Using the Lanczos Algorithm to Approximate Eigenvalues of Large Stochastic Matrices

William J. De Meo

December 16, 1997

Abstract

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 conventional computer memory. Thus, traditional methods for approximating eigenvalues are rendered useless. In particular, Krylov subspace methods (e.g. the Lanczos algorithm) require that we can at least perform matrix vector multiplication.

In this paper we demonstrate that, if the Markov chain is reversible, and we understand the structure of the chain, we can exploit this structure to derive the usual Lanczos coefficients 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 develop the classical theory which relates the information contained in the Lanczos coefficients to the eigenvalues of the Markov chain.

1 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 problem's relevance to the areas 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 (Geman and Geman [4]).

To implement the MCMC algorithm, we invent a Markov chain that converges to the desired distribution (this is often accomplished using the Metropolis algorithm) 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 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 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 [8]. More recent methods, employing logarithmic Sobolev inequalities, are reviewed in [2]. Most of the bounds in common use involve the *sub-dominant*, or second largest, eigenvalue of the Markov chain's *transition probability matrix*. In many applications, however, these matrices are so large that it becomes impossible to store even a single vector of

[1] This is not a precise definition. In particular the chain must have started from its stationary distribution. Full rigor is postponed until Section 2.

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

the matrix in conventional computer memory. These so called *out-of-core* problems and 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 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, if only to become familiar with our notation. Sections 3 and 4 develop the context in which we formulate the new ideas of the paper, and conclude with the statement and proof a key result (Proposition 4.1) on which these ideas are based. Section 5 presents the *Krylov subspace* and explains why this represents our best approximation to a subspace containing extremal eigenvectors of the transition probability matrix. Section 6 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. the hard disk and the cache). Therefore, we discuss an alternative to the usual Lanczos algorithm and demonstrate that the *Lanczos coefficients* are readily available through simulations of the Markov chain, which fact allows us to avoid the standard algorithm altogether.

2 Markov Chains

2.1 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 [3], to which we refer the reader desiring greater detail.

2.2 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 superfluous. To be a bit more precise, if technical, we need the following definitions.

Definition 2.1 Let $(S, \mathcal{S})$ be a measurable space. A sequence X_n, $n \geq 0$, of random variables taking values in S is said to be a *Markov chain* with respect to the filtration $\sigma(X_0, \ldots, X_n)$ if for all $B \in \mathcal{S}$,

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

Equation (1) merely states that, given the present location of X_n, the past is irrelevant for predicting the location of X_{n+1}.

Definition 2.2 A function $p : S \times \mathcal{S} \to \mathbf{R}$ is said to be a *transition probability* if:

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

2. for each $A \in \mathcal{S}$, $x \to p(x, A)$ is a measurable function.

We call X_n a Markov chain with transition probabilities p_n if

$$P(X_{n+1} \in B \mid \sigma(X_n)) = p_n(X_n, B) \tag{2}$$

[3]By "traditional algorithms" we refer to those found, for example, in Golub and Van Loan[5]. For a more recent, "state-of-the-art" discussion, consult Demmel's new book [1], and http://www.netlib.org for the software implementations.

Let us make the foregoing a bit 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. (This is the stochastic analogue of a conservative system.) Next, the Markov chain that we will study on our probability space is simply $X_n(\omega) = \omega_n$, the coordinate maps. We further assume that our state space (outcome space) S is finite, and suppose $p(i,j) \geq 0$, and $\sum_j p(i,j) = 1$ for all i. Intuitively, we write equation (2) as

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

and define our transition probabilities by

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

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

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

The matrix P, which contains the transition probabilities of the Markov chain, is aptly labeled 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, thus 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 2.3 A measure μ is called *reversible* if it satisfies

$$\mu(x)p(x,y) = \mu(y)p(y,x)$$

for all x and y.

We call a Markov chain *reversible* if its *stationary distribution* (defined in Section 2.3) is reversible.

2.3 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 2.4 A measure π is said to be a *stationary measure* if

$$\sum_x \pi(x)p(x,y) = \pi(x) \tag{3}$$

Equation 3 says $P_\pi(X_1 = y) = \pi(y)$. Induction further implies $P_\pi(X_n = y) = \pi(y)$ 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 in mind, we can state the convergence theorem in terms of the transition probability matrix.

