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 μ 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_μ on $(\Omega, \mathcal{F})$ so that the $X_n(\omega)$ have finite dimensional distributions (1.1.3).

If instead of μ, we begin with the initial distribution δ_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_μ, given any initial distribution μ, by

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

That the foregoing construction—which, recall, was derived merely from an initial distribution μ 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 μ, 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 μ 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 π 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 π is a probability measure, then we call π a stationary distribution. It represents an equilibrium for the chain in the sense that, if X_0 has distribution π, 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 π 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* π*. 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 nth 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 π.

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

Definition 1.1.6. Let π 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 ϵ. 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 π 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 ϕ_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 π denote its stationary distribution. If we start the chain from its stationary distribution—i.e., suppose X_0 has distribution π—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 P^s, the sth power of the transition probability matrix. Similarly, we define the *cross-covariance* between the function ϕ_i at time n and ϕ_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 jth 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. □

Chapter 2

Invariant and Approximate Invariant Subspaces

2.1 Invariant Subspaces

Definition 2.1.1. A subspace $S \subseteq \mathbb{R}^n$ with the property that

$$x \in S \quad \Longrightarrow \quad \mathrm{M}\, x \in S$$

is said to be *invariant* for M.

Recall that, having chosen observables $\langle \Phi \rangle = (\phi_1, \ldots, \phi_p)$, we constructed the covariance matrix $\mathrm{C}(s) = \Psi^t \mathrm{M}\, \Psi$. If the column space of Ψ, which we denote by $\mathrm{ran}(\Psi)$, is an invariant subspace for M, the definition implies $\mathrm{M}\, \psi_j \in \mathrm{ran}(\Psi)$. That is, for each ψ_i there exists a vector t of coeicients such that $\mathrm{M}\, \psi_j = \sum_{i=1}^{p} t_i \psi_{ij}$. This is true for all j and, putting each vector of coefficients into a matrix T, we see that $\mathrm{M}\, \Psi = \Psi\, \mathrm{T}$. Conversely, $\mathrm{M}\, \Psi = \Psi\, \mathrm{T}$ implies that $\mathrm{M}\, \psi_j$ is a linear combination

of columns of Ψ, so $\mathrm{ran}(\Psi)$ is invariant. We have thus proved the following

Fact. The subspace $\mathrm{ran}(\Psi)$ is invariant for M if and only if there exists $\mathrm{T} \in \mathbb{R}^{p \times p}$ such that $\mathrm{M}\,\Psi = \Psi\,\mathrm{T}$.

Consequently,

Fact. $\lambda(\mathrm{T}) \subseteq \lambda(\mathrm{M})$.

Proof.

$$\begin{aligned} \lambda \in \lambda(\mathrm{T}) \quad &\iff \quad (\exists v \in \mathbb{R}^p)\, \mathrm{T}\, v = \lambda v \\ &\iff \quad \Psi\, \mathrm{T}\, v = \lambda \Psi v \qquad\qquad \iff \quad \mathrm{M}\, \Psi v = \lambda \Psi v \quad \iff \quad \lambda \in \lambda(M). \end{aligned}$$

The second equivalence follows from Fact 2.1. □

To see why Facts 2.1 and 2.1 are theoretically useful, consider the equation of Fact 2.1:

$$\begin{aligned} & \Psi\, \mathrm{T} = \mathrm{M}\, \Psi \\ \iff \quad & \Psi^t \Psi\, \mathrm{T} = \Psi^t\, \mathrm{M}\, \Psi \\ \iff \quad & \mathrm{T} = (\Psi^t \Psi)^t \Psi^t\, \mathrm{M}\, \Psi. \end{aligned}$$

In this form, we recognize that $\mathrm{T} = \mathrm{C}^{-1}(0)\, \mathrm{C}(1)$. That is, using only the covariance of observables on the state space, we can generate a matrix T with the property $\lambda(\mathrm{T}) \subseteq \lambda(\mathrm{M})$. Recalling that $\mathrm{M} = \Pi^{1/2}\, \mathrm{P}\, \Pi^{-1/2}$, we see that M is similar to our original stochastic matrix P, and thus $\lambda(M) = \lambda(P)$.

