Spectral Bounds for Certain Two-Factor Non-Reversible MCMC Algorithms

by

Jeffrey S. Rosenthal and Peter Rosenthal

Departments of Statistics and Mathematics, University of Toronto

(August, 2015; revised November 2015)

Abstract

We prove that the Markov operator corresponding to the two-variable, non-reversible Gibbs sampler has spectrum which is entirely real and non-negative, thus providing a first step towards the spectral analysis of MCMC algorithms in the non-reversible case. We also provide an extension to Metropolis-Hastings components, and connect the spectrum of an algorithm to the spectrum of its marginal chain.

1 Introduction

This paper is inspired by the earlier paper [23], which discusses the importance of real, non-negative spectra for MCMC algorithms, and proves this property for several different reversible cases. In this paper, we extend that result to some common *non*-reversible MCMC algorithms, as we shall explain.

Markov chain Monte Carlo (MCMC) algorithms, such as the Gibbs sampler [9, 8] and the Metropolis-Hastings algorithm [16, 10, 26], are an extremely active area of modern research, with applications to numerous areas (see e.g. [3] and the references therein). Much of the mathematical analysis of these algorithms centers around their *convergence rate*; i.e., how long they need to be run before they produce accurate samples from the designated target probability distribution (cf. [20]). Some of this analysis uses probabilistic techniques such as coupling and minorisation conditions (e.g. [21, 4]). However, much of the analysis involves considering the *spectrum* of the associated Markov operator (see Section 2.2). In such cases, the Markov operator is nearly always assumed to be *self-adjoint*, corresponding to the Markov chain being *reversible* (see e.g. [13, 24, 6, 5, 12]). The paradigm used is then roughly as follows:

1. Since the Markov operator is self-adjoint, its spectrum must be *real* (not complex), and can often be shown (or forced) to be non-negative, cf. [23].

2. The corresponding *spectral gap* can then be bounded away from zero using various techniques (Cheeger's inequality, quadratic forms, etc.).

3. These spectral gap bounds then imply bounds on the operator's norm, which in turn lead to bounds on the Markov chain's convergence rate.

However, if the Markov chain is *not* reversible, then much of this paradigm breaks down (though the spectral radius formula is still of some relevance to step 3 above; see Section 2.2 below), and the analysis becomes much more difficult (see e.g. [17]). Some authors have attempted to get around this difficulty by *replacing* the non-reversible Markov chain by its "reversibilisation" [7], or by some other chain which provably has the same convergence properties [19]. However, there has been very little success at directly investigating the spectral properties of non-reversible Markov chains themselves, despite the fact that many commonly used MCMC algorithms (such as the systematic-scan Gibbs sampler) are not reversible and thus not amenable to the above paradigm.

In this paper, we make a small start in this direction. We consider one of the simplest common classes of non-reversible MCMC algorithms; namely, those which are a product of two factors each of which is a reversible Markov chain. In particular, we consider the two-variable systematic-scan Gibbs sampler, and prove step 1 of the above paradigm; i.e., that a Markov operator corresponding to such a sampler must have spectrum which is real and non-negative (Theorem 1). This implies (Corollary 2) that the corresponding auto-covariances are also non-negative. We also consider a combination of a Metropolis-Hastings component and a Gibbs Sampler component, and prove that the corresponding spectrum must still be real in that case (Theorem 3). Finally, we consider the relationship between the spectra of certain (non-reversible) systematic scan chains, and their corresponding (reversible) marginal chains (Theorem 5). We hope that these results will lead to further efforts to extend the above spectral analysis paradigm to non-reversible Markov chains.

2 Background

We begin with some background needed for our results.

2.1 Markov Chain

A (time-homogeneous) Markov chain on a measurable space $(\mathcal{X}, \mathcal{F})$ has a Markov kernel $P : \mathcal{X} \times \mathcal{F} \to [0, 1]$, where $P(x, A)$ represents the probability that, if the chain begins in the state $x \in \mathcal{X}$, it will then "move" to a state in $A \in \mathcal{F}$ on the next iteration. Formally, for each fixed $x \in \mathcal{X}$, the mapping $A \mapsto P(x, A)$ is a probability measure on $(\mathcal{X}, \mathcal{F})$, and for each fixed $A \in \mathcal{F}$, the mapping $x \mapsto P(x, A)$ is a measurable function on $\mathcal{X}$. A sequence of $\mathcal{X}$-valued random variables $X_0, X_1, X_2, \ldots$ is a Markov chain following the transitions P if for any $n \geq 0$ and all $A \in \mathcal{F}$, $\mathbf{Prob}[X_{n+1} \in A \,|\, X_0, X_1, \ldots, X_n] = P(X_n, A)$.

