Efficiency of Reversible MCMC Methods: Elementary Derivations and Applications to Composite Methods

Radford M. Neal
radford@utstat.utoronto.ca

Jeffrey S. Rosenthal
jeff@math.toronto.edu

Department of Statistical Sciences, University of Toronto

(May, 2023; revised March 2024)

Abstract. We review criteria for comparing the efficiency of Markov chain Monte Carlo (MCMC) methods with respect to the asymptotic variance of estimates of expectations of functions of state, and show how such criteria can justify ways of combining improvements to MCMC methods. We say that a chain on a finite state space with transition matrix P efficiency-dominates one with transition matrix Q if for every function of state it has lower (or equal) asymptotic variance. We give elementary proofs of some previous results regarding efficiency dominance, leading to a self-contained demonstration that a reversible chain with transition matrix P efficiency-dominates a reversible chain with transition matrix Q if and only if none of the eigenvalues of $Q - P$ are negative. This allows us to conclude that modifying a reversible MCMC method to improve its efficiency will also improve the efficiency of a method that randomly chooses either this or some other reversible method, and to conclude that improving the efficiency of a reversible update for one component of state (as in Gibbs sampling) will improve the overall efficiency of a reversible method that combines this and other updates. It also explains how antithetic MCMC can be more efficient than i.i.d. sampling. We also establish conditions that can guarantee that a method is not efficiency-dominated by any other method.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms (e.g. [6]) estimate the expected value of a function $f : S \to \mathbf{R}$ with respect to a probability distribution π on a state space S, which in this paper we assume to be finite, using an estimator such as

$$\widehat{f}_N \;=\; \frac{1}{N} \sum_{k=1}^{N} f(X_k)\,,$$

where $X_1, X_2, X_3, \ldots$ is a time-homogeneous Markov chain with stationary distribution π, having transition probabilities $P(x, y)$ from state x to state y (often viewed as a matrix P).

An important measure of the efficiency of this estimator is its asymptotic variance:

$$v(f, P) \;:=\; \lim_{N\to\infty} N \operatorname{Var}\Big[\widehat{f}_N\Big] \;=\; \lim_{N\to\infty} N \operatorname{Var}\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big] \;=\; \lim_{N\to\infty} \frac{1}{N} \operatorname{Var}\Big[\sum_{i=1}^{N} f(X_i)\Big]. \qquad (1)$$

For the irreducible Markov chains used for MCMC, the initial state of the chain does not affect the asymptotic variance, and the bias of the estimator converges to zero at rate $1/N$ regardless of initial state. (In practice, an initial portion of the chain is usually simulated but not used for estimation, in order to reduce the bias in a finite-length run.)

If we run the chain for a large number of iterations, N, we therefore expect that $v(f, P)/N$ will be an indication of the likely squared error of the estimate obtained. Indeed, when $v(f, P)$ is finite, one can show (e.g., [25, Theorem 5]) that a Central Limit Theorem applies, with the distribution of $(\widehat{f} - E_\pi(f)) \,/\, \sqrt{v(f, P)/N}$ converging to $N(0, 1)$.

We are therefore motivated to try to modify the chain to reduce $v(f, P)$, ideally for all functions f. We say that one transition matrix, P, efficiency-dominates another one, Q, if

$$v(f, P) \;\le\; v(f, Q) \qquad \text{for all } f : S \to \mathbf{R}\,.$$

Various conditions are known [25, 10, 20] which ensure that P efficiency-dominates Q. One of these, for reversible chains, is the Peskun-dominance condition [22, 26] which on a finite state space is that $P(x, y) \ge Q(x, y)$ for all $x \ne y$. This condition is widely cited and has gotten significant recent attention [16, 9, 27], and even extended to non-reversible chains [1]. But it is a very strong condition, and P might well efficiency-dominate Q even if it does not Peskun-dominate it.

In this paper, we focus on reversible chains with a finite state space. We present several known equivalences of efficiency dominance, whose proofs were previously scattered in the literature, sometimes only hinted at, and sometimes based on very technical mathematical arguments. We provide complete elementary proofs of them in Sections 3, 4, and 8, using little more than simple linear algebra techniques.

In Section 5, we use these equivalences to derive new results, which can show efficiency dominance for some chains constructed by composing multiple component transition matrices, as is done for the Gibbs Sampler. These results are applied to methods for improving Gibbs sampling in a companion paper [21]. In Section 6, we consider eigenvalue connections, and show how one can sometimes prove that a reversible chain cannot be efficiency-dominated by any other reversible chain, and also explain (Corollary 15) how antithetic MCMC can be more efficient than i.i.d. sampling, These results also allow an easy re-derivation, in Section 7, of the fact that Peskun dominance implies efficiency dominance.

2 Background Preliminaries

We assume that the state space S is finite, with $|S| = n$, and let π be a probability distribution on S, with $\pi(x) > 0$ for all $x \in S$, and $\sum_{x \in S} \pi(x) = 1$. For functionals $f, g : S \to \mathbf{R}$, define the $L^2(\pi)$ inner product by

$$\langle f,\, g \rangle \;=\; \sum_{x \in S} f(x)\, g(x)\, \pi(x)\,.$$

That is, $\langle f,\, g \rangle = \mathbf{E}_\pi[f(X)\, g(X)]$. Equivalently, if we let $S = \{1, 2, \ldots, n\}$, represent a function f by the column vector $f = [f(1), \ldots, f(n)]^T$, and let $D = \mathrm{diag}(\pi)$ be the $n \times n$ diagonal matrix with $\pi(1), \ldots, \pi(n)$ on the diagonal, then $\langle f,\, g \rangle$ equals the matrix product $f^T D g$.

We aim to estimate expectations with respect to π by using a time-homogeneous Markov chain $X_1, X_2, X_3, \ldots$ on S, with transition probabilities $P(x, y) = \mathbf{P}(X_{t+1} = y \,|\, X_t = x)$, often written as a matrix P, for which π is a <u>stationary distribution</u> (or <u>invariant distribution</u>):

$$\pi(y) \;=\; \sum_{x \in S} \pi(x) P(x, y)$$

Usually, π is the only stationary distribution, though we sometimes consider transition matrices that are not irreducible (see below), for which this is not true, as building-blocks for other chains.

For $f : S \to \mathbf{R}$, let $(Pf) : S \to \mathbf{R}$ be the function defined by

$$(Pf)(x) \;=\; \sum_{y \in S} P(x, y)\, f(y)\,.$$

Equivalently, if we represent f as a vector of its values for elements of S, then Pf is the product of the matrix P with the vector f. Another interpretation is that $(Pf)(x) = E_P[f(X_{t+1}) | X_t = x]$, where E_P is expectation with respect to the transitions defined by P. We can see that

$$\langle f,\, Pg \rangle \;=\; \sum_{x \in S} \sum_{y \in S} f(x)\, P(x, y)\, g(y)\, \pi(x)\,.$$

Equivalently, $\langle f,\, Pg \rangle$ is the matrix product $f^T D P g$. Also, $\langle f,\, Pg \rangle = \mathbf{E}_{\pi,P}[f(X_t)\, g(X_{t+1})]$, where $\mathbf{E}_{\pi,P}$ means expectation with respect to the Markov chain with initial state drawn from the stationary distribution π and proceeding according to P.

A transition matrix P is called <u>reversible</u> with respect to π if $\pi(x)\, P(x, y) = \pi(y)\, P(y, x)$ for all $x, y \in S$. This implies that π is a stationary distribution for P, since $\sum_x \pi(x) P(x, y) = \sum_x \pi(y) P(y, x) = \pi(y) \sum_x P(y, x) = \pi(y)$.

If P is reversible, $\langle f,\, Pg \rangle = \langle Pf,\, g \rangle$ for all f and g — i.e., P is <u>self-adjoint</u> (or, <u>Hermitian</u>) with respect to $\langle \cdot,\, \cdot \rangle$. Equivalently, P is reversible with respect to π if and only if DP is a symmetric matrix — i.e., DP is self-adjoint with respect to the classical dot-product. This allows us to easily verify some well-known facts about reversible Markov chains:

Lemma 1 *If P is reversible with respect to π then: (a) the eigenvalues of P are real; (b) these eigenvalues can be associated with real eigenvectors; (c) if λ_i and λ_j are eigenvalues of P with $\lambda_i \neq \lambda_j$, and v_i and v_j are real eigenvectors associated with λ_i and λ_j, then $v_i^T D v_j = 0$, where D is the diagonal matrix with π on the diagonal (i.e., $\langle v_i, v_j \rangle = 0$); (d) all the eigenvalues of P are in $[-1, 1]$.*

Proof. Since DP is symmetric and D is diagonal, $DP = (DP)^T = P^T D$. (a) If $Pv = \lambda v$ with v non-zero, then $\overline{\lambda}\,\overline{v}^T = \overline{v}^T P^T$, hence $\overline{\lambda}\,(\overline{v}^T D v) = \overline{v}^T P^T D v = \overline{v}^T D P v = \lambda\,(\overline{v}^T D v)$. Since $\overline{v}^T D v$ is non-zero (because D has positive diagonal elements), it follows that $\overline{\lambda} = \lambda$, and hence λ is real. (b) If $Pv = \lambda v$, with P and λ real and v non-zero, then at least one of $\mathrm{Re}(v)$ and $\mathrm{Im}(v)$ is non-zero and is a real eigenvector associated with λ. (c) $\lambda_i(v_i^T D v_j) = v_i^T P^T D v_j = v_i^T D P v_j = \lambda_j(v_i^T D v_j)$, which when $\lambda_i \neq \lambda_j$ implies that $v_i^T D v_j = 0$. (d) Since rows of P are non-negative and sum to one, the absolute value of an element of the vector Pv can be no larger than the largest absolute value of an element of v. If $Pv = \lambda v$, this implies that $|\lambda| \leq 1$, hence $\lambda \in [-1, 1]$. ∎

