* or *multiple* roots. The simple matrix A_1 in (1.7) has purely imaginary eigenvalues $\lambda_1 = i$ and $\lambda_2 = -i$,

$$A_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 \\ -1 & 2\mu \end{bmatrix}, \tag{1.7}$$

while a minor variant A_2 has, for $\mu < 1$, two complex conjugate eigenvalues

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \mu \pm i\sqrt{1-\mu}.$$

Distinct complex eigenvalues give rise to no real complication; only the corresponding eigenvectors take on complex values. Mathematical software automatically return the available complex eigen-pairs when they exist.

2. Matrix Diagonalization

The situation with a multiple root (real or complex) is different. We are guaranteed at least one eigenvector. If there are as many distinct (linearly independent) eigenvectors as the multiplicity of each root, the matrix is said to be *nondefective*. For example, the three eigenvalues of the *symmetric* matrix

$$A_1 = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix}$$

are

$$\lambda_1 = 8, \qquad \lambda_2 = \lambda_3 = -1$$

so that -1 is an eigenvalue of multiplicity 2. (A square matrix is said to be a *symmetric matrix* if $A^T = A$.) It is straightforward to find

$$\mathbf{v}^{(1)} = (2, 1, 2)^T.$$

For $\lambda_2 = \lambda_3 = -1$, the set of equations in vector form for the components of the corresponding eigenvector is

$$[A_1 - \lambda_2 I\,]\,\mathbf{v} = [A_1 + I\,]\,\mathbf{v} = \begin{bmatrix} 4 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{bmatrix} \mathbf{v} = \mathbf{0}. \tag{2.1}$$

All three equations of the linear homogeneous system (2.1) are effectively the same:

$$4v_1 + 2v_2 + 4p_3 = 0 \quad \text{or} \quad v_2 = -2v_1 - 2v_3$$

from which we get two linearly independent solution vectors for (2.1). These can be taken to be:

$$\mathbf{v}^{(2)} = (1, -2, 0)^T, \qquad \mathbf{v}^{(3)} = (0, -2, 1)^T.$$

The first is by setting $v_1 = 1$ and $v_3 = 0$ and the second by setting $v_1 = 0$ and $v_3 = 1$. By forming $\alpha\mathbf{v}^{(2)} + \beta\mathbf{v}^{(3)} = 0$, we see immediately that both α and β must vanish and the two vectors are linearly independent. We have then a full set of three eigenvectors for A_2 and the matrix is nondefective.

If we now collect the three eigenvectors as columns of a new matrix P (called the modal matrix of A_2 in engineering)

$$P = \begin{bmatrix} 2 & 1 & 0 \\ 1 & -2 & -2 \\ 2 & 0 & 1 \end{bmatrix}. \tag{2.2}$$

Observe that

$$AP = [A\mathbf{v}^{(1)},\ A\mathbf{v}^{(2)},\ A\mathbf{v}^{(3)}] = [\lambda_1\mathbf{v}^{(1)},\ \lambda_2\mathbf{v}^{(2)},\ \lambda_3\mathbf{v}^{(3)}] = P\Lambda \tag{2.3}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} 8 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

is a diagonal matrix with the eigenvalues on its diagonal. Since the eigenvectors are linear independent, we can invert P to get P^{-1}(without actually implementing the inversion) and, with it, write (2.3) as

$$P^{-1}AP = \Lambda. \tag{2.4}$$

This relation shows that the matrix A can be diagonalized by the matrix P by way of a "similarity transformation." The combination of $Q^{-1}GQ$ for any invertible matrix Q is called a *similarity transformation* of the matrix G and G is said to be similar to the diagonal matrix Λ with its eigenvalues on the diagonal. The steps leading to the relation (2.4) for the specific matrix A_2 can be repeated to prove the following *spectral theorem* in linear algebra:

THEOREM 22. *A nondefective square matrix can always be diagonalized by its modal matrix by way of a similarity transformation.*

The ability to diagonalize a matrix is important in science and engineering. For example, it enables us to solve a large system of linear ordinary differential equations with constant coefficients (see Chapter 4 of these notes). Unfortunately, not all matrices are non-defective. An example is the matrix

$$A_3 = \begin{bmatrix} 4 & -2 \\ 2 & 0 \end{bmatrix}$$

which has a double eigenvalue $\lambda_1 = \lambda_1 = 2$ with only one eigenvector $\mathbf{v}^{(1)} = (1,1)^T$. It is therefore useful to know what matrices are nondefective and what matrices have only real eigenvalues (and eigenvectors). The following theorem provides useful information to these questions:

THEOREM 23. *If $A^T = A$ (symmetric matrices have only real eigenvalues and a full set of eigenvectors (and therefore nondefective).*

