ANALYSIS OF TWO-COMPONENT GIBBS SAMPLERS USING THE THEORY OF TWO PROJECTIONS

BY QIAN QIN[1]

[1]*School of Statistics, University of Minnesota, qqin@umn.edu*

The theory of two projections is utilized to study two-component Gibbs samplers. Through this theory, previously intractable problems regarding the asymptotic variances of two-component Gibbs samplers are reduced to elementary matrix algebra exercises. It is found that in terms of asymptotic variance, the two-component random-scan Gibbs sampler is never much worse, and could be considerably better than its deterministic-scan counterpart, provided that the selection probability is appropriately chosen. This is especially the case when there is a large discrepancy in computation cost between the two components. The result contrasts with the known fact that the deterministic-scan version has a faster convergence rate, which can also be derived from the method herein. On the other hand, a modified version of the deterministic-scan sampler that accounts for computation cost can outperform the random-scan version.

1. Introduction. Gibbs samplers are a class of Markov chain Monte Carlo (MCMC) algorithms commonly used in statistics for sampling from intractable distributions (Gelfand and Smith, 1990; Casella and George, 1992). In this work, I will introduce a method for analyzing different variants of two-component Gibbs samplers via the theory of two projections developed by Halmos (1969). As a first application of this new method, I will conduct a detailed comparison between deterministic-scan and random-scan samplers, which I now define.

Let $(X_1, \mathcal{B}_1)$ and $(X_2, \mathcal{B}_2)$ be measurable spaces and let $(X, \mathcal{B}) = (X_1 \times X_2, \mathcal{B}_1 \times \mathcal{B}_2)$. Let π be a probability measure on $(X, \mathcal{B})$ that is the joint distribution of the random element (X_1, X_2), where X_i is X_i-valued for $i = 1, 2$. For $x_1 \in X_1$ and $x_2 \in X_2$, let $\pi_2(\cdot \mid x_2)$ be the conditional distribution of $X_1 \mid X_2 = x_2$, and let $\pi_1(\cdot \mid x_1)$ be that of $X_2 \mid X_1 = x_1$. When one cannot sample from π directly, but can sample from $\pi_1(\cdot|x_1)$ and $\pi_2(\cdot|x_2)$, a two-component Gibbs sampler may be used to produce an approximate sample from π. Although simple, Gibbs algorithms with two components are surprisingly useful in practice (Tanner and Wong, 1987; Albert and Chib, 1993; Polson, Scott and Windle, 2013). There are two basic forms of two-component Gibbs samplers. Each simulates a Markov chain that has π as a stationary distribution. Moreover, under mild conditions, the marginal distribution of the tth element of each Markov chain converges to π in some sense as $t \to \infty$ (Tierney, 1994; Roberts and Rosenthal, 2006).

The first type, called deterministic-scan Gibbs (DG) sampler, simulates a time-inhomogeneous Markov chain $(\tilde{X}_t)$ in the following fashion. (T here is the length of the simulation.)

Algorithm 1: DG sampler

Draw $\tilde{X}_0 = (X_{1,0}, X_{2,0})$ from some initial distribution on $(\mathsf{X}, \mathcal{B})$, and set $t = 0$;
while $t < T$ **do**
    if $t = 2s$ *for some non-negative integer s* **then**
        draw $X_{2,t+1}$ from $\pi_1(\cdot \mid X_{1,t})$, set $X_{1,t+1} = X_{1,t}$, and let
        $\tilde{X}_{t+1} = (X_{1,t+1}, X_{2,t+1})$;
    if $t = 2s + 1$ *for some non-negative integer s* **then**
        draw $X_{1,t+1}$ from $\pi_2(\cdot \mid X_{2,t})$, set $X_{2,t+1} = X_{2,t}$, and let
        $\tilde{X}_{t+1} = (X_{1,t+1}, X_{2,t+1})$;
    set $t = t + 1$;

REMARK 1.1.   A time-homogeneous version of the DG chain can be obtained through thinning. Indeed, it is common to discard $\tilde{X}_t$ when t is odd, and only use $(\tilde{X}_{2s})_{s=0}^{\lfloor T/2 \rfloor}$ as a Monte Carlo sample. However, thinning reduces the efficiency of the sampler (MacEachern and Berliner, 1994), and is often discouraged (Link and Eaton, 2012).

The second type is called the random-scan Gibbs (RG) sampler. To run the algorithm, one needs to specify a selection probability $r \in (0, 1)$. The algorithm then simulates a time-homogeneous Markov chain $(\tilde{X}_t)$ in the following fashion.

Algorithm 2: RG sampler

Draw $\tilde{X}_0 = (X_{1,0}, X_{2,0})$ from some initial distribution on $(\mathsf{X}, \mathcal{B})$, and set $t = 0$;
while $t < T$ **do**
    draw W from a Bernoulli(r) distribution;
    if $W = 0$ **then**
        draw $X_{2,t+1}$ from $\pi_1(\cdot \mid X_{1,t})$, and set $X_{1,t+1} = X_{1,t}$;
    if $W = 1$ **then**
        draw $X_{1,t+1}$ from $\pi_2(\cdot \mid X_{2,t})$, and set $X_{2,t+1} = X_{2,t}$;
    set $\tilde{X}_{t+1} = (X_{1,t+1}, X_{2,t+1})$;
    set $t = t + 1$;

Both samplers generate new elements by updating X_1 using the conditional distribution of $X_1 \mid X_2$, and updating X_2 using the conditional distribution of $X_2 \mid X_1$. They only differ in terms of which component is updated in each iteration.

One important question regarding the DG and RG samplers is which of them performs better. This problem is not unique to the two-component case, but it is in this case where substantial progress has been made, as I now describe.

When comparing MCMC algorithms, there are two main aspects to consider: convergence speed and asymptotic variance (Jones and Hobert, 2001). Moreover, one must also account for computation cost, i.e., the time it takes to run one iteration of each algorithm. Qin and Jones (2022) showed that the DG algorithm is better than the RG algorithm in terms of L^2 convergence rate. Their result takes computation time into account, and holds for every

selection probability $r \in (0, 1)$. As to asymptotic variance, existing comparisons are less conclusive.

Consider a generic MCMC algorithm that simulates a Markov chain $(\tilde{X}_t)_{t=0}^{\infty}$ which converges to its stationary distribution π. Then $(\tilde{X}_0, \ldots, \tilde{X}_{T-1})$ forms an approximate Monte Carlo sample from π. The Monte Carlo sample is usually used to estimate the mean of a function $f : \mathsf{X} \to \mathbb{R}$, i.e., $\pi f := \int_{\mathsf{X}} f(x) \pi(\mathrm{d}x)$. The usual estimator is the sample mean

$$S_T(f) := \frac{1}{T} \sum_{t=0}^{T-1} f(\tilde{X}_t).$$

Under regularity conditions, $S_T(f)$ is subject to the central limit theorem (CLT):

$$\sqrt{T}[S_T(f) - \pi f] \xrightarrow{d} \mathrm{N}[0, V(f)] \quad \text{as } T \to \infty.$$

See Dobrushin (1956), Greenwood, McKeague and Wefelmeyer (1998), and Jones (2004) for Markov chain CLTs under various settings. $V(f)$ is called the asymptotic variance, and can usually be obtained from the formula

$$V(f) = \lim_{T \to \infty} T \operatorname{var}[S_T(f)] = \lim_{T \to \infty} T^{-1} \sum_{t=0}^{T-1} \sum_{t'=0}^{T-1} \operatorname{cov}[f(\tilde{X}_t), f(\tilde{X}_{t'})],$$

where $\tilde{X}_0$ can be assumed to follow π. Disregarding computation cost, one may say the smaller $V(f)$ the better. Define $V_{\mathrm{D}}(f)$ and $V_{\mathrm{R}}(f, r)$ to be the asymptotic variances associated with f for the DG and RG sampler, respectively, where r is the selection probability. Exact formulas for these quantities will be given in Section 2.1. Greenwood, McKeague and Wefelmeyer (1998) showed that $V_{\mathrm{D}}(f) \leq V_{\mathrm{R}}(f, 1/2)$. See also Andrieu (2016) where the result is extended to beyond Gibbs algorithms. As far as I am aware, no similar result existed for $r \neq 1/2$. Moreover, the arguments in Greenwood, McKeague and Wefelmeyer (1998) and Andrieu (2016) heavily rely on the symmetry of the RG sampler that arises only when $r = 1/2$, and it seems unlikely that they can be extended to the $r \neq 0.5$ case. This motivates the current work.

To appreciate the potential benefits of using a selection probability r that is not $1/2$, one needs to think about computation cost. Consider this: If the time it takes to draw from $\pi_2(\cdot | x_2)$ is much longer than $\pi_1(\cdot | x_1)$ for $x_1 \in \mathsf{X}_1$ and $x_2 \in \mathsf{X}_2$, then one has an incentive to update X_2 using $\pi_1(\cdot | x_1)$ more frequently, and may want to use an RG sampler with a small r rather than the DG sampler. To conduct a concrete analysis, I will study the adjusted asymptotic variance, defined below.

Suppose that on average, the aforementioned generic MCMC algorithm takes τ_0 units of time to produce a sample point, so that after running the algorithm for $\tilde{\tau}$ units of time where $\tilde{\tau}$ is large, around $\tilde{\tau}/\tau_0$ sample points are generated. Then the sample mean has variance roughly equal to $\tau_0 V(f)/\tilde{\tau}$. Define the computation time adjusted asymptotic variance to be $V^{\dagger}(f) := \tau_0 V(f)$. Given a fixed amount of computation effort, $V^{\dagger}(f)$, instead of $V(f)$, is what we should focus on. Define $V_{\mathrm{D}}^{\dagger}(f)$ and $V_{\mathrm{R}}^{\dagger}(f, r)$ to be the adjusted asymptotic variance for the DG and RG samplers, respectively. It will be shown that when r is well chosen, $V_{\mathrm{R}}^{\dagger}(f, r)$ can be much smaller than $V_{\mathrm{D}}^{\dagger}(f)$ for certain functions f, especially when the costs of drawing from π_1 and π_2 differ significantly. Moreover, for any function f such that $f(X)$ has finite second moment for $X \sim \pi$, under mild conditions, $V_{\mathrm{R}}^{\dagger}(f, r) \leq 2V_{\mathrm{D}}^{\dagger}(f)$, given that r is well chosen. The appropriate value of r depends on the time it takes to draw from the two conditional distributions, and an explicit formula will be provided. These results mean that the RG sampler can outperform the DG algorithm by a large margin in some scenarios, while never being too much worse. This is in contrast with Qin and Jones's (2022) result on

convergence rate. However, as discussed in Section 3.3, the DG sampler could still compete with RG in terms of adjusted asymptotic variance for certain problems.

The DG sampler can be made more robust through a simple modification. Suppose that it is much less costly draw from π_1 compared to π_2. Then one may consecutively draw from π_1 many times before drawing once from π_2. This modified sampler, defined in Section 3.4, behaves similarly to an RG sampler with some well-chosen r in terms of adjusted asymptotic variance. Moreover, it is possible to parallelize parts of the modified DG algorithm to make it even more efficient.

Aside from acquiring the more specific results described above, a central goal of this paper is to demonstrate how some tools in classical linear algebra can trivialize difficult problems concerning two-component Gibbs. The key is Halmos's (1969) theory of two projection operators. It is known that the two conditional distributions in a two-component Gibbs sampler correspond to two orthogonal projections on some function space (Greenwood, McKeague and Wefelmeyer, 1998; Diaconis, Khare and Saloff-Coste, 2010). Halmos (1969) gave block matrix representations for any given pair of orthogonal projections. Using these representations, one can obtain explicit formulas for the asymptotic variance of a given Gibbs sampler. The framework is powerful because it reduces the problem at hand to simple matrix calculations. I am unaware of previous works that analyze Gibbs samplers using Halmos's (1969) theory. The tools developed here can also be utilized to study the convergence rates of Gibbs chains, and reproduce Qin and Jones's (2022) result. Moreover, they can be used to study other variants of two-component Gibbs samplers, as demonstrated in the Appendix. Indeed, there are many potentially interesting variants of two-component Gibbs sampler besides the ones discussed here, and Halmos's (1969) theory opens an avenue for studying them in a manner that was not possible before.

The rest of the article is organized as follows. Section 2 provides the necessary preliminaries, including a short introduction to Halmos's (1969) theory. Section 3 illustrates the usefulness of Halmos's (1969) theory via a comparison between the (standard and modified) DG and RG samplers in terms of adjusted asymptotic variance. A comparison concerning convergence rate is conducted in Section 4. Also included in this section are some general formulas involving Markov chain convergence rate as described in the previous paragraph. Section 5 contains some discussion. Some technical details are relegated to the Appendix. The Appendix also contains an analysis of a fourth type of two-component Gibbs sampler using the theory of two projections.

2. Preliminaries.

2.1. *Basic properties of two-component Gibbs samplers.* The transition laws of the DG and RG Markov chains are described by their Markov transition kernels, or Mtks. In general, an Mtk on $(\mathsf{X}, \mathcal{B})$ is a function $K : \mathsf{X} \times \mathcal{B} \to [0, 1]$ such that $K(x, \cdot)$ is a probability measure for $x \in \mathsf{X}$, and $K(\cdot, A)$ is a measurable function for $A \in \mathcal{B}$. If K and G are Mtks on $(\mathsf{X}, \mathcal{B})$, then their mixture $aK + (1 - a)G$ with $a \in [0, 1]$ defines an Mtk such that

$$[aK + (1 - a)G](x, A) = aK(x, A) + (1 - a)G(x, A), \quad \forall x \in \mathsf{X}, A \in \mathcal{B},$$

and their product KG is defined to be an Mtk such that

$$(KG)(x, A) = \int_{\mathsf{X}} K(x, \mathrm{d}x')G(x', A), \quad \forall x \in \mathsf{X}, A \in \mathcal{B}.$$

For a non-negative integer t, the tth power of an Mtk K, K^t, is defined to be $\prod_{s=1}^{t} K$, so that $K^1 = K$, and $K^0(x, \cdot)$ is the point mass at x. For a Markov chain $(\tilde{X}_t)$ that is possibly time-inhomogeneous (e.g., the DG chain), its $(t', t' + t)$-Mtk, where t and t' are non-negative

integers, is an Mtk $K_{t',t'+t}$ such that $K_{t',t'+t}(x,\cdot)$ gives the conditional distribution of $\tilde{X}_{t'+t}$ given $\tilde{X}_{t'} = x$. In particular, call $K_t := K_{0,t}$ the t-step Mtk. If the chain is time-homogeneous with a single-step Mtk K, then $K_{t',t'+t} = K_t = K^t$. If the distribution of $\tilde{X}_0$ is given by some probability measure μ, then

$$(\mu K_t)(\cdot) := \int_{\mathsf{X}} \mu(\mathrm{d}x) K_t(x,\cdot)$$

gives the marginal distribution of $\tilde{X}_t$.

