STOCHASTIC MODELS
https://doi.org/10.1080/15326349.2026.2688991





# Estimating MCMC convergence rates using common random number simulations

Sabrina Sixta[a], Jeffrey S. Rosenthal[a], and Austin Brown[b]

[a]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada; [b]Department of Statistics, Texas A&M University, College Station, Texas, USA

**ABSTRACT**

This paper presents how to use common random number (CRN) simulation to evaluate Markov chain Monte Carlo (MCMC) convergence to stationarity. We provide an upper bound on the Wasserstein distance of a Markov chain to its stationary distribution after $N$ steps in terms of averages over CRN simulations. We apply our bound to Gibbs samplers on a model related to James-Stein estimators, a variance component model, and a Bayesian linear regression model. Using our examples, we show that the CRN-based simulation combined with a coalescing condition to generate a total variation bound converges to zero much more faster than the available drift and minorization bounds, while also converging at the same rate as the one-shot coupling bound.

**ARTICLE HISTORY**
Received 27 June 2024
Accepted 10 June 2026

**KEYWORDS**
Common random number; synchronous coupling; Markov chains; MCMC; convergence rate; Gibbs sampler; drift and minorization

## 1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are often used to simulate from a stationary distribution of interest (see e.g.[20]). One of the primary questions when using these Markov chains is, after how many iterations is the distribution of the Markov chain sufficiently close to the stationary distribution of interest, i.e. when should actual sampling begin[25]. The number of iterations it takes for the distribution of the Markov chain to be sufficiently close to stationarity is called the burn-in period. Various informal methods are available for estimating the burn-in period, such as effective sample size estimation, the Gelman-Rubin diagnostic, and visual checks using traceplots or autocorrelation graphs[17,27,43,52]. While these methods "can detect problems with an MCMC simulation, [. . .] they cannot prove that the simulation is generating a representative sample"[39].

Convergence analysis studies the distance of a Markov chain to stationarity and is traditionally measured in terms of total variation distance (e.g.[46,57]), though more recently the Wasserstein distance has been considered[21,30,37,42]. However, finding upper bounds on either distance can be quite difficult to establish[20,25,39], and if an upper bound is known, it is usually based on complicated problem-specific calculations[30,41,54,55]. This motivates the desire

**CONTACT** Sabrina Sixta sabrina.sixta@mail.utoronto.ca Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada.

to instead estimate convergence bounds from actual simulations of the Markov chain combined with less complicated calculations, which we consider here *via* common random number (CRN) simulation.

CRN simulation occurs when we initiate two copies of the chain with different values, but use the same sequence of random variables to simulate a specific random function representation of the chain (see Section 2 for a formal definition). This simulation method is also sometimes referred to as synchronous coupling[4,10,29] and falls under the general framework of "auxiliary simulation"[12], i.e. using extra preliminary Markov chain runs to estimate the convergence time needed in the final run. Estimating Markov chain convergence rates using CRN simulation was first proposed in[32] to find estimates of mixing times in total variation distance. Further, one of the first Wasserstein bounds generated on a Markov chain model used CRN in its construction[21]. More recently,[7] showed how CRN simulation could be used for estimating an upper bound on the Wasserstein distance (their Proposition 3.1), and provided useful applications of the CRN method to high-dimensional and tall data (their Section 4). CRN simulation to estimate convergence rates has been used on a wide range of applications including Langevin algorithms[18], stochastic gradient descent[35], stochastic differential equations (SDE)[4], and Hamiltonian Monte Carlo[8,23].

The use of CRN to generate tight bounds on the Wasserstein and total variation distance is promising. In this paper we show that for the Gibbs sampler for a model related to James-Stein estimators and the variance component model, CRN bounds perform significantly better than drift and minorization (DnM) bounds (see Figures 2 and 5 and commentary in Numerical Example 5.1 for more details). Under certain conditions,[4] and[22] show that CRN is the optimal coupling of the Wasserstein distance between simulated solutions of SDEs and two random variables in the $\mathcal{L}^2$ metric, respectively.[16] shows that CRN simulation on the independent Metropolis-Hastings algorithm is the maximal coupling of the total variation distance.[10] shows that the distance between two reflected Brownian motion processes simulated using CRN converge to 0 almost surely.

In Theorem 3.2, using CRN simulation, we provide an estimated upper bound between the Wasserstein distance of a Markov chain and the corresponding stationary distribution when only the unnormalized density of the stationary distribution is known. Furthermore the 95% confidence intervals of our estimates in the real data examples are fairly narrow. See Lemma 3.3 and its implementation in the Figures 1, 4, 5 and 6 and Tables 5 and 7 for more details.

This paper is organized as follows. In Section 2, we present definitions and notation. In Section 3, we establish convergence bounds of a Markov chain to its corresponding stationary distribution using CRN simulation when the initial distribution is not in stationarity. In Section 4, we apply Theorem 3.2 to a Gibbs

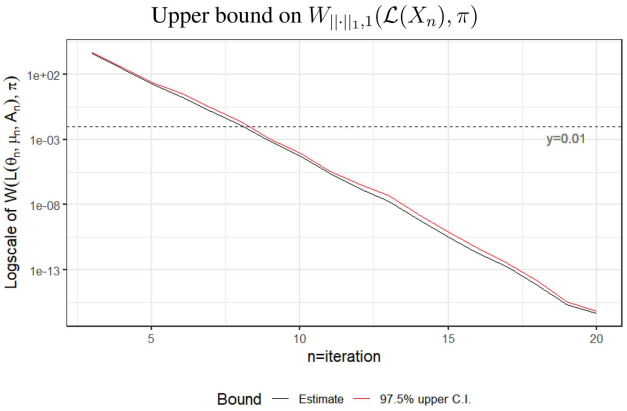


**Figure 1.** The upper bound on the Wasserstein distance for the Gibbs sampler model related to James-Stein estimators. The estimate uses the mean of 1000 simulations and initial value $(\vec{\theta}_0, \mu_0, A_0) = (100)^{q+2}$. The vertical axis is log scaled. $r = 1$.

sampler for a model related to the James-Stein estimator and compare the CRN bound to a DnM and one-shot coupling bound in Real Data Example 4.1. In Section 5 we apply Theorem 3.2 to a Gibbs sampler for a variance component model and compare the CRN bound to a DnM bound in Numerical Example 5.1 and another DnM bound in Real Data Example 5.1. In Section 6 we apply Theorem 3.2 to a Gibbs sampler for a Bayesian linear regression model.

For all of the cases where we compare the CRN bound, which is measured in Wasserstein distance, to the DnM and one-shot coupling bound, which are measured in total variation distance, we must state a coalescing condition. The coalescing condition bounds the total variation distance with respect to the Wasserstein distance and is a part of a one-shot coupling bound[54]. Further, since the one-shot coupling bound uses a contraction condition, which is established using CRN, we expect the empirical CRN simulation bound and the theoretical one-shot coupling bound to approach zero at the same rate (see Figure 2 as an example).

For each numerical example, we also provide the Gelman-Rubin diagnostic and traceplots for comparison (see Section 8.1). The Gelman-Rubin statistic assesses convergence by comparing within-chain and between-chain variability across chains initialized from dispersed starting points. It is not immediately clear under what conditions the Gelman-Rubin statistic will produce a more conservative cutoff than the CRN bound. In three of the four numerical and real-data examples, the Gelman–Rubin statistic suggests an earlier cutoff than the CRN-based criterion. Also, since the initial distributions may vary (Gelman-Rubin statistic requires an overdispersed initial distribution), we are cautious in comparing the two methods. Trace plots provide a visual diagnostic for assessing convergence by illustrating whether chains initialized at different starting values

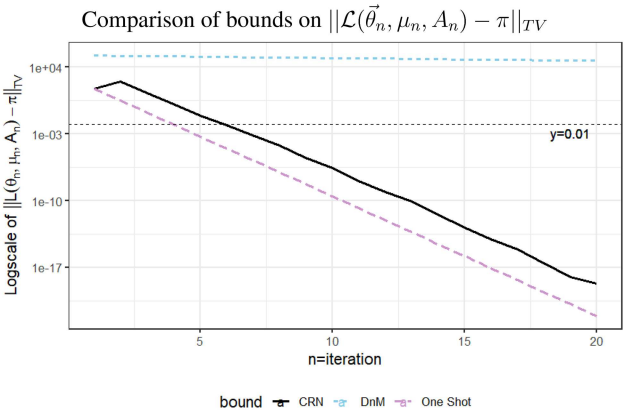


**Figure 2.** Comparison of an upper bound on the total variation distance on the Gibbs sampler for a model related to James-Stein estimators (Real Data Example 4.1) using a CRN simulated bound, a DnM bound and a one-shot coupling bound. The CRN estimate uses the mean of 1000 simulations. For all of the bounds, $(\vec{\theta}_0, \mu_0, A_0) = (100)^{q+2}$.

have converged to the same stationary distribution. The traceplots in our examples use the same initial distributions as our numerical examples. In all of the numerical and real-data examples, the traceplots indicate mixture sooner than the Gelman-Rubin statistic and the CRN bound.

The code used to generate all of the tables and calculations can be found at github.com/sixter/CommonRandomNumber/.

## 2. Background and notation

### 2.1. Metrics on distributions

Let $(\mathcal{X}, d)$ be a Polish metric space and $\mathcal{F}$ the associated Borel $\sigma$-algebra. The $L^p-$ Wasserstein distance between distributions $\pi$ and $\nu$ on $(\mathcal{X}, \mathcal{F})$ is defined as $W_{d,p}(\pi, \nu) = \inf_{X\sim\pi, Y\sim\nu} E[d(X, Y)^p]^{1/p}$. The Wasserstein space $P_p(\mathcal{X})$ is the set of Borel probability measures on the metric space $(\mathcal{X}, d)$ such that for any $\mu \in P_p(\mathcal{X})$, $\int_{\mathcal{X}} d(x_0, x)^p \mu(\mathrm{d}x) < \infty$ for some $x_0 \in \mathcal{X}$. Total variation distance is defined as $||\pi - \nu||_{TV} = \inf_{X\sim\pi, Y\sim\nu} P(X \neq Y) = \sup_{A\in\mathcal{F}} |\pi(A) - \nu(A)|$. The total variation distance can be written as a special case of the Wasserstein distance, $||\pi - \nu||_{TV} = W_{d_I,1}(\pi, \nu)$ where $d_I(X, Y) = I_{X\neq Y}$. If $\mathcal{X} \subset \mathbb{R}^q$, $q \in \mathbb{N}$, define the norm over a vector as $||x||_p = (\sum_{i=1}^q |x_i|^p)^{1/p}$ where $p \in [1, \infty)$. Define the p-norm over a function $\phi : \mathcal{X} \to \mathbb{R}$ as $\|\phi\|_p = \left(\int_{\mathcal{X}} \phi(x)^p \mathrm{d}x\right)^{\frac{1}{p}}$. If $X \perp\!\!\!\perp Y$, then $X$ and $Y$ are independent. If $\pi \ll \nu$, then $\pi$ is absolutely continuous with respect to $\nu$.

### 2.2. CRN and Markov chains

Define a Markov chain $\{X_n\}_{n\geq 1}$ in $\mathcal{X}$ such that $X_n = k_{\theta_n}(X_{n-1})$, where $\{\theta_n\}_{n\geq 1}$ are i.i.d. random variables on some measurable space $\Theta$ and random

measurable mappings $k_{\theta_n} : \mathcal{X} \mapsto \mathcal{X}$. The set of random functions $k_{\theta_1}, k_{\theta_2}, \ldots$ is called an iterated function system. Any time-homogeneous Markov chain can be represented as an iterated function system[62]. We can also write $X_n = k_{\theta_n}(k_{\theta_{n-1}}(k_{\theta_{n-2}} \ldots k_{\theta_1}(x_0))) = k_{\Theta_n}(x_0)$ where $\Theta_n = (\theta_1, \ldots, \theta_n)$ and $\theta_i \perp\!\!\!\perp \theta_j$, $i \neq j$. CRN simulation in this context occurs when we simulate two Markov chains $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ with different initial values $x_0 \neq y_0$, but with the same transition functions. That is, $\theta_n$ are the same in $X_n = k_{\theta_n}(X_{n-1})$ and $Y_n = k_{\theta_n}(Y_{n-1})$, $n \geq 1$. In different notation, the $\Theta_n$ is the same in $X_n = k_{\Theta_n}(x_0)$ and $Y_n = k_{\Theta_n}(y_0)$. See Algorithm 1 for more details. The stationary distribution of $X_n$ is defined as $\mathcal{L}(X_\infty) = \pi$.

### 2.3. Distribution functions

$IG(\alpha, \beta)$ represents the inverse gamma distribution with density function $f_{IG}(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$. $N(\mu, \sigma^2)$ represents the normal distribution with density function $f_N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. For a distribution $\nu$, the density function is denoted as $f_\nu$.

## 3. Estimating Markov chain convergence rates using CRN simulation

We propose bounding the $L^p-$ Wasserstein distance between the $n$th iteration of a Markov chain $\{X_n\}_{n\geq 1}$ and the corresponding stationary distribution $\pi$ through CRN simulation. Our main result is Theorem 3.2. Using CRN simulation as a convergence diagnostic tool was first discussed in[32] for bounding total variation distance.

We first provide a method for bounding the Wasserstein distance between $\mathcal{L}(X_n)$ and the corresponding stationary distribution $\pi$, $W_{d,p}(\mathcal{L}(X_n), \pi)$ by proposing an auxiliary random variable $Y_0 \sim \nu$. That is $W_{d,p}(\mathcal{L}(X_n), \pi)$ is bounded above by the product of the 'distance' between $\pi$ and $\mathcal{L}(Y_0)$ ($K_s(\pi, \nu)$, $s > 1$) and $E[d(X_n, Y_n)]$.

**Theorem 3.1.** *Let $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ be two copies of a Markov chain with stationary distribution $\pi$ and initial distributions $\mathcal{L}(Y_0) = \nu$, $\mathcal{L}(X_0) = \mu$ where $X_0 \perp\!\!\!\perp Y_0$. Assume $\pi \ll \nu$. Then for $r \geq p \geq 1$ and $s > 1$, the following holds where* $K_s(\pi, \nu) = E_{Y_0\sim\nu}\left[\left(\frac{f_\pi(Y_0)}{f_\nu(Y_0)}\right)^{\frac{s}{s-1}}\right]^{\frac{s-1}{s}}$

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq K_s(\pi, \nu)^{1/r} E\left[d(X_n, Y_n)^{rs}\right]^{\frac{1}{rs}} \tag{1}$$

*See Section 8.2 for a proof.*

**Remark 1.** If we take $K_1(\pi, \nu) = \lim_{s \downarrow 1} E_{Y_0 \sim \nu} \left[ \left( \frac{f_\pi(Y_0)}{f_\nu(Y_0)} \right)^{\frac{s}{s-1}} \right]^{\frac{s-1}{s}}$, then from Theorem 3.1 we get that $K_1(\pi, \nu) = ess\sup_{x \in \mathcal{X}} f_\pi(x)/f_\nu(x)$ and

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq K_1(\pi, \nu)^{1/r} E\left[d(X_n, Y_n)^r\right]^{1/r}$$

The case of $s \downarrow 1$ can be proven using the rejection sampler or the separation distance (see the proof of this special case in Section 8.3.1). The use of rejection sampling to generate an upper bound on the expected distance between a proposal density function $\nu$ and a stationary density function $\pi$ was motivated by[32], which used rejection sampling to generate similar upper bounds in total variation distance. This case is used in Lemmas 4.1 and 6.2

**Remark 2.** Applying $d_I(X, Y) = I_{X \neq Y}$ to Remark 1, and setting $p = r = 1$, the total variation distance between a Markov chain and the corresponding stationary distribution is bounded above as follows,

$$||\mathcal{L}(X_n) - \pi||_{TV} \leq K_1(\pi, \nu) P(X_n \neq Y_n) \tag{2}$$

And so, for stopping time $\tau = \min\{n : d(X_n, Y_n) = 0\}$, $||\mathcal{L}(X_n) - \pi||_{TV} \leq K_1(\pi, \nu) P(\tau > n)$.

The following are additional observations regarding $K_s(\pi, \nu)$.

1. $K_s(\pi, \nu)$ is related to the $f$-divergence between the initialized and target distribution. The $f$-divergence suffers from the curse of dimensionality[36,53] and so $K_s(\pi, \nu)$ may become very large as dimension increases.
2. In many cases, the normalizing constant for $f_\pi$ is unknown. That is, we only know of a function $g(x)$ such that $\frac{1}{L} g(x) = f_\pi(x)$ where $L$ is unknown. For cases like this, we note that for all $x \in \mathcal{X}$ and any $B \subset \mathcal{X}$,

$$f_\pi(x) = \frac{g(x)}{\int_{\mathcal{X}} g(x) \mathrm{d}x} \leq \frac{g(x)}{\int_B g(x) \mathrm{d}x} \tag{3}$$

   This upper bound on $f_\pi$ is used to upper bound $K$ in all of the examples, which we calculate analytically.
3. The quantity $K_s(\pi, \nu)$ can also be estimated using importance sampling, since $\left( \frac{f_\pi(Y_0)}{f_\nu(Y_0)} \right)^{\frac{s}{s-1}}$ typically becomes large on low-probability events under $\nu$. As $K_2(\pi, \nu)$ increases, the more computationally expensive it will be to calculate[2].

Next we define Algorithm 1, which generates an estimate of $E[d(X_n, Y_n)^{rs}]$ using CRN simulation.

**Algorithm 1** An estimate using CRN of $E[d(X_N, Y_N)^{rs}] \approx \frac{1}{M}\sum_{i=1}^{M} d(x_{N,i}, y_{N,i})^{rs}$

**for** $i = 1, \ldots, M$ **do**
  $x_{0,i} \sim \mu$, $y_{0,i} \sim \nu$ where $x_{0,i} \perp\!\!\!\perp y_{0,i}$
  **for** $n = 1, \ldots, N$ **do**
    $\theta_n \sim \mathcal{L}(\Theta)$
    $x_{n,i} \leftarrow k_{\theta_n}(x_{n-1,i})$
    $y_{n,i} \leftarrow k_{\theta_n}(y_{n-1,i})$
  **end for**
**end for**
**return** $\frac{1}{M}\sum_{i=1}^{M} d(x_{N,i} - y_{N,i})^{rs}$

Combining Algorithm 1 with Theorem 3.1, we can simulate an upper bound on the Wasserstein distance between the law of a Markov chain at iteration $N$, $\mathcal{L}(X_N)$, and the corresponding stationary distribution, $\pi$, as follows.