PROOF. We prove here the first half of the theorem. Suppose the symmetric matrix A has a complex eigenvalue λ and $\mathbf{v}$ the corresponding eigenvector. Then the conjugates λ^* and $\mathbf{v}^*$also form an eigen-pair since

$$Av^* = A^*v^* = \lambda^*\mathbf{v}^*.$$

Form the scalar products

$$(\mathbf{v}^*)^T A\mathbf{v} = \lambda(\mathbf{v}^*)^T\mathbf{v} = \lambda\left|\mathbf{v}\right|^2, \qquad \mathbf{v}^T A\mathbf{v}^* = \lambda^*\mathbf{v}^T\mathbf{v}^* = \lambda^*\left|\mathbf{v}\right|^2.$$

and observe

$$\left(\mathbf{v}^T A \mathbf{v}^* \right)^T = (\mathbf{v}^*)^T A \mathbf{v}$$

so that

$$(\lambda^* - \lambda) |\mathbf{v}|^2 = 0.$$

It follows that the imaginary part of λ must vanish and the eigenvalue assumed to be complex is in fact real. □

The proof that all symmetric matrices are nondefective and the extension of the spectral theorem to complex matrices can be found in [**16**]. When a matrix is defective and is therefore cannot diagonalized, the best we can do is a reduction to Jordan form by a similarity transformation. A description of Jordan matrices and their applications to solving differential equations can be found in Chapter 4 of these notes. The theory of Jordan form is discussed in [**16**] and other text on linear algebra.

3. Decoupling a Linear System

Similar to the ODE counterpart, single higher order difference equations and a system of more than one linear difference equations are more compactly written in terms of a state vector as we did for Markov chains in the previous chapter:

$$\mathbf{x}(n+1) = M(n)\mathbf{x}(n) + \mathbf{q}(n), \qquad \mathbf{x}(0) = \mathbf{p} \tag{3.1}$$

for $n = 0, 1, 2, ...$Taken in the form (3.1), $\mathbf{x}(n)$, $\mathbf{q}(n)$ and $\mathbf{p}$ are m vectors and M is a known $m \times m$ matrix. Among the vectors, $\mathbf{q}(n)$ and $\mathbf{p}$ are prescribed and $\mathbf{x}(n)$ is to be determined starting with some initial state (distribution) $\mathbf{x}(0) = \mathbf{p}$. If $\mathbf{q}(n) = \mathbf{0}$, the linear system is said to be *homogeneous*. If M does not depend on n then, the system is said to be of *constant coefficients*. The matrix M is said to be *nondefective* if it has a full set of eigenvectors.

THEOREM 24. *The general solution of linear homogeneous systems with a nondefective constant (transition) matrix M may be written as*

$$\mathbf{x}(n) = c_1 \mathbf{v}^{(1)} \lambda_1^n + c_2 \mathbf{v}^{(2)} \lambda_2^n + \ \cdots \ + c_m \mathbf{v}^{(m)} \lambda_m^n$$

where $\{\lambda_k,\ \mathbf{v}^{(k)}\}$ are the eigen-pairs of M and the constants $\{c_1, c_2, \cdots,\ c_m\}$ are determined by the initial condition $\mathbf{x}(0) = \mathbf{p}$.

PROOF. The proof of this theorem is by diagonalizing M similar to what we did for the ODE counterpart in Math 227A. □

The general solution of linear inhomogeneous systems with forcing with a nondefective constant matrix M may be obtained by the method of variation of parameters or, for a simple forcing term $\mathbf{q}(n)$, the method of undetermined coefficients. These methods are analogous to their ODE counterparts and will not be discussed here. The case of a *defective* matrix with a multiple eigenvalue for which there is an inadequate number of eigenvectors, the sure fire method of solution would be to reduce M to Jordan normal form by a suitable similarity transformation analogous to what was done for ODE in the Math 227A course notes.

When M varies with n, then Theorem 24 does not hold though the method of variation of parameters continues to apply if we have a complete set of (complementary) solutions for the corresponding homogeneous equation. Techniques for finding complementary solutions for linear equations with variable coefficients

can be developed similar to their counterparts in ordinary differential equations. However, the solutions obtained by such methods are no more attractive than a repeated execution of a the recurrence relation implied by the difference equation. In this latter approach, we have the following compact expression for $x(n) = x_n$ using the subscript notation to conserve space:

THEOREM 25. *The unique solution of the IVP*

$$\mathbf{x}(n+1) = M(n)\mathbf{x}(n), \qquad \mathbf{x}(0) = \mathbf{p}$$

may be taken in the form

$$\mathbf{x}(n) = \Pi_{k=0}^{n-1}[M_k]\left\{p + \sum_{j=0}^{n-1} \Pi_{k=j}^{0}[M_k]^{-1} q_j\right\}$$

with

$$\Pi_{j=i}^{k}[M_j] = M_k M_{k-1} \cdots M_i, \tag{3.2}$$

PROOF. For $n = 0$ and $n = 1$, we have

$$\begin{aligned}
\mathbf{x}_1 &= M_0\mathbf{x}_0 + \mathbf{q}_0, \\
\mathbf{x}_2 &= M_1\mathbf{x}_1 + \mathbf{q}_1 = M_1[M_0\mathbf{x}_0 + \mathbf{q}_0] + \mathbf{q}_1 \\
&= \mathbf{p}\Pi_{k=0}^{1}[M_k] + \Pi_{k=0}^{1}[M_k][M_0]^{-1}\mathbf{q}_0 + \Pi_{k=0}^{1}[M_k]\Pi_{k=1}^{0}[M_k]^{-1}\mathbf{q}_1
\end{aligned}$$

