Bounding and estimating MCMC convergence rates using common random number simulations

Sabrina Sixta, Jeffrey S. Rosenthal, and Austin Brown

March 22, 2024

Abstract

This paper explores how and when to use common random number (CRN) simulation to evaluate Markov chain Monte Carlo (MCMC) convergence rates. We discuss how CRN simulation is closely related to theoretical convergence rate techniques such as one-shot coupling and coupling from the past. We present conditions under which the CRN technique generates an unbiased estimate of the squared L^2-Wasserstein distance between two random variables. We also discuss how this unbiasedness over a single iteration does not extend to unbiasedness over multiple iterations. 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. Finally, we apply our result to a Bayesian regression Gibbs sampler.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are often used to simulate from a stationary distribution of interest (see e.g. [5]). 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 [20]. 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 [13, 21, 34, 39]. However, none of these methods provide a formal estimate of the distance between the distribution of the Markov chain and the stationary distribution.

From a theoretical perspective, distance to stationarity is traditionally measured in terms of total variation distance (e.g. [35, 46]), though more recently the Wasserstein distance has been considered [17, 24, 29, 33]. However, finding upper bounds on either distance can be quite difficult to establish [15, 20], and if an upper bound is known, it is usually based on complicated problem-specific calculations [24, 32, 42, 44]. This motivates the desire to instead estimate convergence bounds from actual simulations of the Markov chain, which we consider here.

One common method for generating upper bounds on the Wasserstein distance is through a contraction condition (see Definition 2.1). This can often be established using the common random number (CRN) simulation technique, i.e. using the same sequence of random variables to simulate two copies of a Markov chain with different initial values (see Section 2). Estimating Markov chain convergence rates using CRN simulation was first proposed in [25] to find estimates of mixing times in total variation distance; see also [11, 22]. This approach falls under the general framework of "auxiliary simulation" [9], i.e. using extra preliminary Markov chain runs to estimate the convergence time needed in the final run. More recently, [2] 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). Simulation using the CRN technique is useful since for random variables under certain conditions it is an unbiased estimate of the squared Wasserstein distance (see equation 4) and for Markov chains under certain conditions it is a conditionally unbiased estimate of the squared Wasserstein distance (see equation 11). It was shown in [18] that simulated Euclidean distance between two random variables generated using the CRN technique is an unbiased estimate of the squared Wasserstein distance when the random transformation is

an increasing function of the uniform random variable (see propositions 3.1 and 3.2 below).

In this paper, in Theorem 3.3, we generalize the result of [18] and conclude that the CRN technique generates an unbiased estimate of the squared L^2-Wasserstein distance whenever the intervals over which the transformation of a random variable in $\mathbb{R}$ are increasing and decreasing are the same. This theorem should help establish whether the CRN technique is optimal for simulating the L^2-Wasserstein distance, or if another simulation technique such as [3, 9, 26, 30, 47, 50] is merited. Within the context of Markov chains, unbiasedness is only proven over a single iteration. We show how it is more difficult to extend over multiple iterations in Section 3.3.

Then, in Theorem 4.4, we provide an estimated upper bound in terms of CRN simulation on the Wasserstein distance between a Markov chain and the corresponding stationary distribution when only the unnormalized density of the stationary distribution is known. We apply this theorem (Section 5) to a Bayesian regression Gibbs sampler with semi-conjugate priors.

This paper is organized as follows. In Section 2, we present definitions and notation. We also discuss the relationship between the closely related notions of coupling from the past, one-shot coupling, and the CRN technique. In Section 3, we present a set of random functions (of real-valued random variables) that will generate unbiased estimates of the squared L^2-Wasserstein distance when the CRN technique is used. In Section 4, we establish convergence bounds of a Markov chain to its corresponding stationary distribution using the CRN technique when the initial distribution is not in stationarity. Finally, in Section 5, we apply our Theorem 4.4 to a Bayesian regression Gibbs sampler example. The code used to generate all of the tables and calculations can be found at github.com/sixter/CommonRandomNumber.

2 Background

2.1 Distances between measures

Let $(\mathcal{X}, \mathcal{F}, \lambda)$ be a measure space with $\mathcal{F}$ as the Borel σ-algebra and equipped with a metric d so that $(\mathcal{X}, d)$ is a Polish metric space. We define two random variables X, Y on $(\mathcal{X}, \mathcal{F}, \lambda)$ where $\mathcal{L}(X)$ is the distribution of X. Denote $\mathcal{P}(\mathcal{X})$ as the set of probability measures on $(\mathcal{X}, \mathcal{F})$. For $\mathcal{L}(X) = \pi, \mathcal{L}(Y) = \mu \in \mathcal{P}(\mathcal{X})$, define the set of couplings $J(X, Y)$ as

$$J(X,Y) = \{v \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : v(A_1 \times \mathcal{X}) = \pi(A_1), v(\mathcal{X} \times A_2) = \mu(A_2) \text{ for all } A_1, A_2 \in \mathcal{F}\}$$

The L^p-Wasserstein distance can then be defined as follows for $p \in [1, \infty)$

$$W_p(\pi, \mu) = \inf_{(X,Y) \sim J(X,Y)} E[d(X,Y)^p]^{1/p}$$

Total variation is defined as $||\pi - \nu||_{TV} = \sup_{A \in \mathcal{F}} |\pi(A) - \nu(A)|$.

2.2 Essential supremum and infimum

Let $f : \mathcal{X} \to \mathbb{R}$ be a function on the measure space $(\mathcal{X}, \mathcal{F}, \lambda)$. 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, $ess\inf_x f(x) = \sup_{a \in \mathbb{R}}\{a \mid \lambda(x \mid f(x) < a) = 0\}$ [31].

2.3 Iterative function systems: Backward and forward process

Define a Markov chain $\{X_n\}_{n \geq 1}$ in $(\mathcal{X}, \mathcal{F})$ such that $X_n = f_{\theta_n}(X_{n-1})$, where $\{\theta_n\}_{n \geq 1}$ are i.i.d. random variables on some measurable space Θ and random measurable mappings $f_{\theta_n} : \mathcal{X} \mapsto \mathbb{R}$. The set of random functions $f_{\theta_1}, f_{\theta_2}, \ldots$ is called an iterated function system. Any time-homogeneous Markov chain can be represented as an iterated function system [45].

The iterated function system defines the forward and backward processes. The forward

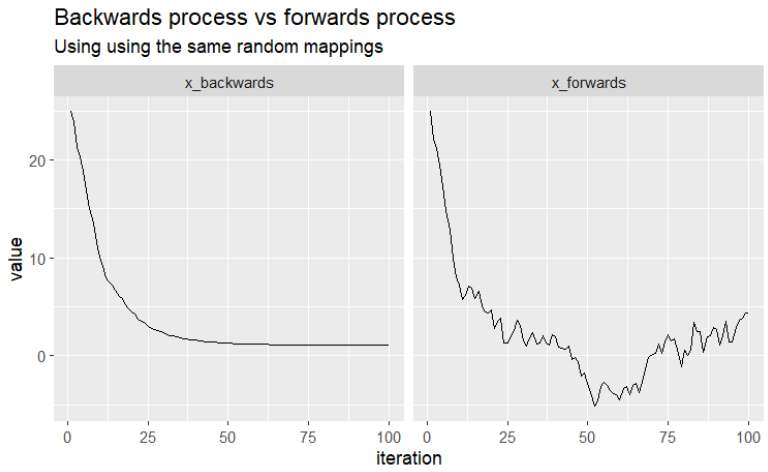

Figure 1: The backwards, $\{\tilde{X}_n\}_{n\geq 1}$, and forwards, $\{X_n\}_{n\geq 1}$, process of the autoregressive normal system where $X_n, \tilde{X}_n$ are simulated using the same random mappings, $X_n = 0.9X_{n-1} + Z_n, Z_n \sim N(0.1)$ and $X_0 = 25$.

process, $\{X_n\}_{n\geq 1}$, which is a Markov chain, is defined as follows,

$$X_n = f_{\theta_n}(f_{\theta_{n-1}}(\dots f_{\theta_1}(X_0)))$$

The backward process, $\{\tilde{X}_n\}_{n\geq 1}$, which is not a Markov chain [44] with respect to filtration for $X_0, \dots, X_n$ but tends to converge pointwise to a limit, is defined as follows,

$$\tilde{X}_n = f_{\theta_1}(f_{\theta_2}(\dots f_{\theta_n}(X_0)))$$

Figure 1 graphs the forward and backward process of an autoregressive process using the same random mappings. The point towards which the backwards process converges is itself random.

