Comparing the Efficiency of General State Space Reversible MCMC Algorithms

Geoffrey T. Salmon
geoffrey.salmon@mail.utoronto.ca

Jeffrey S. Rosenthal
jeff@math.toronto.edu

Department of Statistical Sciences, University of Toronto

August 6, 2024

Abstract

We review and provide new proofs of results used to compare the efficiency of estimates generated by reversible Markov chain Monte Carlo (MCMC) algorithms on a general state space. We provide a full proof of the formula for the asymptotic variance for real-valued functionals on φ-irreducible reversible Markov chains, first introduced by Kipnis and Varadhan in [15]. Given two Markov kernels P and Q with stationary measure π, we say that the Markov kernel P efficiency dominates the Markov kernel Q if the asymptotic variance with respect to P is at most the asymptotic variance with respect to Q for every real-valued functional $f \in L^2(\pi)$. Assuming only a basic background in functional analysis, we prove that for two φ-irreducible reversible Markov kernels P and Q, P efficiency dominates Q if and only if the operator $\mathcal{Q} - \mathcal{P}$, where $\mathcal{P}$ is the operator on $L^2(\pi)$ that maps $f \mapsto \int f(y)P(\cdot, dy)$ and similarly for $\mathcal{Q}$, is positive on $L^2(\pi)$, i.e. $\langle f, (\mathcal{Q} - \mathcal{P})\, f\rangle \geq 0$ for every $f \in L^2(\pi)$. We use this result to show that reversible antithetic kernels are more efficient than i.i.d. sampling, and that efficiency dominance is a partial ordering on φ-irreducible reversible Markov kernels. We also provide a proof based on that of Tierney in [26] that Peskun dominance is a sufficient condition for efficiency dominance for reversible kernels. Using these results, we show that Markov kernels formed by randomly selecting other "component" Markov kernels will always efficiency dominate another Markov kernel formed in this way, as long as the component kernels of the former efficiency dominate those of the latter. These results on the efficiency dominance of combining component kernels generalises the results on the efficiency dominance of combined chains introduced by Neal and Rosenthal in [19] from finite state spaces to general state spaces.

1 Introduction

A common problem in statistics and other areas, is that of estimating the average value of a function $f : \mathbf{X} \to \mathbf{R}$ with respect to a probability measure π. The domain of f, $\mathbf{X}$, is called the state space. When the probability measure π is complicated and the expected value of f with respect to π, $\mathbf{E}_\pi(f)$, can't be computed directly, Markov chain Monte Carlo (MCMC) algorithms are very effective (see [21]). In MCMC, the solution is to find a Markov chain $\{X_k\}_{k \in \mathbf{N}}$ with underlying Markov kernel P, and estimate the expected value of f with respect to π as

$$\mathbf{E}_\pi(f) \approx \frac{1}{N} \sum_{k=1}^{N} f(X_k) =: \overline{f}_N.$$

The advantage is that the Markov kernel P provides much simpler probability measures at each step, making it easier to compute, while the law of the Markov chain, X_k, approaches the probability distribution π.

When the chosen function f is in $L^2(\pi)$, i.e. when $\int_{\mathbf{X}} |f(x)|^2 \pi(dx) = \mathbf{E}_\pi(|f|^2) < \infty$, one measure of the effectiveness of the chosen Markov kernel for the function f, is the *asymptotic variance* of f using the kernel P, $v(f, P)$, defined as

$$v(f,P) := \lim_{N \to \infty} \left[N \mathbf{Var} \left(\frac{1}{N} \sum_{k=1}^{N} f(X_k) \right) \right] = \lim_{N \to \infty} \left[\frac{1}{N} \mathbf{Var} \left(\sum_{k=1}^{N} f(X_k) \right) \right],$$

where $\{X_k\}_{k \in \mathbf{N}}$ is a Markov chain with kernel P, started in stationarity (i.e. $X_1 \sim \pi$).

Thus if $v(f, P)$ is finite, we would expect the variance of the estimate $\overline{f}_N$ to be near $v(f, P)/N$. Furthermore, if P is φ-irreducible and reversible, a central limit theorem (CLT) holds whenever $v(f, P)$ is finite, and the variance of said CLT is the asymptotic variance, i.e. $\sqrt{N} \left(\overline{f}_N - \mathbf{E}_\pi(f) \right) \xrightarrow{d} N(0, v(f, P))$ (see [15]). For further reference, see [13], [11], [25], [21] and [15].

X_1 is not usually sampled from π directly, but if P is φ-irreducible, then we can get very close to sampling from π directly by running the chain for a number of iterations before using the samples for estimation.

In practice, it is common not to know in advance which function f will be needed, or need estimates for multiple functions. In these cases it is useful to have a Markov chain to run estimates for various functions simultaneously. Thus, we would like the variance of our estimates, and thus the asymptotic variance, to be as low as possible for not just one function $f : \mathbf{X} \to \mathbf{R}$, but as low as possible for every function $f : \mathbf{X} \to \mathbf{R}$ simultaneously. This gives rise to the notion of an ordering of Markov kernels based on the asymptotic variance of functions in $L^2(\pi)$.

Given two Markov kernels P and Q with stationary distribution π, we say that P *efficiency dominates* Q if for every $f \in L^2(\pi)$, $v(f,P) \leq v(f,Q)$.

In this paper, we focus our attention on reversible Markov kernels. Many important algorithms, most notably the Metropolis-Hastings (MH) algorithm, are reversible (see [26], [21]). When the target probability density is difficult (or impossible) to calculate for the use of the MH algorithm, an exact approximation of the MH algorithm, an algorithm that uses the MH algorithm with an estimator of the target probability density, could be a viable option. In this case, it's possible this exact approximation algorithm run with one estimator efficiency dominate the same algorithm run with a different estimator. See [2] for more details.

Aside from Section 5, many of the results of this paper are known but are scattered in the literature, have incomplete or unclear proofs, or are missing proofs altogether. We present new clear, complete, and accessible proofs, using basic functional analysis where very technical results were previously used, most notably in the proof of Theorem 4.1. We show how once Theorem 4.1 is established, many further results are vastly simplified. This paper is self-contained assuming basic Markov chain theory and functional analysis.

1.1 Simple Examples

We now present two simple examples, using the theory that will be developped in the rest of the paper, in order to motivate and provide some explicit examples of the material.

Example 1.1. *As a simple example in finite state spaces, take* $\mathbf{X} = \{1,2,3\}$ *and* π *such that* $\pi(\{1\}) = 1/2$, $\pi(\{2\}) = 1/4$, *and* $\pi(\{3\}) = 1/4$. *Then let* P *and* Q *be Markov kernels on* $\mathbf{X}$ *such that*

$$\mathcal{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \textit{and} \qquad \mathcal{Q} = \left(\frac{1}{3}\right) \begin{pmatrix} 1 & 1 & 1 \\ 2 & \frac{1}{2} & \frac{1}{2} \\ 2 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Note that here we are using the expression of P *as*

$$\mathcal{P} = \begin{pmatrix} P(1,\{1\}) & P(1,\{2\}) & P(1,\{3\}) \\ P(2,\{1\}) & P(2,\{2\}) & P(2,\{3\}) \\ P(3,\{1\}) & P(3,\{2\}) & P(3,\{3\}) \end{pmatrix},$$

and similarly for Q.

In order to compare these two Markov kernels, it would be very difficult to calculate the asymptotic variance directly by definition. Even with the formula for the asymptotic variance from Theorem 3.1, which simplifies as $\mathbf{X}$ *is finite to* $v(f,P) = (a_2)^2 \frac{1+\lambda_2}{1-\lambda_2} + (a_3)^2 \frac{1+\lambda_3}{1-\lambda_3}$ *for every* $f : \mathbf{X} \to \mathbf{R}$, *where* λ_2 *and* λ_3 *are* *eigenvalues of $\mathcal{Q} - \mathcal{P}$, and a_2 and a_3 are the coefficients of the second and third eigenvectors of $\mathcal{Q} - \mathcal{P}$ in the eigenvector representation of f (see Proposition 2 of [19]), this is still a task that requires a lot of computation.*

However, by using Theorem 4.7, as $\mathbf{X}$ is finite we can simply calculate the eigenvalues of $\mathcal{Q} - \mathcal{P}$, which are $\{2/3, 0, 0\}$, and as they are all nonnegative, we can conclude that P efficiency dominates P.

Next we consider an example with a continuous state space.

Example 1.2. *Suppose $\mathbf{X} = \mathbf{R}$, and $\pi \sim Exponential(1)$, i.e. $\pi(dy) = f(y)dy$, where $f : \mathbf{R} \to [0, \infty)$ such that $f(y) = e^{-y}$ if $y \geq 0$ and $f(y) = 0$ if $y < 0$.*

Let $h : \mathbf{R} \to [0, \infty)$ be the density function of the $Normal(0, 1)$ distribution, i.e. $h(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$ for every $y \in \mathbf{R}$. Let R be the Markov kernel associated to i.i.d. sampling from the $Normal(0, 1)$ distribution, i.e. $R(x, dy) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = h(y)dy$ for every $x \in \mathbf{R}$.

Let Π denote the Markov kernel associated with i.i.d. sampling from π (i.e. $\Pi(x, dy) = \pi(dy) = e^{-y}dy = f(y)dy$ for every $x \in \mathbf{X}$), and let Q be the Markov kernel associated with the Metropolis-Hastings algorithm with proposal R (see Example 6.3 or [21] for more details). Expicitly, we have

$$Q(x, dy) \ = \ \alpha(x, y)R(x, dy) + r(x)\delta_x(dy),$$

where $\alpha(x, y) \ = \ \min\left\{1, \frac{f(y)h(x)}{f(x)h(y)}\right\}$, $r(x) \ = \ 1 - \int_{-\infty}^{\infty} \min\left\{1, \frac{f(y)h(x)}{f(x)h(y)}\right\} h(y)dy$, and $\delta_x(dy)$ is the point mass at $x \in \mathbf{R}$, for every $x, y \in \mathbf{R}$.

Trying to distinguish which algorithm, if either, of Π and Q efficiency dominates the other would be very difficult given only the definition of the asymptotic variance. However, with Theorem 6.2, we shall see that it is enough to show that for every $x \in \mathbf{R}$ and measurable set $A \subseteq \mathbf{R}$, $\Pi(x, A \setminus \{x\}) \geq Q(x, A \setminus \{x\})$, i.e. that Π Peskun-dominates Q (see the end of Section 1 below), in order to prove that Π efficiency dominates Q. With this in mind, we simply calculate

$$\begin{aligned} Q(x, A \setminus \{x\}) \ &= \ \int_A \alpha(x, y)h(y)dy \ = \ \int_{A \cap [0,\infty)} \alpha(x, y)h(y)dy \\ &\leq \ \int_{A \cap [0,\infty)} h(y)dy \ \leq \ \int_{A \cap [0,\infty)} f(y)dy \ = \ \int_A f(y)dy \\ &= \ \Pi(x, A) \ = \ \Pi(x, A \setminus \{x\}), \end{aligned}$$

as $f(y) = 0$ for every $y < 0$ and $f(y) \leq h(y)$ for every $y \geq 0$, thus proving Π efficiency dominates Q by Theorem 6.2.

Although these examples are simple, they demonstrate how the theory that follows allows for a much easier analysis and comparison of many algorithms, that would otherwise be impossible.

1.2 Outline of the Paper

In Section 3, we provide a full proof of the formula for the asymptotic variance of φ-irreducible reversible Markov kernels established by Kipnis and Varadhan in [15],

$$v(f,P)=\int_{[-1,1)}\frac{1+\lambda}{1-\lambda}\,\mathcal{E}_{f,\mathcal{P}}(d\lambda).$$

We also provide a full proof of a useful characterisation of the asymptotic variance for aperiodic Markov kernels.

In Section 4, we use the above formula as well as some functional analysis from Section 7, to show that efficiency dominance is equivalent to a much simpler condition for φ-irreducible reversible kernels; given Markov kernels P and Q, for every $f\in L_0^2(\pi)$, $\langle f,\mathcal{P}f\rangle\leq\langle f,\mathcal{Q}f\rangle$ (Theorem 4.1). This equivalent condition is sometimes called *covariance dominance* (see [17]). The functional analysis used in the proof of Theorem 4.1 is derived from the basics in Section 7. We use this theorem to show that antithetic Markov kernels are more efficient than i.i.d. sampling (Proposition 4.10), and show that efficiency dominance is a partial ordering (Theorem 4.12).

In Section 5, we generalise the results on the efficiency dominance of combined chains in [19] from finite state spaces to general state spaces. Given reversible Markov kernels $P_1,\ldots,P_l$ and $Q_1,\ldots,Q_l$, we show that if $\mathcal{Q}_k-\mathcal{P}_k$ is a positive operator on $L_0^2(\pi)$ for every k, and $\{\alpha_1,\ldots,\alpha_l\}$ is a set of mixing probabilities such that $P=\sum\alpha_kP_k$ and $Q=\sum\alpha_kQ_k$ are φ-irreducible, then P efficiency dominates Q (Theorem 5.1). This can be used to show that a random-scan Gibbs sampler with more efficient component kernels will always be more efficient. We also show that for two combined kernels differing in one component, one efficiency dominates the other if and only if it's unique component kernel efficiency dominates the other's (Corollary 5.2).

In Section 6, we consider Peskun dominance, or dominance off the diagonal (see [20], [26]). We say that a Markov kernel P *Peskun dominates* another kernel Q, if for π-a.e. $x\in\mathbf{X}$, for every measurable set A, $P(x,A\setminus\{x\})\geq Q(x,A\setminus\{x\})$. In Lemma 6.1, we follow the techniques of Tierney in [26] to show that if P Peskun dominates Q, then $\mathcal{Q}-\mathcal{P}$ is a positive operator. We then show that with Theorem 4.1 established, the proof that Peskun dominance implies efficiency dominance is simplified (Theorem 6.2).

2 Background

We are given the probability space $(\mathbf{X},\mathcal{F},\pi)$, where we assume the state space $\mathbf{X}$ is non-empty.

2.1 Markov Chain Background

A Markov kernel on $(\mathbf{X}, \mathcal{F})$ is a function $P : \mathbf{X} \times \mathcal{F} \to [0, \infty)$ such that $P(x, \cdot)$ is a probability measure for every $x \in \mathbf{X}$, and $P(\cdot, A)$ is a measurable function for every $A \in \mathcal{F}$. A time-homogeneous Markov chain $\{X_n\}_{n\in\mathbf{N}}$ on $(\mathbf{X}, \mathcal{F})$ has P as a Markov kernel if for every $n \in \mathbf{N}$, $\mathbf{P}\left(X_{n+1} \in A | \, X_n = x\right) = P(x, A)$, for every $x \in \mathbf{X}$ and $A \in \mathcal{F}$, i.e. the Markov kernel P describes the transition probabilities of the Markov chain $\{X_n\}_{n\in\mathbf{N}}$. The Markov kernel P is *stationary* with respect to π if $\int_{\mathbf{X}} P(x, A)\pi(dx) = \pi(A)$ for every $A \in \mathcal{F}$. Intuitively, this means that once the chain $\{X_n\}$ has reached the distribution π, $X_n \sim \pi$, then it stays at that distribution π forever, $X_{n+k} \sim \pi$ for every $k \in \mathbf{N}$. Thus if the chain $\{X_n\}$ is started in stationarity, i.e. $X_0 \sim \pi$, then it stays in stationarity, $X_n \sim \pi$ for every $n \in \mathbf{N}$.

The Markov kernel P is *reversible* with respect to π if $P(x, dy)\pi(dx) = P(y, dx)\pi(dy)$, i.e. the probability of starting at x distributed by π and then jumping to y is the same as the probability of starting at y distributed by π and then jumping to x. A Markov kernel P is *φ-irreducible* if there exists a non-zero σ-finite measure φ on $(\mathbf{X}, \mathcal{F})$ such that for every $A \in \mathcal{F}$ with $\varphi(A) > 0$, for every $x \in \mathbf{X}$ there exists $n \in \mathbf{N}$ such that $P^n(x, A) > 0$. Intuitively, the Markov chain has positive probability of eventually reaching every set of positive φ measure.

