2009-07-10

The Expectation of Transition Events on Finite-state Markov Chains

Jeremy Michael West
Brigham Young University - Provo

THE EXPECTATION OF TRANSITION EVENTS ON FINITE-STATE MARKOV CHAINS

by

Jeremy M. West

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Mathematics

Brigham Young University

August 2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Jeremy M. West

This thesis has been read by each of member of the following graduate committee and by majority vote has been found to be satisfactory.

Date — Jeffrey Humpherys, Chair

Date — Kening Lu

Date — C. Shane Reese

Date — H. Dennis Tolley

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Jeremy M. West in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Jeffrey Humpherys Chair, Graduate Committee

Accepted for the Department

Tyler J. Jarvis
Chair, Department of Mathematics

Accepted for the College

Thomas W. Sederberg, Associate Dean
College of Physical and Mathematical Sciences

Abstract

The Expectation of Transition Events on Finite-state Markov Chains

Jeremy M. West

Department of Mathematics

Master of Science

Markov chains are a fundamental subject of study in mathematical probability and have found wide application in nearly every branch of science. Of particular interest are finite-state Markov chains; the representation of finite-state Markov chains by a transition matrix facilitates detailed analysis by linear algebraic methods.

Previous methods of analyzing finite-state Markov chains have emphasized state events. In this thesis we develop the concept of a transition event and define two types of transition events: cumulative events and time-average events. Transition events generalize state events and provide a more flexible framework for analysis. We derive computable, closed-form expressions for the expectation of these two events, characterize the conditioning of transition events, provide an algorithm for computing the expectation of these events, and analyze the complexity and stability of the algorithm. As an application, we derive a construction of composite Markov chains, which we use to study competitive dynamics.

Acknowledgments

This thesis is proof of the support, patience, and encouragement of my wife, Wendy. I thank her, and my two daughters, Ayden and Claire, for living with a husband and father who has not often been home, and for being satisfied with what a graduate student can provide. My advisor, Dr. Jeffrey Humpherys, has been an excellent mentor and friend and has provided me with numerous opportunities to learn and advance. Similarly, the members of my thesis committee, Drs. Kening Lu, C. Shane Reese, and H. Dennis Tolley, have provided invaluable advice, instruction, suggestions, and encouragement. Also, Dr. Brian Ewald, at Florida State University, has contributed to the ideas and analytic techniques of this thesis.

I am grateful to the many students with whom I have worked, primarily the members of the IMPACT program and the graduate program. Particularly, I am grateful to Bryant Angelos, Blake Barker, McKay Heasley, Brent Kerby, and Nathan Perry for their friendship during graduate school. Finally, I am thankful to the BYU Math Department faculty, staff, and students, for their financial, academic, and emotional support – particularly during the challenging time after my accident. It has been my pleasure to study here.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Markov chains are a fundamental subject of study in mathematical probability and have found wide application in nearly every branch of science. Of particular interest are finite-state Markov chains; the representation of finite-state Markov chains by a transition matrix facilitates detailed analysis by linear algebraic methods. The relatively recent development of the theory of generalized inverses of linear transformations has led to the development of many new results on the asymptotic and transient behavior of finite-state Markov chains.

1.1 Previous Work

Much of what is known about the application of generalized inverses to finite-state Markov chains is due to the work of Decell and Odell [2, 13] and the work of Meyer [12]. Decell and Odell introduced the notion of the fundamental matrix. This matrix, derived using the Moore-Penrose inverse, contains many of the fundamental quantities of a finite-state Markov chain. For example, using the fundamental matrix, the number of steps to absorption, the probability of absorption into a particular state, and the mean first passage – or average number of steps until the first visit to a state – can be determined.

Meyer improved upon the work of Decell and Odell by showing that the Drazin inverse, a spectral generalized inverse, was a more natural fit for Markov chains. Using the Drazin inverse, Meyer reproduced and extended the results of Decell and Odell and the development using the Drazin inverse is more natural.

1.2 Contributions

Both of the aforementioned approaches emphasize state events: the number of visits to a state, absorption into a specific state, and mean first passage. The identifying characteristic of a Markov chain is that it is completely described by the transition probabilities. It seems natural that events on Markov chains should be described in terms of transitions, rather than states.

Chapters 2-4 provide the mathematical background for this thesis. In Chapter 5 we rigorously define two types of transition events and give closed-form, computable expressions for the expectation of these events on the broadest possible class of finite-state Markov chains. The expressions for these two types of transition events are nearly identical, differing mainly in the type of generalized inverse used. This suggests that similar expressions may exist for other types of transition events using different generalized inverses. In each section of Chapter 5 we give examples to show how state events may be reproduced by transition events. Additionally, we provide several examples of transition events that are not readily derived in terms of state events.

Our problem was originally motivated by an investigation of competitive stochastic systems. In Chapter 6 we derive a method for constructing a composite system from multiple individual systems. We show how this composite system may be used to analyze multi-agent competition. In particular, we describe some transition events that can be used to study competition in these composite systems.

In Chapter 7 we characterize the conditioning of the expectation of transition events. We give an algorithm and the describe complexity and numerical stability of the algorithm. In Chapter 8, we use this algorithm to perform calculations for a specific Markov chain and compare the results to a Monte Carlo simulation.

1.3 Future Work

The work of Decell and Odell and of Meyer addresses variance in addition to expectation. This is a limitation of our current work, as our results yield only the first moment. Furthermore, we have described only two types of transition events: cumulative events and time-average events. It may be possible to develop results similar to those presented in this thesis for other classes of transition events. In particular, we envision events described in terms of stopping times as being particularly relevant.

Chapter 2. Linear Analysis

Linear algebra is a fundamental tool for analyzing finite state Markov chains. In this chapter we develop some of the less-known results from linear algebra that are used in this thesis. We denote by $\mathbb{R}^n$ the n-dimensional Euclidean space and by $\mathbb{R}^{m\times n}$ the space of $m \times n$ matrices with real entries. For a matrix, A^T is the transpose and $\mathrm{tr}(A)$ is the trace, or sum of the diagonal entries. The (i,j) entry of A is $A_{i,j}$ or for a vector, x_i is the i^{th} entry. We also make use of the notation $|A|$ to denote the matrix whose (i,j) entry is $|A_{i,j}|$. Matrix inequalities are interpreted entry-wise, that is, $A \leq B$ if $A_{i,j} \leq B_{i,j}$ for al i and j.

2.1 Special Products

We make use of two non-traditional matrix products: the Hadamard product and the Kronecker product.

Definition 2.1. The *Hadamard product* of two $m \times n$ matrices A and B is the $m \times n$ matrix whose (i,j) entry is

$$[A \odot B]_{i,j} = A_{i,j} B_{i,j}. \tag{2.1}$$

It is immediate from the definition that the Hadamard product is both commutative and associative. However, it does not necessarily commute, nor associate, with standard matrix multiplication. The following theorem, which may be found in [6, p. 305], relates the Hadamard product to standard matrix-vector multiplication.

Theorem 2.2. *Let $x \in \mathbb{R}^n$ and $A, B \in \mathbb{R}^{m\times n}$ be given and let $D = \mathrm{diag}(x)$. Then*

$$\left[ADB^T\right]_{i,i} = \left[(A \odot B)x\right]_i . \tag{2.2}$$

Proof. Since D is diagonal, the i^{th} diagonal entry of ADB^T is given by

$$\left[ADB^T\right]_{i,i} = \sum_{j=1}^{n} A_{i,j} x_j B^T_{j,i} = \sum_{j=1}^{n} A_{i,j} x_j B_{i,j} = \sum_{j=1}^{n} [A \odot B]_{i,j} x_j = \left[(A \odot B)x\right]_i . \tag{2.3}$$

□

Corollary 2.3. *Let $x \in \mathbb{R}^n$ and $A, B \in \mathbb{R}^{m \times n}$ be given and let $D = \operatorname{diag}(x)$. Then*

$$\sum_{i=1}^{m} \left[(A \odot B)x\right]_i = \operatorname{tr}(ADB^T). \tag{2.4}$$

Definition 2.4. The *Kronecker product* of $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$ is the $mp \times nq$ matrix,

$$A \otimes B = \begin{bmatrix} A_{1,1}B & \dots & A_{1,n}B \\ \vdots & & \vdots \\ A_{m,1}B & \dots & A_{m,n}B \end{bmatrix}. \tag{2.5}$$

Example 2.5. For the matrices

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad \text{and} \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{2.6}$$

we have

$$A \otimes B = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 3 & 0 & 4 & 0 \\ 0 & 3 & 0 & 4 \end{bmatrix}, \qquad \text{and} \qquad B \otimes A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 3 & 4 \end{bmatrix}. \tag{2.7}$$

This example illustrates that the Kronecker product is not commutative.

A convenient indexing scheme for a Kronecker product is using tuples. If $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$ then $A \otimes B$ is an $m \times n$ block matrix, where each block is $p \times q$. Denote by $[A \otimes B]_{(i_1,i_2),(j_1,j_2)}$

the (i_2, j_2) entry of the (i_1, j_1) block of $A \otimes B$. Using this indexing scheme,

$$[A \otimes B]_{m(i_1-1)+i_2, n(j_1-1)+j_2} = [A \otimes B]_{(i_1,i_2),(j_1,j_2)} = A_{i_1,j_1} B_{i_2,j_2}. \tag{2.8}$$

For example, using A and B in Example 2.5,

$$[A \otimes B]_{1,3} = [A \otimes B]_{(1,1),(2,1)} = A_{1,2} B_{1,1} = 2. \tag{2.9}$$

It is easy to verify that Kronecker products are associative. Therefore, we need no parentheses in the expression $A = A_1 \otimes \cdots \otimes A_p$. Furthermore, this allows a definition of Kronecker powers:

$$A^{\otimes p} = A \otimes \cdots \otimes A. \tag{2.10}$$

When the Kronecer product of multiple matrices is formed, we use the same indexing scheme, only a p-wise Kronecker product requires a p-tuple for the row and a p-tuple for the column. For vectors $x_1 \in \mathbb{R}^{n_1}, \ldots, x_p \in \mathbb{R}^{n_p}$, the product $x_1 \otimes \cdots \otimes x_p \in \mathbb{R}^{n_1 \cdots n_p}$ is a vector whose entries may be indexed by a p-tuple $(i_1, \ldots, i_p)$, in which case

$$[x_1 \otimes \cdots \otimes x_p]_{(i_1,\ldots,i_p)} = [x_1]_{i_1} \ldots [x_p]_{i_p}. \tag{2.11}$$

2.2 Generalized Inverses

Any bijection $f : X \to Y$ has a unique inverse $f^{-1} : Y \to X$ satisfying

$$\begin{aligned} f(f^{-1}(y)) &= y \\ f^{-1}(f(x)) &= x, \end{aligned} \tag{2.12}$$

that is, $f^{-1} \circ f$ is the identity on X and $f \circ f^{-1}$ is the identity on Y. If f is injective then it is bijective onto $f(X)$ so that f has a unique inverse $f^{-1} : f(X) \to X$. If f is not injective, there is not a unique inverse. However, if we restrict f to $S \subset X$ on which f is injective, we may obtain an inverse.

Treating $A \in \mathbb{R}^{m \times n}$ as a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, we encounter the same situation. If $m = n$

and A is full rank then A has a unique inverse A^{-1} satisfying $AA^{-1} = A^{-1}A = I$. If $m \neq n$ then A cannot be both injective and surjective. If we restrict to subspaces of $\mathbb{R}^n$ on which A is bijective, the restricted linear operator has an inverse, which we extend to a linear operator on $\mathbb{R}^m \to \mathbb{R}^n$, called a *generalized inverse.* If A is not injective, there are multiple possible definitions for a generalized inverse, each arising from a different choice of subspaces.

2.2.1 The Moore-Penrose Inverse. The Moore-Penrose inverse divides along orthogonal complements. Since $N(A)$ is the problematic subspace, that is, $N(A)$ is nontrivial precisely when A fails to be injective, we decompose $\mathbb{R}^n$ into complementary subspaces $\mathbb{R}^n = N(A) \oplus N(A)^\perp = N(A) \oplus R(A^T)$, noting that A is injective on $R(A^T)$.

Definition 2.6. For a matrix $A \in \mathbb{R}^{m\times n}$, the *Moore-Penrose inverse* of A is the unique matrix $A^\dagger \in \mathbb{R}^{n\times m}$ satisfying

(i) $AA^\dagger A = A$

(ii) $A^\dagger AA^\dagger = A^\dagger$

(iii) $(AA^\dagger)^T = AA^\dagger$

(iv) $(A^\dagger A)^T = A^\dagger A$.

Remark. By a simple examination of the symmetry of properties (i)-(iv) it is immediate that if $A^\dagger$ is the Moore-Penrose inverse of A then A is the Moore-Penrose inverse of $A^\dagger$. That is, $(A^\dagger)^\dagger = A$. Furthermore, it is clear that if A is invertible, then A^{-1} is the Moore-Penrose inverse of A.

We now justify the declaration that a Moore-Penrose inverse exists and is unique. Recall that a *projection* is an idempotent matrix P, that is a matrix satisfying $P^2 = P$. For such a matrix, $R(P)$ and $N(P)$ are complementary subspaces and $Px = x$ for all $x \in R(P)$. We say that P is the projection onto $R(P)$ along $N(P)$. Furthermore, for any two complementary subspaces U and W of $\mathbb{R}^n$, there exists a unique projection onto U along W. What's more, the projection onto U along W is the unique matrix satisfying $Px = x$ for $x \in U$ and $Px = 0$ for $x \in W$. A projection P is symmetric if and only if $N(P) = R(P)^\perp$, in which case, we say that P is the *orthogonal projection* onto $R(P)$. Therefore, a matrix satisfying $Px = x$ for $x \in U$ and $Px = 0$ for $x \in U^\perp$ is the orthogonal projection onto U.

Theorem 2.7. *Let P_U denote the orthogonal projection onto a subspace U. For a given $A \in \mathbb{R}^{m\times n}$, $A^\dagger$ satisfies (i)-(iv) if and only if*

(a) $AA^\dagger = P_{R(A)}$,

(b) $A^\dagger A = P_{R(A^\dagger)}$.

Proof. ($\Rightarrow$) If $A^\dagger$ satisfies (i)-(iv) then by (i), $AA^\dagger AA^\dagger = (AA^\dagger A)A^\dagger = AA^\dagger$ so that $AA^\dagger$ is idempotent (a projection). Property (iii) guarantees that $AA^\dagger$ is also symmetric so that $AA^\dagger$ is an orthogonal projection. It remains to show that $R(AA^\dagger) = R(A)$. By property (i), and the fact that $R(AB) \subseteq R(A)$ for any matrices A and B,

$$R(A) = R(AA^\dagger A) \subseteq R(AA^\dagger) \subseteq R(A), \tag{2.13}$$

by which we see that $R(A) = R(AA^\dagger)$. The same arguments with properties (ii) and (iv) and the roles reversed gives a similar result for $A^\dagger A$.

($\Leftarrow$) If $A^\dagger$ satisfies (a) then $AA^\dagger$ is symmetric which gives (iii). Furthermore,

$$AA^\dagger A = P_{R(A)}A = A, \tag{2.14}$$

so we have (i). By similar reasoning, (b) gives (ii) and (iv). □

Corollary 2.8. *If $A^\dagger$ satisfies (i)-(iv) then* $\operatorname{rank} A^\dagger = \operatorname{rank} A$.