Despite the difference in behaviour of the two processes, the marginal distributions of the forward and backward processes are the same, $X_n \stackrel{d}{=} \tilde{X}_n$, so the backwards process is sometimes used when studying convergence properties of the forwards Markov chain [12, 45, 44].

2.4 Convergence of forward and backward processes

When studying the convergence rates of iterated random functions, we are typically interested in establishing a contraction condition. The 'vanilla' contraction condition, sometimes referred to as global average [45] or strongly [44] contractive, is defined as the supremum over all x of the expected Lipschitz constant.

Definition 2.1 (Global average contraction condition)**.** An iterative function system is global average contractive if there exists a $D \in (0, 1)$ such that,

$$D \geq \sup_{x \neq x'} \frac{E[d(f_\theta(x), f_{\theta'}(x'))]}{d(x, x')}$$

Modifications to the above contraction condition have been widely studied and can be found in [12, 16, 23, 27, 44, 45].

In particular, if two forward processes X_n, X'_n are simulated using the CRN technique (i.e., $X_n = f_{\theta_n}(X_{n-1})$ and $X'_n = f_{\theta_n}(X'_{n-1})$ for $n \in \mathbb{N}$) then the expected distance between the nth iteration of the two forward processes where $X_0 = x$ and $X'_0 = f_{\theta_{n+1}}(x)$ is equal to the expected distance between the nth and $n+1$th iteration of the backwards process, $\tilde{X}_n$.

$$\begin{aligned}
&E[d(X_n, X'_n) \mid X_0 = x] \\
&= E[d[f_{\theta_n}(f_{\theta_{n-1}}(\dots f_{\theta_1}(x)\dots)), f_{\theta_n}(f_{\theta_{n-1}}(\dots f_{\theta_1}(f_{\theta_{n+1}}(x))\dots))] \mid X_0 = x] \\
&= E[d[f_{\theta_1}(f_{\theta_2}(\dots f_{\theta_n}(x)\dots)), f_{\theta_1}(f_{\theta_2}(\dots f_{\theta_n}(f_{\theta_{n+1}}(x))\dots))] \mid \tilde{X}_0 = x] \\
&= E[d(\tilde{X}_n, \tilde{X}_{n+1}) \mid \tilde{X}_0 = x]
\end{aligned}$$

2.5 Common random numbers

Previously we defined the CRN technique to setting $\theta_n = \theta'_n$, i.e. using the same random variables $(\theta_n)_n$ to simulate both Markov chains. This is the intuitive definition of CRN for application [11, 2, 19]. However, we will first restrict our discussion to defining the CRN technique based on using uniform random variables as the common random number. Later on, we will discuss how

expanding the definition of the CRN technique from a uniform random variable to θ_n affects the optimality of the CRN technique.

3 Using the common random number technique as a conditionally unbiased estimate of the squared L^2- Wasserstein distance

3.1 Common random number applied to a random variable

Suppose X, Y are two random variables such that their marginal distributions can be represented as the inverse cumulative distribution function (CDF) of a uniform random variable. That is $X = F_X^{-1}(U)$ where F_X^{-1} is the inverse cumulative distribution function of the marginal and $U \sim Unif(0,1)$. Most joint probability spaces can be represented as such [18, 40, 48, 49]. A sufficient condition is that F_X be continuous (see exercise 1.2.4 of [14]).

We first define CRN to jointly setting $(X,Y) = (F_X^{-1}(U), F_Y^{-1}(U))$ [18]. We will call this definition the InvCDF-CRN (inverse CDF - common random number). The InvCDF-CRN is the joint distribution that minimizes the expected square distance between two random variables. That is, the InvCDF-CRN solves the Monge-Kanterovich problem when $d(x,y) = (x-y)^2$ (see Section 2.3.1 of [22]).

This definition stems from the following proposition which says that the maximum supermodular transform between the joint distribution of (X,Y) is attained when $(X,Y) = (F_X^{-1}(U), F_Y^{-1}(U))$.

Proposition 3.1 (Theorem 2 of [7] and Proposition 2.1 of [18])**.** *Suppose that $\psi : \mathbb{R}^2 \to \mathbb{R}$ is a supermodular function ($\psi(x_1,y_1) + \psi(x_2,y_2) \geq \psi(x_1,y_2) + \psi(x_2,y_1)$ when $x_1 \leq x_2$ and $y_1 \leq y_2$) and a right continuous function. Then for random variables X, Y,*

$$\sup_{(X,Y)\sim J(X,Y)} E[\psi(X,Y)] = E[\psi(F_X^{-1}(U), F_Y^{-1}(U))]$$

where $U \sim Unif(0,1)$.

Some examples of supermodular functions are $\psi(xy) = xy, (x+y)^2, \min\{x,y\}, f(x-y)$ where f is concave and continuous, $-|x-y|^p, p \geq 1$ (Section 4 of [7]). Proposition 3.1 is consistent with Theorem 2.9 of [41].

In this paper, we focus on the supermodular function $\psi(x,y) = xy$. If the CRN generates the supremum expectation for the function $\psi(x,y) = xy$ (i.e., $\sup_{J(X,Y)} E[XY] = E[F_X^{-1}(U)F_Y^{-1}(U)]$), then the CRN generates the supremum covariance when the marginal distributions are fixed as follows:

$$\sup_{(X,Y)\sim J(X,Y)} Cov(X,Y) = \sup_{(X,Y)\sim J(X,Y)} E[XY] - E[X]E[Y] \tag{1}$$

$$= E[F_X^{-1}(U)F_Y^{-1}(U)] - E[F_X^{-1}(U)]E[F_Y^{-1}(U)] \tag{2}$$

$$= Cov(F_X^{-1}(U), F_Y^{-1}(U)). \tag{3}$$

By the same reasoning, if the CRN generates the supremum expectation for the function $\psi(x,y) = xy$, then the CRN generates the L^2-Wasserstein distance (where $d(x,y) = |x-y|$) if X, Y have finite second moments as follows:

$$W_2(\mathcal{L}(X), \mathcal{L}(Y))^2 \coloneqq \inf_{(X,Y)\sim J(X,Y)} E[(X-Y)^2] \tag{4}$$

$$= E[X^2] + E[Y^2] - 2 \sup_{(X,Y)\sim J(X,Y)} E[XY] \tag{5}$$

$$= E[(F_X^{-1}(U) - F_Y^{-1}(U))^2] \tag{6}$$

Note that the above result can be generalized from L^2-Wasserstein distance to L^p-Wasserstein distance [22, 18].