The space $L^2(\pi)$ is defined rigorously as the set of equivalence classes of π-square-integrable real-valued functions, with two functions f and g being equivalent if $f = g$ π-a.e., i.e. $f = g$ with probability 1. Less rigorously, $L^2(\pi)$ is simply the set of π-square-integrable real-valued functions. When this set is endowed with the inner-product $\langle \cdot, \cdot \rangle : L^2(\pi) \times L^2(\pi) \to \mathbf{R}$ such that $f \times g \mapsto \langle f, g \rangle := \int_{\mathbf{X}} f(x)g(x)\pi(dx)$, this space becomes a real Hilbert space (a vector space that is complete with respect to the norm generated by the inner product, $\|\cdot\| := \sqrt{\langle \cdot, \cdot \rangle}$). (When we are also dealing with complex functionals, we define the inner-product instead to be $f \times g \mapsto \langle f, g \rangle := \int_{\mathbf{X}} f(x)\overline{g(x)}\pi(dx)$, where $\overline{\alpha}$ is the complex conjugate of $\alpha \in \mathbf{C}$, and $L^2(\pi)$ becomes a complex Hilbert space. As we are only dealing with real-valued functions, we do not need this distinction.)

Recall from Section 1 that for a function $f \in L^2(\pi)$, it's *asymptotic variance* with respect to the Markov kernel P, denoted $v(f, P)$, is defined as $v(f, P) := \lim_{N\to\infty} \left[N \mathbf{Var}\left(\frac{1}{N}\sum_{k=1}^N f(X_k)\right)\right] = \lim_{N\to\infty} \left[\frac{1}{N}\mathbf{Var}\left(\sum_{k=1}^N f(X_k)\right)\right]$, where $\{X_k\}_{k\in\mathbf{N}}$ is a Markov chain with Markov kernel P started in stationarity. As is clear from our definition, the asymptotic variance is a measure of the variance of an MCMC estimate of the mean of f in the limit using the Markov chain $\{X_k\}$ with kernel P. Also from Section 1, recall that given two Markov kernels P and Q, both with stationary measure π, P *efficiency dominates* Q if $v(f, P) \leq v(f, Q)$

for every $f \in L^2(\pi)$. Thus if P efficiency dominates Q, we should expect to have better estimates using a Markov chain with kernel P rather than a Markov chain with kernel Q. As such, when deciding on a Markov kernel to use for MCMC estimation, if P efficiency dominates Q, we would rather use P (of course, this is if all other aspects of the kernels, like convergence to π, are similar).

For every Markov kernel P, we can define a linear operator $\mathcal{P}$ on the space of $\mathcal{F}$ measurable functions by

$$\mathcal{P}f(\cdot) := \int_{y \in \mathbf{X}} f(y)P(\cdot, dy).$$

For every Markov kernel, we denote the associated linear operator defined above by it's letter in calligraphics. If P is stationary with respect to π, the image of $\mathcal{P}$ restricted to $L^2(\pi)$ is contained in $L^2(\pi)$, as for every $f \in L^2(\pi)$, by Jensen's inequality

$$\int_{x \in \mathbf{X}} |\mathcal{P}f(x)|^2 \pi(dx) \leq \iint_{x,y \in \mathbf{X}} |f(y)|^2 P(x, dy)\pi(dx) = \int_{y \in \mathbf{X}} |f(y)|^2 \pi(dy) < \infty. \tag{2.1}$$

(2.1 is true for every $1 \leq r \leq \infty$, not just $r = 2$. See [3].) In this paper, we will only deal with functions in $L^2(\pi)$, and thus view $\mathcal{P}$ as a map from $L^2(\pi) \to L^2(\pi)$, or a subset thereof.

Notice that the constant function, $\mathbb{1} : \mathbf{X} \to \mathbf{R}$ such that $\mathbb{1}(x) = 1$ for every $x \in \mathbf{X}$, exists in $L^2(\pi)$ (as π is a probability measure), and furthermore, as $P(x, \cdot)$ is a probability measure for every $x \in \mathbf{X}$, $\mathcal{P}\mathbb{1} = \mathbb{1}$. Thus $\mathbb{1}$ is an eigenfunction of $\mathcal{P}$ with eigenvalue 1. We define the space $L_0^2(\pi)$ to be the subspace of $L^2(\pi)$ perpendicular to $\mathbb{1}$, or the subspace of $L^2(\pi)$ functions with zero mean, i.e. $L_0^2(\pi) := \{f \in L^2(\pi) | f \perp \mathbb{1}\} = \{f \in L^2(\pi) | \langle f, \mathbb{1} \rangle = 0\} = \{f \in L^2(\pi) | \mathbf{E}_\pi(f) = 0\}$. Notice that if $\mathcal{P}$ is restricted to $L_0^2(\pi)$ and P is stationary with respect to π, then it's range is contained in $L_0^2(\pi)$, as for every $f \in L_0^2(\pi)$, by Fubini's Theorem, $\langle \mathcal{P}f, \mathbb{1} \rangle = \int_{y \in \mathbf{X}} f(y) \int_{x \in \mathbf{X}} P(x, dy)\pi(dx) = \int_{\mathbf{X}} f(y)\pi(dy) = \langle f, \mathbb{1} \rangle = 0$.

When considering efficiency dominance, it is enough to look at the smaller subspace $L_0^2(\pi)$. If P and Q are Markov kernels with stationary measure π, P efficiency dominates Q if and only if P efficiency dominates Q on the smaller subspace $L_0^2(\pi)$. The forward implication is trivial as $L_0^2(\pi) \subseteq L^2(\pi)$, and for the converse, notice that for every $f \in L^2(\pi)$, $v(f, P) = v(f_0, P)$ and $v(f, Q) = v(f_0, Q)$, where $f_0 := f - \mathbf{E}_\pi(f) \in L_0^2(\pi)$. Thus when talking about efficiency dominance, we lose nothing by restricting ourselves to $L_0^2(\pi)$, and we get rid of the eigenfunction $\mathbb{1}$. Unless stated otherwise, we will consider $\mathcal{P}$ as an operator on and to $L_0^2(\pi)$.

A Markov kernel P is *periodic* with period $d \geq 2$ if there exists $\mathcal{X}_1, \ldots, \mathcal{X}_d \in \mathcal{F}$ such that $\mathcal{X}_k \cap \mathcal{X}_j = \emptyset$ for every $j \neq k$, and for every $i \in \{1, \ldots, d-1\}$,

$P(x, \mathcal{X}_{i+1}) = 1$ for every $x \in \mathcal{X}_i$ and $P(x, \mathcal{X}_1) = 1$ for every $x \in \mathcal{X}_d$. Intuitively, the Markov chain jumps from $\mathcal{X}_i$ to $\mathcal{X}_{i+1}$, then from $\mathcal{X}_{i+1}$ to $\mathcal{X}_{i+2}$ and so on. The sets $\mathcal{X}_1, \dots, \mathcal{X}_d \in \mathcal{F}$ described above are called a periodic decomposition of P. A Markov kernel P is *aperiodic* if it is not periodic.

A common definition related to the efficiency of Markov kernels and a common measure in time-series analysis is the *lag-k autocovariance*. For an $\mathcal{F}$-measurable function f, the lag-k autocovariance, denoted γ_k, is the covariance between $f(X_0)$ and $f(X_k)$, where $\{X_k\}_{k \in \mathbf{N}}$ is a Markov chain run from stationarity with kernel P, i.e. $\gamma_k := \mathbf{Cov}_{\pi,P}(f(X_0), f(X_k))$. When $\{X_k\}$ is a Markov chain run from stationarity, $\{f(X_k)\}$ is a stationary time-series, and the definition of the lag-k autocovariance is simply the regular definition from the time-series approach. When the function f is in $L_0^2(\pi)$, notice $\gamma_k = \mathbf{E}_{\pi,P}(f(X_0)f(X_k)) = \langle f, \mathcal{P}^k f \rangle$.

We denote the Markov kernel associated with i.i.d. sampling from π as Π, i.e. $\Pi : \mathbf{X} \times \mathcal{F} \to [0, \infty)$ such that $\Pi(x, A) = \pi(A)$ for every $x \in \mathbf{X}$ and $A \in \mathcal{F}$. Notice that for every $f \in L_0^2(\pi)$, $\Pi f(x) = \mathbf{E}_\pi(f) = 0$ for every $x \in \mathbf{X}$. Thus Π restricted to $L_0^2(\pi)$ is the zero function on $L_0^2(\pi)$.

2.2 Functional Analysis Background

Here we present some functional analysis that will be used throughout the paper. For a proper introduction to functional analysis, see Rudin's [23], or Conway's [6].

An operator T on a Hilbert space $\mathbf{H}$ (i.e. a linear function $T : \mathbf{H} \to \mathbf{H}$) is called *bounded* if there exists $C > 0$ such that for every $f \in \mathbf{H}$, $\|Tf\| \leq C \|f\|$. An elementary result from functional analysis shows that the operator T is continuous if and only if T is bounded in the above sense. In finite dimensional vector spaces, all linear operators are bounded and hence continuous. The same is not true in general. The *norm* of a bounded operator is defined as the smallest such constant $C > 0$ such that the above holds, i.e. $\|T\| := \inf\{C > 0 : \|Tf\| \leq C \|f\|, \forall f \in \mathbf{H}\}$. A bounded operator T is called *invertible* if it is bijective, and the inverse of T, T^{-1}, is bounded.

Unbounded operators on $\mathbf{H}$ are linear operators T such that there is no $C > 0$ such that $\|Tf\| \leq C \|f\|$ for every $f \in \mathbf{H}$. Oftentimes, unbounded operators T are not defined on the whole space $\mathbf{H}$, but only on a subset of $\mathbf{H}$. An unbounded operator T is *densely defined* if the domain of T is dense in $\mathbf{H}$.

The *adjoint* of a bounded operator T is the unique bounded operator T^* such that $\langle Tf, g \rangle = \langle f, T^* g \rangle$ for every $f, g \in \mathbf{H}$. Similarly, if T is a densely defined operator, then the *adjoint* of T is the linear operator T^* such that $\langle Tf, g \rangle = \langle f, T^* g \rangle$ for every $f \in \text{domain}(T)$ and $g \in \mathbf{H}$ such that $f \mapsto \langle Tf, g \rangle$ is a bounded linear functional on domain(T). Thus we define domain$(T^*) := \{g \in \mathbf{H} | f \mapsto \langle Tf, g \rangle$ is a bounded linear functional on domain$(T)\}$. These two definitions are equivalent when T is bounded.

A bounded operator T is called *normal* if T commutes with it's adjoint, $TT^* = T^*T$, and *self-adjoint* if T equals it's adjoint, $T = T^*$. Equivalently, a bounded operator T self-adjoint if $\langle Tf, g\rangle = \langle f, Tg\rangle$ for every f and $g \in \mathbf{H}$. A densely defined operator T is called *self-adjoint* if $T = T^*$ and $\text{domain}(T) = \text{domain}(T^*)$.

$\mathcal{P}$ restricted to $L^2(\pi)$ is self-adjoint if and only if P is reversible. As $L_0^2(\pi) \subseteq L^2(\pi)$, if P is reversible with respect to π, then $\mathcal{P}$ restricted to $L_0^2(\pi)$ is self-adjoint as well.

The *spectrum* of an operator T is the subset $\sigma(T) := \{\lambda \in \mathbf{C} | T - \lambda\mathcal{I} \text{ is not invertible}\}$ of the complex plane, where $\mathcal{I}$ is the identity operator. Note that in the above definition, invertible is meant in the context of bounded linear operators, i.e. T is bijective and the inverse of T, T^{-1}, is also bounded. If the operator T is self-adjoint, the spectrum of T is real, i.e. $\sigma(T) \subseteq \mathbf{R}$ (see Theorem 12.26 (a) in [23]). It is important to note that in the case the underlying Hilbert space of the operator T is finite-dimensional, as is the case for $L^2(\pi)$ and $L_0^2(\pi)$ when $\mathbf{X}$ is finite, that the spectrum of T is exactly the set of eigenvalues of T. When the Hilbert space is not finite-dimensional, the spectrum also includes limit points and points where $T - \lambda\mathcal{I}$ is not surjective.

An operator T on a Hilbert space $\mathbf{H}$ is called *positive* if for every $f \in \mathbf{H}$, $\langle f, Tf\rangle \geq 0$. As we shall see in Lemma 4.6, if T is bounded and normal, then T is positive if and only if the spectrum of T is positive, i.e. $\sigma(T) \subseteq [0, \infty)$. It is important to note that when the Hilbert space $\mathbf{H}$ is a real Hilbert space, it is not necessarily true that if T is positive and bounded then T is self-adjoint. This is an important distinction, as this is true when $\mathbf{H}$ is a complex Hilbert space.

Furthermore, as shown by inequality (2.1), (which is equivalent to $\|\mathcal{P}f\|^2 \leq \|f\|^2$), when P is stationary with respect to π, the norm of $\mathcal{P}$ on $L^2(\pi)$ is less than or equal to 1. This also bounds the spectrum of $\mathcal{P}$ to $\lambda \in \mathbf{C}$ such that $|\lambda| \leq \|\mathcal{P}\| \leq 1$. If P is reversible, then $\mathcal{P}$ is self-adjoint and thus the spectrum of $\mathcal{P}$ is real, $\sigma(\mathcal{P}) \subseteq \mathbf{R}$. Thus if P is reversible with respect to π, then $\sigma(\mathcal{P}) \subseteq [-1, 1]$.

Given a bounded self-adjoint operator T on the Hilbert space $\mathbf{H}$, by the Spectral Theorem (see Theorem 12.23 of [23], Chapter 9 Theorem 2.2 of [6]), we know that there exists a spectral measure $\mathcal{E}_T : \mathcal{B}(\sigma(T)) \to \mathfrak{B}(\mathbf{H})$ such that $T = \int_{\sigma(T)} \lambda \mathcal{E}_T(d\lambda)$. $\mathcal{B}(\sigma(T))$ denotes the Borel σ-field of $\sigma(T) \subseteq \mathbf{C}$ and $\mathfrak{B}(\mathbf{H})$ denotes the set of bounded operators on the Hilbert space $\mathbf{H}$. So when P is reversible, $\mathcal{P}$ is self-adjoint, and thus by the Spectral Theorem, $\mathcal{P} = \int_{\sigma(\mathcal{P})} \lambda\, \mathcal{E}_{\mathcal{P}}(d\lambda)$, where $\mathcal{E}_{\mathcal{P}}$ is the spectral measure of $\mathcal{P}$.

The spectral measure $\mathcal{E}_T$ satisfies (i) $\mathcal{E}_T(\emptyset) = 0$ and $\mathcal{E}_T({}^{(}T)) = \mathcal{I}$, (ii) for every $A \in \mathcal{B}(\sigma(T))$, $\mathcal{E}_T(A) \in \mathfrak{B}(\mathbf{H})$ is a self-adjoint projection, i.e. $\mathcal{E}_T(A) = \mathcal{E}_T(A)^2 = \mathcal{E}_T(A)^*$, (iii) for every $A_1, A_2 \in \mathcal{B}(\sigma(T))$, $\mathcal{E}_T(A_1 \cap A_2) = \mathcal{E}_T(A_1)\mathcal{E}_T(A_2)$, and (iv) for every sequence of disjoint subsets $\{A_n\}_{n\in\mathbf{N}} \subseteq \mathcal{B}(\sigma(T))$, $\mathcal{E}_T\left(\cup_{n\in\mathbf{N}} A_n\right) = \sum_{n\in\mathbf{N}} \mathcal{E}_T(A_n)$.

Although this definition and the decomposition of T above may seem complicated, recall from linear algebra that when $\mathbf{H}$ is finite-dimensional, we can decompose a self-adjoint operator T as $T = \sum_{k=0}^{n-1} \lambda_k P_k$, where $\{\lambda_k\}_{k=0}^{n-1} \subseteq \mathbf{C}$ are the eigenvalues of T and $P_k : \mathbf{H} \to \mathbf{H}$ is the projection onto the eigenspace of λ_k. This is not so different from the above, except that when $\mathbf{H}$ is allowed to be infinite-dimensional, this sum of projections becomes an integral of projections.