**Theorem 3.2.** *Let $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ be two copies of a Markov chain with stationary distribution $\pi$, initial distributions $\mathcal{L}(Y_0) = \nu$, $\mathcal{L}(X_0) = \mu$ ($X_0 \perp\!\!\!\perp Y_0$) and simulated according to Algorithm 1. Assume $\pi \ll \nu$. Then for $r \geq p \geq 1$ and $s \geq 1$, the following holds almost surely for $N \geq 1$, where $K_s(\pi, \nu) = E\left[\left(\frac{f_\pi(Y_0)}{f_\nu(Y_0)}\right)^{\frac{s}{s-1}}\right]^{\frac{s-1}{s}}$ and $\mathcal{L}(X_N), \mathcal{L}(Y_N) \in P_{rs}(\mathcal{X})$.*

$$W_{d,p}(\mathcal{L}(X_N), \pi) \leq K_s(\pi, \nu)^{1/r} \left(\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} d(X_{N,i}, Y_{N,i})^{rs}\right)^{\frac{1}{rs}} \tag{4}$$

*Proof.* Since $\mathcal{L}(X_N), \mathcal{L}(Y_N) \in P_{rs}(\mathcal{X})$, $E[d(X_N, Y_N)^{rs}] < \infty$ and so by the strong law of large numbers, $\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} d(X_{N,i}, Y_{N,i})^{rs} \overset{a.s.}{=} E[d(X_N, Y_N)^{rs}]$. By Theorem 3.1, $W_{d,p}(\mathcal{L}(X_N), \pi) \leq K_s(\pi, \nu)^{1/r} E[d(X_N, Y_N)^{rs}]^{\frac{1}{rs}} \overset{a.s.}{=} K_s(\pi, \nu)^{1/r} \left(\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} d(X_{N,i}, Y_{N,i})^{rs}\right)^{\frac{1}{rs}}$ □

To optimize this bound we can choose $\nu$ that is close to $\pi$. If $K_s(\pi, \nu)$ is very large (which may happen in high dimension), we can choose an $r \geq p$ such that $K_s(\pi, \nu)^{1/r}$ is a reasonable value. However, choosing a larger $r$ increases the value of $E[d(X_n, Y_n)^{rs}]^{\frac{1}{rs}}$, resulting in a tradeoff between the two effects. Nevertheless, when the CRN simulation produces expected values close to those of an optimal coupling, the distance $d(X_N, Y_N)$ is often small, and the resulting CRN-based bound may still perform well empirically even in high-dimensional settings with large r.

Next we state the confidence interval on our bound. We use this confidence interval in Figures 1, 4, 5 and 6 and Table 5.

**Lemma 3.3.** *The* $1-\alpha$ *confidence interval on the Wasserstein upper bound, based on a simulation estimate of* $E[d(X_N, Y_N)^{rs}]^{\frac{1}{rs}}$ *is given by,*

$$\left[K_s(\pi,\nu)^{\frac{1}{r}}\left(m - z_{\alpha/2}\sigma_m/\sqrt{M}\right)^{\frac{1}{rs}}, K_s(\pi,\nu)^{\frac{1}{r}}\left(m + z_{\alpha/2}\sigma_m/\sqrt{M}\right)^{\frac{1}{rs}}\right]$$

*Where* $m = \frac{1}{M}\sum_{i=1}^{M} d(X_{N,i}, Y_{N,i})^{rs}$, $\sigma_m^2$ *is the sample variance of* $m$ *and* $\mathcal{L}(X_N), \mathcal{L}(Y_N) \in P_{2rs}(\mathcal{X})$.

*Proof.* Since $\mathcal{L}(X_N), \mathcal{L}(Y_N) \in P_{2rs}(\mathcal{X})$, the variances are finite and we can apply the Central Limit Theorem. That is, $P\left(m - \frac{z_{\alpha/2}\sigma_m}{\sqrt{M}} \leq E[d(X_n, Y_n)^{rs}] \leq m + \frac{z_{\alpha/2}\sigma_m}{\sqrt{M}}\right) \approx 1-\alpha$. And so, $P(K_s(\pi,\nu)^{\frac{1}{r}}\left(m - \frac{z_{\alpha/2}\sigma_m}{\sqrt{M}}\right)^{\frac{1}{rs}} \leq K_s(\pi,\nu)^{\frac{1}{r}} E[d(X_n, Y_n)^{rs}]^{\frac{1}{rs}} \leq K_s(\pi,\nu)^{\frac{1}{r}}\left(m + \frac{z_{\alpha/2}\sigma_m}{\sqrt{M}}\right)^{\frac{1}{rs}}) \approx 1-\alpha$. □

Note that since we are applying the Central Limit Theorem, this is an asymptotic confidence interval with respect to $M \to \infty$.

The approach of Biswas and Mackey[7]

Algorithm 1 closely resembles Algorithm 1 of[7]. Biswas and Mackey show in[7] how CRN simulation can be used to estimate the CUB estimator, which is defined as follows for a given $p \in [1, \infty)$,

$$\mathrm{CUB}_{M,N,T} = \left(\frac{1}{M(T-N)}\sum_{i=1}^{M}\sum_{n=N+1}^{T} d(X_{n,i}, Y_{n,i})^p\right)^{1/p} \tag{5}$$

By Corollary 3.2 of[7], under certain regularity conditions, the CUB estimator bounds the Wasserstein distance between the sample means,

$$W_{d,p}\left(\mathcal{L}\left(\frac{1}{(T-N)}\sum_{n=N+1}^{T} X_{n,i}\right), \mathcal{L}\left(\frac{1}{(T-N)}\sum_{n=N+1}^{T} Y_{n,i}\right)\right) \leq \lim_{M\to\infty} \mathrm{CUB}_{M,N,T}$$

If we set $N, T-N \to \infty$ we obtain a bound on the stationary distributions,

$$W_{d,p}(\mathcal{L}(X_\infty), \mathcal{L}(Y_\infty)) \leq \lim_{M,N\to\infty, T-N\to\infty} \mathrm{CUB}_{M,N,T}$$

In[7], the CUB estimator eventually bounds the Wasserstein distance between the corresponding stationary distributions, but does not tell us precisely when it will happen or by how much. That is, if the starting points of the two Markov chains are close in distance but are far from stationarity, the results from[7] would conclude, for some fixed iteration $N$, that the distance between the two chains is close (true), but it would be incorrect to state that the distance to stationarity is close. In comparison, Theorem 3.2 measures distance to stationarity for any initial distribution on $X_0$ and iteration $N \in \mathbb{N}$.

The approach of Johnson[32]

The motivation of Theorem 3.2 comes from[32], which provides a bound in total variation distance between a Markov chain and its corresponding stationary distribution using CRN simulation. The following is a simplification of the results with altered notation to match this paper. For complete details, refer to[32]. For a discussion on when CRN is applicable within the setting of[32], see[13].

Let $\tau$ be a 'stopping' time, $\tau = \min\{n : d(X_n, Y_n) \leq \epsilon\}$. Theorem 2 of[32] states that the total variation between two Markov chains $\{X_n\}_{n\geq 1}, \{Y_n\}_{n\geq 1}$ both with initial distribution $\nu$ and stationary distribution $\pi$ is as follows,

$$||\pi - \mathcal{L}(X_n)||_{TV} \leq 2K_1(\pi, \nu)P(\tau \geq n) + O(\epsilon)$$

The notation $O(\epsilon)$ means that for some fixed $\epsilon_0 > 0$, there exists an $M \in \mathbb{N}$ such that for any $\epsilon \in (0, \epsilon_0)$, $||\pi - \mathcal{L}(X_n)||_{TV} \leq \frac{2P(\tau \geq n)}{1-r(\pi,\nu)} + M\epsilon$.

In[32], $K_1(\pi, \nu) = \frac{1}{1-r(\pi,\nu)}$ where $r(\pi, \nu)$ is the rejection rate in a rejection sampler with stationary distribution $\pi$ and proposal distribution $\nu$ (See Definition 8.2). By Equation 8.2, $\frac{1}{1-r(\pi,\nu)} = ess\sup_z \frac{f_\pi(z)}{f_\nu(z)}$. Thus, both Theorem 3.2 (with $s_1 = 1, s_2 = \infty$) and Theorem 2 of[32] require calculating upper bounds on $r(\pi, \nu)$. However, Theorem 3.2 estimates an exact upper bound on the Wasserstein distance while Theorem 2 of[32] estimates an upper bound on the total variation distance with an unknown component that converges to 0 as $\epsilon \to 0$.

Applying total variation distance to the bound generated by Theorem 3.2 and taking $\epsilon = 0$, we get a similar result to[32], $||\pi - \mathcal{L}(X_n)||_{TV} \leq K_1(\pi, \nu)P(\tau \geq n)$. See Remark 2 for more details. In our simulations, exact simulation of $d(X_n, Y_n) = 0$ was not realistically attainable, however. This is why[32] introduces the error $\epsilon$. This points to CRN simulation being a better fit for bounding Wasserstein distance compared to total variation distance.

## 4. Gibbs sampler for a model related to James-Stein estimators

Consider a Gibbs sampler for a model related to the James-Stein estimator, which we will apply Theorem 3.2 to simulate convergence bounds on the Wasserstein distance. The James-Stein estimator is similar to a variance component model (Example 5.1), but each $\theta_i$ has only one observation, $Y_i$. Its significance in the frequentist statistics literature comes from Stein's paradox, which proves that for $\theta_i$ the estimator, $\left(1 - \frac{I-2}{\sum_{j=1}^{I} Y_j^2}\right)_+ Y_i$, has smaller risk than the intuitive estimator, $Y_i$, when $I \geq 3$[60]. Here, we look at the same model setup, but through a Bayesian statistics lense. See[50] for more details on this example.

**Algorithm 2** The Gibbs sampler algorithm for a model related to James-Stein estimators.

Initialize $(\vec{\theta}_0, \mu_0, A_0)$
**for** $n = 1, \ldots, N$ **do**
For each $i \in 1, \ldots, q$, generate $\theta_{n,i}|\mu_{n-1}, A_{n-1}, Y \sim N(\frac{Y_i A_{n-1}+\mu_{n-1}V}{A_{n-1}+V}, \frac{VA_{n-1}}{V+A_{n-1}})$
Generate $\mu_n|\vec{\theta}_n, A_{n-1}, Y \sim N(\bar{\theta}_n, A_{n-1}/q)$ where $\bar{\theta}_n = \frac{1}{q}\sum_{i=1}^{q}\theta_{n,i}$
Generate $A_n|\vec{\theta}_n, \mu_n, Y \sim IG\left(\alpha + \frac{q-1}{2}, \frac{\sum_{i=1}^{q}(\theta_{n,i}-\mu_n)^2}{2} + \beta\right)$
**end for**

**Example 4.1** (A model related to James-Stein estimators). Suppose that we have the following observed data $Y \in \mathbb{R}^q$ where

$$Y_i|\theta_i \sim N(\theta_i, V)$$

for known $V > 0$ and unknown $\theta_i \sim N(\mu, A)$ ($\theta_i$'s are i.i.d.), where $\mu$ has a flat prior and $A \sim IG(\alpha, \beta)$.

The joint posterior density function of $\vec{\theta}, \mu, A \mid Y$ is proportional to the following equation,

$$g(\vec{\theta}, \mu, A \mid Y) = f_{IG}(A \mid \alpha, \beta) \times \prod_{i=1}^{q} f_N(Y_i \mid \theta_i, V) \times \prod_{i=1}^{q} f_N(\theta_i \mid \mu, A) \quad (6)$$

The Gibbs sampler for a model related to James-Stein estimators.
Algorithm 2 shows how to apply a Gibbs sampler on this example.

We will bound the Wasserstein distance between a Markov chain and its corresponding stationary distribution using Lemma 4.1.

**Lemma 4.1.** *Let* $\{X_n\}_{n\geq 1} = (\vec{\theta}_n, \mu_n, A_n)_{n\geq 1}$ *and* $\{X'_n\}_{n\geq 1} = (\vec{\theta}'_n, \mu'_n, A'_n)_{n\geq 1}$ *be two copies of the Markov chain initialized with* $\vec{\theta}'_0 \sim N_q(\vec{Y}, VI_q)$, $A'_0 \sim IG(\alpha + (q-1)/2, \beta)$, $\mu'_0 \sim N(\bar{\theta}'_0, A'_0)$, *and* $X_0 \perp\!\!\!\perp X'_0$. *Fix* $p \in (1, \infty]$ *and let* $\pi$ *be the corresponding stationary distribution. Then for* $r \geq p$

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq \left(\frac{\Gamma(\alpha + (q-1)/2)(2\pi)^{1/2}}{\beta^{\alpha+(q-1)/2}} \frac{1}{L} E[d(X_n, X'_n)^r]\right)^{1/r}$$

*Where* $L = \int_B \frac{A^{-\alpha-1}e^{-\beta/A}}{(A+V)^{q/2}} e^{\frac{1}{2}\sum_{i=1}^{q}\left[\frac{(\frac{Y_i}{V}+\frac{\mu}{A})^2}{\frac{1}{V}+\frac{1}{A}} - \frac{Y_i^2}{V} - \frac{\mu^2}{A}\right]} \mathrm{d}(\mu, A)$ *and* $B \subset \mathbb{R} \times \mathbb{R}_+$ *is a bounded set.*

*The proof is in section 8.4.*

### 4.1. Real Data Example 4.1

We apply our results to the batting averages ($Y$) for 18 ($q = 18$) major league baseball players in 1970[9,28]. We want to find the estimated upper bound on the Wasserstein and total variation distance for a Gibbs sampler applied to the

James-Stein estimator using the baseball data. We set the priors to $\alpha = 0.01$, $\beta = 2$ and $V = Var(Y) = 0.00485$. The initial value of our Markov chains are independently drawn; $X_0 = (\vec{\theta}_0, \mu_0, A_0) = (100)^{q+2}$ and $X_0'$ is distributed according to Lemma 4.1. Using Lemma 4.1, $K_1(\pi, \nu) = 5.9535$ and

$$W_{||\cdot||_1,1}(\mathcal{L}(X_n), \pi) \leq 5.9535 E[||X_n - X_n'||_1]$$

holds almost surely.

Using the CRN technique, we simulated $||X_n - X_n'||_1$ one thousand times ($M = 1000$) over 20 iterations ($N = 20$). At iteration 1, $\frac{1}{1000}\sum_{i=1}^{1000} ||X_{1,i} - X_{1,i}'||_1 = 1994.712$. By iteration 9, $\frac{1}{1000}\sum_{i=1}^{1000} ||X_{9,i} - X_{9,i}'||_1 = 0.00012$. Figure 1 graphs the upper bound of the Wasserstein distance between $\mathcal{L}(X_n)$ and $\pi$ on a log scale.

At iteration 9, the bound on the Wasserstein distance first hits below 0.01 and $W_{||\cdot||_1,1}(\mathcal{L}(X_9), \pi) \leq 0.00073$.

Before we bound the Markov chain and the corresponding stationary distribution in total variation we introduce Theorem 4.8 of[54], which provides an upper bound in total variation based on the expected differences between the Markov chains through the constant $K_{TV}$. This theorem will also be used to generate the one-shot coupling bound.

**Proposition 4.2** (Theorem 4.8 of[54]). *For two copies of a Gibbs sampler on a model related to the James-Stein estimators with flat priors on $\mu$ and A, the following holds for $q \geq 3$,*

$$\|\mathcal{L}(X_n) - \mathcal{L}(X_n')\|_{TV} \leq K_{TV} E[|A_0 - A_0'|] \left(\frac{1}{q}\right)^{n-1} \tag{7}$$

*where* $K_{TV} = \frac{(S/2)^{\frac{q-1}{2}}}{\Gamma(\frac{q-1}{2})} \left(\frac{S}{q+1}\right)^{-\frac{q-3}{2}} e^{-\frac{q+1}{2}}$ *and* $S = \sum_{i=1}^{q}(Y_i - \bar{Y})^2$.

Using Proposition 4.2, we get $K_{TV} = 0.0047$. Therefore by Lemma 4.11 of[54],

$$\|\mathcal{L}(X_n) - \pi\|_{TV} \leq 0.0282 E[||X_n - X_n'||_1]$$

So, by iteration 6, $\|\mathcal{L}(X_6) - \pi\|_{TV} \leq 0.00866$.

We compare our bound to Proposition 4.2 and Theorem 4 of[50], which use one-shot coupling and DnM arguments, respectively, to generate a theoretical bound in total variation. The methodologies and assumptions are outlined in Table 1.

Note the minor variations in the examples, Proposition 4.2 assumes that the prior on $A$ is flat. In Theorem 4 of[50], the prior on $A \sim IG(-1, 2)$ with the intent of generating a flat distribution. In our real data example we set the prior to $A \sim IG(0.01, 2)$, which could approximate a flat distribution. We will assume the minor variations are comparable.

Table 2 and Figure 2 compare the bounds on total variation distance using three different methods: CRN, DnM, and one-shot coupling. The one-shot

**Table 1.** Theoretical bounds that are benchmarked against the CRN simulated bound.

| Method | Common random number | Drift and minorization | One shot coupling |
|---|---|---|---|
| Source | Lemma 4.1 | Theorem 4 of[50] | Proposition 4.2 and Proposition 6 of[42] |
| Prior | $A \sim IG(0.01, 2)$ and $f_\mu \propto 1$ | $A \sim IG(-1, 2)$ and $f_\mu \propto 1$ | $f_A \propto 1$ and $f_\mu \propto 1$ |
| Bound | $\|\mathcal{L}(X_n) - \pi\|_{TV} \leq 0.0282 E[\|X_n - X_n'\|]$ where $X_n, X_n'$ are simulated using CRN. | $\|\mathcal{L}(X_n) - \pi\|_{TV} \leq 0.967^n + (0.935)^n \times \left(1.17 + \sum_{i=1}^{q} (\theta_i - \bar{Y})^2\right)$ | $\|\mathcal{L}(X_n) - \pi\|_{TV} \leq \frac{0.0282}{17} E[\|A_0 - A_1\|] \left(\frac{1}{18}\right)^n$ |

All bounds have initial value $X_0 = (\vec{\theta}_0, \mu_0, A_0) = (100)^{q+2}$. $f$ denotes the density function.