2.2 Approximate Invariant Subspace

We noted above that Facts 2.1 and 2.1 are theoretically useful. Speaking practically now, when choosing observables on the state space we may not be sure that they will satisfy the primary assumption underlying the two facts. Recall the assumption: $\mathrm{ran}(\Psi)$ is an invariant subspace for M where $\Psi = \Pi^{1/2}(\Phi - \langle\Phi\rangle)$. It may well be the case that there exists $\psi_j \in \mathrm{ran}(\Psi)$ such that $\mathrm{M}\,\psi_j \notin \mathrm{ran}(\Psi)$, thereby violating the assumption. Even if we are lacking an invariant subspace, however, for some applications it is reasonable to expect that observables can be chosen to provide at least an *approximate invariant subspace*, which is defined as follows:

Definition 2.2.1. If the columns of $\Psi \in \mathbb{R}^{d\times p}$ are independent and the norm of the *residual matrix* $\mathrm{E} = \mathrm{M}\,\Psi - \Psi\,\mathrm{T}$ is small for some $\mathrm{T} \in \mathbb{R}^{p\times p}$, then $\mathrm{ran}(\Psi)$ defines an *approximate invariant subspace*.

To see how an approximate invariant subspace can be useful for approximating eigenvalues of M, we recall a theorem from Golub and Van Loan [2].

Theorem 2.2.2. *Suppose* $\mathrm{M} \in R^{d\times d}$ *and* $\mathrm{T} \in \mathbb{R}^{p\times p}$ *are symmetric and let* $\mathrm{E} = \mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}$, *where* $\mathrm{Q} \in \mathbb{R}^{d\times p}$ *is orthonormal (i.e.,* $\mathrm{Q}^t\,\mathrm{Q} = \mathrm{I}$*). Then there exist* $\mu_1, \ldots, \mu_p \in \lambda(\mathrm{T})$ *and* $\lambda_1, \ldots, \lambda_p \in \lambda(\mathrm{M})$ *such that*

$$|\mu_k - \lambda_k| \leq \sqrt{2}\|\,\mathrm{E}\,\|_2, \qquad \textit{for } k = 1, \ldots, p.$$

If the subspace $\mathrm{ran}(\mathrm{Q})$ is an approximate invariant subspace, then the definition implies that there is a choice T rendering the error $\|\,\mathrm{E}\,\|_2$ small, and thus the eigenvalues of T provide a good approximation to those of M.

When considering the foregoing ideas, it is apparent that their application

presents new—but hopefully less prohibitive—problems. As these problems are the focus of the rest of the paper, now is a good time to examine them.

First, the preceding theorem assumes a matrix Q whose columns form an orthonormal basis for the approximate invariant subspace. For our problem, derivation of such a Q is tricky, and we must prepare for this.

Next, having an approximate invariant subspace at our disposal merely tells us that there exists some matrix T which makes the error $\| \mathrm{E} \|_2$ small. We must discover the form of such a T. Moreover, it is natural to seek that T which minimizes $\| \mathrm{E} \|_2$ for a given approximate invariant subspace.

Finally, in order to apply these ideas to realistic eigenvalue problems, we must find a practical way to generate a good approximate invariant subspace.

We now address each of these issues in turn.

Recall the matrix $\Psi = \Pi^{1/2}(\Phi\langle\Phi\rangle)$. The columns of this matrix, though independent (by choice of independent observables), are not necessarily orthonormal. However, consider the polar decomposition $\Psi = \mathrm{Q}\,\mathrm{Z}$, where $\mathrm{Q}^t\,\mathrm{Q} = \mathrm{I}$ and $\mathrm{Z}^2 = \Psi^t\Psi$ is a symmetric positive semidefinite matrix.[1] Note that $\mathrm{Z} = (\Psi^t\Psi)^{1/2}$ is nonsingular, so Q has the form

$$\mathrm{Q} = \Psi\,\mathrm{Z}^{-1} = \Psi(\Psi^t\Psi)^{-1/2},$$

and it is clear that $\mathrm{ran}(\mathrm{Q}) = \mathrm{ran}(\Psi)$. Perhaps $\mathrm{ran}(\mathrm{Q})$ is a useful approximation to the invariant subspace for M. If $\mathrm{ran}(Q)$ is not itself invariant, we have the error matrix $\mathrm{E} = \mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}$. Below we show that the T which minimizes $\| \mathrm{E} \|_2$ is $\mathrm{T} = \mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q}$. This yields the following