Theorem 2.1
Suppose P is irreducible, aperiodic, and has stationary distribution π. Then as $n \to \infty$,

$$p^n(i,j) \to \pi(j)$$

A Markov chain whose transition probability matrix satisfies the hypotheses of Theorem 2.1 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 which is close to π. To make this statement more precise requires that we define "close".

Definition 2.5 Let π be a probability measure on S. The the *total variation distance* at time n with initial state x is given by

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

The rate of convergence is often measured using the function $\tau_x(\epsilon)$, 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 consider the connection between convergence rates and eigenvalues, 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. Furthermore, P is ergodic if and only if $\lambda_{d-1} > -1$. 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 $\lambda_{\max} = \max\{\lambda_1, |\lambda_{d-1}|\}$. The function bounding the rate of convergence is given by the following theorem.

Theorem 2.2
The quantity $\tau_x(\epsilon)$ satisfies

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

As this theorem shows, if we have an upper bound on the subdominant eigenvalue, then we have an upper bound on the function $\tau_x(\epsilon)$.

3 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, so $\phi_i : S \to \mathbf{R}$. It is often useful, as we will see below, 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

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

and *autocovariance function*

$$
\begin{aligned}
\mathbf{C}_\pi\left(\phi_i(X_n), \phi_i(X_{n+s})\right) &\equiv \mathbf{E}_\pi\left[\left(\phi_i(X_n) - \mathbf{E}_\pi \phi_i\right)\left(\phi_i(X_{n+s}) - \mathbf{E}_\pi \phi_i\right)\right] \\
&= \sum_{x,y \in S} \mathbf{P}_\pi(X_n = x, X_{n+s} = y)\left(\phi_i(X_n) - \mathbf{E}_\pi \phi_i\right)\left(\phi_i(X_{n+s}) - \mathbf{E}_\pi \phi_i\right)
\end{aligned} \tag{4}
$$

which, by the definition of conditional probability, we can write as,

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

$$
= \sum_{x,y \in S} \pi(x) p_{xy}^s \left(\phi_i(X_n) - \mathbf{E}_\pi \phi_i\right)\left(\phi_i(X_{n+s}) - \mathbf{E}_\pi \phi_i\right)
$$

Here p_{xy}^s denotes the element in row x and column y of P^s, the sth power of the transition probability matrix. Similarly, we define the *covariance* between the function ϕ_i at time n and ϕ_j at time $n+s$ as

$$
\begin{aligned}
\mathbf{C}_\pi\left(\phi_i(X_n), \phi_j(X_{n+s})\right) &\equiv \mathbf{E}_\pi\left[\left(\phi_i(X_n) - \mathbf{E}_\pi \phi_i\right)\left(\phi_j(X_{n+s}) - \mathbf{E}_\pi \phi_j\right)\right] \\
&= \sum_{x,y \in S} \pi(x) p_{xy}^s \left(\phi_i(X_n) - \mathbf{E}_\pi \phi_i\right)\left(\phi_j(X_{n+s}) - \mathbf{E}_\pi \phi_j\right)
\end{aligned}
$$

Now let $\langle \Phi \rangle$ denote the matrix of mean vectors whose jth column is $\mathbf{E}_\pi \phi_j \mathbf{1}$, 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 $(i,j)th$ element is $\mathbf{C}_\pi\left(\phi_i(X_n), \phi_j(X_{n+s})\right)$, we have

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

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 $\mathrm{M} \equiv \Pi^{1/2} \mathrm{P} \Pi^{-1/2}$. As is easily verified, this allows us to write the covariance matrix as

$$
\begin{aligned}
\mathrm{C}(s) &= (\Phi - \langle \Phi \rangle)^t \Pi^{\frac{1}{2}t} \mathrm{M}^s \Pi^{\frac{1}{2}} (\Phi - \langle \Phi \rangle) \\
&\equiv \Psi^t \mathrm{M}^s \Psi
\end{aligned} \tag{5}
$$

where we have defined $\Psi \equiv \Pi^{1/2}(\Phi - \langle \Phi \rangle)$. Recall that our main motivation is that, for out-of-core problems, traditional eigenvalue algorithms are inadequate. With this fact, and the form (5) in mind, we will consider using the covariance of observables on the state space to implement the (otherwise useless) *Rayleigh-Ritz* procedure, which we describe below. This procedure requires that M be symmetric. As the next fact demonstrates, this required symmetry is the reason for our interest in the reversibility property of the Markov chain.

