# A Lanczos Procedure for Approximating Eigenvalues of Large Stochastic Matrices

by

William J. DeMeo

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Mathematics

New York University

Jan 1999

---

Professor Jonathan Goodman

To my parents, with love and appreciation.

# Acknowledgements

First, I thank my advisor, Professor Jonathan Goodman, for giving me the opportunity to work on this problem, and helping me arrive at the following exposition. Next, I wish to thank Professor Helena Frydman for rst sparking my interest in Markov chains by giving lucid descriptions of their many interesting properties and applications. I thank Professors James Demmel and Beresford Parlett, for answering many questions pertaining to this problem, and Professor Leslie Greengard for agreeing to review this paper.

Finally, I would like to thank my family, for their support of my interests in research (and everything else!), and especially my parents, Barbara and Ted Terry, and Benita and Bill De Meo.

Without them, this paper would not have been written.

# Abstract

A Lanczos Procedure for Approximating Eigenvalues of Large Stochastic Matrices

by

William J. DeMeo

Master of Science in Mathematics

New York University

Professor Jonathan Goodman, Chair

The rate at which a Markov chain converges to a given probability distribution has long been an active area of research. Well known bounds on this rate of convergence involve the subdominant eigenvalue of the chain's underlying transition probability matrix. However, many transition probability matrices are so large that we are unable to store even a vector of the matrix in fast computer memory. Thus, traditional methods for approximating eigenvalues are rendered useless.

In this paper we demonstrate that, if the Markov chain is reversible, and we understand the structure of the chain, we can derive the coefficients of the traditional Lanczos algorithm without storing a single vector. We do this by considering the variational properties of observables on the chain's state space. In the process we present the classical theory which relates the information contained in the Lanczos coefficients to the eigenvalues of the Markov chain.

# Contents

# List of Figures

# List of Tables

# Introduction

The rate at which a Markov chain converges to a given probability distribution has long been an active area of research. This is not surprising considering this problems relevance to the ar- eas of statistics, statistical mechanics, and computer science. Markov Chain Monte Carlo (MCMC) algorithms provide important examples. These algorithms come in handy when we encounter a complicated probability distribution from which we want to draw random samples. In statistical mechanics, we might wish to estimate the phase average of a function on the state space. Goodman and Sokal [6] examine Monte Carlo methods in this context. Examples from statistics occur in the Bayesian paradigm when we are forced to simulate an unwieldy posterior distribution (see, e.g., Geman and Geman

To implement the MCMC algorithm, we invent a Markov chain that converges to the desired distribution (this is often accomplished using the Metropolis algorithm described in Chapter 5). Realizations of the chain will eventually represent samples from this distribution. Sometimes "eventually" – meaning all but finitely many terms of the chain – is just not enough. We need more practical results. In particular, we want to know how many terms of the chain should be discarded before we are sampling from a distribution that is close (in total variation distance) to the distribution of interest. This is the purpose of bounding convergence rates for Markov chains.

Often the Markov chains encountered in this context satisfy a condition known in the physics literature as detailed balance. Probabilists call chains with this property reversible. This simply means that the chain has the same probability law whether time moves forward or backward.[1] In this paper, we consider the rate

[1]This is not a precise definition. In particular the chain must have started from its stationary

at which such chains converge to a *stationary distribution*.[2]

There are a number of different methods in common use for bounding convergence rates of Markov chains, and a good review of these methods with many references can be found in More recently developed methods, employing logarithmic Sobolev inequalities, are reviewed in Most of the bounds in common use involve the subdominant eigenvalue of the Markov chain's transition probability matrix, and thus require good approximations to such eigenvalues. In many applications, however, the transition probability matrix is so large that it becomes impossible to store even a single vector of the matrix in conventional computer memory. These so called out-of-core problems are not amenable to traditional eigenvalue algorithms[3] without modification. This paper develops such a modification for the Markov chain eigenvalue problem. In particular it develops a method for approximating the first few eigenvalues of a transition probability matrix when we know the general structure of the underlying Markov chain. The method does not require storage of large matrices or vectors. Instead we need only simulate the Markov chain, and conduct a statistical analysis of the simulation.