In the case of MCMC algorithms, there is always a fixed probability measure π on $(\mathcal{X}, \mathcal{F})$ which is *stationary* for P, meaning that $(\pi P)(A) := \int_{x \in \mathcal{X}} \pi(dx)\, P(x, A) = \pi(A)$ for all $A \in \mathcal{F}$. Under mild conditions, if the Markov chain is run repeatedly, then it will *converge in distribution* to π. Indeed, this is the main motivation for MCMC algorithms, and indeed the *speed* of this convergence is of great importance (see e.g. [20]).

One condition which guarantees that π is stationary for P is that the Markov chain is *reversible* with respect to π; i.e., that $\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)$ for all $x, y \in \mathcal{X}$.

2.2 Markov Operator

Such a Markov kernel P can also be viewed as a linear operator (see e.g. [22] for basic facts about operators), which acts on functions $f : \mathcal{X} \to \mathbf{C}$ by

$$(Pf)(x) := \int_{y \in \mathcal{X}} f(y)\, P(x, dy)\,,$$

so that $(Pf)(x)$ is the conditional expected value of f when the Markov chain takes one step starting at x.

The stationary probability measure π gives rise to an inner product $\langle f, g \rangle = \int_{x \in \mathcal{X}} f(x)\, \overline{g(x)}\, \pi(dx)$ and norm $\|f\| = \sqrt{\langle f, f \rangle}$ on the Hilbert space

$$L^2(\pi) := \{ f : \mathcal{X} \to \mathbf{C};\ \int_{x \in \mathcal{X}} |f(x)|^2\, \pi(dx) < \infty \}\,.$$

Then P acts on $L^2(\pi)$, and indeed it is easily seen (e.g. [2]) that we always have $\|Pf\| \leq \|f\|$; i.e., $\|P\| \leq 1$; i.e., P is a (weak) contraction on $L^2(\pi)$. Similar comments also apply to P acting on the subspace

$$L_0^2(\pi) \;:=\; \{f : \mathcal{X} \to \mathbf{C};\; f \in L^2(\pi),\; \int_{x \in \mathcal{X}} f(x)\, \pi(dx) = 0\}\,,$$

which is more directly related to MCMC convergence (since it avoids the specific eigenvalue 1 for constant functions, corresponding to the fact that $\pi P = \pi$ since π is a stationary distribution). The operator P is also related to the *auto-covariance* of the chain, which is important in understanding the accuracy of MCMC samplers (see e.g. [15]). Indeed, for $f : \mathcal{X} \to \mathbf{R}$,

$$\langle P^k f, f\rangle \;=\; \int_{x \in \mathcal{X}} P^k f(x)\, f(x)\, \pi(dx) \;=\; \int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} f(y)\, P^k(x, dy)\, f(x)\, \pi(dx)$$

$$=\; \mathbf{E}[f(X_k)\, f(X_0)] \;=\; \mathbf{Cov}[f(X_k), f(X_0)]\,,$$

where the expected value $\mathbf{E}$ is taken with respect to a Markov chain $\{X_n\}$ started in stationary and following the transitions P.

It is easily seen that P is reversible if and only if the operator P is *self-adjoint*; i.e., $\langle Pf, g\rangle = \langle f, Pg\rangle$ for all $f, g \in L^2(\pi)$. An operator P is *positive* if it is self-adjoint and also $\langle Pf, f\rangle \geq 0$ for all $f \in L^2(\pi)$. Any positive operator has a unique positive *square-root*; i.e., a positive operator $S := P^{1/2}$ with $S^2 = P$.

The *spectrum* of the operator P is defined, as usual, by

$$\sigma(P) \;:=\; \{\lambda \in \mathbf{C};\; (\lambda I - P) \text{ is not invertible}\}\,.$$

(Here I is the identity operator on $L^2(\pi)$, and "invertible" means having an inverse within the class of all bounded (i.e., continuous) linear operators on $L^2(\pi)$.) The corresponding *spectral radius* is $r(P) = \sup\{|z|; z \in \sigma(P)\}$. Since $\|P\| \leq 1$, it follows that $r(P) \leq 1$. In general, $\sigma(P)$ consists of complex numbers. However, for self-adjoint operators (corresponding to reversible Markov chains), the spectrum is well-known to contain only real numbers. And, for positive operators, the spectrum is also non-negative; i.e., contained in $[0, \infty)$.