Proof. We have $\operatorname{rank} A = \operatorname{rank} P_{R(A)}$. Since $\dim R(AB) \le \dim R(B)$ for any matrices A and B, this implies that

$$\operatorname{rank} A = \operatorname{rank} P_{R(A)} = \operatorname{rank} AA^\dagger \le \operatorname{rank} A^\dagger. \tag{2.15}$$

On the other hand,

$$\operatorname{rank} A^\dagger = \operatorname{rank} P_{R(A^\dagger)} = \operatorname{rank} A^\dagger A \le \operatorname{rank} A. \tag{2.16}$$

□

Corollary 2.9. *If $A^\dagger$ satisfies (i)-(iv) then $R(A^\dagger) = R(A^T)$.*

Proof. If $x \in N(A) = R(A^T)^\perp$ then

$$P_{R(A^\dagger)}x = A^\dagger Ax = 0. \tag{2.17}$$

Therefore, $N(A) \subseteq N(P_{R(A^\dagger)})$. Since $P_{R(A^\dagger)}$ is the orthogonal projection onto $R(A^\dagger)$, we have $N(A) \perp R(A^\dagger)$. It follows that $R(A^T)^\perp = N(A) \subseteq R(A^\dagger)^\perp$, and so $R(A^\dagger) \subseteq R(A^T)$. Since $\operatorname{rank} A^\dagger = \operatorname{rank} A = \operatorname{rank} A^T$ we obtain $R(A^\dagger) = R(A^T)$. □

Remark. Recall that $\dim R(A^T) = \dim R(A)$, so the two spaces are isomorphic. What's more, $A|_{R(A^T)}$ is an isomorphism, that is, for each $y \in R(A)$, there exists a unique $x \in R(A^T)$ such that $Ax = y$.

Theorem 2.10. *For every matrix $A \in \mathbb{R}^{m\times n}$ there exists a unique Moore-Penrose inverse. That is, there exists a unique $A^\dagger \in \mathbb{R}^{n\times m}$ satisfying (i)-(iv).*

Proof. If $A^\dagger$ satisfies (i)-(iv) and $z \in R(A)^\perp$ then by Theorem 2.7

$$A^\dagger z = A^\dagger AA^\dagger z = A^\dagger P_{R(A)}z = A^\dagger 0 = 0. \tag{2.18}$$

For any $y \in R(A)$, denote by x_y the unique point in $R(A^T)$ such that $Ax_y = y$. Then $y = P_{R(A)}y = AA^\dagger y$. Since by Corollary 2.9, $R(A^\dagger) = R(A^T)$ we must have $A^\dagger y = x_y$.

Thus far we have shown that the only possible definition of $A^\dagger$ that satisfies (i)-(iv) is to have $A^\dagger z = 0$ for $z \in R(A)^\perp$ and $A^\dagger y = x_y$, which establishes uniqueness. For existence, we show that this definition satisfies (a) and (b) in Theorem 2.7.

If $y \in R(A)$ then $AA^\dagger y = Ax_y = y$. Furthermore, for $z \in R(A)^\perp$ we have $AA^\dagger z = 0$. Therefore $AA^\dagger = P_{R(A)}$. For $x_y \in R(A^\dagger)$, where $y \in R(A)$, we have $A^\dagger Ax = x$. If $z \in N(A)$ then it is immediate that $A^\dagger Az = 0$ so $A^\dagger A = P_{R(A^\dagger)}$. □

The Moore-Penrose inverse is often viewed as an equation solving inverse. Consider the system $Ax = b$. If the system is consistent, that is, $b \in R(A)$, then a solution is given by $x = A^\dagger b$ since $Ax = AA^\dagger b = P_{R(A)}b = b$. If the system is inconsistent, then $x = A^\dagger b$ is a least squares solution of $Ax = b$. Note that we have not assumed A has full column rank.

The singular value decomposition gives a convenient method for computing $A^\dagger$. Recall that any $A \in \mathbb{R}^{m\times n}$ has a singular value decomposition

$$A = U\Sigma V^T, \tag{2.19}$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times n}$ has block form

$$\Sigma = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.20}$$

where $\Sigma_0 \in \mathbb{R}^{r\times r}$ is the diagonal matrix $\Sigma_0 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

Theorem 2.11. *Let $A = U\Sigma V^T$ be a singular value decomposition of $A \in \mathbb{R}^{m\times n}$. Then*

$$A^\dagger = V\Sigma^\dagger U^T, \qquad \textit{where} \qquad \Sigma^\dagger = \begin{bmatrix} \Sigma_0^{-1} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times m}. \tag{2.21}$$

Proof. We first verify that $\Sigma^\dagger$ is in fact the Moore-Penrose inverse of Σ. To see this,

$$\Sigma\Sigma^\dagger = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_0^{-1} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{2.22}$$

which is clearly symmetric. Similarly,

$$\Sigma^\dagger\Sigma = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{2.23}$$

is symmetric. Also, using these identities, it is immediate that $\Sigma\Sigma^\dagger\Sigma = \Sigma$ and $\Sigma^\dagger\Sigma\Sigma^\dagger = \Sigma^\dagger$.

We now show that $A^\dagger$ satisfies (i)-(iv). We have

$$\begin{aligned}
AA^\dagger A &= U\Sigma V^T V\Sigma^\dagger U^T U\Sigma V^T = U\Sigma\Sigma^\dagger\Sigma V^T = U\Sigma V^T = A, && (2.24)\\
A^\dagger AA^\dagger &= V\Sigma^\dagger U^T U\Sigma V^T V\Sigma^\dagger U^T = V\Sigma^\dagger\Sigma\Sigma^\dagger U^T = V\Sigma^\dagger U^T = A^\dagger, && (2.25)\\
(AA^\dagger)^T &= (U\Sigma\Sigma^\dagger U^T)^T = U(\Sigma\Sigma^\dagger)^T U^T = U\Sigma\Sigma^\dagger U^T = AA^\dagger, && (2.26)\\
(A^\dagger A)^T &= (V\Sigma^\dagger\Sigma V^T)^T = V(\Sigma^\dagger\Sigma)^T V^T = V\Sigma^\dagger\Sigma V^T = A^\dagger A. && (2.27)
\end{aligned}$$

□

Remark. The singular value decomposition of a matrix is not unique. The matrix Σ is unique, but the matrices U and V are not. In spite of this, the Moore-Penrose inverse is unique. That is, for *any* singular value decomposition of A, the above theorem gives a method for computing the *unique* Moore-Penrose inverse $A^\dagger$.

There are other inverses that are similar to the Moore-Penrose because they share one or more of properties (i)-(iv) from Definition 2.6 on Page 6. These are collectively called (i, j, k)-inverses, referring to an inverse which satisfies the i^{th}, j^{th}, and k^{th} properties but not the remaining property. It is not necessary to have three properties. For instance, in this thesis we make use of a (1,2)-inverse. That is, an inverse satisfying properties (i) and (ii) but not (iii) or (iv).

2.2.2 The Drazin Inverse. There are many spectral properties of square matrices that are not preserved by the Moore-Penrose inverse. For example, If A and B are similar, invertible matrices, it follows that A^{-1} and B^{-1} are similar. This is not true with the Moore-Penrose inverse. Furthermore, it is not always the case that $(A^2)^\dagger = (A^\dagger)^2$. Of course, we cannot hope for the property $(AB)^\dagger = B^\dagger A^\dagger$ either. For these properties we turn to a spectral generalized inverse, called the Drazin inverse. The outline of this section follows [1, Chapter 7].

We use the convention $A^0 = I$. Recall that for any two matrices A and B, $R(AB) \subseteq R(A)$. Therefore, for $A \in \mathbb{R}^{n\times n}$ we obtain the nested sequence

$$\mathbb{R}^n = R(A^0) \supseteq R(A^1) \supseteq R(A^2) \supseteq \cdots . \quad (2.28)$$

Since this is a decreasing sequence and $\mathbb{R}^n$ is finite-dimensional there exists a smallest nonnegative integer k such that $R(A^k) = R(A^{k+1}) = R(A^{k+2}) = \cdots$. Equivalently, $\operatorname{rank} A^k = \operatorname{rank} A^{k+1} = \operatorname{rank} A^{k+2} = \cdots$, which also implies $N(A^k) = N(A^{k+1})$.

Definition 2.12. For a matrix $A \in \mathbb{R}^{n\times n}$, the smallest nonnegative integer k such that $\operatorname{rank} A^k = \operatorname{rank} A^{k+1}$ is called the *index* of A and is denoted $\operatorname{Ind} A$.

Remark. The matrix A is invertible if and only if $\operatorname{Ind} A = 0$ since $\mathbb{R}^n = R(I) = R(A^0) = R(A^1)$ if and only if A is full rank.

Definition 2.13. For $A \in \mathbb{R}^{n\times n}$ with $\operatorname{Ind} A = k$, the *Drazin inverse* of A is the unique matrix A^D satisfying

(i) $A^D A A^D = A^D$,

(ii) $AA^D = A^D A$,

(iii) $A^{k+1}A^D = A^k$.

We are obliged to show that such a matrix exists and is unique.

Proposition 2.14. *Let $A \in \mathbb{R}^{n\times n}$ have index k. Then $R(A^k)$ and $N(A^k)$ are complementary subspaces. That is, $\mathbb{R}^n = R(A^k) \oplus N(A^k)$.*

Proof. The Rank-Nullity Theorem implies that $\dim R(A^k) + \dim N(A^k) = n$. Therefore, to show that $R(A^k)$ and $N(A^k)$ are complementary subspaces, it is sufficient to show that $R(A^k) \cap N(A^k) = \{\, 0 \,\}$. For $k = 0$ this is immediate. If $k \geq 1$ and $y \in R(A^k) \cap N(A^k)$ then $y = Ax$ for some $x \in R(A^{k-1})$. However, since $y \in N(A^k)$, $0 = A^k y = A^{k+1}x$. Thus $x \in N(A^{k+1})$. But $k = \operatorname{Ind} A$ so $N(A^{k+1}) = N(A^k)$. Therefore, $y = A^k x = 0$ and $R(A^k) \cap N(A^k) = \{\, 0 \,\}$. □

Remark. The space $N(A^k)$ is the generalized eigenspace of A corresponding to the eigenvalue $\lambda = 0$ and $R(A^k)$ is the direct sum of the nonzero generalized eigenspaces of A.

Recall that for a matrix $A \in \mathbb{R}^{n\times n}$, a subspace $U \subseteq \mathbb{R}^n$ is *A-invariant* if $x \in U$ implies that $Ax \in U$.

Proposition 2.15. *If $k = \operatorname{Ind} A$ for $A \in \mathbb{R}^{n\times n}$ then $N(A^k)$ and $R(A^k)$ are A-invariant subspaces of $\mathbb{R}^n$.*

Proof. If $x \in R(A^k)$ then $Ax = R(A^{k+1}) = R(A^k)$ since $k = \operatorname{Ind} A$. Therefore $R(A^k)$ is A-invariant. If $x \in N(A^k)$, then $Ax \in N(A^{k-1}) \subseteq N(A^k)$, hence $N(A^k)$ is also A-invariant. □

Recall that a square matrix N is *nilpotent of order* k if k is the smallest nonnegative integer such that $N^k = 0$ and $N^{k-1} \neq 0$. The zero matrix is nilpotent of order 1.

Lemma 2.16. *Let $A \in \mathbb{R}^{n\times n}$ with $\operatorname{Ind} A = k$ be given and let $r = \operatorname{rank} A^k$. Then there exists an invertible $P \in \mathbb{R}^{n\times n}$ such that*

$$A = P \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} P^{-1}, \tag{2.29}$$

where $M \in \mathbb{R}^{r\times r}$ is invertible and $N \in \mathbb{R}^{(n-r)\times(n-r)}$ is nilpotent of order k.

Proof. Let $p_1, \ldots, p_r$ be a basis for $R(A^k)$ and $p_{r+1}, \ldots, p_n$ be a basis for $N(A^k)$. By Proposition 2.14, $R(A^k)$ and $N(A^k)$ are complementary subspaces. It follows that $p_1, \ldots, p_n$ is a basis for $\mathbb{R}^n$ and $P = \begin{bmatrix} p_1 & \ldots & p_n \end{bmatrix}$ is invertible. By Proposition 2.15, $R(A^k)$ and $N(A^k)$ are invariant subspaces of A, therefore, $P^{-1}AP$ has the form

$$P^{-1}AP = \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}. \tag{2.30}$$

Where $M \in \mathbb{R}^{r\times r}$ is the action of A on $R(A^k)$ and $N \in \mathbb{R}^{(n-r)\times(n-r)}$ is the action of A on $N(A^k)$. It is immediate then that M is invertible and N is nilpotent of order k. □

Theorem 2.17. *Every square matrix has a unique Drazin inverse.*

Proof. Given $A \in \mathbb{R}^{n\times n}$, let P, M, and N be the matrices in the decomposition (2.29) guaranteed by Theorem 2.16. Define

$$A^D = P \begin{bmatrix} M^{-1} & 0 \\ 0 & 0 \end{bmatrix} P^{-1}. \tag{2.31}$$

We claim that A^D satisfies (i)-(iii) in Definition 2.13. By inspection, $A^D A A^D = A^D$, so (i) is satisfied. For the second,

$$AA^D = P \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} P^{-1} P \begin{bmatrix} M^{-1} & 0 \\ 0 & N \end{bmatrix} P^{-1} = P \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} P^{-1} = A^D A, \tag{2.32}$$

so (ii) is satisfied. Note that

$$A^k = P \begin{bmatrix} M^k & 0 \\ 0 & 0 \end{bmatrix} P^{-1}, \tag{2.33}$$

since N is nilpotent of order k. Therefore,

$$A^{k+1}A^D = P \begin{bmatrix} M^{k+1} & 0 \\ 0 & 0 \end{bmatrix} P^{-1} P \begin{bmatrix} M^{-1} & 0 \\ 0 & 0 \end{bmatrix} P^{-1} = P \begin{bmatrix} M^k & 0 \\ 0 & 0 \end{bmatrix} P^{-1} = A^k. \tag{2.34}$$

To show uniqueness, suppose X satisfies (i)-(iii) and write

$$P^{-1}XP = \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix}. \tag{2.35}$$

By (iii),

$$P \begin{bmatrix} M^k & 0 \\ 0 & 0 \end{bmatrix} P^{-1} = A^k = A^{k+1}X = P \begin{bmatrix} M^{k+1}X_1 & M^{k+1}X_2 \\ 0 & 0 \end{bmatrix} P^{-1}. \tag{2.36}$$

Since P is invertible, $M^k = M^{k+1}X_1$ and $0 = M^{k+1}X_2$. Since M is invertible, $X_2 = 0$ and $MX_1 = I$, or $X = M^{-1}$. By (ii) and (iii)

$$P \begin{bmatrix} M^k & 0 \\ X_3M^k & 0 \end{bmatrix} P^{-1} = XA^{k+1} = A^k = P \begin{bmatrix} M^k & 0 \\ 0 & 0 \end{bmatrix} P^{-1}, \tag{2.37}$$

from which it follows that $X_3 = 0$. By (i) and (ii), we have that $AX^2 = X$. Therefore, $NX_4^2 = X_4$. Thus, $N^{k-1}X_4 = N^kX_4^2 = 0$. But then $N^{k-2}X_4 = N^{k-1}X_4^2 = 0$. Continuing, we finally arrive at

$X_4 = N^0 X_4 = N^1 X_4^2 = 0$, hence $X = A^D$ is unique. □

Remark. By (2.32) in the proof of the previous theorem, $AA^D x = x$ for $x \in R(A^k)$ and $AA^D x = 0$ for $x \in N(A^k)$. It follows that $AA^D = A^D A$ is the projection onto $R(A^k)$ along $N(A^k)$.

2.3 Limits and Summation

In the study of Markov chains, the series

$$\sum_{k=0}^{\infty} T^k = I + T + T^2 + T^3 + \cdots \tag{2.38}$$

is of fundamental importance. In this section we determine when the series converges and what the limit is.