For every $f \in \mathbf{H}$, we define the induced measure $\mathcal{E}_{f,T}$ on $\mathbf{C}$ as $\mathcal{E}_{f,T}(A) := \langle f, \mathcal{E}_T(A) f\rangle$ for every Borel measurable set $A \subseteq \mathbf{C}$. Note that as $\mathcal{E}_T(A)$ is a self-adjoint projection for every Borel measurable set $A \subseteq \mathbf{C}$, the measure $\mathcal{E}_{f,T}$ on $\mathbf{C}$ is a positive measure. For every Borel measurable function $\phi : \mathbf{C} \to \mathbf{C}$, the operator $\phi(T)$ on $\mathbf{H}$ is defined as $\phi(T) := \int \phi(\lambda)\mathcal{E}_T(d\lambda)$. $\phi(T)$ is a bounded operator whenever the function ϕ is bounded. Putting this together with our definition of $\mathcal{E}_{f,T}$, for every bounded Borel measureable function $\phi : \mathbf{C} \to \mathbf{C}$ and for every $f \in \mathbf{H}$,

$$\langle f, \phi(T) f\rangle = \int_{\sigma(T)} \phi(\lambda)\mathcal{E}_{f,T}(d\lambda).$$

We will also usually assume that the Markov kernel P is φ-irreducible, as when P is φ-irreducible, the constant function is the only eigenfunction (up to a scalar multiple) of $\mathcal{P}$ with eigenvalue 1 (see Lemma 4.7.4 of [10]). Thus if P is φ-irreducible, by restricting ourselves to $L_0^2(\pi)$, we get rid of the eigenvalue 1, i.e. 1 is not an eigenvalue of $\mathcal{P}$ on $L_0^2(\pi)$. Note however that this does not mean that $1 \notin \sigma(\mathcal{P})$ when restricted to $L_0^2(\pi)$, as $(\mathcal{P} - \mathcal{I})^{-1}$ could still be unbounded.

3 Asymptotic Variance

We now provide a detailed proof of the formula for the asymptotic variance of φ-irreducible reversible Markov kernels, originally introduced by Kipnis and Varadhan in [15]. Also in this section, we prove another more familiar and practical characterisation of the asymptotic variance for φ-irreducible reversible aperiodic Markov kernels (see [12], [11], and [13]).

Theorem 3.1. *If P is a φ-irreducible reversible Markov kernel with stationary distribution π, then for every $f \in L_0^2(\pi)$,*

$$v(f,P) \;=\; \int_{\lambda\in[-1,1)} \frac{1+\lambda}{1-\lambda}\,\mathcal{E}_{f,\mathcal{P}}(d\lambda),$$

where $\mathcal{E}_{f,\mathcal{P}}$ is the measure induced by the spectral measure of $\mathcal{P}$ (see Section 2.2). Note however, that this may still diverge to ∞.

Proof. For every $f \in L_0^2(\pi)$, by expanding the squares of $\mathbf{Var}_{\pi,P}\left(\sum_{k=1}^N f(X_k)\right)$, as $\mathbf{E}_\pi(f) = 0$ (by definition of $L_0^2(\pi)$),

$$\frac{1}{N}\mathbf{Var}_{\pi,P}\left(\sum_{k=1}^N f(X_k)\right) \; = \; \|f\|^2 + 2\sum_{k=1}^N \left(\frac{N-k}{N}\right)\langle f, \mathcal{P}^k f\rangle. \tag{3.1}$$

Thus as $\langle f, \mathcal{P}^k f\rangle \; = \; \int_{\sigma(\mathcal{P})} \lambda^k \, \mathcal{E}_{f,\mathcal{P}}(d\lambda)$ for every $k \in \mathbf{N}$,

$$\begin{aligned} v(f,P) \; &= \; \lim_{N\to\infty}\left[\frac{1}{N}\mathbf{Var}\left(\sum_{n=1}^N f(X_n)\right)\right] \\ &= \; \lim_{N\to\infty}\left[\|f\|^2 + 2\sum_{k=1}^N \left(\frac{N-k}{N}\right)\langle f, \mathcal{P}^k f\rangle\right] \\ &= \; \|f\|^2 + 2\lim_{N\to\infty}\left[\int_{\lambda\in\sigma(\mathcal{P})}\sum_{k=1}^\infty \mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k \mathcal{E}_{f,\mathcal{P}}(d\lambda)\right]. \end{aligned} \tag{3.2}$$

In order to deal with the limit in (3.2), we split the integral over $\sigma(\mathcal{P})$ into three subsets, $(0,1)$, $(-1,0]$ and $\{-1\}$. (Recall from Section 2 that $\sigma(\mathcal{P}) \subseteq [-1,1]$, and notice that we do not need to worry about $\{1\}$, as 1 is not an eigenvalue of $\mathcal{P}$ on $L_0^2(\pi)$ as P is φ-irreducible (see Section 2.1), and thus $\mathcal{E}_{f,\mathcal{P}}(\{1\}) = 0$ by Lemma 3.2 below).

For the first two subsets, for every fixed $\lambda \in (-1,0] \cup (0,1) = (-1,1)$, notice that for every $N \in \mathbf{N}$,

$$\sum_{k=1}^\infty \left|\mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k\right| \le \sum_{k=1}^\infty |\lambda^k| = \frac{|\lambda|}{1-|\lambda|} < \infty,$$

so the sum is absolutely summable. Thus we can pull the pointwise limit through and show that for every $\lambda \in (-1,1)$, by the geometric series $\sum_{k=1}^\infty r^k = \frac{r}{1-r}$, $\lim_{N\to\infty}\left[\sum_{k=1}^\infty \mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k\right] = \frac{\lambda}{1-\lambda}$.

By the Monotone Convergence Theorem for $\lambda \in (0,1)$ and the Dominated Convergence Theorem for $\lambda \in (-1,0]$,

$$\lim_{N\to\infty}\left[\int_{\lambda\in(0,1)}\sum_{k=1}^\infty \mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k \mathcal{E}_{f,\mathcal{P}}(d\lambda)\right] \; = \; \int_{\lambda\in(0,1)}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda), \tag{3.3}$$

and

$$\lim_{N\to\infty}\left[\int_{\lambda\in(-1,0]}\sum_{k=1}^\infty \mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k \mathcal{E}_{f,\mathcal{P}}(d\lambda)\right] \; = \; \int_{\lambda\in(-1,0]}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda). \tag{3.4}$$

For the last case, the case of $\{-1\}$, notice that for every $N \in \mathbf{N}$, (simplifying the equation found in [19]), denoting the floor of $x \in \mathbf{R}$ as $\lfloor x \rfloor$,

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbf{1}_{k\leq N}(k)\frac{N-k}{N}(-1)^k \mathcal{E}_{f,\mathcal{P}}(\{-1\}) \;&=\; \mathcal{E}_{f,\mathcal{P}}(\{-1\})N^{-1}\sum_{k=1}^{N}\left[(-1)^k(N-k)\right] \\
=\; \mathcal{E}_{f,\mathcal{P}}(\{-1\})N^{-1}\sum_{m=1}^{\lfloor N/2 \rfloor}\left[(N-2m)-(N-2m+1)\right] \\
=\; \mathcal{E}_{f,\mathcal{P}}(\{-1\})N^{-1}\sum_{m=1}^{\lfloor N/2 \rfloor}(-1) \;=\; \left(\frac{\lfloor -N/2 \rfloor}{N}\right)\mathcal{E}_{f,\mathcal{P}}(\{-1\}).
\end{aligned}$$

Thus as $\lim_{N\to\infty} \frac{\lfloor -N/2 \rfloor}{N} = -1/2$, we have the pointwise limit

$$\begin{aligned}
\lim_{N\to\infty}\sum_{k=1}^{\infty} \mathbf{1}_{k\leq N}(k)\left(\frac{N-k}{N}\right)(-1)^k \mathcal{E}_{f,\mathcal{P}}(\{-1\}) \;&=\; \left(\frac{-1}{2}\right)\mathcal{E}_{f,\mathcal{P}}(\{-1\}) \\
&=\; \left(\frac{\lambda}{1-\lambda}\right)\mathcal{E}_{f,\mathcal{P}}(\{\lambda\})|_{\lambda=-1}. \qquad (3.5)
\end{aligned}$$

In order to split the integral in (3.2) into our three pieces, $(0,1)$, $(-1,0]$ and $\{-1\}$, and pull the limit through, we have to verify that if we do pull the limit through, we are not performing $\infty - \infty$. To verify this, it is enough to show that $\left|\lim_{N\to\infty}\int_{\lambda\in(-1,0]}\sum_{k=1}^{\infty}\mathbf{1}_{k\leq N}(k)\left(\frac{N-k}{N}\right)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right|$ and $\left|\lim_{N\to\infty}\sum_{k=1}^{\infty}\mathbf{1}_{k\leq N}(k)\left(\frac{N-k}{N}\right)\mathcal{E}_{f,\mathcal{P}}(\{-1\})\right|$ are finite. So by (3.4) and (3.5) above, we find that $\left|\lim_{N\to\infty}\int_{\lambda\in(-1,0]}\sum_{k=1}^{\infty}\mathbf{1}_{k\leq N}(k)\left(\frac{N-k}{N}\right)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right| = \left|\int_{\lambda\in(-1,0]}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right| < \infty$ and $\left|\lim_{N\to\infty}\sum_{k=1}^{\infty}\mathbf{1}_{k\leq N}(k)\left(\frac{N-k}{N}\right)\mathcal{E}_{f,\mathcal{P}}(\{-1\})\right| = \left|\left(\frac{-1}{2}\right)\mathcal{E}_{f,\mathcal{P}}(\{-1\})\right| < \infty$.

So, denoting $H(N,k) := \mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)$, by (3.3), (3.4) and (3.5),

$$
\begin{aligned}
&\lim_{N\to\infty}\left[\int_{\lambda\in[-1,1)}\sum_{k=1}^{\infty}\mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right]\\
&=\lim_{N\to\infty}\left[\sum_{k=1}^{\infty}H(N,k)(-1)^k\mathcal{E}_{f,\mathcal{P}}(\{-1\}) + \int_{\lambda\in(-1,0]}\sum_{k=1}^{\infty}H(N,k)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right.\\
&\qquad\left.+\int_{\lambda\in(0,1)}\sum_{k=1}^{\infty}H(N,k)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right]\\
&=\lim_{N\to\infty}\sum_{k=1}^{\infty}H(N,k)(-1)^k\mathcal{E}_{f,\mathcal{P}}(\{-1\}) + \lim_{N\to\infty}\int_{\lambda\in(-1,0]}\sum_{k=1}^{\infty}H(N,k)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\\
&\qquad+\lim_{N\to\infty}\int_{\lambda\in(0,1)}\sum_{k=1}^{\infty}H(N,k)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\\
&=\left(\frac{\lambda}{1-\lambda}\right)\mathcal{E}_{f,\mathcal{P}}(\{\lambda\})|_{\lambda=-1} + \int_{\lambda\in(-1,0]}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda) + \int_{\lambda\in(0,1)}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda)\\
&=\int_{\lambda\in[-1,1)}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda).
\end{aligned}
$$

Plugging this into (3.2), and as $\mathcal{E}_{f,\mathcal{P}}(\{1\}) = 0$ by Lemma 3.2 below, we have

$$
\begin{aligned}
v(f,P) &= \|f\|^2 + 2\lim_{N\to\infty}\left[\int_{\lambda\in\sigma(\mathcal{P})}\sum_{k=1}^{\infty}\mathbf{1}_{k\le N}(k)\left(\frac{N-k}{N}\right)\lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda)\right]\\
&= \int_{\lambda\in[-1,1)}\mathcal{E}_{f,\mathcal{P}}(d\lambda) + 2\int_{\lambda\in[-1,1)}\frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda)\\
&= \int_{\lambda\in[-1,1)}\frac{1+\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda).
\end{aligned}
$$

□

Lemma 3.2. *If P is a φ-irreducible Markov kernel reversible with respect to π, then for every $f \in L_0^2(\pi)$, $\mathcal{E}_{f,\mathcal{P}}(\{1\}) = 0$.*

Proof. As outlined in [10] Lemma 4.7.4, as P is φ-irreducible, the constant function is the only eigenfunction of $\mathcal{P}$ with eigenvalue 1. Thus 1 is not an eigenvalue of $\mathcal{P}$ when restricted to $L_0^2(\pi)$.

As seen in Theorem 12.29 (b) of [23], for every normal bounded operator T on a Hilbert space, if $\lambda \in \mathbf{C}$ is not an eigenvalue of T, then $\mathcal{E}_T(\{\lambda\}) = 0$. As P is reversible with respect to π, $\mathcal{P}$ is self-adjoint and thus also normal.

So applying the above theorem to $\mathcal{P}$, as $1 \in \mathbf{C}$ is not an eigenvalue of $\mathcal{P}$, $\mathcal{E}_{\mathcal{P}}(\{1\}) = 0$. Thus for every $f \in L_0^2(\pi)$, $\mathcal{E}_{f,\mathcal{P}}(\{1\}) = \langle f, \mathcal{E}_{\mathcal{P}}(\{1\})f\rangle = \langle f, 0\rangle = 0$. □

We now show that if P is aperiodic, then -1 cannot be an eigenvalue of $\mathcal{P}$.

Proposition 3.3. *Let P be a Markov kernel reversible with respect to π. If P is aperiodic then -1 is not an eigenvalue of $\mathcal{P} : L_0^2(\pi) \to L_0^2(\pi)$.*

Proof. We show the contrapositive. That is, we assume that -1 is an eigenvalue of $\mathcal{P}$, and show that P is not aperiodic, i.e. periodic.

Let $f \in L_0^2(\pi)$ be an eigenfunction of $\mathcal{P}$ with eigenvalue -1. As $\mathcal{P}$ is self-adjoint (as P is reversible), we can assume that f is real-valued. Let $\mathcal{X}_1 = \{x \in \mathbf{X} : f(x) > 0\} = f^{-1}((0,\infty))$ and $\mathcal{X}_2 = \{x \in \mathbf{X} : f(x) < 0\} = f^{-1}((-\infty,0))$. As f is $(\mathcal{F}, \mathcal{B}(\mathbf{R}))$-measureable where $\mathcal{B}(\mathbf{R})$ is the Borel σ-field on $\mathbf{R}$, $\mathcal{X}_1, \mathcal{X}_2 \in \mathcal{F}$.

As f is an eigenfunction, $f \neq 0$ π-a.e. As $f \in L_0^2(\pi)$,

$$0 = \mathbf{E}_\pi(f) = \int_{\mathbf{X}} f(x)\pi(dx) = \int_{\mathcal{X}_1} f(x)\pi(dx) + \int_{\mathcal{X}_2} f(x)\pi(dx).$$

So, the above combined with the fact that $f \neq 0$ π-a.e. gives us that $\pi(\mathcal{X}_1), \pi(\mathcal{X}_2) > 0$.

So, as $\mathcal{P}f = -f$ for π-a.e. $x \in \mathbf{X}$ and P is reversible with respect to π,

$$\int_{\mathcal{X}_1} f(x)\pi(dx) = -\int_{\mathcal{X}_2} f(x)\pi(dx) = \int_{\mathcal{X}_2} \mathcal{P}f(x)\pi(dx) = \int_{y \in \mathbf{X}} f(y)P(y, \mathcal{X}_2)\pi(dy).$$

Similarly, $\int_{\mathbf{X}} f(x)P(x,\mathcal{X}_1)\pi(dx) = \int_{\mathcal{X}_2} f(x)\pi(dx)$.

We now claim that P is periodic with respect to $\mathcal{X}_1$ and $\mathcal{X}_2$. So assume for a contradiction there exists $E \in \mathcal{F}$ such that $\pi(E) > 0$, $E \subseteq \mathcal{X}_1$ and for every $x \in E$, $P(x, \mathcal{X}_2) < 1$. Then by definition of $\mathcal{X}_2$,

$\int_{\mathbf{X}} f(x)P(x,\mathcal{X}_2)\pi(dx) = \int_{\mathcal{X}_1} f(x)\pi(dx) > \int_{\mathcal{X}_1} f(x)P(x,\mathcal{X}_2)\pi(dx)$
$\geq \int_{\mathbf{X}} f(x)P(x,\mathcal{X}_2)\pi(dx)$, a clear contradiction.