It turns out (see e.g. [18]) that in the MCMC context, the spectral radius $r(P)$ for the operator P on $L_0^2(\pi)$ is of great importance to convergence rates. In the reversible case, this is because $r(P)^n$ then equals the operator norm $\|P^n\|$, and hence provides direct bounds on $\|P^n f\|$ for $f \in L_0^2(\pi)$. For example, if $f(x) = \mathbf{1}_A(x) - \pi(A)$, then $f \in L_0^2(\pi)$, and $\|f\| \leq 1$, and $(P^n f)(x) = P^n(x, A) - \pi(A)$, so $\int_{x \in \mathcal{X}} |P^n(x, A) - \pi(A)|^2\, \pi(dx) \leq \|P^n\| \leq r(P)^n$. In the non-reversible case, that bound does not hold; however by the spectral radius formula (e.g. [22], Theorem 10.13) we still have $r(P) = \lim_{n \to \infty} \|P^n\|^{1/n}$, so the bound still holds asymptotically in this sense.

2.3 Gibbs Sampler

Suppose now that $(\mathcal{X}, \mathcal{F}) = (\mathcal{X}_1, \mathcal{F}_1) \times (\mathcal{X}_2, \mathcal{F}_2) \times \ldots \times (\mathcal{X}_d, \mathcal{F}_d)$ is a d-fold product measurable space, and that λ_i is some σ-finite reference measure on $(\mathcal{X}_i, \mathcal{F}_i)$ for each i. (The most common case is where each λ_i equals Lebesgue measure on $\mathcal{X}_i = \mathbf{R}$.) Suppose further that the stationary probability distribution π has a density ϕ with respect to λ; i.e., $\pi \ll \lambda$ with $\frac{d\pi}{d\lambda} = \phi$. Then the i^{th} component *Gibbs sampler* is the Markov kernel G_i which leaves all coordinates besides i unchanged, and replaces the i^{th} coordinate by a draw from the full

conditional distribution of π conditional on all the other components. That is, for $x \in \mathcal{X}$ and $A_i \in \mathcal{F}_i$, if

$$S_{x,i,A_i} \;:=\; \{y \in \mathcal{X};\; y_j = x_j \text{ for } j \neq i, \text{ and } y_i \in A_i\}\,,$$

then

$$G_i(x,\, S_{x,i,A_i}) \;=\; \frac{\int_{t \in A_i} \phi(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)\, \lambda_i(dt)}{\int_{t \in \mathcal{X}_i} \phi(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)\, \lambda_i(dt)}\,.$$

These single-component Gibbs samplers G_i are easily seen to be reversible Markov chains corresponding to self-adjoint operators. In fact, they are projection operators, i.e. $(G_i)^2 = G_i$, so their spectra consist entirely of the values 0 and 1, and in particular their spectra are real and non-negative.

The single-component Gibbs samplers G_i are then combined together to form a complete MCMC algorithm P. There are two main ways of doing this. The first is the *systematic-scan Gibbs sampler*, defined by $P = G_1 G_2 \ldots G_d$, corresponding to cycling through all of the different coordinates in order. The second is the *random-scan Gibbs sampler*, defined by $\frac{1}{d}(G_1 + G_2 + \ldots + G_d)$, corresponding to choosing a coordinate uniformly at random and updating that coordinate only. Now, it is easily seen that the random-scan Gibbs sampler is reversible, so that its spectrum can be analysed in various ways (see e.g. [23]). However, the systematic-scan Gibbs sampler is more commonly used in applications, and it is definitely *not* reversible. (For example, if $d = 2$ and the support of π is an "L" shape, then with $G_1 G_2$ it is possible to move from the lower-right corner to the upper-left corner, but not to move the other way.)

In this paper, we focus on the two-variable systematic-scan Gibbs sampler; i.e., the case where $d = 2$ and $P = G_1 G_2$ (equivalent to the *data augmentation* algorithm introduced in [25]), which is arguably the simplest common non-reversible MCMC algorithm.