**Table 2.** Comparison of total variation bounds to stationarity for methods outlined in Table 1 of the Gibbs sampler for a model related to James-Stein estimators.

| Iteration | CRN | DnM | One-Shot Coupling |
|---|---|---|---|
| 1 | 56.2430 | 167,409.95 | 58.6595 |
| 2 | 336.0368 | 156,528.33 | 3.2589 |
| 3 | 20.0791 | 146,354.02 | 0.1811 |
| 4 | 1.4003 | 136,841.04 | 0.0101 |
| 5 | 0.0898 | 127,946.40 | **0.0005** |
| 6 | **0.0087** | 119,629.91 | 3.1049e-05 |
| 7 | 0.0007 | 111,853.99 | 1.7247e-06 |
| 8 | 6.0301e-05 | 104,583.51 | 9.5814e-08 |
| 9 | 3.4421e-06 | 97,785.60 | 5.3230e-09 |
| 10 | 2.6439e-07 | 91,429.56 | 2.9572e-10 |
| ... | ... | ... | ... |
| 249 | 1.6957e-307 | **0.0099** | 2.7286e-310 |

The values for each method that first reach less than 0.01 are in bold.

coupling bound indicates total variation distance of less than 0.01 by iteration 5. The DnM bound indicates total variation distance of less than 0.01 by iteration 249. Thus, the one-shot coupling bounds performs the best closely followed by the CRN bound and significantly better than the DnM bound.

The bound generated from CRN simulation and one-shot coupling are expected to be very similar since to generate the geometric convergence rate of the one-shot coupling bound (the 1/18 in Table 1) we use CRN (see[29,54,55,62]). This is confirmed by Figure 2 where the CRN bound and the one-shot bound are approximately parallel.

Finally, we compare our convergence rates with informal methods in Figure 7. The Gelman-Rubin statistic is evaluated for each parameter. By iteration 5, the Gelman-Rubin statistic for all of the parameters graphed is less than 1.05. The traceplot paints a similar picture suggesting that all parameters have converged by iteration 6. Thus, the CRN bound indicating total variation distance convergence by iteration 6, is tight.

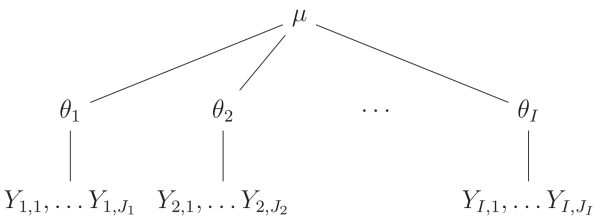


**Figure 3.** Schema of the variance component model where $\theta_i \sim N(\mu, V)$, $Y_{ij} \sim N(\theta_i, W)$. The parameters $W$, $V$, $\mu$ are unknown. $\theta_i|\mu, V \perp\!\!\!\perp \theta_j|\mu, V$ and $Y_{i_1,j_1}|\vec{\theta}, W \perp\!\!\!\perp Y_{i_2,j_2}|\vec{\theta}, W$.

## 5. Gibbs sampler for a variance component model

Consider a Gibbs sampler on the variance component model (see[49]), which we will apply Theorem 3.2 to simulate convergence bounds on the Wasserstein distance. Convergence rate bounds on the total variation metric for the variance component model have already been discussed in Theorems 1 of[12,33,49].

**Example 5.1** (Variance component model a.k.a. random effects model a.k.a. hierarchical model). Suppose that we have a population that has overall mean $\mu$, which consists of $I$ groups of mean $\theta_i \sim N(\mu, V)$ and where each group has $J_i$ observations with distribution $Y_{ij} \sim N(\theta_i, W)$. Conditional on $\vec{\theta}$ and $W$, the $Y_{ij}$'s are independent. Conditional on $\mu$ and $V$, the $\theta_i$'s are independent. The parameters $\mu, \vec{\theta}, V, W$ are unknown. See Figure 3 for a graphical representation.

Propose the following priors on the unknown parameters where $a_1, b_1, a_2, b_2, b_3 > 0$.

$$V \sim IG(a_1, b_1) W \sim IG(a_2, b_2) \mu \sim N(a_3, b_3) \tag{8}$$

The unnormalized posterior distribution of $(\vec{\theta}, V, W, \mu)$ is written as follows,

$$g(\vec{\theta}, v, w, \mu) = f_{IG}(V \mid a_1, b_1) f_{IG}(W \mid a_2, b_2) f_N(\mu \mid a_3, b_3) \left( \prod_{i=1}^{I} f_N(\theta_i \mid \mu, V) \right) \left( \prod_{i=1}^{I} \prod_{j=1}^{J_i} f_N(Y_{ij} \mid \theta_i, W) \right)$$

We would like to sample from the normalized posterior distribution.
The Gibbs sampler for the variance component model.
Algorithm 3 shows how to apply a Gibbs sampler on this example.

Suppose that we have the following proposal density,

$$f_v(\vec{\theta}, V, W, \mu) = f_{IG}(V \mid 2a_1 + 1, 2b_1 - 1) f_{IG}(W \mid 2a_2 + 1, 2b_2) f_N(\mu \mid a_3, b_3) \prod_{i=1}^{I} f_N\left(\theta_i \mid \bar{Y}_i, \frac{W}{2J_i}\right) \tag{9}$$

**Algorithm 3** A Gibbs sampler algorithm for the variance component model

Initialize $(\vec{\theta}_0, V_0, W_0, \mu_0) = (\vec{\theta}_0, v_0, w_0, \mu_0)$
**for** $n = 1, \ldots, N$ **do**
Generate $W_n|\vec{\theta}_{n-1} \sim IG\left(a_2 + JI/2, b_2 + \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\theta_{n-1,i})^2}{2}\right)$
Generate $V_n|\mu_{n-1}, \vec{\theta}_{n-1} \sim IG\left(a_1 + I/2, b_1 + \frac{\sum_{i=1}^{I}(\theta_{n-1,i}-\mu_{n-1})^2}{2}\right)$
Generate $\mu_n|V_n, \vec{\theta}_{n-1} \sim N\left(\frac{a_3 V_n + b_3\sum_{i=1}^{I}\theta_{n-1,i}}{V_n + Ib_3}, \frac{V_n b_3}{V_n + Ib_3}\right)$
For each $i \in I$, generate $\theta_{n,i}|V_n, W_n, \mu_n \sim N\left(\frac{\mu_n W_n + V_n\sum_{j=1}^{J} Y_{ij}}{W_n + JV_n}, \frac{V_n W_n}{W_n + JV_n}\right)$
**end for**

The following lemma bounds the Wasserstein distance between the variance component model Gibbs sampler and the stationary distribution using the above proposal distribution.

**Lemma 5.1.** *Let $X_n = (\vec{\theta}_n, V_n, W_n, \mu_n)$ and $X'_n = (\vec{\theta}'_n, V'_n, W'_n, \mu'_n)$ be two copies of the variance component Gibbs sampler initialized with $X'_0 \sim \nu$ where $f_\nu$ is defined in Equation 9 and $X_0 \perp\!\!\!\perp X'_0$. Fix $p \in [1, \infty)$ and let $\pi$ be the corresponding stationary distribution. Then for $r \geq p$, $b_1 > 1/2$,*

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq K_2(\pi, \nu)^{1/r} E[d(X_n, X'_n)^{2r}]^{\frac{1}{2r}} \tag{10}$$

*Where*

- $K_2(\pi, \nu) = \frac{1}{L} C_1 C_2^{1/2}$
- $L = \int_B (b_1^*)^{-a_1^*} e^{-\frac{(\mu - a_3)^2}{2b_3}} \left(b_2^*\right)^{-(a_2^*)} \mathrm{d}(\vec{\theta}, \mu)$ *where* $B \subset \mathbb{R}^{I+1}$
- $C_1 = \frac{\Gamma(a_1)}{\Gamma(a_1^*) b_1^{a_1}} \frac{\Gamma(a_2)}{\Gamma(a_2^*) b_2^{a_2}} (2\pi)^{\sum_{i=1}^{I} J_i/2 + (I+1)/2} b_3^{1/2}$
- $C_2 = \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}} \frac{\Gamma(I/2-1)}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{I+\sum_{i=1}^{I} J_i} \pi^{T+I/2}} \frac{\Gamma(T-1)}{S^{T-1}}$
- $a_1^* = a_1 + I/2$ *and* $b_1^* = b_1 + \frac{\sum_{i=1}^{I}(\theta_i - \mu)^2}{2}$
- $a_2^* = a_2 + \sum_{i=1}^{I} J_i/2$ *and* $b_2^* = b_2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{(Y_{ij} - \theta_i)^2}{2}$
- $S = \sum_{i=1}^{I}(\sum_{j=1}^{J_i} Y_{ij}^2 - J_i(\bar{Y}_i)^2)$ *and* $T = \sum_{i=1}^{I} J_i - I/2$

*The proof is in Section 8.5*

Note that in Lemma 5.1 the value of $K_2(\pi, \nu)$ still has an integral, which must be bounded from below, $L = \int_B g(\vec{\theta}, v, w, \mu)\mathrm{d}(\vec{\theta}, V, W, \mu)$. To address this, we use the `adaptIntegrate` function in `R`.

For this variance component model, upper bounds in total variation distance have been established. In Real Data Example 5.1, we compare our bounds in Wasserstein distance to the total variation bounds generated in[33].

[49] also provides an upper bound on the total variation distance between the variance component Gibbs sampler model and its corresponding stationary distribution (See Theorem 1 of[49]). However the constants required to obtain an explicit upper bound are difficult to calculate (see Remark 6 of[49]). An upper

**Table 3.** Summary of simulated data used in[33] (see their Table 1) and in Numerical Example 5.1. $I = 5$, $J = 10$, $\sum_{i=1}^{5}\sum_{j=1}^{10}(y_{ij} - \bar{y}_i)^2 = 32.990$.

| Cell | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\bar{Y}_i$ | −0.80247 | −1.0014 | −0.69090 | −1.1413 | −1.0125 |

**Table 4.** Hyperparameters used in Numerical Example 5.1 and Table 1 of[33]. $\bar{Y} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$.

| Hyperparameter | $a_1$ | $b_1$ | $a_2$ | $b_2$ | $a_3$ | $b_3$ |
|---|---|---|---|---|---|---|
| From[33] and this paper | 2.5 | 1 | 1 | 1 | $\bar{Y}$ | 1 |

bound in total variation distance for this example is also established in[12]. Discussions of convergence diagnostics on total variation distance for variations of this example can be found in[1,24,31,47,56].

Upper bounds in Wasserstein distance of the variance component model have also already been discussed.[42] analyzes the variance component model when the prior on $\mu$ is $\mu \propto 1$ instead of $\mu \sim N(a_3, b_3)$. In particular Proposition 25 of[42] states that if $\lim_{I\to\infty} J^2/(I^{3+\delta}) = \infty$ for some $\delta > 0$ and the model is properly specified (condition (E2) in[42]), then the geometric rate of convergence for this model goes to 0. Proposition 24 of[42] also bounds the total variation distance in terms of the Wasserstein distance. Discussions of convergence rates for the variance component model in Wasserstein distance are also presented in[15,61].

### 5.1. Numerical Example 5.1

We are going to compare our convergence bounds to the bounds from[33]. The data for this example is simulated in[33] and summarized in Table 3.[33] generates an upper bound on the total variation distance between a Markov chain and its corresponding stationary distribution using a DnM bound stated in[48] (see Theorem 12 of[48] and Theorem 3.1 of[33]). The hyperparameters for the DnM bound are defined in Table 4.

Applying Lemma 5.1, $K_2(\pi, \nu) = 48.7818$ and the following holds for $r = 1$,

$$W_{||\cdot||_1,1}(\mathcal{L}(X_n), \pi) \leq 48.7818 E\left[(||X_n - X_n'||_1)^2\right]^{\frac{1}{2}}$$

The initial value of $X_0 = (\vec{\theta}_0, V_0, W_0, \mu_0)$ is $V_0 \sim IG(a_1, b_1)$, $W_0 \sim IG(a_2, b_2)$ and $(\vec{\theta}_0, \mu_0) =$ the initial value defined in[33], Remark 4.1 of[33]. The initial value of $X_0'$ satisfies Equation 9.

We simulated $(||X_n - X_n'||_1)^2$ 1000 times ($M = 1000$) using CRN. Figure 4 graphs the estimated upper bound on the Wasserstein distance between the Markov chain and its corresponding stationary distribution while Table 5

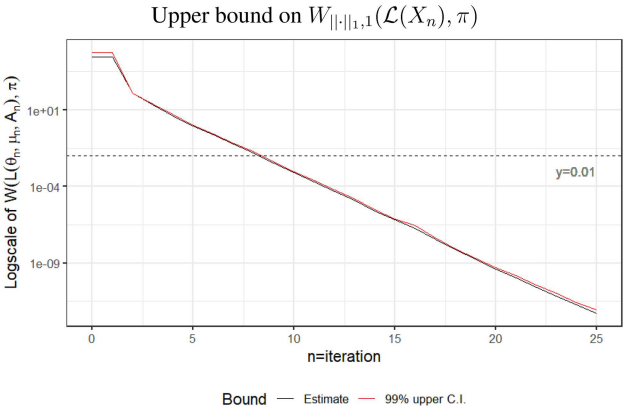


**Figure 4.** A Wasserstein distance upper bound on the Gibbs sampler for a variance component model (Numerical Example 5.1) using CRN. The estimate uses the mean of 1000 simulations with initial values of $\mu_0, \vec{\theta}_0$ satisfying the initial values defined in[33] (see their Remark 4.1). The vertical axis is a log scale.

**Table 5.** Wasserstein bound estimates and confidence intervals for the variance component model, Numerical Example 5.1.

| $n$ | $\frac{1}{1000}\sum_{m=1}^{1000}(\|\|X_{n,m} - X'_{n,m}\|\|_1)^2$ | 99% lower C.I. | Estimated bound | 99% upper C.I. |
|---|---|---|---|---|
| 1 | 397506.3 | 0 | 3075.601 | 55941.57 |
| 2 | 6.0002 | 112.7212 | 119.4924 | 125.8998 |
| 3 | 0.1835 | 17.3169 | 20.8980 | 23.9495 |
| 4 | 0.0076 | 3.2230 | 4.2734 | 5.1123 |
| 5 | 0.0004 | 0.7855 | 0.9132 | 1.0251 |
| 6 | 2.4539e-05 | 0.1837 | 0.2417 | 0.2881 |
| 7 | 1.4015e-06 | 0.0441 | 0.0578 | 0.0687 |
| 8 | 1.0485e-07 | 0.0114 | 0.0158 | 0.0192 |
| 9 | 5.4138e-09 | 0.0027 | 0.0036 | 0.0043 |

provides the bound estimates and the 99% confidence intervals using Lemma 3.3.

The CRN bound indicates Wasserstein distance of less than 0.0099 by iteration 9 ($W_{||\cdot||_1,1}(\mathcal{L}(X_9), \pi) \leq 0.00359$). The bound for[33] indicates total variation distance of less than 0.00999 by iteration 3415 ($||\mathcal{L}(X_{3415}) - \pi||_{TV} \leq 0.00999$) (See Table 2 of[33]). Note that the CRN bound is measured in Wasserstein distance while the bound in[33] is measured in total variation distance. The large disparity between the DnM and the CRN bound is consistent with the conclusions of[40] that say that the convergence rate bounds based on single-step DnM tend to be overly conservative.

Finally we compare our convergence rates with informal methods. The trace-plot (Figure 8) suggests that convergence of $\mu$, $V$, and $W$ is nearly immediate. The Gelman-Rubin statistic is evaluated at different iterations in Table 8. By iteration 10, the Gelman-Rubin statistic is close to or less than 1.05 for $V$, $W$, and $\mu$.

**Table 6.** Hyperparameters used in Real Data Example 5.1 and[12] (see their Section 4).

| Hyperparameter | $a_1$ | $b_1$ | $a_2$ | $b_2$ | $a_3$ | $b_3$ |
|---|---|---|---|---|---|---|
| From[12] | 0.5 | 1 | 1 | 0 | 0 | $10^{12}$ |
| This paper | 0.5 | 1 | 1 | 1 | 0 | $10^{12}$ |

The hyperparameters follow the same notation as in Equation 8.

### 5.2. Real Data Example 5.1

We are given the yield of dyestuff from 6 batches of 5 yield measurements in grams (i.e. $I = 6$ and $J = 5$). This data was published in Davies[14]. The data can be structured as a variance component model where $\theta_i$ represents the mean yield for batch $i$.

We are going to compare our Wasserstein bound to the total variation bound from[12] generates an upper bound on the total variation distance between a Markov chain and its corresponding stationary distribution using a modified version of the popular DnM condition stated in[48] (see Theorem 12 of[48] and Proposition 1 of[12]) and estimates the required DnM parameters $(\Lambda, \lambda, \epsilon)$ using simulation. The hyperparameters for the DnM bound are defined in Table 6.

In our comparison, we will first generate an upper bound on the Wasserstein distance between the Markov chain and the corresponding stationary distribution. Next, we will generate an upper bound on the total variation distance based on the Wasserstein distance using Proposition 24 of[42].

Proposition 24 of[42] will be used to bound the total variation distance given the Wasserstein distance. Note that in[42], the prior on $\mu \sim 1$ is flat. Since in this example $\mu \sim N(0, 10^{12})$, we assume that they are the same.

**Proposition 5.2** (Proposition 24 of[42]). *Let $X_n$ be a variance component Gibbs sampler generated according to the priors of $V \sim IG(a_1, b_1)$, $W \sim IG(a_2, b_2)$, and $\mu \propto 1$. Assume that $I \geq 2$. Then for each $n \geq 1$ and $X_0 = (\vec{\theta_0}, V_0, W_0, \mu_0) \in \mathbb{R}^I \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$,*

$$||\mathcal{L}(X_n) - \pi||_{TV} \leq K_{TV} W_{||\cdot||_1,1}(\mathcal{L}(X_{n-1}), \pi) \tag{11}$$

*where $K_{TV} = (c_1 + c_2)J^{3/2}I$ and*

- $c_1 = \frac{2}{I}\left(\frac{I}{2} + a_1\right)\left(1 + \sqrt{\frac{2}{b_1}}\frac{1}{I} + \frac{1}{2b_1 I^2}\right)^{I/2+a_1-1}\left(\sqrt{\frac{2}{b_1}} + \frac{1}{b_1 I}\right)$
- $c_2 = \frac{2}{IJ^{3/2}}\left(\frac{IJ}{2} + a_2\right)\left(1 + \frac{2}{\sqrt{b_2}IJ} + \frac{1}{b_2 I^2 J^2}\right)^{IJ/2+a_2-1}\left(2\sqrt{\frac{J}{b_2}} + \frac{2}{b_2\sqrt{JI}}\right)$

We bound the whole chain in Proposition 5.2 $(\vec{\theta_n}, V_n, W_n, \mu_n)$ rather than the marginal chain $(\vec{\theta_n}, \mu_n)$ as presented in Proposition 24 of[42] due to de-initialization properties[45]. See Section 5.1 of[42] for more details.