Let us now write down the Mtks of two-component Gibbs samplers. For a point x in a generic measurable space, use $\delta_x(\cdot)$ to denote the point mass concentrated at x. Define, for $(x_1, x_2) \in \mathsf{X}$ and $A \in \mathcal{B}$,

(1)
$$P_1((x_1, x_2), A) = \int_A \pi_1(\mathrm{d}x_2' \mid x_1) \delta_{x_1}(\mathrm{d}x_1'),$$

$$P_2((x_1, x_2), A) = \int_A \pi_2(\mathrm{d}x_1' \mid x_2) \delta_{x_2}(\mathrm{d}x_2').$$

Then P_1 characterizes the transition rule for updating X_2 through the conditional distribution of $X_2 \mid X_1$, and P_2 does the same for updating X_1 through $X_1 \mid X_2$. In other words, for $i \in \{1, 2\}$, P_i leaves X_i the same and updates the other component.

Consider first the Markov chain associated with the DG algorithm. Let $\mathcal{D}_{t',t'+t}$ be its $(t', t'+t)$-Mtk, and let $\mathcal{D}_t = \mathcal{D}_{0,t}$. Then, for a non-negative integer s' and positive integer s, the following hold:

(2)
$$\mathcal{D}_{2s',2(s'+s)-1} = \mathcal{D}_{2s-1} = (P_1 P_2)^{s-1} P_1, \quad \mathcal{D}_{2s',2(s'+s)} = \mathcal{D}_{2s} = (P_1 P_2)^s,$$

$$\mathcal{D}_{2s'+1,2(s'+s)} = \mathcal{D}_{1,2s} = (P_2 P_1)^{s-1} P_2, \quad \mathcal{D}_{2s'+1,2(s'+s)+1} = \mathcal{D}_{1,2s+1} = (P_2 P_1)^s.$$

On the other hand, the RG sampler simulates a time-homogeneous Markov chain, whose t-step Mtk is

(3)
$$\mathcal{R}_t := [(1-r)P_1 + rP_2]^t.$$

One of the main goals of the current work is to conduct a comparison between the two algorithms in terms of asymptotic variance. As a shorthand notation, for any signed measure μ and measurable function f on $(\mathsf{X}, \mathcal{B})$, let $\mu f = \int_{\mathsf{X}} f(x)\mu(\mathrm{d}x)$ whenever the integral is well-defined. Fix a function $f : \mathsf{X} \to \mathbb{R}$ such that $\pi f^2 < \infty$, where $f^2(x) := f(x)^2$ for $x \in \mathsf{X}$. Consider the asymptotic variance of $S_T(f)$, where $S_T(f)$ is the Monte Carlo sample mean based on either the DG or RG sampler; see the Introduction. Without loss of generality, assume that $\pi f = 0$. Consider first the DG sampler. Denote by $(\tilde{X}_t)$ a DG Markov chain with $\tilde{X}_0 \sim \pi$. By Lemma 24 in Maire, Douc and Olsson (2014), the asymptotic variance can be calculated as follows:

(4)
$$V_{\mathsf{D}}(f) = E[f(\tilde{X}_0)^2] + \sum_{s=1}^{\infty} E[f(\tilde{X}_0)f(\tilde{X}_{2s-1})] + \sum_{s=1}^{\infty} E[f(\tilde{X}_0)f(\tilde{X}_{2s})]$$

$$+ \sum_{s=1}^{\infty} E[f(\tilde{X}_1)f(\tilde{X}_{2s})] + \sum_{s=1}^{\infty} E[f(\tilde{X}_1)f(\tilde{X}_{2s+1})].$$

See also Greenwood, McKeague and Wefelmeyer (1998). For the RG algorithm with selection probability $r \in (0, 1)$, which simulates a time-homogeneous Markov chain, the asymptotic variance is a lot simpler. See, e.g., Jones (2004), Corollary 1. Let $(\tilde{X}_t)$ instead be an RG chain with $\tilde{X}_0 \sim \pi$. Then the asymptotic variance is

(5)
$$V_{\mathsf{R}}(f, r) = E[f(\tilde{X}_0)^2] + 2\sum_{t=1}^{\infty} E[f(\tilde{X}_0)f(\tilde{X}_t)].$$

These two quantities will be analyzed in detail in the next subsection. As mentioned in the Introduction, one needs to take into account the computation costs of the two algorithms. Suppose that the time it takes to sample from $\pi_1(\cdot \mid x_1)$ is a constant (i.e., independent of x_1), and set this to be unit time. Suppose that the time it takes to sample from $\pi_2(\cdot \mid x_2)$ is also a constant, denoted by τ. Then, on average, the DG algorithm takes $(1 + \tau)/2$ units of time to produce a sample point, whereas the RG algorithm takes $r\tau + 1 - r$ to do so. Although the above assumption may not always hold in practice, the simplification is necessary for analyses herein. Multiplying the asymptotic variance of a sampler by the average time it takes to produce a sample point gives the computation time adjusted asymptotic variance. For the DG sampler, the adjusted asymptotic variance is

$$V_{\mathrm{D}}^{\dagger}(f) = \frac{(1 + \tau)}{2} V_{\mathrm{D}}(f).$$

For the RG sampler, the adjusted asymptotic variance is

$$V_{\mathrm{R}}^{\dagger}(f, r) = (r\tau + 1 - r) V_{\mathrm{R}}(f, r).$$

2.2. *Markov operators.* Suppose that an Mtk K on $(\mathsf{X}, \mathcal{B})$ has π as a stationary distribution in the sense that $\pi K = \pi$. Consider $L^2(\pi)$, the linear space of measurable functions $f : \mathsf{X} \to \mathbb{R}$ such that

$$\|f\| := \sqrt{\pi f^2} < \infty.$$

For $f, g \in L^2(\pi)$, say $f = g$ if $\|f - g\| = 0$, i.e., $f(x) = g(x)$ π-almost everywhere. Define an inner product

$$\langle f, g \rangle = \int_{\mathsf{X}} f(x) g(x) \pi(\mathrm{d}x).$$

Then $(L^2(\pi), \langle \cdot, \cdot \rangle)$ forms a real Hilbert space and $\|\cdot\|$ is the L^2 norm. It is convenient to define $L_0^2(\pi)$, the subspace of $L^2(\pi)$ that is orthogonal to constant functions. In other words, $L_0^2(\pi)$ consists of $f \in L^2(\pi)$ such that

$$\langle f, 1 \rangle = \pi f = 0.$$

We can define a linear operation $f \mapsto Kf$ for $f \in L_0^2(\pi)$ in the following way:

$$Kf(x) = \int_{\mathsf{X}} K(x, \mathrm{d}x') f(x'), \quad x \in \mathsf{X}.$$

One can verify that $Kf \in L_0^2(\pi)$ whenever $f \in L_0^2(\pi)$. Moreover, by Cauchy-Schwarz, $\|Kf\| \leq \|f\|$. Therefore, K can be regarded as a bounded linear operator on $L_0^2(\pi)$. This is called a Markov operator. The operator associated with the mixture (product) of two Mtks is simply the mixture (product) of their operators.

The Mtks P_1 and P_2, as defined in (1), satisfy $\pi P_1 = \pi P_2 = \pi$, and thus give rise to the following operators on $L_0^2(\pi)$:

(6)
$$P_1 f(x_1, x_2) = \int_{\mathsf{X}_2} f(x_1, x_2') \pi_1(\mathrm{d}x_2' \mid x_1),$$

$$P_2 f(x_1, x_2) = \int_{\mathsf{X}_1} f(x_1', x_2) \pi_2(\mathrm{d}x_1' \mid x_2).$$

The Mtks of the DG and RG samplers, as given in (2) and (3), can then be treated as linear operators on $L_0^2(\pi)$ as well. It is straightforward to verify that P_1 and P_2 are self-adjoint, i.e., for $i = \{1, 2\}$ and $f, g \in L_0^2(\pi)$,

$$\langle P_i f, g \rangle = \langle f, P_i g \rangle.$$

Moreover, P_1 and P_2 are idempotent, i.e., $P_1^2 = P_1$ and $P_2^2 = P_2$. Indeed, for $i \in \{1, 2\}$, P_i is the orthogonal projection onto the space of functions in $L_0^2(\pi)$ that only depend on x_i.

Various quantities of interest can be studied within the above operator theoretic framework. For instance, if $K(x, \cdot)$ is the conditional distribution of some random element Y given $X = x$, where $X \sim \pi$, then, for any $f, g \in L_0^2(\pi)$,

$$\langle f, Kg \rangle = \text{cov}(f(X), g(Y)).$$

In particular, for $f \in L_0^2(\pi)$, the asymptotic variances in (4) and (5) can be written as

$$
\begin{aligned}
V_{\text{D}}(f) = \|f\|^2 &+ \sum_{s=1}^{\infty} \langle f, \mathcal{D}_{0,2s-1} f \rangle + \sum_{s=1}^{\infty} \langle f, \mathcal{D}_{0,2s} f \rangle \\
&+ \sum_{s=1}^{\infty} \langle f, \mathcal{D}_{1,2s} f \rangle + \sum_{s=1}^{\infty} \langle f, \mathcal{D}_{1,2s+1} f \rangle \\
= \|f\|^2 &+ \sum_{s=1}^{\infty} \langle f, (P_1 P_2)^{s-1} P_1 f \rangle + \sum_{s=1}^{\infty} \langle f, (P_1 P_2)^s f \rangle \\
&+ \sum_{s=1}^{\infty} \langle f, (P_2 P_1)^{s-1} P_2 f \rangle + \sum_{s=1}^{\infty} \langle f, (P_2 P_1)^s f \rangle,
\end{aligned}
$$
(7)

and

$$(8) \qquad V_{\text{R}}(f, r) = \|f\|^2 + 2\sum_{t=1}^{\infty} \langle f, \mathcal{R}_t f \rangle = \|f\|^2 + 2\sum_{t=1}^{\infty} \langle f, [(1-r)P_1 + rP_2]^t f \rangle.$$

These asymptotic variances will be carefully compared in Section 3.

To continue, I will describe the algebraic foundation of this work.

2.3. *Halmos's Theory of Two Projections.* Just for this subsection, let P_1 and P_2 be two orthogonal projection operators on a generic Hilbert space $(H, \langle \cdot, \cdot \rangle)$. Let H_1 be the range of P_1, and H_2, that of P_2. Then H has the following orthogonal decomposition:

$$H = M_{00} \oplus M_{01} \oplus M_{10} \oplus M_{11} \oplus M_{\text{R}},$$

where $\oplus$ denotes direct sum, $M_{00} = H_1 \cap H_2$, $M_{01} = H_1 \cap H_2^{\perp}$, $M_{10} = H_1^{\perp} \cap H_2$, $M_{11} = H_1^{\perp} \cap H_2^{\perp}$, and M_{R} is the rest. Here, $\perp$ denotes the orthogonal complement of a subspace. P_1 and P_2 leave each of M_{00}, M_{01}, M_{10}, M_{11}, and M_{R} invariant. Thus, M_{R} can be further decomposed into $M_{\text{R}} = (P_1 M_{\text{R}}) \oplus (I - P_1) M_{\text{R}}$, where I is the identity operator on M_{R}. The behavior of P_1 and P_2 on M_{ij} for $i, j \in \{0, 1\}$ is simple. How the two projections act on $M_{\text{R}} = P_1 M_{\text{R}} \oplus (I - P_1) M_{\text{R}}$ is, on the other hand, non-trivial.

I will make use of an important result in linear algebra by Halmos (1969). Some additional concepts are needed to state the result. Let G' and H' be subspaces of some linear space, and let G'' and H'' be subspaces of another. Suppose that $G' \cap H' = 0$ and $G'' \cap H'' = 0$. A linear transformation A from $G' \oplus H'$ to $G'' \oplus H''$ has a 2×2 matrix representation. To see this, let P be the linear operator that maps any $g_0 + g_1 \in G'' \oplus H''$ (where $g_0 \in G''$ and $g_1 \in H''$) to $g_0 + 0$. Let $A_{00} = PA|_{G'}$ (where $A|_{G'}$ means A restricted to G'), $A_{01} = PA|_{H'}$, $A_{10} = (I - P)A|_{G'}$, $A_{11} = (I - P)A|_{H'}$. Then, for $g_0 + g_1 \in G' \oplus H'$,

$$A \sum_{i=0}^{1} g_i = \sum_{i=0}^{1} \sum_{j=0}^{1} A_{ij} g_j,$$

and one may write

$$A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix}.$$

Linear combinations and products of linear transformations can be expressed as those of their matrix representations. Finally, a self-adjoint operator B on a generic Hilbert space is called a positive contraction if both B and identity minus B are positive semi-definite.

Consider the matrix representations of P_1 and P_2 restricted to $M_R = P_1 M_R \oplus (I - P_1) M_R$. They are characterized by the following famous result.

LEMMA 2.1. *(Halmos, 1969) Assume that $M_R \neq \{0\}$. Then $P_1 M_R$ and $(I - P_1) M_R$ are both nontrivial and have the same dimension. Moreover, there exist a unitary transformation $W : (I - P_1) M_R \to P_1 M_R$ and positive contractions C and S on $P_1 M_R$ that have the following properties:*

1. $C^2 + S^2 = I_0$, *where I_0 the identity operator on $P_1 M_R$;*
2. $Ker(C) = Ker(S) = \{0\}$, *where Ker denotes the kernel of an operator;*
3.

$$P_1|_{M_R} = \Gamma^* \begin{pmatrix} I_0 & 0 \\ 0 & 0 \end{pmatrix} \Gamma,$$

$$P_2|_{M_R} = \Gamma^* \begin{pmatrix} C^2 & CS \\ CS & S^2 \end{pmatrix} \Gamma,$$

where

$$\Gamma = I_0 \oplus W := \begin{pmatrix} I_0 & 0 \\ 0 & W \end{pmatrix},$$

and $\Gamma^* = I_0 \oplus W^*$ is its adjoint.

REMARK 2.2. S is but a compact way of writing $\sqrt{I_0 - C^2}$. In particular, $CS = SC$.