2.4 Metropolis-Hastings Algorithm

Let $d, \mathcal{X}_i, \mathcal{F}_i, \lambda_i, \phi$ be as above. When some of the Gibbs sampler kernels G_i cannot be feasibly implemented, practitioners sometimes instead use *Metropolis-Hastings* components, defined as follows. Let Q_i be an arbitrary Markov kernel on $\mathcal{X}$ which leaves all coordinates besides the i^{th} one unchanged; i.e., such that in the above notation $Q_i(S_{x,i,\mathcal{X}_i}) = 1$. Assume that $Q_i(x, \cdot)$ has a density $q_{i,x}(t)$ with respect to λ_i, in the sense that

$$Q_i(x,\, S_{x,i,A_i}) \;=\; \int_{t \in A_i} q_{i,x}(t)\, \lambda_i(dt)\,.$$

Then the i^{th} component Metropolis-Hastings algorithm is the Markov kernel M_i corresponding to "proposing" a new state $y \in \mathcal{X}$ according to Q_i, and then accepting this new state with probability $\alpha_i(x; y) := \min(1,\; \frac{\phi(y)\, q_{i,y}(x_i)}{\phi(x)\, q_{i,x}(y_i)})$, otherwise with probability $1 - \alpha_i(x, y)$ the new state is rejected so the Markov chain remains at the state x. In terms of Markov operators, writing $x[i,t] := (x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_d)$, this corresponds to setting

$$(M_i f)(x) \;=\; r(x)\, f(x) + \int_{t \in \mathcal{X}_i} q_{i,x}(t)\, \alpha_i(x, x[i,t])\, f(x[i,t])\, \lambda_i(dt)\,,$$

where $r(x) = 1 - \int_{t \in \mathcal{X}_i} q_{i,x}(t)\, \alpha_i(x, x[i,t])\, \lambda_i(dt)$ is the overall probability of rejecting the proposal.

Now, the acceptance probabilities $\alpha_i(x, y)$ have been chosen precisely (see e.g. [26, 20]) to ensure that each kernel M_i is reversible with respect to π, so π is stationary for M_i. Hence, the operator M_i is self-adjoint, though it might not be a positive operator.

Remark. It is also possible to define a full-dimensional Metropolis-Hastings algorithm, which acts on all components simultaneously. In the above notation, that corresponds to the case $d = 1$; i.e., to letting $\mathcal{X}_1$ be the entire state space and setting $P = M_1$. This approach is quite common, though we do not pursue it here.

3 Main Results

In terms of the above background, our first main result is as follows.

Theorem 1. *Consider a two-variable systematic-scan Gibbs sampler $P = G_1G_2$ as above (or any other product $P = G_1G_2$ for any positive Markov operators G_1 and G_2). Then the spectrum of P is real and non-negative, with $\sigma(P) \subseteq [0, 1]$.*

As discussed in the Introduction, this theorem extends step 1 of the reversible Markov chain paradigm to a non-reversible case.

Then, since $\langle P^k f, f\rangle = \mathbf{Cov}[f(X_k), f(X_0)]$ for real-valued f as noted above, it follows immediately that:

Corollary 2. *Let $\{X_n\}$ be a random sequence started in stationary and following the transitions $P = G_1G_2$ of a two-variable systematic-scan Gibbs sampler as above. Then for any real-valued $f \in L^2(\pi)$ and $k \in \mathbf{N}$, $\mathbf{Cov}[f(X_k), f(X_0)] \geq 0$.*

We also consider the case of a combination of a Gibbs sampler component and a Metropolis-Hastings component, as follows.

Theorem 3. *Consider a two-variable systematic-scan combination of a Metropolis-Hastings component and a Gibbs sampler component, of the form $P = M_1G_2$ or $P = G_1M_2$, with G_i and M_i as above (or any other positive Markov operator G_i and any other reversible Markov operator M_i). Then the spectrum of P is real, with $\sigma(P) \subseteq [-1, 1]$.*

4 Proofs of Main Results

Our proofs rely on the following known operator theory facts, following [11].

Proposition 4. (i) *Let A and B be two self-adjoint operators on a Hilbert space $\mathcal{H}$, with B positive. Then the spectra of the product operators AB and BA are equal and real; i.e., $\sigma(AB) = \sigma(BA) \subseteq \mathbf{R}$.*
(ii) *If, in addition to the above, A is also positive, then the spectra of the product operators are non-negative; i.e., $\sigma(AB) = \sigma(BA) \subseteq [0, \infty)$.*

Proof. By Proposition 1 of [11], $\sigma(AB) = \sigma(BA) = \sigma(SAS)$, where $S = B^{1/2}$ is the (unique) positive square root of the operator B (see Appendix for a discussion of the proof from [11]). But SAS is self-adjoint by inspection. Hence, $\sigma(AB) = \sigma(BA) = \sigma(SAS) \subseteq \mathbf{R}$, proving (i). Furthermore, if A is also positive, then $\langle SASf, f\rangle = \langle ASf, Sf\rangle \geq 0$ by the positivity of A, so that $\sigma(AB) = \sigma(SAS) \subseteq [0, \infty)$, proving (ii). ∎