Fact 3.1 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:

$$
\begin{array}{rrcl}
 & \multicolumn{3}{l}{\mathrm{M} \text{ is symmetric}} \\
\Leftrightarrow & (\Pi^{1/2} \mathrm{P} \Pi^{-1/2})^t & = & \Pi^{1/2} \mathrm{P} \Pi^{-1/2} \\
\Leftrightarrow & (\Pi \mathrm{P})^t & = & \Pi \mathrm{P}
\end{array}
$$

Element-wise, the final equality implies that $\pi_i p_{ij} = \pi_j p_{ji}$. According to Definition 2.3, this states that p is a reversible measure.

4 Invariant and Approximate Invariant Subspaces

Definition 4.1 A subspace $S \subseteq \mathbf{R}^n$ with the property that

$$x \in S \Rightarrow \mathrm{M}x \in S$$

is said to be *invariant* for M.

Recall that, having chosen observables $\Phi = (\phi_1, \ldots, \phi_p)$, we constructed the covariance matrix $\mathrm{C}(s) = \Psi^t \mathrm{M}^s \Psi$. 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_j there exists a vector t of coefficients 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 4.1 The subspace $\mathrm{ran}(\Psi)$ is invariant for M if and only if there exists $\mathrm{T} \in \mathbf{R}^{p\times p}$ such that $\mathrm{M}\Psi = \Psi\mathrm{T}$.

Consequently,

Fact 4.2 $\lambda(\mathrm{T}) \subseteq \lambda(\mathrm{M})$

Facts 4.1 and 4.2 are theoretically useful. To see why, write the equation of fact 4.1 as follows:

$$\begin{aligned} \Psi\mathrm{T} &= \mathrm{M}\Psi \\ \Psi^t\Psi\mathrm{T} &= \Psi^t\mathrm{M}\Psi \\ \mathrm{T} &= (\Psi^t\Psi)^{-1}\Psi^t\mathrm{M}\Psi \end{aligned}$$

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 $\lambda(\mathrm{M}) = \lambda(\mathrm{P})$.

We qualified the foregoing by stating that the facts are only theoretically useful. This is because when choosing observables of the state space we aren't sure that they will satisfy the primary assumption underlying facts 4.1 and 4.2. 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 is a $\psi_j \in \mathrm{ran}(\Psi)$ such that $\mathrm{M}\psi_j \notin \mathrm{ran}(\Psi)$, thereby violating the assumption. However, lacking an invariant subspace, we might hope that our observables at least provide an *approximate invariant subspace*.

Definition 4.2 If the columns of $\Psi \in \mathbf{R}^{d\times p}$ are independent and the *residual matrix* $\mathrm{E} \equiv \mathrm{M}\Psi - \Psi\mathrm{T}$ is small for some $\mathrm{T} \in \mathbf{R}^{p\times p}$, then $\mathrm{ran}(\Psi)$ defines an *approximate invariant subspace.*

To see how an approximate invariant subspace can be useful in approximating eigenvalues of M, we extract a theorem from Golub [5].

Theorem 4.1
Suppose $\mathrm{M} \in \mathbf{R}^{d\times d}$ *and* $\mathrm{T} \in \mathbf{R}^{p\times p}$ *are symmetric and let* $\mathrm{E} \equiv \mathrm{MQ} - \mathrm{QT}$, *where* $\mathrm{Q} \in \mathbf{R}^{d\times p}$ *satisfies* $\mathrm{Q}^t\mathrm{Q} = \mathrm{I}$. *Then there exist* $\mu_1, \ldots, \mu_p \in \lambda(\mathrm{T})$ *and* $\lambda_1, \ldots, \lambda_p \in \lambda(\mathrm{M})$ *such that*

$$|\mu_k - \lambda_k| \leq \sqrt{2}||\mathrm{E}||$$

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 preceding ideas, it becomes apparent that their application will present new – but hopefully less prohibitive – problems. As these problems summarize 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 $||E||_2$ small. We must discover the form of such a T. Moreover, it is natural to seek that T which minimizes $||E||_2$ for a given approximate invariant subspace. Finally, in order to apply these ideas to real examples of our eigenvalue problem, we must find a practical way to generate the most appropriate approximate invariant subspace. We now address each of these issues in the order raised.