Here is a look at what follows. Section 2.1 contains a review of the relevant Markov chain theory. Readers conversant in the asymptotic theory of Markov chains might wish to at least skim Section 2.1, if only to become familiar with our notation. Section 2.2 describes functions on the state space of the Markov process. This section and Chapter 3 develop the context in which we formulate the new ideas of the paper. In the last section of Chapter 3, Section 3.3, we present the familiar Krylov subspace and explain why this represents our best approximation to

distribution. Full rigor is postponed until Section 2.1.

[2]This and other italicized terms are defined in Section 2.1.

[3]By "traditional eigenvalue algorithms" we refer to those found, for example, in Golub and Van Loan[5]. See also the book by Demmel [1] for a more recent discussion.

a subspace containing extremal eigenvectors of the transition probability matrix.[4] The first section of Chapter 4 describes the *Lanczos algorithm* for generating an orthonormal basis for the Krylov subspace. As it stands, this algorithm is useless for an out-of-core problem such as ours since, by definition of such problems, it requires too much data movement; all the computing time is spent swapping data between slow and fast memory (e.g. between the hard disk and cache). Therefore, we discuss alternatives to Lanczos and demonstrate that the *Lanczos coeficients* are readily available through simulations of the Markov chain, which fact allows us to avoid the standard algorithm altogether. Following this is a chapter describing the Metropolis algorithm used to produce a reversible stochastic matrix. It is here that we experiment with the procedure described in Section 4.2 and approximate the extremal eigenvalues of the matrix, without storing any of its vectors. Finally, Chapter 6 concludes the paper.

[4] or, more precisely, a similarity transformation of this matrix.

# Chapter 1

# Markov Chains

## 1.1 General Theory

This review of Markov chain theory can be found in any good probability text. The present discussion is most similar to that of Durrett [1], to which we refer the reader desiring greater detail.

### 1.1.1 The Basic Setup

Heuristically, a Markov chain is a stochastic process with a lack of memory property. Here this means that the future of the process, given its past behavior and its present state, depends only on its present state. This is the probabilistic analogue of a familiar property of classical particle systems. Given the position and velocities of all particles at time t, the equations of motion can be completely solved for the future evolution of the system. Thus, information describing the behavior of the process prior to time t is superuous. To be a bit more precise, if technical, we need the following definitions.

**Definition 1.1.1.** Let $(\mathcal{S}, \mathcal{S})$ be a measurable space. A sequence $X_n$, $n \geq 0$, of random variables taking values in $\mathcal{S}$ is said to be a Markov chain with respect to the filtration $\sigma(X_0, \ldots, X_n)$ if for all $B \in \mathcal{S}$,

$$P(X_{n+1} \in B \mid \sigma(X_0, \ldots, X_n)) = P(X_{n+1} \in B \mid \sigma(X_n)). \tag{1.1.1}$$

Equation (1.1.1) merely states that if we know the present location or state of $X_n$, then information about earlier locations or states is irrelevant for predicting $X_{n+1}$.

**Definition 1.1.2.** A function $p : S \times \mathcal{S} \to \mathbb{R}$ is said to be a *transition probability* if:

1. for each $x \in S$, $A \mapsto p(x, A)$ is a probability measure on $(S, \mathcal{S})$.
2. for each $A \in \mathcal{S}$, $x \mapsto p(x, A)$ is a measurable function.

We call $X_n$ a Markov chain with transition probabilities $p_n$ if

$$P(X_{n+1} \in B \mid \sigma(X_n)) = p_n(X_n, B) \tag{1.1.2}$$

The spaces $(S, \mathcal{S})$ that we encounter below are standard Borel spaces, so the existence of the transition probabilities follows from the existence of regular conditional probabilities on Borel spaces—a standard measure theory result (see e.g.[1]).