So for π-a.e. $x \in \mathcal{X}_1$, $P(x,\mathcal{X}_2) = 1$. Similarly, for π-a.e. $x \in \mathcal{X}_2$, $P(x,\mathcal{X}_1) = 1$.

Thus P is periodic with period $d \geq 2$. □

This proposition, combined with Theorem 3.1, gives us a characterization of $v(f,P)$ as sums of autocovariances, γ_k (recall from Section 2.1), when P is aperiodic (see [11], [12], [13]). Though this characterization will not be used in this paper, it is perhaps more common and easier to interpret from a statistical point of view.

Proposition 3.4. *If P is an aperiodic φ-irreducible Markov kernel reversible with respect to π, then for every $f \in L_0^2(\pi)$,*

$$v(f,P) \;=\; \|f\|^2 + 2\sum_{k=1}^{\infty}\langle f, \mathcal{P}^k f\rangle \;=\; \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k.$$

Remark. *Even if P is periodic, Proposition 3.4 is still true for all $f \in L_0^2(\pi)$ that are perpendicular to the eigenfunctions of $\mathcal{P}$ with eigenvalue -1. (This ensures $\mathcal{E}_{f,\mathcal{P}}(\{-1\}) = 0$).*

Proof. Let $f \in L_0^2(\pi)$. By Proposition 3.3, -1 is not an eigenvalue of $\mathcal{P}$. So, just as in the proof of Lemma 3.2, again by Theorem 12.29 (b) of [23], as $\mathcal{P}$ is self-adjoint and thus normal, $\mathcal{E}_{f,\mathcal{P}}(\{-1\}) = 0$. So by Theorem 3.1, $v(f,P) = \int_{\lambda\in[-1,1)} \frac{1+\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda) = \|f\|^2 + 2\int_{\lambda\in(-1,1)} \frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda)$.

Recalling the geometric series $\sum_{k=1}^{\infty}\lambda^k = \frac{\lambda}{1-\lambda}$ for $\lambda \in (-1,1)$, by the Monotone Convergence Theorem for $\lambda \in (0,1)$ and by the Dominated Convergenve Theorem for $\lambda \in (-1,0]$, we have

$$\int_{\lambda\in(-1,1)} \frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda) = \int_{\lambda\in(-1,1)} \sum_{k=1}^{\infty} \lambda^k \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \sum_{k=1}^{\infty} \int_{\lambda\in(-1,1)} \lambda^k \mathcal{E}_{f,\mathcal{P}}(d\lambda).$$

So the asymptotic variance becomes $v(f,P) = \|f\|^2 + 2\int_{\lambda\in(-1,1)} \frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda) = \|f\|^2 + 2\sum_{k=1}^{\infty}\int_{\lambda\in(-1,1)} \lambda^k\mathcal{E}_{f,\mathcal{P}}(d\lambda) = \|f\|^2 + 2\sum_{k=1}^{\infty}\langle f, \mathcal{P}^k f\rangle = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k$, as $\mathbf{E}_\pi(f) = 0$. □

Remark. *In the non-reversible case, it is not guaranteed that the asymptotic variance exists (either a real number or infinite). However, the existence of the "λ-asymptotic variance," $v_\lambda(f,P) := \|f\|^2 + 2\sum_{k=1}^{\infty}\lambda^k\langle f, \mathcal{P}^k f\rangle$, for $\lambda \in [0,1)$ (where λ is simply a parameter, not an element of the spectrum of $\mathcal{P}$), is guaranteed to exist (though may be infinite). As such, progress has been made comparing the λ-asymptotic variance of non-reversible kernels, rather than the asymptotic variance, as in [1], and we are left with the problem of showing $v_\lambda(f,P)$ converges to $v(f,P)$, i.e. $\lim_{\lambda\uparrow 1} v_\lambda(f,P) = v(f,P)$. Noting that as in [26] we can write $v_\lambda(f,P) = \langle f, (\mathcal{I}-\lambda\mathcal{P})^{-1}(\mathcal{I}-\lambda\mathcal{P})f\rangle$, an easy application of the Spectral Theorem and the Dominated and Monotone Convergence Theorems as we have done in the proofs of Theorem 3.1 and Proposition 3.4 shows that $\lim_{\lambda\uparrow 1} v_\lambda(f,P) = v(f,P)$ whenever P is φ-irreducible and reversible (not necessarily aperiodic).*

4 Efficiency Dominance Equivalence

In this section, we use some basic functional analysis to prove our most useful equivalent condition for efficiency dominance for φ-irreducible reversible Markov kernels, simplifying the proof and staying away from overly technical arguments. We then use this equivalent condition to show how reversible antithetic Markov kernels are more efficient than i.i.d. sampling, and show that efficiency dominance is a partial ordering on φ-irreducible reversible kernels.

We state the equivalent condition theorem here, and then cover a brief example before introducing the lemmas we need from functional analysis proving each direction of the equivalence. We prove said lemmas in Section 7.

Theorem 4.1. *If P and Q are φ-irreducible Markov kernels reversible with respect to π, then P efficiency dominates Q if and only if*

$$\langle f, \mathcal{P}f\rangle \leq \langle f, \mathcal{Q}f\rangle, \qquad \forall f \in L_0^2(\pi). \tag{4.1}$$

Remark. *The condition that P and Q be φ-irreducible can actually be dropped, i.e. if P and Q are simply Markov kernels reversible with respect to π, then P efficiency dominates Q if and only if (4.1) holds.*

This follows by recognizing that every function in the eigenspace of $\mathcal{P}$ with respect to the eigenvalue 1*, i.e. every $f \in E_{\mathcal{P},1} := \{f \in L_0^2(\pi) : \mathcal{P}f = f\} =$ $\mathrm{null}(\mathcal{P}-\mathcal{I})$, has unbounded asymptotic variance (by (3.1), $v(f,P) = \lim_{N\to\infty}[\|f\|^2$ $+ 2\sum_{k=1}^{N}(\frac{N-k}{N})\langle f, \mathcal{P}^k f\rangle] = \|f\|^2 [\lim_{N\to\infty}(1 + 2\sum_{k=1}^{N} \frac{N-k}{N})] = \infty$). Thus as eigenspaces are closed, $L_0^2(\pi) = E_{\mathcal{P},1} \bigoplus E_{\mathcal{P},1}^{\perp}$, where $E_{\mathcal{P},1}^{\perp} = \{g \in L_0^2(\pi) : \langle f, g\rangle =$ $0, \forall f \in E_{\mathcal{P},1}\}$ and $\bigoplus$ is the direct product of Hilbert spaces, and we can follow the same arguments as the φ-irreducible case on restricted subspcaes.*

Example 4.2 (Data Augmentation Algorithms and their Sandwich Variants)**.** *Say we want to estimate samples from the probability density $f_X : \mathbf{X} \to [0,\infty)$ on $(\mathbf{X}, \mathcal{X}, \mu)$. If we have access to conditional densities $f_{X|Y}(\cdot|y)$ and $f_{Y|X}(\cdot|x)$ on the spaces $(\mathbf{X}, \mathcal{X}, \mu)$ and another space $(\mathbf{Y}, \mathcal{Y}, \nu)$ respectively (where there exists a density $f : \mathbf{X} \times \mathbf{Y} \to [0,\infty)$ with $f_X(x) = \int_{\mathbf{Y}} f(x,y)\nu(dy)$) then we can implement a data augmentation (DA) MCMC algorithm to estimate from f_X.*

The DA MCMC algorithm works as follows. If $\{X_k\}_{k=0}^{\infty}$ is our Markov chain, after starting X_0 according to some initial distribution, given $X_k = x \in \mathbf{X}$, we sample from $f_{Y|X}(\cdot|x)$ to get a $y \in \mathbf{Y}$, and then sample again from $f_{X|Y}(\cdot|y)$ to get our value $x' \in \mathbf{X}$, and set $X_{k+1} = x'$. The Markov kernel according to this algorithm is $P_{DA}(x, dx') = \int_{\mathbf{Y}} f_{X|Y}(x'|y) f_{Y|X}(y|x)\nu(dy)\mu(dx')$

Oftentimes the DA algorithm can be inefficient, and we can improve performance by adding a middle step. Given a Markov kernel R on $(\mathbf{Y}, \mathcal{Y})$, after we have our sample $y \in \mathbf{Y}$ by sampling from $f_{Y|X}(\cdot|x)$, we sample from $R(y, \cdot)$ to get another sample in $\mathbf{Y}$, $y' \in \mathbf{Y}$, and then use this value to finish the loop by sampling from $f_{X|Y}(\cdot|y')$. This algorithm is dubbed the sandwich DA algorithm, as the middle step of sampling from $R(y, \cdot)$ is "sandwiched" between the outer steps of the DA algorithm. This MCMC algorithm has Markov kernel $P_S(x, dx') =$ $\iint_{\mathbf{Y}} f_{X|Y}(x'|y') R(y, dy') f_{Y|X}(y|x)\nu(dy)\mu(dx')$.

The following diagram illustrates the DA and sandwich DA MCMC algorithms, *with the DA algorithm on the left and the sandwich DA algorithm on the right.*

$$\begin{tikzcd}
\mathbf{Y} \arrow[r, bend right, "f_{X|Y}"'] & \mathbf{X} \arrow[l, bend right, "f_{Y|X}"'] \arrow[r, bend left, "f_{Y|X}"] & \mathbf{Y} \arrow[l, bend left, "f_{X|Y}"] \arrow[loop right, "R"]
\end{tikzcd}$$

This general form of the sandwich DA algorithm was first introduced in [14]. For a more detailed analysis and background on the comparison between the DA and sandwich DA algorithms, see [14]. We follow their analysis here to show that under suitable conditions, the sandwich DA algorithm efficiency dominates the DA algorithm.

Using the definition of conditional densities, it is not hard to see that $P_{DA}(x, dx')f_X(x)\mu(dx) = P_{DA}(x', dx)f_X(x')\mu(dx')$*, i.e. that the DA algorithm is reversible with respect to* $f_X\mu$*. If the Markov kernel* R *is reversible with respect to* $f_Y\nu$*, a similar argument will show that the sandwich DA algorithm is reversible with respect to* $f_X\mu$ *as well.*

We let $T_X : L^2_0(\mathbf{Y}, \mathcal{Y}, f_Y\nu) \to L^2_0(\mathbf{X}, \mathcal{X}, f_X\mu)$ *such that*

$$Th(x) = \int_{y\in\mathbf{Y}} h(y)f_{Y|X}(y|x)\nu(dy), \quad \forall h \in L^2_0(f_Y), \forall x \in \mathbf{X}$$

(where $L^2_0(f_Y) := L^2_0(\mathbf{Y}, \mathcal{Y}, f_Y\nu)$*). We similarly define* $T_X : L^2_0(f_X) \to L^2_0(f_Y)$*. Then notice that given the DA kernel* P_{DA} *and the sandwich DA kernel* P_S*, we have*

$$\mathcal{P}_{DA} = T_X T_Y, \quad \mathcal{P}_S = T_X \mathcal{R} T_Y, \quad T_X^* = T_Y. \tag{4.2}$$

A common condition on the Markov kernel R *is that it is idempotent, i.e.* $\int_{z\in\mathbf{Y}} R(y, dz)R(z, dy') = R(y, dy')$ *for every* $y \in \mathbf{Y}$*, or equivalently,* $\mathcal{R}^2 = \mathcal{R}$*. So, if the sandwich algorithm as given is such that* R *is reversible and idempotent, then notice that by (4.2), for every* $g \in L^2_0(f_X)$*,*

$$\begin{aligned}\langle g, \mathcal{P}_S g\rangle &= \langle g, T_X\mathcal{R}T_Y g\rangle = \langle T_Y g, \mathcal{R}T_Y g\rangle = \langle \mathcal{R}T_Y g, \mathcal{R}T_Y g\rangle \\ &\leq \langle T_Y g, T_Y g\rangle = \langle g, T_X T_Y g\rangle = \langle g, \mathcal{P}_{DA} g\rangle.\end{aligned}$$

Thus $\mathcal{P}_S$ *and* $\mathcal{P}_{DA}$ *satisfy (4.1), and thus be Theorem 4.1 the sandwich DA algorithm efficiency dominates the DA algorithm.*

To prove the “if” direction of Theorem 4.1, we follow an approach similar to the one seen in [17], but with some notable differences. The most important difference, is the use of the following lemma instead of some overly technical results from [4].

The following lemma, Lemma 4.3, is a generalisation of some results found in Chapter V of [5] from finite dimensional vector spaces to general Hilbert spaces. The finite dimensional version of Lemma 4.3 is also presented in [19] as Lemma 24.

Lemma 4.3. *If T and N are self-adjoint bounded linear operators on a Hilbert space $\mathbf{H}$, such that $\sigma(T), \sigma(N) \subseteq (0, \infty)$, then $\langle f, Tf\rangle \leq \langle f, Nf\rangle$ for every $f \in \mathbf{H}$, if and only if $\langle f, T^{-1}f\rangle \geq \langle f, N^{-1}f\rangle$, for ever $f \in \mathbf{H}$.*

Lemma 4.3 is proven in Section 7.1, where we discuss the differences in Lemma 4.3 between finite dimensional (as shown in [19]) and general Hilbert spaces.

We now prove the "if" direction of Theorem 4.1 with the help of Lemma 4.3.

Proof of "if" Direction of Theorem 4.1.

Say $\mathcal{P}$ and $\mathcal{Q}$ satisfy (4.1). For every $\eta \in [0, 1)$, let $T_{\mathcal{P},\eta} = \mathcal{I} - \eta\mathcal{P}$ and $T_{\mathcal{Q},\eta} = \mathcal{I} - \eta\mathcal{Q}$. Then as $\|\mathcal{P}\|, \|\mathcal{Q}\| \leq 1$, by the Cauchy-Schwartz inequality, for every $f \in L_0^2(\pi)$, $|\langle f, \mathcal{P}f\rangle| \leq \|f\|^2$ and $|\langle f, \mathcal{Q}f\rangle| \leq \|f\|^2$. So again by the Cauchy-Schwartz inequality, for every $f \in L_0^2(\pi)$,

$$\begin{aligned}\|T_{\mathcal{P},\eta}f\| \, \|f\| &\geq |\langle T_{\mathcal{P},\eta}f, f\rangle| \\ &= \left|\|f\|^2 - \eta\langle f, \mathcal{P}f\rangle\right| \\ &\geq |1-\eta| \, \|f\|^2 .\end{aligned}$$

Thus for every $f \in L_0^2(\pi)$, $\|T_{\mathcal{P},\eta}f\|, \|T_{\mathcal{Q},\eta}f\| \geq |1-\eta| \, \|f\|$. As $\eta \in [0, 1)$, $|1 - \eta| > 0$, and as $T_{\mathcal{P},\eta}$ and $T_{\mathcal{Q},\eta}$ are both normal (as they are self-adjoint) $T_{\mathcal{P},\eta}$ and $T_{\mathcal{Q},\eta}$ are both invertible, in the sense of bounded linear operators, i.e. the inverse of $T_{\mathcal{P},\eta}$ and $T_{\mathcal{Q},\eta}$ are bounded, and $0 \notin \sigma(T_{\mathcal{P},\eta}), \sigma(T_{\mathcal{Q},\eta})$ (by Lemma 7.2).

As $\|\mathcal{P}\|, \|\mathcal{Q}\| \leq 1$ and $\mathcal{P}$ and $\mathcal{Q}$ are self-adjoint (as P and Q are reversible), $\sigma(\mathcal{P}), \sigma(\mathcal{Q}) \subseteq [-1, 1]$. Thus for every $\eta \in [0, 1)$, $\sigma(T_{\mathcal{P},\eta}), \sigma(T_{\mathcal{Q},\eta}) \subseteq (0, 2) \subseteq (0, \infty)$.