Recall the matrix $\Psi \equiv \Pi^{1/2}(\Phi - \langle\Phi\rangle)$. The columns of this matrix, though independent (by choice of independent observables), are not necessarily orthonormal. However, consider the polar decomposition $\Psi = \mathrm{QZ}$ where $\mathrm{Q}^t\mathrm{Q} = \mathrm{I}$ and $\mathrm{Z}^2 = \Psi^t\Psi$ is symmetric positive semidefinite.[4] 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 ran(Q) = ran(Ψ). Perhaps ran(Q) is a useful approximation to the invariant subspace for M. If ran(Q) is not itself invariant, we have the error E = MQ − QT. It is easy to show, as we do in the following proof, that the T which minimizes $||E||_2$ is $\mathrm{T} = \mathrm{Q}^t\mathrm{MQ}$. This yields the following,

Proposition 4.1 *If* $\Psi = \mathrm{QZ}$ *is the polar decomposition of* Ψ*, then the matrix*

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

minimizes $||\mathrm{E}||_2 = ||\mathrm{MQ} - \mathrm{QT}||_2$

Proof:
We prove the result in two steps. First, we verify the following
Claim: Given $\mathrm{M} \in \mathbf{R}^{d\times d}$ suppose $\mathrm{Q} \in \mathbf{R}^{d\times p}$ satisfies $\mathrm{Q}^t\mathrm{Q} = \mathrm{I}$. Then,

$$\min_{T\in\mathbf{R}^{p\times p}} ||\mathrm{MQ} - \mathrm{QT}||_2 = ||(\mathrm{I} - \mathrm{QQ}^t)\mathrm{MQ}||_2$$

and $\mathrm{T} = \mathrm{Q}^t\mathrm{MQ}$ is the minimizer.
To verify the claim, observe the following application of the Pythagorean theorem:
For any $\mathrm{T} \in \mathbf{R}^{p\times p}$ we have

$$\begin{aligned} ||\mathrm{MQ} - \mathrm{QT}||_2^2 &= ||\mathrm{MQ} - \mathrm{QQ}^t\mathrm{MQ} + \mathrm{QQ}^t\mathrm{MQ} - \mathrm{QT}||_2^2 \\ &= ||(\mathrm{I} - \mathrm{QQ}^t)\mathrm{MQ}||_2^2 + ||\mathrm{Q}(\mathrm{Q}^t\mathrm{MQ} - \mathrm{T})||_2^2 \qquad (6) \\ &\geq ||(\mathrm{I} - \mathrm{QQ}^t)\mathrm{MQ}||_2^2 \end{aligned}$$

Equality (6) holds since $\mathrm{I} - \mathrm{QQ}^t$ projects MQ onto the subspace orthogonal to ran(Q). Thus, the two terms in the expression on the right are orthogonal, and the Pythagorean theorem yields the equality. The concluding inequality establishes that the minimizing T is that which annihilates the second term in (6); i.e. $\mathrm{T} = \mathrm{Q}^t\mathrm{MQ}$.

Having established the claim, proof of Proposition 4.1 proceeds by mere definition chasing: recall, the polar decomposition produced

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

Therefore,

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

is the minimizer, as we set out to prove.

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

5 The Optimal Subspace

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

1. $T = Q^t M Q$ minimizes $\|E\|_2 = \|MQ - QT\|_2$ and
2. $\exists\, \mu_1, \ldots, \mu_p \in \lambda(T)$ and $\lambda_1, \ldots, \lambda_p \in \lambda(M)$ such that

$$|\mu_k - \lambda_k| \le \sqrt{2}\|E\|$$

   for $k = 1, \ldots, p$.