Applying R simulation to Lemma 5.1, we calculate $K_2(\pi, \nu) = 0.00198$. The following holds,

$$W_{||\cdot||_1,1}(\mathcal{L}(X_n), \pi) \leq 0.00198 E\left[(||X_n - X_n'||_1)^2\right]^{\frac{1}{2}}$$

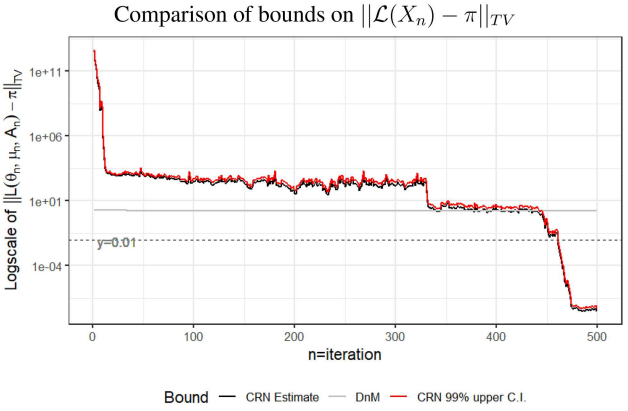


**Figure 5.** Comparison of total variation distance upper bounds on the Gibbs sampler for a variance component model (Real Data Example 5.1) using CRN simulation and a DnM bound. The CRN estimate uses the mean of 500 simulations. The initial values of $X_0$ satisfy Equation 9 in[12].

**Table 7.** Total variation bound estimates and confidence intervals for the variance component model, Real Data Example 5.1.

| $n$ | $\frac{1}{500}\sum_{m=1}^{500}(\|\|X_{n,m} - X'_{n,m}\|\|_1)^2$ | 99% lower C.I. | CRN | 99% upper C.I. | DnM |
|---|---|---|---|---|---|
| 10 | 3.1983e + 11 | 0 | 686202.5 | 1210249 | 1.7888 |
| 50 | 574231.3 | 413.0225 | 919.4749 | 1232.997 | 1.7821 |
| 100 | 61787.44 | 0 | 301.6105 | 483.7103 | 1.7737 |
| 250 | 10979.74 | 0 | 127.1430 | 231.2048 | 1.7490 |
| 461 | 9.0992e-06 | 0 | **0.0037** | 0.0067 | 1.7149 |

Comparison of the estimated total variation bound using CRN with the total variation bound generated from[12], which was calculated using DnM.

Applying Proposition 5.2, $K_{TV} = 611.4339$ and so,

$$||\mathcal{L}(X_n) - \pi||_{TV} \leq 1.21338E\left[(||X_n - X'_n||_1)^2\right]^{\frac{1}{2}}$$

The initial value of $X_0$ is $\vec{\theta}_0 =$ the initial values defined in Equation 9 of[12], $V_0 \sim IG(a1, b1)$, $W_0 \sim IG(a2, b2)$ and $\mu_0 = \bar{Y}$. The initial value of $X'_0$ is distributed according to Equation 9.

We simulated $X_n$ and $X'_n$ 500 times ($M = 500$) using CRN. Figure 5 graphs the estimated upper bound on the total variation distance between the Markov chain and its corresponding stationary distribution. Table 7 provides the total variation bound estimates and the 99% confidence intervals using Lemma 3.3.

The CRN bound indicates total variation distance of less than 0.01 by iteration 461 ($||\mathcal{L}(X_{461}) - \pi||_{TV} \leq 0.0037$). The DnM bound indicates total variation distance of less than 0.01 by iteration 98, 750 ($||\mathcal{L}(X_{98,750}) - \pi||_{TV} \leq 0.0072$). The two bounds are compared in Figure 5 and Table 7. The large disparity between the DnM and the CRN bound is consistent with the conclusions of[40] that say that the convergence rate bounds based on single-step DnM tend to be overly conservative. See Figure 5 for a graphical comparison between the two bounds.

**Table 8.** Gelman-Rubin statistic at various iterations for Numerical Example 5.1.

| Iteration | V | W | $\mu$ |
|---|---|---|---|
| 10 | 1.064556 | 1.046249 | 1.062693 |
| 20 | 1.026606 | 1.021977 | 1.034042 |
| 30 | 1.017409 | 1.013104 | 1.021825 |
| 40 | 1.013610 | 1.011070 | 1.014756 |
| 50 | 1.010736 | 1.008834 | 1.011396 |
| 60 | 1.009296 | 1.007351 | 1.010150 |

The range of $\mu$ is -2 to 0.

**Table 9.** Gelman-Rubin statistic at various iterations for Real Data Example 5.1.

| $n$ | $V$ | $W$ | $\mu$ | $\theta_1$ |
|---|---|---|---|---|
| 50 | 1.186270 | 2.052273 | 2.984139 | 2.792903 |
| 100 | 1.095662 | 1.591515 | 1.597947 | 1.587649 |
| 150 | 1.083923 | 1.454056 | 1.241374 | 1.215334 |
| 200 | 1.062822 | 1.208700 | 1.086050 | 1.097018 |
| 250 | 1.054723 | 1.074812 | 1.038007 | 1.056482 |

The range of $V$ is (0.5, 3.5), $W$ is 1000, 20000, and $\mu, \theta_1$ is (1400, 1600). Note that because the range of initial values to calculate the Gelman-Rubin statistic is larger than the initial values in our bound, the convergence rates cannot be blindly compared. Rather, the Gelman-Rubin statistic indicates that the chains move relatively slowly.

Finally, we compare our convergence rates with informal methods. The trace-plot (Figure 9) suggests that convergence of $\mu$, $V$, and $W$ to stationarity occurs after the first few iterations (the variance of $\mu$ seems to slightly increase with $n$, however). The Gelman-Rubin statistic is evaluated at different iterations in Table 9. By iteration 250, the Gelman-Rubin statistic is close to or less than 1.05 for $V$, $W$, $\mu$, and $\theta_1$.

## 6. Gibbs sampler for a Bayesian linear regression model

Consider a Gibbs sampler for the Bayesian linear regression model with semi-conjugate priors (see Chapter 5 of[43]) for which we will apply Theorem 3.2 to simulate convergence bounds on the Wasserstein distance. For this particular example, we can also provide an upper bound on the total variation distance given the Wasserstein distance using similar calculations as[54].

**Example 6.1** (Bayesian linear regression model). Suppose that we have the following observed data $Y \in \mathbb{R}^k$ and $X \in \mathbb{R}^{k\times q}$ where

$$Y|\beta, \sigma^2 \sim N_k(X\beta, \sigma^2 I_k)$$

for unknown parameters $\beta \in \mathbb{R}^q, \sigma^2 \in \mathbb{R}_+$. Suppose that we apply the prior distributions on the unknown parameters,

$$\beta \sim N_q(\beta_0, \Sigma_\beta)\sigma^2 \sim IG\left(\frac{\upsilon_0}{2}, \frac{\upsilon_0 c_0^2}{2}\right).$$

The joint posterior density function of $\beta, \sigma^2|Y, X$ is proportional to the following equation,

$$g(\beta, \sigma^2) = \frac{1}{(\sigma^2)^{(k+\upsilon_0)/2+1}} \tag{12}$$

$$\exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) - \frac{1}{2}(\beta - \beta_0)^T\Sigma_\beta^{-1}(\beta - \beta_0) - \frac{\upsilon_0 c_0^2}{2\sigma^2}\right). \tag{13}$$

The Gibbs sampler on the Bayesian linear regression model is based on the conditional posterior distributions of $\beta_n, \sigma_n^2$ and is defined as follows initialized at $\beta_0, \sigma_0^2$:

(1) $\beta_n|\sigma_{n-1}^2, Y, X \sim N_q(\tilde{\beta}_{\sigma_{n-1}^2}, V_{\beta,\sigma_{n-1}^2})$
(2) $\sigma_n^2|\beta_n, Y, X \sim IG\left(\frac{k+\upsilon_0}{2}, \frac{1}{2}\left[\upsilon_0 c_0^2 + (Y - X\beta_n)^T(Y - X\beta_n)\right]\right)$

where

$$V_{\beta,\sigma_{n-1}^2} = \left(\frac{1}{\sigma_{n-1}^2}X^TX + \Sigma_\beta^{-1}\right)^{-1}, \quad \tilde{\beta}_{\sigma_{n-1}^2} = V_{\beta,\sigma_{n-1}^2}\left(\frac{1}{\sigma_{n-1}^2}X^TY + \Sigma_\beta^{-1}\beta_0\right).$$

Replacing $\beta_n = \tilde{\beta}_{\sigma_{n-1}^2} + V_{\beta,\sigma_{n-1}^2}^{1/2} Z_n$, where $Z_n \sim N_q(0, I_q)$ into the equation for $\sigma_n^2$ and with $IG_n \sim IG(\frac{k+\upsilon_0}{2}, 1)$, we get that

$$\sigma_n^2|\sigma_{n-1}^2, Y, X = \left[\begin{array}{c} \frac{\upsilon_0 c_0^2}{2} + \frac{(X\tilde{\beta}_{\sigma_{n-1}^2} - Y + XV_{\beta,\sigma_{n-1}^2}^{1/2} Z_n)^T}{2} \\ \qquad\qquad (X\tilde{\beta}_{\sigma_{n-1}^2} - Y + XV_{\beta,\sigma_{n-1}^2}^{1/2} Z_n) \end{array}\right] IG_n \tag{14}$$

where $(Z_n, IG_n)_n$ are independent for all $n$.

Although the Markov chain may be high-dimensional in $\beta_n$, special properties of this Gibbs sampler allow us to upper bound the total variation distance between the law of two joint Markov chains, $\mathcal{L}(\sigma_n^2, \beta_n)$ and $\mathcal{L}(\sigma_\infty^2, \beta_\infty) = \pi$, in terms of the $L^1$ distance between $\sigma_n^2$ and $\sigma_\infty^2$ only.

**Lemma 6.1.** *The total variation distance between $\mathcal{L}(\beta_n, \sigma_n^2)$ and the corresponding stationary distribution, $\pi$, can be bounded by the expected distance as follows,*

$$\left\|\mathcal{L}(\beta_n, \sigma_n^2) - \pi\right\|_{\mathrm{TV}} \leq \frac{(k + \upsilon_0)^2}{2\upsilon_0 c_0^2} E[|\sigma_{n-1}^2 - \sigma_\infty^2|]$$

*The proof is in Section 8.6.*

Lemma 6.1 shows that total variation distance can ignore the difference between $\beta_n$ and $\beta_n'$. The Wasserstein distance, on the other hand, must take into account the distances between both $\beta_n$, $\beta_n'$ and $\sigma_n^2$, $\sigma_n^{2'}$ as shown below.

**Lemma 6.2.** *Let $(\beta_n, \sigma_n^2)_{n\geq 1}$ and $(\beta_n', \sigma_n^{'2})_{n\geq 1}$ be two copies of the Markov chain with $\sigma_0^{'2} \sim IG(\alpha', \beta')$ and $\beta_0' \sim N_q(\beta_0, \Sigma_\beta)$ where $\alpha' = (k + \upsilon_0)/2 - 2$ and $\beta' = \upsilon_0 c_0^2/2$. Fix $p \in [1, \infty)$ and let $\pi$ be the corresponding stationary distribution. Then for $r > p$*

$$W_{d,p}(\mathcal{L}(\beta_n, \sigma_n^2), \pi) \leq K_1(\pi, \nu)^{1/r} E[d((\beta_n, \sigma_n^2), (\beta_n', \sigma_n^{'2}))^r]^{1/r}$$

*Where $K_1(\pi, \nu) = \frac{C}{L}$*

- $C = \frac{16\Gamma((k+\upsilon_0)/2-2)(2\pi)^{q/2}\det(\Sigma_\beta)^{1/2}}{\Gamma((k+\upsilon_0)/2)((Y-X\hat{\beta})^T(Y-X\hat{\beta}))^2(\upsilon_0 c_0^2/2)^{(k+\upsilon_0)/2-2}e^2}$
- $L = \int_B \frac{e^{-\frac{1}{2}(\beta-\beta_0)^T\Sigma_\beta^{-1}(\beta-\beta_0)}}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}}\mathrm{d}(\beta)$ *and $B \subset \mathbb{R}^q \times \mathbb{R}_+$ is a bounded set.*

*The proof is in Section 8.6*

Convergence diagnostics for variations of the Bayesian linear regression Gibbs sampler can be found in[5,6,19,26,44,54,58]. However, we could not find any explicit bounds on this version of the Gibbs sampler for a Bayesian linear regression (i.e. the bounds were applied to more simpler examples or the analysis did not provide an explicit bound). Using example 6.1 notation, the papers[44,54,58] assume that $\Sigma_\beta = \sigma^2\Sigma$ where $\Sigma$ is known. That is, the variance of $Y$ and the variance of $\beta$ are proportionally the same. Putting the results of[54] and[44] together, it is shown that the geometric convergence rate that is generated when using one-shot coupling is the actual geometric convergence rate (see Theorem 3.1 of[44] compared with Theorem 4.4 of[54]). The actual geometric convergence rate is calculated using the spectral gap in[44]. The papers ([5,26]) assume that the error of $Y|\beta, \sigma^2$ are non-Gaussian (to take into account outliers) and provide sufficient conditions for geometric ergodicity with respect to $k$ (dimension of $Y$) and $q$ (dimension of $\beta$).[19] tells us that the Bayesian regression Gibbs sampler with semi-conjugate priors is geometrically ergodic regardless of the size of $k$ compared to $q$. Finally,[6] generates upper bounds in total variation between two copies of a linear regression Gibbs sampler model. The model in[6] adds slightly more complexity to example 6.1 by taking $\Sigma_\beta$ as random. Similar to our results, the chains converge quickly.

Using Lemma 6.2 we show how an upper bound on the convergence rate in Wasserstein distance can be simulated in Real Data Example 6.1. We further provide an estimate of the upper bound in total variation using Lemma 6.1.

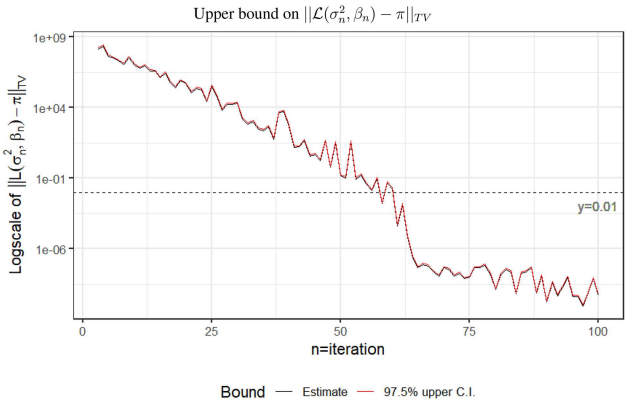


**Figure 6.** Total variation distance upper bound on the Gibbs sampler for a linear regression model (Real Data Example 6.1) using CRN. The estimate uses the mean of 10000 simulations with initial values $\sigma_0^2 = 100$ and $\beta_0 \sim N(0, I_q)$. The vertical axis is log-scaled and $r = 5$.

### 6.1. Real Data Example 6.1

Suppose that we are interested in evaluating the carbohydrate consumption (Y) by age, relative weight, and protein consumption (X) for twenty male insulin-dependent diabetics. For more information on this example, see Section 6.3.1 of[17].

We want to find the estimated upper bound on the Wasserstein and total variation distance to stationarity for a Bayesian linear regression Gibbs sampler. In this case, there are 20 observed values and 4 parameters ($k = 20, q = 4$). We set the priors to $\beta_0 = \vec{0}$, $\Sigma_\beta = I_4$, $\upsilon_0 = 6$, $c_0^2 = 140$. The initial values of our Markov chains are $\sigma_0^2 = 100$, $\beta_0 \sim N_4(0, I_4)$, $\sigma_0'^2 \sim IG(11, 420)$ and $\beta_0' \sim N_4(0, I_4)$. We can apply Lemma 6.2 as follows, where $K_1(\pi, \nu) = 1025971$ and $r = 5$.

$$W_{||\cdot||_1,1}(\mathcal{L}(\sigma_n^2, \beta_n), \pi) \leq 15.93E[||(\sigma_n^2, \beta_n) - (\sigma_n^{2'}, \beta_n)||_1^5]^{1/5}$$

holds almost surely.

By Lemma 6.1 the total variation distance is bounded above by $\frac{(k+\upsilon_0)^2}{2\upsilon_0 c_0^2} = 0.40238$ times the expected distance between $\sigma_n^2$ and $\sigma_\infty^2$ and so,

$$||\mathcal{L}(\sigma_n^2, \beta_n) - \pi||_{TV} \leq 6.41E\left[||(\sigma_n^2, \beta_n) - (\sigma_n^{2'}, \beta_n)||_1^5\right]^{1/5}$$

Using CRN simulation, we simulated $||(\sigma_n^2, \beta_n) - (\sigma_n^{2'}, \beta_n)||_1$ ten thousand times ($M = 10000$) over 100 iterations ($N = 100$). Figure 6 graphs the upper bound on the total variation distance.

By iteration 58, the total variation distance first hits below 0.01, $||\mathcal{L}(\sigma_{58}^2, \beta_{58}) - \pi||_{TV} \leq 0.0016$. In comparison, the Gelman-Rubin statistic first hits below 1.05 at iteration 49 (see Figure 10) and the traceplot suggests convergence after the first few iterations (see Figure 11).