REMARK 2.3. For more background on the theory of two projections, see Böttcher and Spitkovsky (2010). For a proof of Halmos's result, see Böttcher and Spitkovsky (2018).

Using Lemma 2.1, one can obtain useful representations of mixtures of two projections. For $r \in (0, 1)$,

$$(1 - r)P_1|_{M_R} + rP_2|_{M_R} = \Gamma^* \begin{pmatrix} (1 - r)I_0 + rC^2 & rCS \\ rCS & rS^2 \end{pmatrix} \Gamma.$$

Let

$$\Delta(C) = (1 - 2r)^2 I_0 + 4r(1 - r)C^2.$$

The following is a result given by Nishio (1985).

LEMMA 2.4. *(Nishio, 1985) Assume that $M_R \neq \{0\}$. For $r \in (0, 1)$, there exists a unitary operator $U : M_R \to M_R$ such that*

$$(1 - r)P_1|_{M_R} + rP_2|_{M_R} = \Gamma^* U \begin{pmatrix} [I_0 + \sqrt{\Delta(C)}]/2 & 0 \\ 0 & [I_0 - \sqrt{\Delta(C)}]/2 \end{pmatrix} U^* \Gamma.$$

An elementary proof of Lemma 2.4 is provided in Appendix B, along with a matrix representation of U which is not given in Nishio (1985).

In the context of Gibbs samplers, $H = L_0^2(\pi)$, and P_1 and P_2 are projection operators given by (6). H_1 consists of functions $f \in L_0^2(\pi)$ such that $f(x_1, x_2)$ depends only on x_1, while H_2 consists of functions $f \in L_0^2(\pi)$ such that $f(x_1, x_2)$ depends only on x_2. The DG algorithm can be viewed as an alternating projection algorithm (Diaconis, Khare and Saloff-Coste, 2010).

The subspaces M_{ij}, $i, j \in \{0, 1\}$, also have interpretations, although they are not always easy to identify in practice. The space M_{00} consists of $f \in L_0^2(\pi)$ such that $f(x_1, x_2)$ can be written as a function of just $x_1 \in X_1$ as well as one of just $x_2 \in X_2$. The space M_{01} consists of $f \in L_0^2(\pi)$ such that $f(x_1, x_2)$ depends only on x_1 and that $E[f(X_1, X_2)|X_2] = E[f(X_1, X_2)] = 0$, where $(X_1, X_2) \sim \pi$. The space M_{10} consists of $f \in L_0^2(\pi)$ such that $f(x_1, x_2)$ depends only on x_2 and that $E[f(X_1, X_2)|X_1] = 0$. The space M_{11} consists of $f \in L_0^2(\pi)$ such that $E[f(X_1, X_2)|X_1] = 0$ and $E[f(X_1, X_2)|X_2] = 0$. Note that, for $f \in L_0^2(\pi)$, to say that $E[f(X_1, X_2)|X_i] = 0$ is to say that $f(X_1, X_2)$ is uncorrelated with any square integrable function of X_i.

Starting from the next section, I will assume that $M_{00} = \{0\}$. To see what this assumption entails, consider the contrapositive, that is, there exist a nonzero $L_0^2(\pi)$ function f, some $g : X_1 \to \mathbb{R}$, and some $h : X_2 \to \mathbb{R}$ such that $f(x, y) = g(x) = h(y)$ for π-a.e. (x, y). Then there exists measurable $B \subset \mathbb{R}$ such that $\pi[f^{-1}(B)] > 0$ and $\pi[f^{-1}(B^c)] > 0$. Moreover, $f^{-1}(B)$ and $g^{-1}(B) \times h^{-1}(B)$ differ only by a π-measure zero set. The same goes for $f^{-1}(B^c)$ and $g^{-1}(B^c) \times h^{-1}(B^c)$. Then π has positive mass on $B' = g^{-1}(B) \times h^{-1}(B)$ and $B'' = g^{-1}(B^c) \times h^{-1}(B^c)$, and zero mass on everything else. This is a "reducible" structure that would render two-component Gibbs algorithms ineffective, as any chain that starts in B' cannot enter B'', and any chain that starts in B'' cannot enter B'.

Another assumption that will be made is that $M_R \neq \{0\}$. When $M_{00} = \{0\}$, this assumption states that H_1 and H_2 are not orthogonal. In the context of Gibbs samplers, this assumption is equivalent to X_1 not being independent of X_2, where (X_1, X_2) has joint distribution π. Indeed, for $f, g \in L_0^2(\pi)$,

$$\langle f, g \rangle = \mathrm{cov}[f(X_1, X_2), g(X_1, X_2)].$$

Therefore, there exist $f \in H_1$ and $g \in H_2$ such that $\langle f, g \rangle \neq 0$ if and only if X_1 and X_2 are not independent.

Finally, consider $\|C\|$, where $\|\cdot\|$ denotes the L^2 operator norm. (There is a slight abuse of notation since $\|\cdot\|$ is also used to denote the L^2 norm on $H = L_0^2(\pi)$.) It is well-known that $\|C\|$ is the cosine of the Friedrichs angle between H_1 and H_2, given by

$$\sup \left\{ |\langle f, g \rangle| : f \in H_1 \cap (H_1 \cap H_2)^\perp, \ g \in H_2 \cap (H_1 \cap H_2)^\perp, \ \|f\| = \|g\| = 1 \right\}.$$

See, e.g., Böttcher and Spitkovsky (2010), Example 3.10. For $(X_1, X_2) \sim \pi$, this is the maximal correlation between X_1 and X_2. By Theorem 3.2 in Liu, Wong and Kong (1994) along with Lemma 3.2, Theorem 4.1, and Proposition 3.5 in Qin and Jones (2022), when $\mathcal{B}$ is countably generated, the DG/RG Markov chain is geometrically ergodic only if $\|C\| < 1$.

EXAMPLE 2.5. Suppose that $X_1 = \mathbb{R}^p$ for some positive integer p, $X_2 = \mathbb{R}$, and π is the distribution of a $(p + 1)$-dimensional normal vector

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{p+1} \left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} A & b \\ b^\top & 1 \end{pmatrix}^{-1} \right),$$

where $m_1 \in \mathbb{R}$, $m_2 \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$ is positive definite, and $b \in \mathbb{R}^p$. To ensure that the precision matrix is positive definite, assume that $b^\top A^{-1} b < 1$. Assume also that $b \neq 0$, so that X_1

and X_2 are dependent. MCMC is not needed to sample from π, but pretend that we are to apply two-component Gibbs anyway. The full conditional $\pi_1(\cdot|x_1)$ is a normal distribution with mean $m_2 - b^\top(x_1 - m_1)$ and variance 1, and $\pi_2(\cdot|x_2)$ is a p-variate normal distribution with mean $m_1 - A^{-1}b(x_2 - m_2)$ and covariance matrix A^{-1}. Evidently, $M_{00} = \{0\}$. The subspace M_R may be difficult to conceptualize, but the following calculation gives two non-trivial elements of M_R. Let $f \in L_0^2(\pi)$ be such that

$$(9) \qquad f(x_1, x_2) = x_2 - m_2 + b^\top(x_1 - m_1).$$

Claim: $f \in (I - P_1)M_R$. To see this, consider $P_1 P_2 f$, given by

$$P_1 P_2 f(x_1, x_2) = -(1 - b^\top A^{-1}b)b^\top(x_1 - m_1).$$

Then $P_1 P_2 f \in H_1 = M_{01} \oplus P_1 M_R$. For any $g \in M_{01} = H_1 \cap H_2^\perp$, $\langle P_1 P_2 f, g \rangle = \langle P_2 f, g \rangle = 0$. Thus $P_1 P_2 f \in P_1 M_R$. It's easy to verify that f is a linear combination of $P_1 P_2 f$ and $P_2 P_1 P_2 f$, and $P_1 f = 0$. Since M_R is invariant under P_2, it must hold that $f \in (I - P_1)M_R$. Finally, standard convergence analysis shows that the associated DG Markov chain is geometrically ergodic, so $\|C\| < 1$. In fact, the squared maximal correlation $\|C\|^2$ between X_1 and X_2, is known to be $b^\top A^{-1}b$ (see, e.g., Amit, 1991, Lemma 3).

3. Comparing Asymptotic Variances. We now use matrix representations of P_1 and P_2 to study the DG and RG algorithms. Throughout this section, assume that $M_{00} = \{0\}$, $M_R \neq \{0\}$, and that $\|C\| < 1$.

3.1. *Characterizing $V_D(f)$ and $V_R(f, r)$.* Fix $f \in L_0^2(\pi)$. Then one has the orthogonal decomposition

$$(10) \qquad f = f_{00} + f_{01} + f_{10} + f_{11} + f_0 + f_1,$$

where $f_{ij} \in M_{ij}$, $f_0 \in P_1 M_R$, and $f_1 \in (I - P_1)M_R$. Note that $f_{00} = 0$. Moreover, $P_1 f_{01} = f_{01}$, $P_1 f_{10} = P_1 f_{11} = 0$, $P_2 f_{10} = f_{10}$, and $P_2 f_{01} = P_2 f_{11} = 0$.

Let us characterize the asymptotic variance of the DG sampler using Lemma 2.1.

PROPOSITION 3.1.

$$(11) \qquad V_D(f) = 2\|f_{01}\|^2 + 2\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_D(f_0 + f_1) \rangle,$$

where

$$(12) \qquad \Sigma_D = \Gamma^* \begin{pmatrix} 2I_0 + 4C^2 S^{-2} & 2CS^{-1} \\ 2CS^{-1} & 2I_0 \end{pmatrix} \Gamma.$$

PROOF. Recall from (7) that

$$V_D(f) = \|f\|^2 + \sum_{s=1}^{\infty} \langle f, (P_1 P_2)^{s-1} P_1 f \rangle + \sum_{s=1}^{\infty} \langle f, (P_1 P_2)^s f \rangle$$

$$+ \sum_{s=1}^{\infty} \langle f, (P_2 P_1)^{s-1} P_2 f \rangle + \sum_{s=1}^{\infty} \langle f, (P_2 P_1)^s f \rangle.$$

By the decomposition of f, this can be written as

$$(13) \qquad V_D(f) = 2\|f_{01}\|^2 + 2\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma(f_0 + f_1) \rangle,$$

where

$$\Sigma = I + \sum_{s=1}^{\infty} (P_1|_{M_R} P_2|_{M_R})^{s-1} P_1|_{M_R} + \sum_{s=1}^{\infty} (P_1|_{M_R} P_2|_{M_R})^s +$$

$$\sum_{s=1}^{\infty} (P_2|_{M_R} P_1|_{M_R})^{s-1} P_2|_{M_R} + \sum_{s=1}^{\infty} (P_2|_{M_R} P_1|_{M_R})^s.$$

For $s \geq 1$, we have the matrix representations

$$(P_1|_{M_R} P_2|_{M_R})^s = \Gamma^* \begin{pmatrix} C^{2s} & C^{2s-1}S \\ 0 & 0 \end{pmatrix} \Gamma, \qquad (P_2|_{M_R} P_1|_{M_R})^s = \Gamma^* \begin{pmatrix} C^{2s} & 0 \\ C^{2s-1}S & 0 \end{pmatrix} \Gamma.$$

It follows that

$$\Sigma = \Gamma^* \begin{pmatrix} 2I_0 + 4\sum_{s=1}^{\infty} C^{2s} & 2CS \sum_{s=0}^{\infty} C^{2s} \\ 2CS \sum_{s=0}^{\infty} C^{2s} & I_0 + S^2 \sum_{s=0}^{\infty} C^{2s} \end{pmatrix} \Gamma.$$

Since $\|C\| < 1$,

$$\sum_{s=0}^{\infty} C^{2s} = (I_0 - C^2)^{-1} = S^{-2},$$

and $\Sigma = \Sigma_D$. (13) then gives the desired result. $\qquad\square$

Next, consider the RG sampler with selection probability $r \in (0,1)$.

PROPOSITION 3.2.

$$(14) \qquad V_R(f,r) = \frac{2-r}{r} \|f_{01}\|^2 + \frac{1+r}{1-r} \|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_R(r)(f_0 + f_1) \rangle,$$

where

$$(15) \qquad \Sigma_R(r) = \Gamma^* \begin{pmatrix} \frac{2-r}{r} I_0 + \frac{2}{r(1-r)} C^2 S^{-2} & \frac{2}{1-r} CS^{-1} \\ \frac{2}{1-r} CS^{-1} & \frac{1+r}{1-r} I_0 \end{pmatrix} \Gamma.$$

PROOF. Note that

$$[(1-r)P_1 + rP_2]f = (1-r)f_{01} + rf_{10} + [(1-r)P_1|_{M_R} + rP_2|_{M_R}](f_0 + f_1),$$

where

$$(1-r)P_1|_{M_R} + rP_2|_{M_R} = \Gamma^* \begin{pmatrix} (1-r)I_0 + rC^2 & rCS \\ rCS & rS^2 \end{pmatrix} \Gamma.$$

By (8), the asymptotic variance is then
(16)
$$V_R(f,r) = \|f\|^2 + 2\sum_{t=1}^{\infty} \langle f, [(1-r)P_1 + rP_2]^t f \rangle$$

$$= \|f\|^2 + 2\sum_{t=1}^{\infty} \left[(1-r)^t \|f_{01}\|^2 + r^t \|f_{10}\|^2 + \langle f_0 + f_1, [(1-r)P_1|_{M_R} + rP_2|_{M_R}]^t (f_0 + f_1) \rangle \right]$$

$$= \frac{2-r}{r} \|f_{01}\|^2 + \frac{1+r}{1-r} \|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma(f_0 + f_1) \rangle,$$

where

$$\Sigma = I + 2\sum_{t=1}^{\infty}[(1-r)P_1|_{M_R} + rP_2|_{M_R}]^t.$$

By Lemma 2.4,

$$\|(1-r)P_1|_{M_R} + rP_2|_{M_R}\| = \frac{1}{2}\left\|I_0 + \sqrt{\Delta(C)}\right\|.$$

It follows that the Neumann series in the expression of Σ is convergent whenever $\|I_0 + \Delta(C)\| < 2$, which is equivalent to $\|C\| < 1$. Σ can then be written as

$$\Sigma = 2\left[I - (1-r)P_1|_{M_R} - rP_2|_{M_R}\right]^{-1} - I$$

$$= 2\Gamma^*\left(\begin{array}{cc} rS^2 & -rCS \\ -rCS & I_0 - rS^2 \end{array}\right)^{-1}\Gamma - I.$$

Evaluating the inverse may seem daunting, but it is completely analogous to calculating the inverse of a 2×2 matrix. It is straightforward to verify that for $x \in [0,1]$,

$$\left(\begin{array}{cc} r(1-x^2) & -rx\sqrt{1-x^2} \\ -rx\sqrt{1-x^2} & 1 - r(1-x^2) \end{array}\right)^{-1}$$

$$= \frac{1}{r(1-r)}\left(\begin{array}{cc} -r + (1-x^2)^{-1} & rx/\sqrt{1-x^2} \\ rx/\sqrt{1-x^2} & r \end{array}\right).$$

Analogously,

$$\left(\begin{array}{cc} rS^2 & -rCS \\ -rCS & I_0 - rS^2 \end{array}\right)^{-1} = \frac{1}{r(1-r)}\left(\begin{array}{cc} -rI_0 + S^{-2} & rCS^{-1} \\ rCS^{-1} & rI_0 \end{array}\right).$$

Routine calculations show that $\Sigma = \Sigma_R(r)$, and the desired result follows from (16). □

EXAMPLE 3.3. This is a continuation of Example 2.5. Recall that f, as given in (9), is in $(I - P_1)M_R$. By Propositions 3.1 and 3.2, $V_D(f) = 2\|f\|^2$, while $V_R(f,r) = (1 + r)\|f\|^2/(1-r)$.