So for every $\eta \in [0, 1)$, as $T_{\mathcal{P},\eta}$ and $T_{\mathcal{Q},\eta}$ are both self-adjoint, and for every $f \in L_0^2(\pi)$, $\langle f, T_{\mathcal{Q},\eta}f\rangle = \|f\|^2 - \eta\langle f, \mathcal{Q}f\rangle \leq \|f\|^2 - \eta\langle f, \mathcal{P}f\rangle = \langle f, T_{\mathcal{P},\eta}f\rangle$, by Lemma 4.3, $\langle f, T_{\mathcal{Q},\eta}^{-1}f\rangle \geq \langle f, T_{\mathcal{P},\eta}^{-1}f\rangle$, for every $f \in L_0^2(\pi)$.

Notice that for every $\eta \in [0, 1)$, $T_{\mathcal{P},\eta}^{-1} = (\mathcal{I} - \eta\mathcal{P})(\mathcal{I} - \eta\mathcal{P})^{-1} + \eta\mathcal{P}(\mathcal{I} - \eta\mathcal{P})^{-1} = \mathcal{I} + \eta\mathcal{P}(\mathcal{I} - \eta\mathcal{P})^{-1}$. So, for every $f \in L_0^2(\pi)$,

$$\begin{aligned}\|f\|^2 + \eta\langle f, \mathcal{P}(\mathcal{I} - \eta\mathcal{P})^{-1} f\rangle &= \langle f, T_{\mathcal{P},\eta}^{-1}f\rangle \\ &\leq \langle f, T_{\mathcal{Q},\eta}^{-1}f\rangle = \|f\|^2 + \eta\langle f, \mathcal{Q}(\mathcal{I} - \eta\mathcal{Q})^{-1} f\rangle,\end{aligned}$$

so for every $f \in L_0^2(\pi)$, $\langle f, \mathcal{P}(\mathcal{I} - \eta\mathcal{P})^{-1} f\rangle \leq \langle f, \mathcal{Q}(\mathcal{I} - \eta\mathcal{Q})^{-1} f\rangle$.

Thus by the Monotone Convergence Theorem for $\lambda \in (0, 1)$ and the Dominated Convergence Theorem for $\lambda \in [-1, 0]$, for every $f \in L_0^2(\pi)$, $\lim_{\eta\to1^-} \int_{[-1,1)} \frac{\lambda}{1-\eta\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda) = \int_{[-1,1)} \frac{\lambda}{1-\lambda}\mathcal{E}_{f,\mathcal{P}}(d\lambda)$, and similarly for $\mathcal{Q}$.

As P and Q are φ-irreducible and reversible with respect to π, by Theorem 3.1, for every $f \in L_0^2(\pi)$,

$$\begin{aligned} v(f, P) &= \int_{[-1,1)} \frac{1+\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \|f\|^2 + 2\int_{[-1,1)} \frac{\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda) \\ &= \|f\|^2 + 2 \lim_{\eta \to 1^-} \int_{[-1,1)} \frac{\lambda}{1-\eta\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \|f\|^2 + 2 \lim_{\eta \to 1^-} \langle f, \mathcal{P}\left(\mathcal{I} - \eta\mathcal{P}\right)^{-1} f\rangle \\ &\leq \|f\|^2 + 2 \lim_{\eta \to 1^-} \langle f, \mathcal{Q}\left(\mathcal{I} - \eta\mathcal{Q}\right)^{-1} f\rangle = \|f\|^2 + 2 \lim_{\eta \to 1^-} \int_{[-1,1)} \frac{\lambda}{1-\eta\lambda} \mathcal{E}_{f,\mathcal{Q}}(d\lambda) \\ &= \|f\|^2 + 2\int_{[-1,1)} \frac{\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{Q}}(d\lambda) = \int_{[-1,1)} \frac{1+\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{Q}}(d\lambda) = v(f, Q). \end{aligned}$$

So as $v(f, P) \leq v(f, Q)$ for every $f \in L_0^2(\pi)$, $v(f, P) \leq v(f, Q)$ for every $f \in L^2(\pi)$ (see Section 2.1), thus P efficiency dominates Q. □

Remark. *For the other direction of Theorem 4.1, it is hard to make use of any arguments utilising Lemma 4.3. If P effieciency dominates Q, we are given that each* individual $f \in L_0^2(\pi)$ *satisfies* $\lim_{\eta \to 1^-} \langle f, T_{\mathcal{P},\eta}^{-1} f\rangle \leq \lim_{\eta \to 1^-} \langle f, T_{\mathcal{Q},\eta}^{-1} f\rangle$. *As we can't apply Lemma 4.3 to the limits above, it seems the most we can do is fix an $\epsilon > 0$, and by the above limit for every $f \in L_0^2(\pi)$ there exists $\eta_f \in [0, 1)$ such that for every $\eta_f \leq \eta < 1$, $\langle f, T_{\mathcal{P},\eta}^{-1} f\rangle \leq \langle f, T_{\mathcal{Q},\eta}^{-1} f\rangle + \epsilon \|f\|^2 = \langle f, \left(T_{\mathcal{Q},\eta_f}^{-1} + \epsilon\mathcal{I}\right) f\rangle$. However, this results in a possible different η_f for every $f \in L_0^2(\pi)$, and it's not obvious that* $\sup\{\eta_f : f \in L_0^2(\pi)\} < 1$, *leaving us with no single $\eta \in [0, 1)$ such that the above inequality holds for every $f \in L_0^2(\pi)$ to allow us to apply Lemma 4.3.*

Due to the difficulties in applying Lemma 4.3 in the only if direction of Theorem 4.1, we make use of the following lemma, which appears as Corollary 3.1 from [17].

Lemma 4.4. *Let T and N be injective self-adjoint positive bounded linear operators on the Hilbert space $\mathbf{H}$ (though $T^{-1/2}$ and $N^{-1/2}$ may be unbounded). If* $\text{domain}(T^{-1/2}) \subseteq \text{domain}(N^{-1/2})$ and *for every $f \in$* $\text{domain}(T^{-1/2})$ *we have* $\left\|N^{-1/2} f\right\| \leq \left\|T^{-1/2} f\right\|$, *then $\langle f, Tf\rangle \leq \langle f, Nf\rangle$ for every $f \in \mathbf{H}$.*

See Section 7.2 for the proof of Lemma 4.4.

We must also make use of the fact that the space of functions with finite asymptotic variance is the domain of $(\mathcal{I} - \mathcal{P})^{-1/2}$. This was first stated in [15], using a test function argument. We take a different approach and provide a proof using the ideas of the Spectral Theorem. In particular, it uses some ideas from the proof of Theorem X.4.7 from [6].

Lemma 4.5. *If P is a φ-irreducible Markov kernel reversible with respect to π, then*

$$\left\{f \in L_0^2(\pi) : v(f, P) < \infty\right\} = \text{domain}\left((\mathcal{I} - \mathcal{P})^{-1/2}\right).$$

Proof. Let $\phi : [-1,1] \to \mathbf{R}$ such that $\phi(\lambda) = (1-\lambda)^{-1/2}$ for every $\lambda \in [-1,1)$ and $\phi(1) = 0$.

Even though ϕ is unbounded, it still follows from spectral theory that $\int \phi d\mathcal{E}_{\mathcal{P}} = \int (1-\lambda)^{-1/2} \mathcal{E}_{\mathcal{P}}(d\lambda) = (\mathcal{I} - \mathcal{P})^{-1/2}$, the inverse operator of $(\mathcal{I} - \mathcal{P})^{1/2}$, including equality of domains (for a formal argument, see [8] Theorem XII.2.9). Thus our problem reduces to showing that $\{f \in L_0^2(\pi) : v(f,P) < \infty\} = \text{domain}\left(\int \phi d\mathcal{E}_{\mathcal{P}}\right)$.

Now notice that for every $f \in L_0^2(\pi)$, as $\mathcal{E}_{f,\mathcal{P}}(\{1\}) = 0$ by Lemma 3.2,

$$\int_{[-1,1)} \left(\frac{1}{1-\lambda}\right) \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \int_{[-1,1)} |\phi(\lambda)|^2 \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \int |\phi|^2 d\mathcal{E}_{f,\mathcal{P}}.$$

Thus by Theorem 3.1, for every $f \in L_0^2(\pi)$, $v(f,P) = \int_{[-1,1)} \left(\frac{1+\lambda}{1-\lambda}\right) \mathcal{E}_{f,\mathcal{P}}(d\lambda)$, so

$$v(f,P) < \infty \qquad \textit{if and only if} \qquad \int_{[-1,1)} \left(\frac{1}{1-\lambda}\right) \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \int |\phi|^2 d\mathcal{E}_{f,\mathcal{P}} < \infty.$$

Thus we would like to show that $\int |\phi|^2 d\mathcal{E}_{f,\mathcal{P}}$ is finite if and only if $f \in \text{domain}\left(\int \phi d\mathcal{E}_{\mathcal{P}}\right)$.

For every $n \in \mathbf{N}$, let $\phi_n := \mathbf{1}_{(|\phi|<n)}\, \phi$ and $\Delta_n := \phi^{-1}(-n,n) = \phi_n^{-1}(\mathbf{R})$. Then notice $\cup_{k=1}^{\infty} \Delta_n = \mathbf{R}$ and Δ_n is a Borel set for every n as ϕ is Borel measurable.

As ϕ is positive, notice $\phi_n \leq \phi_{n+1}$ for every n. Thus as $\phi_n \to \phi$ pointwise for every $\lambda \in \sigma(\mathcal{P})$, by the Monotone Convergence Theorem,

$$\int |\phi_n|^2 d\mathcal{E}_{f,\mathcal{P}} \to \int |\phi|^2 d\mathcal{E}_{f,\mathcal{P}}. \tag{4.3}$$

As ϕ_n is bounded for every n, by definition of ϕ_n we have

$$\begin{aligned}
\int |\phi_n|^2 d\mathcal{E}_{f,\mathcal{P}} &= \left\| \left(\int \phi_n d\mathcal{E}_{\mathcal{P}} \right) f \right\|^2 \\
&= \left\| \left(\int \phi d\mathcal{E}_{\mathcal{P}} \right) \mathcal{E}_{\mathcal{P}} \left(\cup_{k=1}^{n} \Delta_k \right) f \right\|^2 \\
&= \left\| \mathcal{E}_{\mathcal{P}} \left(\cup_{k=1}^{n} \Delta_k \right) \left(\int \phi d\mathcal{E}_{\mathcal{P}} \right) f \right\|^2 .
\end{aligned}$$

Thus as $\mathcal{E}_{\mathcal{P}}(\cup_{k=1}^{n} \Delta_k) \to \mathcal{E}_{\mathcal{P}}(\mathbf{R}) = \mathcal{I}$ as $n \to \infty$ in the strong operator topology (i.e. $\|\mathcal{E}_{\mathcal{P}}(\cup_{k=1}^{n} \Delta_k) f - f\| \to 0$ for every $f \in L_0^2(\pi)$), we have

$$\int |\phi_n|^2 d\mathcal{E}_{f,\mathcal{P}} = \left\| \mathcal{E}_{\mathcal{P}} \left(\cup_{k=1}^{n} \Delta_k \right) \left(\int \phi d\mathcal{E}_{\mathcal{P}} \right) f \right\|^2 \to \left\| \left(\int \phi d\mathcal{E}_{\mathcal{P}} \right) f \right\|^2 . \tag{4.4}$$

Thus by (4.3) and (4.4) we have

$$\int |\phi|^2 \, d\mathcal{E}_{f,\mathcal{P}} \;=\; \lim_{n\to\infty} \int |\phi_n|^2 \, d\mathcal{E}_{f,\mathcal{P}} \;=\; \left\| \left(\int \phi d\mathcal{E}_{\mathcal{P}} \right) f \right\|^2 . \tag{4.5}$$

Thus as $\left\| \left(\int \phi d\mathcal{E}_{\mathcal{P}} \right) f \right\|^2 < \infty$ *if and only if* $f \in$ domain $\left(\int \phi d\mathcal{E}_{\mathcal{P}} \right)$, this completes the proof. □

Remark. *In [15], Kipnis and Varadhan state that* $\{f \in L_0^2(\pi) : v(f,P) < \infty\} =$ range $\left[(\mathcal{I} - \mathcal{P})^{1/2} \right]$. *As* range $\left[(\mathcal{I} - \mathcal{P})^{1/2} \right]$ $=$ domain $\left[(\mathcal{I} - \mathcal{P})^{-1/2} \right]$, *these are equivalent.*

Now we are ready to prove the "only if" direction of Theorem 4.1, as outlined in [17]. Recall the "only if" direction of Theorem 4.1; if P and Q are φ-irreducible Markov kernels reversible with respect to π, such that P efficiency dominates Q, then $\langle f, \mathcal{P} f \rangle \le \langle f, \mathcal{Q} f \rangle$, for every $f \in L_0^2(\pi)$.

Proof of "only if" Direction of Theorem 4.1.

As P efficiency dominates Q, if $f \in L_0^2(\pi)$ such that $v(f,Q) < \infty$, then $v(f,P) \le v(f,Q) < \infty$. Thus $\{f \in L_0^2(\pi) : v(f,Q) < \infty\} \subseteq \{f \in L_0^2(\pi) : v(f,P) < \infty\}$. So by Lemma 4.5, we have

$$\text{domain} \left[(\mathcal{I} - \mathcal{Q})^{-1/2} \right] \subseteq \text{domain} \left[(\mathcal{I} - \mathcal{P})^{-1/2} \right]. \tag{4.6}$$

By Theorem 3.1, for every $f \in L_0^2(\pi)$, $v(f,P) = \int_{[-1,1)} \frac{1+\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda)$ and $v(f,Q)$ $= \int_{[-1,1)} \frac{1+\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{Q}}(d\lambda)$. Thus as $\int_{[-1,1)} \frac{1+\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda) \;=\; \|f\|^2 + 2 \int_{[-1,1)} \frac{\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda)$ and $\int_{[-1,1)} \frac{1}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda) \;=\; \|f\|^2 + \int_{[-1,1)} \frac{\lambda}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda)$ and similarly for Q, as P efficiency dominates Q, for every $f \in L_0^2(\pi)$,

$$\int_{[-1,1)} \frac{1}{1-\lambda} \mathcal{E}_{f,\mathcal{P}}(d\lambda) \le \int_{[-1,1)} \frac{1}{1-\lambda} \mathcal{E}_{f,\mathcal{Q}}(d\lambda). \tag{4.7}$$

Furthermore, by (4.5) in the proof of Lemma 4.5, (recall that $\phi(\lambda) = (1 - \lambda)^{-1/2}$ in the proof of Lemma 4.5), for every $f \in$ domain $\left[(\mathcal{I} - \mathcal{P})^{-1/2} \right]$,

$$\int_{[-1,1)} \left(\frac{1}{1-\lambda} \right) \mathcal{E}_{f,\mathcal{P}}(d\lambda) \;=\; \left\| (\mathcal{I} - \mathcal{P})^{-1/2} f \right\|^2,$$

and similarly $\int_{[-1,1)} \left(\frac{1}{1-\lambda} \right) \mathcal{E}_{f,\mathcal{Q}}(d\lambda) = \left\| (\mathcal{I} - \mathcal{Q})^{-1/2} f \right\|^2$. Thus by (4.6) and (4.7), for every $f \in$ domain $\left[(\mathcal{I} - \mathcal{Q})^{-1/2} \right]$,

$$\left\| (\mathcal{I} - \mathcal{P})^{-1/2} f \right\|^2 \le \left\| (\mathcal{I} - \mathcal{Q})^{-1/2} f \right\|^2 .$$

So, by Lemma 4.4, with $T = (\mathcal{I} - \mathcal{Q})$ and $N = (\mathcal{I} - \mathcal{P})$, for every $f \in L_0^2(\pi)$, $\langle f, (\mathcal{I} - \mathcal{Q})\, f\rangle \leq \langle f, (\mathcal{I} - \mathcal{P})\, f\rangle$, and thus

$$\langle f, \mathcal{P} f\rangle \leq \langle f, \mathcal{Q} f\rangle.$$

□

Condition (4.1) is clearly equivalent to $\langle f, (\mathcal{Q} - \mathcal{P})\, f\rangle \geq 0$ for every $f \in L_0^2(\pi)$, i.e. equivalent to $\mathcal{Q} - \mathcal{P}$ being a positive operator. We can relate this back to the spectrum of the operator $\mathcal{Q} - \mathcal{P}$ with the following lemma.