The self-adjoint property implies that P is a "normal operator", which guarantees (e.g., [13, Theorem 2.5.3]) the existence of an <u>orthonormal basis</u>, $v_1, v_2, \ldots, v_n$, of eigenvectors for P, with $Pv_i = \lambda_i v_i$ for each i, and $\langle v_i, v_j \rangle = \delta_{ij}$. (In particular, this property implies that P is <u>diagonalisable</u> or <u>non-defective</u>, but it is stronger than that.) Without loss of generality, we can take $\lambda_1 = 1$, and $v_1 = \mathbf{1} := [1, 1, \ldots, 1]^T$, so that $v_1(x) = \mathbf{1}(x) = 1$ for all $x \in S$, since $P\mathbf{1} = \mathbf{1}$ due to the transition probabilities in P summing to one. We can assume for convenience that all of P's eigenvalues (counting multiplicity) satisfy $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

In terms of orthonormal eigenvectors of P, any functions $f, g : S \to \mathbf{R}$ can be written as linear combinations $f = \sum_{i=1}^n a_i v_i$ and $g = \sum_{j=1}^n b_j v_j$. It then follows from orthonormality of these eigenvectors that

$$\langle f, g \rangle = \sum_i a_i b_i, \quad \langle f, f \rangle = \sum_i (a_i)^2, \quad \langle f, Pg \rangle = \sum_i a_i b_i \lambda_i, \quad \langle f, Pf \rangle = \sum_i (a_i)^2 \lambda_i.$$

In particular, $\langle f, v_i \rangle = a_i$, so the a_i are the projections of f on each of the v_i. This shows that $\sum_{i=1}^n v_i v_i^T D$ is equal to the identity matrix, since for all f,

$$\Big(\sum_{i=1}^n v_i v_i^T D\Big) f \;=\; \sum_{i=1}^n v_i (v_i^T D f) \;=\; \sum_{i=1}^n v_i \langle v_i, f \rangle \;=\; \sum_{i=1}^n a_i v_i \;=\; f\,.$$

Furthermore, any self-adjoint A whose eigenvalues are all zero must be the zero operator, since we can write any f as $f = \sum_{i=1}^n a_i v_i$, from which it follows that $Af = \sum_{i=1}^n \lambda_i a_i v_i = 0$.

A matrix A that is self-adjoint with respect to $\langle \cdot, \cdot \rangle$ has a <u>spectral representation</u> in terms of its eigenvalues and eigenvectors as $A = \sum_{i=1}^n \lambda_i v_i v_i^T D$. If $h : \mathbf{R} \to \mathbf{R}$ we can define $h(A) := \sum_{i=1}^n h(\lambda_i) v_i v_i^T D$, which is easily seen to be self-adjoint. Using $h(\lambda) = 1$ gives the identity matrix. One can also easily show that $h_1(A) + h_2(A) = (h_1 + h_2)(A)$ and

$h_1(A)h_2(A) = (h_1h_2)(A)$, and hence (when all $\lambda_i \neq 0$) that $A^{-1} = \sum_{i=1}^n \lambda_i^{-1} v_i v_i^T D$ and (when all $\lambda_i \geq 0$) that $A^{1/2} = \sum_{i=1}^n \lambda_i^{1/2} v_i v_i^T D$, so both of these are self-adjoint. Finally, note that if A and B are self-adjoint, so is ABA.

We say that P is <u>irreducible</u> if movement from any x to any y in S is possible via some number of transitions that have positive probability under P. An irreducible chain will have only one stationary distribution. A reversible irreducible P will have $\lambda_i < 1$ for $i \geq 2$. (As an aside, this implies that P is <u>variance bounding</u>, which in turn implies that $v(f, P)$ from (1) must be finite for each f [23, Theorem 14].) For MCMC estimation, we want our chain to be irreducible, but irreducible chains are sometimes built using transition matrices that are not irreducible — for example, by letting $P = (1/2)P_1 + (1/2)P_2$, where P_1 and/or P_2 are not irreducible, but P is irreducible.

An irreducible P is <u>periodic</u> with period p if S can be partitioned into $p > 1$ subsets $S_0, \ldots, S_{p-1}$ such that $P(x, y) = 0$ if $x \in S_a$ and $y \notin S_b$, where $b = a+1 \bmod p$ (and this is not true for any smaller p). Otherwise, P is <u>aperiodic</u>. An irreducible periodic chain that is reversible must have period 2, and will have $\lambda_n = -1$ and $\lambda_i > -1$ for $i \neq n$. A reversible aperiodic chain will have all $\lambda_i > -1$.

Since $v(f, P)$ as defined in (1) only involves variance, we can subtract off the mean of f without affecting the asymptotic variance. Hence, we can always assume without loss of generality that $\pi(f) = 0$, where $\pi(f) := \mathbf{E}_\pi(f) = \sum_{x \in S} f(x)\,\pi(x) = \langle f,\, \mathbf{1}\rangle$. In other words, we can assume that $f \in L_0^2(\pi) := \{f : \pi(f) = 0,\ \pi(f^2) < \infty\}$, where the condition that $\pi(f^2) = \mathbf{E}_\pi(f^2)$ be finite is automatically satisfied when S is finite, and hence can be ignored. Also, if $\pi(f) = 0$, then $\langle f,\, \mathbf{1}\rangle = \langle f,\, v_1\rangle = 0$, so f is orthogonal to v_1, and hence its coefficient a_1 is zero.

Next, note that

$$\langle f,\, P^k g\rangle \;=\; \sum_{x \in S} f(x)\,(P^k g)(x)\,\pi(x) \;=\; \sum_{x \in S}\sum_{y \in S} f(x) P^k(x, y) g(y) \pi(x) \;=\; \mathbf{E}_{\pi,P}[f(X_t)\,g(X_{t+k})]\,.$$

where $P^k(x, y)$ is the k-step transition probability from x to y. If $f \in L_0^2(\pi)$ (i.e., the mean of f is zero), this is the <u>covariance</u> of $f(X_t)$ and $g(X_{t+k})$, when the chain is started in stationarity (and hence is the same for all t). We define the <u>lag-k autocovariance</u>, γ_k, as:

$$\gamma_k \;:=\; \mathrm{Cov}_{\pi,P}[f(X_t), f(X_{t+k})] \;:=\; \mathbf{E}_{\pi,P}[f(X_t) f(X_{t+k})] \;=\; \langle f,\, P^k f\rangle, \quad \text{for } f \in L_0^2(\pi).$$

If $f = \sum_{i=1}^n a_i v_i$ as above (with $a_1 = 0$ since the mean of f is zero), then using orthonormality of the eigenvectors v_i,

$$\gamma_k \;=\; \langle f,\, P^k f\rangle \;=\; \sum_{i=2}^n \sum_{j=2}^n \langle a_i v_i,\, P^k(a_j v_j)\rangle \;=\; \sum_{i=2}^n \sum_{j=2}^n \langle a_i v_i,\, (\lambda_j)^k a_j v_j)\rangle \;=\; \sum_{i=2}^n (a_i)^2 (\lambda_i)^k\,.$$

In particular, $\gamma_0 = \langle f,\, f\rangle := \|f\|_{L^2(\pi)} = \sum_i (a_i)^2$. (If the state space S were not finite, we would need to require $f \in L^2(\pi)$, but finite variance is guaranteed with a finite state space.)

One particular example of a transition matrix P, useful for comparative purposes, is Π, the operator corresponding to i.i.d. sampling from π. It is defined by $\Pi(x,y) = \pi(y)$ for all $x \in S$. This operator satisfies $\Pi\mathbf{1} = \mathbf{1}$, and $\Pi f = 0$ whenever $\pi(f) = 0$. Hence, its eigenvalues are $\lambda_1 = 1$ and $\lambda_i = 0$ for $i \neq 1$.

3 Relating Asymptotic Variance to Eigenvalues

In this section, we consider some expressions for the asymptotic variance, $v(f,P)$, of (1), beginning with a result relating the asymptotic variance to the eigenvalues of P. This result (as observed by [10]) can be obtained (at least in the aperiodic case) as a special case of the more technical results of Kipnis and Varadhan [15, eqn (1.1)].

Proposition 2 *If P is an irreducible (but possibly periodic) Markov chain on a finite state space S, which is reversible with respect to π, with orthonormal basis $v_1, v_2, \ldots, v_n$ of eigenvectors, and corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and $f \in L_0^2(\pi)$ with $f = \sum_i a_i v_i$, then the limit $v(f,P)$ in (1) exists, and*

$$v(f,P) \;=\; \sum_{i=2}^{n} (a_i)^2 \;+\; 2\sum_{i=2}^{n} (a_i)^2 \,\frac{\lambda_i}{1-\lambda_i} \;=\; \sum_{i=2}^{n} (a_i)^2 \,\frac{1+\lambda_i}{1-\lambda_i}\,.$$

Proof. First, by expanding the square, using stationarity, and collecting like terms, we obtain the well-known result that for $f \in L_0^2(\pi)$,

$$\begin{aligned}
\frac{1}{N}\,\mathrm{Var}\left(\sum_{i=1}^{N} f(X_i)\right) &= \frac{1}{N}\,\mathbf{E}_{\pi,P}\left[\left(\sum_{i=1}^{N} f(X_i)\right)^2\right] \\
&= \frac{1}{N}\left(N\,\mathbf{E}_{\pi,P}[f(X_j)^2] \;+\; 2\sum_{k=1}^{N-1}(N-k)\;\mathbf{E}_{\pi,P}[f(X_j)\;f(X_{j+k})]\right) \\
&= \gamma_0 \;+\; 2\sum_{k=1}^{N-1}\frac{N-k}{N}\;\gamma_k\,,
\end{aligned}$$

where $\gamma_k = \mathrm{Cov}_{\pi,P}[f(X_j), f(X_{j+k})] = \langle f,\, P^k f\rangle$ is the lag-k autocovariance in stationarity.