3.2. *The comparison.* Propositions 3.1 and 3.2 allow one to compare $V_D^{\dagger}(f)$ and $V_R^{\dagger}(f,r)$ (as defined at the end of Section 2.1) for $f \in L_0^2(\pi)$, the asymptotic variances of the two samplers after adjusting for computation time.

Consider first the unadjusted asymptotic variances. The key result of this section is as follows.

PROPOSITION 3.4. *For $r \in (0,1)$ and $f \in L_0^2(\pi)$,*

$$V_D(f) \leq k_1(r)V_R(f,r),$$
$$V_R(f,r) \leq k_2(r)V_D(f,r),$$

where

$$k_1(r) = 1 - r + r^2 + \sqrt{r^2(1-r)^2 + (1-2r)^2},$$

$$k_2(r) = \frac{1 - r + r^2}{2r(1-r)} + \frac{\sqrt{r^2(1-r)^2 + (1-2r)^2}}{2r(1-r)}.$$
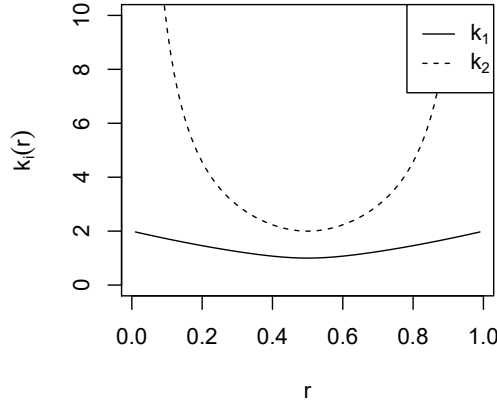
Fig 1: Graph of $r \mapsto k_i(r)$ for $i = 1, 2$.

PROOF. Decompose $f \in L_0^2(\pi)$ as in (10). Recall (11) and (14):

$$V_{\mathrm{D}}(f) = 2\|f_{01}\|^2 + 2\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_{\mathrm{D}}(f_0 + f_1)\rangle,$$

$$V_{\mathrm{R}}(f,r) = \frac{2-r}{r}\|f_{01}\|^2 + \frac{1+r}{1-r}\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_{\mathrm{R}}(r)(f_0 + f_1)\rangle.$$

By Lemma A.3 in the Appendix, $\Sigma_{\mathrm{D}} \leq k_1(r)\Sigma_{\mathrm{R}}(r)$, and $\Sigma_{\mathrm{R}}(r) \leq k_2(r)\Sigma_{\mathrm{D}}$, where $\leq$ denotes Loewner ordering. Then

$$V_{\mathrm{D}}(f) \leq \max\left\{\frac{2r}{2-r}, \frac{2(1-r)}{1+r}, 1, k_1(r)\right\} V_{\mathrm{R}}(f,r),$$

$$V_{\mathrm{R}}(f,r) \leq \max\left\{\frac{2-r}{2r}, \frac{1+r}{2(1-r)}, 1, k_2(r)\right\} V_{\mathrm{D}}(f).$$

The result then follows from Lemma A.1.   □

The bounds in Proposition 3.4 are sharp in the sense that they cannot be improved without additional knowledge on π and f. A more detailed explanation is given in Proposition A.5. It is easy to verify that $1 \leq k_1(r) \leq 2$, and $k_1(1/2) = 1$; see Figure 1. This means that when computation time is not taken into account, the DG algorithm is never much less efficient compared to the RG algorithm in terms of asymptotic variance.

REMARK 3.5. Applying (11)-(15) for $r = 1/2$ yields

$$0 \leq V_{\mathrm{R}}(f, 1/2) - \|f\|^2 = 2[V_{\mathrm{D}}(f) - \|f\|^2].$$

This result can be found in Greenwood, McKeague and Wefelmeyer (1998).

In practice, one needs to compare $V_{\mathrm{D}}^\dagger(f) = (1+\tau)V_{\mathrm{D}}(f)/2$ and $V_{\mathrm{R}}^\dagger(f,r) = (r\tau + 1 - r)V_{\mathrm{R}}(f,r)$, where τ is the time it takes to sample from $\pi_2(\cdot \mid x_2)$ if it takes unit time to sample from $\pi_1(\cdot \mid x_1)$. The following result follows immediately from Proposition 3.4.

COROLLARY 3.6. *For $r \in (0,1)$ and $f \in L_0^2(\pi)$,*

$$V_D^\dagger(f) \leq \kappa(\tau, r) k_1(r) V_R^\dagger(f, r),$$

where $k_1(r)$ is given in Proposition 3.4, and

$$\kappa(\tau, r) = \frac{\tau + 1}{2(r\tau + 1 - r)}.$$

When $\tau \approx 1$, i.e., the times it takes to update X_1 and X_2 are roughly the same, $\kappa(\tau, r) \approx 1$, and $V_D^\dagger(f)$ is never much larger than $V_R^\dagger(f, r)$. On the other hand, if $\tau \ll 1$ or $\tau \gg 1$, i.e., the times it takes to update X_1 and X_2 are significantly different, then for some values of r, e.g., $r = 1/\sqrt{\tau}$ when $\tau \gg 1$ and $r = 1 - \sqrt{\tau}$ when $\tau \ll 1$, $\kappa(\tau, r)$ can be very large. Intuitively, if $\tau \gg 1$, it makes sense to sample from $\pi_2(\cdot \mid x_2)$ less frequently, i.e., pick a smaller r. The following toy example shows that when $\tau \gg 1$ and $r \ll 1$, $V_R^\dagger(f, r)$ can be considerably smaller than $V_D^\dagger(f)$ for some function f.

EXAMPLE 3.7. This is a continuation of Example 3.3. Recall that for the particular choice of f, $V_D(f) = 2\|f\|^2$ and $V_R(f, r) = (1 + r)\|f\|^2/(1 - r)$. Suppose that p is large, and it takes a long time to update X_1 compared to X_2. Then $\tau \gg 1$. Avoid updating X_1 frequently by setting $r \ll 1$. Then

$$\frac{V_R^\dagger(f, r)}{V_D^\dagger(f)} = \frac{r\tau + 1 - r}{\tau + 1} \frac{1 + r}{1 - r} \ll 1.$$

A general statement is given by the following result.

COROLLARY 3.8. *Let r be a function of τ such that $r \to 0$ if $\tau \to \infty$. Then, for any nonzero $f \in M_{10} \oplus M_{11} \oplus (I - P_1)M_R$, as $\tau \to \infty$,*

$$\frac{V_R^\dagger(f, r)}{V_D^\dagger(f)} \to 0.$$

By symmetry, a similar result holds as $\tau \to 0$ and $r \to 1$.

PROOF. Let $f = f_{10} + f_{11} + f_1 \in M_{10} \oplus M_{11} \oplus (I - P_1)M_R$ be a nonzero function. By (11)-(15),

$$V_D(f) = 2\|f_{10}\|^2 + \|f_{11}\|^2 + 2\|f_1\|^2 \geq \|f\|^2,$$

$$V_R(f, r) = \frac{1 + r}{1 - r}\|f_{10}\|^2 + \|f_{11}\|^2 + \frac{1 + r}{1 - r}\|f_1\|^2 \leq \frac{1 + r}{1 - r}\|f\|^2.$$

Then

$$\frac{V_R^\dagger(f, r)}{V_D^\dagger(f)} \leq \frac{2(r\tau + 1 - r)}{\tau + 1} \frac{1 + r}{1 - r}.$$

The result follows. $\square$

REMARK 3.9. The space $M_{10} \oplus M_{11} \oplus (I - P_1)M_R$ consists of functions that are orthogonal to $M_{01} \oplus P_1 M_R$, which is the space of functions $f \in L_0^2(\pi)$ such that $f(x_1, x_2)$ depends only on x_1. In other words, $f \in M_{10} \oplus M_{11} \oplus (I - P_1)M_R$ if and only if $f(X_1, X_2)$ is uncorrelated with any L^2 function of (X_1, X_2) that only depends on X_1, where $(X_1, X_2) \sim \pi$.

The question remains whether an RG sampler with a well-chosen r performs well compared to the DG sampler for an arbitrary function f. This is answered by the following.

COROLLARY 3.10.   *For $r \in (0,1)$ and $f \in L_0^2(\pi)$,*

(17)
$$V_R^\dagger(f,r) \le \frac{k_2(r)}{\kappa(\tau,r)} V_D^\dagger(f) = \frac{2(r\tau + 1 - r)}{\tau + 1} \frac{1 - r + r^2 + \sqrt{r^2(1-r)^2 + (1-2r)^2}}{2r(1-r)} V_D^\dagger(f),$$

where $k_2(r)$ is given in Proposition 3.4, and $k(\tau,r)$ is given in Corollary 3.6. In particular, if we let $r = 0.5$ when $\tau = 1$, and

(18)
$$r = \frac{-2\tau - 1 + \sqrt{\tau(2\tau + 1)(\tau + 2)}}{\tau^2 - 1} \in (0,1)$$

when $\tau \ne 1$, then $k_2(r)/\kappa(\tau,r) = 2$, and thus

$$V_R^\dagger(f,r) \le 2V_D^\dagger(f).$$

PROOF.   (17) follows from Proposition 3.4.

When $\tau = 1$ and $r = 0.5$, the result clearly holds. Fix $\tau \ne 1$ and let r be as in (18). It's straightforward to verify that $r \in (0,1)$. Moreover, solving for τ yields

$$\tau = \frac{1 - 2r + \sqrt{(1-2r)^2 + r^2(1-r)^2}}{r^2}.$$

Then

$$\frac{1}{\kappa(\tau,r)} = \frac{2r[1 - r - r^2 + \sqrt{(1-2r)^2 + r^2(1-r)^2}]}{(1-r)^2 + \sqrt{(1-2r)^2 + r^2(1-r)^2}}.$$

Routine calculations show that $k_2(r)/\kappa(\tau,r) = 2$.   □
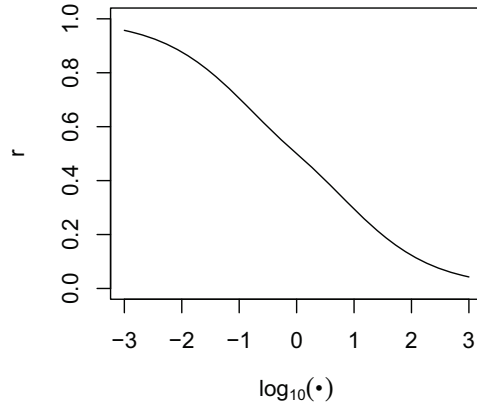

Fig 2: Relationship between τ and r as given in (18). τ is given in log scale.

The relationship between r and τ as given by (18) is plotted in Figure 2. It can be shown that this choice of r is not much worse than optimal in terms of minimizing $k_2(r)/\kappa(\tau,r)$.

Moreover, as $\tau \to \infty$, $r \to 0$, and as $\tau \to 0$, $r \to 1$. In other words, this choice of r also satisfies the conditions in Corollary 3.8.

Corollaries 3.8 and 3.10 show that, in terms of adjusted asymptotic variance, when τ is extremely large or extremely small, the RG sampler with a good choice of r is never significantly worse than the DG sampler, and can be considerably better in some cases. One can say that the RG sampler is more robust in this regard.

One may also use other choices of r. By Corollaries 3.8 and 3.10, similar effects can be achieved whenever r is a decreasing function of τ such that $1 \ll 1/r = O(\tau)$ when $\tau \to \infty$ and $1 \ll 1/(1-r) = O(1/\tau)$ when $\tau \to 0$. Here, $1/r = O(\tau)$ means $1/(r\tau)$ is bounded.

3.3. *Data augmentation.* Studying the asymptotic variance $V(f)$ for all $f \in L_0^2(\pi)$ is relevant only if we care about integrating all L^2 functions, which is not always the case. For example, in a data augmentation setting (Tanner and Wong, 1987; van Dyk and Meng, 2001), one usually only wishes to estimate πf for $f \in L^2(\pi)$ such that $f(x_1, x_2)$ just depends on x_1. In this case, it is only interesting to compare $V_{\mathrm{D}}^\dagger(f)$ and $V_{\mathrm{R}}^\dagger(f, r)$ for $f \in H_1 = M_{01} \oplus P_1 M_{\mathrm{R}}$.

Let $f = f_{01} + f_0 \in M_{01} \oplus P_1 M_{\mathrm{R}}$. By (11)-(15),

$$V_{\mathrm{D}}(f) = 2\|f_{01}\|^2 + \langle f_0, (2I_0 + 4C^2 S^{-2})f_0 \rangle,$$

(19)
$$V_{\mathrm{R}}(f, r) = \frac{2-r}{r}\|f_{01}\|^2 + \left\langle f_0, \left(\frac{2-r}{r}I_0 + \frac{2}{r(1-r)}C^2 S^{-2} \right) f_0 \right\rangle.$$

Then

$$V_{\mathrm{D}}(f) \leq \max\left\{ \frac{2r}{2-r}, 2r(1-r) \right\} V_{\mathrm{R}}(f, r).$$

Let $\kappa(\tau, r)$ be defined as in Corollary 3.6. Then

$$V_{\mathrm{D}}^\dagger(f) \leq \kappa(\tau, r) \max\left\{ \frac{2r}{2-r}, 2r(1-r) \right\} V_{\mathrm{R}}^\dagger(f, r).$$

Now,

$$\kappa(\tau, r) \max\left\{ \frac{2r}{2-r}, 2r(1-r) \right\} \leq \frac{\tau+1}{2r\tau} \max\left\{ \frac{2r}{2-r}, 2r(1-r) \right\} \leq \frac{\tau+1}{\tau}.$$

Thus, when τ is not too small, the DG algorithm can compete with the RG algorithm in terms of adjusted asymptotic variance. When $\tau \ll 1$, depending on the structure of π, a result like Corollary 3.8 could still hold for some $f \in H_1$.