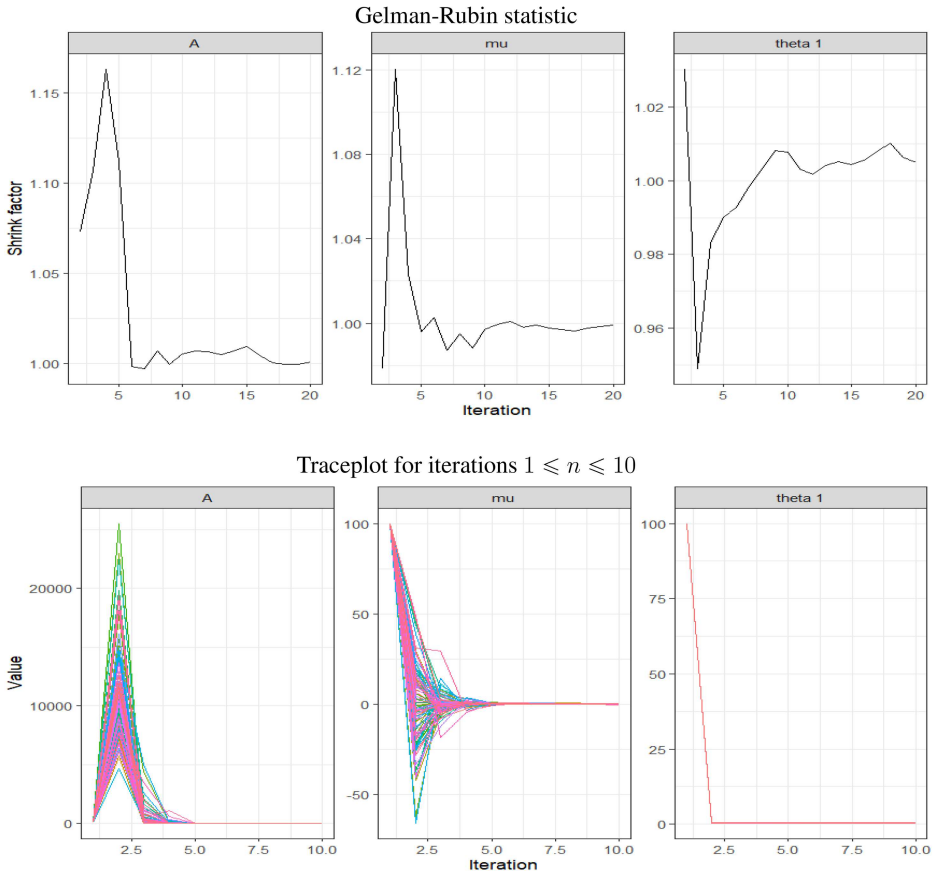


**Figure 7.** Gelman-Rubin statistic and traceplot for Real Data Example 4.1.

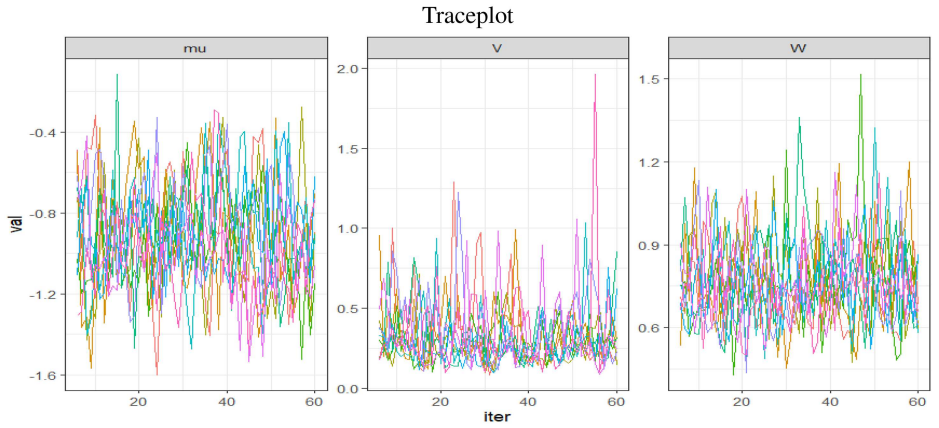


**Figure 8.** Traceplot for $\mu$, $V$, and $W$ for Numerical Example 5.1.

*Proof of Theorem 3.1.* Let $X_n, Y_n$ be two copies of the Markov chain such that $X_n = k_{\Theta_n}(X_0)$ and $Y_n = k_{\Theta'_n}(Y_0)$ (see Section 2 for more details on the notation). If $X_\infty \sim \pi$, then $k_{\Theta'_n}(X_\infty) \sim \pi$.

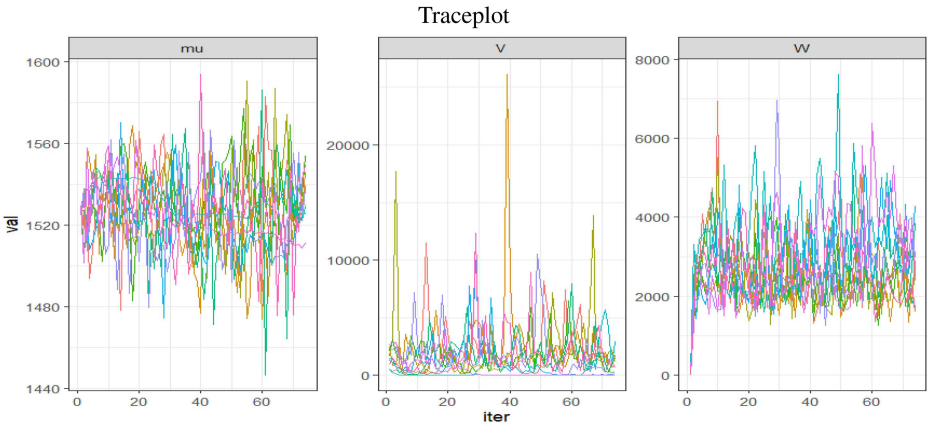


**Figure 9.** Traceplot for $\mu$, $V$, and $W$ for Real Data Example 5.1.

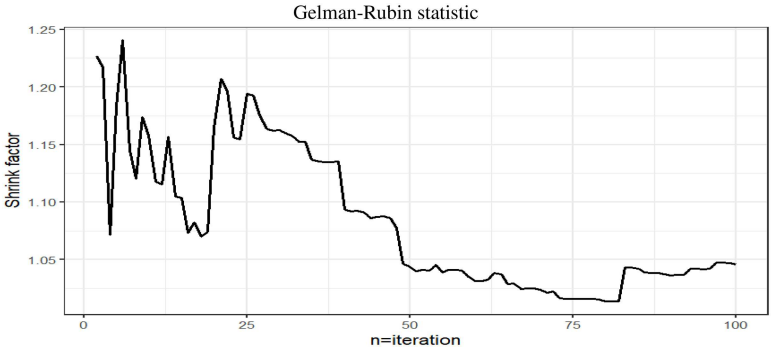


**Figure 10.** Gelman-Rubin statistic of $\sigma_n^2$.

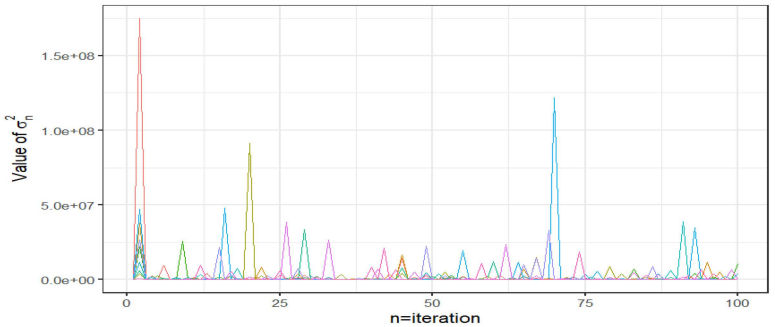


**Figure 11.** Traceplot of 10 simulations of $\sigma_n^2$ where $\sigma_0^2 = 100$.

First we show that $E[d(X_n, X_\infty)^r] = E\left[d(X_n, Y_n)^r \frac{\mathrm{d}\pi(Y_0)}{\mathrm{d}\nu(Y_0)}\right]$ where $\mathrm{d}\pi(Y_0)/\mathrm{d}\nu(Y_0) = f_\pi(Y_0)/f_\nu(Y_0)$ and $\mathrm{d}\pi/\mathrm{d}\nu$ is the Radon-Nikodym derivative.

$$E[d(X_n, X_\infty)^r] \tag{15}$$

$$= E[E[d(k_{\Theta_n}(X_0), k_{\Theta_n'}(X_\infty))^r \mid X_0, \Theta_n, \Theta_n']] \tag{16}$$

$$= E\left[\int_{\mathcal{X}} d(k_{\Theta_n}(X_0), k_{\Theta'_n}(y))^r \, \mathrm{d}\pi(y)\right] \tag{17}$$

$$= E\left[\int_{\mathcal{X}} d(k_{\Theta_n}(X_0), k_{\Theta'_n}(y))^r \frac{\mathrm{d}\pi(y)}{\mathrm{d}\nu(y)} \mathrm{d}\nu(y)\right] \tag{18}$$

$$= E\left[E\left[d(k_{\Theta_n}(X_0), k_{\Theta'_n}(Y_0))^r \frac{\mathrm{d}\pi(Y_0)}{\mathrm{d}\nu(Y_0)} \mid X_0, \Theta_n, \Theta'_n\right]\right] \tag{19}$$

$$= E\left[d(X_n, Y_n)^r \frac{\mathrm{d}\pi(Y_0)}{\mathrm{d}\nu(Y_0)}\right] \tag{20}$$

By Holder's inequality, for $s > 1$ and $K_s(\pi, \nu) = E\left[\left(\frac{\mathrm{d}\pi(Y_0)}{\mathrm{d}\nu(Y_0)}\right)^{\frac{s}{s-1}}\right]^{\frac{s-1}{s}}$.

$$E[d(X_n, X_\infty)^r] = E\left[d(X_n, Y_n)^r \frac{\mathrm{d}\pi(Y_0)}{\mathrm{d}\nu(Y_0)}\right] \leq K_s(\pi, \nu) E\left[d(X_n, Y_n)^{rs}\right]^{\frac{1}{s}} \tag{21}$$

Thus, the $L^r$-Wasserstein distance is bounded above as follows,

$$W_{d,r}(\mathcal{L}(X_n), \pi) \leq E[d(X_n, X_\infty)^r]^{\frac{1}{r}} \leq K_s(\pi, \nu)^{\frac{1}{r}} E\left[d(X_n, Y_n)^{rs}\right]^{\frac{1}{rs}} \tag{22}$$

Finally by Remark 6.6 of[59], for $p \leq r$, $W_{d,p}(\mathcal{L}(X_n), \pi) \leq W_{d,r}(\mathcal{L}(X_n), \pi)$ and so, by Equation 22,

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq W_{d,r}(\mathcal{L}(X_n), \pi) \leq K_s(\pi, \nu)^{\frac{1}{r}} E\left[d(X_n, Y_n)^{rs}\right]^{\frac{1}{rs}}$$

□

## 7. Alternative proof of Theorem 3.1 when s = 1

First we define essential supremum and infimum.

Let $f : \mathcal{X} \to \mathbb{R}$ be a function and $\lambda$ be the Lebesgue measure. The essential supremum is the smallest value $a \in \mathbb{R}$ such that $\lambda(x \mid f(x) < a) = 1$. More formally, $ess\sup_x f(x) = \inf_{a\in\mathbb{R}}\{a \mid \lambda(x \mid f(x) > a) = 0\}$. The essential infimum is likewise the largest value $a \in \mathbb{R}$ such that $\lambda(x \mid f(x) > a) = 1$. Or, $ess\inf_x f(x) = \sup_{a\in\mathbb{R}}\{a \mid \lambda(x \mid f(x) < a) = 0\}$ [38].

### 7.1. Rejection rate and separation distance

We will define rejection sampling and separation distance and show how they are related.

**Definition 8.1** (Rejection sampling). Suppose that we have a target distribution $\pi$, which we want to sample from, but is difficult to do, and we have a proposal distribution $\nu$ that is easier to sample from. Suppose also that for the corresponding density functions $f_\pi(x)$ and $f_\nu(x)$, $f_\nu(x) = 0 \implies f_\pi(x) = 0$, where

$x \in \mathcal{X}$ and that $K_1(\pi, \nu) = ess\sup_{x\in\mathcal{X}} f_\pi(x)/f_\nu(x)$. To generate a random variable $X_\infty$, $\mathcal{L}(X_\infty) = \pi$ we do the following,

(1) Sample $X \sim \nu$ and $U \sim Unif(0, 1)$ independently.
(2) If $U \leq \frac{1}{K_1(\pi,\nu)} \frac{f_\pi(X)}{f_\nu(X)}$ then accept $X$ as a draw from $\pi$. Otherwise reject $X$ and restart from step 1.

**Definition 8.2** (Rejection sampler rejection rate). For the rejection sampler algorithm defined above, denote the event $A = \{X \text{ is accepted as a draw from } \pi\}$. The rejection rate is $r(\pi, \nu) = 1 - P(A) = 1 - 1/K_1(\pi, \nu)$ where $K_1(\pi, \nu) = ess\sup_{x\in\mathcal{X}} f_\pi(x)/f_\nu(x)$.

*Proof.* See Section 11.2.2 of[51]. □

We further define the separation distance on the continuous state space $\mathcal{X}$ as follows. Separation distance was first defined in[3] for discrete state spaces. We use the definition of separation distance defined in[11] where the density functions are known.

**Definition 8.3** (Separation distance (Remark 5 of[11])). Let $\nu$ and $\pi$ be two probability distributions defined on the same measurable space $(\mathcal{X}, \mathcal{F})$ with densities $f_\nu$ and $f_\pi$ such that for $x \in \mathcal{X}$, $f_\pi(x) = 0 \implies f_\nu(x) = 0$. The separation distance is $s(\pi, \nu) = ess\sup_x \left(1 - \frac{f_\nu(x)}{f_\pi(x)}\right)$.

It turns out that the separation distance and the rejection rate of the rejection sampler are the same.

**Lemma 8.1.** *Let $\nu$ and $\pi$ be two distributions defined on the same measurable space $(\mathcal{X}, \mathcal{F})$. If $f_\nu$ is the proposal density and $f_\pi$ is the target density in a rejection sampler, then the rejection rate equals the separation distance,*

$$s(\pi, \nu) = r(\pi, \nu)$$

*Proof.*

$$\begin{aligned} s(\pi, \nu) &= ess\sup_{x\in\mathcal{X}} \left(1 - \frac{f_\nu(x)}{f_\pi(x)}\right) = 1 - ess\inf_{x\in\mathcal{X}} \frac{f_\nu(x)}{f_\pi(x)} \\ &= 1 - \frac{1}{ess\sup_{x\in\mathcal{X}} \frac{f_\pi(x)}{f_\nu(x)}} = 1 - \frac{1}{K_1(\pi, \nu)} = r(\pi, \nu) \end{aligned}$$

□

*Alternative proof of Theorem 3.1 for when $s_1 = 1, s_2 = \infty$.* In this proof, we use rejection sampling, which is similar to the proof method

presented in[32]. Let $X_n, Y_n$ be two copies of the Markov chain and $A = \{\text{Accept } Y_0 \text{ as a draw from } \pi\}$. Recall from Definition 8.2 that $1 - 1/K_1(\pi, \nu) = r(\pi, \nu) = 1 - P(A)$ and so $P(A) = 1/K_1(\pi, \nu)$. Note that $\mathcal{L}(Y_0|A) = \pi$ and so, $\mathcal{L}(Y_n|A) = \pi$.

$$\begin{aligned} E[d(X_n, Y_n)^p] &= E[d(X_n, Y_n)^p \mid A]P(A) + E[d(X_n, Y_n)^p \mid A^c]P(A^c) \\ &\geq E[d(X_n, X_\infty)^p]P(A) \\ &\geq E[d(X_n, X_\infty)^p]\frac{1}{K_1(\pi, \nu)} \end{aligned}$$

Thus,

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq E[d(X_n, X_\infty)^p] \leq K_1(\pi, \nu)E[d(X_n, Y_n)^p]$$

□

Further note that by combining Definition 8.2 and Lemma 8.1 we can write $K$ as follows

$$K_1(\pi, \nu) = \frac{1}{1 - s(\pi, \nu)} = \frac{1}{1 - r(\pi, \nu)}$$

### 7.2. Proof of lemma related to the Gibbs sampler for a variance component model

*Proof of Lemma 5.1.* First note that the unnormalized joint density function of $(\vec{\theta}, V, W, \mu)$ is written as follows ($f_\pi(\vec{\theta}, V, W, \mu) = \frac{1}{L}g(\vec{\theta}, V, W, \mu)$ where $L$ is unknown),

$$\begin{aligned} g(\vec{\theta}, V, W, \mu) &= \left(\frac{b_1^{a_1}}{\Gamma(a_1)}e^{-b_1/V}V^{-a_1-1}\right)\left(\frac{b_2^{a_2}}{\Gamma(a_2)}e^{-b_2/W}W^{-a_2-1}\right) \\ &\quad \left(\frac{1}{\sqrt{2\pi b_3}}e^{-\frac{(\mu-a_3)^2}{2b_3}}\right)\left(\prod_{i=1}^{I}\frac{1}{\sqrt{2\pi V}}e^{-\frac{(\theta_i-\mu)^2}{2V}}\right)\left(\prod_{i=1}^{I}\prod_{j=1}^{J_i}\frac{1}{\sqrt{2\pi W}}e^{-\frac{(Y_{ij}-\theta_i)^2}{2W}}\right) \\ &= f_{IG}(V \mid a_1, b_1)f_{IG}(W \mid a_2, b_2)f_N(\mu \mid a_3, b_3)\left(\prod_{i=1}^{I}f_N(\theta_i|\mu, V)\right) \\ &\quad \left(\prod_{i=1}^{I}\prod_{j=1}^{J_i}f_N(Y_{ij}|\theta_i, W)\right) \end{aligned}$$

For the variance component model, we will use Theorem 3.1 and set $s = 2$ (so, $\frac{s}{s-1} = 2$). Denote $L$ as a lower bound on $\int_{\mathbb{R}^I\times\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}} g(\vec{\theta}, V, W, \mu)$ $\mathrm{d}(\vec{\theta}, V, W, \mu)$. We want to find

$$K_s(\pi, \nu) = \left(\int_{\mathbb{R}^{I+1}\times\mathbb{R}_+^2}\left(\frac{f_\pi(\vec{\theta}, V, W, \mu)}{f_\nu(\vec{\theta}, V, W, \mu)}\right)^2 f_\nu(\vec{\theta}, V, W, \mu)\mathrm{d}(\vec{\theta}, V, W, \mu)\right)^{1/2}$$

$$= \left( \int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \left( \frac{1}{L} \frac{g(\vec{\theta}, V, W, \mu)}{f_\nu(\vec{\theta}, V, W, \mu)} \right)^2 f_\nu(\vec{\theta}, V, W, \mu) \mathrm{d}(\vec{\theta}, V, W, \mu) \right)^{1/2}$$

$$= \frac{1}{L} \left( \int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \frac{g(\vec{\theta}, V, W, \mu)^2}{f_\nu(\vec{\theta}, V, W, \mu)} \mathrm{d}(\vec{\theta}, V, W, \mu) \right)^{1/2}$$

**Step 1: Find an upper bound on** $\int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \frac{g(\vec{\theta}, V, W, \mu)^2}{f_\nu(\vec{\theta}, V, W, \mu)} \mathrm{d}(\vec{\theta}, V, W, \mu)$