Lemma 4.6. *If T is a bounded self-adjoint linear operator on a Hilbert space* $\mathbf{H}$*, then T is positive if and only if $\sigma(T) \subseteq [0, \infty)$.*

Proof. For the forward direction, if $\lambda < 0$, then for every $f \in \mathbf{H}$ such that $f \neq 0$, by the Cauchy-Schwartz inequality,

$$\begin{aligned}
\|(T - \lambda)f\|\, \|f\| &\geq |\langle (T - \lambda)f, f\rangle| \\
&= |\langle Tf, f\rangle - \lambda \|f\|^2 | \\
&\geq \langle Tf, f\rangle + |\lambda|\, \|f\|^2 && \text{(as } \lambda < 0 \text{ and by assumption)} \\
&\geq |\lambda|\, \|f\|^2 . && \text{(by assumption)}
\end{aligned}$$

Thus as $f \neq 0$ and $\lambda \neq 0$,

$$\|(T - \lambda)f\| \geq |\lambda|\, \|f\| > 0.$$

Thus as $T - \lambda$ is normal (as it is self-adjoint), $T - \lambda$ is invertible (by Lemma 7.2), and $\lambda \notin \sigma(T)$ by definition.

For the converse, if $\sigma(T) \subseteq [0, \infty)$, then as $\mathcal{E}_{f,T}$ is a positive measure for every $f \in \mathbf{H}$ (as $\mathcal{E}_T(A)$ is a self-adjoint projection for every Borel set A),

$$\langle f, Tf\rangle = \int_{\lambda \in \sigma(T)} \lambda \mathcal{E}_{f,T}(d\lambda) = \int_{\lambda \in [0,\infty)} \lambda \mathcal{E}_{f,T}(d\lambda) \geq 0, \qquad f \in \mathbf{H}.$$

□

This gives us the following.

Theorem 4.7. *If P and Q are φ-irreducible Markov kernels reversible with respect to π, then P efficiency dominates Q if and only if $\sigma(\mathcal{Q} - \mathcal{P}) \subseteq [0, \infty)$.*

Proof. By Theorem 4.1, P efficiency-dominates Q *if and only if* $\mathcal{P}$ and $\mathcal{Q}$ satisfy (4.1). As $\mathcal{P}$ and $\mathcal{Q}$ are both bounded linear operators, this is equivalent to

$$\langle f, (\mathcal{Q} - \mathcal{P})f\rangle \geq 0, \quad \text{for every } f \in L_0^2(\pi),$$

or in other words equivalent to $\mathcal{Q} - \mathcal{P}$ being a positive operator.

Thus as $\mathcal{Q} - \mathcal{P}$ is a bounded self-adjoint linear operator on $L_0^2(\pi)$, by Lemma 4.6, $\mathcal{Q} - \mathcal{P}$ is a positive operator *if and only if* $\sigma(\mathcal{Q} - \mathcal{P}) \subseteq [0, \infty)$. □

Proposition 4.8. *If P and Q are φ-irreducible Markov kernels reversible with respect to π such that $\sup\sigma(\mathcal{P}) \leq \inf\sigma(\mathcal{Q})$, then P efficiency dominates Q.*

Proof. Firstly, notice that for every $f \in L^2_0(\pi)$,

$$\langle f, \mathcal{P}f\rangle = \int_{\sigma(\mathcal{P})} \lambda \mathcal{E}_{f,\mathcal{P}}(d\lambda) \leq \sup\sigma(\mathcal{P}) \int_{\sigma(\mathcal{P})} \mathcal{E}_{f,\mathcal{P}}(d\lambda) = \sup\sigma(\mathcal{P}) \cdot \|f\|^2\,,$$

and similarly $\langle f, \mathcal{Q}f\rangle \geq \inf\sigma(\mathcal{Q}) \cdot \|f\|^2$. Thus for every $f \in L^2_0(\pi)$,

$$\langle f, (\mathcal{Q} - \mathcal{P})\, f\rangle \;\geq\; (\inf\sigma(\mathcal{Q}) - \sup\sigma(\mathcal{P}))\, \|f\|^2 \;\geq\; 0,$$

and thus by Theorem 4.1 P efficiency dominates Q. □

Example 4.9. *In [24], the authors prove positivity (i.e. that $\sigma(\mathcal{P}) \subseteq [0,\infty)$) for many general state space MCMC algorithms. In particular, they use the fact that conjugation of a positive operator by a bounded operator is again a positive operator, i.e. they show that the corresponding Markov operator $\mathcal{P}$ of each algorithm is of the form MTM^*, where $M : \mathbf{H} \to L^2_0(\pi)$ is a bounded linear operator from some Hilbert space $\mathbf{H}$, $M^* : L^2_0(\pi) \to \mathbf{H}$ is it's adjoint operator (see Section 2.2), and $T : \mathbf{H} \to \mathbf{H}$ is a self-adjoint positive operator on $\mathbf{H}$.*

For context, let $n \geq 1$, $K \subseteq \mathbf{R}^n$ with nonempty interior, and $f : K \to [0,\infty)$ be a possibly unnormalised probability density. The authors in [24] prove positivity for the following algorithms.

1. *Hit-and-Run Algorithm.*

   - *Starting at $x_k \in K$, choose a direction θ in the $n-1$ dimensional unit sphere uniformly at random, and then sample $x_{k+1} \in K$ from f restricted to the one dimensional subset $\{x + \delta\theta : \delta \in \mathbf{R}$ such that $x + \delta\theta \in K\}$.*

2. *Random Scan Gibbs Sampler Algorithm.*

   - *Starting at $x_k \in K$, choose an axis $j \in \{1, \ldots, n\}$ unformly at random, and then sample $x_{k+1} \in K$ from f restricted to the one dimensional subset $\{x + \delta e_j : \delta \in \mathbf{R}$ such that $x + \delta e_j \in K\}$, where e_j is the jth coordinate vector of $\mathbf{R}^n$.*

3. *Slice Sampler Algorithm.*

   - *(Simple Slice Sampler) Starting at $x_k \in K$, choose a level $t \in (0, f(x_k)]$ uniformly at random, and then sample $x_{k+1} \in K$ uniformly at random from the set $\{x \in K : f(x) \geq t\} =: f^{-1}([t,\infty))$. (In [24], they prove positivity for more general slice sampler algorithms, allowing for more general transitions once a level t is chosen, not only uniform transitions.)*

4. Metropolis Algorithm.

- *Given any Markov kernel Q that is reversible with respect to Lebesgue measure and positive, given $x_k \in K$, generate a sample $y \in K$ from $Q(x_k, \cdot)$, and accept the proposal and set $x_{k+1} = y$ with probability $\alpha(x_k, y) = \min\{1, \frac{f(y)}{f(x)}\}$ and reject the proposal and set $x_{k+1} = x_k$ with probability $1 - \alpha(x_k, y)$.*

Recall from Section 2.1 that $\Pi \equiv 0$, where Π is the Markov kernel corresponding to i.i.d. sampling from π. Thus clearly $\sigma(\Pi) = \{0\}$, and as the above algorithms are positive, denoting their Markov operator as $\mathcal{P}$, $\sigma(\mathcal{P}) \subseteq [0, \infty)$. So, in particular $\inf \sigma(\mathcal{P}) \geq 0 = \sup \sigma(\Pi)$, and thus by Proposition 4.8, we find that i.i.d. sampling efficiency dominates each of the above algorithms.

As seen in [12], antithetic methods can lead to improved efficiency of MCMC methods. In this paper, we define *antithetic* Markov kernels as Markov kernels P such that $\sigma(\mathcal{P}) \subseteq [-1, 0]$ when restricted to $L_0^2(\pi)$. We will show here that antithetic reversible Markov kernels are more efficient than i.i.d. sampling from π directly.

Proposition 4.10. *Let P be a φ-irreducible Markov kernel reversible with respect to π. Then P is antithetic if and only if P efficiency dominates Π (the Markov kernel corresponding to i.i.d. sampling from π).*

Proof. Recall from Section 2.1 that $\Pi \equiv 0$ on $L_0^2(\pi)$, and thus $\sigma(\Pi) = \{0\}$.

So say P is antithetic. Then $\sup \sigma(\mathcal{P}) \leq 0 = \inf \sigma(\Pi)$, and by Proposition 4.8, P efficiency dominates Π.

On the other hand if P efficiency dominates Π, then by Proposition 4.8, $\sup \sigma(\mathcal{P}) \leq \inf \sigma(\Pi) = 0$, and thus $\sigma(\mathcal{P}) \subseteq (-\infty, 0] \cap [-1, 1] = [-1, 0]$, so P is antithetic. □

Example 4.11. *Take $(\mathbf{X}, \mathcal{F}, \pi) = ([0,1], \mathcal{B}, m)$ where $\mathcal{B}$ is the Borel σ-field on $[0, 1]$ and m is the Lebesgue measure on $[0, 1]$, i.e. $(\mathbf{X}, \mathcal{F}, \pi)$ is the uniform probability space.*

Denote the left and right half of $[0, 1]$ by L and R respectively, i.e. $L = [0, 1/2]$ and $R = (1/2, 1]$. Then let P be the Markov kernel that jumps from the left half L to the right half R uniformly and similarly from the right half R to the left half L uniformly, i.e. $P(x, dy) = 2\left(\mathbf{1}_L(x) m|_R(dy) + \mathbf{1}_R(x) m|_L(dy)\right)$.

P is clearly reversible (and φ-irreducible). Furthermore, notice that for every $f \in L_0^2(m)$, as $0 = \mathbf{E}_m(f) = \int f dm = \int_L f dm + \int_R f dm$, $-\int_L f dm = \int_R f dm$,

and so

$$\begin{aligned}\langle f, \mathcal{P}f\rangle &= 2\iint_{[0,1]} f(x)f(y)\left(\mathbf{1}_L(x)m|_R(dy) + \mathbf{1}_R(x)m|_L(dy)\right)m(dx)\\ &= 2\left(\int_{x\in L}\int_{y\in R} f(x)f(y)m(dy)m(dx) + \int_{x\in R}\int_{y\in L} f(x)f(y)m(dy)m(dx)\right)\\ &= 2\left(\left(\int_L fdm\right)\left(\int_R fdm\right) + \left(\int_R fdm\right)\left(\int_L fdm\right)\right)\\ &= 2\left(\left(\int_L fdm\right)\left(-\int_L fdm\right) + \left(-\int_L fdm\right)\left(\int_L fdm\right)\right)\\ &= -4\left(\int_L fdm\right)^2 \leq 0.\end{aligned}$$

Thus we see that $-\mathcal{P}$ is a positive operator, and thus by Lemma 4.6, $-\sigma(\mathcal{P}) = \sigma(-\mathcal{P}) \subseteq [0,\infty)$, and thus $\sigma(\mathcal{P}) \subseteq (-\infty, 0]$. As $\mathcal{P}$ is an operator arising from a Markov kernel, we have $\sigma(\mathcal{P}) \subseteq [-1, 0]$, so P is antithetic.

Thus by Proposition 4.10, P efficiency dominates i.i.d. sampling. So, for every square-integrable Borel function f on the interval $[0,1]$, we can achieve a lower variance in our Monte Carlo estimate using a Markov chain with P as it's kernel than by i.i.d. sampling.

The above example illustrates even in very simple scenarios, we can improve our estimates using MCMC algorithms over i.i.d. sampling.

We now show that efficiency dominance is partial ordering on the set of φ-irreducible reversible Markov kernels reversible with respect to π.

Theorem 4.12. *Efficiency dominance is a partial order on φ-irreducible reversible Markov kernels, reversible with respect to π (reversible with respect to the same probability measure).*

Proof. Reflexivity is trivial.

Suppose P and Q are φ-irreducible reversible Markov kernels reversible with respect to π such that P efficiency dominates Q and Q efficiency dominates P. Then by Theorem 4.1, for every $f \in L_0^2(\pi)$,

$$\langle f, \mathcal{P}f\rangle \leq \langle f, \mathcal{Q}f\rangle \quad \text{and} \quad \langle f, \mathcal{P}f\rangle \geq \langle f, \mathcal{Q}f\rangle,$$

so $\langle f, \mathcal{P}f\rangle = \langle f, \mathcal{Q}f\rangle$ for every $f \in L_0^2(\pi)$. Thus $\langle f, (\mathcal{Q}-\mathcal{P})f\rangle = 0$ for every $f \in L_0^2(\pi)$. So for every $g, h \in L_0^2(\pi)$, as $\mathcal{Q}$ and $\mathcal{P}$ are self-adjoint,

$$\begin{aligned}0 &= \langle g+h, (\mathcal{Q}-\mathcal{P})(g+h)\rangle\\ &= \langle g, (\mathcal{Q}-\mathcal{P})g\rangle + \langle h, (\mathcal{Q}-\mathcal{P})h\rangle + 2\langle g, (\mathcal{Q}-\mathcal{P})h\rangle\\ &= 2\langle g, (\mathcal{Q}-\mathcal{P})h\rangle.\end{aligned}$$

So for every $g, h \in L^2_0(\pi)$, $\langle g, (\mathcal{Q} - \mathcal{P})h\rangle = 0$. Thus $\mathcal{Q} - \mathcal{P} = 0$, so $\mathcal{P} = \mathcal{Q}$, and thus $P = Q$. So the relation is antisymmetric.

Suppose P, Q and R are φ-irreducible reversible Markov kernels reversible with respect to π, such that P efficiency dominates Q and Q efficiency dominates R. Then by Theorem 4.1, for every $f \in L^2_0(\pi)$, $\langle f, (\mathcal{R} - \mathcal{Q})f\rangle \geq 0$ and $\langle f, (\mathcal{Q} - \mathcal{P})f\rangle \geq 0$. So, for every $f \in L^2_0(\pi)$,

$$\langle f, (\mathcal{R} - \mathcal{P})f\rangle = \langle f, (\mathcal{R} - \mathcal{Q})f\rangle + \langle f, (\mathcal{Q} - \mathcal{P})f\rangle \geq 0.$$

And thus by Theorem 4.1, we have P efficiency dominates R, so the relation is transitive. □

5 Combining Chains

In this section, we generalise the results of Neal and Rosenthal in [19] on the efficiency dominance of combined chains from finite state spaces to general state spaces. We state the most general result first, a sufficient condition for the effiency dominance of combined kernels, and then a simple Corollary following from it.

Theorem 5.1. *Let $P_1, \ldots, P_l$ and $Q_1, \ldots, Q_l$ be Markov kernels reversible with respect to π. Let $\alpha_1, \ldots, \alpha_l$ be mixing probabilities (i.e. $\alpha_k \geq 0$ for every k, and $\sum \alpha_k = 1$) such that $P = \sum \alpha_k P_k$ and $Q = \sum \alpha_k Q_k$ are φ-irreducible.*

If $\mathcal{Q}_k - \mathcal{P}_k$ is a positive operator (i.e. $\sigma(\mathcal{Q}_k - \mathcal{P}_k) \subseteq [0, \infty)$) for every k, then P efficiency dominates Q.