[1] Recall that the polar decompostion is derived from the singular value decomposition, $\Psi = \mathrm{U}\,\Sigma\,\mathrm{V}^t$, by letting $\mathrm{Q} = \mathrm{U}\,\mathrm{V}^t$ and $\mathrm{Z} = \mathrm{V}\,\Sigma\,\mathrm{V}^t$.

Theorem 2.2.3. *If $\Psi = \mathrm{Q}\,\mathrm{Z}$ is the polar decomposition of Ψ, then the matrix*

$$\mathrm{T} = \mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q} = (\Psi^t\Psi)^{-1/2}\Psi^t\,\mathrm{M}^t\,\Psi(\Psi^t\Psi)^{-1/2}$$

minimizes $\|\,\mathrm{E}\,\|_2 = \|\,\mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}\,\|_2$.

Proof. We prove the result by establishing the following

Claim: Given $\mathrm{M} \in \mathbb{R}^{d\times d}$ suppose $Q \in \mathbb{R}^{d\times d}$ satisfies $\mathrm{Q}^t\,\mathrm{Q} = \mathrm{I}$. Then,

$$\min_{\mathrm{T}\in\mathbb{R}^{p\times p}} \|\,\mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}\,\|_2 = \|(\mathrm{I} - \mathrm{Q}\,\mathrm{Q}^t)\,\mathrm{M}\,\mathrm{Q}\,\|_2,$$

and $\mathrm{T} = \mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q}$ is the minimizer. The claim is verified by an easy application of the Pythagorean theorem: For any $\mathrm{T} \in \mathbb{R}^{p\times p}$ we have

$$\begin{aligned}
\|\,\mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}\,\|_2^2 &= \|\,\mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q} + \mathrm{Q}\,\mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}\,\|_2^2 \\
&= \|(\mathrm{I} - \mathrm{Q}\,\mathrm{Q}^t)\,\mathrm{M}\,\mathrm{Q} + \mathrm{Q}(\mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q} - \mathrm{T})\|_2^2 \\
&= \|(\mathrm{I} - \mathrm{Q}\,\mathrm{Q}^t)\,\mathrm{M}\,\mathrm{Q}\,\|_2^2 + \|\,\mathrm{Q}(\mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q} - \mathrm{T})\|_2^2 \qquad (2.2.1) \\
&= \|(\mathrm{I} - \mathrm{Q}\,\mathrm{Q}^t)\,\mathrm{M}\,\mathrm{Q}\,\|_2^2
\end{aligned}$$

Equality (2.2.1) holds since $\mathrm{I} - \mathrm{Q}\,\mathrm{Q}^t$ projects $\mathrm{M}\,\mathrm{Q}$ onto the subspace orthogonal to ran(Q). Thus, the two terms in the expression on the right are orthogonal, and the Pythagorean theorem yields equality. The concluding inequality establishes that the minimizing T is that which annihilates the second term in (2.2.1), that is, $\mathrm{T} = \mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q}$.

The second equality in Theorem 2.2.3 is a consequence of the polar decomposition, in which $\mathrm{Q} = \Psi\,\mathrm{Z}^{-1} = \Psi(\Psi^t\Psi)^{-1/2}$. Thus, $\mathrm{T} = (\Psi^t\Psi)^{-1/2}\Psi^t\,\mathrm{M}^t\,\Psi(\Psi^t\Psi)^{-1/2}$

is the minimizer, as claimed. □

2.2.1 The Krylov Subspace

Suppose the columns of a matrix $\mathrm{Q} \in \mathbb{R}^{d\times p}$ give an orthonormal basis for an approximate invariant subspace. Then, as we have seen,