Now, $f = \sum_{i=1}^{n} a_i v_i$, with $a_1 = 0$ since $\pi(f) = 0$, so $\gamma_k \;=\; \langle f,\, P^k f\rangle \;=\; \sum_{i=2}^{n}(a_i)^2(\lambda_i)^k$ and $\gamma_0 \;=\; \sum_{i=2}^{n}(a_i)^2$. The above then gives that

$$\frac{1}{N}\,\mathrm{Var}\left(\sum_{i=1}^{N} f(X_i)\right) \;=\; \sum_{i=2}^{n}(a_i)^2 + 2\sum_{k=1}^{N-1}\frac{N-k}{N}\;\sum_{i=2}^{n}(a_i)^2(\lambda_i)^k\,, \tag{2}$$

i.e.

$$\frac{1}{N}\,\mathrm{Var}\left(\sum_{i=1}^{N} f(X_i)\right) \;=\; \sum_{i=2}^{n}(a_i)^2 + 2\sum_{k=1}^{\infty}\sum_{i=2}^{n} I_{k\leq N-1}\;\frac{N-k}{N}\;(a_i)^2(\lambda_i)^k\,.$$

If P is aperiodic, then $\Lambda := \max_{i\geq 2}|\lambda_i| \;<\; 1$, hence $\sum_{k=1}^{\infty}\left|\sum_{i=2}^{n} I_{k\leq N-1}\,\frac{N-k}{N}\;(a_i)^2(\lambda_i)^k\right| \leq \sum_{k=1}^{\infty}\sum_{i=2}^{n}\left|I_{k\leq N-1}\,\frac{N-k}{N}\;(a_i)^2(\lambda_i)^k\right| \;\leq\; \sum_{k=1}^{\infty}\sum_{i=2}^{n}(a_i)^2(\Lambda)^k \;=\; \gamma_0\,\Lambda/(1-\Lambda) < \infty$, so the above

sum is <u>absolutely summable</u>. This lets us exchange the limit and summations to obtain

$$v(f,P) \;:=\; \lim_{N\to\infty}\frac{1}{N}\,\mathrm{Var}\left(\sum_{i=1}^{N} f(X_i)\right) \;=\; \sum_{i=2}^{n}(a_i)^2 + 2\sum_{i=2}^{n}\sum_{k=1}^{\infty}\lim_{N\to\infty}\left[I_{k\le N-1}\;\frac{N-k}{N}\;(a_i)^2(\lambda_i)^k\right]$$

$$= \sum_{i=2}^{n}(a_i)^2 + 2\sum_{i=2}^{n}\sum_{k=1}^{\infty}(a_i)^2(\lambda_i)^k \;=\; \sum_{i=2}^{n}(a_i)^2 + 2\sum_{i=2}^{n}(a_i)^2\frac{\lambda_i}{1-\lambda_i} \;=\; \sum_{i=2}^{n}(a_i)^2\,\frac{1+\lambda_i}{1-\lambda_i}\,.$$

If P is periodic, with $\lambda_n = -1$, then the above $\Lambda = 1$, and $\sum_{k=1}^{\infty}(\lambda_n)^k$ is not even defined, so the above argument does not apply. Instead, separate out the $i = n$ term in (2) to get

$$\frac{1}{N}\,\mathrm{Var}\left(\sum_{i=1}^{N} f(X_i)\right) \;=\; \sum_{i=2}^{n}(a_i)^2 \;+\; 2\sum_{k=1}^{N-1}\frac{N-k}{N}\;\sum_{i=2}^{n-1}(a_i)^2(\lambda_i)^k \;+\; 2\sum_{k=1}^{N-1}\frac{N-k}{N}(a_n)^2(-1)^k\,.$$

Since $\Gamma := \max\{|\lambda_2|, |\lambda_3|, \ldots, |\lambda_{n-1}|\} < 1$, the previous argument applies to the middle double-sum term to show that

$$\lim_{N\to\infty} 2\sum_{k=1}^{N-1}\frac{N-k}{N}\;\sum_{i=2}^{n-1}(a_i)^2(\lambda_i)^k \;=\; 2\sum_{i=2}^{n-1}(a_i)^2\,\frac{\lambda_i}{1-\lambda_i}\,.$$

As for the final term, writing values for k as $2m-1$ or $2m$, we have

$$\sum_{k=1}^{N-1}\frac{N-k}{N}(-1)^k \;=\; \frac{1}{N}\sum_{m=1}^{\lfloor (N-1)/2\rfloor}\Big[-(N-2m+1)+(N-2m)\Big] \;-\; \frac{1}{N}\,I_{N-1\text{ is odd}}$$

$$= \frac{1}{N}\sum_{m=1}^{\lfloor (N-1)/2\rfloor}\Big[-1\Big] \;-\; \frac{1}{N}\,I_{N-1\text{ is odd}} \;=\; -\frac{\lfloor (N-1)/2\rfloor}{N} \;-\; \frac{1}{N}\,I_{N-1\text{ is odd}}\,,$$

which converges as $N\to\infty$ to $-\frac{1}{2} = \frac{-1}{1-(-1)} = \frac{\lambda_n}{1-\lambda_n}$. So, we again obtain that

$$v(f,P) \;=\; \lim_{N\to\infty}\frac{1}{N}\,\mathrm{Var}\left(\sum_{i=1}^{N} f(X_i)\right) \;=\; \sum_{i=2}^{n}(a_i)^2 \;+\; 2\sum_{i=2}^{n-1}(a_i)^2\frac{\lambda_i}{1-\lambda_i} \;+\; 2\,(a_n)^2\frac{\lambda_n}{1-\lambda_n}$$

$$= \sum_{i=2}^{n}(a_i)^2 \;+\; 2\sum_{i=2}^{n}(a_i)^2\frac{\lambda_i}{1-\lambda_i} \;=\; \sum_{i=2}^{n}(a_i)^2\,\frac{1+\lambda_i}{1-\lambda_i}\,. \qquad \blacksquare$$

Note that when P is periodic, λ_n will be -1, and the final term in the expression for $v(f,P)$ will be zero. Such a periodic P will have zero asymptotic variance when estimating the expectation of a function f for which a_n is the only non-zero coefficient.

When P is aperiodic, we can obtain from Proposition 2 the more familiar [2, 5, 7, 12, 14, 25] expression for $v(f,P)$ in terms of sums of autocovariances, though it is not needed for this paper (and actually still holds without the reversibility condition [5, Theorem 20.1]):

Proposition 3 *If P is a reversible, irreducible, aperiodic Markov chain on a finite state space S with stationary distribution π, and $f \in L_0^2(\pi)$, then*

$$v(f,P) \;=\; \lim_{N\to\infty}\frac{1}{N}\,\mathrm{Var}_\pi\left(\sum_{i=1}^{N} f(X_i)\right) \;=\; \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k\,, \tag{3}$$

where $\gamma_k = \mathrm{Cov}_{\pi,P}[f(X_t), f(X_{t+k})]$ is the lag-k autocovariance of f in stationarity.

Proof. Since $\gamma_k = \langle f,\, P^k f\rangle$, and as above in the aperiodic case $\Lambda := \sup_{i\geq 2}|\lambda_i| < 1$, the double-sum is again absolutely summable, and we compute directly that if $f = \sum_i a_i v_i$ then

$$\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \;=\; \langle f,\, f\rangle + 2\sum_{k=1}^{\infty}\langle f,\, P^k f\rangle \;=\; \sum_i (a_i)^2 + 2\sum_{k=1}^{\infty}\sum_i (a_i)^2(\lambda_i)^k$$

$$= \sum_i (a_i)^2 + 2\sum_i (a_i)^2 \sum_{k=1}^{\infty}(\lambda_i)^k \;=\; \sum_i (a_i)^2 + 2\sum_i (a_i)^2\,\frac{\lambda_i}{1-\lambda_i}\,,$$

so the result follows from Proposition 2. ∎

Proposition 2 gives the following formula for $v(f,P)$ (see also [19, Lemma 3.2]):

Proposition 4 *The asymptotic variance, $v(f,P)$, for the functional $f \in L^2_0(\pi)$ using an* <u>*irreducible*</u> *Markov chain P which is* <u>*reversible*</u> *with respect to π satisfies the equation*

$$v(f,P) \;=\; \langle f,\, f\rangle \;+\; 2\,\langle f,\, P(I-P)^{-1}f\rangle\,,$$

which we can also write as $v(f,P) = \langle f,\, f\rangle \;+\; 2\,\langle f,\, \frac{P}{I-P}\,f\rangle$, or as $v(f,P) = \langle f,\, \frac{I+P}{I-P}\,f\rangle$.

Proof. Let P have an orthonormal basis $v_1, v_2, \ldots, v_n$, with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and using this basis let $f = \sum_i a_i v_i$. Note that $a_1 = 0$, since f has mean zero, so we can ignore v_1 and λ_1. Define $h(\lambda) := \lambda(1-\lambda)^{-1}$. As discussed in Section 2, applying h to the eigenvalues of P will produce another self-adjoint matrix, with the same eigenvectors, which will equal $P(I-P)^{-1}$. Using this, we can write

$$\begin{aligned}
\langle f,\, f\rangle \;+\; 2\,\langle f,\, P(I-P)^{-1}f\rangle \;&=\; \sum_i (a_i)^2 \;+\; 2\sum_i\sum_j \langle a_i v_i,\, P(I-P)^{-1}(a_j v_j)\rangle \\
&=\; \sum_i (a_i)^2 \;+\; 2\sum_i\sum_j \langle a_i v_i,\, \lambda_j(1-\lambda_j)^{-1}(a_j v_j)\rangle \\
&=\; \sum_i (a_i)^2 \;+\; 2\sum_i (a_i)^2\,\lambda_i\,(1-\lambda_i)^{-1}\,,
\end{aligned}$$

so the result follows from Proposition 2. ∎

Remark. If we write $P = \sum_{i=1}^n \lambda_i v_i v_i^T D$, so $I-P = \sum_{i=1}^n (1-\lambda_i) v_i v_i^T D$, then on $L^2_0(\pi)$ this becomes $I-P = \sum_{i=2}^n (1-\lambda_i) v_i v_i^T D$, so $(I-P)^{-1} = \sum_{i=2}^n (1-\lambda_i)^{-1} v_i v_i^T D$.