Given an approximate invariant subspace $\mathrm{ran}(Q)$ of dimension p, these facts tell us what matrix we should use to approximate p elements of M's spectrum. Now all we lack is a description of $\mathrm{ran}(Q)$. That is, we have not specified which approximate invariant subspace would best suit our objective. To do so requires the following

Definition 5.1 The *Raleigh quotient* of a symmetric matrix M and a nonzero vector x is $\rho(x, M) \equiv (x^t M x)/(x^t x)$.

We will also denote this 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. Thus, start with one observable ϕ on the state space, and let $\psi_1 = \Pi^{1/2}(\phi - \mathbf{E}_\pi \phi)$. That is, ψ_1 represents a centered (mean zero) observable whose ith coordinate is weighted by $\sqrt{\pi(i)}$. First note that

$$\rho(\psi_1, M) \le \max_{x\neq 0} \rho(x, M) = \lambda_{\max}(M) \tag{7}$$

and recall that our interest centers on the first few elements of the spectrum of M, and especially on the subdominant eigenvalue. Therefore, we would like a ψ_1 which makes the left hand side of (7) 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, at best, only comes close to lying in the subspace of interest. Now, given ψ_1, a judicious choice for the second dimension ψ_2, and hence $\Psi_2 \equiv [\psi_1\ \psi_2]$, would be that which makes $\max_{a\neq 0} \rho(\Psi_2 a)$ is as large as possible. To establish that this is indeed the pertinent objective, note the following:

$$\max_{a\neq 0} \rho(\Psi_2 a, M) = \max_{x\in \mathrm{ran}(\Psi_2)} \rho(x, M) \le \max_{x\neq 0} \rho(x, M) = \lambda_{\max}(M)$$

In words, we wish to chose $\Psi_2 = [\psi_1\, \psi_2]$ so that there is a vector $a \in \mathbf{R}^2$ making $\rho(\Psi_2 a, M)$ close to $\lambda_{\max}(M)$.

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

$$\begin{aligned}\nabla\rho(\psi_1) &\equiv \left(\frac{\partial\rho(\psi_1)}{\partial\psi_1(1)}, \ldots, \frac{\partial\rho(\psi_1)}{\partial\psi_1(d)}\right)\\ &= \frac{2}{\psi_1^t\psi_1}(M\psi_1 - \rho(\psi_1)\psi_1) \end{aligned}\tag{8}$$

So 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

$$\nabla\rho(\psi_1) \in \mathrm{ran}\{\psi_1, \psi_2\} \equiv \mathrm{ran}(\Psi_2) \tag{9}$$

Clearly, equation (8) implies $\nabla\rho(\psi_1) \in \mathrm{ran}\{\psi_1, M\psi_1\}$. Therefore, if $\mathrm{ran}\{\psi_1, \psi_2\} = \mathrm{ran}\{\psi_1, M\psi_1\}$, then (9) is satisfied.

In general, having chosen $\Psi_k \equiv [\psi_1\ \cdots\ \psi_k]$ so that

$$\mathrm{ran}\{\psi_1, \psi_2, \ldots, \psi_k\} = \mathrm{ran}\{\psi_1, M\psi_1, \ldots, M^{k-1}\psi_k\}$$

we must chose ψ_{k+1} so that, for any non-zero vector $a \in \mathbf{R}^k$,

$$\nabla\rho(\Psi_k a) \in \mathrm{ran}\{\psi_1, \psi_2, \ldots, \psi_{k+1}\} \tag{10}$$

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, requirement (10) is met when $\mathrm{ran}\{\psi_1, \psi_2, \ldots, \psi_{k+1}\} = \mathrm{ran}\{\psi_1, \mathrm{M}\psi_1, \ldots, \mathrm{M}^k\psi_1\}$.

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

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

This is known as the *Krylov subspace*. Therefore, in answer to 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 $\mathrm{T} = \mathrm{Q}^t\mathrm{MQ}$ are our best estimates of the p largest eigenvalues of M.

6 The Lanczos Algorithm

The following will review the usual procedure for generating an orthonormal basis for $\mathcal{K}(\mathrm{M}, \psi_1, p)$, and conclude by showing that, for our problem, we can find the matrix $\mathrm{Q}^t\mathrm{MQ}$ without actually carrying out this procedure. Note that this conclusion is essential since the procedure requires matrix vector multiplication - an operation we have assumed impossible for large enough d.