1. $\mathrm{T} = \mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q}$ minimizes $\|\,\mathrm{E}\,\|_2 = \|\,\mathrm{M}\,\mathrm{Q} - \mathrm{Q}\,\mathrm{T}\,\|_2$ and

2. $\exists \mu_1, \ldots, \mu_p \in \lambda(\mathrm{T})$ and $\lambda_1, \ldots, \lambda_p \in \lambda(\mathrm{M})$ such that

$$\|\mu_k - \lambda_k\| \leq \sqrt{2}\|\,\mathrm{E}\,\|_2, \qquad \text{for } k = 1, \ldots, p.$$

Given an approximate invariant subspace ran(Q) of dimension p, these facts tell us what matrix we should use to approximate p elements of the spectrum of M. Now all we lack is a description of ran(Q). That is, we have not specified which approximate invariant subspace would best suit our objective of approximating the subdominant eigenvalue $\lambda_{\max}(\mathrm{M})$. For this purpose the following definition is useful:

Definition 2.2.4. The *Raleigh quotient* of a symmetric matrix M and a nonzero vector x is

$$\rho(x, \mathrm{M}) = \frac{x^t\,\mathrm{M}\,x}{x^t x}.$$

We will denote Raleigh quotient by $\rho(x)$ when the context makes clear what matrix is involved.

To find the approximate invariant subspace most appropriate for our problem, we choose each dimension successively, providing justification at each step. We start with one observable ϕ on the state space, and let $\psi_1 = \Pi^{1/2}(\phi - \mathrm{E}_\pi\,\phi)$. That

is, ψ_1 is a centered (mean zero) observable whose ith coordinate is weighted by $\sqrt{\pi(i)}$. Notice that the definition (which comes from the definition of Ψ following Equation (1.2.3)) is such that ψ_1 is a constant vector if and only if ϕ is oonstant on the state space, in which case ψ_1 is the constant zero function. (Observables that are constant on the state space are not interesting.) Second, recall that the row sums of the matrix P are all one, therefore the eigenvector corresponding to the eigenvalue $\lambda_0(\mathrm{P}) = 1$ is the constant vector. By definition of $\mathrm{M} = \Pi^{1/2}\,\mathrm{P}\,\Pi^{-1/2}$, we see that the constant vector is also the eigenvector corresponding to λ_0. This will play an important role in what follows, as it allows us to focus on the subdominant eigenvalue $\lambda_{\max}(\mathrm{M}) = \max\{\lambda_1(\mathrm{M}), |\lambda_{d-1}(\mathrm{M})|\}$ rather than on $\lambda_0(M)$ (which we already know is 1).

Now, notice that

$$|\rho(\psi_1, \mathrm{M})| \leq \max |\rho(x, \mathrm{M})| = \lambda_{\max}(\mathrm{M}) \tag{2.2.2}$$

where the max is taken over all nonconstant vectors. Since our interest centers on $\lambda_{\max}(\mathrm{M})$, we would like a ψ_1 that makes the left hand side of (2.2.2) large. This would be achieved if ψ_1 were to lie in the space spanned by, say, the first two eigenvectors of the Markov chain (sometimes referred to as the *slowest modes* of the process). However, this subspace spans only two dimensions of the entire d-dimensional space, and it is more likely that ψ_1 only comes close, at best, to lying in the subspace of interest. Now, given ψ_1, a judicious choice for the second dimension ψ_2, and hence $\Psi_2 = [\psi_1\ \psi_2]$, would be that which makes $\max_{a \neq 0} |\rho(\Psi_2 a)|$ as large as possible. To establish that this is indeed the right objective, note the

following:

$$\begin{aligned}\max_{a\neq 0}|\rho(\Psi_2 a,\mathrm{M})| &= \max_{a\neq 0}\left|\frac{a^t\Psi_2^t\,\mathrm{M}\,\Psi_2 a}{a^t\Psi_2^t\Psi_2 a}\right| \\ &= \max_{x\in\mathrm{ran}(\Psi_2)}|\rho(x,\mathrm{M})| \\ &= \max|\rho(x,\mathrm{M})| \\ &= \lambda_{\max}(\mathrm{M})\end{aligned} \tag{2.2.3}$$

Again, the max on the right side of (2.2.3) is over nonconstant vectors. In other words, we wish to chose $\Psi_2 = [\psi_1\ \psi_2]$ so that there is a vector $a \in \mathbb{R}^2$ making $|\rho(\Psi_2 a, \mathrm{M})|$ close to $\lambda_{\max}(\mathrm{M})$.