Suppose we are given an initial probability distribution $\mu$ on $(S, \mathcal{S})$ and a sequence $p_n$ of transition probabilities. We can define a consistent set of finite di-

mensional distributions by

$$P(X_j \in B_j, 0 \leq j \leq n) = \int_{B_0} \mu(dx_0) \int_{B_1} p_0(x_0, dx_1) \int_{B_2} p_1(x_1, dx_2) \cdots \int_{B_n} p_{n-1}(x_{n-1}, dx_n). \tag{1.1.3}$$

Furthermore, denote our probability space by

$$(\Omega, \mathcal{F}) = (S^\omega, \mathcal{S}^\omega), \quad \text{where} \omega = \{0, l, ...\}.$$

We call this *sequence space* and it is defined more explicitly by

$$S^\omega = \{(\omega_0, \omega_1, \dots) : \omega_i in S\} \quad \text{and} \quad \mathcal{S}^\omega = \sigma(\omega : \omega_i \in A_i \in \mathcal{S}).$$

The Markov chain that we will study on this space is simply $X_n(\omega) = \omega$, the coordinate maps. Then, by the Kolmogorov extension theorem, there exists a unique probability measure $P_\mu$ on $(\Omega, \mathcal{F})$ so that the $X_n(\omega)$ have finite dimensional distributions (1.1.3).

If instead of $\mu$, we begin with the initial distribution $\delta_x$, i.e., point mass at $x$, then we denote the probability measure by $P_x$. With such measures defined for each $x$, we can in turn define distributions $P_\mu$, given any initial distribution $\mu$, by

$$P_\mu(A) = \int \mu(dx) P_x(A).$$

That the foregoing construction—which, recall, was derived merely from an initial distribution $\mu$ and a sequence $p_n$ of transition probabilities—satisfies Definition (1.1.2) of a Markov chain is not obvious, and a proof can be found in [1].

To state the converse of the foregoing, if $X_n$ is a Markov chain with transition

probabilities $p_n$ and initial distribution $\mu$, then its finite dimensional distributions are given by (1.1.3). Proof of this is also found in [1].

Now that we have put the theory on a firm, if abstract, foundation, we can bring the discussion down to earth by making the forgoing a little more concrete. First, we specialize our study of Markov chains by assuming that our chain is *temporally homogeneous*, which means that the transition probabilities do not depend on time; i.e., $p_n(\omega_n, B) = p(\omega_n, B)$. (This is the stochastic analogue of a conservative system.) Next we assume that our state space $S$ is finite, and suppose for all states $i, j \in S$ that $p(i, j) \geq 0$, and $\sum_j p(i, j) = 1$ for all $i$. In this case, equation (1.1.2) takes a more intuitive form,

$$P(X_{n+1} = j \mid X_n = i) = p(i, j),$$

and our transition probabilities become

$$p(i, A) = \sigma_{j \in A} p(i, j).$$

If P is a matrix whose $(i, j)$ element is the transition probability $p(i, j)$ then P is a *stochastic matrix*; that is, a matrix with elements $p_{ij}$ satisfying

$$p_{ij} \geq 0, \quad \sum_j p_{ij} = 1, \quad (i, j = 1, 2, \ldots, d).$$

We also refer to P as the transition probability matrix.

Without loss of generality, we can further suppose our Markov chain is *irreducible*. This means that, for any states $i, j$, starting in state $i$ the chain will make a transition to state $j$ at some future time with positive probability. This state of

affairs is often described by saying that all states *communicate*. We lose no generality with this assumption because any *reducible* Markov chain can be factored into irreducible classes of states which can each be studied separately.