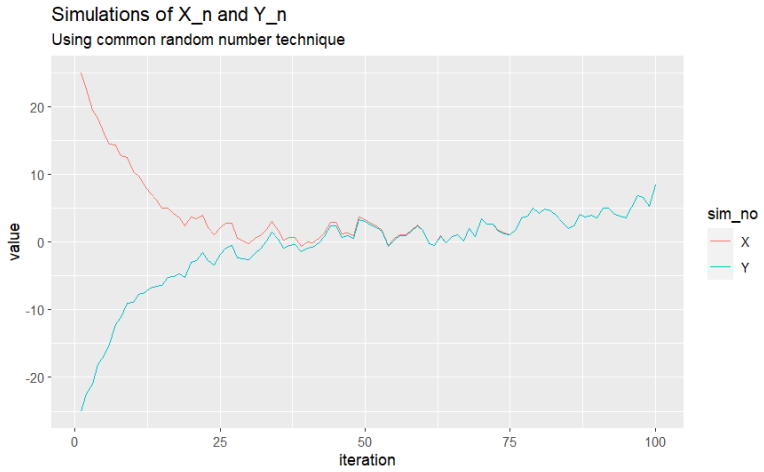

Figure 2: Two copies of the autoregressive process where X_n, Y_n are simulated using CRN, $X_n = 0.9X_{n-1} + Z_n, Z_n \sim N(0,1)$, and $X_0 = 25, Y_0 = -25$

3.2 Common random number in a Markov chain setting.

Within the context of a Markov chain, we define the CRN technique as follows. Let $\{X_n\}_{n\geq 0}$ be a Markov chain such that X_n is defined as an iterated function system and $X_n = f_{\theta_n}(X_{n-1}) = f(\theta_n, X_{n-1})$ where $\theta_1, \theta_2 \ldots$ are i.i.d. random variables. We assume that θ can be constructed from a uniform random variable, $\theta = F_\theta^{-1}(U)$, $U \sim Unif(0,1)$. Note that if $\vec{\theta} \in \mathbb{R}^q$ is a vector of independent random variables, then each coordinate can be constructed from a uniform random variable $\theta_i = F_{\theta_i}^{-1}(U_i), i \in \{1, \ldots q\}$ where $U_i \sim Unif(0,1)$ are i.i.d. (this is consistent with equation 3 of [18]). When used in simulation, the CRN technique visibly shows how two copies of a Markov chain converge. See figure 2 for an example of two autoregressive processes that converge.

Next, we extend Proposition 3.1 to non-decreasing functions of uniform random variables.

Proposition 3.2 (Proposition 2.2 of [18] for $\theta \in \mathbb{R}$)**.** *Fix $x, y \in \mathbb{R}$ and let $X_1 = f(\theta_U, x)$ and $Y_1 = f(\theta_V, y)$ with measurable $f : \mathbb{R}^2 \to \mathbb{R}$ where $\theta_U = F_\theta^{-1}(U)$, $\theta_V = F_\theta^{-1}(V)$ and $U, V \sim Unif(0,1)$. Suppose for $z = x, y$, $f(\cdot, z)$ is a non-decreasing continuous function with*

$E[X_1^2], E[Y_1^2] < \infty$. *Then*

$$\sup_{(U,V)\sim J(U,V)} E[f(\theta_U, x)f(\theta_V, y)] \tag{7}$$

is attained by setting $U = V$.

The original Proposition 2.2 of [18] is generalized to $\theta \in \mathbb{R}^q$.

We relaxed the assumption that $f(\cdot, x)$ be increasing to assuming that $f(\cdot, x)$ is non-decreasing. The relaxed assumption generates the same conclusion since $\psi(f(\cdot, x), f(\cdot, x))$ remains supermodular if ψ is supermodular and $f(\cdot, x), f(\cdot, y)$ are non-decreasing (which is required in the proof of Proposition 2.2 of [18]). Note, for example that the distribution function F^{-1} in Proposition 2.1 of [18] is a non-decreasing function, not a strictly increasing function.

Similar to Proposition 3.1, the supremum covariance (equation 1) and infimum Euclidean distance (equation 4) are attained when $U = V$.

In practise, however, $f(\theta, x)$ and $f(\theta, y)$ are not always non-decreasing functions with respect to θ. Rather, the following theorem shows that the supremum is attained when $U = V$ if the functions $f(\theta, x)$ and $f(\theta, y)$ are both increasing and decreasing for the same values of $\theta \in \mathbb{R}$.

Theorem 3.3. *For fixed* $x, y \in \mathcal{X}$, *let* $X_1 = f(\theta_U, x)$ *and* $Y_1 = f(\theta_V, y)$ *be random functions with measurable* $f : \mathbb{R}^2 \mapsto \mathbb{R}$ *and* $\theta_U = F_\theta^{-1}(U)$ *and* $\theta_V = F_\theta^{-1}(V)$ *with* $U, V \sim Unif(0, 1)$. *For* $z = x, y$, *assume* $\theta \mapsto f(\theta, z)$ *is of bounded variation (see section 3.1 of [43] for definition) and has finite second moments. Let* I_z *be the area over which* $\theta \mapsto f(\theta, z)$ *is a non-decreasing function and* D_z *is the area over which* $\theta \mapsto f(\theta, z)$ *is a non-increasing function. Define the set* $A = (I_x \cap I_y) \cup (D_x \cap D_y)$.

- *If* $P(A) = 1$ *(the intervals of positive measure over which* θ *is increasing and decreasing on* $f(\cdot, x)$ *and* $f(\cdot, y)$ *are the same),*

$$W_2(\mathcal{L}(X_1), \mathcal{L}(Y_1))^2 = E[(f(\theta_U, x) - f(\theta_U, y))^2] \tag{8}$$

i.e. $(f(\theta_U, x) - f(\theta_U, y))^2$ *is an unbiased estimator of* $W_2(\mathcal{L}(X_1), \mathcal{L}(Y_1))^2$.

- *If $P(A) = 0$ (that is the function $f(\cdot, x)$ is increasing and $f(\cdot, y)$ is decreasing over the same intervals of positive measure or vice versa),*

$$W_2(\mathcal{L}(X_1), \mathcal{L}(Y_1))^2 = E[(f(\theta_{1-U}, x) - f(\theta_U, y))^2] \tag{9}$$

i.e. $(f(\theta_{1-U}, x) - f(\theta_U, y))^2$ is an unbiased estimator of $W_2(\mathcal{L}(X_1), \mathcal{L}(Y_1))^2$.

- *If $0 < P(A(x, y)) < 1$ (there are intervals of positive measure over which θ is increasing on $f(\cdot, x)$ and decreasing on $f(\cdot, y)$ or vice versa), then*

$$E[(f(\theta_U, x) - f(\theta_U, y))^2] - e \leq W_2(\mathcal{L}(X_1), \mathcal{L}(Y_1))^2 \leq E[(f(\theta_U, x) - f(\theta_U, y))^2] \tag{10}$$

where $e = 2E[(f(\theta_{1-U}, x) - f(\theta_U, x))f(\theta_U, y)1_{A^C}]$.

The proof is in Section 6.

The set $A \equiv A(x, y)$ is simpler to visualize. It is the values of θ for which either both $f(\theta, x)$ and $f(\theta, y)$ are increasing or decreasing for fixed x, y. See figure 3 for an example.

3.3 Conditionally unbiased estimates of the L^2-Wasserstein distance

Denote $X_n = f(\theta_n, X_{n-1})$ and $Y_n = f(\theta'_n, Y_{n-1})$ where $\theta_n, \theta'_n \in \mathbb{R}$ are random variables. Theorem 3.3 implies that if $P(A) = 1$ for all $X_{n-1} = x$ and $Y_{n-1} = y$, then applying the CRN technique to simulate $E[(X_n - Y_n)^2]$ will generate conditionally unbiased estimates of the squared L^2-Wasserstein distance. That is, for two copies of a Markov chain, X_n and Y_n, that were simulated by CRN ($X_n = f(\theta_n, X_{n-1})$ and $Y_n = f(\theta_n, Y_{n-1})$), the squared L^2-Wasserstein distance will be conditionally unbiased on X_{n-1} and Y_{n-1} as follows:

$$W_2(\mathcal{L}(X_n), \mathcal{L}(Y_n) \mid X_{n-1} = x, Y_{n-1} = y)^2 \tag{11}$$

$$\coloneqq \inf_{(X_n, Y_n) \sim J(X_n, Y_n)} E[(X_n - Y_n)^2 \mid X_{n-1} = x, Y_{n-1} = y] \tag{12}$$

$$= E[(f(\theta_n, X_{n-1}) - f(\theta_n, Y_{n-1}))^2 \mid X_{n-1} = x, Y_{n-1} = y] \tag{13}$$