Proof. By definition of a positive operator (Section 2.2), for every $k \in \{1, \ldots, l\}$, we have

$$\langle f, (\mathcal{Q}_k - \mathcal{P}_k)\, f\rangle \geq 0, \qquad \forall f \in L^2_0(\pi).$$

So, for every $f \in L^2_0(\pi)$,

$$\langle f, (\mathcal{Q} - \mathcal{P})\, f\rangle = \sum \alpha_k \langle f, (\mathcal{Q}_k - \mathcal{P}_k)\, f\rangle \geq 0.$$

Thus as P and Q are also φ-irreducible (and reversible as each P_k and Q_k are reversible), $\mathcal{P}$ and $\mathcal{Q}$ satisfy (4.1), and thus by Theorem 4.1, P effieciency dominates Q. □

Remark. *If the Markov kernels $P_1, \cdots, P_l$ and $Q_1, \ldots, Q_l$ are φ-irreducible, then Theorem 5.1 can be restated as follows. Let $P_1, \ldots, P_l$ and $Q_1, \ldots, Q_l$ be φ-irreducible Markov kernels reversible with respect to π and $\alpha_1, \ldots, \alpha_l$ be mixing probabilities. If P_k efficiency dominates Q_k for every k, then $P = \sum \alpha_k P_k$ efficiency dominates $Q = \sum \alpha_k Q_k$.*

The converse of Theorem 5.1 is not true, even in the case where $P_1, \dots, P_l$ and $Q_1, \dots, Q_l$ are φ-irreducible. For a simple counter example, take $l = 2$, and let P_1 and P_2 be any φ-irreducible Markov kernels, reversible with respect to a probability measure π such that P_1 efficiency dominates P_2, but P_2 does not efficiency dominate P_1. Then by taking $Q_1 = P_2$, $Q_2 = P_1$ and $\alpha_1 = \alpha_2 = 1/2$, we have $P = 1/2\,(P_1 + P_2)$ and $Q = 1/2\,(Q_1 + Q_2) = 1/2\,(P_1 + P_2)$, so $P = Q$. Thus as efficiency dominance is reflexive by Theorem 4.12, P efficiency dominates Q. However, by assumption P_2 does not efficiency dominates $P_1 = Q_2$, thus the components do not efficiency dominate each other.

What is true is the following.

Corollary 5.2. *Let P, Q and R be Markov kernels reversible with respect to π, such that P and Q are φ-irreducible. Then for every $\alpha \in (0,1)$, P efficiency dominates Q if and only if $\alpha P + (1-\alpha)R$ efficiency dominates $\alpha Q + (1-\alpha)R$.*

Proof. As P and Q are φ-irreducible, $\alpha P + (1-\alpha)R$ and $\alpha Q + (1-\alpha)R$ are also φ-irreducible (to see this use the same σ-finite measure and the fact that for every $x \in \mathbf{X}$ and $A \in \mathcal{F}$, $(\alpha P + (1-\alpha)R)^n (x, A) \geq \alpha^n P^n(x, A)$).

If P efficiency dominates Q, by Theorem 5.1, $\alpha P + (1-\alpha)R$ efficiency dominates $\alpha Q + (1-\alpha)R$.

If $\alpha P + (1-\alpha)R$ efficiency dominates $\alpha Q + (1-\alpha)R$, by Theorem 4.7,

$$\sigma(\mathcal{Q} - \mathcal{P}) = \sigma(\alpha^{-1}\left[\alpha\mathcal{Q} + (1-\alpha)\mathcal{R} - (\alpha\mathcal{P} + (1-\alpha)\mathcal{R})\right]) \subseteq [0, \infty),$$

so by Theorem 4.7 again, P efficiency dominates Q. □

Thus when swapping only one component, the new combined chain efficiency dominates the old combined chain if and only if the new component efficiency dominates the old component.

Example 5.3. *For $n \geq 1$, say we are interested in a possibly unnormalised probability density $f : \mathbf{R}^n \to [0, \infty)$ (here we assume that $\mathbf{R}^n$ is equipped with the usual Borel σ-field and Lebesgue measure). Then recall from Example 4.9, that the random scan Gibbs sampler algorithm works by randomly choosing an axis in $\mathbf{R}^n$, and updating only the jth coordinate using the marginal distribution of f on that axis given the last point, i.e. given $x_k = (x_k^{(1)}, \dots, x_k^{(n)}) \in \mathbf{R}^n$, choose an axis $j \in \{1, \dots, n\}$ uniformly at random, pick $x_{k+1}^{(j)}$ according to the one-dimensional probability distribution $f_j(\cdot|(x_k^{(1)}, \dots, x_k^{(j-1)}, x_k^{(j+1)}, \dots, x_k^{(n)})$, where f_j is the marginal distribution in the jth coordinate, and set $x_{k+1} = (x_k^{(1)}, \dots, x_k^{(j-1)}, x_{k+1}^{(j)}, x_k^{(j+1)}, \dots, x_k^{(n)})$.*

For every measurable set $A \subseteq \mathbf{R}^n$, $x = (x^{(1)}, \dots, x^{(n)}) \in \mathbf{R}^n$ and $j \in \{1, \dots, n\}$, let $A_j(x) := \{y^{(j)} \in \mathbf{R}^n : (x^{(1)}, \dots, x^{(j-1)}, y^{(j)}, x^{(j+1)}, \dots, x^{(n)}) \in A\}$. Then this *algorithm has Markov kernel*

$$Q(x,A) = \frac{1}{n}\sum_{j=1}^{n} Q_j(x,A) = \frac{1}{n}\sum_{j=1}^{n} \frac{\int_{A_j(x)} f(x^{(1)},\dots,x^{(j-1)},t,x^{(j+1)},\dots,x^{(n)})dt}{\int_{\mathbf{R}} f(x^{(1)},\dots,x^{(j-1)},t,x^{(j+1)},\dots,x^{(n)})dt},$$

for every $x \in \mathbf{R}^n$ and measurable $A \subseteq \mathbf{R}^n$, where Q_j is the Markov kernel associated to updating the jth coordinate. Notice that for every $j \in \{1,\dots,n\}$, the Markov kernel Q_j corresponding to updating the jth coordinate, is simply i.i.d. sampling from the one-dimensional probability distribution f_j given the other $n-1$ coordinates (as described in the above paragraph).

Thus in order to find a more efficient algorithm than Gibbs sampling, by Theorem 5.1, it suffices to simply find an algorithm that efficiency dominates i.i.d. sampling in one dimension, and then the sum of this new algorithm applied to each coordinates marginal distribution, as described above, will efficiency dominate Gibbs sampling. Some work in this direction can be found in [18] for discrete state spaces.

6 Peskun Dominance

In this section, we show that Theorem 4.7, once established, simplifies the proof that Peskun dominance implies efficiency dominance. Peskun dominance is another widely used condition, introduced by Peskun in [20] for finite state spaces, and generalized to general state spaces by Tierney in [26]. We shall see that it is a stronger condition than efficiency dominance. We follow the techniques of Tierney (see [26]) to establish a key lemma, then show that with Theorem 4.7 already established, this lemma immediately gives us our result. For a different proof of the fact that Peskun dominance implies efficiency dominance in the finite state space case, see [19].

We start with our key lemma.

Lemma 6.1. *If P and Q are Markov kernels reversible with respect to π, such that P Peskun dominates Q, then $\mathcal{Q} - \mathcal{P}$ is a positive operator.*

Proof. For every $x \in \mathbf{X}$, let $\delta_x : \mathcal{F} \to \{0,1\}$ be the measure such that
$\delta_x(E) = \begin{cases} 1, & x \in E \\ 0, & o.w. \end{cases}$ for every $E \in \mathcal{F}$.

Then notice that as P and Q are reversible with respect to π,

$$\pi(dx)(\delta_x(dy) + P(x,dy) - Q(x,dy)) = \pi(dy)(\delta_y(dx) + P(y,dx) - Q(y,dx)).$$

Thus for every $f \in L_0^2(\pi)$, we have

$$\begin{aligned}\langle f, (\mathcal{Q} - \mathcal{P})f\rangle &= \iint_{x,y\in\mathbf{X}} f(x)f(y)(Q(x,dy) - P(x,dy))\pi(dx) \\ &= \int_{x\in\mathbf{X}} f(x)^2\pi(dx) \\ &\qquad - \iint_{x,y\in\mathbf{X}} f(x)f(y)(\delta_x(dy) + P(x,dy) - Q(x,dy))\pi(dx) \\ &= \frac{1}{2}\iint_{x,y\in\mathbf{X}} (f(x) - f(y))^2 \left(\delta_x(dy) + P(x,dy) - Q(x,dy)\right)\pi(dx).\end{aligned}$$

As P Peskun dominates Q, $(\delta_x(\cdot) + P(x,\cdot) - Q(x,\cdot))$ is a positive measure for π-almost every $x \in \mathbf{X}$. Thus

$$\frac{1}{2}\iint_{x,y\in\mathbf{X}} (f(x) - f(y))^2 \left(\delta_x(dy) + P(x,dy) - Q(x,dy)\right)\pi(dx) \geq 0.$$

As $f \in L_0^2(\pi)$ is arbitrary, $\mathcal{Q} - \mathcal{P}$ is a positive operator. □

Now with Theorem 4.7 we can easily show the following.

Theorem 6.2. *If P and Q are φ-irreducible Markov kernels reversible with respect to π, such that P Peskun dominates Q, then P efficiency dominates Q.*

Proof. By Lemma 6.1, $\mathcal{Q} - \mathcal{P}$ is a positive operator (i.e. $\sigma(\mathcal{Q} - \mathcal{P}) \subseteq [0,\infty)$), and thus by Theorem 4.7, P efficiency dominates Q. □

In [26], Theorem 6.2 is used to show that given a set of proposal kernels, the Metropolis-Hastings algorithm that samples from a weighted sum of the proposal kernels and then performs the accept/reject step efficiency dominates the weighted sum of Metropolis-Hastings algorithms with said proposal kernels. We consider this example next.

Example 6.3. *Say we are interested in sampling from a possibly unnormalised probability density $f : \mathbf{X} \to [0,\infty)$ on $(\mathbf{X}, \mathcal{X}, \mu)$. One of the most common algorithms we can use is the Metropolis-Hastings (MH) algorithm.*

Let $Q(x,dy) = q(x,y)\mu(dy)$ be any other proposal Markov kernel that is absolutely continuous with respect to μ. Then if $\{X_k\}_{k=0}^\infty$ is the Markov chain run by the Metropolis-Hastings algorithm, then given $X_k = x \in \mathbf{X}$, we first use $Q(x,\cdot)$ to generate a sample $y \in \mathbf{X}$, and with probability $\alpha(x,y) = \min\left\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right\}$ we set $X_{k+1} = y$, otherwise the chain stays at x, i.e. $X_{k+1} = x$. This chain has Markov kernel $P(x,dy) = \alpha(x,y)Q(x,dy) + r(x)\delta_x(dy)$, where $r(x) := 1 -$

$\int_{\mathbf{X}} \alpha(x,y)Q(x,dy)$ and δ_x is the point mass measure at $x \in \mathbf{X}$, i.e. $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ if $x \notin A$ for every $A \in \mathcal{X}$. Thus for every $x \in \mathbf{X}$ and every subset $A \in \mathcal{X}$ such that $x \notin A$, $P(x,A) = \int_{y \in A} \alpha(x,y)Q(x,dy) = \int_{y \in A} \alpha(x,y)q(x,y)\mu(dy)$.

When deciding on a proposal kernel Q to use, we oftentimes consider a sum of simpler kernels, i.e. given $\{Q_n\}_{n=0}^{N-1}$ and $\{\beta_n\}_{n=0}^{N-1}$ such that $\sum_{n=0}^{N-1} \beta_n = 1$, we might take $Q = \sum_{n=0}^{N-1} \beta_n Q_n$. Now we have two natural options. We can either use the MH algorithm of the kernel $Q = \sum \beta_n Q_n$, where we denote the kernel of this algorithm as P, or we can use the sum of MH algorithms $\sum_{n=0}^{N-1} \beta_n P_n$, where each P_n is the kernel arising from the MH algorithm with proposal Q_n. We will show, as is shown in [26], that the MH algorithm run from the combined proposal Q, P, efficiency dominates the sum of MH algorithms.

For every $x \in \mathbf{X}$ and $A \in \mathcal{X}$,

$$\begin{aligned}
P(x, A \setminus \{x\}) &= \int_{A\setminus\{x\}} \alpha(x,y)q(x,y)\mu(dy) \\
&= \int_{A\setminus\{x\}} \min\left\{q(x,y), \left(\frac{f(y)}{f(x)}\right) q(y,x)\right\} \mu(dy) \\
&\geq \int_{A\setminus\{x\}} \sum_{n=0}^{N-1} \beta_n \min\left\{q_n(x,y), \left(\frac{f(y)}{f(x)}\right) q_n(y,x)\right\} \mu(dy) \\
&= \sum_{n=0}^{N-1} \beta_n \int_{A\setminus\{x\}} \alpha_n(x,y)q_n(x,y)\mu(dy) \;=\; \sum_{n=0}^{N-1} \beta_n P_n(x, A \setminus \{x\}),
\end{aligned}$$

where $\alpha(x,y) = \min\left\{1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right\}$ and $\alpha_n(x,y) = \min\left\{1, \frac{f(y)q_n(y,x)}{f(x)q_n(x,y)}\right\}$ are the acceptance probabilities of their respective MH algorithm.

As $A \in \mathcal{X}$ was arbitrary, P Peskun dominates $\sum_{n=0}^{N-1} \beta_n P_n$, and thus by Theorem 6.2, P efficiency dominates $\sum_{n=0}^{N-1} \beta_n P_n$.

The converse of Theorem 6.2 is not true. In fact, we have already seen a simple example of a kernel that efficiency dominates but doesn't Peskun dominate another kernel.

Example 6.4 (Example 4.11, Continued)**.** *We have already seen that the Markov kernel P of Example 4.11 efficiency dominates i.i.d. sampling on the interval* $[0,1]$*. It is also easy to see that P does not Peskun dominate i.i.d. sampling.*

Take for example $x = 0 \in [0,1]$ and $A = [0,1/2] = L \in \mathcal{B}$. Then as P jumps from the left half of the interval L to the right half of the interval R uniformly, as 0 is in the left half, clearly $P(0, L \setminus \{0\}) \leq P(0,L) = 0$. Conversely, as single points have zero mass in Lebesgue measure, clearly $M(0, L \setminus \{0\}) = m(L \setminus \{0\}) = m(L) = 1/2$ (where M is the operator associated to i.i.d. sampling on $([0,1], \mathcal{B}, m)$*).*

Even in finite state spaces Peskun dominance is not a necessary condition for efficiency dominance. For a simple example in the finite state space case, see Section 7 of [19]. Although Peskun dominance can be an easier condition to check, as is clear here, efficiency dominance is a much more general condition.

7 Functional Analysis Lemmas

We seperate this section into two subsections. In the first subsection, we follow a parrallel approach to that of Neal and Rosenthal in [19] in the finite case, substituting linear algebra for functional analysis where appropriate, to prove Lemma 4.3. In the second subsection, we follow the techniques of Mira and Geyer in [17] to prove Lemma 4.4.

7.1 Proof of Lemma 4.3

As shown in [17], Lemma 4.3 follows from some more general results in [4]. However, these general results are very technical, and require much more than basic functional analysis to prove. So we present a different approach using basic functional analysis. These techniques are similar to what has been done in Chapter V of [5], as presented by Neal and Rosenthal in [19], but generalized for general Hilbert spaces rather than finite dimensional vector spaces.

We begin with some lemmas about bounded self-adjoint linear operators on a Hilbert space $\mathbf{H}$.

Lemma 7.1. *If X, Y, and Z are bounded linear operators on a Hilbert space $\mathbf{H}$ such that $\langle f, Xf \rangle \leq \langle f, Yf \rangle$ for every $f \in \mathbf{H}$, and Z is self-adjoint, then $\langle f, ZXZf \rangle \leq \langle f, ZYZf \rangle$ for every $f \in \mathbf{H}$.*

Proof. For every $f \in \mathbf{H}$, $Zf \in \mathbf{H}$, so

$$\langle f, ZXZf \rangle = \langle Zf, XZf \rangle \leq \langle Zf, YZf \rangle = \langle f, ZYZf \rangle.$$

□

This is where the finite state space case differs from the general case. In the finite state space case, $L_0^2(\pi)$ is a finite dimensional vector space, and thus in order to prove that $\langle f, Tf \rangle \leq \langle f, Nf \rangle$ for every $f \in \mathbf{V}$ if and only if $\langle f, T^{-1}f \rangle \geq \langle f, N^{-1}f \rangle$ for every $f \in \mathbf{V}$, when T and N are self-adjoint operators, the only additional assumption needed is that T and N are *strictly positive*, i.e. that for every $f \neq 0 \in \mathbf{V}$, $\langle f, Tf \rangle, \langle f, Nf \rangle > 0$. This is presented by Neal and Rosenthal in [19], Section 8. However, in the general case, as $L_0^2(\pi)$ may not be finite dimensional, T and N being strictly positive is not a strong enough assumption. In the general case, it is possible for T to be strictly positive and self-adjoint, but not be invertible in the bounded sense. Thus it is possible that $0 \in \sigma(T)$. So, we must use a slightly stronger assumption. We must assume that $\sigma(T), \sigma(N) \subseteq (0, \infty)$. In the finite case, this is equivalent to being strictly positive, however it is stronger in general.