Remark. The inverse $(I-P)^{-1}$ in Proposition 4 is on the restricted space $L^2_0(\pi)$ of functions f with $\pi(f) = 0$. That is, $(I-P)^{-1}\,(I-P)\,f \;=\; (I-P)\,(I-P)^{-1}\,f \;=\; f$ for any f in $L^2_0(\pi)$. By contrast, $I-P$ will <u>not</u> be invertible on the full space $L^2(\pi)$ of <u>all</u> functions on S, since, for example, $(I-P)\mathbf{1} = \mathbf{1}-\mathbf{1} = \mathbf{0}$, so if $(I-P)^{-1}$ existed on all of $L^2(\pi)$ then we would have the contradiction that $\mathbf{1} = (I-P)^{-1}(I-P)\,\mathbf{1} = (I-P)^{-1}\,\mathbf{0} = \mathbf{0}$.

4 Efficiency Dominance Equivalences

Combining Proposition 4 with the definition of efficiency dominance proves:

Proposition 5 *For reversible irreducible Markov chain transition matrices P and Q, P efficiency-dominates Q <u>if and only if</u> $\langle f,\, P(I-P)^{-1}f\rangle \le \langle f,\, Q(I-Q)^{-1}f\rangle$ for all $f \in L_0^2(\pi)$, or informally that $\langle f,\, \frac{P}{I-P}f\rangle \le \langle f,\, \frac{Q}{I-Q}f\rangle$ for all $f \in L_0^2(\pi)$.*

Next, we need the following fact:

Lemma 6 *If P and Q are reversible and irreducible Markov chain transition matrices, $\langle f,\, P(I-P)^{-1}f\rangle \le \langle f,\, Q(I-Q)^{-1}f\rangle$ for all $f \in L_0^2(\pi)$ <u>if and only if</u> $\langle f,\, Pf\rangle \le \langle f,\, Qf\rangle$ for all $f \in L_0^2(\pi)$.*

Lemma 6 follows from the very technical results of Bendat and Sherman [3]. It is somewhat subtle since the equivalence is only for <u>all</u> f at once, not for individual f; see the discussion after Lemma 23 below. In Section 8 below, we present an <u>elementary</u> proof. (For alternative direct proofs of Lemma 6 and related facts, see also [4, Chapter V].)

Combining Lemma 6 and Proposition 5 immediately shows the following, which is also shown by Mira and Geyer [19, Theorem 4.2]:

Theorem 7 *For Markov chain transition matrices P and Q that are reversible and irreducible, P efficiency-dominates Q <u>if and only if</u> $\langle f,\, Pf\rangle \le \langle f,\, Qf\rangle$ for all $f \in L_0^2(\pi)$, i.e. <u>if and only if</u> $\langle f,\, (Q-P)f\rangle \ge 0$ for all $f \in L_0^2(\pi)$.*

Remark. Here the restriction that $f \in L_0^2(\pi)$, i.e. that $\pi(f) = 0$, can be omitted, since if $c := \pi(f) \ne 0$ then $f = f_0 + c$ where $\pi(f_0) = 0$, and $\langle f,\, Pf\rangle = \langle f_0 + c,\, P(f_0 + c)\rangle = \langle f_0,\, Pf_0\rangle + c^2$, and similarly for Q. But we do not need this fact here.

Remark. Some authors (e.g. [18]) say that P <u>covariance-dominates</u> Q if $\langle f,\, Pf\rangle \le \langle f,\, Qf\rangle$ for all $f \in L_0^2(\pi)$, or equivalently if $\mathrm{Cov}_{\pi,P}[f(X_t),\, f(X_{t+1})]$ is always smaller under P than under Q. The surprising conclusion of Theorem 7 is that for reversible chains this is <u>equivalent</u> to efficiency dominance — i.e., to $v(f,P) \le v(f,Q)$ for all $f \in L_0^2(\pi)$. So, there is no need to consider the two concepts separately.

To make the condition $\langle f,\, (Q-P)f\rangle \ge 0$ for all f more concrete, we have the following:

Lemma 8 *Any self-adjoint matrix J satisfies $\langle f,\, Jf\rangle \ge 0$ for all f <u>if and only if</u> the eigenvalues of J are all non-negative, which is <u>if and only if</u> the eigenvalues of DJ are all non-negative where $D = \mathrm{diag}(\pi)$.*

Proof. Let J have orthonormal basis of eigenvectors $v_1, v_2, \ldots, v_n$ as in Section 2, so any f can be written as $f = \sum_{i=1}^n a_i v_i$. Then

$$\langle f,\, Jf\rangle \;=\; \langle \sum_i a_i v_i,\, \sum_j a_j J v_j\rangle \;=\; \langle \sum_i a_i v_i,\, \sum_j a_j \lambda_j v_j\rangle \;=\; \sum_i (a_i)^2 \lambda_i\,.$$

If each $\lambda_i \geq 0$, then this expression must be ≥ 0. Conversely, if some $\lambda_i < 0$, then choosing $f = v_i$ gives $\langle f,\, Jf\rangle = \lambda_i < 0$. This proves the first statement.

For the second statement, recall that DJ is self-adjoint with respect to the classical dot-product. Hence, by the above, the matrix product $f^T DJf \geq 0$ for all f <u>if and only if</u> the eigenvalues of DJ are all non-negative. So, since $f^T DJf = \langle f,\, Jf\rangle$, the two statements are equivalent. ∎

Combining Lemma 8 (with J replaced by $Q-P$) with Theorem 7 shows:

Theorem 9 *If P and Q are reversible irreducible Markov chain transitions, P efficiency-dominates Q <u>if and only if</u> the operator $Q - P$ (equivalently, the matrix $Q - P$) has all eigenvalues non-negative, which is <u>if and only if</u> the matrix $D\,(Q - P)$ has all eigenvalues non-negative.*

Remark. By Theorem 9, if $Q - P$ has even a single negative eigenvalue, say $(Q - P)\,z = -cz$ where $c > 0$, then there must be some $f \in L_0^2(\pi)$ such that $v(f, Q) < v(f, P)$. By following through our proof of Lemma 6 in Section 8 below, it might be possible to construct such an f <u>explicitly</u> in terms of z and c. We leave this as an open problem.

Remark. It might be possible to give another alternative proof of Theorem 9 using the step-wise approach of [20], by writing $Q-P = R_1 + R_2 + \ldots + R_\ell$ where each R_i is of rank one (e.g., $R_i = \lambda_i v_i v_i^T D$ with λ_i and v_i an eigenvalue and eigenvector of $Q-P$). We leave this as another open problem.

Theorem 9 allows us to prove the following, which helps justify the phrase "efficiency-dominates" (see also [19, Section 4]):

Theorem 10 *Efficiency dominance is a <u>partial order</u> on reversible chains, i.e.:*

(a) It is <u>reflexive</u>: P always efficiency-dominates P;

(b) It is <u>antisymmetric</u>: if P efficiency-dominates Q, and Q efficiency-dominates P, then $P = Q$;

(c) It is <u>transitive</u>: if P efficiency-dominates Q, and Q efficiency-dominates R, then P efficiency-dominates R.

Proof. Statement (a) is trivial. Statement (c) is true because $v(f,P) \leq v(f,Q)$ and $v(f,Q) \leq v(f,R)$ imply $v(f,P) \leq v(f,R)$. For statement (b), Theorem 9 implies that both $Q-P$ and $P-Q$ have all eigenvalues non-negative, hence their eigenvalues must all be zero, which implies (since $Q-P$ is self-adjoint) that $Q-P=0$, and hence $P=Q$. ∎

Remark. Statement (b) of Theorem 10 does not hold if we do not assume reversibility. For example, if $S=\{1,2,3\}$, and $\pi = \text{Uniform}(S)$, and $P(1,2)=P(2,3)=P(3,1)=1$, and $Q(1,3)=Q(3,2)=Q(2,1)=1$, then $v(f,P)=v(f,Q)=0$ for all $f: S \to \mathbf{R}$, so they each (weakly) efficiency-dominate the other, but $P \neq Q$.

5 Efficiency Dominance of Combined Chains

Using Theorems 7 and 9, we can now prove some new results about efficiency dominance that are useful when Markov chains are constructed by combining two or more chains.