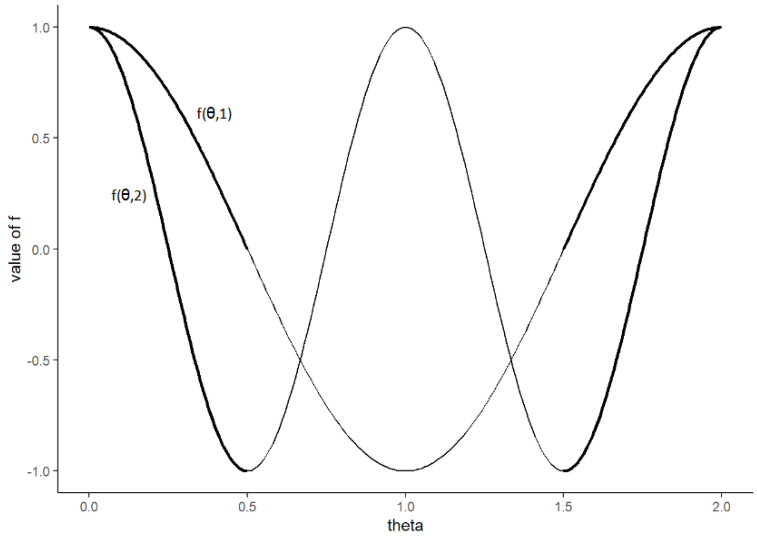

Figure 3: Denote $f(\theta, x) = \cos(\pi x\theta)$. The above graphs two functions: $f(\theta, 1) = \cos(\pi\theta)$ and $f(\theta, 2) = \cos(2\pi\theta)$. The set A in this case is the value of θ where the function is bold. That is, $A = (0, 0.5) \cup (1.5, 2)$.

This is the optimal coupling for the L^2-Wasserstein distance given the previous iteration.

Remark. Note that Markov chains that satisfy $P(A) = 1$ for all $X_{n-1} = x$ and $Y_{n-1} = y$ and are simulated by CRN may not be unconditionally unbiased. Unconditional unbiasedness is defined as follows for $X_0 = x$ and $Y_0 = y$:

$$W_2(\mathcal{L}(X_n), \mathcal{L}(Y_n))^2 = E[(f(\theta_n, f(\theta_{n-1}, \dots f(\theta_1, x)\dots)) - f(\theta_n, f(\theta_{n-1}, \dots f(\theta_1, y)\dots)))^2].$$

To illustrate our point, fix $n = 2$ and initial values $X_0 = x, Y_0 = y$. Then

$$W_2(\mathcal{L}(X_2), \mathcal{L}(Y_2))^2 \coloneqq \inf_{(\theta_1,\theta_2,\theta_1',\theta_2') \sim J(\theta_1,\theta_1') \times J(\theta_2,\theta_2')} E[(f(\theta_2, f(\theta_1, x)) - f(\theta_2', f(\theta_1', y)))^2].$$

However, the function $f(\theta_2, \cdot)$ may be increasing for $f(\theta_1, x)$ and decreasing for $f(\theta_1, y)$ (or vice versa), which by Theorem 3.3 means that the function might not generate the infimum expectation.

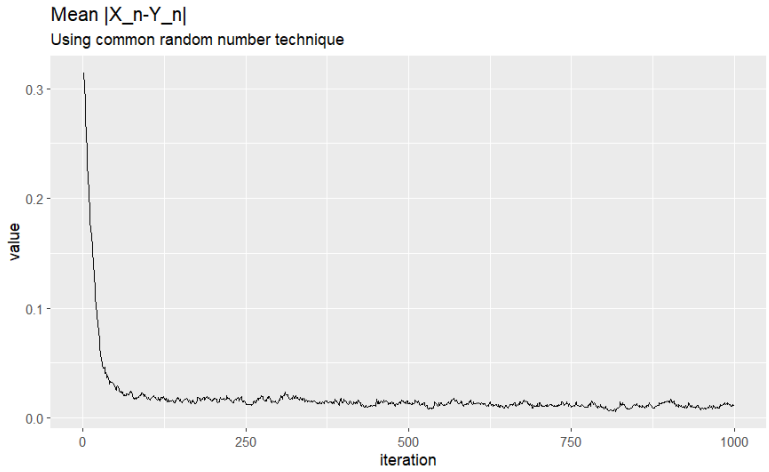

Figure 4: We apply a random-walk Metropolis algorithm on the unnormalized ($c > 0$) target density, $\pi(x) = cg(x)$ where $g(x) = x^3 \sin y^4 \cos y^5$ with proposal distribution $N(X_{n-1}, 0.01)$. The next iteration can be written as a function of $Z_n \sim N(0,1)$ and $U_n \sim Unif(0,1)$ as follows $X_n = (X_{n-1} + 0.1Z_n)I\{U_n < \pi(X_{n-1} + 0.1Z_n)/\pi(X_{n-1})\} + X_{n-1}I\{U_n \geq \pi(X_{n-1} + 0.1Z_n)/\pi(X_{n-1})\}$. When we use CRN on Z_n and U_n to simulate X_n and Y_n, $|X_n - Y_n|$ does not converge even though the Wasserstein distance does converge.

Remark. Even if a Markov chain does not present itself as conditionally unbiased it might still be conditionally unbiased under different function construction. This is because the CRN technique depends on the function f. To quote [47] "In most [...] scenarios the user must construct a coupling tailored to the problem at hand." For example, if X_1 and Y_1 have continuous distribution functions, then they can be written as transformations of the uniform distribution and thus a common random number that generates the maximum covariance exists. Generating such a function g such that $X_n = g(U), U \sim Unif(0,1)$ may not be easy to calculate, however. For example, the Metropolis Hastings algorithm does not appear to converge in expectation when the heuristic algorithm for using the common random number on the proposal and accept/reject random variables is used [10]. Figure 4 shows that when we apply the heuristic algorithm on the CRN technique of a Metropolis algorithm, sample mean convergence is bounded away from 0. In [30], asymptotically optimal estimates using a variation of the CRN are provided for the Metropolis-Hastings algorithm when the target distribution is elliptical normal.

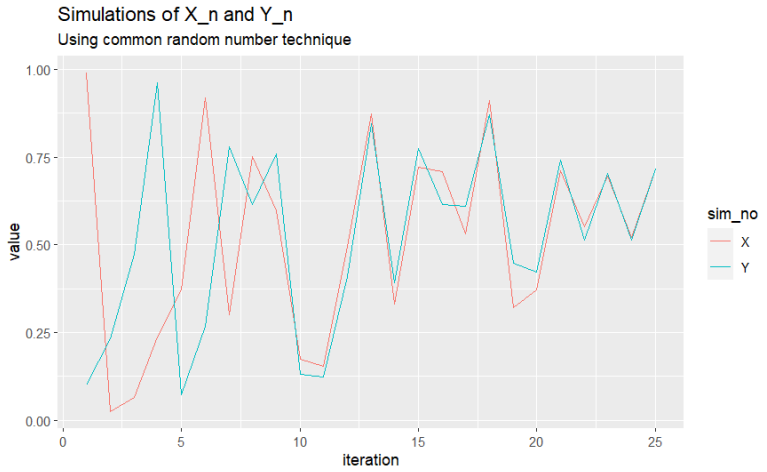

Figure 5: Two copies of the random logistic map where X_n, Y_n are simulated using CRN, $X_n = 4\theta_n X_{n-1}(1 - X_{n-1}), \theta \sim Beta(1.5, 0.5)$, and $X_0 = 0.99, Y_0 = 0.1$

The following two examples use Theorem 3.3 to show that the CRN technique generates conditionally unbiased estimates of the squared L^2-Wasserstein distance.

Example 3.1 (Random logistic map)**.** Define $X_n \in [0, 1]$ to be a random logistic map. That is, $X_n = f(\theta_n, X_{n-1}) = 4\theta_n X_{n-1}(1 - X_{n-1})$ and $\theta_n \sim Beta(a + 1/2, a - 1/2)$ for $a > 1/2$. Since $f(\theta_n, X_{n-1})$ is a non-decreasing function of θ_n for all values of $X_{n-1} \in [0, 1]$, then by Theorem 3.3 the CRN technique provides simulated estimates of the Euclidean distance that are conditionally unbiased to the squared L^2-Wasserstein distance.