A matrix norm is a norm on $\mathbb{R}^{m\times n}$ that satisfies the submultiplicative property $\|AB\| \leq \|A\|\|B\|$. It follows that $\|A^k\| \leq \|A\|^k$. If $\|T\| < 1$ then $\lim_{k\to\infty} \|T^k\| \leq \lim_{k\to\infty} \|T\|^k = 0$. For any matrix norm, $|T_{i,j}| \leq \|T\|$ so$\|T\| < 1$ implies that $\lim_{k\to\infty} T^k = 0$.

Proposition 2.18. *If $\|T\| < 1$ for some matrix norm $\|\cdot\|$ then* (2.38) *converges to $(I-T)^{-1}$.*

Proof. If we multiply the partial sums of (2.38) by $(I-T)$ we obtain

$$(I-T)\sum_{k=0}^{N} T^k = I - T^{N+1}. \tag{2.39}$$

Since $T^k \to 0$ as $k \to \infty$, we obtain the desired result. □

Let $\sigma(T)$ be the spectrum, or set of eigenvalues, of T. The spectral radius is

$$\rho(T) = \max_{\lambda\in\sigma(T)} |\lambda|. \tag{2.40}$$

Note that for any induced matrix norm, the inequality, $\|Tx\| \leq \|T\|\|x\|$ implies that $|\lambda| \leq \|T\|$. Therefore, $\rho(T) \leq \|T\|$. Thus, $\|T\| < 1$ implies $\rho(T) < 1$ but the converse is not necessarily true. If $\rho(T) < 1$, we can show that (2.38) converges to $(I-T)^{-1}$. To do so, we appeal to Jordan forms. We give a brief review of Jordan forms here and refer the reader to [11] for more details.

The Jordan form of a diagonalizable matrix is the diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. If a matrix is not diagonalizable, the Jordan form becomes more complicated. This can only occur if the matrix has repeated eigenvalues. If A is $n \times n$ with $r < n$ distinct eigenvalues $\lambda_1, \ldots, \lambda_r$, then A is similar to a sum of the form

$$P^{-1}AP = D + J, \tag{2.41}$$

where D and J are block diagonal matrices $D = \operatorname{diag}(D_1, \ldots, D_r)$ and $J = \operatorname{diag}(J_1, \ldots, J_r)$. Here D_i and J_i are both $m_i \times m_i$ matrices, where m_i is the algebraic multiplicity of λ_1 and D_i is the scalar matrix $D_i = \lambda_i I$.

The blocks J_i, called Jordan segments, have a block structure $J_i = \operatorname{diag}(J_i^{(1)}, \ldots, J_i^{(p_i)})$. Each block $J_i^{(k)}$ is called a Jordan block. It is a nilpotent matrix of the form

$$J_i^{(k)} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{2.42}$$

Each Jordan segment $J_i^{(k)}$ is a $n_i^{(k)} \times n_i^{(k)}$ matrix and is nilpotent of order $n_i^{(k)} - 1$. Furthermore, the ones in $J_i^{(k)}$ advance up diagonals with each power. That is, the ones in $J_i^{(k)}$ lie on the first super-diagonal. $(J_i^{(k)})^2$ has a similar structure but with ones on the second super-diagonal, etc. If a Jordan block is 1×1, it is just the scalar zero. If an entire Jordan segment is zero, then for that eigenvalue, A is diagonalizable. This occurs when the geometric multiplicity of λ_i equals its algebraic multiplicity, in which case we say that λ_i is *semisimple*.

Jordan forms are an essential tool for analyzing the convergence of (2.38). Note that $T^k = P(D+J)^k P^{-1}$. Therefore, T^k converges if and only if $(D+J)^k$ converges. Similarly,

$$\sum_{k=0}^{\infty} T^k = \sum_{k=0}^{\infty} \left(P(D+J)P^{-1}\right)^k = P\left(\sum_{k=0}^{\infty}(D+J)^k\right)P^{-1}. \tag{2.43}$$

Therefore, (2.38) converges for T if and only if it converges for the Jordan form $D + J$ of T.

Lemma 2.19. *The limit* $\lim_{k\to\infty} T^k = 0$ *if and only if* $\rho(T) < 1$.

Proof. Let $D + J$ be the Jordan form of T and let $r = \rho(T)$. Note that J is nilpotent of order m. By the binomial theorem,

$$(D+J)^k = \sum_{p=0}^{k} \binom{k}{p} D^{k-p} J^p. \tag{2.44}$$

For $p \geq m$, $J^p = 0$, therefore,

$$(D+J)^k = \sum_{p=0}^{m} \binom{k}{p} D^{k-p} J^p. \tag{2.45}$$

For all $p < m$, $J^p_{i,j}$ is either one or zero. Since D is diagonal, $\left[D^{k-p}J^p\right]_{i,j} = D^{k-p}_{i,i} J^p_{i,j}$, which is either $D^{k-p}_{i,i}$ or 0. Thus, $(D+J)^k \to 0$ is possible only when $r < 1$. If this is the case, then since $p \leq m$ and m is fixed, for large enough k, $\binom{k}{p} \leq k^{m+1}$. Therefore,

$$|[(D+J)^k]_{i,j}| \leq (m+1)k^{m+1}r^{k-p} \to 0, \qquad \text{as } k \to \infty. \tag{2.46}$$

□

Corollary 2.20. *If* $\rho(T) < 1$ *then* (2.38) *converges to* $(I - T)^{-1}$.

Proof. Multiplying $(I - T)$ by partial sums of (2.38) we obtain

$$(I - T)\sum_{k=0}^{N} T^k = I - T^{N+1}. \tag{2.47}$$

By the previous proposition, $T^{N+1} \to 0$ as $N \to \infty$ so we obtain the desired result. □

It should be clear that (2.38) converges only when $\lim T^k = 0$. This implies that $\rho(T) < 1$. However, if $\rho(T) = 1$, it is possible that T is *Cesaro summable.* A *unit eigenvalue* of T is any eigenvalue satisfying $|\lambda| = 1$.

Theorem 2.21 (see [11])**.** *The series*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} T^k \tag{2.48}$$

converges if and only if $\rho(T) < 1$, or if $\rho(T) = 1$, it converges if and only the unit eigenvalues of T are semisimple. When it converges, the limit is the spectral projection onto $N(I-T)$ along $R(I-T)$. If $S = I - T$, this may be written in terms of the Drazin inverse as follows.

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} T^k = I - SS^D. \tag{2.49}$$

2.4 Stochastic Matrices

A Markov chain is represented by its transition matrix (see Section 3.3), which is a stochastic matrix. A matrix T is stochastic if $T_{i,j} \geq 0$ and $\|T_j\|_1 = 1$ for all j, where T_j is the j^{th} column of T. Sometimes the term stochastic matrix refers to nonnegative matrices whose rows sum to 1. This may be clarified by indicating whether the matrix is column stochastic or row stochastic, however, in this thesis we deal only with column stochastic matrices and we refer to them simply as stochastic.

The theory of nonnegative matrices, also called Perron-Frobenius theory, is a rich area of matrix analysis. We address only a few ideas and refer the reader to [10] and [11] for more details. Nonnegative matrices fall into two basic categories: reducible and irreducible.

Definition 2.22. Let T be a square nonnegative matrix. T is *reducible* if there exists a permutation P such that

$$P^T T P = \begin{bmatrix} X & 0 \\ Y & Z \end{bmatrix}, \tag{2.50}$$

where X and Z are both square. If no such permutation exists then T is *irreducible.*

The way to interpret this definition is using graph theory. A nonnegative $n \times n$ matrix may be interpreted as the adjacency matrix of a directed graph with n nodes. The entry $T_{i,j}$ corresponds to the weight from node j to node i and a zero entry indicates that no edge exists. If P is a permutation matrix then $P^T T P$ is the adjacency matrix of an isomorphic graph, that is, a graph obtained by relabeling the nodes.

Suppose that T is reducible, and that it has already been transformed into the form (2.50). Then there are no paths from any of the nodes corresponding to the block Z to the nodes corresponding to block X. If Y is nonzero then there is some path from X to Z, but once a node in the Z group has been entered, there is no path back to the X group. In contrast, if T is irreducible, no such subdivision exists. Such a graph is said to be *strongly connected.* That is, there exists a path from any node to any other node.

If T is reducible then the submatrix X is also a square, nonnegative matrix and it is also either reducible or irreducible. If X is reducible, we may apply another permutation to T to produce a matrix of the form.

$$T \sim \begin{bmatrix} X_1 & 0 & 0 \\ X_2 & X_3 & 0 \\ Y_1 & Y_2 & Z \end{bmatrix}. \tag{2.51}$$

Continuing in this manner we obtain the *canonical form for reducible matrices*,

$$T = \left[\begin{array}{cccc|cccc} T_{11} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ T_{21} & T_{22} & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ T_{r1} & T_{r2} & \dots & T_{rr} & 0 & 0 & \dots & 0 \\ \hline T_{r+1,1} & T_{r+1,2} & \dots & T_{r+1,r} & T_{r+1,r+1} & 0 & \dots & 0 \\ T_{r+2,1} & T_{r+2,2} & \dots & T_{r+2,r} & 0 & T_{r+2,r+2} & & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ T_{m1} & T_{m2} & \dots & T_{mr} & 0 & 0 & \dots & T_{mm} \end{array}\right]. \tag{2.52}$$

Each diagonal block T_{ii} is an irreducible matrix. The blocks $T_{11}, \dots, T_{rr}$ are called the transient classes. The blocks $T_{r+1,r+1}, \dots, T_{mm}$ are called the ergodic classes. In terms of the graph of T, once the nodes corresponding to the block T_{11} have been left, they cannot be re-entered. Therefore, they are transient nodes. A path through the graph then travels down the transient classes until arriving in one of the ergodic classes. Once a node in an ergodic class has been reached, there is no path to another ergodic class or back to a transient class.

Proposition 2.23. *If T is a stochastic matrix then $\rho(T) = 1$ and $1 \in \sigma(T)$.*

Proof. Since $\|T_j\|_1 = 1$ for all j, it follows that $\|T\|_1 = 1$. Therefore, $\rho(T) \leq 1$. However, since all the columns of T, or all the rows of T^T sum to 1, $\lambda = 1$ is an eigenvalue of T^T, and therefore T. We conclude that $\rho(T) = 1$. □

The previous result and the analysis of Section 2.3 indicate that (2.38) does not converge. However, Perron-Frobenius theory does guarantee that every unit eigenvalue of a stochastic matrix is semisimple. Recall that an eigenvalue is semisimple if the algebraic multiplicity equals the geometric multiplicity. This leads us to state, without proof, the following theorem.

Theorem 2.24. *For every stochastic matrix T, the series*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} T^k = G \tag{2.53}$$

converges to $G = I - SS^D$ where $S = I - T$.

Chapter 3. Stochastic Analysis

Finite state Markov chains are a specific class of stochastic processes. In this chapter we develop the basic theory of probability spaces and stochastic processes.

3.1 Probability Spaces

Probability theory is the principle tool for analyzing models with uncertainty. The key concept in probability theory is the experiment, for example, rolling a die or tossing a coin. Experiments are characterized by the fact that the outcome is uncertain, that is, each realization of the experiment may yield different results. Probability spaces provide a mathematical formalism for probability theory so that the tools of measure theory may be applied. In this section we give a brief introduction; for a more complete reference, see [4] and [8].

Recall that a σ-algebra $\mathcal{F}$ on a set Ω is a collection of subsets of Ω satisfying

(i) $\emptyset \in \mathcal{F}$.

(ii) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, where $A^c = \Omega \setminus A$ is the complement of A in Ω.

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$ is countable then $\cup_i A_i \in \mathcal{F}$.

Given a σ-algebra on Ω, a measure is a function $\mu : \mathcal{F} \to [0, \infty]$ satisfying

(i) $\mu(\emptyset) = 0$,

(ii) If $A_1, A_2, \ldots$ is a countable disjoint collection of sets in $\mathcal{F}$ then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i). \tag{3.1}$$

Definition 3.1. A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where Ω, called the *sample space*, consists of all possible outcomes, $\mathcal{F}$ is a σ-algebra of measurable subsets of Ω, called *events*, and P is a measure $P : \mathcal{F} \to [0, 1]$ satisfying $P(\Omega) = 1$.

Definition 3.2. If $\mathcal{A}$ is a collection of subsets of a sample space Ω, the σ-algebra generated by $\mathcal{A}$, denoted $\sigma(\mathcal{A})$, is the smallest σ-algebra over Ω containing $\mathcal{A}$.

Implicit in the definition is a claim that such a σ-algebra exists and is unique. The power set of Ω is a σ-algebra over Ω containing every subset. Therefore, a σ-algebra containing $\mathcal{A}$ exists. It is straightforward to show that the intersection of σ-algebras is a σ-algebra. Therefore, the intersection of all σ-algebras on Ω that contain $\mathcal{A}$ is the unique smallest σ-algebra $\sigma(\mathcal{A})$.

Definition 3.3. Given $A, B \in \mathcal{F}$, the *conditional probability* $P(A|B)$ defined by

$$P(A|B)P(B) = P(A \cap B) \tag{3.2}$$

is the probability that event A occurs given that event B occurs.

Theorem 3.4 (Bayes' Rule)**.** *For $A, B \in \mathcal{F}$,*

$$P(A|B)P(B) = P(B|A)P(A). \tag{3.3}$$

Proof. From the definition,

$$P(A|B)P(B) = P(A \cap B) = P(B \cap A) = P(B|A)P(A). \tag{3.4}$$

□

Theorem 3.5 (Law of Total Probability)**.** *Let $A_1, A_2, \ldots \in \mathcal{F}$ be a countable partition of Ω, that is, the A_i are pairwise disjoint and $\cup_i A_i = \Omega$. Then for any $A \in \mathcal{F}$,*

$$P(A) = \sum_{i=1}^{\infty} P(A|A_i)P(A_i). \tag{3.5}$$

Proof. Since the A_i are a partition of Ω,

$$\begin{aligned} \sum_{i=1}^{\infty} P(A|A_i)P(A_i) = \sum_{i=1}^{\infty} P(A \cap A_i) &= P\left(\bigcup_{i=1}^{\infty} A \cap A_i\right) \\ &= P\left(A \bigcup_{i=1}^{\infty} A_i\right) = P(A \cap \Omega) = P(A). \end{aligned} \tag{3.6}$$

□

Corollary 3.6. *Let $A_1, A_2, \ldots \in \mathcal{F}$ be a countable partition of Ω. Then for any $A, B \in \mathcal{F}$,*

$$P(A|B) = \sum_{i=1}^{\infty} P(A|A_i)P(A_i|B). \tag{3.7}$$

3.2 Random Variables

Definition 3.7. A *random variable* is an $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$. That is, for any Borel set $B \subseteq \mathbb{R}$, the preimage under X of B is measurable: $X^{-1}(B) \in \mathcal{F}$.

We have insisted that X be measurable with respect to a σ-algebra $\mathcal{F}$. This requirement is somewhat superfluous as every function $X : \Omega \to \mathbb{R}$ is measurable with respect to some σ-algebra on Ω. In fact, one can show that

$$\sigma(X) = \left\{ X^{-1}(B) \,\middle|\, B \text{ is a Borel set} \right\} \tag{3.8}$$

is a σ-algebra on Ω. Clearly X is measurable with respect to $\sigma(X)$. In fact, $\sigma(X)$ is the smallest σ-algebra for which X is measurable and is called the *σ-algebra generated by X*. Therefore, unless the σ-algebra is important, we assume $\mathcal{F} = \sigma(X)$.