To find an upper bound on the integral of the ratio, note the following ratios,

$$\frac{f_{IG}(V \mid a_1, b_1)^2}{f_{IG}(V \mid 2a_1 + 1, 2b_1 - 1)} = \frac{\left( \frac{b_1^{a_1}}{\Gamma(a_1)} e^{-b_1/V} V^{-a_1 - 1} \right)^2}{\frac{(2b_1 - 1)^{2a_1 + 1}}{\Gamma(2a_1 + 1)} e^{-(2b_1 - 1)/V} V^{-(2a_1 + 1) - 1}}$$

$$= \frac{\frac{b_1^{2a_1}}{\Gamma(a_1)^2} e^{-2b_1/V} V^{-2a_1 - 2}}{\frac{(2b_1 - 1)^{2a_1 + 1}}{\Gamma(2a_1 + 1)} e^{-(2b_1 - 1)/V} V^{-2a_1 - 2}}$$

$$= \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1 + 1)}{(2b_1 - 1)^{2a_1 + 1}} e^{-1/V}$$

$$\frac{f_{IG}(W \mid a_2, b_2)^2}{f_{IG}(W \mid 2a_2 + 1, 2b_2)} = \frac{\left( \frac{b_2^{a_2}}{\Gamma(a_2)} e^{-b_2/W} W^{-a_2 - 1} \right)^2}{\frac{(2b_2)^{2a_2 + 1}}{\Gamma(2a_2 + 1)} e^{-2b_2/W} W^{-2a_2 - 1 - 1}}$$

$$= \frac{\frac{b_2^{2a_2}}{\Gamma(a_2)^2} e^{-2b_2/W} W^{-2a_2 - 2}}{\frac{(2b_2)^{2a_2 + 1}}{\Gamma(2a_2 + 1)} e^{-2b_2/W} W^{-2a_2 - 2}}$$

$$= \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2 + 1)}{(2b_2)^{2a_2 + 1}}$$

$$\frac{\prod_{i=1}^I \prod_{j=1}^{J_i} f_N(Y_{ij} \mid \theta_i, W)^2}{\prod_{i=1}^I f_N(\theta_i \mid \bar{Y}_i, \frac{W}{2J_i})} = \frac{\prod_{i=1}^I \prod_{j=1}^{J_i} \frac{1}{2\pi W} e^{-\frac{2(Y_{ij} - \theta_i)^2}{2W}}}{\prod_{i=1}^I \frac{1}{\sqrt{2\pi \frac{W}{2J_i}}} e^{-\frac{2J_i(\bar{Y}_i - \theta_i)^2}{2W}}}$$

$$= \frac{\prod_{i=1}^I \frac{1}{(2\pi W)^{J_i}} e^{-\frac{\sum_{j=1}^{J_i} (Y_{ij} - \theta_i)^2}{W}}}{\prod_{i=1}^I \frac{1}{\sqrt{\pi W / J_i}} e^{-\frac{J_i(\bar{Y}_i - \theta_i)^2}{W}}}$$

$$= \prod_{i=1}^I \frac{\frac{1}{(2\pi W)^{J_i}} e^{-\frac{\sum_{j=1}^{J_i} Y_{ij}^2 - 2Y_{ij}\theta_i + \theta_i^2}{W}}}{\frac{1}{(\pi W / J_i)^{1/2}} e^{-\frac{J_i(\bar{Y}_i)^2 - 2J_i\bar{Y}_i\theta_i + J_i\theta_i^2}{W}}}$$

$$= \prod_{i=1}^{I} \frac{1}{(J_i)^{1/2} 2^{J_i} (\pi W)^{J_i - 1/2}} e^{\frac{J_i(\bar{Y}_i)^2 - \sum_{j=1}^{J_i} Y_{ij}^2}{W}}$$

$$= \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{\sum_{i=1}^{I} J_i} (\pi W)^{\sum_{i=1}^{I} J_i - I/2}} e^{\frac{\sum_{i=1}^{I} (J_i(\bar{Y}_i)^2 - \sum_{j=1}^{J_i} Y_{ij}^2)}{W}}$$

$$= \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{\sum_{i=1}^{I} J_i} (\pi W)^{T}} e^{\frac{-S}{W}}$$

where $S = \sum_{i=1}^{I} (\sum_{j=1}^{J_i} Y_{ij}^2 - J_i(\bar{Y}_i)^2)$ and $T = \sum_{i=1}^{I} J_i - I/2$.

We evaluate the following integral, which will be used evaluating the integral of $\frac{g(\vec{\theta}, V, W, \mu)^2}{f_\nu(\vec{\theta}, V, W, \mu)}$

$$\int_{\mathbb{R}^I} \prod_{i=1}^{I} f_N(\theta_i \mid \mu, V)^2 \mathrm{d}\vec{\theta} = \prod_{i=1}^{I} \int_{\mathbb{R}} \left( \frac{1}{\sqrt{2\pi V}} e^{-\frac{(\theta_i - \mu)^2}{2V}} \right)^2 \mathrm{d}\theta_i$$

$$= \frac{1}{(2\pi V)^I} \prod_{i=1}^{I} \int_{\mathbb{R}} e^{-\frac{(\theta_i - \mu)^2}{V}} \mathrm{d}\theta_i$$

$$= \frac{1}{(2\pi V)^I} \prod_{i=1}^{I} \int_{\mathbb{R}} e^{-\frac{(\theta_i - \mu)^2}{2(V/2)}} \mathrm{d}\theta_i$$

$$= \frac{1}{(2\pi V)^I} \prod_{i=1}^{I} \sqrt{2\pi (V/2)}$$

$$= \frac{1}{(2\pi V)^I} (\pi V)^{I/2}$$

$$= \frac{1}{2^I (\pi V)^{I/2}}$$

Next, we put all of the ratios together.

We first integrate with respect to $\vec{\theta}$.

$$\int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \frac{g(\vec{\theta}, V, W, \mu)^2}{f_\nu(\vec{\theta}, V, W, \mu)} \mathrm{d}(\vec{\theta}, V, W, \mu)$$

$$= \int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \frac{(f_{IG}(V \mid a_1, b_1) f_{IG}(W \mid a_2, b_2) f_N(\mu \mid a_3, b_3) \left( \prod_{i=1}^{I} f_N(\theta_i \mid \mu, V) \right))^2}{f_{IG}(V \mid 2a_1 + 1, 2b_1 - 1) f_{IG}(W \mid 2a_2 + 1, 2b_2) f_N(\mu \mid a_3, b_3)}$$

$$\frac{\prod_{i=1}^{I} \prod_{j=1}^{J_i} f_N(Y_{ij} \mid \theta_i, W)^2}{\prod_{i=1}^{I} f_N(\theta_i \mid \bar{Y}_i, W/2J_i)} \mathrm{d}(\vec{\theta}, V, W, \mu)$$

$$= \int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1 + 1)}{(2b_1 - 1)^{2a_1 + 1}} e^{-1/V} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2 + 1)}{(2b_2)^{2a_2 + 1}}$$

$$
\begin{aligned}
& f_N(\mu \mid a_3, b_3) \prod_{i=1}^{I} f_N(\theta_i \mid \mu, V)^2 \\
& \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{\sum_{i=1}^{I} J_i} (\pi W)^T} e^{\frac{-S}{W}} \mathrm{d}(\vec{\theta}, V, W, \mu) \\
= & \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}} \\
& \int_{\mathbb{R}^1 \times \mathbb{R}_+^2} \left( \int_{\mathbb{R}^I} \prod_{i=1}^{I} f_N(\theta_i \mid \mu, V)^2 \mathrm{d}\vec{\theta} \right) e^{-1/V} f_N(\mu \mid a_3, b_3) \\
& \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{\sum_{i=1}^{I} J_i} (\pi W)^T} e^{\frac{-S}{W}} \mathrm{d}(V, W, \mu)
\end{aligned}
$$

Next we integrate with respect to $V$, $W$, $\mu$

$$
\begin{aligned}
= & \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}} \\
& \int_{\mathbb{R}^1 \times \mathbb{R}_+^2} \frac{1}{2^I (\pi V)^{I/2}} e^{-1/V} f_N(\mu \mid a_3, b_3) \\
& \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{\sum_{i=1}^{I} J_i} (\pi W)^T} e^{\frac{-S}{W}} \mathrm{d}(V, W, \mu) \\
= & \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}} \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{I+\sum_{i=1}^{I} J_i} \pi^{T+I/2}} \\
& \int_{\mathbb{R}^1 \times \mathbb{R}_+^2} \frac{1}{V^{I/2}} e^{-1/V} f_N(\mu \mid a_3, b_3) \frac{1}{W^T} e^{\frac{-S}{W}} \mathrm{d}(V, W, \mu) \\
= & \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}} \frac{1}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{I+\sum_{i=1}^{I} J_i} \pi^{T+I/2}} \\
& \Gamma(I/2-1) \frac{\Gamma(T-1)}{S^{T-1}}
\end{aligned}
$$

To summarize, we get,

$$
\begin{aligned}
C_2 &= \int_{\mathbb{R}^{I+1} \times \mathbb{R}_+^2} \frac{g(\vec{\theta}, V, W, \mu)^2}{f_v(\vec{\theta}, V, W, \mu)} \mathrm{d}(\vec{\theta}, v, w, \mu) \\
&= \frac{b_1^{2a_1}}{\Gamma(a_1)^2} \frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}} \frac{b_2^{2a_2}}{\Gamma(a_2)^2} \frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}} \\
&\quad \frac{\Gamma(I/2-1)}{(\prod_{i=1}^{I} J_i)^{1/2} 2^{I+\sum_{i=1}^{I} J_i} \pi^{T+I/2}} \frac{\Gamma(T-1)}{S^{T-1}}
\end{aligned}
$$

**Step 2: Find a lower bound on** $\int_{\mathbb{R}^I \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}} g(\vec{\theta}, v, w, \mu) \mathrm{d}(\vec{\theta}, v, w, \mu)$

To find a lower bound over the integral we estimate the above integral over a closed set $B \subset \mathbb{R}^I \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$.

$$
\begin{aligned}
g(\vec{\theta}, V, W, \mu) &= \left(\frac{b_1^{a_1}}{\Gamma(a_1)} e^{-b_1/V} V^{-a_1-1}\right)\left(\frac{b_2^{a_2}}{\Gamma(a_2)} e^{-b_2/W} W^{-a_2-1}\right) \\
&\quad \left(\frac{1}{\sqrt{2\pi b_3}} e^{-\frac{(\mu-a_3)^2}{2b_3}}\right)\left(\prod_{i=1}^{I} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(\theta_i-\mu)^2}{2V}}\right)\left(\prod_{i=1}^{I}\prod_{j=1}^{J_i} \frac{1}{\sqrt{2\pi W}} e^{-\frac{(Y_{ij}-\theta_i)^2}{2W}}\right) \\
&= \left(\frac{b_1^{a_1}}{\Gamma(a_1)} e^{-b_1/V} V^{-a_1-1}\right)\left(\frac{b_2^{a_2}}{\Gamma(a_2)} e^{-b_2/W} W^{-a_2-1}\right) \\
&\quad \left(\frac{1}{\sqrt{2\pi b_3}} e^{-\frac{(\mu-a_3)^2}{2b_3}}\right)\left(\frac{1}{(2\pi V)^{I/2}} e^{-\frac{\sum_{i=1}^{I}(\theta_i-\mu)^2}{2V}}\right) \\
&\quad \left(\frac{1}{(2\pi)^{\sum_{i=1}^{I} J_i/2} W^{\sum_{i=1}^{I} J_i/2}} e^{-\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{(Y_{ij}-\theta_i)^2}{2}\right)/W}\right) \\
&= \left(\frac{b_1^{a_1}}{\Gamma(a_1)(2\pi)^{I/2}} e^{-\left(b_1+\frac{\sum_{i=1}^{I}(\theta_i-\mu)^2}{2}\right)/V} V^{-(a_1+I/2)-1}\right)\left(\frac{1}{\sqrt{2\pi b_3}} e^{-\frac{(\mu-a_3)^2}{2b_3}}\right) \\
&\quad \left(\frac{b_2^{a_2}}{\Gamma(a_2)(2\pi)^{\sum_{i=1}^{I} J_i/2}} W^{-(a_2+\sum_{i=1}^{I} J_i/2)-1} e^{-\left(b_2+\sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{(Y_{ij}-\theta_i)^2}{2}\right)/W}\right)
\end{aligned}
$$

Define

$$
\begin{aligned}
a_1^* &= a_1 + I/2 & b_1^* &= b_1 + \frac{\sum_{i=1}^{I}(\theta_i-\mu)^2}{2} \\
a_2^* &= a_2 + \sum_{i=1}^{I} J_i/2 & b_2^* &= b_2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{(Y_{ij}-\theta_i)^2}{2}
\end{aligned}
$$

Taking the integral of $g$ with respect to $V$ and $W$, we get,

$$
\begin{aligned}
&\int_{\mathbb{R}^I\times\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}} g(\vec{\theta}, v, w, \mu)\mathrm{d}(\vec{\theta}, v, w, \mu) \\
&= \int_{\mathbb{R}^I\times\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}} \left(\frac{b_1^{a_1}}{\Gamma(a_1)(2\pi)^{I/2}} e^{-b_1^*/V} V^{-a_1^*-1}\right)\left(\frac{1}{\sqrt{2\pi b_3}} e^{-\frac{(\mu-a_3)^2}{2b_3}}\right) \\
&\left(\frac{b_2^{a_2}}{\Gamma(a_2)(2\pi)^{\sum_{i=1}^{I} J_i/2}} W^{-a_2^*-1} e^{-b_2^*/W}\right)\mathrm{d}(\vec{\theta}, v, w, \mu) \\
&= \int_{\mathbb{R}^{I+1}} \left(\frac{\Gamma(a_1^*) b_1^{a_1}}{\Gamma(a_1)(2\pi)^{I/2}} \left(b_1^*\right)^{-a_1^*}\right)\left(\frac{1}{\sqrt{2\pi b_3}} e^{-\frac{(\mu-a_3)^2}{2b_3}}\right) \\
&\left(\frac{\Gamma(a_2^*) b_2^{a_2}}{\Gamma(a_2)(2\pi)^{\sum_{i=1}^{I} J_i/2}} \left(b_2^*\right)^{-a_2^*}\right)\mathrm{d}(\vec{\theta}, \mu)
\end{aligned}
$$

The integral can then be defined as follows,

$$\begin{aligned}
&\int_{\mathbb{R}^I\times\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}} g(\vec{\theta}, v, w, \mu)\mathrm{d}(\vec{\theta}, v, w, \mu)\\
&= \frac{\Gamma(a_1^*)b_1^{a_1}}{\Gamma(a_1)}\frac{\Gamma(a_2^*)b_2^{a_2}}{\Gamma(a_2)}\frac{1}{(2\pi)^{\sum_{i=1}^I J_i/2+(I+1)/2}b_3^{1/2}}\\
&\int_{\mathbb{R}^{I+1}} (b_1^*)^{-a_1^*}e^{-\frac{(\mu-a_3)^2}{2b_3}}\left(b_2^*\right)^{-(a_2^*)}\mathrm{d}(\vec{\theta}, \mu)\\
&= C_1^{-1}\int_{\mathbb{R}^{I+1}} (b_1^*)^{-a_1^*}e^{-\frac{(\mu-a_3)^2}{2b_3}}\left(b_2^*\right)^{-(a_2^*)}\mathrm{d}(\vec{\theta}, \mu)
\end{aligned}$$

Where $C_1^{-1} = \frac{\Gamma(a_1^*)b_1^{a_1}}{\Gamma(a_1)}\frac{\Gamma(a_2^*)b_2^{a_2}}{\Gamma(a_2)}\frac{1}{(2\pi)^{\sum_{i=1}^I J_i/2+(I+1)/2}b_3^{1/2}}$

Combining step 1 and step 2, we get,

$$\frac{1}{\int_B g(\vec{\theta}, V, W, \mu)\mathrm{d}(\vec{\theta}, V, W, \mu)}\left(\int_{\mathbb{R}^{I+1}\times\mathbb{R}_+^2}\frac{g(\vec{\theta}, V, W, \mu)^2}{f_\nu(\vec{\theta}, V, W, \mu)}\mathrm{d}(\vec{\theta}, v, w, \mu)\right)^{1/2}$$
$$= \frac{1}{L}C_1C_2^{1/2}$$

Where $L = \int_{\mathbb{R}^{I+1}}(b_1^*)^{-a_1^*}e^{-\frac{(\mu-a_3)^2}{2b_3}}\left(b_2^*\right)^{-(a_2^*)}\mathrm{d}(\vec{\theta}, \mu)$

$C_1 = \frac{\Gamma(a_1)}{\Gamma(a_1^*)b_1^{a_1}}\frac{\Gamma(a_2)}{\Gamma(a_2^*)b_2^{a_2}}(2\pi)^{\sum_{i=1}^I J_i/2+(I+1)/2}b_3^{1/2}$

$C_2 = \frac{b_1^{2a_1}}{\Gamma(a_1)^2}\frac{\Gamma(2a_1+1)}{(2b_1-1)^{2a_1+1}}\frac{b_2^{2a_2}}{\Gamma(a_2)^2}\frac{\Gamma(2a_2+1)}{(2b_2)^{2a_2+1}}\frac{\Gamma(I/2-1)}{(\prod_{i=1}^I J_i)^{1/2}2^{I+\sum_{i=1}^I J_i}\pi^{T+I/2}}\frac{\Gamma(T-1)}{S^{T-1}}$ □

### 7.3. Proof of lemma related to the Gibbs sampler for a James-Stein estimator

*Proof of Lemma 4.1.* First note that the density function of $\nu$ is as follows,

$$\begin{aligned}
&f_\nu(\vec{\theta}, \mu, A)\\
&= \left(\prod_{i=1}^q \frac{1}{\sqrt{2\pi V}}e^{-(Y_i-\theta_i)^2/(2V)}\right)\left(\frac{\beta^{\alpha+(q-1)/2}}{\Gamma(\alpha+(q-1)/2)}A^{-\alpha-(q-1)/2-1}e^{-\beta/A}\right)\\
&\left(\frac{1}{\sqrt{2\pi A}}e^{-(\mu-\bar{\theta})/(2A)}\right)
\end{aligned}$$

Next, we want to calculate the supremum of the ratios where the function $g$ is defined as in Equation 6.