3.4. *A Modified DG Sampler.* The RG sampler achieves robustness by updating the less costly component more frequently. Naturally, one can use this idea to modify the DG sampler. Suppose that $\tau \geq 1$, i.e., it is more costly to sample from $\pi_2(\cdot \mid x_2)$ (which updates X_1) than $\pi_1(\cdot \mid x_1)$ (which updates X_2). Consider the following algorithm, which updates X_2 ℓ times consecutively before updating X_1 once, where $\ell \geq 1$. It simulates a Markov chain that again has π as its stationary distribution.

Algorithm 3: Modified DG sampler with ℓ repeated draws from π_1

Draw $\tilde{X}_0 = (X_{1,0}, X_{2,0})$ from some initial distribution on $(\mathsf{X}, \mathcal{B})$, and set $t = 0$;

while $t < T$ **do**

    if $t = s(\ell+1) + q$ *for some non-negative integer s and $q \in \{0, \ldots, \ell-1\}$* **then**

        draw $X_{2,t+1}$ from $\pi_1(\cdot \mid X_{1,t})$, set $X_{1,t+1} = X_{1,t}$, and let
        $\tilde{X}_{t+1} = (X_{1,t+1}, X_{2,t+1})$;

    if $t = s(\ell+1) + \ell$ *for some non-negative integer s* **then**

        draw $X_{1,t+1}$ from $\pi_2(\cdot \mid X_{2,t})$, set $X_{2,t+1} = X_{2,t}$, and let
        $\tilde{X}_{t+1} = (X_{1,t+1}, X_{2,t+1})$;

    set $t = t + 1$;

Let $f \in L_0^2(\pi)$. Let $V_{\mathrm{M}}(f, \ell)$ be the unadjusted asymptotic variance associated with this function and the modified DG algorithm. By Proposition 2 of Greenwood, McKeague and Wefelmeyer (1998),

$$V_{\mathrm{M}}(f, \ell) = \|f\|^2 + \frac{2}{\ell+1}\left[\langle f, P_2 f \rangle + \sum_{s=1}^{\infty} \langle f, (P_2 P_1)^s P_2 f \rangle + \ell \sum_{s=1}^{\infty} \langle f, (P_2 P_1)^s f \rangle \right.$$
$$\left. + \frac{\ell(\ell+1)}{2} \langle f, P_1 f \rangle + \ell \sum_{s=1}^{\infty} \langle f, (P_1 P_2)^s f \rangle + \ell^2 \sum_{s=1}^{\infty} \langle f, (P_1 P_2)^s P_1 f \rangle \right].$$

Assume that $M_{00} = \{0\}$, $M_{\mathrm{R}} \neq \{0\}$, and that $\|C\| < 1$. Similarly to how Propositions 3.1 and 3.2 are proved, one can establish the following.

PROPOSITION 3.11.
$$V_M(f, \ell) = (\ell+1)\|f_{01}\|^2 + \frac{\ell+3}{\ell+1}\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_M(\ell)(f_0 + f_1) \rangle,$$

where

(20) $$\Sigma_M(\ell) = \Gamma^* \begin{pmatrix} (\ell+1)I_0 + 2(\ell+1)C^2 S^{-2} & 2CS^{-1} \\ 2CS^{-1} & \frac{\ell+3}{\ell+1}I_0 \end{pmatrix} \Gamma.$$

REMARK 3.12. If instead one updates X_2 once before updating X_1 ℓ times, then the asymptotic variance is

$$V_{\mathrm{M}}(f, \ell^{-1}) = \frac{\ell+3}{\ell+1}\|f_{01}\|^2 + (\ell+1)\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_{\mathrm{M}}(\ell^{-1})(f_0 + f_1) \rangle,$$

where

$$\Sigma_{\mathrm{M}}(\ell^{-1}) = \Gamma^* \begin{pmatrix} \frac{\ell+3}{\ell+1}I_0 + 2(\ell+1)C^2 S^{-2} - \frac{\ell(\ell-1)}{\ell+1}C^2 & 2\ell CS^{-1} - \frac{\ell(\ell-1)}{\ell+1}CS \\ 2\ell CS^{-1} - \frac{\ell(\ell-1)}{\ell+1}CS & \frac{2\ell^2 + \ell + 1}{\ell+1}I_0 - \frac{\ell(\ell-1)}{\ell+1}S^2 \end{pmatrix} \Gamma.$$

To proceed, let us consider the computational cost of the modified DG algorithm. If the algorithm is run verbatim, then, on average, each iteration takes $(\tau + \ell)/(\ell+1)$ units of time. It is, however, possible to save time through a parallelization scheme. Note that, in Algorithm 3, we may first simulate the subchain

$$\tilde{X}_0, \tilde{X}_\ell, \tilde{X}_{\ell+1}, \tilde{X}_{(\ell+1)+\ell}, \tilde{X}_{2(\ell+1)}, \ldots, \tilde{X}_{(s-1)(\ell+1)+\ell}, \tilde{X}_{s(\ell+1)}, \ldots$$

through the standard DG sampler (Algorithm 1). After simulating the subchain at a given length, for $t = s(\ell+1) + q$ where s is a non-negative integer and $q \in \{1, \ldots, \ell-1\}$,

$\tilde{X}_t = (X_{1,t}, X_{2,t})$ can be generated by setting $X_{1,t} = X_{1,s(\ell+1)}$, and drawing $X_{2,t}$ from $\pi_1(\cdot \mid X_{1,s(\ell+1)})$. Hence, the repeated sampling from π_1 can be done through post-processing. Moreover, in an ideal setting, this can be executed in parallel, resulting in the overall runtime of the modified DG algorithm being essentially equal to the time it takes to simulate the subchain. A modified DG chain of length T corresponds to a subchain of length roughly $2T/(\ell+1)$. Thus, depending on how much parallelization is feasible, the per-iteration cost (in terms of time) of the modified DG algorithm can vary between $2/(\ell+1) \times (\tau+1)/2 = (\tau+1)/(\ell+1)$ and $(\tau+\ell)/(\ell+1)$.

I will first analyze the modified DG sampler when assuming that there is no parallelization. It will be shown that even in the absence of parallelization, this sampler is competitive against the standard DG and RG samplers.

The adjusted asymptotic variance for the modified DG algorithm is

$$V_{\mathrm{M}}^{\dagger}(f,\ell) = \frac{\ell+\tau}{\ell+1} V_{\mathrm{M}}(f,\ell).$$

Let us now compare $V_{\mathrm{M}}^{\dagger}(f,\ell)$ to $V_{\mathrm{D}}^{\dagger}(f)$ and $V_{\mathrm{R}}^{\dagger}(f,r)$.

One would expect that the modified DG sampler behaves similarly to the RG sampler with selection probability $r = 1/(\ell+1)$. The following result shows that the modified DG sampler is better than, but comparable to the RG sampler with $r = 1/(\ell+1)$ in terms of adjusted asymptotic variance.

COROLLARY 3.13.   *For $\ell \geq 1$ and $f \in L_0^2(\pi)$,*

$$V_M^{\dagger}(f,\ell) \leq V_R^{\dagger}\left(f, \frac{1}{\ell+1}\right) \leq \max\left\{\frac{2\ell+1}{\ell+1}, \frac{\ell+1}{\ell}\right\} V_M^{\dagger}(f,\ell).$$

PROOF. Fix $\ell \geq 1$ and $f \in L_0^2(\pi)$. By Lemma A.4,

$$\Sigma_{\mathrm{M}}(\ell) \leq \Sigma_{\mathrm{R}}\left(\frac{1}{\ell+1}\right) \leq \max\left\{\frac{2\ell+1}{\ell+1}, \frac{\ell+1}{\ell}\right\} \Sigma_{\mathrm{M}}(\ell).$$

It then follows from Propositions 3.2 and 3.11 that

$$V_{\mathrm{M}}(f,\ell) \leq V_{\mathrm{R}}\left(f, \frac{1}{\ell+1}\right) \leq \max\left\{\frac{2\ell+1}{\ell+1}, \frac{\ell+1}{\ell}\right\} V_{\mathrm{M}}(f,\ell).$$

The desired result is obtained by noting that $1 - r + r\tau = (\ell+\tau)/(\ell+1)$ if $r = 1/(\ell+1)$. □

A comparison between $V_{\mathrm{R}}^{\dagger}(r)$ and $V_{\mathrm{M}}^{\dagger}(\ell)$ for a general r that is not $1/(\ell+1)$ can be conducted on a case-by-case basis. The key step is verifying whether operators of the form $k\Sigma_{\mathrm{M}}(\ell) - \Sigma_{\mathrm{R}}(r)$ and $k\Sigma_{\mathrm{R}}(r) - \Sigma_{\mathrm{M}}(\ell)$ are positive semi-definite. A tool for this is provided in Lemma A.2.

Naturally, just like the RG sampler, the modified DG sampler enjoys robustness over the vanilla DG sampler, given that ℓ is well-chosen. The next two results are derived from Propositions 3.1, 3.11, and Lemma A.4. The proofs are extremely similar to those of Corollaries 3.8 and 3.10, so they will be omitted.

COROLLARY 3.14.   *Let ℓ be a function of τ such that $\ell \to \infty$ if $\tau \to \infty$. Then, for any nonzero $f \in M_{10} \oplus M_{11} \oplus (I - P_1)M_R$, as $\tau \to \infty$,*

$$\frac{V_M^{\dagger}(f,\ell)}{V_D^{\dagger}(f)} \to 0.$$

COROLLARY 3.15. *For $\ell \geq 1$ and $f \in L_0^2(\pi)$,*

$$V_{\mathrm{M}}^\dagger(f,\ell) \leq \frac{2(\ell+\tau)}{(\ell+1)(\tau+1)} \frac{\ell^2 + \ell + 2 + (\ell-1)\sqrt{(\ell+1)^2+1}}{2(\ell+1)} V_D^\dagger(f).$$

In particular, if

$$(21) \qquad \qquad \ell \leq \frac{1+\sqrt{1+4(\tau+1)}}{2},$$

then

$$V_{\mathrm{M}}^\dagger(f,\ell) \leq 2V_D^\dagger(f).$$

If $\tau \gg 1$, then letting $\ell \gg 1$ under the restriction of (21) would give the desired robustness. In fact, having $\ell \gg 1$ and $\ell = O(\tau)$ would yield a similar effect.

When post-processing and parallelization is fully exploited, $V_{\mathrm{M}}^\dagger(f,\ell)$ can be further diminished. Indeed, if we assume full parallelization and consider only the cost of computation time, then $V_{\mathrm{M}}^\dagger(f,\ell) = (\tau+1)V_{\mathrm{M}}(f,\ell)/(\ell+1)$, rather than $(\tau+\ell)V_{\mathrm{M}}(f,\ell)/(\ell+1)$ as before. One may then multiply a factor of $(\tau+1)/(\tau+\ell)$ to the any upper or lower bounds on $V_{\mathrm{M}}^\dagger(\ell)$ in Corollaries 3.13 and 3.15.

4. Convergence Rate. I will now use Halmos's (1969) theory to study the convergence properties of two-component Gibbs samplers, and provide an alternative proof of Qin and Jones (2022) main result, which, loosely speaking, states that the DG Markov chain converges faster than the RG chain.

I will first provide some general results regarding the convergence rates of possibly time-inhomogeneous Markov chains. Define $L_*^2(\pi)$ to be the set of probability measures μ such that μ is absolutely continuous with respect to π, and that $\mathrm{d}\mu/\mathrm{d}\pi \in L^2(\pi)$. For $\mu, \nu \in L_*^2(\pi)$, one can define their L^2 distance

$$\|\mu - \nu\|_* = \sup_{f \in L_0^2(\pi), \|f\|=1} (\mu f - \nu f) = \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right\|.$$

This distance is often used in the convergence analysis of Markov chains. Let K_t be the t-step transition kernel of a possibly time-inhomogeneous Markov chain such that π is invariant, i.e., $\pi K_t = \pi$ for each positive integer t. It can be checked that $\mu K_t \in L_*^2(\pi)$ whenever $\mu \in L_*^2(\pi)$. If $(\tilde{X}_t)$ is a chain associated with this Mtk and $\tilde{X}_0 \sim \mu \in L_*^2(\pi)$, then μK_t is the marginal distribution of $\tilde{X}_t$. The L^2 distance between μK_t and π is

$$
\begin{aligned}
\|\mu K_t - \pi\|_* &= \sup_{f \in L_0^2(\pi), \|f\|=1} (\mu - \pi)K_t f \\
&= \sup_{f \in L_0^2(\pi), \|f\|=1} \left\langle \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1, K_t f \right\rangle \\
&= \left\| K_t^* \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right) \right\| \\
&= \left\langle \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1, K_t K_t^* \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right) \right\rangle^{1/2},
\end{aligned}
$$

(22)

where K_t^* is the adjoint of K_t.

Evidently, the behavior of the positive semi-definite operator $K_t K_t^*$ as $t \to \infty$ is closely related to the convergence properties of the chain. To make this relationship more explicit, consider the convergence rate

$$\rho_0 = \exp \left(\sup_{\mu \in L_*^2(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu K_t - \pi\|_* \right)$$

(Roberts and Tweedie, 2001; Qin and Hobert, 2020). It can be shown that $\rho_0 \in [0, 1]$. If $\rho_0 < \rho \in (0, 1]$, then for each $\mu \in L_*^2(\pi)$ there exists $M_\mu < \infty$ such that

(23) $$\|\mu K_t - \pi\|_* < M_\mu \rho^t$$

for every positive integer t. Conversely, if (23) holds for each t, then $\rho_0 \leq \rho$. Hence, the smaller the rate, the faster the convergence.

The following result is proved using standard techniques like those from Fill (1991); Roberts and Rosenthal (1997); Paulin (2015), with details given in Appendix C. Readers are also referred to Kontorovich and Ramanan (2008), which establishes concentration inequalities for time-inhomogeneous chains.