We first consider the situation where we randomly choose to apply transitions defined either by P or by Q. For example, P might move about one region of the state space well, while Q moves about a different region well. Randomly choosing either P or Q may produce a chain that moves well over the entire state space. The following theorem says that if in this situation we can improve P to P', then the random combination will also be improved:

Theorem 11 *Let P, P', and Q be reversible with respect to π, with P and P' irreducible, and let $0 < a < 1$. Then P' efficiency-dominates P <u>if and only if</u> $aP' + (1-a)Q$ efficiency-dominates $aP + (1-a)Q$.*

Proof. Since P and P' are irreducible, so are $aP' + (1-a)Q$ and $aP + (1-a)Q$. So by Theorem 9, P' efficiency-dominates P <u>if and only if</u> $P - P'$ has all non-negative eigenvalues, which is clearly <u>if and only if</u> $a(P-P') = [aP + (1-a)Q] - [aP' + (1-a)Q]$ has all non-negative eigenvalues, which is <u>if and only if</u> $aP'+(1-a)Q$ efficiency-dominates $aP+(1-a)Q$. ∎

The next result applies to Markov chains built using component transition matrices that are not necessarily irreducible, such as single-variable updates in a random-scan Gibbs sampler, again showing that improving one of the components will improve the combination, assuming the combination is irreducible:

Theorem 12 *Let $P_1, \ldots, P_\ell$ and $P'_1, \ldots, P'_\ell$ be reversible with respect to π (though not necessarily irreducible). Let $a_1, \ldots, a_\ell$ be mixing probabilities, with $a_k > 0$ and $\sum_k a_k = 1$, and let $P = a_1 P_1 + \ldots + a_\ell P_\ell$ and $P' = a_1 P'_1 + \ldots + a_\ell P'_\ell$. Then if P and P' are irreducible, and for each k the eigenvalues of $P_k - P'_k$ (or of $D(P_k - P'_k)$ where $D = \text{diag}(\pi)$) are all non-negative, then P' efficiency-dominates P.*

Proof. Choose any $f \in L_0^2(\pi)$. Since P_k and P'_k are self-adjoint, we have from Lemma 8 that $\langle f, (P_k - P'_k)f\rangle \geq 0$ for each k. Then, by linearity,

$$\langle f, (P - P')f\rangle \;=\; \langle f, \sum_k a_k\,(P_k - P'_k)f\rangle \;=\; \sum_k a_k\,\langle f, (P_k - P'_k)f\rangle \;\geq\; 0\,,$$

too. Hence, by Theorem 7, P' efficiency-dominates P. ∎

In the Gibbs sampling application, the state is composed of ℓ components, so that $S = S_1 \times S_2 \times \cdots \times S_\ell$, and P_k is the transition that samples a value for component k, independent of its current value, from its conditional distribution given the values of other components, while leaving the values of these other components unchanged. Since it leaves other components unchanged, such a P_k will not be irreducible. P_k will be a block-diagonal matrix, in a suitable ordering of states (different for each k), with $B = |S|/|S_k|$ blocks, each of size $K = |S_k|$.

For example, suppose $\ell = 2$, $S_1 = \{1, 2\}$, $S_2 = \{1, 2, 3\}$, and $\pi(x) = 1/9$ except that $\pi((1, 2)) = 4/9$. With lexicographic ordering, the Gibbs sampling transition matrix for the second component, P_2, will be

$$P_2 \;=\; \begin{pmatrix} 1/6 & 4/6 & 1/6 & 0 & 0 & 0 \\ 1/6 & 4/6 & 1/6 & 0 & 0 & 0 \\ 1/6 & 4/6 & 1/6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}$$

Each block of P_k can be regarded as the $K \times K$ transition matrix for a Markov chain having S_k as its state space, which is reversible with respect to the conditional distribution on S_k given the values for other components associated with this block. For each block, the eigenvalues and eigenvectors of this transition matrix give rise to corresponding eigenvalues and eigenvectors of P_k, after prepending and appending zeros to the eigenvector according to how many blocks precede and follow this block. If the transition matrix for each block is irreducible, there will be B eigenvalues of P_k equal to one, with eigenvectors of the form $[0, \ldots, 0, 1, \ldots, 1, 0 \ldots, 0]^T$, which are zero except for a series of K ones corresponding to one of the blocks.

The overall transition matrix when using Gibbs sampling to update a component randomly chosen with equal probabilities will be $P = (1/\ell)(P_1 + \cdots + P_\ell)$. We can try to improve the efficiency of P by modifying one or more of the P_k. An improvement to P_k can take the form of an improvement to one of its blocks, each of which corresponds to particular values of components of the state other than component k. With each P_k changed to P'_k, the modified overall transition matrix is $P' = (1/\ell)(P'_1 + \cdots + P'_\ell)$.

For the example above, we could try to improve P by improving P_2, with the improvement to P_2 taking the form of an improvement to how the second component is changed when the

first component has the value 1, as follows:

$$P_2' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 2/4 & 1/4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}$$

The change to the 3×3 upper-left block still leaves it reversible with respect to the conditional distribution for the second component given the value 1 for the first component (which has probabilities of 1/6, 4/6, 1/6), but introduces an antithetic aspect to the sampling.

If we leave P_1 unchanged, so $P_1' = P_1$, Theorem 12 can be used to show that the P' built with this modified P_2' efficiency-dominates P built with the original P_1 and P_2. The difference $P_2 - P_2'$ will also be block diagonal, and its eigenvalues will be those of the differences in the individual blocks (which are zero for blocks that have not been changed). In the example above, the one block in the upper-left that changed has difference:

$$\begin{pmatrix} 1/6 & 4/6 & 1/6 \\ 1/6 & 4/6 & 1/6 \\ 1/6 & 4/6 & 1/6 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 2/4 & 1/4 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} +2/12 & -4/12 & +2/12 \\ -1/12 & +2/12 & -1/12 \\ +2/12 & -4/12 & +2/12 \end{pmatrix}$$

The eigenvalues of this difference matrix are 1/2, 0, and 0. The eigenvalues of $P_2 - P_2'$ will be these plus three more zeros. If $P_1' = P_1$, Theorem 12 then guarantees that P' efficiency-dominates P, the original Gibbs sampling chain, since these eigenvalues are all non-negative. Note that here one cannot show efficiency-dominance using Peskun-dominance, since the change reduces some off-diagonal transition probabilities.

More generally, suppose a Gibbs sampling chain is changed by modifying one or more of the blocks of one or more of the P_k, with the new blocks efficiency-dominating the old Gibbs sampling blocks (seen as transition matrices reversible with respect to the conditional distribution for that block). Then by Theorem 9, the eigenvalues of the differences between the old and new blocks are all non-negative, which implies that the eigenvalues of $P_k - P_k'$ all non-negative for each k, which by Theorem 12 implies that the modified chain efficiency-dominates the original Gibbs sampling chain. The practical applications of this are developed further in the companion paper [21].

6 Efficiency Dominance and Eigenvalues

We will now present some results relating eigenvalues of transition matrices to efficiency dominance, which can sometimes be used to show that a reversible transition matrix <u>cannot</u> be efficiency-dominated by any other reversible transition matrix.

Say that P <u>eigen-dominates</u> Q if both P and Q are reversible and the eigenvalues of P are no greater than the corresponding eigenvalues of Q — that is, when the eigenvalues of P (counting multiplicities) are written non-increasing as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, and the eigenvalues of Q are written non-increasing as $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_n$, then $\lambda_i \leq \beta_i$ for each i. Then we have (see also [19, Theorem 3.3]):

Proposition 13 *If P and Q are irreducible and reversible with respect to π, and P efficiency-dominates Q, then P eigen-dominates Q.*

Proof. By Theorem 7, $\langle f,\, Pf\rangle \leq \langle f,\, Qf\rangle$ for all $f \in L_0^2(\pi)$. Hence, the result follows from the "min-max" characterisation of eigenvalues (e.g. [13, Theorem 4.2.6]) that

$$\lambda_i \;=\; \inf_{g_1,\ldots,g_{i-1}} \; \sup_{\substack{f\in L_0^2(\pi)\\ \langle f,\, f\rangle=1\\ \langle f,\, g_j\rangle=0\,\forall\, j}} \langle f,\, Pf\rangle\,.$$

Intuitively, $g_1, \ldots, g_{i-1}$ represent the first $i-1$ eigenvectors (excluding the eigenvector $\mathbf{1}$ associated with the eigenvalue 1, since $\mathbf{1} \notin L_0^2(\pi)$), so that the new eigenvector f will be orthogonal to them. However, since the formula is stated in terms of <u>any</u> vectors $g_1, \ldots, g_{i-1}$, the same formula applies for both P and Q, thus giving the result. ∎

The <u>converse</u> of Proposition 13 does <u>not</u> hold, contrary to a claim in [18, Theorem 2]. For example, suppose the state space is $S = \{1, 2, 3\}$. Let $e = 0.05$, and let

$$P \;=\; \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2}-e & e \\ 0 & e & 1-e \end{pmatrix}, \quad Q \;=\; \begin{pmatrix} 1-e & e & 0 \\ e & \frac{1}{2}-e & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad R \;=\; \begin{pmatrix} 1-e & e & 0 \\ e & \frac{1}{2} & \frac{1}{2}-e \\ 0 & \frac{1}{2}-e & \frac{1}{2}+e \end{pmatrix}$$

These all are reversible with respect to $\pi = \text{Uniform}(S)$, and are irreducible and aperiodic. One can see that P eigen-dominates Q (and vice versa), since P and Q are equivalent upon swapping states 1 and 3, and so have the same eigenvalues, which are equal (to four decimal places) to 1, 0.9270, -0.0270. However, P does <u>not</u> efficiency-dominate Q, since $Q - P$ has eigenvalues 0.7794, 0, -0.7794 which are <u>not</u> all non-negative. (Nor does Q efficiency dominate P, analogously.)

Intuitively, in this example, Q moves easily between states 2 and 3, but only infrequently to or from state 1, while P moves easily between states 1 and 2 but not to or from state 3. Hence, if, for example, $f(1) = 2$ and $f(2) = 1$ and $f(3) = 3$ so that $f(1) = \frac{1}{2}[f(2) + f(3)]$, then $v(f, Q) < v(f, P)$, since Q moving slowly between $\{1\}$ and $\{2, 3\}$ doesn't matter, but P moving slowly between $\{1, 2\}$ and $\{3\}$ <u>does</u> matter.

R is a slight modification to Q that has two smaller off-diagonal elements, and hence is Peskun-dominated (and efficiency-dominated) by Q. It's eigenvalues are 1, 0.9272, 0.0728, the later two of which are strictly larger than those of P, so P eigen-dominates R. But the

eigenvalues of $R - P$ are 0.7865, 0, -0.6865, which are not all non-negative, so P does not efficiency-dominate R, despite strictly eigen-dominating it.