upon observing the notation (3.2). By induction, we get for general n

$$\begin{aligned}
\mathbf{x}_n &= M_{n-1}\mathbf{x}_{n-1} + \mathbf{q}_{n-1} \\
&= \Pi_{k=0}^{n-1}[M_k]\left\{\Pi_{j=0}^{0}[M_j]^{-1}\mathbf{q}_0 + \Pi_{k=1}^{0}[M_k]^{-1}\mathbf{q}_1 + \cdots + \Pi_{k=n-1}^{0}[M_k]^{-1}\mathbf{q}_{n-1}\right\} \\
&\quad + \mathbf{p}\Pi_{k=0}^{n-1}[M_k] \\
&= \Pi_{k=0}^{n-1}[M_k]\left\{\mathbf{p} + \sum_{j=0}^{n-1} \Pi_{k=j}^{0}[M_k]^{-1}\mathbf{q}_j\right\}.
\end{aligned}$$

□

Bibliography

[1] R.H. Bartels and G.W. Stewart, "*Solution of the matrix equation $AX + XB = C$*," Comm. ACM, vol. 15, 1972, 820-826.

[2] W.E. Boyce & R.C. DiPrima, *Elementary Differential Equations*. 7th ed.,Wiley, 2001

[3] L. Edelstein-Keshet, *Mathematical Models in Biology*, SIAM ed., 2005.

[4] C.P. Fall, E.S. Marland, J.M. Wagner and J.J. Tyson, *Computational cell Biology*, Springer, 2002

[5] W. Fuller, *An Introduction to Probability Theory and Its Applications*, vol. I, 2^{nd} ed., John Wiley and Sons, Inc. 1957.

[6] G. Golub, S. Nash and C. Van Loan, "*A Hessenberg-Schur Method for the Problem $AX + XB = C$*," Tras. Automatic Control, vol AC-24, (6), 1979, 909-913.

[7] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhauser, Boston, 1982

[8] W. Kelley & A. Peterson, *The Theory of Differential Equations, Classical and Qualitative*, Pearson Prentice Hall, 2004.

[9] H.B. Keller, *Numerical Methods for Two-Point Boundary Value Problems*, Ginn/Blaisdell, Waltham, Mass., 1968.

[10] F. Kozin, "*On the probability densities of the output of some random systems*," J. App. Mech., vol. 28, 1961, 161-165.

[11] Q. Nie, F.Y.M. Wan, Y.-T. Zhang and X.-F. Liu, "*Compact integration factor methods in high spatial dimension*," J. Comp. Phys., vol. 227, 2008, 5238-5525.

[12] M.A. Nowak & R.M. May, *Virus Dynamics*, Oxford, 2000.

[13] L. Perko, *Differential Equations and Dynamical Systems*, 3rd ed., Springer, 2001

[14] L.A. Segel, *Modeling dynamics phenomena in molecular and cellular biology*, Cambridge University Press, 1984.

[15] R.T. Smith and R.B. Hinton, *Calculus*, 3rd ed., McGraw-Hill, 2008.

[16] G. Strang, Linear Algebra and Its Applications, 3rd ed., Harcourt, 1988.

[17] S.H. Strogatz, *Nonlinear Dynamics and Chaos*, Paperback ed., Perseus, 2000.

[18] H. C. Tuckwell, *Introduction to theoretical neurobiology*, vol. 1 - Linear cable theory and dendritic structure, Cambridge University Press, 1988

[19] H. C. Tuckwell, *Introduction to theoretical neurobiology*, vol. 1I - Nonlinear and stochastic theories, Cambridge University Press, 1988

[20] H. C. Tuckwell, F.Y.M. Wan and J.-P.Rospars, "*A spatial stochastic neuronal model with Ornstein-Uhlenbeck input current,"* Biol. Cybern. vol. 86, 2002, 137-145.

[21] H. C. Tuckwell & F.Y.M. Wan, "Nature of equilibria and effects of drug treatments in some viral population dynamical models," IMA J. Math. Appl. Med. & Biol., Vol. 17, 2000, pp. 311-327.

[22] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed., Elsevier, 2008

[23] C. F. Van Loan, *Introduction to Scientific Computing*, 2nd ed., Prentice Hall, 2000.

[24] E. O. Voit, *A First Course in Systems Biology,* Garland Science, Taylor & Francis Group, New York and London, 2013

[25] F. Y. M. Wan, "*A Direct Method for Linear Dynamical Problems in Continuum Mechanics with Random Loads*," Studies in Appl. Math. vol. 52, 1973, 259-276.

[26] F. Y. M. Wan, *Introduction to the Calculus of Variations and Its Applications*, Chapman & Hall, 1995.

[27] F.Y.M. Wan, *Mathematical Models and Their Analysis*, Harper and Row, 1989.