See figure 5 for an example of two random logistic map processes that converge. See Section 7.4 of [44] for theoretical convergence diagnostics in total variation distance and example 2 of [29] for upper bounds in L^2-Wasserstein distance as a function of total variation distance.

Example 3.2. Define $X_n \in [-1, 1]$ to be a Markov chain such that

$$X_n = f(\theta_n, X_{n-1}) = \sin[(1 - |X_{n-1}|)\cos(\theta_n)]$$

where $\theta_n \sim Unif(-\pi/2, 3\pi/2)$. Since $f(\theta_n, X_{n-1})$ is increasing and decreasing over the same

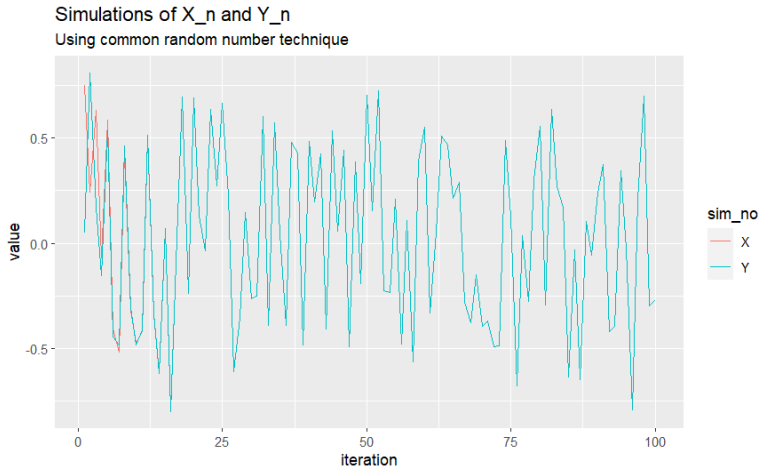

Figure 6: Two copies of a Markov chain $X_n = f(\theta_n, X_{n-1}) = \sin[(1-|X_{n-1}|)\cos(\theta_n)]$ where $\theta_n \sim Unif(-\pi/2, 3\pi/2)$. X_n, Y_n are simulated using CRN, with initial values and $X_0 = 0.75, Y_0 = 0.05$

regions of θ for fixed $X_{n-1}, Y_{n-1} \in [-1, 1]$, then by Theorem 3.3 the CRN technique provides simulated estimates of the Euclidean distance that are conditionally unbiased to the squared L^2-Wasserstein distance.

Figure 6 provides an example of copies of this Markov chain converging when the CRN technique is used. Figure 7 shows that for initial values $X_0 = 0.75$ and $Y_0 = 0.05$ the functions $f(X_0, \theta)$ and $f(Y_0, \theta)$ are both increasing and decreasing over the same regions of θ.

Finally, note that Theorem 3.3 is only applicable for $\theta \in \mathbb{R}$. Further research needs to be done to extend the above results to $\theta \in \mathbb{R}^q$. However, we believe that a proof can be established using Proposition 2.2 of [18] and assuming that $f(\theta, x)$ is of bounded Arzelà-variation (see Definition 3.2.1 of [4]). If a function defined in $\mathbb{R}^d$ is of bounded Arzelà-variation, then it can be written as the difference of two coordinate-wise increasing functions (see Theorem 3.4.1 of [4]) where Proposition 2.2 of [18] can then be applied. Example 3.3 is a Markov chain of Dirichlet process means where $\theta \in \mathbb{R}^2$ and $f(\theta, x)$ is not always an increasing function of θ, so none of the theorems in this text apply. The CRN technique still appears to generate converging Markov chains, however (see figure 8).

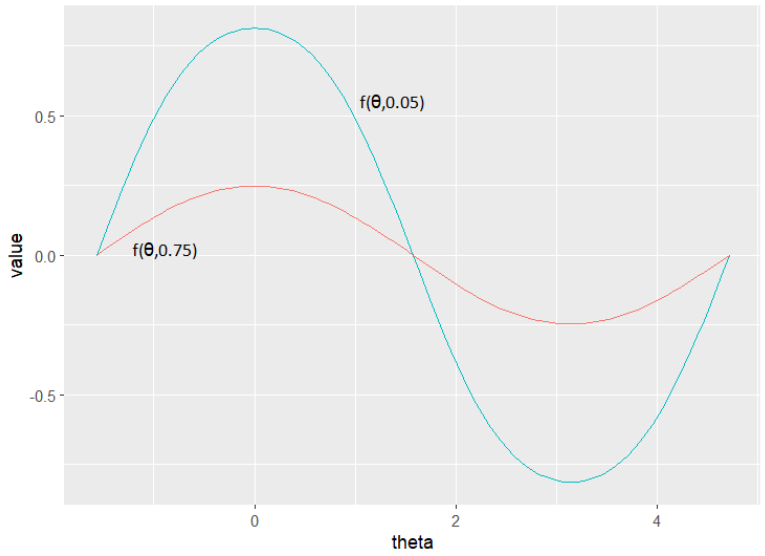

Figure 7: Denote $f(\theta, x) = \sin[(1 - |x|)\cos(\theta)]$. The above graphs two functions: $f(\theta, 0.75) = \sin[0.25\cos(\theta)]$ and $f(\theta, 0.05) = \sin[0.95\cos(\theta)]$. Note that both functions are decreasing over the same region, $\theta = (0, \pi)$.

Example 3.3 (Dirichlet process means)**.** Define $X_n \in [0, 1]$ to be a Markov chain of Dirichlet process means. That is $X_n = f(\theta_n, Z_n, X_{n-1}) = (1 - \theta_n)Z_n + \theta_n X_{n-1}$ and $\theta_n \sim Beta(a, 1)$ for $a > 0$ and $Z_n \sim N(0, 1)$. In this case $(\theta_n, Z_n) \in \mathbb{R}^2$, so Theorem 3.3 does not apply. Further, since $f(\theta_n, Z_n, X_{n-1})$ may not necessarily be an increasing function of (θ_n, Z_n), so Proposition 2.2 of [18] also does not apply. The Markov chain of Dirichlet process means still seems to converge when the CRN technique is used. See figure 8 for an example of two Markov chains with Dirichlet processes means that converge. See section 7 of [37] for theoretical convergence diagnostics.

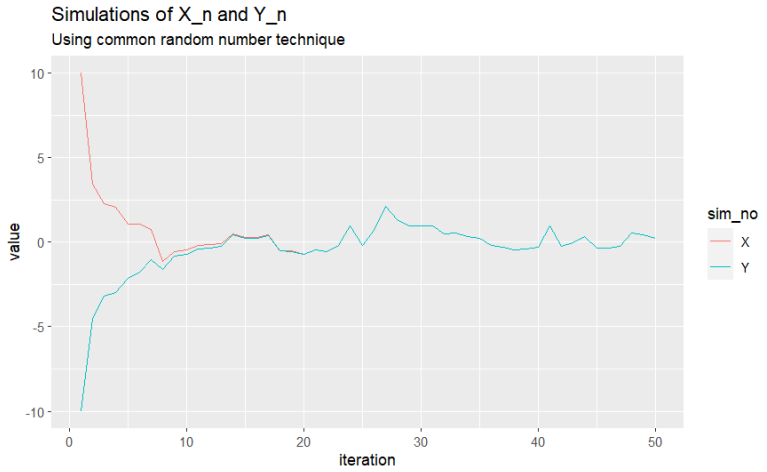

Figure 8: Two copies of a Markov chain with Dirichlet process means where X_n, Y_n are simulated using CRN, $X_n = (1-\theta_n)Z_n + \theta_n Z_{n-1}, \theta \sim Beta(1.5, 1), Z_n \sim N(0,1)$, and $X_0 = 10, Y_0 = -10$

4 Common random number as a method of simulating Markov chain convergence rates

We propose estimating the L^p-Wasserstein distance between the nth iteration of a Markov chain X_n and the corresponding stationary distribution X_∞ through simulation using the CRN technique. Our method is outlined in Theorem 4.4. Using CRN simulation as a convergence diagnostic tool was first discussed in [25] for bounding total variation distance.