6.1 The General Lanczos Algorithm

Let $\psi^{(k)} \equiv \mathrm{M}^k\psi$ and consider the $d \times d$ matrix

$$\Psi \equiv [\psi_1, \mathrm{M}\psi_1, \ldots, \mathrm{M}^{d-1}\psi_1] \equiv [\psi^{(0)}, \psi^{(1)}, \ldots, \psi^{(d-1)}]$$

We notice that $\mathrm{M}\Psi = [\psi^{(1)}, \ldots, \psi^{(d-1)}, \mathrm{M}^d\psi_1]$, and assuming for the moment that Ψ is non-singular, we can compute the vector $h = -\Psi^{-1}M^d\psi_1$. Thus,

$$\mathrm{M}\Psi = \Psi[e_2, \ldots, e_d, -h] \equiv \Psi\mathrm{H}$$

Equivalently,

$$\mathrm{H} \equiv \begin{pmatrix} 0 & 0 & \cdots & 0 & -h_1 \\ 1 & 0 & \cdots & 0 & -h_2 \\ & \ddots & \ddots & \vdots & \vdots \\ & & \ddots & 0 & -h_{d-1} \\ & & & 1 & -h_d \end{pmatrix} = \Psi^{-1}\mathrm{M}\Psi \tag{11}$$

H is a *companion matrix* which means that its characteristic polynomial is $p(x) = x^d + \sum_{i=1}^d h_i x^{i-1}$. Since H is similar to M, finding the eigenvalues of M is equivalent to finding the roots of $p(x)$. However, this is of little practical use since finding h, constructing $p(x)$, and finding its roots is probably a harder problem than the one we started with. Instead, the value of decomposition (11) derives from its upper Hessenberg form. We exploit this property below.

Let $\Psi = \mathrm{QR}$ be the QR decomposition of Ψ. Since Ψ is assumed nonsingular,

$$\Psi_d^{-1}\mathrm{M}\Psi_d = (\mathrm{R}^{-1}\mathrm{Q}^t)\mathrm{M}(\mathrm{QR}) = \mathrm{H}$$

Therefore,

$$\mathrm{Q}^t\mathrm{MQ} = \mathrm{RHR}^{-1} \equiv \mathrm{T}$$

Since R and R^{-1} are both upper triangular and H is upper Hessenberg, $T = RHR^{-1}$ is also upper Hessenberg. Furthermore, since M is symmetric, it is clear that $T^t = Q^t MQ = T$. That is, T is an upper Hessenberg matrix that is symmetric. Therefore, T is tridiagonal and we can write it as follows:

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & & \cdots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \beta_{d-1} \\ 0 & \cdots & & \beta_{d-1} & \alpha_d \end{pmatrix}$$

Equating columns j on both sides of MQ = QT yields

$$Mq_j = \beta_{j-1} q_{j-1} + \alpha_j q_j + \beta_j q_{j+1}$$

Since the columns of Q are orthonormal, multiplying both sides of (6.1) by q_j yields $q_j^t M q_j = \alpha_j$.

The foregoing justifies what is called the *Lanczos algorithm*, performed as follows:

The Lanczos algorithm for (partial) reduction to symmetric tridiagonal form

$q_1 = \psi/||\psi||_2$, $\beta_0 = 0$, $q_0 = 0$

for $j = 1$ to p
- $z = Mq_j$
- $\alpha_j = q_j^t z$
- $z = z - \alpha_j q_j - \beta_{j-1} q_{j-1}$
- $\beta_j = ||z||_2$
- if $\beta_j = 0$, quit
- $q_{j+1} = z/\beta_j$

end for

The q_j computed by the Lanczos algorithm are often called the *Lanczos vectors*. If the loop in the algorithm is terminated because $\beta_k = 0$, this indicates that an exact invariant subspace has been computed, and is given by ran$\{q_1 \cdots q_k\}$. However, an exact zero is rarely found in practice, so we usually concern ourselves with the case in which we halt the this algorithm after p steps, and consider what can be said about p eigenvalues of M. For our problem, we are mainly interested in the first two eigenvalues of M, so we could stop after only two steps.