Although a random variable is a function, there is some notation that is unique to random variables. Often, we drop the argument, denoting $X = X(\omega)$, and think of X as a variable in its own right. The notation $X \in A$ for $A \subseteq \mathbb{R}$ is used to denote the subset of Ω:

$$X \in A = X^{-1}(A) = \{\, \omega \in \Omega \mid X(\omega) \in A \,\}\,. \tag{3.9}$$

Definition 3.8. The *distribution* of a random variable X is the measure μ on $\mathbb{R}$ defined by

$$\mu(A) = P(X \in A). \tag{3.10}$$

Remark. Because the preimage under X of any Borel set is an $\mathcal{F}$-measurable set in Ω, μ is a probability measure on a σ-algebra containing all of the Borel sets of $\mathbb{R}$.

There are two fundamental results from measure theory that have great application in probability theory: the Monotone Convergence Theorem and the Dominated Convergence Theorem. Before stating them, we remind the reader of a few results from measure theory.

Definition 3.9. Given $A \in \mathcal{F}$, The *indicator random variable* of A, denoted 1_A is the random variable

$$1_A(\omega) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \notin A. \end{cases} \tag{3.11}$$

Definition 3.10. A *simple random variable* is a random variable S that can be written as a weighted sum of a finite number of indicator random variables. That is

$$S = \sum_{i=1}^{k} \alpha_i 1_{A_i}, \tag{3.12}$$

where $\alpha_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$.

Remark. The A_i may be chosen to be disjoint and we assume that this is the case.

Definition 3.11. The *expectation* of a random variable X, denoted EX is the integral of X on Ω

with respect to P. That is,

$$EX = \int_\Omega X dP. \tag{3.13}$$

Recall from measure theory that an integral of the type used in Definition 3.11 is defined in four phases. First, for an indicator random variable, $E1_A = P(A)$ is the obvious definition. For a simple function, the desirable linearity property dictates that $ES = \sum_{i=1}^{k} \alpha_i P(A_i)$. For a nonnegative random variable X, that is $X(\omega) \geq 0$ for all $\omega \in \Omega$, we define EX to be

$$EX = \sup \{ \, ES \mid 0 \leq S \leq X \text{ and } S \text{ is a simple random variable.} \, \} \tag{3.14}$$

This expectation may be infinite. However, since $P(\Omega) = 1 < \infty$, if X is bounded, so is EX. For general random variables X, we define the *positive part* of X by

$$X^+(\omega) = \begin{cases} X(\omega) & X(\omega) \geq 0 \\ 0 & X(\omega) < 0 \end{cases} \tag{3.15}$$

and the *negative part* of X by

$$X^-(\omega) = \begin{cases} 0 & X(\omega) \geq 0 \\ -X(\omega) & X(\omega) < 0. \end{cases} \tag{3.16}$$

Therefore, X^+ and X^- are nonnegative random variables and $X = X^+ - X^-$. If either of EX^+ or EX^- are finite, we define $EX = EX^+ - EX^-$. By convention, for $\alpha \in \mathbb{R}$, $\infty - \alpha = \infty$ and $\alpha - \infty = -\infty$. Only the indeterminant case $\infty - \infty$ is left undefined. A necessary and sufficient condition for EX to exist and be finite is $E|X| < \infty$.

We now state the Monotone and Dominated Convergence Theorems and a couple of their corollaries. There are analogous results in measure theory, so we do not prove them here. We invite the reader to review [4] and [8].

Theorem 3.12 (Monotone Convergence Theorem). *Let* $0 \leq X_1 \leq X_2 \leq \cdots$ *be a sequence of* *nonnegative, increasing random variables. Then* $\lim X_k$ *exists and*

$$E \lim_{k\to\infty} X_k = \lim_{k\to\infty} EX_k. \tag{3.17}$$

Corollary 3.13. *Let* $0 \le X_1, X_2, \ldots$ *be a sequence of nonnegative random variables. Then*

$$E \sum_{k=1}^{\infty} X_k = \sum_{k=1}^{\infty} EX_k. \tag{3.18}$$

Theorem 3.14 (Dominated Convergence Theorem)**.** *Let* $X_1, X_2, \ldots$ *be a sequence of random variables. If there exists a nonnegative random variable* Y *with* $EY < \infty$ *and* $|X_k| \le Y$ *for all* k *then* $\lim X_k$ *exists and*

$$E \lim_{k\to\infty} X_k = \lim_{k\to\infty} EX_k. \tag{3.19}$$

Corollary 3.15. *Let* $X_1, X_2, \ldots$ *be a sequence of random variables. If there exists a nonnegative random variable* Y *with* $EY < \infty$ *and*

$$\sum_{k=0}^{N} |X_k| \le Y \tag{3.20}$$

for all N *then* $\sum X_k$ *exists and*

$$E \sum_{k=0}^{\infty} X_k = \sum_{k=0}^{\infty} EX_k. \tag{3.21}$$

3.3 Markov Chains

In this section we develop the fundamentals of Markov chains. We begin with the general notion of a stochastic process and then focus in on temporally-homogeneous, finite-state Markov chains.

Definition 3.16. A *stochastic process* is a sequence of random variables $X_k : \Omega_k \to \mathbb{R}$, $k = 0, 1, \ldots$.

Generally we think of k as a time variable. At each discrete instant of time, an experiment occurs that is modeled by the random variable X_k. Presumably, the experiments are in some way related. Often they are the same experiment. In this case $\Omega_0 = \Omega_1 = \cdots$.

For a stochastic process, we are generally interested in the sequence of outcomes $X_0, X_1, \ldots$. A result of Kolmogorov allows us to construct a space $\Omega = \Omega_0 \times \Omega_1 \times \cdots$ consisting of all possible sequences of outcomes. In this case $\omega \in \Omega$ is a sequence $(\omega_0, \omega_1, \ldots)$, where each $\omega_k \in \Omega_k$. We may think of X_k as a function on Ω by setting

$$X_k(\omega) = X_k(\omega_k). \tag{3.22}$$

In this setting, we often define $X = (X_0, X_1, \ldots)$ to be the sequence of random variables. Thus, X is in fact a random variable on Ω.

Definition 3.17. A *filtration* on a sample space Ω is a sequence of σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$.

A σ-algebra describes the measurable events. In that sense, it is a description of the amount of information available at time k. For a stochastic process, we choose $\mathcal{F}_k$ so that $X_0, \ldots, X_k$ are all $\mathcal{F}_k$-measurable. Therefore, unless otherwise specified, we assume that $\mathcal{F}_k = \sigma(X_0, \ldots, X_k)$ is the smallest σ-algebra on Ω for which $X_0, \ldots, X_k$ are all measurable.

Definition 3.18. Let X be a random variable on Ω and let $\mathcal{F}$ be a σ-algebra on Ω. The *conditional expectation* of X with respect to $\mathcal{F}$, denoted $E(X|\mathcal{F})$ is the unique $\mathcal{F}$-measurable random variable Y satisfying for all $A \in \mathcal{F}$,

$$\int_A X dP = \int_A Y dP. \tag{3.23}$$

Remark. While it is true that the conditional expectation of X with respect to $\mathcal{F}$ exists and is unique, the proof of these facts is beyond the scope of this thesis; see [4]. By way of notation, $E(X|Y)$, where X and Y are random variables, is defined to be $E(X|\sigma(Y))$.

Note that conditional expectation differs from conditional probability in that conditional probability is a number between 0 and 1, whereas the conditional expectation of a random variable is another random variable. The idea is that Y is the best approximation of X that is measurable with respect to the σ-algebra $\mathcal{F}$. If X is already measurable with respect to $\mathcal{F}$ then $E(X|\mathcal{F}) = X$.

The context of temporally-homogeneous, finite-state Markov chains makes the abstract concepts of filtrations and conditional expectation much more concrete. We only treat temporally homogeneous Markov chains in this thesis. Therefore, when referring to a Markov chain, it is implicit that we mean a temporally homogeneous Markov chain.

Definition 3.19. A stochastic process X_k on $\Omega = \Omega_0 \times \Omega_1 \times \cdots$ is a *Markov chain* if for $k = 1, 2, \ldots$,

$$E(X_k | \mathcal{F}_{k-1}) = E(X_k | X_{k-1}). \tag{3.24}$$

The equality (3.24) is called the *Markov property.* It says that if we know the outcome of $X_0, \ldots, X_{k-1}$ then we are no better off for predicting X_k than if we just knew X_{k-1}.

Definition 3.20. A *finite-state* Markov chain is a Markov chain in which each X_k takes on values in some finite subset of $\mathbb{R}$, which we denote by $\mathcal{S} = \{ s_1, \ldots, s_n \}$. Each s_i is called a *state.* If $X_k = s_i$ we say that the Markov chain is in state s_i at time k.

By the Markov property,

$$P(X_k = s_i | X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_{k-1} = s_{i_{k-1}}) = P(X_k = s_i | X_{k-1} = s_{i_{k-1}}), \tag{3.25}$$

that is, the probability of *transitioning* into state s_i at time k depends only on the state at time $k-1$.

Definition 3.21. The transition matrix of a finite-state Markov chain X_k with n states is the $n \times n$ stochastic matrix T whose entries are

$$T_{i,j} = P(X_k = s_i | X_{k-1} = s_j). \tag{3.26}$$

Remark. Note that $T_{i,j}$ is independent of k. That is, we are assuming that the probability of moving from state s_j to state s_i in a single step is the same regardless of the time. This is precisely what we mean by temporally homogeneous. This assumption simplifies much of the analysis while still describing a surprising breadth of phenomena. A simple application of Theorem 3.5 shows that T is stochastic.

Proposition 3.22. *For all $m \geq 0$,*

$$P(X_{k+m} = s_i | X_k = s_j) = \left[T^m\right]_{i,j}. \tag{3.27}$$

Proof. The proof is by induction on m. By Corollary 3.6,

$$\begin{aligned} P(X_{k+m} = s_i | X_k = s_j) &= \sum_{h=1}^{n} P(X_{k+m} = s_i | X_{k+m-1} = s_h) P(X_{k+m-1} = s_h | X_k = s_j) \\ &= \sum_{h=1}^{n} T_{i,h} \left[T^{m-1}\right]_{h,j} = \left[T^m\right]_{i,j}. \end{aligned} \tag{3.28}$$

□

Let μ be the distribution of X_k. Since X_k takes on finitely many values in $\mathbb{R}$, μ is completely described by point-masses at $s_1, \ldots, s_n$. Therefore, there is a unique correspondence between distributions of X_k and stochastic vectors in $\mathbb{R}^n$ and we use these two representations interchangeably. Thus, when we refer to the *initial-distribution* of a finite-state Markov chain, we mean the stochastic vector $\mu \in \mathbb{R}^n$ defined by

$$\mu_i = P(X_0 = s_i). \tag{3.29}$$

Given an initial distribution μ, there is a unique probability measure P_μ on Ω satisfying

$$P_\mu(X_0 = s_i) = \mu_i, \qquad \text{and} \qquad P_\mu(X_k = s_i | X_{k-1} = s_j) = T_{i,j}, \tag{3.30}$$

see, for example, [4]. Let E_μ denote expectation with respect to P_μ:

$$E_\mu Y = \int_\Omega Y dP_\mu. \tag{3.31}$$

Theorem 3.23. *Let X_k be a finite-state Markov chain with initial distribution μ and transition matrix T. Then for $k = 0, 1, 2, \ldots,$*

$$P_\mu(X_k = s_i) = \left[T^k \mu\right]_i. \tag{3.32}$$

Proof. We proceed by induction on k. For $k = 0$ the result is obvious since $P_\mu(X_0 = s_i) =$

$\left[T^0\mu\right]_i = \mu_i$. By Theorem 3.5,

$$\begin{aligned} P_\mu(X_k = s_i) &= \sum_{h=1}^{n} P_\mu(X_k = s_i | X_{k-1} = s_h) P_\mu(X_{k-1} = s_h) \\ &= \sum_{h=1}^{n} T_{i,h} \left[T^{k-1}\mu\right]_h = \left[T^k\mu\right]_i . \end{aligned} \tag{3.33}$$

□

Chapter 4. Numerical Analysis

Modern computers have made it possible to solve increasingly large and computationally complex problems. However, computers suffer from the limitations of time, space, and precision. This necessitates a careful numerical analysis of algorithms that are intended to solve applied problems. In this chapter we review some of the fundamental concepts of numerical analysis.

4.1 Complexity

An analysis of the complexity of an algorithm is generally split into two pieces: the temporal complexity and the spatial complexity. Temporal complexity refers to the amount of time required to complete an operation. This is typically measured in terms of floating point operations (FLOPs). Spatial complexity is the amount of space, or memory, required by the algorithm and is generally measured in bytes.

Since the exact number of operations and the time each takes to complete varies by machine, as does the amount of memory, we typically describe complexity using the less detailed and more practical big-O notation.

Definition 4.1. Let $f(x)$ and $g(x)$ be real-valued functions defined on some subset of the reals. We say that $f(x)$ is *big-O* of $g(x)$, denoted $f(x) = O(g(x))$, as $x \to x_0$ if there exists $M > 0$ and

$\delta > 0$ such that

$$|f(x)| \leq M|g(x)|, \qquad x \in (x_0 - \delta, x_0 + \delta). \tag{4.1}$$

Example 4.2. Let $f : \mathbb{N} \to \mathbb{N}$ be given by $f(n) = 3n^3 + 3n^2 + 2$. Then $f(n) = O(n^3)$ as $n \to \infty$ since for $M = 4$ and sufficiently large n, $|f(n)| \leq 4n^3$.

Using big-O notation, we say an algorithm has temporal complexity $O(t(n))$ and spatial complexity $O(s(n))$ if the number of FLOPs for a problem of size n is a function which is $O(t(n))$ as $n \to \infty$ and the number of bytes required for a problem of size n is $O(s(n))$ as $n \to \infty$.

Example 4.3. Consider the problem of adding n floating point numbers $x_1, \ldots, x_n$. The basic algorithm is

Algorithm 4.4. (i) Initialize $s = 0$.

(ii) For $i = 1, \ldots, n$, set $s = s + n_i$.

This algorithm requires $n + 1$ floating point operations. One to initialize $s = 0$ and then one for each addition $s = s + x_i$. It also requires $n + 1$ “pieces” of memory, one for s and one for each x_i. The number of bytes depends on the machine and the precision, however, it is clearly some constant multiple of $n + 1$. Therefore, we say that the algorithm has temporal complexity $O(n)$ and spatial complexity $O(n)$.

4.2 Conditioning

Conditioning is a measure of the sensitivity of a function to changes in the inputs. The outline of this section follows [14, Chapter 12].

Definition 4.5. For a given function $f : \mathbb{R}^n \to \mathbb{R}^m$, the *absolute condition number* of f at a point x is

$$a(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta f(x)\|}{\|\delta x\|}, \tag{4.2}$$

where $\delta f(x) = f(x + \delta x) - f(x)$.

Remark. We hasten to point out that the conditioning of a problem has absolutely nothing to do with any specific algorithm for computing f. Conditioning is a property of the function itself, not how the function is implemented on a computer. We also note that on a finite-dimensional vector space, the choice of norm is often unimportant since any two norms differ by at most a constant multiple independent of x. Therefore, we typically use the norm that is convenient for the problem.

Definition 4.6. For a given function $f : \mathbb{R}^n \to \mathbb{R}^m$, the *relative condition number* of f at a point x is

$$\kappa(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta f(x)\|}{\|f(x)\|} \bigg/ \frac{\|\delta x\|}{\|x\|}\ , \tag{4.3}$$

where $\delta f(x) = f(x + \delta x) - f(x)$.

Remark. Relative errors are generally more informative that absolute errors because they are invariant of scale. Furthermore, relative errors are typically used in defining and implementing finite-precision arithmetic, see Section 4.3.