PROPOSITION 4.1. *Let K_t be the t-step transition kernel of a Markov chain with stationary distribution π. Then*

$$\sup_{\mu \in L_*^2(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu K_t - \pi\|_* = \log \left(\sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} \right).$$

That is,

$$\rho_0 = \sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)}.$$

Let ρ_D and $\rho_R(r)$ be the convergence rates of the DG and RG chains, respectively, where r is the selection probability. That is,

$$\rho_D := \exp \left(\sup_{\mu \in L_*^2(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu \mathcal{D}_t - \pi\|_* \right),$$

$$\rho_R(r) := \exp \left(\sup_{\mu \in L_*^2(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu R_t - \pi\|_* \right).$$

Let us now use Proposition 4.1 to find the relationship between the two rates. The following lemma is useful.

LEMMA 4.2. *Suppose that in Proposition 4.1, $K_t K_t^* = A^{m(t)}$ for t sufficiently large, where A is a positive contraction on $L_0^2(\pi)$ and $m(t)$ is a non-negative function of t. Then*

$$\sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} = \|A\|^m,$$

where

$$m = \liminf_{t \to \infty} (2t)^{-1} m(t).$$

PROOF. Because A is positive semi-definite and $m(t)$ is non-negative, for $f \in L_0^2(\pi)$,

$$\langle f, A^{m(t)} f \rangle \leq \|A^{m(t)}\| = \|A\|^{m(t)}.$$

Then, since $\|A\| \leq 1$,

$$\sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} = \sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, A^{m(t)} f \rangle^{1/(2t)} \leq \|A\|^m.$$

Consider the reverse inequality, which obviously holds when $\|A\| = 0$. Suppose that $\|A\| > 0$, and let $\varepsilon \in (0, \|A\|)$ be arbitrary. A has a spectral decomposition $A = \int_{-\infty}^{\infty} \lambda E_A(\mathrm{d}\lambda)$, where E_A is the projection-valued measure associated with A (see, e.g., Kubrusly, 2012, Theorem 3.15). There exists a function $f_{(\varepsilon)}$ in the range of $E_A((\|A\| - \varepsilon, \|A\|])$ such that $\|f_{(\varepsilon)}\| = 1$. Then

$$\sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} \geq \limsup_{t \to \infty} \langle f_{(\varepsilon)}, A^{m(t)} f_{(\varepsilon)} \rangle^{1/(2t)}$$

$$= \limsup_{t \to \infty} \left(\int_{-\infty}^{\infty} \lambda^{m(t)} \langle f_{(\varepsilon)}, E_A(\mathrm{d}\lambda) f_{(\varepsilon)} \rangle \right)^{1/(2t)}$$

$$\geq (\|A\| - \varepsilon)^m.$$

Since ε is arbitrary,

$$\sup_{f \in L_0^2(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} \geq \|A\|^m.$$

$\square$

Consider the DG and RG Markov chains. Recall from (2) and (3) that their Mtks are as follows: $\mathcal{D}_{2s-1} = (P_1 P_2)^{s-1} P_1$ and $\mathcal{D}_{2s} = (P_1 P_2)^s$ for any positive integer s, while $\mathcal{R}_t = [(1-r)P_1 + rP_2]^t$ for any positive integer t. Note that for $t \geq 2$, $\mathcal{D}_t \mathcal{D}_t^* = (P_1 P_2 P_1)^{t-1}$. In light of Proposition 4.1 and Lemma 4.2,

$$\rho_{\mathrm{D}} = \|P_1 P_2 P_1\|^{1/2}, \quad \rho_{\mathrm{R}}(r) = \|(1-r)P_1 + rP_2\|.$$

Let $f \in L_0^2(\pi)$ be decomposed as in (10). That is, $f = \sum_{i=0}^{1} \sum_{j=0}^{1} f_{ij} + f_0 + f_1$. Assume that $M_{00} = \{0\}$ and $M_{\mathrm{R}} \neq \{0\}$. One can use Lemma 2.1 to obtain $P_1 P_2 P_1 f = C^2 f_0$. Then $\|P_1 P_2 P_1\| = \|C\|^2$. On the other hand,

$$[(1-r)P_1 + rP_2]f = (1-r)f_{01} + rf_{10} + [(1-r)P_1|_{M_{\mathrm{R}}} + rP_2|_{M_{\mathrm{R}}}](f_0 + f_1).$$

It then follows from Lemma 2.4 that

$$\|(1-r)P_1 + rP_2\| = \max\left\{ 1-r,\, r,\, \frac{\|I_0 + \sqrt{(1-2r)^2 + 4r(1-r)C^2}\|}{2} \right\}$$

$$= \max\left\{ 1-r,\, r,\, \frac{1 + \sqrt{(1-2r)^2 + 4r(1-r)\|C\|^2}}{2} \right\}$$

$$= \frac{1 + \sqrt{(1-2r)^2 + 4r(1-r)\|C\|^2}}{2}.$$

Note that the second equality holds since $0 \leq \|C\| \leq I_0$, and

$$x \mapsto \frac{1 + \sqrt{(1-2r)^2 + 4r(1-r)x^2}}{2}$$

is an increasing function on $[0, 1]$. In summary, we have obtained the following result, which is Qin and Jones's (2022) Theorem 4.1.

COROLLARY 4.3.   *(Qin and Jones, 2022) Assume that $M_{00} = \{0\}$ and $M_R \neq \{0\}$. Then*

$$\rho_D = \|C\|, \quad \rho_R(r) = \frac{1 + \sqrt{(1-2r)^2 + 4r(1-r)\|C\|^2}}{2}.$$

REMARK 4.4.   The formula $\rho_D = \|C\|$ can be viewed as a special case of well-known results concerning the convergence rates of alternating projections (Kayalar and Weinert, 1988; Badea, Grivaux and Müller, 2012).

Corollary 4.3 allows for a comparison between $\rho_R(r)$ and ρ_D. To adjust for computation time, one should raise a convergence rate to the power of the number of iterations that the associated algorithm can simulate in unit time. To be precise, assume as before that one iteration of the DG and RG algorithms take $(1+\tau)/2$ and $r\tau + 1 - r$ units of time respectively, and let

$$\rho_D^\dagger = \rho_D^{2/(1+\tau)}, \quad \rho_R(r)^\dagger = \rho_R(r)^{1/(r\tau+1-r)}.$$

By Corollary 4.3 and Young's inequality, for $r \in (0,1)$,

$$\rho_R(r)^\dagger \geq \|C\|^{2r(1-r)/(r\tau+1-r)} \geq \|C\|^{2/(1+\tau)} = \rho_D^\dagger.$$

Moreover, equality holds only if $\|C\| = 1$. It is in this sense that the DG chain has a faster rate of convergence than the RG chain, after adjusting for computational cost. It is worth mentioning that this does not imply that

$$\|\mu \mathcal{D}_{\lfloor 2t/(1+\tau)\rfloor} - \pi\|_* \leq \|\mu \mathcal{R}_{\lfloor t/(r\tau+1-r)\rfloor} - \pi\|_*$$

for every $\mu \in L_*^2(\pi)$ and t large enough. That is, for certain starting distributions μ, it is still possible for the RG chain to approach π faster than the DG chain.

Finally, consider the modified DG algorithm with ℓ consecutive draws from π_1, as described in Section 3.4. Denote its convergence rate by $\rho_M(\ell)$. For $t = (\ell+1)s + q$, where s is a non-negative integer, and $q \in \{0, \ldots, \ell\}$, the t-step Mtk of the modified DG chain is $\mathcal{M}_{(\ell+1)s+q} = (P_1 P_2)^s$ if $q = 0$, and $\mathcal{M}_{(\ell+1)s+q} = (P_1 P_2)^s P_1$ if $q \in \{1, \ldots, \ell\}$. Using Lemma 2.1, Proposition 4.1, and Lemma 4.2, one can then establish the following.

PROPOSITION 4.5.   *Assume that $M_{00} = \{0\}$ and $M_R \neq \{0\}$. Then*

$$\rho_M(\ell) = \|C\|^{2/(\ell+1)}.$$

When there is no post-processing and parallelization, it takes $(\ell + \tau)/(\ell + 1)$ units of time to simulate one step of the chain. To adjust for computation cost, let $\rho_M^\dagger(\ell) = \rho_M(\ell)^{(\ell+1)/(\ell+\tau)}$. Then we have the following comparison result:

$$\rho_R^\dagger\left(\frac{1}{\ell+1}\right) \geq \|C\|^{2\ell/[(\ell+1)(\ell+\tau)]} \geq \rho_M^\dagger(\ell) = \|C\|^{2/(\ell+\tau)} \geq \|C\|^{2/(1+\tau)} = \rho_D^\dagger.$$

This, along with Corollary 3.13, shows that if $r = 1/(\ell + 1)$, the modified DG algorithm is better than, but comparable to the RG algorithm in terms of adjusted asymptotic variance and convergence rate. The standard DG algorithm is the best out of the three algorithms in terms of convergence rate, but least robust in terms of asymptotic variance.

If post-processing and parallelization is employed, then, in the best case scenario, the modified DG algorithm would have the same convergence rate as the standard DG algorithm. Indeed, with full parallelization, the per-iteration cost of the modified DG algorithm is $(\tau + 1)/(\ell + 1)$. In this case, $\rho_M^\dagger(\ell) = \rho_M(\ell)^{(\ell+1)/(\tau+1)} = \|C\|^{2/(\tau+1)}$, so $\rho_M^\dagger(\ell) = \rho_D^\dagger$.

5. Discussion. The methods herein can be used to analyze other variants of two-component Gibbs samplers. Indeed, let $(\tilde{X}_t)$ be a Markov chain with $(t', t'+t)$-Mtk $K_{t',t'+t}$. Suppose that for non-negative integers t and t', the operator $K_{t',t'+t}$ is in an algebra of finite type generated by P_1, P_2, and the identity operator I. Then $K_{t',t'+t}$ leaves M_{00}, M_{01}, M_{10}, M_{11} and M_R invariant. Moreover, $K_{t',t'+t}|_{M_R}$ has a matrix representation in accordance with Lemma 2.1. In principle, one can analyze the convergence rate and asymptotic variance of the associated MCMC algorithm through elementary matrix calculations. In Appendix D, this idea is applied to a random sequence scan Gibbs sampler that has Mtk $(P_1 P_2 + P_2 P_1)/2$. There are many other interesting variants of two-component Gibbs samplers that can be studied in future works. It'd be interesting to know which variant is optimal in terms of asymptotic variance and/or convergence rate.

In applications of MCMC, one often needs to evaluate $\pi f = \int_X f(x)\pi(\mathrm{d}x)$ for some vector-valued $f : X \to \mathbb{R}^p$, where p is some positive integer (Vats, Flegal and Jones, 2019). Based on a Markov chain $(\tilde{X}_t)$, the sample mean $S_T(f) = T^{-1}\sum_{t=0}^{T-1} f(\tilde{X}_t)$ is then a random vector. Under regularity conditions, the multivariate central limit theorem holds:

$$\sqrt{T}[S_T(f) - \pi f] \xrightarrow{d} N_p(0, \tilde{V}(f)) \quad \text{as } T \to \infty,$$

where $\tilde{V}(f)$ is an asymptotic covariance matrix. If we use $\tilde{V}_D(f)$ and $\tilde{V}_R(f, r)$ to denote the asymptotic covariance matrices associated with the DG and RG samplers, respectively. Then for $a \in \mathbb{R}^p$,

$$V_D(a^\top f) = a^\top \tilde{V}_D(f) a, \quad V_R(a^\top f, r) = a^\top \tilde{V}_R(f, r) a.$$

Thus, Proposition 3.4 provides a Loewner ordering for $V_D(f)$ and $V_R(f)$. That is,

$$\tilde{V}_D(f) \le k_1(r)\tilde{V}_R(f, r), \quad \tilde{V}_R(f, r) \le k_2(r)\tilde{V}_D(f),$$

where k_1 and k_2 are defined in the said proposition.

One future research avenue is to study the efficiency of the RG sampler with $r \ne 1/2$ and the modified DG sampler with $\ell > 1$ when there is a large discrepancy in the variability of X_1 and X_2. For instance, assume that we wish to estimate πf and πg where $f(x_1, x_2)$ depends only on x_1 and $g(x_1, x_2)$ depends only on x_2. If the variance of $f(X_1, X_2)$ is much larger than that of $g(X_1, X_2)$, how much can be gained from more frequent updates of X_1?

A natural question is whether the techniques herein can be used to study Gibbs sampler with more than two components. Gibbs samplers with n components are associated with Markov operators generated by n orthogonal projections. Unfortunately, results like Lemma 2.1 do not extend to the $n \ge 3$ case in general. See Section 11 of Böttcher and Spitkovsky (2010) for a review. Whether it is possible to extend results in this work to the $n \ge 3$ case is a subject for future work. See Greenwood, McKeague and Wefelmeyer (1998), Andrieu (2016), Roberts and Rosenthal (2016), and the recent work Chlebicka, Łatuszyński and Miasojedow (2023) for studies on this topic based on other tools.

APPENDIX A: TECHNICAL RESULTS

LEMMA A.1.   *Let $k_1(r)$ and $k_2(r)$ be defined as in Proposition 3.4, i.e.,*

$$k_1(r) = 1 - r + r^2 + \sqrt{r^2(1-r)^2 + (1-2r)^2},$$

$$k_2(r) = \frac{1 - r + r^2}{2r(1-r)} + \frac{\sqrt{r^2(1-r)^2 + (1-2r)^2}}{2r(1-r)}.$$

Then, for $r \in (0, 1)$, each of the following holds.

$$(i)\ k_1(r) \geq 1; \quad (ii)\ k_1(r) > 2r/(2-r); \quad (iii)\ k_1(r) > 2(1-r)/(1+r);$$
$$(iv)\ k_2(r) \geq 2; \quad (v)\ k_2(r) > (2-r)/(2r); \quad (vi)\ k_2(r) > (1+r)/[2(1-r)].$$

PROOF. It's obvious that

$$k_1(r) \geq 1 - r + r^2 + |r - r^2|,$$

so (i) holds. Moreover,

$$k_1(r) \geq 1 - r + r^2 + |1 - 2r|.$$

Thus,

$$k_1(r) \geq r^2 + r = r(1+r) > \frac{2r}{2-r},$$

and

$$k_1(r) \geq 2 - 3r + r^2 = (1-r)(2-r) > \frac{2(1-r)}{1+r}.$$

This proves (i)-(iii).