However, the next result is in a sense a converse of Proposition 13 for the special case where <u>all</u> of the non-trivial eigenvalues for P are smaller than <u>all</u> of those for Q:

Theorem 14 *Let P and Q be irreducible and reversible with respect to π, with eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$ for P and $1 = \beta_1 \geq \beta_2 \geq \beta_3 \geq \ldots \geq \beta_n$ for Q (counting multiplicities). Suppose $\max_{i \geq 2} \lambda_i \leq \min_{i \geq 2} \beta_i$, i.e. $\lambda_2 \leq \beta_n$, i.e. $\lambda_i \leq \beta_j$ for <u>any</u> $i, j \geq 2$. Then P efficiency-dominates Q.*

Proof. Let a be the maximum eigenvalue of P restricted to $L_0^2(\pi)$, and let b be the minimum eigenvalue of Q restricted to $L_0^2(\pi)$, which for both P and Q will exclude the eigenvalue of 1 associated with $\mathbf{1}$. The assumptions imply that $a \leq b$. But the "min-max" characterisation of eigenvalues (described in the proof of Proposition 13 above), applied to the largest eigenvalue on $L_0^2(\pi)$ (i.e., excluding the eigenvalue 1), implies that

$$a \;=\; \sup_{\substack{f \in L_0^2(\pi) \\ \langle f,\, f \rangle = 1}} \langle f,\, Pf \rangle \,.$$

Also, since $-b$ is the largest eigenvalue of $-Q$,

$$b \;=\; -\sup_{\substack{f \in L_0^2(\pi) \\ \langle f,\, f \rangle = 1}} \langle f,\, -Qf \rangle \;=\; \inf_{\substack{f \in L_0^2(\pi) \\ \langle f,\, f \rangle = 1}} \langle f,\, Qf \rangle \,.$$

Since $a \leq b$, this implies that $\langle f,\, Pf \rangle \leq \langle f,\, Qf \rangle$ for any $f \in L_0^2(\pi)$ with $\langle f,\, f \rangle = 1$, and hence (by linearity) for <u>any</u> $f \in L_0^2(\pi)$. It then follows from Theorem 7 that P efficiency-dominates Q. ∎

A chain is called "antithetic" (cf. [11]) if all its eigenvalues (except $\lambda_1 = 1$) are non-positive, with at least one negative. Our next result shows that such antithetic samplers always efficiency-dominate i.i.d. sampling:

Corollary 15 *If P is irreducible and reversible with respect to π, and has eigenvalues $\lambda_1 = 1$ and $\lambda_2, \lambda_3, \ldots, \lambda_n \leq 0$, then P efficiency-dominates Π (the operator corresponding to i.i.d. sampling from π).*

Proof. By assumption, $\max_{i \geq 2} \lambda_i \leq 0$. Also, if $1 = \beta_1 \geq \beta_2 \geq \beta_3 \geq \ldots \geq \beta_n$ are the eigenvalues for Π, then $\beta_i = 0$ for all $i \geq 2$, so $\min_{i \geq 2} \beta_i = 0$. Hence, $\max_{i \geq 2} \lambda_i \leq \min_{i \geq 2} \beta_i$. The result then follows from Theorem 14. ∎

Remark. Theorem 1 of [8] shows that if $\pi_{\min} = \min_x \pi(x)$, the maximum eigenvalue (other than λ_1) of a transition matrix reversible with respect to π must be greater than or equal to $-\pi_{\min} / (1-\pi_{\min})$, which must be greater than or equal to $-1/(n-1)$ since $\pi_{\min} \leq 1/n$. Now, if $f = v_i$ where $\lambda_i \geq -1/(n-1)$, then Proposition 2 gives $v(f, P) = (1+\lambda_i)/(1-\lambda_i) \geq (n-2)/n$, which for large n is only slightly smaller than $v(f, \Pi) = (1+0)/(1-0) = 1$. On the other hand, we can still have e.g. $\lambda_1 = 1$, $\lambda_n = -1$, and all the other $\lambda_i = 0$, and then if $f = v_n$ then $v(f, P) = (1+(-1))/(1-(-1)) = 0/2 = 0$, which is significantly less than $v(f, \Pi) = 1$. Hence, the improvement in Corollary 15 could be large for some functions f, but small for some others.

Since practical interest focuses on whether or not some chain, P, efficiency-dominates another chain, Q, Proposition 13 is perhaps most useful in its contrapositive form — if P and Q are reversible, and P does <u>not</u> eigen-dominate Q, then P does <u>not</u> efficiency-dominate Q. That is, if Q has at least one eigenvalue less than the corresponding eigenvalue of P, then P does not efficiency-dominate Q. If both chains have an eigenvalue less than the corresponding eigenvalue of the other chain, then neither efficiency-dominates the other.

But what if two different chains have exactly the same ordered set of eigenvalues — that is, they both eigen-dominate the other? In that case, neither efficiency-dominates the other. To show that, we first prove a result about strict trace comparisons:

Theorem 16 *If P and Q are both irreducible transitions matrices, reversible with respect to π, and P efficiency-dominates Q, and $P \neq Q$, then trace$(P) <$ trace(Q), i.e. the trace (or equivalently the sum of eigenvalues) of P is <u>strictly</u> smaller than that of Q.*

Proof. By Theorem 9, if P efficiency-dominates Q, then $Q-P$ has no negative eigenvalues. And it cannot have all zero eigenvalues, since then $Q-P = 0$ (since $Q-P$ is self-adjoint), contradicting the premise that $P \neq Q$. So, $Q-P$ has at least one positive eigenvalue, and no negative eigenvalues, and hence the sum of eigenvalues of $Q-P$ is strictly positive. But the sum of the eigenvalues of a matrix is equal to its trace [13, p. 51], so this is equivalent to trace$(Q-P) > 0$. Since trace is linear, this implies that trace$(Q)-$trace$(P) > 0$, and hence trace$(P) <$ trace(Q). ∎

Corollary 17 *If P and Q are both irreducible transitions matrices, reversible with respect to π, and the eigenvalues (counting multiplicity) for both are identical, and $P \neq Q$, then P does not efficiency-dominate Q, and Q does not efficiency-dominate P.*

Proof. If P and Q have identical eigenvalues, then trace$(P) =$ trace(Q), so this follows immediately from Theorem 16. ∎

We next present a lemma about minimal values of trace(P).

Lemma 18 *For any transition matrix P on a finite state space S, for which π is a stationary distribution, the sum of the diagonal elements of P — that is, trace(P) — must be at least* $\max(0,\ (2\pi_{\max} - 1)/\pi_{\max})$*, where* $\pi_{\max} = \max_x \pi(x)$*. Furthermore, any P attaining this minimum value will have at most one non-zero value on its diagonal, and any such non-zero diagonal value will be for a state x^* for which $\pi(x^*) = \pi_{\max} > 1/2$.*

Proof. The statement is trivial when $\pi_{\max} \leq 1/2$, since the lower limit on trace(P) is then zero, and any such P has all zeroes on the diagonal. Otherwise, if x^* is such that $\pi(x^*) = \pi_{\max} > 1/2$, then stationarity implies that

$$\begin{aligned}
\pi_{\max} \;=\; \pi(x^*) \;&=\; \sum_{x\in S} \pi(x)P(x,x^*) \;=\; \pi(x^*)P(x^*,x^*) \;+\; \sum_{x\in S,\, x\neq x^*} \pi(x)P(x,x^*) \\
&\leq\; \pi(x^*)P(x^*,x^*) \;+\; \sum_{x\in S,\, x\neq x^*} \pi(x) \;=\; \pi(x^*)P(x^*,x^*) \;+\; (1-\pi(x^*)) \\
&=\; \pi_{\max}\, P(x^*,x^*) \;+ (1-\pi_{\max})\,.
\end{aligned}$$

It follows that $P(x^*,x^*) \geq (2\pi_{\max} - 1)/\pi_{\max}$, hence trace($P$) $\geq \max(0,\ (2\pi_{\max} - 1)/\pi_{\max})$. Furthermore, if trace($P$) $= \max(0,\ (2\pi_{\max} - 1)/\pi_{\max})$, then trace($P$) $= P(x^*,x^*)$, and hence all other values on the diagonal of P must be zero. ∎

Remark. For any π, the minimum value of trace(P) in Lemma 18 is attainable, and can indeed be attained by a P that is reversible. Several methods for constructing such a P are discussed in the companion paper [21], including, for example, the "shifted tower" method of [24], which produces a reversible P when the shift is by 1/2.

We can now state a criterion for a reversible chain to not be efficiency-dominated by any other reversible chain:

Theorem 19 *If P is the transition matrix for an irreducible Markov chain on a finite state space that is reversible with respect to π, and the sum of the eigenvalues of P (equivalently, the trace of P) equals* $\max(0,\ (2\pi_{\max} - 1)/\pi_{\max})$*, where* $\pi_{\max} = \max_x \pi(x)$*, then no other reversible chain can efficiency-dominate P.*

Proof. A reversible chain, Q, not equal to P, that efficiency-dominates P, must by Theorem 16 have trace(Q) $<$ trace(P). But by Lemma 18, trace(P) is as small as possible. So there can be no reversible Q that efficiency-dominates P. ∎

As an example of how this theorem can be applied, if the state space is $S = \{1,2,3\}$, with $\pi(1) = \pi(2) = 1/5$ and $\pi(3) = 3/5$, for which $\pi_{\max} = 3/5$, then the transition matrix

$$P_1 \;=\; \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

is reversible with respect to π, and has eigenvalues of $1, 0, -2/3$, which sum to $1/3$ (the trace). By Theorem 19, P_1 cannot be efficiency-dominated by any other reversible transition matrix, since its sum of eigenvalues is equal to $(2\pi_{\max} - 1) / \pi_{\max}$.