Proof of Theorem 1. Applying Proposition 4(ii) with $A = G_1$ and $B = G_2$ shows that $\sigma(P) = \sigma(G_1 G_2) \subseteq [0, \infty)$. But we know that $r(P) \leq 1$, whence $\sigma(P) \subseteq [0, 1]$, as claimed. ∎

Remark. Theorem 1 does not extend directly to Gibbs samplers with $d > 2$ coordinates. Indeed, we have checked numerically that if $\mathcal{X} = \{1, 2\}^3$, with $\pi(i, j, k) \propto i + j + k$, then the corresponding three-variable systematic-scan Gibbs sampler has non-real eigenvalues $0.0002515 \pm 0.0014018\, i$, among others. Indeed, it is well-known (see [1]) that even Proposition 4 does not extend to three operators. Daniel Rosenthal has pointed out a simple example: if

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \text{ and } C = \begin{pmatrix} 2 & i \\ -i & 2 \end{pmatrix},$$

then A and B and C are each positive matrices, but the product ABC has complex eigenvalues $\frac{1}{2}\left(10 + i \pm \sqrt{75 + 20i}\right)$.

Proof of Theorem 3. Applying Proposition 4(i) with $A = M_1$ and $B = G_2$ shows that $\sigma(M_1 G_2) \subseteq \mathbf{R}$, or with $A = M_2$ and $B = G_1$ shows that $\sigma(G_1 M_2) \subseteq \mathbf{R}$, so either way we have $\sigma(P) \subseteq \mathbf{R}$. But we know that $r(P) \leq 1$, whence $\sigma(P) \subseteq [-1, 1]$, as claimed. ∎

5 The Marginal Chain

We now consider the connection between the spectrum of P, and the spectrum of the *marginal* chain $\widetilde{P}$, defined as follows.

For the two-variable systematic-scan Gibbs sampler $P = G_1 G_2$, the Markov chain proceeds by first (via G_1) "replacing" the first coordinate by a fresh value depending only on the second coordinate. This means that $P(x, A)$ does not depend on the first coordinate of x; i.e., $P((y, x_2), A) = P((z, x_2), A)$ for all $y, z \in \mathcal{X}_1$. Hence, also the function Pf depends only on x_2. That in turn implies the existence of a "marginal" Markov chain which only keeps track of the second coordinate; i.e., which has state space $(\mathcal{X}_2, \mathcal{F}_2)$, and transition kernel $\widetilde{P}$ defined by $\widetilde{P}(x_2, A_2) = P(x, \{(y_1, y_2) \in \mathcal{X}; y_2 \in A_2\})$ for $x_2 \in \mathcal{X}_2$ and $A_2 \in \mathcal{F}_2$. (Usually, a function of a Markov chain will not itself be a Markov chain, but rather a *hidden Markov model*.) In this case, it turns out [15, 18, 12] that $\widetilde{P}$ is reversible with respect to the marginal distribution of π on $\mathcal{X}_2$, defined by $\widetilde{\pi}(A_2) = \pi\{(x_1, x_2) \in \mathcal{X}; x_2 \in A_2\}$, and furthermore the convergence rate of $\widetilde{P}$ to $\widetilde{\pi}$ is identical to the convergence rate of P to π. So, that provides a different avenue to studying convergence of two-variable Gibbs samplers, using the methodology of reversible chains.

The above facts for the two-variable Gibbs sampler also extend ([14], Section 2.4) to the case $P = G_1 M_2$ of a combination of a Gibbs sampler component followed by a Metropolis-Hastings component; i.e., it also has a marginal chain $\widetilde{P}$ which is reversible with respect to $\widetilde{\pi}$ with the same convergence rate.

The identical convergence rates of the full and the marginal chain in these cases suggest that there might be a connection between their spectra. Indeed, we have the following.

Theorem 5. *Let $P = G_1 G_2$ or $P = G_1 M_2$ as above, and let $\widetilde{P}$ be the corresponding (reversible) marginal chain as above. Then $\sigma(P) = \sigma(\widetilde{P}) \cup \{0\}$; i.e., P and $\widetilde{P}$ have identical (real) spectra except perhaps for $\lambda = 0$.*

To prove Theorem 5, we require another operator theory result.