The final two conditions we place on the Markov chains considered below will cost us some generality. Nonetheless, there remain many examples of chains meeting these conditions and making the present study worthwhile. Furthermore, it may be the case that, with a little more work, we will be able to drop these conditions in future studies. The first condition is that the chain is *aperiodic*. If we let $I_x = \{n \geq 1 : p^n(x,x) > 0\}$, we call a Markov chain *aperiodic* if, for any state $x$, the greatest common divisor of $I_x$ is 1. The second assumption is that our chain is *reversible*. This characterization is understood in terms of the following definition.

**Definition 1.1.3.** A measure $\mu$ is called *reversible* if it satisfies

$$\mu(x)p(x,y) = \mu(y)p(y,x), \qquad \text{for all } x \text{ and } y.$$

We call a Markov chain *reversible* if its stationary distribution (defined in Section 1.1.2) is reversible.

### 1.1.2 A Convergence Theorem

In succeeding arguments, we use some results concerning the asymptotic behavior of Markov chains. These results require a few more definitions.

**Definition 1.1.4.** A measure $\pi$ is said to be a *stationary measure* if

$$\sum_x \pi(x)p(x,y) = \pi(x). \tag{1.1.4}$$

Equation (1.1.4) says $P_\pi(X_1 = y) = \pi(y)$, and by induction that $P_\pi(X_n = y) = \pi(y)$ for all $n \geq 1$. If $\pi$ is a probability measure, then we call $\pi$ a stationary distribution. It represents an equilibrium for the chain in the sense that, if $X_0$ has distribution $\pi$, then so does $X_n$ for all $n$.

When the Markov chain is irreducible and aperiodic, the distribution of the state at time $n$ converges pointwise to $\pi$ as $n \to \infty$, regardless of the initial state. It is convenient to state this convergence result in terms of the Markov chain's transition probability matrix P. Before doing so, we note that irreducibility of a Markov chain is equivalent to irreducibility (in the usual matrix theory sense) of its transition probability matrix. Furthermore, it turns out that a transition probability matrix of an aperiodic Markov chain falls into that class of matrices often called acyclic, but for simplicity we will call such stochastic matrices aperiodic. With this terminology, we can state the convergence theorem in terms of the transition probability matrix P.

**Theorem 1.1.5.** *Suppose* P *is irreducible, aperiodic, and has stationary distribution* $\pi$*. Then as* $n \to \infty$*,* $p^n(i, j) \to \pi(j)$*.*

The notation $p^n(i, j)$ means the $(i, j)$ element of the $n$th power of P.

A Markov chain whose transition probability matrix satisfies the hypotheses of Theorem 1.1.5 is called *ergodic*. If we simulate an ergodic chain for sufficiently many steps, having begun in any initial state, the final state is a sample point from a distribution that is close to $\pi$.

To make this statement more precise requires that we define "close."

**Definition 1.1.6.** Let $\pi$ be a probability measure on $S$. Then the *total variation*

*distance* at time $n$ with initial state $x$ is given by

$$\Delta_x(n) = \| \mathrm{P}^n(x, A) - \pi(A)\|_{TV} = \max_{A \in \mathcal{S}} | \mathrm{P}^n(x, A) - \pi(A)|.$$

In what follows, we will measure rate of convergence using the function $\tau_x(\epsilon)$, defined as the first time after which the total variation distance is always less than $\epsilon$. That is,

$$\tau_x(\epsilon) = \min\{m : \Delta_x(n) \leq \epsilon \text{ for all } n \geq m\}.$$

To begin our consideration of the connection between convergence rates of Markov chains and eigenvalues, we first note that an aperiodic stochastic matrix P has real eigenvalues $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d-1} \geq -1$, where $d = |S|$ is the dimension of the state space. For an ergodic chain, the rate of convergence to the stationary distribution $\pi$ is bounded by a function of the *subdominant* eigenvalue. By subdominant eigenvalue we mean that eigenvalue which is second largest in absolute value, and we denote this eigenvalue by $\lambda_{\max} = \max \lambda_1, |\lambda_{d-1}|$. The function bounding the rate of convergence of a Markov chain is given by the following theorem (log denotes the natural logarithm):

**Theorem 1.1.7.** *The quantity $\tau_x(\epsilon)$ satisfies*

1. $\tau_x(\epsilon) \leq (1 - \lambda_{\max})^{-1}(\log \pi(x)^{-1} + \log \epsilon^{-1})$;

2. $\max_{x \in S} \tau_x(\epsilon) \geq \frac{1}{2}\lambda_{\max}(1 - \lambda_{\max})^{-1} \log(2\epsilon)^{-1}$.