6.2 The Lanczos Algorithm for Markov Chains

Notice that the Lanczos algorithm requires the computation of Mq_j. Throughout this paper we have assumed that $M \in \mathbf{R}^{d \times d}$, and that d is so large as to make the matrix vector multiplication Mq_j impossible. Now, the Lanczos algorithm constructs an orthonormal basis for the Krylov subspace. For our purposes, this construction is not essential. What is essential is that we find a way to generate the *Lanczos coefficients* α_j, β_j, $(j = 1, \ldots, p)$, and whence the matrix T, without performing the operation Mq_j required by the Lanczos algorithm. We now address this problem.

Begin with the centered and weighted observable $\psi_1 = \Pi^{1/2}(\phi - \mathbf{E}_\pi \phi)$, and let $q_1 = \psi_1/||\psi_1||_2$. To simplify notation, let $\phi(n) \equiv \phi(X_n)$. The form of the first Lanczos coefficient α_1 is straight forward:

$$\begin{aligned} \alpha_1 &= q_1^t M q_1 \\ &= \frac{\psi_1^t M \psi_1}{\psi_1^t \psi_1} \end{aligned}$$

and now the special properties of the Markov chain setting play an important role. For, regard the following

$$\begin{aligned} \psi_1^t M \psi_1 &= (\phi - \mathbf{E}_\pi \phi)^t \Pi^{\frac{1}{2}t} M \Pi^{\frac{1}{2}} (\phi - \mathbf{E}_\pi \phi) \\ &= \mathbf{C}_\pi (\phi(n), \phi(n+1)) \end{aligned} \tag{12}$$

and

$$\begin{aligned}\psi_1^t\psi_1 &= (\phi - \mathbf{E}_\pi\phi)^t\Pi(\phi - \mathbf{E}_\pi\phi) \\ &= \mathrm{Var}_\pi(\phi(n))\end{aligned} \tag{13}$$

Recall that the definition of M implies $\Pi^{\frac{1}{2}t}\mathrm{M}\Pi^{\frac{1}{2}} = \Pi\mathrm{P}$, and this justifies equations (12) and (13). Putting these equations together, we have a nice form for α_1:

$$\alpha_1 = \frac{\mathbf{C}_\pi\left(\phi(n), \phi(n+1)\right)}{\mathrm{Var}_\pi(\phi(n))}$$

The next Lanczos coefficient, β_1, is only slightly more complicated.

$$\begin{aligned}\beta_1 &= \|\mathrm{M}q_1 - \alpha_1 q_1\|_2 \\ &= \left[(\mathrm{M}q_1 - \alpha_1 q_1)^t(\mathrm{M}q_1 - \alpha_1 q_1)\right]^{1/2} \\ &= \left[q_1^t\mathrm{M}^2 q_1 - \alpha_1^2\right]^{1/2}\end{aligned} \tag{14}$$

Again recalling the definition $\mathrm{M} \equiv \Pi^{1/2}\mathrm{P}\Pi^{-1/2}$, a simple manipulation verifies $\Pi^{\frac{1}{2}t}\mathrm{M}^s\Pi^{\frac{1}{2}} = \Pi\mathrm{P}^s$, whereby it follows that

$$\begin{aligned}q_1^t\mathrm{M}^2 q_1 &= \frac{(\phi - \mathbf{E}_\pi\phi)^t\Pi\mathrm{P}^2(\phi - \mathbf{E}_\pi\phi)}{(\phi - \mathbf{E}_\pi\phi)^t\Pi(\phi - \mathbf{E}_\pi\phi)} \\ &= \frac{\mathbf{C}_\pi\left(\phi(n), \phi(n+2)\right)}{\mathrm{Var}_\pi(\phi(n))}\end{aligned}$$

Inserting the last expression into equation (14) gives

$$\beta_1 = \left[\frac{\mathbf{C}_\pi\left(\phi(n), \phi(n+2)\right)}{\mathrm{Var}_\pi(\phi(n))} - \left(\frac{\mathbf{C}_\pi\left(\phi(n), \phi(n+1)\right)}{\mathrm{Var}_\pi(\phi(n))}\right)^2\right]^{1/2}$$

Notice that we can estimate these expressions of α_1 and β_1 by running simulations of the Markov chain. To do so, we run the Markov chain starting from a stationary distribution, compute the value of the observable at each step of the chain, and from these data, compute estimates of the covariances and variance of the observable.