The proofs for (iv)-(vi) are similar. □

LEMMA A.2. *Let A be a linear operator on M_R with matrix representation*

$$A = \Gamma^* \begin{pmatrix} dI_0 + aC^2 S^{-2} & bCS^{-1} \\ bCS^{-1} & cI_0 \end{pmatrix} \Gamma, \quad a, b, c, d \in \mathbb{R}.$$

Then A is positive semi-definite if each of the following conditions holds:

(i) $a, c, d \geq 0$;
(ii) $ac - b^2 \geq 0$.

PROOF. If $c = 0$, then (ii) implies that $b = 0$, and the proof is trivial. Assume that $c > 0$. Then $A \geq B$, where $\geq$ denotes Loewner ordering, and

$$B = \Gamma^* \begin{pmatrix} \frac{b^2}{c} C^2 S^{-2} & bCS^{-1} \\ bCS^{-1} & cI_0 \end{pmatrix} \Gamma.$$

Recall that $\Gamma = I_0 \oplus W$ where $W : (I - P_1)M_R \to M_R$ is unitary. For $f_0 \in P_1 M_R$ and $f_1 \in (I - P_1)M_R$,

$$\langle f_0 + f_1, B(f_0 + f_1) \rangle = \frac{b^2}{c} \langle f_0, C^2 S^{-2} f_0 \rangle + b\langle f_0, CS^{-1} W f_1 \rangle + b\langle f_1, W^* CS^{-1} f_0 \rangle + c\|f_1\|^2$$

$$= \left\| \frac{b}{\sqrt{c}} CS^{-1} f_0 + \sqrt{c} W f_1 \right\|^2$$

$$\geq 0.$$

Thus, B is positive semi-definite, and so is A. □

LEMMA A.3. *For $r \in (0,1)$, let Σ_D and $\Sigma_R(r)$ be defined as in (12) and (15), and let $k_1(r)$ and $k_2(r)$ be defined as in Lemma A.1. Then $\Sigma_D \leq k_1(r)\Sigma_R(r)$, and $\Sigma_R(r) \leq k_2(r)\Sigma_D$, where $\leq$ denotes Loewner ordering.*

PROOF. I will establish $\Sigma_D \leq k_1(r)\Sigma_R(r)$. The other inequality can be proved in a similar way.

$$k_1(r)\Sigma_R(r) - \Sigma_D = \Gamma^* \begin{pmatrix} d(r)I_0 + a(r)C^2S^{-2} & b(r)CS^{-1} \\ b(r)CS^{-1} & c(r)I_0 \end{pmatrix} \Gamma,$$

where

$$a(r) = \frac{2}{r(1-r)}k_1(r) - 4,$$

$$b(r) = \frac{2}{1-r}k_1(r) - 2,$$

$$c(r) = \frac{1+r}{1-r}k_1(r) - 2,$$

$$d(r) = \frac{2-r}{r}k_1(r) - 2.$$

By Lemma A.1, $a(r), c(r), d(r) \geq 0$. Straightforward calculations reveal that

$$a(r)c(r) - b(r)^2 = \frac{2}{r(1-r)}k_1(r)^2 - \frac{4(r^2-r+1)}{r(1-r)}k_1(r) + 4 = 0.$$

It follows from Lemma A.2 that $\Sigma_D \leq k_1(r)\Sigma_R(r)$. $\qquad\square$

The following result can be proved in a similar fashion.

LEMMA A.4. *Let Σ_D, $\Sigma_R(r)$, $\Sigma_M(\ell)$ be defined as in* (12), (15), *and* (20). *Then, for $\ell \geq 1$,*

$$\Sigma_M(\ell) \leq \Sigma_R\left(\frac{1}{\ell+1}\right) \leq \left\{\frac{2\ell+1}{\ell+1}, \frac{\ell+1}{\ell}\right\}\Sigma_M(\ell),$$

and

$$\Sigma_M(\ell) \leq \frac{\ell^2 + \ell + 2 + (\ell-1)\sqrt{(\ell+1)^2 + 1}}{2(\ell+1)}\Sigma_D,$$

$$\Sigma_D \leq \frac{\ell^2 + \ell + 2 + (\ell-1)\sqrt{(\ell+1)^2 + 1}}{(\ell+1)^2}\Sigma_M(\ell).$$

In Section 3 it was claimed that the bounds in Proposition 3.4 are sharp. The following proposition makes this precise.

PROPOSITION A.5. *Suppose that $M_{00} = \{0\}$ and $M_R \neq \{0\}$. Let $r \in (0, 1)$ be given, and let $k_1(r)$, $k_2(r)$ be defined as in Lemma A.1. Then, for any $\eta < k_1(r)$, if $\|C\|$ is smaller than but sufficiently close to 1, one can find a non-zero $f \in L_0^2(\pi)$ such that*

$$V_D(f) > \eta V_R(f, r).$$

Moreover, for any $\eta < k_2(r)$, if $\|C\|$ is smaller than but sufficiently close to 1, one can find a non-zero $f \in L_0^2(\pi)$ such that

$$V_R(f, r) > \eta V_D(f).$$

PROOF. I'll prove the first assertion. The proof of the second is similar.

Let $\eta < k_1(r)$ be arbitrary. For $k > 0$, define

$$\alpha(k) = \frac{2}{r(1-r)}k - 4,$$

$$\beta(k) = \frac{2}{1-r}k - 2,$$

$$\gamma(k) = \frac{1+r}{1-r}k - 2.$$

Then

$$\alpha(k)\gamma(k) - \beta(k)^2 = \frac{2}{r(1-r)}k^2 - \frac{4(r^2 - r + 1)}{r(1-r)}k + 4.$$

$k = k_1(r)$ is the larger root of the equation $\alpha(k)\gamma(k) - \beta(k)^2 = 0$, which always has two roots. Moreover, by Lemma A.1, $\gamma(k_1(r)) > 0$. Without loss of generality, assume that η is sufficiently close to $k_1(r)$ so that

$$\alpha(\eta)\gamma(\eta) - \beta(\eta)^2 < 0, \quad \gamma(\eta) > 0.$$

One can show that
(24)
$$\eta\Sigma_{\mathrm{R}}(r) - \Sigma_{\mathrm{D}}$$

$$= \Gamma^* \begin{pmatrix} \frac{\beta(\eta)^2}{\gamma(\eta)}C^2 S^{-2} & \beta(\eta)CS^{-1} \\ \beta(\eta)CS^{-1} & \gamma(\eta)I_0 \end{pmatrix} \Gamma + \Gamma^* \begin{pmatrix} \left(\frac{2-r}{r}\eta - 2\right)I_0 + \left[\alpha(\eta) - \frac{\beta(\eta)^2}{\gamma(\eta)}\right]C^2 S^{-2} & 0 \\ 0 & 0 \end{pmatrix} \Gamma.$$

Basic properties regarding continuous functions of self-adjoint operators imply that

$$\|C^2 S^{-2}\| = \|C^2(I_0 - C^2)^{-1}\| = \|C\|^2/(1 - \|C\|^2).$$

Thus as $\|C\| \to 1$ from below, $\|C^2 S^{-2}\| \to \infty$. Moreover, $C^2 S^{-2}$ is positive semi-definite. When $\|C\|$ is sufficiently close to 1, one can find a non-constant function $f_0 \in P_1 M_{\mathrm{R}}$ such that

$$\left(\frac{2-r}{r}\eta - 2\right)\|f_0\|^2 + \left[\alpha(\eta) - \frac{\beta(\eta)^2}{\gamma(\eta)}\right]\langle f_0, C^2 S^{-2} f_0\rangle < 0.$$

Let $f = f_0 - \beta(\eta)\gamma(\eta)^{-1}W^* CS^{-1} f_0$, where W^* is the lower right block of Γ^*. Then by Propositions 3.1 and 3.2 along with (24),

$$\eta V_{\mathrm{R}}(f, r) - V_{\mathrm{D}}(f) = \langle f, [\eta\Sigma_{\mathrm{R}}(r) - \Sigma_{\mathrm{D}}]f\rangle$$

$$= \left\langle f_0, \left[\frac{2-r}{r}\eta - 2\right]f_0 + \left[\alpha(\eta) - \frac{\beta(\eta)^2}{\gamma(\eta)}\right]C^2 S^{-2} f_0\right\rangle$$

$$< 0.$$

$\square$

APPENDIX B: PROOF OF LEMMA 2.4

Recall that the lemma states that for $r \in (0, 1)$,

$$(1-r)P_1|_{M_{\mathrm{R}}} + rP_2|_{M_{\mathrm{R}}} = \Gamma^* U \begin{pmatrix} [I_0 + \sqrt{\Delta(C)}]/2 & 0 \\ 0 & [I_0 - \sqrt{\Delta(C)}]/2 \end{pmatrix} U^* \Gamma,$$

where $U : M_{\mathrm{R}} \to M_{\mathrm{R}}$ is unitary, and

$$\Delta(C) = (1 - 2r)^2 I_0 + 4r(1-r)C^2.$$

Consider first the case $r = 1/2$. When this is the case,

$$(1-r)P_1|_{M_R} + rP_2|_{M_R} = \Gamma^* \begin{pmatrix} I_0/2 + C^2/2 & CS/2 \\ CS/2 & S^2/2 \end{pmatrix} \Gamma,$$

$\Delta(C) = C^2$, and the result holds with

$$U = \begin{pmatrix} \sqrt{(I_0+C)/2} & \sqrt{(I_0-C)/2} \\ \sqrt{(I_0-C)/2} & -\sqrt{(I_0+C)/2} \end{pmatrix}.$$

Assume that $r \neq 1/2$. For $x \in (0,1)$, define

$$\Delta(x) = (1-2r)^2 + 4r(1-r)x^2.$$

Also, for $x \in (0,1)$, let

$$h_{00}(x) = \frac{[1+\sqrt{\Delta(x)}]x}{\sqrt{4(1-r)[r+\sqrt{\Delta(x)}]x^2 + [2r-1+\sqrt{\Delta(x)}]^2}},$$

$$h_{10}(x) = \frac{[2r-1+\sqrt{\Delta(x)}]\sqrt{1-x^2}}{\sqrt{4(1-r)[r+\sqrt{\Delta(x)}]x^2 + [2r-1+\sqrt{\Delta(x)}]^2}},$$

$$h_{01}(x) = \frac{[1-\sqrt{\Delta(x)}]x}{\sqrt{4(1-r)[r-\sqrt{\Delta(x)}]x^2 + [2r-1-\sqrt{\Delta(x)}]^2}},$$

$$h_{11}(x) = \frac{[2r-1-\sqrt{\Delta(x)}]\sqrt{1-x^2}}{\sqrt{4(1-r)[r-\sqrt{\Delta(x)}]x^2 + [2r-1-\sqrt{\Delta(x)}]^2}}.$$

When $r < 1/2$, let $h_{00}(0) = 1$, $h_{10}(0) = 0$, $h_{01}(0) = 0$, $h_{11}(0) = -1$, and let $h_{00}(1) = 1$, $h_{10}(1) = 0$, $h_{01}(1) = 0$, $h_{11}(1) = -1$. When $r > 1/2$, let $h_{00}(0) = 0$, $h_{10}(0) = 1$, $h_{01}(0) = 1$, $h_{11}(0) = 0$, and let $h_{00}(1) = 1$, $h_{10}(1) = 0$, $h_{01}(1) = 0$, $h_{11}(1) = -1$. Then, for $i \in \{0, 1\}$, h_{ij} is continuous on $[0, 1]$. For $r \in (0, 1)$ and $x \in [0, 1]$, the matrix

$$u(x) = \begin{pmatrix} h_{00}(x) & h_{01}(x) \\ h_{10}(x) & h_{11}(x) \end{pmatrix}$$

is unitary since $u(x)^\top u(x) = I$. Moreover,

$$\begin{pmatrix} 1-r+rx^2 & rx\sqrt{1-x^2} \\ rx\sqrt{1-x^2} & r(1-x^2) \end{pmatrix} u(x) = u(x) \begin{pmatrix} [1+\sqrt{\Delta(x)}]/2 & 0 \\ 0 & [1-\sqrt{\Delta(x)}]/2 \end{pmatrix},$$

so

$$\begin{pmatrix} 1-r+rx^2 & rx\sqrt{1-x^2} \\ rx\sqrt{1-x^2} & r(1-x^2) \end{pmatrix} = u(x) \begin{pmatrix} [1+\sqrt{\Delta(x)}]/2 & 0 \\ 0 & [1-\sqrt{\Delta(x)}]/2 \end{pmatrix} u(x)^\top.$$

The spectrum of C is a subset of $[0, 1]$. Using polynomial approximation (see, e.g., Helmberg, 2014, §32), one can define $h_{ij}(C)$ for $i, j \in \{1, 2\}$. Let

$$U = \begin{pmatrix} h_{00}(C) & h_{01}(C) \\ h_{10}(C) & h_{11}(C) \end{pmatrix}.$$

For any two real continuous functions h_1 and h_2 on $[0, 1]$, it holds that $h_1(C) + h_2(C) = (h_1 + h_2)(C)$, and $h_1(C)h_2(C) = (h_1 h_2)(C)$. By previous calculations, one can then conclude that U is unitary, and that Lemma 2.4 holds.