$$\begin{aligned}
&\sup_{(\vec{\theta},\mu,A)\in\mathbb{R}^{q+1}\times\mathbb{R}_+}\frac{g(\vec{\theta}, \mu, A)}{f_\nu(\vec{\theta}, \mu, A)}\\
&= \sup_{(\vec{\theta},\mu,A)\in\mathbb{R}^{q+1}\times\mathbb{R}_+}\frac{\left(\prod_{i=1}^q \frac{1}{\sqrt{2\pi A}}e^{-(\theta_i-\mu)^2/(2A)}\right)\frac{\beta^\alpha}{\Gamma(\alpha)}A^{-\alpha-1}e^{-\beta/A}}{\left(\frac{1}{\sqrt{2\pi A}}e^{-(\mu-\bar{\theta})^2/(2A)}\right)\frac{\beta^{\alpha+(q-1)/2}}{\Gamma(\alpha+(q-1)/2)}A^{-\alpha-q/2+1/2-1}e^{-\beta/A}}
\end{aligned}$$

$$= \frac{\Gamma(\alpha + (q-1)/2)}{\Gamma(\alpha)} \frac{1}{(2\pi\beta)^{(q-1)/2}} \sup_{(\vec{\theta},\mu,A)\in\mathbb{R}^{q+1}\times\mathbb{R}_+} e^{[(\bar{\theta}-\mu)^2 - \sum_{i=1}^{q}(\theta_i-\mu)^2]/(2A)}$$
$$= \frac{\Gamma(\alpha + (q-1)/2)}{\Gamma(\alpha)(2\pi\beta)^{(q-1)/2}}$$

The last equality follows since $(\bar{\theta}-\mu)^2 \leq \sum_{i=1}^{q}(\theta_i-\mu)^2$ by Jensen's inequality. Next we partially calculate $\int_{\mathbb{R}^{q+1}\times\mathbb{R}_+} g(\vec{\theta}, \mu, A \mid Y)\mathrm{d}(\vec{\theta}, \mu, A)$.

$$\begin{aligned}
&\int_{\mathbb{R}^{q+1}\times\mathbb{R}_+} g(\vec{\theta}, \mu, A \mid Y)\mathrm{d}(\vec{\theta}, \mu, A) \\
&= \int_{\mathbb{R}^{q+1}\times\mathbb{R}_+} \frac{\beta^\alpha}{\Gamma(\alpha)} A^{-\alpha-1} e^{-\beta/A} \frac{1}{(2\pi V)^{q/2}(2\pi A)^{q/2}} \\
&\quad \times \left( \prod_{i=1}^{q} e^{-(\theta_i - Y_i)^2/(2V) - (\theta_i-\mu)^2/(2A)} \right) \mathrm{d}(\vec{\theta}, \mu, A) \\
&= \int_{\mathbb{R}^{q+1}\times\mathbb{R}_+} \frac{\beta^\alpha}{\Gamma(\alpha)} A^{-\alpha-1} e^{-\beta/A} \frac{1}{(2\pi)^q (AV)^{q/2}} \\
&\quad \times \left( \prod_{i=1}^{q} e^{-0.5[(\theta_i^2 - 2\theta_i Y_i + Y_i^2)/V + (\theta_i^2 - 2\theta_i\mu + \mu^2)/A]} \right) \mathrm{d}(\vec{\theta}, \mu, A) \\
&= \int_{\mathbb{R}^{q+1}\times\mathbb{R}_+} \frac{\beta^\alpha}{\Gamma(\alpha)} A^{-\alpha-1} e^{-\beta/A} \frac{1}{(2\pi)^q (AV)^{q/2}} \\
&\quad \times \left( \prod_{i=1}^{q} e^{-0.5\left[\theta_i^2\left(\frac{1}{V}+\frac{1}{A}\right) - 2\theta_i\left(\frac{Y_i}{V}+\frac{\mu}{A}\right) + \left(\frac{Y_i^2}{V}+\frac{\mu^2}{A}\right)\right]} \right) \mathrm{d}(\vec{\theta}, \mu, A) \\
&= \int_{\mathbb{R}^{q+1}\times\mathbb{R}_+} \frac{\beta^\alpha}{\Gamma(\alpha)} A^{-\alpha-1} e^{-\beta/A} \frac{e^{-\frac{1}{2}\sum_{i=1}^{q}\left(\frac{Y_i^2}{V}+\frac{\mu^2}{A}\right)}}{(2\pi)^q (AV)^{q/2}} \\
&\quad \times \left( \prod_{i=1}^{q} e^{-0.5\left(\frac{1}{V}+\frac{1}{A}\right)\left[\theta_i^2 - 2\theta_i \frac{\frac{Y_i}{V}+\frac{\mu}{A}}{\frac{1}{V}+\frac{1}{A}}\right]} \right) \mathrm{d}(\vec{\theta}, \mu, A) \\
&= \int_{\mathbb{R}\times\mathbb{R}_+} \frac{\beta^\alpha}{\Gamma(\alpha)} A^{-\alpha-1} e^{-\beta/A} \frac{1}{(2\pi)^q (AV)^{q/2}} e^{-\frac{1}{2}\sum_{i=1}^{q}\left(\frac{Y_i^2}{V}+\frac{\mu^2}{A}\right)} \\
&\quad \times e^{\frac{1}{2}\sum_{i=1}^{q} \frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}}} \left( 2\pi \frac{1}{(\frac{1}{V}+\frac{1}{A})} \right)^{q/2} \mathrm{d}(\mu, A) \\
&= \int_{\mathbb{R}\times\mathbb{R}_+} \frac{\beta^\alpha}{\Gamma(\alpha)} A^{-\alpha-1} e^{-\beta/A} \frac{1}{(2\pi)^q (AV)^{q/2}} e^{-\frac{1}{2}\sum_{i=1}^{q}\left(\frac{Y_i^2}{V}+\frac{\mu^2}{A}\right)}
\end{aligned}$$

$$
\begin{aligned}
&\times e^{\frac{1}{2}\sum_{i=1}^{q}\frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}}}\left(2\pi\frac{AV}{A+V}\right)^{q/2}\mathrm{d}(\mu,A)\\
&=\int_{\mathbb{R}\times\mathbb{R}_+}\frac{\beta^\alpha}{\Gamma(\alpha)}A^{-\alpha-1}e^{-\beta/A}\frac{1}{(2\pi)^{q/2}(A+V)^{q/2}}\\
&\quad\times e^{-\frac{1}{2}\sum_{i=1}^{q}\left(\frac{Y_i^2}{V}+\frac{\mu^2}{A}\right)}e^{\frac{1}{2}\sum_{i=1}^{q}\frac{(\frac{Y_i}{V}+\frac{\mu}{A})^2}{\frac{1}{V}+\frac{1}{A}}}\mathrm{d}(\mu,A)\\
&=\int_{\mathbb{R}\times\mathbb{R}_+}\frac{\beta^\alpha}{\Gamma(\alpha)}A^{-\alpha-1}e^{-\beta/A}\frac{1}{(2\pi)^{q/2}(A+V)^{q/2}}\\
&\quad\times e^{\frac{1}{2}\sum_{i=1}^{q}\left[\frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}}-\frac{Y_i^2}{V}-\frac{\mu^2}{A}\right]}\mathrm{d}(\mu,A)
\end{aligned}
$$

Thus we get that

$$
\begin{aligned}
&\frac{1}{\int_{\mathbb{R}^{q+1}\times\mathbb{R}_+}g(\vec{\theta},\mu,A\mid Y)}\mathrm{d}(\vec{\theta},\mu,A)\\
&=\frac{\Gamma(\alpha)(2\pi)^{q/2}}{\beta^\alpha}\frac{1}{\int_{\mathbb{R}\times\mathbb{R}_+}\frac{A^{-\alpha-1}e^{-\beta/A}}{(A+V)^{q/2}}e^{\frac{1}{2}\sum_{i=1}^{q}\left[\frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}}-\frac{Y_i^2}{V}-\frac{\mu^2}{A}\right]}\mathrm{d}(\mu,A)}
\end{aligned}
$$

So

$$
\begin{aligned}
K_1(\pi,\nu)&=\frac{1}{\int_B g(\vec{\theta},\mu,A)\mathrm{d}(\vec{\theta},\mu,A)}\sup_{(\vec{\theta},\mu,A)\in\mathbb{R}^{q+1}\times\mathbb{R}_+}\frac{g(\vec{\theta},\mu,A)}{f_\nu(\vec{\theta},\mu,A)}\\
&\le\frac{\Gamma(\alpha)(2\pi)^{q/2}}{\beta^\alpha}\frac{1}{\int_{\mathbb{R}\times\mathbb{R}_+}\frac{A^{-\alpha-1}e^{-\beta/A}}{(A+V)^{q/2}}e^{\frac{1}{2}\sum_{i=1}^{q}\left[\frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}}-\frac{Y_i^2}{V}-\frac{\mu^2}{A}\right]}\mathrm{d}(\mu,A)}\\
&\quad\frac{\Gamma(\alpha+(q-1)/2)}{\Gamma(\alpha)(2\pi\beta)^{(q-1)/2}}\\
&=\frac{\Gamma(\alpha+(q-1)/2)(2\pi)^{1/2}}{\beta^{\alpha+(q-1)/2}}\\
&\quad\frac{1}{\int_{\mathbb{R}\times\mathbb{R}_+}\frac{A^{-\alpha-1}e^{-\beta/A}}{(A+V)^{q/2}}e^{\frac{1}{2}\sum_{i=1}^{q}\left[\frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}}-\frac{Y_i^2}{V}-\frac{\mu^2}{A}\right]}\mathrm{d}(\mu,A)}
\end{aligned}
$$

Let $L = \int_{\mathbb{R}\times\mathbb{R}_+} \frac{A^{-\alpha-1}e^{-\beta/A}}{(A+V)^{q/2}} e^{\frac{1}{2}\sum_{i=1}^{q}\left[\frac{\left(\frac{Y_i}{V}+\frac{\mu}{A}\right)^2}{\frac{1}{V}+\frac{1}{A}} - \frac{Y_i^2}{V} - \frac{\mu^2}{A}\right]} \mathrm{d}(\mu, A)$. Then by Equation 3 coupled with Remark 1, for $r \geq p$

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq \left(\frac{\Gamma(\alpha + (q-1)/2)(2\pi)^{1/2}}{\beta^{\alpha+(q-1)/2}} \frac{1}{L} E[d(X_n, Y_n)^r]\right)^{1/r}$$

□

### 7.4. Proof of lemmas related to the Gibbs sampler of a Bayesian linear regression model

*Proof of Lemma 6.1.* It follows from the de-initializing property of the Markov chain [45, Example 3] that

$$\left\|\mathcal{L}(\beta_{n+1}, \sigma_{n+1}^2) - \mathcal{L}(\beta'_{n+1}, \sigma_{n+1}^{'2})\right\|_{\mathrm{TV}} \leq \left\|\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_n^{'2})\right\|_{\mathrm{TV}}.$$

Using CRN, let $G_n = G'_n$ where $G_n \sim \mathrm{Gamma}(\alpha, 1)$ $Z_n \sim N(0, I_q)$ are independent and denote

$$W_n = \frac{\upsilon_0 c_0^2}{2} + \frac{\left\|X\tilde{\beta}_{\sigma_{n-1}^2} - Y + XV^{1/2}_{\sigma_{n-1}^2} Z_n\right\|^2}{2}$$

$$W'_n = \frac{\upsilon_0 c_0^2}{2} + \frac{\left\|X\tilde{\beta}_{\sigma_{n-1}^{'2}} - Y + XV^{1/2}_{\sigma_{n-1}^{'2}} Z_n\right\|^2}{2},$$

so that $\sigma_n^2|\sigma_{n-1}^2 = W_n\frac{1}{G_n}$ and $\sigma_n^{'2}|\sigma_{n-1}^{'2} = W'_n\frac{1}{G'_n}$. Denote $\Delta = W'_n - W_n$ and without loss of generality, assume $W'_n > W_n$. Since $G_n \sim \mathrm{Gamma}(\alpha, 1)$ where $\alpha = \frac{k+\upsilon_0}{2}$, let $\pi_{1/G_n}$ denote the density of $1/G_n$ and similarly denote the density $\pi_{(1+\Delta/W_n)/G_n}$ for $(1 + \Delta/W_n)/G_n$. So we have

$$\pi_{1/G_n}(x) \propto x^{-\alpha-1}e^{-1/x}$$

$$\pi_{(1+\Delta/W_n)/G_n}(x) \propto \frac{1}{1+\Delta/W_n}\left(\frac{x}{1+\Delta/W_n}\right)^{-\alpha-1} e^{-(1+\Delta/W_n)/x}$$

$$\propto (1+\Delta/W_n)^{\alpha} x^{-\alpha-1} e^{-(1+\Delta/W_n)/x}.$$

Using the coupling characterization of total variation

$$||\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_n^{'2})||_{TV}$$
$$\leq E\left[||\mathcal{L}(W_n/G_n \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}) - \mathcal{L}(W'_n/G_n \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2})||_{TV}\right]$$

By Proposition 2.2 of [54]

$$= E\left[||\mathcal{L}(W_n/G_n \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}) - \mathcal{L}((W_n+\Delta)/G_n \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2})||_{TV}\right]$$

$$= E\left[||\mathcal{L}\left(\frac{1}{G_n} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right) - \mathcal{L}\left(\left(1+\frac{\Delta}{W_n}\right)\frac{1}{G_n} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right)||_{TV}\right]$$

By Proposition 2.1 of [54]

$$\leq E\left[E\left[\sup_{x>0}\left\{1 - \frac{\pi_{1/G_n}(x)}{\pi_{(1+\Delta/W_n)/G_n}(x)}\right\} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right]\right]$$

By Lemma 6.16 of of [34]

This implies that,

$$||\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_n^{'2})||_{TV} \leq E\left[\sup_{x>0}\left\{1 - \frac{x^{-\alpha-1}e^{-1/x}}{(1+\Delta/W_n)^{\alpha}x^{-\alpha-1}e^{-(1+\Delta/W_n)/x}}\right\}\right]$$

$$= E\left[\sup_{x>0}\left\{1 - \frac{e^{\Delta/W_n/x}}{(1+\Delta/W_n)^{\alpha}}\right\}\right]$$

$$= E\left[1 - \frac{1}{(1+\Delta/W_n)^{\alpha}}\right].$$

Define $f(x) = \frac{1}{x^{\alpha}}$, $f'(x) = -\alpha\frac{1}{x^{\alpha+1}}$. By the mean value theorem, $f(1+\Delta/W_n) = f(1) - \frac{\Delta}{W_n}\frac{\alpha}{\xi^{\alpha+1}}$, $\xi \in (1, 1+\Delta/W_n)$. So, $f(1+\Delta/W_n) \geq 1 - \alpha\frac{\Delta}{W_n}$. Now,

$$\begin{aligned}
&||\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_n^{'2})||_{TV} \\
&\leq E\left[f(1) - f(1+\frac{\Delta}{W_n})\right] \\
&\leq E\left[1 - (1-\alpha\frac{\Delta}{W_n})\right] = E\left[\alpha\frac{\Delta}{W_n}\right] \\
&\leq E\left[\frac{k+\upsilon_0}{2}\Delta\frac{2}{\upsilon_0 c_0^2}\right] \text{since } \alpha = \frac{k+\upsilon_0}{2} \text{ and } W_n \geq \frac{\upsilon_0 c_0^2}{2} \\
&= \frac{k+\upsilon_0}{\upsilon_0 c_0^2}E[|W_n - W_n'|] \\
&= \frac{k+\upsilon_0}{\upsilon_0 c_0^2}E[|W_n - W_n'|]E[G_n^{-1}]E[G_n^{-1}]^{-1} \\
&= \frac{k+\upsilon_0}{\upsilon_0 c_0^2}E[|W_n/G_n - W_n'/G_n|]E[G_n^{-1}]^{-1} \text{ by independence} \\
&\leq \frac{(k+\upsilon_0)^2}{2\upsilon_0 c_0^2}E[|W_n/G_n - W_n'/G_n|] \\
&= \frac{(k+\upsilon_0)^2}{2\upsilon_0 c_0^2}E[|\sigma_n^2 - \sigma_n^{'2}|].
\end{aligned}$$

□

*Proof of Lemma 6.2.* Define $\alpha' = (k+\upsilon)/2 - 2$ and $\beta' = \upsilon_0 c_0^2/2$. Note that

$$f_\nu(\beta,\sigma^2) = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')}\frac{1}{(\sigma^2)^{\alpha'+1}}e^{-\beta'/\sigma^2} \times (2\pi)^{-q/2} \det(\Sigma_\beta)^{-1/2} e^{-\frac{1}{2}(\beta-\beta_0)^T\Sigma_\beta^{-1}(\beta-\beta_0)}$$

This implies that,

$$\begin{aligned}
&\sup_{(\beta,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+}\frac{g(\beta,\sigma^2)}{f_\nu(\beta,\sigma^2)} \\
&= \frac{\frac{1}{(\sigma^2)^{(k+\upsilon_0)/2+1}}\exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta) - \frac{1}{2}(\beta-\beta_0)^T\Sigma_\beta^{-1}(\beta-\beta_0) - \frac{\upsilon_0 c_0^2}{2\sigma^2}\right)}{\frac{\beta'^{\alpha'}}{\Gamma(\alpha')}\frac{1}{(\sigma^2)^{(k+\upsilon)/2-2+1}}e^{-\upsilon_0 c_0^2/2\sigma^2}(2\pi)^{-q/2}\det(\Sigma_\beta)^{-1/2}e^{-\frac{1}{2}(\beta-\beta_0)^T\Sigma_\beta^{-1}(\beta-\beta_0)}} \\
&= \frac{(\sigma^2)^{-2}\exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right)}{\frac{\beta'^{\alpha'}}{\Gamma(\alpha')}(2\pi)^{-q/2}\det(\Sigma_\beta)^{-1/2}}
\end{aligned}$$

The numerator of the last equality is proportional to an inverse gamma density function with $\alpha^* = 1$ and $\beta^* = \frac{1}{2}(Y-X\beta)^T(Y-X\beta)$. Thus, the mode occurs when $\sigma^{2*} = \frac{\beta^*}{\alpha^*+1} = \frac{(Y-X\beta)^T(Y-X\beta)}{4}$ and so,

$$\begin{aligned}
&\sup_{(\beta,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+}\frac{g(\beta,\sigma^2)}{f_\nu(\beta,\sigma^2)} \\
&\le \frac{16}{((Y-X\beta)^T(Y-X\beta))^2}\frac{\exp\left(-\frac{4}{2(Y-X\beta)^T(Y-X\beta)}(Y-X\beta)^T(Y-X\beta)\right)}{\frac{\beta'^{\alpha'}}{\Gamma(\alpha')}(2\pi)^{-q/2}\det(\Sigma_\beta)^{-1/2}} \\
&= \frac{16}{((Y-X\beta)^T(Y-X\beta))^2}\frac{\Gamma(\alpha')(2\pi)^{q/2}\det(\Sigma_\beta)^{1/2}}{\beta'^{\alpha'}e^2} \\
&\le \frac{16}{((Y-X\hat\beta)^T(Y-X\hat\beta))^2}\frac{\Gamma(\alpha')(2\pi)^{q/2}\det(\Sigma_\beta)^{1/2}}{\beta'^{\alpha'}e^2}
\end{aligned}$$

Where $\hat\beta = (X^TX)^{-1}X^TY$ is the maximum likelihood estimator. Next we find a lower bound on $\int_B g(\beta,\sigma^2)\mathrm{d}(\beta,\sigma^2)$

$$\begin{aligned}
&\int_{\mathbb{R}^q\times\mathbb{R}_+} g(\beta,\sigma^2)\mathrm{d}(\beta,\sigma^2) \\
&= \int_{\mathbb{R}^q\times\mathbb{R}_+}\frac{1}{(\sigma^2)^{(k+\upsilon_0)/2+1}} \\
&\exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta) - \frac{1}{2}(\beta-\beta_0)^T\Sigma_\beta^{-1}(\beta-\beta_0) - \frac{\upsilon_0 c_0^2}{2\sigma^2}\right)\mathrm{d}(\beta,\sigma^2)
\end{aligned}$$

$$= \int_{\mathbb{R}^q} \left( \int_{\mathbb{R}_+} \frac{1}{(\sigma^2)^{(k+\upsilon_0)/2+1}} e^{-(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)/\sigma^2} \mathrm{d}(\sigma^2) \right)$$
$$e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)} \mathrm{d}(\beta)$$
$$= \int_{\mathbb{R}^q} \frac{\Gamma((k+\upsilon_0)/2)}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}} e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)} \mathrm{d}(\beta)$$
$$= \Gamma((k+\upsilon_0)/2) \int_{\mathbb{R}^q} \frac{e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)}}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}} \mathrm{d}(\beta)$$
$$\leq \Gamma((k+\upsilon_0)/2) \int_B \frac{e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)}}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}} \mathrm{d}(\beta)$$

where $B \subset \mathbb{R}^q \times \mathbb{R}_+$ is a bounded set.