As this theorem shows, if we have an upper bound on the subdominant eigenvalue, then we have an upper bound on the function $\tau_x(\epsilon)$. In what follows, we will derive an approximation to the subdominant eigenvalue and supply error bounds.

Together, an approximation and error bounds for $\lambda_{\max}$ provide enough information to make Theorem 1.1.7 useful.

## 1.2 Functions on the State Space

Recall that $X_n(\omega) = \omega_n \in S$ denotes the state in which the Markov chain exists at time $n$. Suppose that $\Phi = \{\phi_1, \ldots, \phi_p\}$ is a collection of $p$ *observables*, or functions defined on the state space $S$. Furthermore, let these observables be real valued, $\phi_i : S \to \mathbb{R}$. It is often useful to assume that none of the observables is a constant function. Suppose now that the state space $S$ is finite with $d$ possible states. Then, since an observable is simply a map of the state space, we can think of each $\phi_i$ as a vector of $d$ real numbers—the $d$ values that it takes on at the different states.

Now assume the Markov chain is irreducible, and let $\pi$ denote its stationary distribution. If we start the chain from its stationary distribution—i.e., suppose $X_0$ has distribution $\pi$—then $X_n$ is a stationary process. Furthermore, for each $i$, $\phi_i(X_n)$ is a stationary stochastic process with *mean*

$$\mathrm{E}_\pi\, \phi_i = \sum_{x \in S} \pi(x)\phi_i(x)$$

and *autocovariance function*

$$\begin{aligned} \mathrm{C}_\pi(\phi_i(X_n), \phi_i(X_{n+s})) &= \mathrm{E}_\pi[(\phi_i(X_n) - \mathrm{E}_\pi\, \phi_i)(\phi_i(X_{n+s}) - \mathrm{E}_\pi\, \phi_i)] \qquad (1.2.1) \\ &= \sum_{x,y \in S} \mathrm{P}_\pi(X_n = x, X_{n+s} = y)(\phi_i(x) - \mathrm{E}_\pi\, \phi_i)(\phi_i(y) - \mathrm{E}_\pi\, \phi_i). \end{aligned}$$

By the definition of conditional probability, we can write (1.1.4) as follows:

$$\sum_{x,y\in S} \mathrm{P}_\pi(X_n = x)\,\mathrm{P}_\pi(X_{n+s} = y \mid X_n = x)(\phi_i(x) - \mathrm{E}_\pi\,\phi_i)(\phi_i(y) - \mathrm{E}_\pi\,\phi_i).$$

Equivalently,

$$\sum_{x,y\in S} \pi(x)p^s_{xy}(\phi_i(x) - \mathrm{E}_\pi\,\phi_i)(\phi_i(y) - \mathrm{E}_\pi\,\phi_i).$$

Here $p^s_{xy}$ denotes the element in row $x$ and column $y$ of $\mathrm{P}^s$, the $s$th power of the transition probability matrix. Similarly, we define the *cross-covariance* between the function $\phi_i$ at time $n$ and $\phi_j$ at time $n+s$ as

$$\begin{aligned}\mathrm{C}_\pi(\phi_i(X_n), \phi_j(X_{n+s})) &= \mathrm{E}_\pi[(\phi_i(X_n) - \mathrm{E}_\pi\,\phi_i)(\phi_j(X_{n+s}) - \mathrm{E}_\pi\,\phi_j)] \\ &= \sum_{x,y\in S} \pi(x)p^s_{xy}(\phi_i(x) - \mathrm{E}_\pi\,\phi_i)(\phi_j(y) - \mathrm{E}_\pi\,\phi_j). \qquad (1.2.2)\end{aligned}$$