We consider one more coefficient in detail, as we will need it when applying these ideas in the next section. Recall that the second basis element q_2 in the Lanczos algorithm is given by $q_2 = (\mathrm{M}q_1 - \alpha_1 q_1)/\beta_1$. From this we can derive an expression for α_2 which, though not as neat as those for α_1 and β_1, is nonetheless tractable.

$$\begin{aligned}\alpha_2 &= q_2^t\mathrm{M}q_2 \\ &= \frac{1}{\beta_1^2}(\mathrm{M}q_1 - \alpha_1 q_1)^t\mathrm{M}(\mathrm{M}q_1 - \alpha_1 q_1) \\ &= \frac{1}{\beta_1^2}(q_1^t\mathrm{M}^3 q_1 - 2\alpha_1 q_1^t\mathrm{M}^2 q_1 + \alpha^2 q_1^t\mathrm{M}q_1) \\ &= \frac{1}{\beta_1^2}(q_1^t\mathrm{M}^3 q_1 - 2\alpha_1\beta_1^2 - \alpha_1^3)\end{aligned} \tag{15}$$

The last equality follows from equation (14), which can be written as $q_1^t\mathrm{M}^2 q_1 = \alpha_1^2 + \beta_1^2$. The preceding paragraphs define all but one of the quantities in the expression (15). The one remaining is $q_1^t\mathrm{M}^3 q_1$, which can also be expressed in terms of covariances of ϕ, and the analysis is almost identical to that for $q_1^t\mathrm{M}^2 q_1$:

$$\begin{aligned}q_1^t\mathrm{M}^3 q_1 &= \frac{(\phi - \mathbf{E}_\pi\phi)^t\Pi\mathrm{P}^3(\phi - \mathbf{E}_\pi\phi)}{(\phi - \mathbf{E}_\pi\phi)^t\Pi(\phi - \mathbf{E}_\pi\phi)} \\ &= \frac{\mathbf{C}_\pi\left(\phi(n), \phi(n+3)\right)}{\mathrm{Var}_\pi(\phi(n))}\end{aligned}$$

It should be clear that the foregoing could be continued for higher order Lanczos coefficients. The expressions become more complicated, but the key point is that each coefficient α_k, β_k can be expressed as a function of the covariance between $\phi(n)$ and $\phi(n+j)$, for $j = 0, \ldots, k+1$. For bounding convergence rates of Markov chains, Theorem 2.2 shows that the subdominant eigenvalue is the only one required. The foregoing development justifies approximating this eigenvalue with an eigenvalue of the matrix

$$\mathrm{T} = \begin{pmatrix} \alpha_1 & \beta_1 \\ \beta_1 & \alpha_2 \end{pmatrix}$$

We apply this idea in the following section.

7 Conclusion

We conclude by noting a few aspects of the special cases to which the foregoing applies. We are interested in the first few eigenvalues of a large transition probability matrix. The foregoing can be used to approximate these eigenvalues when the following conditions are satisfied:

- The Markov chain is reversible (detailed balance condition).
- We can easily simulate realizations of the chain as well as functions (observables) of these realizations.

When these conditions hold, we can side step the traditional Lanczos algorithm, and get at the values of the Lanczos coefficients via the variational properties of observables on the state space.

Current research involves constructing the computer programs to simulate these situations, and the foregoing methods. As preliminary evidence suggests, there is little doubt that the theory will prove useful.

References

[1] J. Demmel. *Applied Numerical Linear Algebra.* Society for Industrial and Applied Mathematics, 1997.

[2] P. Diaconis and L. Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6:695–750, 1996.

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

[4] S. Geman and D. Geman. Stochastic relaxation. gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

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

[6] G. Goodman and A. Sokal. Multigrid monte carlo method. conceptual foundations. *Physical Review D*, 40:2035–2071, 1989.

[7] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods.* Chapman and Hall, 1964.

[8] J. Rosenthal. Convergence rates for markov chains. *Siam Review*, 37:387–405, 1995.

[9] N. Metropolis A. W. Rosenbluth M. N. Rosenbluth A. H. Teller E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.