Thus we get that

$$K_1(\pi, \nu) = \frac{1}{\int_{\mathbb{R}^q \times \mathbb{R}_+} g(\beta, \sigma^2) \mathrm{d}(\beta, \sigma^2)} \sup_{(\beta,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+} \frac{g(\beta, \sigma^2)}{f_\nu(\beta, \sigma^2)}$$
$$\leq \frac{1}{\Gamma((k+\upsilon_0)/2) \int_B \frac{e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)}}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}} \mathrm{d}(\beta)}$$
$$\frac{16\Gamma((k+\upsilon_0)/2-2)(2\pi)^{q/2} \det(\Sigma_\beta)^{1/2}}{((Y-X\hat{\beta})^T(Y-X\hat{\beta}))^2 \left(\upsilon_0 c_0^2/2\right)^{(k+\upsilon_0)/2-2} e^2}$$
$$= \frac{1}{\int_B \frac{e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)}}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}} \mathrm{d}(\beta)}$$
$$\frac{16\Gamma((k+\upsilon_0)/2-2)(2\pi)^{q/2} \det(\Sigma_\beta)^{1/2}}{\Gamma((k+\upsilon_0)/2)((Y-X\hat{\beta})^T(Y-X\hat{\beta}))^2 \left(\upsilon_0 c_0^2/2\right)^{(k+\upsilon_0)/2-2} e^2}$$

So by Equation 3 combined with Theorem 3.2, for $r \geq p$

$$W_{d,p}(\mathcal{L}(X_n), \pi) \leq \left( \frac{C}{L} E[d(X_n, Y_n)^r] \right)^{1/r}$$

Where $C = \frac{16\Gamma((k+\upsilon_0)/2-2)(2\pi)^{q/2} \det(\Sigma_\beta)^{1/2}}{\Gamma((k+\upsilon_0)/2)((Y-X\hat{\beta})^T(Y-X\hat{\beta}))^2 \left(\upsilon_0 c_0^2/2\right)^{(k+\upsilon_0)/2-2} e^2}$

and $L = \int_B \frac{e^{-\frac{1}{2}(\beta-\beta_0)^T \Sigma_\beta^{-1}(\beta-\beta_0)}}{(\upsilon_0 c_0^2/2+(Y-X\beta)^T(Y-X\beta)/2)^{(k+\upsilon_0)/2}} \mathrm{d}(\beta)$ □

**Remark 3.** For Real Data Example 6.1, we want to prove that $E[||(\beta_n, \sigma_n^2) - (\beta_n', \sigma_n'^2)||_1] < \infty$ for $n \in \mathbb{N}$.

Since $\sigma_0^2 = 100$ and $\sigma_0'^2 \sim IG((k+\upsilon_0)/2, \upsilon_0 c_0^2/2)$, $E[|\sigma_0^2|] < \infty$ and $E[|\sigma_0'^2|] < \infty$.

Suppose that $E[|\sigma_n^2|] < \infty$. By Equation 14, $\sigma_n^2 = k_{Z_n, IG_n}(\sigma_{n-1}^2)$, so $E[|\sigma_n^2|] = E[E[k_{Z_n, IG_n}(\sigma_{n-1}^2)|\sigma_{n-1}^2]] < \infty$.

Further, since $\beta_n \sim N_q(\tilde{\beta}_{\sigma_n^2}, V_{\beta,\sigma_n^2})$ and $\beta_n' \sim N_q(\tilde{\beta}_{\sigma_n'^2}, V_{\beta,\sigma_n'^2})$, $E[|\beta_n|] < \infty$ and $E[|\beta_n'|] < \infty$.

Thus, we have established that $E[||(\sigma_n^2, \beta_n)||_1] < \infty$ and $E[||(\sigma_n'^2, \beta_n')||_1] < \infty$. By the triangle inequality, $E[||(\sigma_n^2, \beta_n) - (\sigma_n'^2, \beta_n')||_1] < \infty$.

## Acknowledgements

We thank the referees for their many excellent comments and suggestions.

## Author contributions

CRediT: **Sabrina Sixta**: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Jeffrey S. Rosenthal**: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing; **Austin Brown**: Supervision, Validation, Writing – review & editing.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This research was partially funded by the National Science and Engineering Research Council of Canada (NSERC).

## ORCID

Sabrina Sixta http://orcid.org/0000-0001-5410-8226
Jeffrey S. Rosenthal http://orcid.org/0000-0002-5118-6808
Austin Brown http://orcid.org/0000-0003-1576-8381

## References

[1] Abrahamsen, T.; Hobert, J. P. Convergence analysis of block Gibbs samplers for Bayesian linear mixed models with PN. *Bernoulli* **2017**, *23*, 459–478. DOI: 10.3150/15-BEJ749.
[2] Agapiou, S.; Papaspiliopoulos, O.; Sanz-Alonso, D.; Stuart, A. M. Importance sampling: intrinsic dimension and computational cost. *Statistical Science* **2017**, *32*, 405–431. DOI: 10.1214/17-STS611.
[3] Aldous, D.; Diaconis, P. Strong uniform times and finite random walks. *Adv. Appl. Math.* **1987**, *8*, 69–97. DOI: 10.1016/0196-8858(87)90006-6.
[4] Julio Backhoff-Veraguas, Sigrid Källblad, and Benjamin A. Robinson. Adapted wasserstein distance between the laws of sdes, **2024**.
[5] Backlund, G.; Hobert, J. P. A note on the convergence rate of MCMC for robust Bayesian multivariate linear regression with proper priors. *Comput. Math. Methods* **2020**, *2*, e1086. DOI: 10.1002/cmm4.1086.

[6] Biswas, N.; Bhattacharya, A.; Jacob, P. E.; Johndrow, J. E. Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 03**2022**, *84*, 973–996. DOI: 10.1111/rssb.12495.

[7] Biswas, N.; Mackey, L. Bounding Wasserstein distance with couplings. *J. Am. Stat. Assoc.* **2024**, *119*, 2947–2958.): DOI: 10.1080/01621459.2023.2287773.

[8] Bou-Rabee, N.; Eberle, A.; Zimmer, R. Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability* **2020**, *30*, 1209–1250. DOI: 10.1214/19-AAP1528.

[9] Efron, C. M. B. Data analysis using Stein's estimator and its generalizations. *J. Am. Stat. Assoc.* **1975**, *70*, 311–319.

[10] Burdzy, K.; Chen, Z.-Q.; Jones, P. Synchronous couplings of reflected Brownian motions in smooth domains. *Illinois Journal of Mathematics* **2006**, *50*, 189–268. DOI: 10.1215/ijm/1258059475.

[11] Caputo, P.; Labbé, C.; Lacoin, H. Mixing time of the adjacent walk on the simplex. *The Annals of Probability* **2020**, *48*, 2449–2493. DOI: 10.1214/20-AOP1428.

[12] Cowles, M. K.; Rosenthal, J. S. A simulation approach to convergence rates for Markov Chain Monte Carlo algorithms. *Stat. Comput.* **1998**, *8*, 115–124. DOI: 10.1023/A:1008982016666.

[13] Cowles, M. K.; Carlin, B. P. Markov Chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* **1996**, *91*, 883–904. DOI: 10.1080/01621459.1996.10476956.

[14] Owen L. Davies. *Statistical methods in research and production*. Oliver and Boyd, Edinburgh and London, 3 edition, **1967**.

[15] Davis, B.; Hobert, J. P. On the convergence complexity of Gibbs samplers for a family of simple Bayesian random effects models. *Methodol. Comput. Appl. Probab.* **2021**, *23*, 1323–1351. DOI: 10.1007/s11009-020-09808-8.

[16] Deligiannidis, G.; Jacob, P. E.; Khribch, E. M.; Wang, G. On importance sampling and independent metropolis-hastings with an unbounded weight function. **2024**.

[17] Dobson, A. J.; Barnett, A. G. *An Introduction to Generalized Linear Models. Texts in Statistical Science*. Chapman and Hall, New York, 3 edition, **2008**.

[18] Eberle, A.; Guillin, A.; Zimmer, R. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability* **2019**, *47*, 1982–2010. DOI: 10.1214/18-AOP1299.

[19] Oskar Ekvall, K.; Jones, G. L. Convergence analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions. *Electron. J. Stat.* **2021**, *15*, 691–721. DOI: 10.1214/21-EJS1800.

[20] Geyer, C. J. *Introduction to Markov Chain Monte Carlo*, pages 1–46 Chapman and Hall/CRC, New York, **2011**.

[21] Gibbs, A. L. Convergence in the Wasserstein Metric for Markov Chain Monte Carlo algorithms with applications to image restoration. *Stoch. Model.* **2004**, *20*, 473–492. DOI: 10.1081/STM-200033117.

[22] Glasserman, P.; Yao, D. D. Some guidelines and guarantees for common random numbers. *Manage. Sci.* **1992**, *38*, 884–908. DOI: 10.1287/mnsc.38.6.884.

[23] Heng, J.; Jacob, P. E. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* **2019**, *106*, 287–302. DOI: 10.1093/biomet/asy074.

[24] Hobert, J. P.; Geyer, C. J. Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *J. Multivariate Anal.* **1998**, *67*, 414–430. DOI: 10.1006/jmva.1998.1778.

[25] Hobert, J. P.; Jones, G. L. Honest exploration of intractable probability distributions via Markov Chain Monte Carlo. *Statistical Science* **2001**, *16*, 312–334. DOI: 10.1214/ss/1015346317.

[26] Hobert, J. P.; Jung, Y. J.; Khare, K.; Qin, Q. Convergence analysis of MCMC algorithms for Bayesian multivariate linear regression with non-Gaussian errors. *Scand. J. Stat.* **2018**, *45*, 513–533. DOI: 10.1111/sjos.12310.

[27] Hoff, P. D. *A First Course in Bayesian Statistical Methods*. Springer, New York, **2009**.

[28] Jackman, S. Pscl: Classes and methods for R developed in the Political Science Computational Laboratory. *R Package Version* **2024**, *1*.

[29] Jacob, P. E. Lecture notes for couplings and Monte Carlo. Available at https://sites.google.com/site/pierrejacob/cmclectures?authuser=0. **2021**/09/17).

[30] Jin, Z.; Hobert, J. P. Dimension free convergence rates for Gibbs samplers for Bayesian linear mixed models. *Stochastic Processes Appl.* **2022**, *148*, 25–67. DOI: 10.1016/j.spa.2022.02.003.

[31] Johnson, A. A.; Jones, G. L. Geometric ergodicity of random scan Gibbs samplers for hierarchical one-way random effects models. *J. Multivariate Anal.* **2015**, *140*, 325–342. DOI: 10.1016/j.jmva.2015.06.002.

[32] Johnson, V. E. Studying convergence of Markov Chain Monte Carlo algorithms using coupled sample paths. *J. Am. Stat. Assoc.* **1996**, *91*, 154–166. DOI: 10.1080/01621459.1996.10476672.

[33] Jones, G. L.; Hobert, J. P. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Stat.* **2004**, *32*, 784–817.

[34] Levin, D. A.; Peres, Y.; Wilmer, E. L. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2 edition, **2017**.

[35] Li, X.; Xie, Q. Coupling-Based convergence diagnostic and stepsize scheme for stochastic gradient descent. **2024**.

[36] Litvinenko, A.; Marzouk, Y.; Matthies, H. G.; Scavino, M.; Spantini, A. Computing f-divergences and distances of high-dimensional probability density functions. *Numer. Linear Algebra Appl.* **2023**, *30(3):e2467*, DOI: 10.1002/nla.2467.

[37] Madras, N.; Sezer, D. Quantitative bounds for Markov Chain convergence: Wasserstein and total variation distances. *Bernoulli* **2010**, *16*, 882–908. DOI: 10.3150/09-BEJ238.

[38] Poznyak, A. S. *Advanced Mathematical Tools for Automatic Control Engineers: Deterministic Techniques*. Elsevier, Oxford, **2008**.

[39] Qin, Q. Convergence bounds for Monte Carlo Markov Chains. **2024**.

[40] Qin, Q.; Hobert, J. P. On the limitations of single-step drift and minorization in Markov Chain convergence analysis. *The Annals of Applied Probability* **2021**, *31*, 1633–1659. DOI: 10.1214/20-AAP1628.

[41] Qin, Q.; Hobert, J. P. Geometric convergence bounds for Markov Chains in Wasserstein distance based on generalized drift and contraction conditions. Annales de L'Institut Henri Poincaré, Probabilités et. *Statistiques* **2022**, *58*, 872–889.

[42] Qin, Q.; Hobert, J. P. Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *The Annals of Applied Probability* **2022**, *32*, 124–166. DOI: 10.1214/21-AAP1673.

[43] Rachev, S. T.; Hsu, J. S. J.; Bagasheva, B. S.; Fabozzi, F. J. *Bayesian Methods in Finance*. Wiley and Sons, New York, **2008**.

[44] Rajaratnam, B.; Sparks, D. MCMC-based inference in the era of big data: a fundamental analysis of the convergence complexity of high-dimensional chains. **2015**.

[45] Roberts, G. O.; Rosenthal, J. S. Markov Chains and de-initializing processes. *Scand. J. Stat.* **2001**, *28*, 489–504. DOI: 10.1111/1467-9469.00250.

[46] Roberts, G. O.; Rosenthal, J. S. General state space Markov Chains and MCMC algorithms. *Probability Surveys* **2004**, *1*, 20–71. DOI: 10.1214/154957804100000024.
[47] Roman, J. C.; Hobert, J. P. Convergence analysis of the Gibbs sampler for Bayesian general linear mixed models with improper priors. *Ann. Stat.* **2012**, *40*, 2823–2849.
[48] Rosenthal, J. S. Minorization conditions and convergence rates for Markov Chain Monte Carlo. *J. Am. Stat. Assoc.* **1995**, *90*, 558–566. DOI: 10.1080/01621459.1995.10476548.
[49] Rosenthal, J. S. Rates of convergence for Gibbs sampling for variance component models. *The Annals of Statistics* **1995**, *23*, 740–761. DOI: 10.1214/aos/1176324619.
[50] Rosenthal, J. S. Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Stat. Comput.* **1996**, *6*, 269–275. DOI: 10.1007/BF00140871.
[51] Ross, S. M. *Introduction to Probability Models*. Academic Press, Boston, 10 edition, **2010**.
[52] Roy, V. Convergence diagnostics for Markov Chain Monte Carlo. *Annu. Rev. Stat. Appl.* **2020**, *7*, 387–412. DOI: 10.1146/annurev-statistics-031219-041300.
[53] Rubenstein, P. K.; Bousquet, O.; Djolonga, J.; Riquelme, C.; Tolstikhin, I. *Practical and Consistent Estimation of Undefined-Divergences*. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, Curran Associates Inc., **2019**.
[54] Sixta, S.; Rosenthal, J. S. Convergence rate bounds for iterative random functions using one-shot coupling. *Stat. Comput.* **2022**, *32*, 1573–1375. DOI: 10.1007/s11222-022-10134-x.
[55] Steinsaltz, D. Locally contractive iterated function systems. *Ann. Probab.* **1999**, *27*, 1952–1979. DOI: 10.1214/aop/1022677556.
[56] Tan, A.; Hobert, J. P. Block Gibbs sampling for Bayesian random effects models with improper priors: convergence and regeneration. *Journal of Computational and Graphical Statistics* **2009**, *18*, 861–878. DOI: 10.1198/jcgs.2009.08153.
[57] Tierney, L. Markov Chains for exploring posterior distributions. *Ann. Stat.* **1994**, *22*, 1701–1728.
[58] Vats, D. Geometric ergodicity of Gibbs samplers in bayesian penalized regression models. *Electron. J. Stat.* **2017**, *11*, 4033–4064. DOI: 10.1214/17-EJS1351.
[59] Villani, C. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 1 edition, **2008**.
[60] Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer New York, New York, 1 edition, **2004**.
[61] Yang, J.; Rosenthal, J. S. Complexity results for MCMC derived from quantitative bounds. *The Annals of Applied Probability* **2023**, *33*, 1459–1500. DOI: 10.1214/22-AAP1846.
[62] Stenflo, Ö. A survey of average contractive iterated function systems. *Journal of Difference Equations and Applications* **2012**, *18*, 1355–1380. DOI: 10.1080/10236198.2011.610793.