Proposition 6. *Let A be an operator on a Hilbert space $\mathcal{H}$. Suppose $\mathcal{M}$ is a proper closed linear subspace of $\mathcal{H}$ which contains the range of A; i.e., such that $Af \in \mathcal{M}$ whenever $f \in \mathcal{H}$. Let B be the restriction of A to $\mathcal{M}$; i.e., $B = A\big|_{\mathcal{M}}$. Then $\sigma(A) = \sigma(B) \cup \{0\}$.*

Proof. Let $\mathcal{M}^\perp = \{f \in \mathcal{H}; \langle f, g\rangle = 0 \ \forall g \in \mathcal{M}\}$ be the subspace of functions "perpendicular" to $\mathcal{M}$. Then the entire space $\mathcal{H}$ can be written as the direct sum $\mathcal{M} \oplus \mathcal{M}^\perp$. Hence any operator D can be decomposed in block-matrix form as

$$D \ = \ \left(\begin{array}{c|c} D_{11} & D_{12} \\ \hline D_{21} & D_{22} \end{array}\right)$$

meaning that $D(f_1 \oplus f_2) = (D_{11}f_1 + D_{12}f_2) \oplus (D_{21}f_1 + D_{22}f_2)$. With respect to this decomposition, we must have (since $\mathcal{M}$ contains the range of A) that

$$A \ = \ \left(\begin{array}{c|c} B & C \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)$$

for some operator $C : \mathcal{M}^\perp \to \mathcal{M}$. Then

$$\lambda I - A \ = \ \left(\begin{array}{c|c} \lambda I_{\mathcal{M}} - B & -C \\ \hline \mathbf{0} & \lambda I_{\mathcal{M}^\perp} \end{array}\right)$$

where $I_{\mathcal{M}}$ and $I_{\mathcal{M}^\perp}$ are the identity operators on $\mathcal{M}$ and $\mathcal{M}^\perp$ respectively.

Now, if $\lambda \neq 0$ and $\lambda \notin \sigma(B)$, then it can be checked directly that

$$(\lambda I - A)^{-1} \ = \ \left(\begin{array}{c|c} (\lambda I_{\mathcal{M}} - B)^{-1} & X \\ \hline \mathbf{0} & \lambda^{-1} I_{\mathcal{M}^\perp} \end{array}\right),$$

where $X = (\lambda I_{\mathcal{M}} - B)^{-1} C (\lambda^{-1} I_{\mathcal{M}^\perp})$. So, $\lambda I - A$ is invertible, and hence $\lambda \notin \sigma(A)$. This shows that $\sigma(A) \subseteq \sigma(B) \cup \{0\}$.

Also, since $\text{range}(A) \subseteq \mathcal{M}$, A is not surjective, and therefore $0 \in \sigma(A)$.

Finally, suppose $\lambda \notin \sigma(A)$. Then $(\lambda I - A)$ has an inverse, of the form

$$(\lambda I - A)^{-1} \ = \ \left(\begin{array}{c|c} W & X \\ \hline Y & Z \end{array}\right).$$

Then

$$I \ = \ (\lambda I - A)(\lambda I - A)^{-1} \ = \ \left(\begin{array}{c|c} (\lambda I_{\mathcal{M}} - B)W - CY & (\lambda I_{\mathcal{M}} - B)X - CZ \\ \hline \lambda Y & \lambda Z \end{array}\right).$$

It follows that $\lambda Y = \mathbf{0}$, so $Y = \mathbf{0}$ (since $\lambda \notin \sigma(A)$ so $\lambda \neq 0$). It then follows that $(\lambda I_{\mathcal{M}} - B)W - CY = I_{\mathcal{M}}$; i.e., that $(\lambda I_{\mathcal{M}} - B)W = I_{\mathcal{M}}$. Also,

$$I \ = \ (\lambda I - A)^{-1}(\lambda I - A) \ = \ \left(\begin{array}{c|c} W(\lambda I_{\mathcal{M}} - B) & WC - \lambda X \\ \hline \lambda Y - YB & YC - Z \end{array}\right),$$

from which it follows that $W(\lambda I_{\mathcal{M}} - B) = I_{\mathcal{M}}$. Combining these two facts, $(\lambda I_{\mathcal{M}} - B)W = W(\lambda I_{\mathcal{M}} - B) = I_{\mathcal{M}}$, so $(\lambda I_{\mathcal{M}} - B)$ is invertible (with inverse W). Hence, $\lambda \notin \sigma(B)$. This

shows that $\sigma(B) \subseteq \sigma(A)$. The result follows. ∎

Proof of Theorem 5. Let $\mathcal{J}$ be the set of all functions which depend only on the second coordinate; i.e., $\mathcal{J} = \{f \in L^2(\pi); f(x_1, x_2) = g(x_2)\ \forall x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, for some $g : \mathcal{X}_2 \to \mathbf{C}\}$. Then as discussed above, due to the nature of P we must have $Pf \in \mathcal{J}$ for all $f \in L^2(\pi)$. Hence, we can apply Proposition 6 with $A = P$ and $\mathcal{M} = \mathcal{J}$, to obtain that $\sigma(P) = \sigma(P\big|_{\mathcal{J}}) \cup \{0\}$.