The condition number is a unitless quantity which reflects how perturbations in inputs are magnified by the function f. The larger the condition number, the greater the change caused by a fixed-sized perturbation. Whether a particular condition number is acceptable depends largely on the problem.

The relative condition number satisfies

$$\kappa(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta f(x)\|}{\|\delta x\|} \frac{\|x\|}{\|f(x)\|} \frac{\|x\|}{\|f(x)\|} a(x). \tag{4.4}$$

Often we are interested in uniform bounds for $\kappa = \kappa(x)$ or $a = a(x)$. Note that

$$\kappa(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta f(x)\|}{\|\delta x\|} \frac{\|x\|}{\|f(x)\|} \frac{\|x\|}{\|f(x)\|} a(x). \tag{4.5}$$

Thus, the relative condition number may be determined in terms of the absolute condition number. For a function $f(x, y)$ of two variables, we often evaluate κ_x and κ_y separately. That is, we fix y and treat f as a function of x to determine κ_x, then we reverse roles for κ_y. The next few examples illustrate these ideas.

Example 4.7. Consider the problem of matrix-vector multiplication for an invertible matrix. That is, consider the function $f(A, x) = Ax$ where $A \in \mathbb{R}^{n \times n}$ is invertible and $x \in \mathbb{R}^n$. For the condition number with respect to A we have

$$(A + \delta A)x = b + \delta b. \tag{4.6}$$

That is, perturbing A by δA produces the perturbed output $b + \delta b$ where $b = Ax$ is the solution to the unperturbed problem. In other words, $\delta f = \delta b$. Therefore,

$$\delta A x = \delta b. \tag{4.7}$$

Therefore, $\|\delta b\| = \|\delta A x\| \leq \|\delta A\| \|x\|$. Hence,

$$a_A = \lim_{\delta \to 0} \sup_{\|\delta A\| \leq \delta} \frac{\|\delta b\|}{\|\delta A\|} \leq \lim_{\delta \to 0} \sup_{\|\delta A\| \leq \delta} \frac{\|\delta A\| \|x\|}{\|\delta A\|} = \|x\|. \tag{4.8}$$

It follows that

$$\kappa_A = a_A \frac{\|A\|}{\|b\|} \leq \frac{\|A\| \|x\|}{\|Ax\|} = \frac{\|A\| \|A^{-1} Ax\|}{\|Ax\|} \leq \|A\| \|A^{-1}\|. \tag{4.9}$$

Example 4.8. Consider the same problem: $f(A, x) = Ax = b$, but now treat f as a function of x for fixed A. A perturbation δx yields the system

$$A(x + \delta x) = b + \delta b, \tag{4.10}$$

where again, $Ax = b$. Therefore, $A\delta x = \delta b$ and it follows that $\|\delta b\| \leq \|A\| \|\delta x\|$. Hence,

$$a_x = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta b\|}{\|\delta x\|} \leq \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \|A\| \frac{\|\delta x\|}{\|\delta x\|} = \|A\|. \tag{4.11}$$

Notice the symmetry between a_A and a_x. For the relative condition number,

$$\kappa_x = a_x \frac{\|x\|}{\|b\|} \leq \|A\| \frac{\|x\|}{\|Ax\|} \leq \|A\| \|A^{-1}\|. \tag{4.12}$$

Example 4.9. Now consider the function $g(A, b) = x$ where $Ax = b$ and A is invertible. Since $g(A, b) = f(A^{-1}, b)$,

$$a_A \leq \|b\|, \quad \kappa_A \leq \|A\|\|A^{-1}\|, \quad a_b \leq \|A^{-1}\|, \quad \kappa_b \leq \|A\|\|A^{-1}\|. \tag{4.13}$$

Remark. It is a fortunate (and somewhat miraculous) coincidence that the relative condition number in every case is bounded by $\|A\|\|A^{-1}\|$. That is, whether we are performing matrix-vector multiplication or solving a linear system, whether we treat A, x, or b as the input, we get the same bound $\kappa \leq \|A\|\|A^{-1}\|$. Because of this, $\|A\|\|A^{-1}\|$ is called the condition number of the matrix A. In fact, our analysis does not actually require A to be invertible. We could instead use the Moore-Penrose inverse defined in Section 2.2.1 to get $\kappa_A = \|A\|\|A^{\dagger}\|$. In terms of the 2-norm this is $\kappa_A = \frac{\sigma_1}{\sigma_r}$ where σ_1 and σ_r are the largest and smallest singular values of A respectively.

4.3 Finite-Precision Arithmetic

Digital computers use finite-precision arithmetic. For scientific computing applications, this is typically implemented using IEEE floating point arithmetic; see, for example [3, 5, 14]. We do not go into much depth on finite-precision arithmetic here, although a basic understanding is essential to analyze the stability of numerical algorithms.

Any finite-precision arithmetic system necessarily has bounds on the largest magnitude it can represent. For IEEE double-precision arithmetic, this is approximately 10^{308}. That is, a double-precision floating-point number can represent a number in the range of $\pm 10^{308}$. More importantly, not every number in this range can be represented. For any x in the representable range of a floating point system, let $\mathrm{fl}(x)$ denote the best floating point representation. That is, $\mathrm{fl}(x)$ is the representable number that is closest to x. The standard model for floating-point arithmetic specifies that there exists a number u, called the unit roundoff, such that for any x in the representable range, there exists δ with $|\delta| \leq u$ such that

$$\mathrm{fl}(x) = (1 + \delta)x. \tag{4.14}$$

That is, the relative error in $\mathrm{fl}(x)$ is no more than u. If "op" denotes any of the standard operations

$+, -, \times, /$, then for x, y and $x \operatorname{op} y$ in the representable range,

$$\mathrm{fl}(x \operatorname{op} y) = (1+\delta)(x \operatorname{op} y), \tag{4.15}$$

where $|\delta| \leq u$. That is, the relative error in computing $x \operatorname{op} y$ is at most u. Any reasonable algorithm involves more than a single operation. Let

$$\gamma_k = \frac{ku}{1-ku}, \qquad \text{and} \qquad \tilde{\gamma}_k = \frac{cku}{1-cku}, \tag{4.16}$$

where c is a small integer constant independent of k. The following result shows how errors from multiple operations combine.

Lemma 4.10 (see [5, pp. 67]). *If* $|\delta| \leq \gamma_k$ *and* $|\epsilon| \leq \gamma_j$ *then* $(1+\delta)(1+\epsilon) = (1+\xi)$ *where* $|\xi| \leq \gamma_{k+j}$.

4.4 Stability

Whereas conditioning is a measure of the sensitivity of a function to perturbation, stability attempts to measure the susceptibility of an algorithm to roundoff. Of course, the roundoff that actually occurs depends on the inputs. Furthermore, it is possible for roundoff in different stages of an algorithm to cancel, yielding a result that is much more accurate than a stability analysis would suggest. The outline of this section follows [5] and [14].

A variable wearing a hat will denote a computed quantity. For example $\hat{f}(x)$ denotes the computed value of $f(x)$. Our primary concern is the *accuracy* of an algorithm. The *relative error* of an algorithm is

$$\frac{\|\hat{f}(x) - f(x)\|}{\|f(x)\|}. \tag{4.17}$$

The smaller the magnitude of the relative error, the more accurate the algorithm.

On a finite-precision machine, roundoff errors are unavoidable. In fact, since we may not even be able to precisely express the inputs, there is no hope for an algorithm to be very accurate if a problem is ill-conditioned. With an ill-conditioned problem, the slightest roundoff error in the input may result in large errors in the output, even if the algorithm is exact. Stability is a more realistic measure of the performance of an algorithm.

For a function $f : \mathbb{R}^n \to \mathbb{R}^m$, let $\hat{f}(x)$ represented the computed value of $f(x)$. Choose $\Delta x \in \mathbb{R}^n$ and $\Delta f \in \mathbb{R}^m$ satisfying

$$\hat{f}(x) = f(x + \Delta x) + \Delta f. \tag{4.18}$$

We call Δf the *forward error* and Δx is called the *backward error*.

The forward errors are what people generally think of when they think of computational errors. Forward errors are merely variation from the actual value. Backward errors are errors attributed to the input. Of course, there is no way of knowing where the errors actually occurred, nor is the decomposition generally unique. However, a decomposition clearly exists by setting $\Delta x = 0$ and $\Delta f = \hat{f}(x) - f(x)$.

Definition 4.11. An algorithm is *forward stable* if for fixed n and arbitrary $x \in \mathbb{R}^n$, there exists a forward error Δf and a backward error Δx satisfying (4.18) and

$$\frac{\|\Delta f\|}{\|f\|} = O(u), \qquad \frac{\|\Delta x\|}{\|x\|} = O(u), \tag{4.19}$$

as $u \to 0$.

Definition 4.12. An algorithm is *backward stable* if for fixed n and arbitrary $x \in \mathbb{R}^n$ there exists a backward error Δx such that $\hat{f}(x) = f(x + \Delta x)$ and

$$\frac{\|\Delta x\|}{\|x\|} = O(u), \tag{4.20}$$

as $u \to 0$.

Loosely speaking, a forward stable algorithm gives almost the correct answer to almost the correct question, whereas a backward stable algorithm gives exactly the correct answer to almost the correct question. Setting $\Delta f = 0$ we see that backward stability implies forward stability. Backward stability is significant for two reasons. First, backward stable algorithms are desirable because the answer is plausible. Consider the function $f(x) = e^x$. Suppose we have two algorithms for computing $f(x)$ and suppose that on the input $x = -10$, the first produces the result $\hat{f}(x) = -4.5 \times 10^{-05}$ and the second produces the output $\hat{f}(x) = 4.5 \times 10^{-4}$. The correct answer is approximately 4.5×10^{-5}, therefore, the relative error of the first algorithm is only 2, whereas the second algorithm has a relative error of 9. In fact, the answer is off by an order of magnitude. However, the quantity -4.5×10^{-5} does not make sense for the function $f(x) = e^x > 0$. In a physical problem, this quantity may not even have a meaningful interpretation. Meanwhile, $4.5 \times 10^{-4} \approx e^{-7.7}$ is at least the correct value of e^x for some value of x close to -10.

The second reason is that backward stability isolates the conditioning of the problem. It is possible for an algorithm for an ill-conditioned problem to be backward stable. This is because the large errors in the output may be attributed to small errors in the input. Thus, backward stability is a better measure of the algorithm, whereas conditioning measures the problem. In fact, we can make the interplay between conditioning, stability, and accuracy more precise.

Theorem 4.13 (see [14]). *Suppose the relative condition number of the function $f(x)$ is κ and an algorithm $\hat{f}$ for f is backward stable. Then for any x, the relative error of $\hat{f}$ satisfies*

$$\frac{\|\hat{f}(x) - f(x)\|}{\|f(x)\|} = O(\kappa u). \tag{4.21}$$

The theorem indicates that the accuracy of a backward stable algorithm is as good as the precision of the arithmetic system and the conditioning of the problem allow. We conclude our discussion of stability by stating two results that we shall use in our stability analysis. Recall that $|A|$ is the matrix whose (i, j) entry is $|A_{i,j}|$.

Theorem 4.14 (see [5, p. 63]). *Let x and y be vectors in $\mathbb{R}^n$. Then the computed inner product $\mathrm{fl}(x^T y)$ satisfies*

$$\mathrm{fl}(x^T y) = (x + \Delta x)^T y = x^T (y + \Delta y), \quad |\Delta x| \le \gamma_n |x|, \quad |\Delta y| \le \gamma_n |y|. \tag{4.22}$$

Theorem 4.15 (see [5, p. 361]). *For nonsingular $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$, solving the system $Ax = b$ using the QR factorization computed by the Householder algorithm satisfies the following* *error bounds for the computed quantity $\hat{x}$,*

$$(A + \Delta A)\hat{x} = b, \tag{4.23}$$

where

$$\|a_j\|_2 \leq \tilde{\gamma}_{n^2}\|a_j\|_2, \quad j = 1, \ldots, n. \tag{4.24}$$

Chapter 5. Transition Events

In this chapter we make precise the notion of a transition event. We define two types of transition events: cumulative transition events and time-average transition events. We then derive expressions for computing the expectation of these transition events in terms of the transition matrix T, a mask M, and the initial distribution μ.

A cumulative transition event describes the transient behavior of a reducible Markov chain. To simplify the analysis, we begin with the common case of an absorbing Markov chain, a specific case of reducible Markov chains. We then generalize the analysis to reducible Markov chains. In every case we provide examples of transition events.

5.1 Cumulative Events on Absorbing Chains

A mask is a matrix $M \in \mathbb{R}^{n \times n}$ that describes the weights assigned to the transitions of a Markov chain. Here $M_{i,j}$ is the weight assigned to the transition from s_j to s_i. The transition event for M is the random variable whose value on any realization is the sum of the mask entries,

$$Y_M = \sum_{k=0}^{\infty} M_{X_{k+1},X_k}. \tag{5.1}$$

Lemma 5.1. *Given $M \in \mathbb{R}^{n\times n}$,*

$$E_\mu M_{X_{k+1},X_k} = \sum_{i=1}^{n} \left[(M \odot T)T^k \mu\right]_i . \tag{5.2}$$

Proof. By total probability (see Theorem 3.5)

$$\begin{aligned} E_\mu M_{X_{k+1},X_k} &= \sum_{i,j=1}^{n} M_{i,j} P_\mu(X_{k+1} = s_i, X_k = s_j) \\ &= \sum_{i,j=1}^{n} M_{i,j} P_\mu(X_{k+1} = s_i | X_k = s_j) P_\mu(X_k = s_j) \\ &= \sum_{i,j=1}^{n} M_{i,j} T_{i,j} \left[T^k \mu\right]_j = \sum_{i=1}^{n} \left[(M \odot T)T^k \mu\right]_i . \end{aligned}$$

□

Let $\mathcal{A} \subset \mathcal{S}$ denote the absorbing states of X_k; that is, $s_j \in \mathcal{A}$ if $P(X_{k+1} = s_j \mid X_k = s_j) = 1$, or equivalently, $T_{j,j} = 1$. The Markov chain X_k is absorbing if $\mathcal{A} \neq \emptyset$ and there exists $k \in \mathbb{N}$ such that

$$P(X_k \in \mathcal{A} \mid X_0 = s_j) > 0, \qquad j = 1, \ldots, n. \tag{5.3}$$

In other words, an absorbing chain is a reducible chain in which the ergodic classes are single states; see Section 2.4. Without loss of generality, the transition matrix of an absorbing chain assumes the form

$$T = \begin{bmatrix} A_T & 0 \\ B_T & I \end{bmatrix}, \tag{5.4}$$

where $A_T \in \mathbb{R}^{t\times t}$ and $t = n - |\mathcal{A}|$ is the number of transient states. Thus, A_T and B_T are the transitions leaving the t transient states. In particular, the diagonal entries of A_T are strictly less than 1. Furthermore,

$$T^k = \begin{bmatrix} A_T^k & 0 \\ B_T \sum_{m=0}^{k-1} A_T^m & I \end{bmatrix}. \tag{5.5}$$

Lemma 5.2. *If T is the transition matrix of an absorbing Markov chain then the spectral radius of A_T satisfies $\rho(A_T) < 1$. Moreover, $(I - A_T)^{-1}$ exists and*

$$(I - A_T)^{-1} = \sum_{k=0}^{\infty} A_T^k. \tag{5.6}$$

Proof. Since there is a finite path with positive probability from any state to an absorbing state, it follows that for some $k \geq 0$, B_{T^k} has a nonzero entry in each column. That is, the block of transitions from the transient states to the absorbing states is nonzero for each transient state. Since T^k is stochastic, it follows that $\|A_{T^k}\|_1 = \|A_T^k\|_1 < 1$. Therefore, $\rho(A_T^k) < 1$, which implies that $\rho(A_T) < 1$. □

Lemma 5.3. *Let $M, T \in \mathbb{R}^{n \times n}$ be given, where T is the transition matrix of an absorbing Markov chain. If $M_{j,j} = 0$ whenever $s_j \in \mathcal{A}$ then*

$$\sum_{k=0}^{\infty} (M \odot T) T^k = (M \odot T) \begin{bmatrix} (I - A_T)^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.7}$$

Proof. If $s_j \in \mathcal{A}$ then $(M \odot T)_{i,j} = 0$ for $i = 1, \ldots, n$. Using the block form (5.4) for M,

$$M \odot T = \begin{bmatrix} A_M \odot A_T & 0 \\ B_M \odot B_T & 0 \end{bmatrix}. \tag{5.8}$$

Combining this with (5.5),

$$(M \odot T) T^k = \begin{bmatrix} (A_M \odot A_T) A_T^k & 0 \\ (B_M \odot B_T) A_T^k & 0 \end{bmatrix} = (M \odot T) \begin{bmatrix} A_T^k & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.9}$$

Hence,

$$\sum_{k=0}^{\infty} (M \odot T) T^k = (M \odot T) \sum_{k=0}^{\infty} \begin{bmatrix} A_T^k & 0 \\ 0 & 0 \end{bmatrix} = (M \odot T) \begin{bmatrix} (I - A_T)^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

□

Throughout the paper, let

$$Q = \begin{bmatrix} I - A_T & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad Q^- = \begin{bmatrix} (I - A_T)^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.10}$$

Note that Q^- satisfies $(I-T)Q^-(I-T) = (I-T)$ and $Q^-(I-T)Q^- = Q^-$ so Q^- is a (1,2)-inverse of $I - T$; see Section 2.2. However, it is not always the case that $((I - T)Q^-)^T = (I - T)Q^-$ or that $(Q^-(I-T))^T = Q^-(I-T)$. Hence, Q^- is not the Moore-Penrose inverse; it is not the Drazin inverse of $I - T$ since $I - T$ and Q^- do not necessarily commute. However, it is straightforward to show that Q^- is both the Moore-Penrose inverse and the Drazin inverse of Q.