The following lemma is Theorem 12.12 from [23]. We present a more detailed proof below.

Lemma 7.2. *If T is a normal bounded linear operator on a Hilbert space $\mathbf{H}$, then there exists $\delta > 0$ such that $\delta \|f\| \leq \|Tf\|$ for every $f \in \mathbf{H}$ if and only if T is invertible.*

Proof. For the forward implication, we will show that as T is normal, by the assumption it will follow that T is bijective, and then by the assumption once more the inverse of T is bounded.

Firstly, notice that for every $f \in \mathbf{H}$ such that $f \neq 0$, $\|Tf\| \geq \delta \|f\| > 0$, so $Tf \neq 0$. As $Tf \neq 0$ for every $f \neq 0 \in \mathbf{H}$, T is injective.

As T is normal and injective, T^* is also injective, and as T is normal $\text{range}(T)^\perp = \text{null}(T^*) = \{0\}$, so the range of T is dense in $\mathbf{H}$.

Now we will show that the range of T is closed, and thus T is surjective as the range of T is also dense in $\mathbf{H}$. For any $f \in \overline{\text{range}(T)}$, there exists $\{g_n\}_{n \in \mathbf{N}} \subseteq \mathbf{H}$ such that $Tg_n \to f$. So for every $m, n \in \mathbf{N}$, by our assumption

$$\|g_n - g_m\| = \delta^{-1}\delta \|g_n - g_m\| \leq \delta^{-1} \|Tg_n - Tg_m\|,$$

so $\{g_n\} \subseteq \mathbf{H}$ is Cauchy as $\{Tg_n\}$ converges. Thus as $\mathbf{H}$ is complete (as it is a Hilbert space), there exists $g \in \mathbf{H}$ such that $g_n \to g$. As T is bounded, it is also continuous, and thus $Tg_n \to Tg$, and as the limits are unique and $Tg_n \to f$ as well, $Tg = f$ and $f \in \text{range}(T)$. So $\text{range}(T)$ is closed.

So as T is bijective, there exists an operator T^{-1} such that $TT^{-1}f = f$ for every $f \in \mathbf{H}$. By our assumption, letting $C = \delta^{-1}$, we have

$$C \|f\| = C \left\|TT^{-1}f\right\| \geq \left\|T^{-1}f\right\|$$

for every $f \in \mathbf{H}$, and so T^{-1} is bounded.

For the converse, say T is invertible. Then let $\delta = \|T^{-1}\|^{-1}$. Then for every $f \in \mathbf{H}$, by definition of δ,

$$\delta \|f\| = \delta \left\|T^{-1}Tf\right\| \leq \delta \left\|T^{-1}\right\| \|Tf\| = \|Tf\|.$$

□

Remark. *The assumption that T be normal in the preceding lemma is only to show us that T is bijective in the "only if" direction. In general, if T is a bounded linear operator, not necessarily normal, if T is bijective and there exists $\delta > 0$ such that $\|Tf\| \geq \delta \|f\|$ for every $f \in \mathbf{H}$, then T is invertible. Furthermore, the "if" direction of Lemma 7.2 does not require that T be normal.*

And now we can prove Lemma 4.3.

Proof of Lemma 4.3. Say $\langle f, Tf\rangle \leq \langle f, Nf\rangle$ for every $f \in \mathbf{H}$.

As $\sigma(N) \subseteq (0,\infty)$, N is invertible, and $N^{-1/2}$ is a well defined bounded self-adjoint linear operator. Similarly, $T^{1/2}$ is also a well defined bounded self-adjoint linear operator.

So, for every $f \in \mathbf{H}$, we have

$$\langle f, N^{-1/2}TN^{-1/2}f\rangle = \langle T^{1/2}N^{-1/2}f, T^{1/2}N^{-1/2}f\rangle = \left\|T^{1/2}N^{-1/2}f\right\|^2 \geq 0.$$

Furthermore, as $\sigma(T) \subseteq (0,\infty)$, T is invertible, so by Lemma 7.2, there exists $\delta_T > 0$ such that $\|Tf\| \geq \delta_T \|f\|$ for every $f \in \mathbf{H}$. Also, notice that $\sigma(N^{-1/2}) \subseteq (0,\infty)$, thus by Lemma 7.2, there exists $\delta_1 > 0$ such that $\left\|N^{-1/2}f\right\| \geq \delta_1 \|f\|$ for every $f \in \mathbf{H}$. So, for every $f \in \mathbf{H}$,

$$\left\|N^{-1/2}TN^{-1/2}f\right\| \geq \delta_1 \left\|TN^{-1/2}f\right\| \geq \delta_1\delta_T \left\|N^{-1/2}f\right\| \geq \delta_1\delta_T\delta_1 \|f\|,$$

so by Lemma 7.2, $N^{-1/2}TN^{-1/2}$ is invertible, and thus $0 \notin \sigma(N^{-1/2}TN^{-1/2})$.

By using Lemma 7.1 with $X = T$, $Y = N$ and $Z = N^{-1/2}$, for every $f \in \mathbf{H}$,

$$\langle f, N^{-1/2}TN^{-1/2}f\rangle \leq \langle f, N^{-1/2}NN^{-1/2}f\rangle = \|f\|^2.$$

So if $\lambda > 1$, for any $f \in \mathbf{H}$, as $0 \leq \langle N^{-1/2}TN^{-1/2}f, f\rangle \leq \|f\|^2$, by the Cauchy-Schwartz inequality, $\left\|(N^{-1/2}TN^{-1/2} - \lambda)f\right\| \geq |1-\lambda| \|f\|$, and as $|1-\lambda| > 0$, by Lemma 7.2, $(N^{-1/2}TN^{-1/2} - \lambda)$ is invertible, so $\lambda \notin \sigma(N^{-1/2}TN^{-1/2})$.

Thus we have $\sigma(N^{-1/2}TN^{-1/2}) \subseteq (0,1]$.

Let K denote the inverse of $N^{-1/2}TN^{-1/2}$, i.e. let $K = (N^{-1/2}TN^{-1/2})^{-1}$. Furthermore, we have $\sigma(K) \subseteq [1,\infty)$. So for every $f \in \mathbf{H}$, $\|f\|^2 \leq \langle f, Kf\rangle$.

So by using Lemma 7.1, with $X = \mathcal{I}$, $Y = K$ and $Z = N^{-1/2}$, for every $f \in \mathbf{H}$,

$$\begin{aligned}
\langle f, N^{-1}f\rangle &= \langle f, N^{-1/2}\mathcal{I}N^{-1/2}f\rangle \\
&\leq \langle f, N^{-1/2}KN^{-1/2}f\rangle \\
&= \langle f, N^{-1/2}(N^{-1/2}TN^{-1/2})^{-1}N^{-1/2}f\rangle \\
&= \langle f, N^{-1/2}N^{1/2}T^{-1}N^{1/2}N^{-1/2}f\rangle \\
&= \langle f, T^{-1}f\rangle.
\end{aligned}$$

For the other direction, replace N with T^{-1} and T with N^{-1}. □

7.2 Proof of Lemma 4.4

Here we follow the same steps of [17] to prove Lemma 4.4.

Lemma 7.3. *If T is a self-adjoint, injective, positive and bounded operator on the Hilbert space* $\mathbf{H}$*, then* $\operatorname{domain}(T^{-1}) \subseteq \operatorname{domain}(T^{-1/2})$*.*

Proof. Let $f \in \operatorname{domain}(T^{-1}) = \operatorname{range}(T)$. Then there exists $g \in \mathbf{H}$ such that $Tg = f$. So, as T is positive, $T^{1/2}$ is well-defined, so $T^{1/2}g = h \in \mathbf{H}$. Thus notice $T^{1/2}h = T^{1/2}T^{1/2}g = Tg = f$, so $f \in \operatorname{range}(T^{1/2}) = \operatorname{domain}(T^{-1/2})$. □

The next lemma is a generalization of Lemma 3.1 of [17] from real Hilbert spaces to possibly complex ones. This generalization is simple but unnecessary for us, as we are dealing with real Hilbert spaces anyways.

Lemma 7.4. *If T is a self-adjoint, injective, positive and bounded linear operator on the Hilbert space* $\mathbf{H}$*, then for every* $f \in \mathbf{H}$*,*

$$\langle f, Tf \rangle = \sup_{g \in \operatorname{domain}(T^{-1/2})} \left[\langle g, f \rangle + \langle f, g \rangle - \langle T^{-1.2}g, T^{-1/2}g \rangle \right].$$

Proof. As T is injective and self-adjoint, the inverse of T, $T^{-1} : \operatorname{range}(T) \to \mathbf{H}$, is densely defined and self-adjoint (see Proposition X.2.4 (b) of [6]).

For every $f \in \operatorname{range}(T) = \operatorname{domain}(T^{-1})$, there exists $g \in \mathbf{H}$ such that $Tg = f$. Thus as T is positive and self-adjoint,

$$\langle f, T^{-1}f \rangle = \langle Tg, g \rangle = \langle g, Tg \rangle \geq 0,$$

so T^{-1} is also positive. In particular, this means that $T^{1/2}$ and $T^{-1/2}$ are well-defined.

By Lemma 7.3, $\operatorname{domain}(T^{-1}) \subseteq \operatorname{domain}(T^{-1/2})$. So, let $f \in \mathbf{H}$. Let $h = Tf$. Then for every $g \in \operatorname{domain}(T^{-1/2}) = \operatorname{range}(T^{1/2})$,

$$\begin{aligned}
&\langle f, Tf \rangle - \big(\langle g, f \rangle + \langle f, g \rangle - \langle T^{-1/2}g, T^{-1/2}g \rangle \big) \\
&\quad = \langle T^{1/2}f, T^{1/2}f \rangle - \langle g, T^{-1}h \rangle - \langle T^{-1}h, g \rangle + \langle T^{-1/2}g, T^{-1/2}g \rangle \\
&\quad = \langle T^{-1/2}h, T^{-1/2}h \rangle - \langle T^{-1/2}g, T^{-1/2}h \rangle - \langle T^{-1/2}h, T^{-1/2}g \rangle + \langle T^{-1/2}g, T^{-1/2}g \rangle \\
&\quad = \langle T^{-1/2}(h-g), T^{-1/2}(h-g) \rangle \\
&\quad = \left\| T^{-1/2}(h-g) \right\|^2 \\
&\quad \geq 0.
\end{aligned}$$

As $h \in \operatorname{domain}(T^{-1})$ and $\operatorname{domain}(T^{-1}) \subseteq \operatorname{domain}(T^{-1/2})$, $h \in \operatorname{domain}(T^{-1/2})$. So, as T is self-adjoint,

$$\langle h, f \rangle + \langle f, h \rangle - \langle T^{-1/2}h, T^{-1/2}h \rangle = \langle Tf, f \rangle + \langle f, Tf \rangle - \langle Tf, T^{-1}Tf \rangle = \langle f, Tf \rangle.$$

□

With Lemma 7.4 established, the proof of Lemma 4.4 is straightforward.

Proof of Lemma 4.4. Let $f \in \mathbf{H}$. Then by Lemma 7.4,

$$\begin{aligned}
\langle f, Tf \rangle &= \sup_{g \in \text{domain}(T^{-1/2})} \langle g, f \rangle + \langle f, g \rangle - \langle T^{-1/2} g, T^{-1/2} g \rangle \\
&\le \sup_{g \in \text{domain}(N^{-1/2})} \langle g, f \rangle + \langle f, g \rangle - \langle N^{-1/2} g, N^{-1/2} g \rangle \\
&= \langle f, Nf \rangle.
\end{aligned}$$

□

Acknowledgements. We thank Austin Brown and Heydar Radjavi for some very helpful discussions about the Spectral Theorem which allowed us to prove Lemma 4.5. We also thank the editor for their valuable comments and input, from which this paper benefitted substantially. This work was partially funded by NSERC of Canada.

References

[1] C. Andrieu and S. Livingstone (2021), Peskun-Tierny Ordering for Markovian Monte Carlo: Beyond the Reversible Scenario. Annals of Statistics 49(4), 1958-1981.

[2] C. Andrieu and M. Vihola (2016), Establishing Some Order Amongst Exact Approximations of MCMCs. Annals of Applied Probability 26(5), 2661-2696.

[3] J.R. Baxter and J.S. Rosenthal (1995), Rates of Convergence for Everywhere-Positive Markov Chains. Statistics and Probability Letters 22(4), 333-338.

[4] J. Bendat and S. Sherman (1955), Monotone and Convex Operator Functions. Transactions of the American Mathematical Society 79, 58-71.

[5] R. Bhatia (1997), *Matrix Analysis.* Graduate Texts in Mathematics 169. Springer, New York.

[6] J.B. Conway (1990), *A Course in Functional Analysis, 2nd Ed.* Graduate Texts in Mathematics 96. Springer-Verlag, New York.

[7] R. Douc, E. Moulines, P. Priouret and P. Soulier (2018), *Markov Chains.* Springer Series in Operations Research and Financial Engineering. Springer Nature, Switzerland.

[8] N. Dunford and J.T. Schwartz (1963), *Linear Operators, Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space.* Interscience Publishers. John Wiley & Sons, New York.

[9] G.B. Folland (1999), *Real Analysis: Modern Techniques and Their Applications, 2nd Ed.* Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. John Wiley & Sons, New York.

[10] M.A. Gallegos-Herrada, D. Ledvinka and J.S. Rosenthal (2024), Equivalences of Geometric Ergodicity of Markov Chains. Journal of Theoretical Probability 27, 1230-1256.

[11] C. Geyer (1992), Practical Markov Chain Monte Carlo. Statistical Science 7(4), 473-483.

[12] P.J. Green and X. Han (1992) Metropolis Methods, Gaussian Proposals and Antithetic Variables. In Stochastic Models, Statistical Methods and Algorithms in Image Analysis (eds P. Barone, A. Frigessi & M. Piccioni), 142-164. Lecture Notes in Statistics 74. Springer-Verlag, Berlin.

[13] G.L. Jones (2004), On the Markov Chain Central Limit Theorem. Probability Surveys 1, 299-320.

[14] K. Khare and J.P. Hobert (2011), A Spectral Analytic Comparison of Trace-Class Data Augmentation Algorithms and Their Sandwich Variants. The Annals of Statistics 39(5), 2585-2606.

[15] C. Kipnis and S.R.S. Varadhan (1986), Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions. Communications in Mathematical Physics 104(1), 1-19.

[16] S.P. Meyn and R.L. Tweedie (1993), *Markov Chains and Stochastic Stability.* Communications and Control Engineering. Springer-Verlag, London.

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

[18] R.M. Neal (2024), Modifying Gibbs Sampling to Avoid Self Transitions. Preprint: `https://arxiv.org/abs/2403.18054.`

[19] R.M. Neal and J.S. Rosenthal (2024), Efficiency of Reversible MCMC Methods: Elementary Derivations and Applications to Composite Methods. Journal of Applied Probability, to appear.

[20] P.H. Peskun (1973), Optimum Monte Carlo Sampling Using Markov Chains. Biometrika 60(3), 607-612.

[21] G.O. Roberts and J.S. Rosenthal (2004), General State Space Markov Chains and MCMC Algorithms. Probability Surveys 1, 20-71.

[22] G.O. Roberts and J.S. Rosenthal (2008), Variance Bounding Markov Chains. The Annals of Applied Probability 18(3), 1201-1214.

[23] W. Rudin (1991), *Functional Analysis, 2nd Ed.* International Series in Pure and Applied Mathematics. McGraw-Hill, New York.

[24] D. Rudolf and M. Ullrich (2013), Positivity of Hit-and-run and related algorithms. Electric Communications in Probability 18, 1-8.

[25] L. Tierney (1994), Markov Chains for Exploring Posterior Distributions. The Annals of Statistics 22(4), 1701-1728.

[26] L. Tierney (1998), A Note on Metropolis-Hastings Kernels for General State Spaces. The Annals of Applied Probability 8(1), 1-9.