But $P\big|_{\mathcal{J}}$ is essentially the same as $\widetilde{P}$: if $f \in \mathcal{J}$, with $f(x_1, x_2) = g(x_2)$ for all x_1 and x_2, then $(\widetilde{P}g)(x_2) = (Pf)(x_1, x_2)$. More formally, let $\widetilde{\mathcal{J}} = L^2(\widetilde{\pi})$ be the collection of square-integrable functions on $\mathcal{X}_2$, and x_* be any fixed element of $\mathcal{X}_1$, and define $S : \widetilde{\mathcal{J}} \to \mathcal{J}$ by $(Sf)(x_2) = f(x_*, x_2)$, with inverse $S^{-1} : \mathcal{J} \to \widetilde{\mathcal{J}}$ by $(S^{-1}g)(x_1, x_2) = g(x_2)$. Then $\widetilde{P} = S^{-1}P\big|_{\mathcal{J}}S$, so $\widetilde{P}$ is *similar* to $P\big|_{\mathcal{J}}$. In particular, $\sigma(\widetilde{P}) = \sigma(P\big|_{\mathcal{J}})$. The result follows. ∎

Remark. It is known that for the two-variable systematic-scan Gibbs sampler $P = G_1G_2$, the marginal chain is positive and thus has positive spectrum [15]; and for the combined chain $P = G_1M_2$, the marginal chain is reversible and thus has real spectrum [14]. Using this, Theorem 5 in turn provides an alternative proof of Theorems 1 and 3 – though it also strengthens them by providing a specific description (of sorts) of the spectra $\sigma(P)$ in those two cases.

6 A Self-Contained Operator Theory Proof

Our Proposition 4 above, which is essential to the proofs of Theorems 1 and 3, makes heavy use of Proposition 1 of [11]. The corresponding proof presented in [11] is brief, but it relies on several other operator theory concepts and theorems, and hence is not easily accessible to non-experts. For completeness, we provide here a self-contained proof, following [11].

Proposition 7. *([11]) Let A and B be two self-adjoint operators on a Hilbert space $\mathcal{H}$, with B positive. Let $S := B^{1/2}$ be the (unique) positive square root of B. Then $\sigma(AB) = \sigma(BA) = \sigma(SAS)$.*

We prove this Proposition using a few simple lemmas. The first was proved by Nathan Jacobson years ago; James Fulford has pointed out that there is a nice discussion of this topic at [27].

Lemma 8. *For any operators C and D on a Hilbert space $\mathcal{H}$, the spectra $\sigma(CD)$ and $\sigma(DC)$ differ by at most $\{0\}$; i.e., if $\lambda \in \mathbf{C}$ and $\lambda \neq 0$, then $\lambda \in \sigma(CD)$ if and only if $\lambda \in \sigma(DC)$.*

Proof. By replacing C by C/λ, it suffices to assume that $\lambda = 1$. Thus, it suffices to prove that $I - DC$ is invertible if and only if $I - CD$ is invertible. But this follows from the identity

$$(I - DC)^{-1} \;=\; I + D(I - CD)^{-1}C\,,$$

which can be verified by multiplying $I + D(I - CD)^{-1}C$ by $I - DC$ (on either the left or the right side) and getting the result I. ∎

Remark. The displayed identity in the proof of Lemma 8 is suggested intuitively (see e.g. [27]) by substituting in the (unjustified) expansions

$$(I - CD)^{-1} \;=\; \frac{1}{1 - CD} \;=\; 1 + CD + (CD)^2 + (CD)^3 + \dots$$

and

$$(I - DC)^{-1} \;=\; \frac{1}{1 - DC} \;=\; 1 + DC + (DC)^2 + (DC)^3 + \dots .$$

Lemma 9. *For any operators C and D on a Hilbert space $\mathcal{H}$, if D is self-adjoint, and CD is invertible, then C and D and DC are each invertible.*

Proof. Since CD is invertible, it must be injective; i.e., if $f \neq 0$ then $(CD)f \neq 0$. Hence also $Df \neq 0$. So, D is also injective.