Now, $\rho(\psi_1)$ changes most rapidly in the direction of the gradient $\nabla\rho(\psi_1)$.

$$\nabla\rho(\psi_1) = \left(\frac{\partial\rho(\psi_1)}{\partial\psi_1(1)},\ldots,\frac{\partial\rho(\psi_1)}{\partial\psi_1(d)}\right) = \frac{2}{\psi_1^t\psi_1}(\mathrm{M}\,\psi_1 - \rho(\psi_1)\psi_1). \tag{2.2.4}$$

So, to maximize the left hand side of (2.2.3), Ψ_2 should be chosen so that the subspace $\mathrm{ran}(\Psi_2)$ contains the gradient vector. That is, we must choose ψ_2 so that

$$\nabla\rho(\psi_1) \in \mathrm{ran}\{\psi_1,\psi_2\} = \mathrm{ran}(\Psi_2). \tag{2.2.5}$$

Clearly, Equation (2.2.4) implies $\nabla\rho(\psi_1) \in \mathrm{ran}\{\psi_1, \mathrm{M}\,\psi_1\}$. Thus, if $\mathrm{ran}\{\psi_1,\psi_2\} = \mathrm{ran}\{\psi_1, \mathrm{M}\,\psi_1\}$, then (2.2.5) is satisfied. In general, having chosen $\Psi_k = [\psi_1\ \ldots\ \psi_k]$ so that $\mathrm{ran}\{\psi_1,\psi_2,\ldots,\psi_k\} = \mathrm{ran}\{\psi_1, \mathrm{M}\,\psi_1,\ldots,\mathrm{M}\,\psi_k\}$, we must chose ψ_{k+1} so that for any nonzero vector $a \in \mathbb{R}^k$,

$$\nabla\rho(\Psi_k a) \in \mathrm{ran}\{\psi_1,\psi_2,\ldots,\psi_k\}. \tag{2.2.6}$$

Now,

$$\nabla\rho(\Psi_k a) == \frac{2}{a^t \Psi_k^t \Psi_k a}(\mathrm{M}\,\Psi_k - \rho(\Psi_k a)\Psi_k a).$$

and therefore,

$$\nabla\rho(\Psi_k a) \in \mathrm{ran}(\mathrm{M}\,\Psi_k) \cup \mathrm{ran}(\Psi_k) = \mathrm{ran}\{\psi_1, \mathrm{M}\,\psi_1, \ldots, \mathrm{M}^k\,\psi_1\}.$$

Thus, the requirement (2.2.6) is satisfied when

$$\mathrm{ran}\{\psi_1, \psi_2, \ldots, \psi_k\} = \mathrm{ran}\{\psi_1, \mathrm{M}\,\psi_1, \ldots, \mathrm{M}^k\,\psi_1\}.$$

In conclusion, the p-dimensional approximate invariant subspace that is most suitable for our objective is

$$\mathcal{K}(\mathrm{M}, \psi_1, p) = \mathrm{ran}\{\psi_1, \mathrm{M}\,\psi_1, \ldots, \mathrm{M}^{p-1}\,\psi_1\}.$$

This is known as the *Krylov subspace*. Therefore, to answer the problem posed at the outset of this section, if we take the columns of Q to be an orthonormal basis for $\mathcal{K}(\mathrm{M}, \psi_1, p)$, then the eigenvalues of $\mathrm{T} = \mathrm{Q}^t\,\mathrm{M}\,\mathrm{Q}$ should provide good estimates of p extremal eigenvalues of M.

Bibliography

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

[2] G. Golub and Charles Van Loan. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.