Theorem 5.4. *Let $M, T \in \mathbb{R}^{n\times n}$ and $\mu \in \mathbb{R}^n$ be given, where T is the transition matrix of an absorbing Markov chain and μ is stochastic. Set $D = \operatorname{diag}(Q^-\mu)$. If $M_{j,j} = 0$ for all $s_j \in \mathcal{A}$ then the random variable* (5.1) *has expectation*

$$E_\mu Y_M = \operatorname{tr}(MDT^T). \tag{5.11}$$

Proof. Suppose that $M_{i,j} \geq 0$ for all i, j so that Y_M is an increasing series. Then by the Monotone Convergence Theorem, see Section 3.2, we may exchange the order of summation and expectation,

$$E_\mu Y_M = \sum_{k=0}^{\infty} E_\mu M_{X_{k+1},X_k}. \tag{5.12}$$

Applying Lemma 5.1,

$$E_\mu Y_M = \sum_{k=0}^{\infty}\sum_{i=1}^{n} \left[(M \odot T)T^k\mu\right]_i = \sum_{i=1}^{n} \left[\sum_{k=0}^{\infty}(M \odot T)T^k\mu\right]_i. \tag{5.13}$$

By Lemma 5.3 and Corollary 2.3,

$$E_\mu Y_M = \sum_{i=1}^{n} \left[(M \odot T)Q^-\mu\right]_i = \operatorname{tr}(MDT^T). \tag{5.14}$$

For general M, let Z be the random variable $Z = \sum_{k=0}^{\infty} |M_{X_{k+1},X_k}|$. For all $m \in \mathbb{N}$,

$$\left|\sum_{k=0}^{m} M_{X_{k+1},X_k}\right| \leq \sum_{k=0}^{\infty} |M_{X_{k+1},X_k}| = Z. \tag{5.15}$$

The nonnegative case indicates that $E_\mu|Z| = E_\mu Z < \infty$ so that the Dominated Convergence Theorem allows us to exchange the order of summation with expectation. The remainder of the argument is identical to the nonnegative case. □

Remark. Theorem 5.4 indicates that the condition $M_{j,j} = 0$ for $s_j \in \mathcal{A}$ is sufficient to guarantee that $E_\mu|Y_M| < \infty$. This condition is practically necessary; if $s_j \in \mathcal{A}$ satisfies $P_\mu(X_k = s_j) > 0$ for some $k \in \mathbb{N}$ then $M_{j,j} \neq 0$ implies that $E_\mu|Y_M| = \infty$. Thus, $M_{j,j} = 0$ is required of all absorbing states that are "reachable" from the initial distribution μ.

Example 5.5. Consider an object that moves between n states with transition probabilities $T_{i,j}$ and suppose that s_n is absorbing. Let $d(s_j, s_i)$ be the distance between s_j and s_i and set

$$M_{i,j} = \begin{cases} 0 & j = n \\ d(s_j, s_i) & \text{otherwise.} \end{cases} \tag{5.16}$$

The random variable (5.1) describes the distance traveled on any realization. If the initial position of the object has distribution μ then Theorem 5.4 indicates that the expected distance traveled is given by (5.11). Notice that the quantity depends on the transitions traversed, not the states visited, so that this event is most naturally a transition event.

5.2 Cumulative Events on Reducible Chains

We now generalize to a reducible Markov chain; see Section 2.4. We assume that the transition matrix T is in canonical form (2.52). We generalize the block form (5.4) for T to

$$T = \begin{bmatrix} A_T & 0 \\ B_T & E_T \end{bmatrix}, \tag{5.17}$$

where A_T and B_T correspond to the transient states and E_T is block diagonal containing the ergodic classes. Let $\mathcal{E}$ denote the set of ergodic states.

Theorem 5.6. *Set* $D = \operatorname{diag}(Q^{-}\mu)$. *For a reducible Markov chain* T, *if* $M_{i,j} = 0$ *whenever* s_i *and* s_j *are in the same ergodic class, then the random variable* (5.1) *has expectation*

$$E_\mu Y_M = \operatorname{tr}(MDT^T). \tag{5.18}$$

Proof. Since $\rho(T_{ii}) < 1$ for all the transient classes it follows that $\rho(A_T) < 1$ as in Lemma 5.2. The condition $M_{i,j} = 0$ for s_i and s_j in the same ergodic class guarantees the result of Lemma 5.3. With these results, the remainder of the proof is identical to the proof of Theorem 5.4. □

Example 5.7. Meyer showed that the following quantities may be obtained for absorbing chains using the Drazin inverse:

(i) The probability of being absorbed into state $s_i \in \mathcal{A}$ when initially in state $s_j \notin \mathcal{A}$.

(ii) The expected number of times the chain will be in state $s_i \notin \mathcal{A}$ when initially in state $s_j \notin \mathcal{A}$.

(iii) The expected number of steps until absorption when initially in state $s_j \notin \mathcal{A}$.

For general reducible chains, Meyer suggests representing each ergodic class by a single absorbing state and using the above results to determine the same quantities. We can express these quantities in terms of transition events. Furthermore, we may do so on any reducible chain without having to convert to an absorbing representation. For any ergodic class $\mathcal{E}_m$, let

$$M_{i,j} = \begin{cases} 1 & s_i \in \mathcal{E}_m,\ s_j \notin \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.19}$$

Then Y_M is 1 on any realization which enters $\mathcal{E}_m$ and zero elsewhere. Thus, $E_\mu Y_M$ is the probability of absorption into $\mathcal{E}_m$ which gives (i) for any reducible chain.

For (ii), given $s_h \notin \mathcal{E}$, let

$$M_{i,j} = \begin{cases} 1 & i = h,\ s_j \notin \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.20}$$

Then $E_\mu Y_M$ is the expected number of arrivals at state s_h given the initial distribution μ. Setting $M_{i,j} = 1$ when $j = h$ instead of $i = h$ gives the expected number of departures from state s_h. These quantities may differ depending on the initial distribution.

To find (iii) let

$$M_{i,j} = \begin{cases} 1 & s_j \notin \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.21}$$

Then $E_\mu Y_M$ is the expected number of steps until absorption into some ergodic class.

5.3 Time-Average Events

A *time-average transition event* is the average sum of the transition weights

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} M - X_{k+1}, X_k. \tag{5.22}$$

By Theorem 2.24, this limit converges to $G = I - SS^D$, where $S = I - T$.

Theorem 5.8. *Set $D = \operatorname{diag}(G\mu)$. Then for any stochastic T, the random variable*

$$Y_M = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} M_{X_{k+1},X_k} \tag{5.23}$$

has expectation

$$E_\mu Y_M = \operatorname{tr}(MDT^T). \tag{5.24}$$

Proof. For all $N \in \mathbb{N}$,

$$\frac{1}{N} \sum_{k=0}^{N} M_{X_{k+1},X_k} \leq 2 \max \{ |M_{i,j}| \mid 1 \leq i, j \leq n \} \tag{5.25}$$

so that we may apply the Dominated Convergence Theorem. This and the linearity of expectation give

$$
\begin{aligned}
E_\mu Y_m &= \lim_{N\to\infty} \frac{1}{N}\sum_{k=0}^{N} E_\mu M_{X_{k+1},X_k} \\
&= \sum_{i=1}^{n}\left[(M\odot T)\left(\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N}T^k\right)\mu\right]_i \\
&= \sum_{i=1}^{n}\left[(M\odot T)G\mu\right]_i = \operatorname{tr}(MDT^T).
\end{aligned} \tag{5.26}
$$

□

Remark. If T is reducible, the value of M on the transitions leaving transient states is irrelevant; the value of Y_M on any realization depends only on the ergodic class that is entered. Thus, Y_M represents the steady-state behavior of T in this case. For example, in the case of an absorbing chain

$$
E_\mu Y_M = \sum_{s_j\in\mathcal{A}} P_\mu(X_k \to s_j)M_{j,j}. \tag{5.27}
$$

If we fix $s_j \in \mathcal{A}$ and set $M_{j,j} = 1$ with all other entries zero, then $E_\mu Y_M$ is the probability of absorption into s_j given the initial distribution μ.

Example 5.9. Consider a hydrogen atom that is excited by an external energy source so that the atom's electron is perpetually changing energy states. Let $\{s_1,\dots,s_n\}$ be the various allowable energy levels and $T_{i,j}$ be the probability that the atom's electron moves from s_j to s_i. Also, let μ be the distribution on the electron's initial position. To determine the portion of light emitted by the hydrogen atom that is in a particular range, say the visible light range, we set $M_{i,j} = 1$ for any transition that emits visible light and $M_{i,j} = 0$ otherwise. Then the portion of light that is visible in any realization is the time-average random variable given by (5.23). Applying Theorem 5.8, the expected portion of visible light is given by (5.24).

Chapter 6. Composite Markov Chains

One of the motivating questions for this thesis was to determine the expected number of lead changes in a turn-based, competitive, stochastic system. Transition events are an excellent way to answer this question, but in order to describe the transition event, we need a Markov chain that describes a competitive system.

6.1 Construction

Let $T_1 \in \mathbb{R}^{n_1 \times n_1}$ and $T_2 \in \mathbb{R}^{n_2 \times n_2}$ be stochastic matrices and let $T = T_1 \otimes T_2 \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$ be the Kronecker product of T_1 and T_2; see Section 2.1. Recall that the entries of T may be labeled by $T_{(i_1,i_2),(j_1,j_2)}$, which represents the (i_2, j_2) entry of the (i_1, j_1) block of T and is equal to $T_{(i_1,i_2),(j_1,j_2)} = [T_1]_{i_1,j_1}[T_2]_{i_2,j_2}$.

Proposition 6.1. *If $T_1 \in \mathbb{R}^{n\times n}$ and $T_2 \in \mathbb{R}^{m\times m}$ are stochastic then so is $T = T_1 \otimes T_2$.*

Proof. It is immediate that T is nonnegative. To show that $\|T_j\|_1 = 1$ for every column T_j, we write T_j as $T_{(\cdot,\cdot),(j_1,j_2)}$ and then sum

$$\begin{aligned} \|T_{(\cdot,\cdot),(j_1,j_2)}\|_1 &= \sum_{i_1=1}^{n}\sum_{i_2=1}^{m} T_{(i_1,i_2),(j_1,j_2)} = \sum_{i_1=1}^{n}\sum_{i_2=1}^{m}[T_1]_{i_1,j_1}[T_2]_{i_2,j_2} \\ &= \sum_{i_1=1}^{n}[T_1]_{i_1,j_1}\sum_{i_2=1}^{m}[T_2]_{i_2,j_2} = \sum_{i_1=1}^{n}[T_1]_{i_1,j_1} = 1, \end{aligned} \tag{6.1}$$

since T_1 and T_2 are stochastic. □

Recall that the Kronecker product is associative. It follows that $T_1 \otimes \cdots \otimes T_p$ is also stochastic whenever $T_1, \ldots, T_p$ are stochastic. The stochastic matrix $T = T_1 \otimes T_2$ has an meaningful interpretation. If X_k is the Markov chain of $T_1 \in \mathbb{R}^{n\times n}$ with states $\{s_1, \ldots, s_n\}$ and Y_k is the Markov chain of $T_2 \in \mathbb{R}^m$ with states $\{t_1, \ldots, t_m\}$ then

$$T_{(i_1,i_2),(j_1,j_2)} = P(X_{k+1} = s_{i_1}, Y_{k+1} = t_{i_2} \mid X_k = s_{j_1}, Y_k = t_{j_2}). \tag{6.2}$$

Similarly, given stochastic $\mu_1 \in \mathbb{R}^n$ and $\mu_2 \in \mathbb{R}^m$, the vector $\mu = \mu_1 \otimes \mu_2 \in \mathbb{R}^{nm}$ is stochastic and the same indexing scheme applies:

$$P_\mu(X_0 = s_{i_1}, Y_0 = t_{i_2}) = \mu_{(i_1,i_2)}. \tag{6.3}$$

6.2 Lead Changes

Suppose T_0 is an absorbing Markov chain representing the progression of an agent to the absorbing "goal" state. Suppose further that the states are ordered such that higher indices represent being closer to winning. Then the p-wise Kronecker product $T = T_0^{\otimes p}$ represents competition between p players taking turns. It is natural to ask what the expected number of lead changes is.

For clarity, let $p = 2$. We count a lead change if on any turn a player comes from behind and ends in the lead. If a tie is either created or broken on a turn, we count a half a lead change. The mask for two-player lead changes is given by

$$M_{(i_1,i_2),(j_1,j_2)} = \begin{cases} 0 & s_{j_1} \in \mathcal{A} \text{ or } s_{j_2} \in \mathcal{A} \\ 1 & j_2 < j_1 \text{ and } i_2 > i_1 \\ 1 & j_2 > j_1 \text{ and } i_2 < i_1 \\ 1/2 & j_2 = j_1 \text{ and } i_2 \neq i_1 \\ 1/2 & j_2 \neq j_1 \text{ and } i_2 = i_1 \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

When $s_{j_1} \in \mathcal{A}$ the (i_1, j_1) block is zero, since the first agent has already won. For $s_{j_1} \notin \mathcal{A}$ the

(i_1, j_1) block is

$$M_{(i_1,\cdot),(j_1,\cdot)} = \begin{bmatrix} 0 & \dots & 0 & 1/2 & 1 & \dots & 1 & \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \\ 0 & \dots & 0 & 1/2 & 1 & \dots & 1 & \vdots \\ 1/2 & \dots & 1/2 & 0 & 1/2 & \dots & 1/2 & 0 \\ 1 & \dots & 1 & 1/2 & 0 & \dots & 0 & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \\ 1 & \dots & 1 & 1/2 & 0 & \dots & 0 & \end{bmatrix}. \tag{6.5}$$

When $p > 2$, there are at least two natural ways to define a lead change. The first is to count a lead change whenever the player in the lead is passed by another. We count a half lead change for breaking or establishing a tie in the leading position. The second way to extend lead changes for $p > 2$ is to count the permutations in the players positions. For example, if $j_1 > j_2 > \cdots > j_p$ and $i_1 < i_2 < \cdots < i_p$, then this complete lead change gets a weight of $M_{(i_1,\dots,i_p),(j_1,\dots,j_p)} = 1 + \cdots + p = p(p+1)/2$.