APPENDIX C: PROOF OF PROPOSITION 4.1

Let $\mu \in L^2_*(\pi)$ be such that $\mu \neq \pi$, and set $g = \mathrm{d}\mu/\mathrm{d}\pi - 1 \in L^2_0(\pi)$. Then, by (22),

$$\limsup_{t \to \infty} t^{-1} \log \|\mu K_t - \pi\|_* = \log \left(\limsup_{t \to \infty} \|g\|^{1/t} \langle g/\|g\|, K_t K_t^* g/\|g\| \rangle^{1/(2t)} \right)$$

$$= \log \left(\limsup_{t \to \infty} \langle g/\|g\|, K_t K_t^* g/\|g\| \rangle^{1/(2t)} \right).$$

It follows that

$$\sup_{\mu \in L^2_*(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu K_t - \pi\|_* \leq \log \left(\sup_{f \in L^2_0(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} \right).$$

To prove the reverse inequality, let $f \in L^2_0(\pi)$ be such that $\|f\| = 1$, and decompose it into positive and negative parts: $f = f_+ - f_-$. Then

$$c := \pi f_+ = \pi f_- > 0.$$

Let μ_+ and μ_- be probability measures such that $\mathrm{d}\mu_+/\mathrm{d}\pi = c^{-1} f_+$ and $\mathrm{d}\mu_-/\mathrm{d}\pi = c^{-1} f_-$. Then $\mu_+, \mu_- \in L^2_*(\pi)$, and neither of the two is equal to π. By (22),

$$\langle f, K_t K_t^* f \rangle^{1/2} = c \|K_t^*[(c^{-1}f_+ - 1) - (c^{-1}f_- - 1)]\|$$

$$\leq c \|K_t^*(c^{-1}f_+ - 1)\| + c \|K_t^*(c^{-1}f_- - 1)\|$$

$$= c \|\mu_+ K_t - \pi\|_* + c \|\mu_- K_t - \pi\|_*.$$

For $a, b > 0$, $\log(a + b) \leq \log 2 + \max\{\log a, \log b\}$. It follows that

$$\log \left(\langle f, K_t K_t^* f \rangle^{1/(2t)} \right) \leq t^{-1} \log(2c) + \max \left\{ t^{-1} \log \|\mu_+ K_t - \pi\|_*,\ t^{-1} \log \|\mu_- K_t - \pi\|_* \right\}$$

For two sequences of real numbers (a_t) and (b_t),

$$\limsup_{t \to \infty} \max\{a_t,\, b_t\} = \max \left\{ \limsup_{t \to \infty} a_t,\ \limsup_{t \to \infty} b_t \right\}.$$

Thus,

$$\log \left(\limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} \right) = \limsup_{t \to \infty} \log \left(\langle f, K_t K_t^* f \rangle^{1/(2t)} \right)$$

$$\leq \sup_{\mu \in L^2_*(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu K_t - \pi\|_*.$$

Taking supremum with respect to f shows that

$$\log \left(\sup_{f \in L^2_0(\pi),\, \|f\|=1} \limsup_{t \to \infty} \langle f, K_t K_t^* f \rangle^{1/(2t)} \right) \leq \sup_{\mu \in L^2_*(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu K_t - \pi\|_*.$$

APPENDIX D: ANOTHER VARIANT OF TWO-COMPONENT GIBBS

One can use the theory of two projections to analyze other variants of two-component Gibbs samplers. As a demonstration, consider the following random sequence scan Gibbs sampler.

Algorithm 4: Random sequence scan Gibbs sampler

Draw $\tilde{X}_0 = (X_{1,0}, X_{2,0})$ from some initial distribution on $(\mathsf{X}, \mathcal{B})$, and set $t = 0$;
while $t < T$ **do**
    draw W from a Bernoulli$(1/2)$ distribution;
    if $W = 0$ **then**
        draw $X_{2,t+1}$ from $\pi_1(\cdot \mid X_{1,t})$, then draw $X_{1,t+1}$ from $\pi_2(\cdot \mid X_{2,t+1})$;
    if $W = 1$ **then**
        Draw $X_{1,t+1}$ from $\pi_2(\cdot \mid X_{2,t})$, then draw $X_{2,t+1}$ from $\pi_1(\cdot \mid X_{1,t+1})$;
    set $\tilde{X}_{t+1} = (X_{1,t+1}, X_{2,t+1})$;
    set $t = t + 1$;

The underlying Markov chain $\tilde{X}_t$ is time-homogeneous, and the associated t-step Mtk is

$$\mathcal{S}_t = (P_1 P_2/2 + P_2 P_1/2)^t.$$

It is clear that $\pi \mathcal{S}_t = \pi$, so the chain has π as a stationary distribution. Assume that $M_{00} = \{0\}$ and that $M_{\mathsf{R}} \neq \{0\}$. Then, for $f \in L_0^2(\pi)$ with orthogonal decomposition $f = \sum_{i=0}^1 \sum_{j=0}^1 f_{ij} + f_0 + f_1$ as in (10),

$$(25) \qquad \mathcal{S}_1 f = (P_1|_{M_{\mathsf{R}}} P_2|_{M_{\mathsf{R}}}/2 + P_2|_{M_{\mathsf{R}}} P_1|_{M_{\mathsf{R}}}/2)(f_0 + f_1).$$

By Lemma 2.1, the following matrix representation holds:

$$(26) \qquad P_1|_{M_{\mathsf{R}}} P_2|_{M_{\mathsf{R}}}/2 + P_2|_{M_{\mathsf{R}}} P_1|_{M_{\mathsf{R}}}/2 = \Gamma^* \begin{pmatrix} C^2 & CS/2 \\ CS/2 & 0 \end{pmatrix} \Gamma.$$

For $f \in L_0^2(\pi)$ defined above, the associated asymptotic variance is

$$V_{\mathsf{S}}(f) = E[f(\tilde{X}_0)^2] + 2 \sum_{t=1}^{\infty} E[f(\tilde{X}_0) f(\tilde{X}_t)],$$

where $\tilde{X}_0 \sim \pi$ (see, e.g., Jones, 2004).

PROPOSITION D.1. *Assume that $\|C\| < 1$. Then*

$$V_S(f) = \|f_{01}\|^2 + \|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_S(f_0 + f_1) \rangle,$$

where

$$\Sigma_S = \Gamma^* A_S B_S \Gamma,$$

$$A_S = \begin{pmatrix} (I_0 - C^2/4)^{-1} & 0 \\ 0 & (I_0 - C^2/4)^{-1} \end{pmatrix}, \quad B_S = \begin{pmatrix} I_0 + 2C^2 S^{-2} + C^2/4 & CS^{-1} \\ CS^{-1} & I_0 + C^2/4 \end{pmatrix}.$$

PROOF. By (25),

$$(27) \qquad \begin{aligned} V_{\mathsf{S}}(f) &= \|f\|^2 + 2 \sum_{t=1}^{\infty} \langle f, (P_1 P_2/2 + P_2 P_1/2)^t f \rangle \\ &= \|f_{01}\|^2 + \|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma(f_0 + f_1) \rangle, \end{aligned}$$

where

$$\Sigma = I + 2\sum_{t=1}^{\infty} \left(P_1|_{M_R} P_2|_{M_R}/2 + P_2|_{M_R} P_1|_{M_R}/2 \right)^t.$$

By the matrix representation,

$$\Sigma = I + 2\Gamma^* \sum_{t=1}^{\infty} \left(\begin{array}{cc} C^2 & CS/2 \\ CS/2 & 0 \end{array} \right)^t \Gamma$$

(28)
$$= 2\Gamma^* \left(\begin{array}{cc} S^2 & -CS/2 \\ -CS/2 & I_0 \end{array} \right)^{-1} \Gamma - I$$

$$= \Sigma_{\mathrm{S}}.$$

The desired result then follows from (27). □

The following lemma holds for Σ_{S}.

LEMMA D.2. *Let Σ_D and Σ_S be as in (12) and (28), respectively. Then*

$$\Sigma_D \le 2\Sigma_S.$$

PROOF. Recall that

$$\Sigma_{\mathrm{S}} = \Gamma^* A_{\mathrm{S}} B_{\mathrm{S}} \Gamma,$$

where A_{S} and B_{S} are given in Proposition D.1. Since $0 \le \|C\| < 1$, $A_{\mathrm{S}} \ge I$. Moreover, $(A_{\mathrm{S}} - I)^{1/2}$ commutes with B_{S}. It is straightforward to verify that B_{S} is positive semi-definite. Then

$$\Sigma_{\mathrm{S}} - \Gamma^* B_{\mathrm{S}} \Gamma = \Gamma^* (A_{\mathrm{S}} - I) B_{\mathrm{S}} \Gamma = \Gamma^* (A_{\mathrm{S}} - I)^{1/2} B_{\mathrm{S}} (A_{\mathrm{S}} - I)^{1/2} \Gamma.$$

Hence, $\Sigma_{\mathrm{S}} \ge \Gamma^* B_{\mathrm{S}} \Gamma$. On the other hand,

$$\Gamma^* B_{\mathrm{S}} \Gamma \ge \Gamma^* \left(\begin{array}{cc} I_0 + 2C^2 S^{-2} & CS^{-1} \\ CS^{-1} & I_0 \end{array} \right) \Gamma = \frac{1}{2} \Sigma_{\mathrm{D}}.$$

The desired result then follows. □

Recall from Proposition 3.1 that the asymptotic variance for the DG sampler is

$$V_{\mathrm{D}}(f) = 2\|f_{01}\|^2 + 2\|f_{10}\|^2 + \|f_{11}\|^2 + \langle f_0 + f_1, \Sigma_{\mathrm{D}}(f_0 + f_1) \rangle.$$

It then follows from Proposition D.1 and Lemma D.2 that

$$V_{\mathrm{D}}(f) \le 2V_{\mathrm{S}}(f).$$

On average, one iteration of the random sequence scan algorithm takes twice the time one iteration of the DG algorithm takes. Let $V_{\mathrm{S}}^{\dagger}(f)$ be the computation time adjusted asymptotic variance of the random sequence scan algorithm, i.e., $V_{\mathrm{S}}^{\dagger}(f) = (\tau + 1)V_{\mathrm{S}}(f)$. Then

$$V_{\mathrm{D}}^{\dagger}(f) \le V_{\mathrm{S}}^{\dagger}(f).$$

Thus, the random sequence scan sampler is no more efficient than the DG sampler in terms of adjusted asymptotic variance (and thus could be much worse than the RG sampler).

Consider now the convergence rate of the chain. Define

$$\rho_S = \exp\left(\sup_{\mu \in L^2_*(\pi),\, \mu \neq \pi} \limsup_{t \to \infty} t^{-1} \log \|\mu \mathcal{S}_t - \pi\|_*\right).$$

By Proposition 4.1 and Lemma 4.2,

$$\rho_S = \|P_1 P_2/2 + P_2 P_1/2\| = \left\|P_1|_{M_R} P_2|_{M_R}/2 + P_2|_{M_R} P_1|_{M_R}/2\right\|,$$

where the second equality follows from (25). Consider the matrix representation (26) and note that

$$\begin{pmatrix} C^2 & CS/2 \\ CS/2 & 0 \end{pmatrix} = U_1 \begin{pmatrix} (I_0 + C)C/2 & 0 \\ 0 & -C(I_0 - C)/2 \end{pmatrix} U_1^*,$$

where

$$U_1 = \begin{pmatrix} \sqrt{(I_0 + C)/2} & \sqrt{(I_0 - C)/2} \\ \sqrt{(I_0 - C)/2} & -\sqrt{(I_0 + C)/2} \end{pmatrix}$$

is unitary and $U_1^* = U_1$ is its adjoint. Then

$$\rho_S = \left\|P_1|_{M_R} P_2|_{M_R}/2 + P_2|_{M_R} P_1|_{M_R}/2\right\| = (1 + \|C\|)\|C\|/2.$$

After adjusting for computation time, the convergence rate is

$$\rho_S^\dagger := \rho_S^{1/(1+\tau)} = [(1 + \|C\|)\|C\|/2]^{1/(1+\tau)}.$$

It is obvious that $\rho_S^\dagger \geq \rho_D^\dagger$, where $\rho_D^\dagger = \|C\|^{2/(1+\tau)}$ is given in Section 4. Depending on the value of $\|C\|$, r, and τ, $\rho_S^\dagger$ can be larger or smaller than $\rho_R^\dagger(r)$.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679.

AMIT, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *Journal of Multivariate Analysis* **38** 82–99.

ANDRIEU, C. (2016). On random-and systematic-scan samplers. *Biometrika* **103** 719–726.

BADEA, C., GRIVAUX, S. and MÜLLER, V. (2012). The rate of convergence in the method of alternating projections. *St. Petersburg Mathematical Journal* **23** 413–434.

BÖTTCHER, A. and SPITKOVSKY, I. M. (2010). A gentle guide to the basics of two projections theory. *Linear Algebra and its Applications* **432** 1412–1459.

BÖTTCHER, A. and SPITKOVSKY, I. M. (2018). Robert Sheckley's answerer for two orthogonal projections. In *The Diversity and Beauty of Applied Operator Theory* 125–138. Springer.

CASELLA, G. and GEORGE, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46** 167–174.

CHLEBICKA, I., ŁATUSZYŃSKI, K. and MIASOJEDOW, B. (2023). Solidarity of Gibbs Samplers: the spectral gap. arXiv preprint.

DIACONIS, P., KHARE, K. and SALOFF-COSTE, L. (2010). Stochastic alternating projections. *Illinois Journal of Mathematics* **54** 963–979.

DOBRUSHIN, R. L. (1956). Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications* **1** 65–80.

FILL, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Annals of Applied Probability* 62–87.

GELFAND, A. E. and SMITH, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* **85** 398–409.

GREENWOOD, P. E., McKEAGUE, I. W. and WEFELMEYER, W. (1998). Information bounds for Gibbs samplers. *Annals of Statistics* **26** 2128–2156.

HALMOS, P. R. (1969). Two subspaces. *Transactions of the American Mathematical Society* **144** 381–389.

HELMBERG, G. (2014). *Introduction to Spectral Theory in Hilbert Space*. Elsevier.

JONES, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys* **1** 299–320.

JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–334.

KAYALAR, S. and WEINERT, H. L. (1988). Error bounds for the method of alternating projections. *Mathematics of Control, Signals and Systems* **1** 43–59.

KONTOROVICH, L. and RAMANAN, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *Annals of Applied Probability* **36** 2126–2158.

KUBRUSLY, C. S. (2012). *Spectral theory of operators on Hilbert spaces*. Springer Science & Business Media.

LINK, W. A. and EATON, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution* **3** 112–115.

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes. *Biometrika* **81** 27–40.

MACEACHERN, S. N. and BERLINER, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician* **48** 188–190.

MAIRE, F., DOUC, R. and OLSSON, J. (2014). Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Annals of Statistics* **42** 1483–1510.

NISHIO, K. (1985). The structure of a real linear combination of two projections. *Linear Algebra and Its Applications* **66** 169–176.

PAULIN, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability* **20** 1–32.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108** 1339–1349.

QIN, Q. and HOBERT, J. P. (2020). On the limitations of single-step drift and minorization in Markov chain convergence analysis. *Annals of Applied Probability* **31** 1633–1659.

QIN, Q. and JONES, G. L. (2022). Convergence rates of two-component MCMC samplers. *Bernoulli* **28** 859–885.

ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2** 13–25.

ROBERTS, G. O. and ROSENTHAL, J. S. (2006). Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Annals of Applied Probability* **16** 2123–2139.

ROBERTS, G. O. and ROSENTHAL, J. S. (2016). Surprising convergence properties of some simple Gibbs samplers under various scans. *International Journal of Statistics and Probability* **5** 51–60.

ROBERTS, G. O. and TWEEDIE, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability* **38** 37–41.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82** 528–540.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics* **22** 1701–1728.

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10** 1-50.

VATS, D., FLEGAL, J. M. and JONES, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* **106** 321–337.