Before providing a method for simulating an upper bound on $E[|X_n - X_\infty|^p]$, we must first provide a method of bounding the expected distance between two copies of a Markov chain ($E[|X_n - Y_n|^p]$) and their corresponding expected distance to stationarity, ($E[|X_n - X_\infty|^p]$). To do so, we will first define rejection sampling and separation distance and show how they are related.

Definition 4.1 (Rejection sampling)**.** Suppose that we have a target density π, which we want to sample from, but is difficult to do, and we have a proposal density ν that is easier to sample from. Suppose also that $\pi << \nu$ (i.e., $\nu(A) = 0 \implies \pi(A) = 0$, where $A \in \mathcal{F}$) and $K \geq$

$\pi(x)/\nu(x), x \in \mathcal{X}$ for some known K. To generate a random variable $X_\pi, \mathcal{L}(X_\pi) = \pi$ we do the following,

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

Lemma 4.1 (Rejection sampler rejection rate)**.** *Denote the event*

$$A = \{X \text{ is accepted as a draw from } \pi\}$$

in the rejection sampler algorithm defined above. The rejection rate denoted as r is as follows where $K = ess\sup_{x\in\mathcal{X}} \pi(x)/\nu(x)$.

$$r(\pi, \nu) = 1 - P(A) = 1 - 1/K$$

Proof. See Section 11.2.2 of [38]. □

We further define the separation distance on the continuous state space $\mathcal{X}$ as follows. Separation distance was first defined in [1] for discrete state spaces. As far as we know, separation distance was only recently defined on a continuous state space in [8]. We use the definition of separation distance defined in [8] where the density functions are known.

Definition 4.2 (Separation distance (Remark 5 of [8]))**.** Let ν and π be two probability density functions defined on the same measurable space $(\mathcal{X}, \mathcal{F})$ such that $\nu << \pi$. The separation distance is,

$$s = s(\pi, \nu) = ess\sup_x \left(1 - \frac{\nu(x)}{\pi(x)}\right)$$

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

Lemma 4.2. *Let ν and π be two density functions defined on the same measurable space $(\mathcal{X}, \mathcal{F})$. If ν is the proposal density and π is the target density in a rejection sampler, then the rejection rate equals the separation distance,*

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

Proof.

$$s(\pi, \nu) = ess \sup_{x \in \mathcal{X}} \left(1 - \frac{\nu(x)}{\pi(x)}\right) = 1 - ess \inf_{x \in \mathcal{X}} \frac{\nu(x)}{\pi(x)} = 1 - \frac{1}{ess \sup_{x \in \mathcal{X}} \frac{\pi(x)}{\nu(x)}} = 1 - \frac{1}{K} = r(\pi, \nu)$$

□

Now that we have defined separation distance and the rejection rate of the rejection sampler, we can apply upper bounds on the distance to stationarity as follows. The use of rejection sampling to generate an upper bound on the expected distance between a proposal density function ν and a stationary density function π was inspired by [25], which used rejection sampling to generate similar upper bounds in total variation distance.

Theorem 4.3. *Let $\{X_n\}_{n \geq 1}$ and $\{Y_n\}_{n \geq 1}$ be two copies of a Markov chain in $\mathcal{X}$ with initial distribution $\mathcal{L}(X_0) = \mu, \mathcal{L}(Y_0) = \nu$. Assume π and ν are defined on the same support and let X_∞ be the corresponding stationary random variable with distribution π. Then for $p \in [1, \infty)$*

$$E[|Y_n - X_\infty|^p] \leq K E[|X_n - Y_n|^p] \tag{14}$$

where the expectation is taken with respect to coupling using common random numbers $\{\theta_n\}_{n \geq 1}$ such that X_0 and Y_0 are independent and $K = \frac{1}{1 - s(\pi, \nu)} = \frac{1}{1 - r(\pi, \nu)}$.

Proof. 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 $r(\pi,\nu) = 1 - P(A)$. Note that $\mathcal{L}(Y_0|A) = \pi$ and so, $\mathcal{L}(Y_n|A) = \pi$.

$$\begin{aligned} E[|X_n - Y_n|^p] &= E[|X_n - Y_n|^p : A]P(A) + E[|X_n - Y_n|^p : A^c]P(A^c) \\ &\geq E[|X_n - X_\infty|^p]P(A) \\ &\geq E[|X_n - X_\infty|^p]\frac{1}{K} \end{aligned}$$

Further, $1/K = 1 - r(\pi,\nu) = 1 - s(\pi,\nu)$ and so equation 14 follows. □

In many cases the normalizing constant for π is unknown. That is, we only know of a function $g(x)$ such that $cg(x) = \pi(x)$ where c is unknown. In this case we must find a constant $L > 0$ such that $\int_{\mathcal{X}} g(x)dx \geq L$, which implies that $c \leq \frac{1}{L}$ and,

$$K \leq \frac{1}{L}\sup_{x\in\mathcal{X}}\frac{g(x)}{\nu(x)}. \tag{15}$$

See Example 5.1 for a way of estimating K when the normalizing constant of π is unknown. In Section 5 we estimate L as the integral over a bounded subset of $\mathcal{X}$, $B \subset \mathcal{X}$. That is, $L = \int_B g(x)dx$. [6] presents an alternative approach to estimating the rejection rate, K, when the normalizing constant for π is unknown.

Next we define Algorithm 1 for generating an estimate of $\inf_{(X_N,Y_N)\sim J(X_N,Y_N)} E[|X_N - Y_N|^p]$, $N \geq 1$ where $\mathcal{L}(X_0) = \mu, \mathcal{L}(Y_0) = \nu$. The algorithm generates a conditionally unbiased estimate of the Wasserstein distance provided the conditions in Theorem 3.3 hold. This algorithm is similar to Algorithm 1 in [2]. Combining algorithm 1 with Theorem 4.3, we can simulate an upper bound between a Markov chain at iteration N, X_N, and the corresponding stationary random variable, X_∞, as follows,

Theorem 4.4. *Suppose that the Markov chain $\{X_n\}_{n\geq 0}$ with stationary distribution $\mathcal{L}(X_\infty) = \pi$ can be written as an iterated function system $X_{n+1} = f_{\theta_{n+1}}(X_n)$ where $(\theta_n)_{n\geq 1}$ are i.i.d. random variables. Let (X_n, Y_n) be two copies of the Markov chain coupled using common random numbers $\{\theta_n\}_{n\geq 1}$ with $\mathcal{L}(X_0) = \nu$ and $\mathcal{L}(Y_0) = \mu$. Suppose for each $n \geq 0$, $E[|X_n|^p], E[|Y_n|^p] < \infty, p \in$*

Algorithm 1 An estimate of $E[|X_N - Y_N|^p] \approx \frac{1}{I}\sum_{i=1}^{I}|x_{N,i} - y_{N,i}|^p$ using CRN

for $i = 1, \ldots, I$ **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**
    $\theta_n \sim \Theta$
    $x_{n,i} \leftarrow f_{\theta_n}(x_{n-1,i})$
    $y_{n,i} \leftarrow f_{\theta_n}(y_{n-1,i})$
  end for
end for
return $\frac{1}{I}\sum_{i=1}^{I}|x_{N,i} - y_{N,i}|^p$

$[1, \infty)$ *and that the initial distribution of* X_n, $\mathcal{L}(X_0) = \nu$ *is defined on the same support as* π. *Then*

$$\inf_{(X_N, X_\infty) \sim J(X_N, X_\infty)} E[|X_N - X_\infty|^p]^{1/p} \leq (K \lim_{I \to \infty} \frac{1}{I}\sum_{i=1}^{I}|x_{N,i} - y_{N,i}|^p)^{1/p} \tag{16}$$

holds almost surely where $\frac{1}{I}\sum_{i=1}^{I}|x_{N,i} - y_{N,i}|^p$ *is defined as in Algorithm 1 and* $K = \frac{1}{1-s(\pi,\nu)} = \frac{1}{1-r(\pi,\nu)}$.