6.3 Competitive Advantage

There are other questions about competitive systems that can be answered using transition events. We address only $p = 2$, that is two-player systems. However, these examples have extensions for $p > 2$. Consider the mask

$$M_{(i_1,i_2),(j_1,j_2)} = \begin{cases} 0 & s_{j_1} \in \mathcal{A} \text{ or } s_{j_2} \in \mathcal{A}, \\ 1 & s_{i_1} \in \mathcal{A} \text{ and } s_{i_2} \notin \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

The cumulative transition event Y_M for this mask is the indicator event for player 1 winning. That is, $Y_M = 1$ whenever player 1 wins and 0 otherwise. Thus, $E_\mu Y_M$ is the probability of player 1 winning. Obviously, a similar mask gives the probability that player 2 wins.

Consider the mask

$$M_{(i_1,i_2),(j_1,j_2)} = \begin{cases} 0 & s_{j_1} \in \mathcal{A} \text{ or } s_{j_2} \in \mathcal{A}, \\ 1 & s_{i_1} \in \mathcal{A} \text{ and } s_{i_2} \in \mathcal{A}, \\ 0 & \text{ otherwise.} \end{cases} \tag{6.7}$$

The cumulative transition event Y_M for M is the indicator event for a tie. That is $Y_M = 1$ if player 1 and player 2 reach the absorbing state on the same turn and zero otherwise. $E_\mu Y_M$ is the probability of a tie.

In many turn-based competitive systems, a tie is broken by declaring player 1 the winner. In fact, in many games, if player 1 reaches the absorbing state, player 2 is not granted his last turn. In this case $E_\mu Y_M$ is the probability that player 1 wins merely for being first in the turn ordering and is a measure of the advantage of being the first player.

Chapter 7. Computation

In this chapter we characterize the conditioning of cumulative transition events on reducible chains. The conditioning of time-average events is similar; the only differences arise from the computation of the Drazin inverse, rather than the (1,2)-inverse for cumulative events. After addressing conditioning, we provide an algorithm for computing the expectation of cumulative transition events on reducible chains and discuss the complexity and stability of this algorithm.

7.1 Conditioning

Recall the definition of the relative condition number κ given in Section 4.2. In this section, we give bounds on κ for the function $f(T, M, \mu) = \operatorname{tr}(MDT^T)$ defined by (5.18). We treat this as three separate conditioning problems by analyzing the conditioning of f with respect to each input M, T, and μ individually. This affords an understanding of the sensitivity of (5.18) to perturbations in each input.

Although the one-norm is a natural choice for column-stochastic matrices, M and D are not stochastic and the trace in (5.18) corresponds more naturally to the Frobenius inner product on the space of matrices. Therefore, we give bounds on the condition number κ in terms of the Frobenius norm $\|A\|_F = \sqrt{\operatorname{tr}(A^TA)}$. By Cauchy-Schwarz, $|\operatorname{tr}(A^TB)| \leq \|A\|_F\|B\|_F$. Furthermore, the Frobenius norm satisfies the submultiplicative property, that is, $\|AB\|_F \leq \|A\|_F\|B\|_F$. Therefore,

$$|\operatorname{tr}(MDT^T)| = |\operatorname{tr}(T^TMD)| \leq \|T\|_F\|M\|_F\|D\|_F. \tag{7.1}$$

Theorem 7.1. *Set*

$$\kappa = \frac{\|M\|_F\|T\|_F\|(I - A_T)^{-1}\|_2}{|\operatorname{tr}(MDT^T)|}. \tag{7.2}$$

The relative condition numbers for the expectation of transition events have the following bounds:

$$\begin{aligned} \kappa_M &\leq \kappa, && \text{(7.3a)} \\ \kappa_T &\leq \kappa(1 + \|T\|_F\|(I - A_T)^{-1}\|_2), && \text{(7.3b)} \\ \kappa_\mu &\leq \kappa. && \text{(7.3c)} \end{aligned}$$

Proof. Recall from Theorem 5.6 and (5.10) that $D = \operatorname{diag}(\nu)$, where $\nu = Q^-\mu$. Since μ is stochastic and $\|\cdot\|_2 \leq \|\cdot\|_1$ we obtain the bound $\|\mu\|_2 \leq \|\mu\|_1 = 1$. Therefore,

$$\|D\|_F = \left(\sum_{i=1}^{n} d_{ii}^2\right)^{1/2} = \|\nu\|_2 = \|Q^-\mu\|_2 \leq \|(I - A_T)^{-1}\|_2. \tag{7.4}$$

For κ_M, fix T and μ and treat $f(M) = \operatorname{tr}(MDT^T)$ as a function of M only. We remark that D is independent of M. Therefore, for a perturbation δM of M we obtain

$$\begin{aligned} \|\delta f(M)\|_F &= |\operatorname{tr}((M + \delta M)DT^T) - \operatorname{tr}(MDT^T)| \\ &= |\operatorname{tr}(\delta MDT^T)| \leq \|\delta M\|_F\|T\|_F\|(I - A_T)^{-1}\|_2 \end{aligned}$$

by (7.1) and (7.4). Hence,

$$\lim_{\delta\to 0}\sup_{\|\delta M\|_F\le\delta}\frac{\|\delta f(M)\|_F}{\|\delta M\|_F}\le\|T\|_F\|(I-A_T)^{-1}\|_2. \tag{7.5}$$

Multiplying by $\|M\|_F/\|f(M)\|_F = \|M\|_F/|\operatorname{tr}(MDT^T)|$ we obtain (7.3a).

The matrix D depends on both T and μ. We denote by $D_{T+\delta T}$ and $D_{\mu+\delta\mu}$ the diagonal matrix obtained from $T+\delta T$ and $\mu+\delta\mu$, respectively, and use a similar notation for ν. A perturbation δT of T causes a perturbation in Q^-. If δA_T is the submatrix of δT corresponding to A_T, then

$$\begin{bmatrix}(I-A_T-\delta A_T)^{-1} & 0\\ 0 & 0\end{bmatrix}=(Q-\delta Q)^-, \tag{7.6}$$

where δQ is δA_T padded with zeros. In the limit as $\delta\to 0$, $\|\delta A_T\|_F\le\|\delta T\|_F\le\delta$ implies that the inverse $(I-A_T-\delta A_T)^{-1}$ exists. Note that

$$\nu=Q^-\mu=\begin{bmatrix}(I-A_T)^{-1} & 0\\ 0 & 0\end{bmatrix}\mu=\begin{bmatrix}\tilde\nu\\ 0\end{bmatrix}, \tag{7.7}$$

where $\tilde\nu\in\mathbb{R}^t$ is the transient portion of ν. If $\tilde\mu$ and $\delta\tilde\nu_T$ also represent the transient portions of μ and $\delta\nu_T$, respectively, then

$$(I-A_T-\delta A_T)(\tilde\nu+\delta\tilde\nu_T)=\tilde\mu. \tag{7.8}$$

Since $(I-A_T)\tilde\nu=\tilde\mu$, it follows that

$$\delta\tilde\nu_T=(I-A_T-\delta A_T)^{-1}\delta A_T\tilde\nu. \tag{7.9}$$

Consider $f(T)=\operatorname{tr}(MDT^*)$ as a function of T only, where M and μ are fixed. Then

$$\begin{aligned}\|\delta f(T)\|_F &= |\operatorname{tr}(MD_{T+\delta T}(T+\delta T)^T)-\operatorname{tr}(MDT^T)|\\ &\le |\operatorname{tr}(M(D_{T+\delta T}-D)T^T)|+|\operatorname{tr}(MD_{T+\delta T}\delta T^T)|\\ &\le \|M\|_F\|D_{T+\delta T}-D\|_F\|T\|_F+\|M\|_F\|D_{T+\delta T}\|_F\|\delta T\|_F,\end{aligned} \tag{7.10}$$

by (7.1). We require bounds on $\|D_{T+\delta T} - D\|_F$ and $\|D_{T+\delta T}\|_F$. In terms of ν we have $\|D_{T+\delta T} - D\|_F = \|\nu + \delta\nu_T - \nu\|_2 = \|\delta\nu_T\|_2$. Applying (7.9), and noting that $\|(Q-\delta Q)^-\|_2 = \|(I - A_T - \delta A_T)^{-1}\|_2$ and $\|\delta Q\|_2 = \|\delta A_T\|_2$,

$$\|\delta\nu_T\|_2 = \|(Q-\delta Q)^-\delta Q\nu\|_2 \le \|(I - A_T - \delta A_T)^{-1}\|_2\|\delta A_T\|_2\|\nu\|_2. \tag{7.11}$$

Clearly, $\|\delta A_T\|_2 \le \|\delta T\|_2 \le \|\delta T\|_F$. Combining this fact with (7.4) we obtain,

$$\|D_{T+\delta T} - D\|_F = \|\delta\nu_T\|_2 \le \|(I - A_T - \delta A_T)^{-1}\|_2\|(I-A_T)^{-1}\|_2\|\delta T\|_F. \tag{7.12}$$

We now turn our attention to $\|D_{T+\delta T}\|_F = \|\nu + \delta\nu_T\|_2 \le \|\nu\|_2 + \|\delta\nu_T\|_2$. Using (7.4) and (7.12),

$$\|D_{T+\delta T}\|_F \le \|(I-A_T)^{-1}\|_2 + \|(I - A_T - \delta A_T)^{-1}\|_2\|(I-A_T)^{-1}\|_2\|\delta T\|_F. \tag{7.13}$$

Putting (7.10), (7.12), and (7.13) together, we have

$$\begin{aligned}
&\text{(7.14a)} & \frac{\|\delta f(T)\|_F}{\|\delta T\|_F} &\le \|M\|_F\|(I-A_T)^{-1}\|_2 \\
&\text{(7.14b)} & &+ \|M\|_F\|T\|_F\|(I - A_T - \delta A_T)^{-1}\|_2\|(I-A_T)^{-1}\|_2 \\
&\text{(7.14c)} & &+ \|M\|_F\|(I - A_T - \delta A_T)^{-1}\|_2\|(I-A_T)^{-1}\|_2\|\delta T\|_F.
\end{aligned}$$

In the limit as $\delta \to 0$, (7.14c) is zero and $(I - A_T - \delta A_T)^{-1} = (I-A_T)^{-1}$ in (7.14b), hence

$$\lim_{\delta\to 0}\sup_{\|\delta T\|_F \le \delta} \frac{\|\delta f(T)\|_F}{\|\delta T\|_F} \le \|M\|_F\|(I-A_T)^{-1}\|_2\left(1 + \|T\|_F\|(I-A_T)^{-1}\|_2\right). \tag{7.15}$$

Multiplying by $\|T\|_F/\|f(T)\|_F = \|T\|_F/|\operatorname{tr}(MDT^T)|$ we obtain (7.3b).

Denote by $\delta\nu_\mu$ the change in ν due to a perturbation $\delta\mu$ of μ. This satisfies

$$\nu + \delta\nu_\mu = Q^-(\mu + \delta\mu). \tag{7.16}$$

By multiplying and canceling equal terms, we obtain

$$\delta\nu_\mu = Q^{-}\delta\mu. \tag{7.17}$$

We now consider $f(\mu) = \operatorname{tr}(MDT^T)$ as a function of μ, where M and T are fixed. Applying (7.1) we obtain

$$\|\delta f(\mu)\|_F = |\operatorname{tr}(MD_{\mu+\delta\mu}T^T) - \operatorname{tr}(MDT^T)| \leq \|M\|_F\|T\|_F\|\delta\nu_\mu\|_2. \tag{7.18}$$

Using (7.17) we have $\|\delta\nu_\mu\|_2 \leq \|(I - A_T)^{-1}\|_2\|\delta\mu\|_2$. Thus,

$$\lim_{\delta\to 0} \sup_{\|\delta\mu\|\leq\delta} \frac{\|\delta f(\mu)\|_F}{\|\delta\mu\|_2} \leq \|M\|_F\|T\|_F\|(I - A_T)^{-1}\|_2. \tag{7.19}$$

Since μ is stochastic, $\|\mu\|_F = \|\mu\|_2 \leq \|\mu\|_1 = 1$. Hence, multiplying (7.19) by $\|\mu\|_F/\|f(\mu)\|_F \leq 1/|\operatorname{tr}(MDT^T)|$ we obtain (7.3c). □

Since T is stochastic, $\|T\|_F \leq \sqrt{n}$. In all the examples given in chapters 5 and 6, $\|M\|_F$ is no more than order n^2. Therefore, the magnitude of κ depends primarily on two factors: $\|(I-A_T)^{-1}\|_2$ and $|\operatorname{tr}(MDT^T)|$. As $I - A_T$ becomes singular, $\|(I - A_T)^{-1}\|_2$ is unbounded. In this case, the conditioning may be poor, which is to be expected since the conditioning of the linear system $(I - A_T)\nu = \mu$ is also poor.

The conditioning may also be poor if $\operatorname{tr}(MDT^T)$ is close to zero, particularly when $\|M\|_F\|T\|_F\|(I-A_T)^{-1}\|_2$ is relatively large. As the trace is a summation, cancelation of large magnitude terms with opposite signs results in poor conditioning. However, in all the examples in chapters 5 and 6, M is nonnegative. Since D and T are always nonnegative, cancellation is not a problem in this case, although the order of summation may affect roundoff errors; see [5, p. 63].

Even when M is nonnegative, $\operatorname{tr}(MDT^T)$ may be small due to orthogonality. Recall that $\operatorname{tr}(A^TB)$ is the Frobenius inner product on $\mathbb{R}^{m\times n}$. Therefore, $\operatorname{tr}(MDT^T) = \operatorname{tr}(DT^TM) = \langle TD, M\rangle_F = \|TD\|_F\|M\|_F\cos\theta$ where θ is the angle between TD and M. If these matrices are nearly orthogonal, the condition number may be large. This orthogonality often results from measuring events that are very unlikely to occur.

The quadratic dependence on $\|T\|_F\|(I-A_T)^{-1}\|_2$ in the upper bound for κ_T is to be expected since $E_\mu Y_M = \operatorname{tr}(MDT^T)$ depends on T in two places: the product DT^T and the computation of ν.

7.2 Implementation

In this section we provide an algorithm for computing (5.18). Let $\tilde{\mu}$ and $\tilde{\nu}$ be the first t entries of μ and ν, respectively, where t is the number of transient states and Let M_j and T_j denote the j^{th} columns of M and T. Then (5.18) may be expressed as

$$\operatorname{tr}(MDT^T) = \sum_{i=1}^{n}[(M \odot T)\nu]_i = \sum_{j=1}^{t} \nu_j M_j^T T_j. \tag{7.20}$$

Algorithm 7.2. The following computes (7.20) for the inputs T, M, and μ where T is in canonical form (2.52).

(i) Solve $(I-A_T)\tilde{\nu} = \tilde{\mu}$ by forming the QR factorization of $(I-A_T)$ using Householder reflections; see, for example [5, 14].

(ii) Compute the first t columns of $R = TD$, where $D = \operatorname{diag}(\nu)$ by scaling the j^{th} column of T by ν_j.