Then, since CD is invertible, so is its adjoint $(CD)^*$. In particular, its adjoint must be surjective; i.e., for each $g \in \mathcal{H}$ there is $f \in \mathcal{H}$ with $(CD)^* f = g$. But $(CD)^* = D^* C^* = DC^*$ since D is self-adjoint. So, $D(C^* f) = g$. Hence, D is also surjective.

Thus, D is both injective and surjective, and hence invertible as a linear mapping $\mathcal{H} \to \mathcal{H}$. It then follows from the Open Mapping Theorem (see e.g. Corollary 2.12(b) on page 49 of [22]) that its inverse is a continuous (i.e., bounded) linear operator; i.e., D is invertible as a bounded linear operator on $\mathcal{H}$.

The remaining claims then follow from the fact that the product of invertible operators is invertible. ∎

Corollary 10. *([11]) For any operators C and D on a Hilbert space $\mathcal{H}$, if D is self-adjoint, then $\sigma(CD) = \sigma(DC)$.*

Proof. Lemma 8 above shows that $\sigma(CD)$ and $\sigma(DC)$ agree except possibly for the value 0, and Lemma 9 shows that $0 \in \sigma(CD)$ if and only if $0 \in \sigma(DC)$. ∎

Proof of Proposition 7. The first equality follows directly from Corollary 10. The second equality also follows from Corollary 10, by noting that $\sigma(AB) = \sigma(AS^2) = \sigma((AS)S) = \sigma(S(AS))$ since S is also self-adjoint. ∎

References

[1] C.S. Ballantine (1968), Products of positive definite matrices. II. Pacific J. Math. **24(1)**, 7–17.

[2] J.R. Baxter and J.S. Rosenthal (1995), Rates of convergence for everywhere-positive Markov chains. Stat. Prob. Lett. **22**, 333–338.

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

[4] R. Douc, E. Moulines, and J.S. Rosenthal (2004), Quantitative bounds on convergence of time-inhomogeneous Markov chains. Ann. Appl. Prob. **14**, 1643–1665.

[5] P. Diaconis and L. Saloff-Coste (1993), Comparison theorems for reversible Markov chains. Ann. Appl. Prob. **3**, 696–730.

[6] P. Diaconis and D.W. Stroock (1991), Geometric bounds for reversible Markov chains. Ann. Appl. Prob. **1**, 36–61.

[7] J.A. Fill (1991), Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. Ann. Appl. Prob. **1**, 64–87.

[8] A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.

[9] S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.

[10] W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.

[11] M. Hladnik and M. Omladic (1988), Spectrum of the product of operators. Proc. Am. Math. Soc. **102(2)**, 300–302.

[12] J.P. Hobert, V. Roy, and C.P. Robert (2011), Improving the Convergence Properties of the Data Augmentation Algorithm, with an Application to Bayesian Mixture Modeling. Statistical Science **26(3)**, 332-351.

[13] M. Jerrum and A. Sinclair (1989), Approximating the permanent. SIAM J. Comput. **18**, 1149-1178.

[14] G.L. Jones, G.O. Roberts, and J.S. Rosenthal (2014), Convergence of conditional Metropolis-Hastings samplers. Adv. Appl. Prob. **46(2)**, 422–445.

[15] J.S. Liu, W. Wong, and A. Kong (1994), Covariance structure of the Gibbs sampler, with applications to the comparisons of estimators and augmentation schemes. Biometrika **81**, 27–40.

[16] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

[17] A. Mira and C.J. Geyer (2000), On non-reversible Markov chains, Fields Institute Communications **26**: Monte Carlo Methods, pp. 95–110.

[18] G.O. Roberts and J.S. Rosenthal (1997), Geometric ergodicity and hybrid Markov chains. Elec. Comm. Prob. **2**, 13–25.

[19] G.O. Roberts and J.S. Rosenthal (2001), Markov chains and de-initialising processes. Scand. J. Stat. **28**, 489–504.

[20] G.O. Roberts and J.S. Rosenthal (2004), General state space Markov chains and MCMC algorithms. Prob. Surv. **1**, 20–71.

[21] J.S. Rosenthal (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

[22] W. Rudin (1991), Functional Analysis, 2$^{\text{nd}}$ ed. McGraw-Hill, New York.

[23] D. Rudolf and M. Ullrich (2012), Positivity of hit-and-run and related algorithms. Elec. Comm. Prob. **18**, 1–8.

[24] A. Sinclair (1992), Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Prob., Comput. **1**, 351–370.

[25] M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528-550.

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

[27] Q. Yuan (2012), ab, ba, and the spectrum. Blog post, available at:
https://qchu.wordpress.com/2012/06/05/ab-ba-and-the-spectrum/