Proof. Fix $N \geq 1$. By the strong law of large numbers, $\lim_{I\to\infty}\frac{1}{I}\sum_{i=1}^{I}|x_{N,i} - y_{N,i}|^p \overset{a.s.}{=} E[|X_N - Y_N|^p]$. By Theorem 4.3, $E[|X_N - X_\infty|^p] \leq KE[|X_N - Y_N|^p]$ and so equation 16 follows. □

The consistency of $\lim_{I\to\infty}\frac{1}{I}\sum_{i=1}^{I}|x_{N,i} - y_{N,i}|^p$ is similarly proven in Proposition 3.1 of [2].

5 Bayesian regression Gibbs sampler example

Consider the Bayesian linear regression model with semi-conjugate priors (see Chapter 5 of [34]) for which we will apply Theorem 4.4 to simulate convergence bounds on the Wasserstein distance.

Example 5.1 (Bayesian regression Gibbs sampler with semi-conjugate priors)**.** Suppose we have the following observed data $Y \in \mathbb{R}^k$ and $X \in \mathbb{R}^{k\times q}$ where

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

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

unknown parameters,

$$\beta \sim N_q(\beta_0, \Sigma_\beta) \qquad\qquad \sigma^2 \sim \text{Inv-}\chi^2(\upsilon_0, c_0^2).$$

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

$$g(\beta, \sigma^2) \tag{17}$$

$$= \frac{1}{(\sigma^2)^{(k+\upsilon_0)/2+1}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{1}{2}(\beta - \beta_0)^T\Sigma_\beta^{-1}(\beta - \beta_0) - \frac{\upsilon_0 c_0^2}{2\sigma^2}\right). \tag{18}$$

The Bayesian regression Gibbs sampler is based on the conditional posterior distributions of β_n, σ_n^2 and is defined as follows initialized at β_0, σ_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. $\sigma_n^2|\beta_n, Y, X \sim \Gamma^{-1}\left(\frac{k+\upsilon_0}{2}, \frac{1}{2}\left[\upsilon_0 c_0^2 + (Y - X\beta_n)^T(Y - X\beta_n)\right]\right)$

where

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

Here $\Gamma^{-1}(\alpha, \beta)$ represents the inverse gamma distribution with shape parameter α and rate parameter β. Replacing $\beta_n = \tilde{\beta}_{\sigma_{n-1}^2} + V_{\beta,\sigma_{n-1}^2}^{1/2} Z_n$, where $Z_n \sim N(0, I_q)$ into the equation for σ_n^2 and with $G_n \sim \Gamma(\frac{k+\upsilon_0}{2}, 1)$, we get that

$$\sigma_n^2|\sigma_{n-1}^2, Y, X = \left[\frac{\upsilon_0 c_0^2}{2} + \frac{(X\tilde{\beta}_{\sigma_{n-1}^2} - Y + XV_{\beta,\sigma_{n-1}^2}^{1/2} Z_n)^T(X\tilde{\beta}_{\sigma_{n-1}^2} - Y + XV_{\beta,\sigma_{n-1}^2}^{1/2} Z_n)}{2}\right]\frac{1}{G_n} \tag{19}$$

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

Although the joint Markov chain may be high-dimensional in both coordinates, special properties of this Gibbs sampler allow us to upper bound the total variation between two joint Markov chains, (σ_n^2, β_n) and $(\sigma_\infty^2, \beta_\infty)$, in terms of only σ_n^2 and σ_∞^2.

Lemma 5.1. *Let $\beta_\infty, \sigma^2_\infty$ denote the stationary Gibbs Markov chain. The total variation distance can be bounded by the expected distance as follows,*

$$\left\|\mathcal{L}(\beta_{n+1}, \sigma^2_{n+1}) - \mathcal{L}(\beta_\infty, \sigma^2_\infty)\right\|_{TV} \leq \frac{(k+\upsilon_0)^2}{2\upsilon_0 c_0^2} E[|\sigma^2_n - \sigma^2_\infty|]$$

where the expectation is with respect to coupling using common random numbers $(Z_n, G_n)_n$ for σ^2_n and $\sigma^2_\infty = \sigma^{'2}_n$ such that $\sigma^2_0 \sim \mu$ and $\sigma^{'2}_0 \sim \pi$ are independent.

Proof. It follows from the de-initializing property of the Markov chain [36, Example 3] that

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

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

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

so that $\sigma^2_n|\sigma^2_{n-1} = W_n \frac{1}{G_n}$ and $\sigma^{'2}_n|\sigma^{'2}_{n-1} = W'_n \frac{1}{G'_n}$. 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 π_{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

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

Using the coupling characterization of total variation

$$
\begin{aligned}
&||\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_n^{'2})||_{TV} \\
&\leq E\left[||\mathcal{L}(W_n\frac{1}{G_n}) - \mathcal{L}(W_n'\frac{1}{G_n})||_{TV} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right] \\
&= E\left[||\mathcal{L}(W_n\frac{1}{G_n}) - \mathcal{L}((W_n+\Delta)\frac{1}{G_n})||_{TV} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right] \qquad \text{By Proposition 2.2 of [42]} \\
&= E\left[||\mathcal{L}\left(\frac{1}{G_n}\right) - \mathcal{L}\left(\left(1+\frac{\Delta}{W_n}\right)\frac{1}{G_n}\right)||_{TV} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right] \quad \text{By Proposition 2.1 of [42]} \\
&\leq E\left[\sup_{x>0}\left\{1 - \frac{\pi_{1/G_n}(x)}{\pi_{(1+\Delta/W_n)/G_n}(x)}\right\} \mid Z_n, \sigma_{n-1}^2, \sigma_{n-1}^{'2}\right] \qquad \text{By Lemma 6.16 of of [28]}
\end{aligned}
$$

So,

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

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,

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

□

Using Lemma 5.1, we can apply Theorem 4.4 to only the one-dimensional initialization of the marginal Gibbs Markov chain (σ_n^2) to estimate the total variation of the possibly high-dimensional joint Markov chain.

Lemma 5.2. *Let $(\sigma_n^2)_n$ and $(\sigma_n^{'2})_n$ be two copies of the Markov chain initialized with $\sigma^2 \sim \nu$ and $\sigma^{'2} \sim \mu$. Assume $\sigma_0^2 \sim \nu$ is a distributed $\Gamma^{-1}(\alpha', \beta')$ with parameters $\alpha' = (k+\upsilon_0)/2$ and $\beta' = \upsilon_0 c_0^2$. Let $\sigma_\infty^2 \sim \pi$ be from stationary Markov chain started from the the marginal posterior distribution. Then*

$$E[|\sigma_n^{'2} - \sigma_\infty^2|] \leq \frac{(2\pi)^{q/2}\det(\Sigma_\beta)^{1/2}}{\int_{\mathcal{R}^q \times \mathcal{R}_+} g(\beta, \sigma^2) d(\beta \times \sigma^2)} \frac{\Gamma(\alpha')}{\beta'^{\alpha'}} E[|\sigma_n^2 - \sigma_n^{'2}]$$

where the expectation is taken with respect to coupling using common random numbers and σ_0^2 and $\sigma_0^{'2}$ are independent.

Proof. Define f as the corresponding density function of ν, $f(\sigma^2) = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')}\frac{1}{(\sigma^2)^{\alpha'+1}}e^{-\beta'/\sigma^2}$ where

$\alpha' = (k+\upsilon)/2$ and $\beta' = \upsilon_0 c_0^2/2$. Note that

$$\begin{aligned}\int g(\beta,\sigma^2)d\beta &\leq (2\pi)^{q/2}\det(\Sigma_\beta)^{1/2}\frac{1}{(\sigma^2)^{(k+\upsilon_0)/2+1}}\exp\left(-\frac{\upsilon_0 c_0^2}{2\sigma^2}\right)\\ &= (2\pi)^{q/2}\det(\Sigma_\beta)^{1/2}\frac{\Gamma(\alpha')}{\beta'^{\alpha'}}f(\sigma^2).\end{aligned}$$

By equation 15 the value for K follows. By equation 14 the inequality follows. □

Given Theorem 4.4 and Lemma 5.2 we show how an upper bound on the convergence rate in Wasserstein distance can be simulated for a numerical example of the Bayesian regression Gibbs sampler with semi-conjugate priors, Example 5.1. We further provide an estimate of the upper bound in total variation using Lemma 5.1.