(iii) Compute $\psi = \sum_{j=1}^{t} M_j^T R_j$.

We refer to Steps 1-2 as the *setup*. This portion of the algorithm depends only on T and μ. Furthermore, Step 3 depends only on M. If several transition events are to be determined for the same chain and initial distribution, the setup need only be computed once.

Remark. The matrix $I - A_T$ is invertible and diagonally dominant by columns. Gaussian Elimination on such a system requires no pivots and is stable [5]. However, the theoretical bounds for Gaussian Elimination are insufficient to provide satisfactory bounds for Algorithm 7.2 beyond $n \approx 2300$. It is well-known that Gaussian Elimination generally performs much better in practice than numerical analysis suggests [5]. This does not change the asymptotic complexity of Algorithm 7.2 but does improve the constants. A MATLAB implementation for absorbing chains which uses Gaussian Elimination is provided in Figure 7.1.

```
% T is the transition matrix
% mu is the initial distribution
% preserves spartsity if T is sparse

% Compute mask independent portion: R

n = size(T,1);
a = sum(diag(T == 1));
t = n - a;
nu = (speye(t)-T(1:t,1:t))\mu((1:t)');
R = spdiags([nu;zeros(a,1)],0,n,n)*T';

% now for any masks M1, M2,...
EY1 = full(sum(sum(M1.*R)));
EY2 = full(sum(sum(M2.*R)));
```

Figure 7.1: A MATLAB implementation of Algorithm 7.2 for absorbing chains in canonical form.

7.3 Complexity

Recall that $I - A_T \in \mathbb{R}^{t \times t}$, where t is the number of transient states. It is well known that the temporal complexity of Step 1 is $O(t^3)$ and the spatial complexity is $O(t^2)$; see, for example [3, 5, 14]. Steps 2 and 3 both have temporal and spatial complexity $O(nt)$. Therefore, the setup requires $O(t^3 + nt)$ time and $O(nt)$ space. Once the setup is completed, (7.20) may be computed in $O(nt)$ time and space for each mask representing a transition event.

7.4 Stability

In this section we give bounds on the backward errors introduced in the computation of Algorithm 7.2. Recall from Section 4.3 that u denotes the unit roundoff and

$$\gamma_k = \frac{ku}{1-ku}, \qquad \text{and} \qquad \tilde{\gamma}_k = \frac{cku}{1-cku}, \tag{7.21}$$

where c is a small integer constant independent of k. The following result from Section 4.3 has been reproduced as a reference.

Lemma 7.3 (see [5, pp. 67]). *If* $|\delta| \le \gamma_k$ *and* $|\epsilon| \le \gamma_j$ *then* $(1+\delta)(1+\epsilon) = (1+\xi)$ *where* $|\xi| \le \gamma_{k+j}$.

Theorem 7.4. *Given* T, M *and* μ *the value* $\hat{\psi}$ *computed by Algorithm 7.2 is the exact solution for* *the inputs $T + \Delta T, M + \Delta M$, and μ, where ΔT and ΔM satisfy the following column-wise bounds*

$$\|\Delta T_j\|_2 \le 2\sqrt{n}\tilde{\gamma}_{n^2}\|T_j\|_2, \qquad \textit{and} \qquad \|\Delta M_j\|_2 \le \frac{(1+2\sqrt{n})\tilde{\gamma}_{n^2}}{\sqrt{1-4\sqrt{n}\tilde{\gamma}_{n^2}}}\|M_j\|_2, \tag{7.22}$$

provided $1 - 4\sqrt{n}\tilde{\gamma}_{n^2} > 0$.

Proof. The computed solution obtained in Step 1 satisfies the following column-wise backward error bounds [5, p. 361]:

$$(I - A_T - \Delta A_T)\hat{\nu} = \tilde{\mu} + \Delta\tilde{\mu}, \quad \text{where } \|\Delta A_{Tj}\|_2 \le \tilde{\gamma}_{n^2}\|(I - A_T)_j\|_2, \quad 1 \le j \le t, \tag{7.23}$$

Since T_j is stochastic, $1 = \|T_j\|_1 \le \sqrt{n}\|T_j\|_2$, hence

$$\|(I - A_T)_j\|_2 \le 1 + \|T_j\|_2 \le (\sqrt{n}+1)\|T_j\|_2 \le 2\sqrt{n}\|T_j\|_2. \tag{7.24}$$

Setting

$$\Delta T = \begin{bmatrix} \Delta A_T & 0 \\ 0 & 0 \end{bmatrix}, \tag{7.25}$$

we obtain the bound

$$\|\Delta T_j\|_2 = \|\Delta A_{Tj}\|_2 \le 2\sqrt{n}\tilde{\gamma}_{n^2}\|T_j\|_2. \tag{7.26}$$

Since D is diagonal, the computation in Step 2 to produce the matrix $R = TD$ involves only a single multiplication in each entry of R. Therefore, the computed result satisfies,

$$\hat{R}_{i,j} = (1 + \delta_{i,j})\hat{\nu}_j T_{i,j}, \qquad |\delta_{i,j}| \le u, \quad 1 \le i \le n, 1 \le j \le t, \tag{7.27}$$

where $\delta_{i,j}$ is the relative error caused by roundoff in the multiplication $\hat{\nu}_j T_{i,j}$. Step 3 is the inner product of two $nt \times 1$ vectors: $(\text{vec } M)^T(\text{vec } \hat{R})$. The computed result satisfies the following bound on backward errors,

$$\hat{\psi} = \sum_{j=1}^{t} \sum_{i=1}^{n} (1+\epsilon_{i,j}) M_{i,j} \hat{R}_{i,j} = \sum_{j=1}^{t} \hat{\nu}_j \sum_{i=1}^{n} (1+\epsilon_{i,j})(1+\delta_{i,j}) M_{i,j} T_{i,j}, \tag{7.28}$$

where $\epsilon_{i,j}$ is the backward error of the (i,j) entry that results from the computation of the inner product and satisfies $|\epsilon_{i,j}| \leq \gamma_{nt}$. This error bound is independent of the order of summation; the bounds may be improved by a careful ordering of the terms [5, p. 63]. Since $|\delta_{i,j}| \leq u \leq \gamma_1$, Lemma 4.10 guarantees that $(1+\epsilon_{i,j})(1+\delta_{i,j}) = (1+\xi_{i,j})$ where $|\xi_{i,j}| \leq \gamma_{nt+1}$. To obtain (7.22), we require a perturbation ΔM satisfying

$$\sum_{i=1}^{n} (1+\xi_{i,j}) M_{i,j} T_{i,j} = \sum_{i=1}^{n} (M+\Delta M)_{i,j} (T+\Delta T)_{i,j}, \qquad 1 \leq j \leq t. \tag{7.29}$$

Recall that ΔT was fixed above when solving the system $(I - A_T)\tilde{\nu} = \tilde{\mu}$. Canceling the term $M_{i,j}T_{i,j}$ from the summation and regrouping,

$$\sum_{i=1}^{n} \left(\xi_{i,j} M_{i,j} T_{i,j} - M_{i,j} \Delta T_{i,j}\right) = \sum_{i=1}^{n} \Delta M_{i,j} (T_{i,j} + \Delta T_{i,j}), \qquad 1 \leq j \leq t, \tag{7.30}$$

Let ξ_j be the j^{th} column of the matrix $\xi = (\xi_{i,j})$. For each j, the left hand side of (7.30) is the scalar quantity

$$b_j = (\xi_j \odot M_j)^T T_j - \Delta T_j^T M_j, \tag{7.31}$$

where, $\xi_j \odot M_j$ is the Hadamard, or entry-wise product. The system (7.30) is equivalent to $(T_j + \Delta T_j)^T \Delta M_j = b_j$, which, for nonzero $T_j + \Delta T_j$, has as a solution

$$\Delta M_j = \frac{b_j}{\|T_j + \Delta T_j\|_2^2} (T_j + \Delta T_j). \tag{7.32}$$

Using our bound on ΔT, Cauchy-Schwarz guarantees

$$\begin{aligned} \|T_j + \Delta T_j\|_2^2 &= \|T_j\|_2^2 + 2T_j^T \Delta T_j + \|\Delta T_j\|_2^2 \geq \|T_j\|_2^2 - 2\|T_j\|_2 \|\Delta T_j\|_2 \\ &\geq \|T_j\|_2^2 - 4\sqrt{n}\tilde{\gamma}_{n^2} \|T_j\|_2^2 = (1 - 4\sqrt{n}\tilde{\gamma}_{n^2}) \|T_j\|_2^2 > 0, \end{aligned} \tag{7.33}$$

under the assumption $1 - 4\sqrt{n}\tilde{\gamma}_{n^2} > 0$. Therefore, the computed $\hat{\psi}$ is the exact solution (7.20) for the inputs $M + \Delta M, T + \Delta T$, and μ. In (7.26) we gave bounds for ΔT. By Cauchy-Schwarz,

$$\begin{aligned} \|\Delta M_j\|_2 &= \frac{|b_j| \|T_j + \Delta T_j\|_2}{\|T_j + \Delta T_j\|_2^2} \le \frac{|(\xi_j \odot M_j)^T T_j| + |M_j^T \Delta T_j|}{\|T_j + \Delta T_j\|_2} \\ &\le \frac{\gamma_{nt+1} \|M_j\|_2 \|T_j\|_2 + \|M_j\|_2 \|\Delta T_j\|_2}{\|T_j + \Delta T_j\|_2} \le \frac{\gamma_{nt+1} + 2\sqrt{n}\tilde{\gamma}_{n^2}}{\sqrt{1 - 4\sqrt{n}\tilde{\gamma}_{n^2}}} \|M_j\|_2, \end{aligned}$$

by (7.26) and (7.33) and the observation $\|T_j\|_2 \le \|T_j\|_1 = 1$, since T_j is stochastic. Finally, the bounds $t \le n-1$ and $n \ge 1$ imply that $nt+1 \le n^2$, so $\gamma_{nt+1} \le \gamma_{n^2} \le \tilde{\gamma}_{n^2}$ and we obtain (7.22). □

Remark. For fixed n, (7.22) simplifies to

$$\frac{(1 + 2\sqrt{n})\tilde{\gamma}_{n^2}}{\sqrt{1 - 4\sqrt{n}\tilde{\gamma}_{n^2}}} \le \frac{4\sqrt{n}\tilde{\gamma}_{n^2}}{1 - 4\sqrt{n}\tilde{\gamma}_{n^2}} = O(\sqrt{n}\tilde{\gamma}_{n^2}), \tag{7.34}$$

as $u \to 0$. The quantity ΔM_j obtained in (7.32) is the solution to the optimization problem

$$\begin{aligned} &\text{minimize} && \|\Delta M_j\|_2 \\ &\text{subject to} && (T_j + \Delta T_j)^T \Delta M_j = b_j. \end{aligned} \tag{7.35}$$

Chapter 8. Simulations

We conducted a numerical study by computing expectations and comparing them to a Monte Carlo simulation. We used the game Chutes and Ladders (or Snakes and Ladders), which is characterized by a substantial number of states (82) and exhibits a gradual drift towards the absorbing state combined with occasional large jumps. Furthermore, this game is a good illustration of composite Markov chains as discussed in §6. The MATLAB script used for computing expectations and the code for the simulations can be found in [7]. We simulated the following events in 100 million games and determined the sample mean for each. The results are summarized in Table 8.1.

8.1 Simulated Events

- Second-To-Last Square: This is the number of times that a player gets "stuck" on the second-to-last square. Remember that if a player spins a number that would place him beyond the last square, he forfeits his turn.
- Large Ladder Traversal: The number of times a player traverses the largest ladder from square 28 to square 84.
- Game Length: The number of turns in the game.

In addition to the above events the following were simulated for a two-player game.

- Lead Changes: The number of lead changes in the game as discussed in §6.
- First-player Advantage: This is the indicator event for the first player winning when both players finish on the same turn. In expectation, it is the probability that the first player wins by virtue of being the first player.
- First-player Win Frequency: This is the indicator event for the first player winning. In expectation, this is the probability that the first player wins.

We note that nearly every one of these events is a transition event. Game length for Chutes and Ladders has been studied using various techniques. However, the second-to-last square event, or its generalize to the turn forfeiture event, is most logically described by a transition event. The large ladder traversal event cannot be described solely in terms of states. We have shown how lead changes is also a transition event, since it depends on the ordering of the players at the beginning and end of a transition. Also, player advantage depends on knowing *when* the first player reaches the winning state, not just whether he does it first.

8.2 Simulation Results

As can be seen in Table 8.1, the expectation computed using the expression $\mathrm{tr}(MDT^T)$ agrees with the sample mean for at least three significant digits in every case. This is approximately what would be expected by the Central Limit Theorem after simulating each event $100,000,000 = (10^4)^2$

Table 8.1: Comparison of Monte Carlo simulations with computed expectations

Event	**Sample Mean**	**Computed** $\mathrm{tr}(MDT^T)$	**Computation Time (sec)**
Single-Player Events			
Setup			1.8(-3)
Second-To-Last Square	1.2954	1.2958	1.3(-4)
Large Ladder	0.5895	0.5896	1.0(-4)
Game Length	39.596	39.598	2.9(-4)
Two-Player Events			
Setup			2.5
Second-To-Last Square	1.1159	1.1166	8.1(-3)
Large Ladder	0.8181	0.8180	3.2(-2)
Game Length	26.513	26.513	3.1
Lead Changes	3.9679	3.9679	3.4
First-Player Advantage	0.0156	0.0156	6.2(-3)
First-Player Wins	0.5078	0.5078	1.4(-1)

times. The execution time for computing expectations, shown in the last column of the table, indicates that even moderately large problems can feasibly be solved using this approach; the 2-player Chutes and Ladders matrix has over 6500 rows. Parallelization would permit much larger problems, however, we expect that for large n, simulation will be faster, just as Monte Carlo integration is more efficient than quadrature for high-dimensional problems. For this problem, there may be some structure in the composite matrix $T = T_0^{\otimes p}$ that can be exploited to reduce the high-dimensional sum to a smaller problem, although we were unable to discover any.

References

[1] S. L. Campbell and C. D. Meyer, Jr. *Generalized inverses of linear transformations.* Dover Publications Inc., New York, 1991. Corrected reprint of the 1979 original.

[2] Henry P. Decell, Jr. and P. L. Odell. On the fixed point probability vector of regular or ergodic transition matrices. *J. Amer. Statist. Assoc.*, 62:600–602, 1967.

[3] James W. Demmel. *Applied numerical linear algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[4] Richard Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[5] Nicholas J. Higham. *Accuracy and stability of numerical algorithms.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2002.

[6] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis.* Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[7] Jeffrey Humpherys. http://math.byu.edu/˜jeffh/mathematics/games/chutes/index.htm. Markov simulation code, 2008.

[8] Frank Jones. *Lebesgue integration on Euclidean space.* Jones and Bartlett Publishers, Boston, MA, 1993.

[9] John G. Kemeny and J. Laurie Snell. *Finite Markov chains.* Springer-Verlag, New York, 1976. Reprinting of the 1960 original, Undergraduate Texts in Mathematics.

[10] Amy N. Langville and Carl D. Meyer. *Google's PageRank and beyond: the science of search engine rankings.* Princeton University Press, Princeton, NJ, 2006.

[11] Carl Meyer. *Matrix analysis and applied linear algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. With 1 CD-ROM (Windows, Macintosh and UNIX) and a solutions manual (iv+171 pp.).

[12] Carl D. Meyer, Jr. The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Rev.*, 17:443–464, 1975.

[13] P. L. Odell and H. P. Decell. On computing the fixed-point probability vector of ergodic transition matrices. *J. Assoc. Comput. Mach.*, 14:765–768, 1967.

[14] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.