On the other hand, consider the following transition matrix, reversible with respect to the same π:

$$P_2 = \begin{pmatrix} 0 & 1/4 & 3/4 \\ 1/4 & 0 & 3/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

P_2 has eigenvalues of $1, -1/4, -1/4$, which sum to $1/2$, greater than $(2\pi_{\max} - 1) / \pi_{\max}$, so Theorem 19 does not apply. However, P_2 is an instance of a transition matrix constructed according to a procedure of Frigessi, Hwang, and Younes [8, Theorem 1], which they prove has the property that the transition matrix produced has the smallest possible value for λ_2, and subject to having that value for λ_2, the smallest possible value for λ_3, etc. We can therefore again conclude from Proposition 13 and Corollary 17 that no other reversible chain can efficiency-dominate P_2.

It's easy to see that any reversible P with at least two non-zero diagonal elements, say $P(x, x)$ and $P(y, y)$, can be efficiency-dominated by a chain, Q, that is the same as P except that these diagonal elements are reduced, allowing $Q(x, y)$ and $Q(y, x)$ to be greater than $P(x, y)$ and $P(y, x)$, so that Q Peskun-dominates P. Theorem 19 shows that <u>some</u> reversible P in which only a single diagonal element is non-zero cannot be efficiency-dominated by any other reversible chain. We know of no examples of a reversible P with only one non-zero diagonal element that is dominated by another reversible chain, but we do not have a proof that this is impossible. This leads to:

Open Problem. Does there exists a reversible P with only one non-zero diagonal element, which is efficiency-dominated by some other reversible chain?

7 Re-deriving Peskun's Theorem

Recall that P <u>Peskun-dominates</u> Q if $P(x, y) \geq Q(x, y)$ for all $x \neq y$ — i.e., that $Q - P$ has all non-positive entries off the diagonal (and hence also that $Q - P$ has all non-negative entries on the diagonal). It is known through several complicated proofs [22, 26, 20] that if P Peskun-dominates Q, then P efficiency-dominates Q. We will see here that once Theorem 9 has been established, this fact can be shown easily.

Proposition 20 *If P and Q are irreducible, and both are reversible with respect to some π, and P Peskun-dominates Q, then P efficiency-dominates Q.*

To prove Proposition 20, we begin with a simple eigenvalue lemma.

Lemma 21 *If Z is an $n \times n$ matrix with $z_{ii} \geq 0$ and $z_{ij} \leq 0$ for all $i \neq j$, and row-sums $\sum_j z_{ij} = 0$ for all i, then all eigenvalues of Z must be non-negative.*

Proof. Suppose $Zv = \lambda v$. Find the index j which maximizes $|v_j|$, i.e. such that $|v_j| \geq |v_k|$ for all k. We can assume $v_j > 0$ (if not, replace v by $-v$), so $v_j \geq |v_k|$ for all k. Then

$$\lambda\, v_j \;=\; (Zv)_j \;=\; \sum_i z_{ji} v_i \;=\; z_{jj} v_j + \sum_{i \neq j} z_{ji} v_i \;\geq\; z_{jj} v_j - \sum_{i \neq j} |z_{ji}|\, |v_i|$$

$$\geq\; z_{jj} v_j - \sum_{i \neq j} |z_{ji}|\, v_j \;=\; v_j \Big(z_{jj} + \sum_{i \neq j} z_{ji} \Big) \;=\; v_j(0) \;=\; 0\,.$$

So, $\lambda\, v_j \geq 0$. Hence, since $v_j > 0$, we must have $\lambda \geq 0$. ∎

Proof of Proposition 20. Let $Z = Q - P$. Since P Peskun-dominates Q, $z_{ii} = Q(i,i) - P(i,i) \geq 0$ and $z_{ij} = Q(i,j) - P(i,j) \leq 0$ for all $i \neq j$. Also $\sum_j z_{ij} = \sum_j P(i,j) - \sum_j Q(i,j) = 1 - 1 = 0$. Hence, by Lemma 21, $Z = Q - P$ has all eigenvalues non-negative. Hence, by Theorem 9, P efficiency-dominates Q. ∎

Remark. Proposition 20 can also be proven by transforming Q into P one step at a time, in the sequence $Q, Q', Q'', \ldots, P$, with each matrix in the sequence efficiency-dominating the previous matrix. At each step, say from Q' to Q'', two of the off-diagonal transition probabilities that differ between Q and P, say those involving states x and y, will be increased from $Q(x,y)$ to $P(x,y)$ and from $Q(y,x)$ to $P(y,x)$, while $Q''(x,x)$ and $Q''(y,y)$ will decrease compared to $Q'(x,x)$ and $Q'(y,y)$. The difference $Q' - Q''$ will be zero except for a 2×2 submatrix involving states x and y, which will have the form $\begin{pmatrix} a & -a \\ -b & b \end{pmatrix}$ for some $a, b > 0$, which has non-negative eigenvalues of 0 and $a + b$. Hence, by Theorem 9, Q'' efficiency-dominates Q'. Since this will be true for all the steps from Q to P, transitivity (see Theorem 10(c)) implies that P efficiency-dominates Q. ∎

Note that the converse to Proposition 20 is <u>false</u>. For example, let $S = \{1, 2, 3\}$, and

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}, \quad Q - P = \begin{pmatrix} +1/2 & -1/4 & -1/4 \\ -1/2 & +1/4 & +1/4 \\ -1/2 & +1/4 & +1/4 \end{pmatrix}.$$

Here, P does <u>not</u> Peskun-dominate Q, since, for example, $Q(2,3) = 1/4 > 0 = P(2,3)$. However, the eigenvalues of $Q - P$ are $1, 0, 0$, all of which are non-negative, so P <u>does</u> efficiency-dominate Q. Furthermore, Theorem 10(b) implies that P is strictly better than Q — there is some f for which $v(f, P) < v(f, Q)$. (For example, the indicator function for the first state, which has asymptotic variance zero using P, and asymptotic variance $1/4$ using

Q.) Peskun dominance therefore does not capture all instances of efficiency dominance that we are interested in, which motivates our investigation here.

One should note, however, that all our results concern only reversible chains. Non-reversible chains are often used, either in a deliberate attempt to improve performance (see [20]), or somewhat accidentally, as a result of combining methods sequentially rather than by random selection. Extensions of Peskun ordering to non-reversible chains are considered in [1].

8 Elementary Proof of Lemma 6

We conclude by presenting the promised elementary proof of Lemma 6 above, which states the surprising fact that, for any reversible irreducible transition matrices P and Q, $\langle f,\, Pf\rangle \le \langle f,\, Qf\rangle$ for all f <u>if and only if</u> $\langle f,\, P(I-P)^{-1}f\rangle \le \langle f,\, Q(I-Q)^{-1}f\rangle$ for all f.

As observed in [19, pp. 16–17], this proposition follows from the more general result of Bendat and Sherman [3, p. 60], using results of Löwner [17], which states that that if $h(x) = \frac{ax+b}{cx+d}$ where $ad - bc > 0$, and J and K are any two self-adjoint operators with spectrum contained in $(-\infty, -d/c)$ or in $(-d/c, \infty)$, then if $\langle f,\, Jf\rangle \le \langle f,\, Kf\rangle$ for all f, then also $\langle f,\, h(J)f\rangle \le \langle f,\, h(K)f\rangle$ for all f. In particular, choosing $a = d = 1$, $b = 0$, and $c = -1$ gives that $h(J) = \frac{J}{I-J}$, so if $\langle f,\, Jf\rangle \le \langle f,\, Kf\rangle$ for all f then $\langle f,\, \frac{J}{I-J}\, f\rangle \le \langle f,\, \frac{K}{I-K}\, f\rangle$ for all f. Conversely, choosing $a = c = d = 1$ and $b = 0$ gives that $h(\frac{J}{I-J}) = J$, so if $\langle f,\, \frac{J}{I-J}\, f\rangle \le \langle f,\, \frac{K}{I-K}\, f\rangle$ for all f then $\langle f,\, Jf\rangle \le \langle f,\, Kf\rangle$ for all f, finishing the proof.

However, the proof in [3] is very technical, requiring analytic continuations of transition functions into the complex plane. Instead, we now present an elementary proof of Lemma 6. (See also [4, Chapter V].) We begin with some lemmas about operators on a finite vector space $\mathcal{V}$, e.g. $\mathcal{V} = L_0^2(\pi)$.

Lemma 22 *If X, Y, Z are operators on a finite vector space $\mathcal{V}$, with Z self-adjoint, and $\langle f,\, Xf\rangle \le \langle f,\, Yf\rangle$ for all $f \in \mathcal{V}$, then $\langle f,\, ZXZf\rangle \le \langle f,\, ZYZf\rangle$ for all $f \in \mathcal{V}$.*

Proof. Since Z is self-adjoint, making the substitution $g = Zf$ gives

$$\langle f,\, ZXZf\rangle \;=\; \langle Zf,\, XZf\rangle \;=\; \langle g,\, Xg\rangle \;\le\; \langle g,\, Yg\rangle \;=\; \langle Zf,\, YZf\rangle \;=\; \langle f,\, ZYZf\rangle\,. \quad \blacksquare$$

Next, say a self-adjoint matrix J is <u>strictly positive</u> if $\langle f,\, Jf\rangle > 0$ for all non-zero $f \in \mathcal{V}$. Since $\langle f,\, Jf\rangle = \sum_i (a_i)^2 \lambda_i$ (see Section 2), this is equivalent to J having all eigenvalues positive.