Now let $\langle\Phi\rangle$ denote the matrix of mean vectors whose $j$th column is $\mathrm{E}_\pi\,\phi_j\mathbf{1}$, where $\mathbf{1} = (1,\ldots,1)^t$, and let $\Pi = \mathrm{diag}(\pi(\omega_1),\ldots,\pi(\omega_d))$ be the $d \times d$ diagonal matrix with stationary probabilities $\pi(\omega)$ on the main diagonal and zeros elsewhere. Finally, denoting by $\mathrm{C}(s)$ the $p \times p$ covariance matrix whose $(i,j)$ element is $\mathrm{C}_\pi(\phi_i(X_n), \phi_j(X_{n+s}))$, we have

$$\begin{aligned}\mathrm{C}(0) &= \mathrm{E}(\Phi(X_n) - \langle\Phi\rangle)(\Phi(X_n) - \langle\Phi\rangle)^t \\ &= (\Phi - \langle\Phi\rangle)^t\Pi(\Phi - \langle\Phi\rangle), \\ \mathrm{C}(s) &= \mathrm{E}(\Phi(X_n) - \langle\Phi\rangle)(\Phi(X_{n+s}) - \langle\Phi\rangle)^t \\ &= (\Phi - \langle\Phi\rangle)^t\Pi\,\mathrm{P}^s(\Phi - \langle\Phi\rangle).\end{aligned}$$

Below, we will also find it useful to have at our disposal a new matrix that is

*similar* to the transition probability matrix. We have in mind the matrix $\mathrm{M} = \Pi^{1/2}\,\mathrm{P}\,\Pi^{-1/2}$. As is easily verified, this allows us to write the covariance matrix as

$$\begin{aligned} \mathrm{C}(s) &= (\Phi - \langle\Phi\rangle)^t \Pi^{\frac{1}{2}t}\,\mathrm{M}^s\,\Pi^{\frac{1}{2}}(\Phi - \langle\Phi\rangle) \\ &= \Psi^t\,\mathrm{M}^s\,\Psi, \end{aligned} \tag{1.2.3}$$

where we have defined $\Psi = \Pi^{\frac{1}{2}}(\Phi - \langle\Phi\rangle)$. Recall that our main motivation is that for out-of-core problems traditional eigenvalue algorithms are inadequate. With this fact and the form (1.2.3) in mind, we will consider using the covariance of observables on the state space to implement the Rayleigh-Ritz procedure, which we describe below. This procedure requires that M be symmetric. As the next fact demonstrates, this need for symmetry is the reason we insist that the Markov chain be reversible.

**Fact.** The matrix M is symmetric if and only if the Markov chain is reversible (i.e., iff the process satisfies the *detailed balance* condition).

*Proof.*

$$\begin{aligned} \mathrm{M} \text{ is symmetric} \quad &\iff \quad (\Pi^{1/2}\,\mathrm{P}\,\Pi^{-1/2})^t = \Pi^{1/2}\,\mathrm{P}\,\Pi^{-1/2} \\ &\iff \quad \Pi^{-\frac{1}{2}t}\,\mathrm{P}^t\,\Pi^{\frac{1}{2}t} = \Pi^{1/2}\,\mathrm{P}\,\Pi^{-1/2} \\ &\iff \quad \mathrm{P}^t\,\Pi^t = \Pi\,\mathrm{P}\,. \end{aligned}$$

Elementwise, the final equality is $\pi_i p_{ij} = \pi_j p_{ji}$. According to Definition 1.1.3, this states that $p$ is a reversible measure. □

# Bibliography

[1] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, second edition, 1996.