Lemma 23 *If J and K are strictly positive self-adjoint operators on a finite vector space $\mathcal{V}$, and $\langle f,\, Jf\rangle \le \langle f,\, Kf\rangle$ for all $f \in \mathcal{V}$, then $\langle f,\, J^{-1}f\rangle \ge \langle f,\, K^{-1}f\rangle$ for all $f \in \mathcal{V}$.*

Proof. Note (see Section 2) that $K^{-1/2}$ and $K^{-1/2}JK^{-1/2}$ are self-adjoint, and have eigenvalues that are positive, since J and K are strictly positive. Applying Lemma 22 with $X = J$, $Y = K$, and $Z = K^{-1/2}$ then gives that for all $f \in \mathcal{V}$,

$$\langle f,\, K^{-1/2}JK^{-1/2}f\rangle \;\le\; \langle f,\, K^{-1/2}KK^{-1/2}f\rangle \;=\; \langle f,\, If\rangle \;=\; \langle f,\, f\rangle\,.$$

It follows that all the eigenvalues of $K^{-1/2}JK^{-1/2}$ are in $(0,1]$ (since $\langle v,\, Av\rangle \le \langle v,\, v\rangle$ and $Av = \lambda v$ with $v \neq 0$ imply $\lambda \le 1$). Hence, its inverse $(K^{-1/2}JK^{-1/2})^{-1}$ has eigenvalues all ≥ 1, so $\langle f,\, (K^{-1/2}JK^{-1/2})^{-1}f\rangle \;\ge\; \langle f,\, If\rangle$, and therefore

$$\langle f,\, If\rangle \;\le\; \langle f,\, (K^{-1/2}JK^{-1/2})^{-1}f\rangle \;=\; \langle f,\, K^{1/2}J^{-1}K^{1/2}f\rangle\,.$$

Then, applying Lemma 22 again with $X = I$, $Y = K^{1/2}J^{-1}K^{1/2}$, and $Z = K^{-1/2}$ gives

$$\langle f,\, K^{-1/2}IK^{-1/2}f\rangle \;\le\; \langle f,\, K^{-1/2}\,(K^{1/2}J^{-1}K^{1/2})\,K^{-1/2}f\rangle\,.$$

That is, $\langle f,\, K^{-1}f\rangle \;\le\; \langle f,\, J^{-1}f\rangle$ for all $f \in \mathcal{V}$, giving the result. ∎

We emphasise that the equivalence in Lemma 23 is only for <u>all</u> f at once, not for individual f. For example, if $J = I$ and $K = \mathrm{diag}(5, \frac{1}{5})$ and $f = (1,1)$, then $\langle f,\, Jf\rangle = 2 \le 5 + \frac{1}{5} = \langle f,\, Kf\rangle$, but $\langle f,\, J^{-1}f\rangle = 2 \not\ge \frac{1}{5} + 5 = \langle f,\, K^{-1}f\rangle$. This illustrates why the proofs of Lemma 23 and Lemma 6 are not as straightforward as one might think.

Remark. Lemma 23 can be partially proven more directly. If $f = \sum_i a_i v_i$, <u>Jensen's Inequality</u> gives $\left(\sum_i (a_i)^2\lambda_i\right)^{-1} \le \sum_i (a_i)^2\,(\lambda_i)^{-1}$, so we always have $\frac{1}{\langle f,\, Jf\rangle} \le \langle f,\, J^{-1}f\rangle$. Hence, if $\langle f,\, Jf\rangle \le \langle f,\, Kf\rangle$, then $\langle f,\, J^{-1}f\rangle \ge \frac{1}{\langle f,\, Kf\rangle}$. If f is an <u>eigenvector</u> of K, then $\langle f,\, K^{-1}f\rangle = \frac{1}{\langle f,\, Kf\rangle}$, so this shows directly that $\langle f,\, J^{-1}f\rangle \ge \langle f,\, K^{-1}f\rangle$. However, it is unclear how to extend this argument to other f. ∎

Applying Lemma 23 twice gives a (stronger) two-way equivalence:

Lemma 24 *If J and K are strictly positive self-adjoint operators on a finite vector space $\mathcal{V}$, then $\langle f,\, Jf\rangle \le \langle f,\, Kf\rangle$ for all $f \in \mathcal{V}$ <u>if and only if</u> $\langle f,\, J^{-1}f\rangle \ge \langle f,\, K^{-1}f\rangle$ for all $f \in \mathcal{V}$.*

Proof. The forward implication is Lemma 23. And, the reverse implication follows from Lemma 23 by replacing J with K^{-1} and replacing K with J^{-1}. ∎

Using Lemma 24, we easily obtain:

Proof of Lemma 6. Recall that we can restrict to $f \in L_0^2(\pi)$, so $\pi(f) = 0$, and f is orthogonal to the eigenvector corresponding to eigenvalue 1. On that restricted subspace, the eigenvalues of P and Q are contained in $[-1, 1)$. Hence, the eigenvalues of $I - P$ and

$I-Q$ are contained in $(0,2]$, and in particular are all strictly positive. So, $I-P$ and $I-Q$ are strictly positive self-adjoint operators.

Now, $\langle f,\, Pf\rangle \leq \langle f,\, Qf\rangle$ for all $f \in L_0^2(\pi)$ is equivalent to

$$\langle f,\, (I-P)f\rangle = \langle f,\, f\rangle - \langle f,\, Pf\rangle \;\geq\; \langle f,\, f\rangle - \langle f,\, Qf\rangle = \langle f,\, (I-Q)f\rangle \;\text{ for all } f \in L_0^2(\pi).$$

Then, by Lemma 24 with $J = I-Q$ and $K = I-P$, this is equivalent to

$$\langle f,\, (I-P)^{-1}f\rangle \;\leq\; \langle f,\, (I-Q)^{-1}f\rangle \;\text{ for all } f \in L_0^2(\pi).$$

Since $(I-P)^{-1} = P(I-P)^{-1} + (I-P)(I-P)^{-1} = P(I-P)^{-1} + I$ and similarly $(I-Q)^{-1} = Q(I-Q)^{-1} + I$, this latter is equivalent to

$$\langle f,\, P(I-P)^{-1}f\rangle \;\leq\; \langle f,\, Q(I-Q)^{-1}f\rangle \;\text{ for all } f \in L_0^2(\pi),$$

which completes the proof. ∎

Acknowledgements. We are very grateful to Heydar Radjavi for the proof of Lemma 23 herein, and thank Rajendra Bhatia, Gareth Roberts, Daniel Rosenthal, and Peter Rosenthal for several very helpful discussions about eigenvalues. We thank the anonymous referee for a number of useful suggestions which helped to improve the paper.

References

[1] C. Andrieu and S. Livingstone (2021), Peskun-Tierney ordering for Markovian Monte Carlo: beyond the reversible scenario. Ann. Stat. **49(4)**, 1958–1981.

[2] M. S. Bartlett (1946), On the theoretical specification and sampling properties of autocorrelated time-series. Suppl. J. Roy. Stat. Soc. **8(1)**, 27–41.

[3] J. Bendat and S. Sherman (1955), Monotone and convex operator functions. Trans. Amer. Math. Soc. **79**, 58–71.

[4] R. Bhatia (1997), Matrix Analysis. Graduate texts in mathematics **169**. Springer, New York.

[5] P. Billingsley (1968), *Convergence of Probability Measures*, New York: Wiley.

[6] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. (2011), Handbook of Markov chain Monte Carlo. Chapman & Hall, New York.

[7] K. S. Chan and C. J. Geyer (1994), Discussion to reference [25]. Ann. Stat. **22**, 1747–1758.

[8] A. Frigessi, C.-R. Hwang, and L. Younes (1992), Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. Ann. Appl. Prob. **2(3)**, 610–628.

[9] P. Gagnon and F. Maire (2023), An asymptotic Peskun ordering and its application to lifted samplers. Bernoulli, to appear.

[10] C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci. **7(4)**, 473–483.

[11] P. J. Green and X.-L. Han (1992), Metropolis methods, Gaussian proposals, and antithetic variables. In Stochastic Models, Statistical Methods and Algorithms in Image Analysis (P. Barone et al., Eds.), Springer, Berlin.

[12] O. Häggström and J. S. Rosenthal (2007), On variance conditions for Markov chain CLTs. Elec. Comm. Prob. **12**, 454–464.

[13] R. A. Horn and C. R. Johnson (2013), *Matrix Analysis*, 2nd edition, Cambridge University Press.

[14] L.-J. Huang and Y.-H. Mao (2023), Variational formulas for asymptotic variance of general discrete-time Markov chains. Bernoulli **29(1)**, 300–322.

[15] C. Kipnis and S. R. S. Varadhan (1986), Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Comm. Math. Phys. **104**, 1–19.

[16] G. Li, A. Smith, and Q. Zhou (2023), Importance is Important: A Guide to Informed Importance Tempering Methods. arXiv preprint arXiv:2304.06251.

[17] K. Löwner (1934), Über monotone Matrixfunktionen. Mathematische Zeitschrift **38**, 177–216.

[18] A. Mira (2001), Ordering and improving the performance of Monte Carlo Markov chains. Stat. Sci. **16**, 340–350.

[19] A. Mira and C. J. Geyer (1999), Ordering Monte Carlo Markov chains. Technical Report No. 632, School of Statistics, University of Minnesota.

[20] R. M. Neal (2004), Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. `https://arxiv.org/abs/math/0407281`

[21] R. M. Neal (2024), Modifying Gibbs sampling to avoid self transitions. `https://arxiv.org/abs/2403.18054`

[22] P. H. Peskun (1973), Optimum Monte Carlo sampling using Markov chains. Biometrika **60**, 607–612.

[23] G. O. Roberts and J. S. Rosenthal (2008), Variance bounding Markov chains. Ann. Appl. Prob. **18(3)**, 1201–1214.

[24] H. Suwa (2022), "Reducing rejection exponentially improves Markov chain Monte Carlo sampling", `https://arxiv.org/abs/2208.03935`

[25] L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.

[26] L. Tierney (1998), A note on Metropolis-Hastings kernels for general state spaces. Ann. Appl. Prob. **8**, 1–9.

[27] G. Zanella (2019), Informed Proposals for Local MCMC in Discrete Spaces. J. Amer. Stat. Assoc. **115**, 852–865.