Numerical Example 5.1. 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 [13].

We want to find the estimated upper bound on the total variation distance for a Bayesian regression Gibbs sampler with semi-conjugate priors fitted to this model. In this case, there are 20 observed values and 4 parameters ($k = 20, p = 4$). We set the priors to $\beta_0 = \vec{0}$, $\Sigma_\beta = I_4$, $\upsilon_0 = 1$, $c_0^2 = 10$. Applying R simulation, we can set $g(\beta,\sigma^2) \geq L = 0.9687$, where L is defined in equation 15. Using Lemma 5.2 $K \leq 2.1150$ and,

$$E[|\sigma_n^2 - \sigma_\infty^{'2}|] \leq 2.1150 \lim_{I\to\infty}\frac{1}{I}\sum_{i=1}^{I}|\sigma_{n,i}^2 - \sigma_{n,i}^{'2}|$$

holds almost surely.

Using the CRN technique, we simulated $E[|\sigma_{n,i}^2 - \sigma_{n,i}^{'2}|]$ one thousand times ($I = 1000$) over 100 iterations ($N = 100$). Figure 9 graphs the 1000 simulations and shows that the absolute differences spike at the second iteration, $|\sigma_{2,i}^2 - \sigma_{2,i}^{'2}|$, but converge quite quickly after this. At iteration 25, $E[|\sigma_{25}^2 - \sigma_\infty^2|] \leq K\frac{1}{1000}\sum_{i=1}^{1000}|\sigma_{25,i}^2 - \sigma_{25,i}^{'2}| = 2.1150 \times 0.0014 \approx 0.00291$. Figure 10 graphs the histogram of $|\sigma_{25,i}^2 - \sigma_{25,i}^{'2}|$. Further, by Lemma 5.1 the total variation distance is

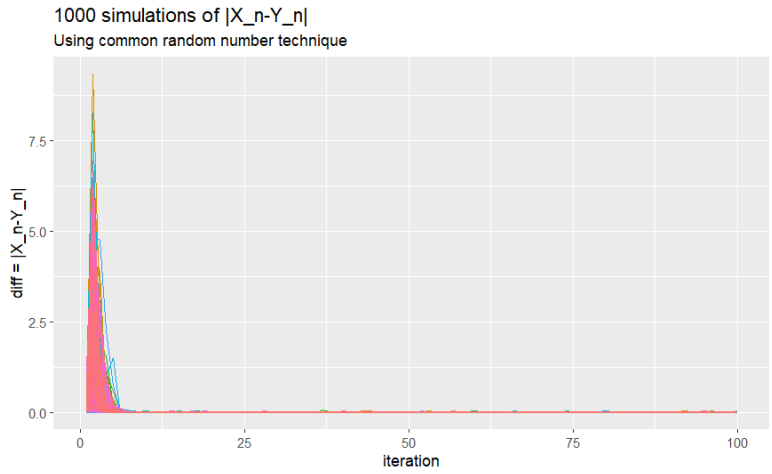

Figure 9: 1000 simulations of $|\sigma^2_{n,i} - \sigma'^2_{n,i}|$ where n =iteration.

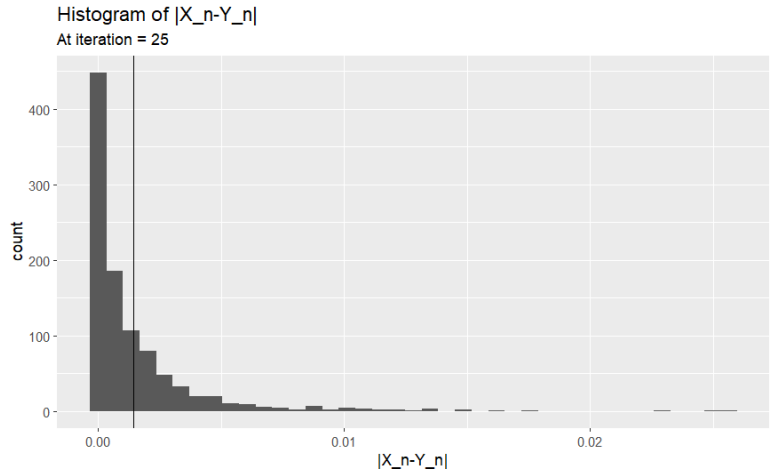

Figure 10: The following histogram graphs 1000 simulations of $|\sigma^2_{25,i} - \sigma'^2_{25,i}|$. The vertical line and sample mean difference is $\frac{1}{1000}\sum_{k=1}^{1000} |\sigma^2_{25,i} - \sigma'^2_{25,i}| = 0.0014$

bounded above by $\frac{(k+\upsilon_0)^2}{2\upsilon_0 c_0^2} = 22.05$ times the expected distance,

$$||\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_\infty^{'2})||_{TV} \leq 22.05E[|\sigma_n^2 - \sigma_\infty^2|]$$

So at the 25th iteration, $||\mathcal{L}(\sigma_{25}^2) - \mathcal{L}(\sigma_\infty^{'2})||_{TV} \leq 0.0642$.

6 Proof of Theorem 3.3

Proof of Theorem 3.3. Fix x, y and denote $g(U) = f(F_\theta^{-1}(U), x)$ and $h(V) = f(F_\theta^{-1}(V), y)$. We write $J(U, V)$ to be the set of joint distributions such that $\mathcal{L}(U) = Unif(0, 1), \mathcal{L}(V) = Unif(0, 1)$. Finally, note that since $\theta = F^{-1}(U)$ where F is invertible, we interchangeably write the set A to signify A (defined on θ) and $F(A)$ (defined on U, V).

Also note that since $g(U), h(V)$ are assumed to be of bounded variation, the sets I_x, I_y, D_x, D_y can be written as the union of countable intervals by corollary 3.6 of [43].

First we show that

$$\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_A] = E[g(U)h(U)1_A] \tag{20}$$

Denote $I_x \cap I_y = \cup_{k\geq 1} J_k$ and $A \setminus (I_x \cap I_y) = \cup_{k\geq 1} L_k$ where J and L are intervals. We have

$$
\begin{aligned}
&\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_A] \\
&\leq \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{I_x\cap I_y}] + \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{A\setminus(I_x\cap I_y)}] \\
&= \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{I_x\cap I_y}] + \sup_{J(U,V)} E[(-g(U))(-h(V))1_{A\setminus(I_x\cap I_y)}] \\
&= \sup_{(U,V)\sim J(U,V)} E[\sum_{k\geq 1} g(U)h(V)1_{J_k}] + \sup_{(U,V)\sim J(U,V)} E[\sum_{k\geq 1}(-g(U))(-h(V))1_{L_k}] \\
&= \sup_{(U,V)\sim J(U,V)} \sum_{k\geq 1} E[g(U)h(V)1_{J_k}] + \sup_{(U,V)\sim J(U,V)} \sum_{k\geq 1} E[(-g(U))(-h(V))1_{L_k}] \\
&\leq \sum_{k\geq 1} \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{J_k}] + \sum_{k\geq 1} \sup_{(U,V)\sim J(U,V)} E[(-g(U))(-h(V))1_{L_k}] \\
&= \sum_{k\geq 1} E[g(U)h(U)1_{J_k}] + \sum_{k\geq 1} E[(-g(U))(-h(U))1_{L_k}] \\
&= E[g(U)h(U)1_{I_x\cap I_y}] + E[(-g(U))(-h(U))1_{A\setminus(I_x\cap I_y)}] \\
&= E[g(U)h(U)1_A]
\end{aligned}
$$

The third equality is by the Dominated Convergence Theorem, Theorem 1.5.8 of [14], and the second last equality is by Proposition 3.2.

Since $E[g(U)h(U)] \leq \sup_{J(U,V)} E[g(U)h(V)] \leq E[g(U)h(U)]$, equality follows. Note that Proposition 3.2 can still be applied even if $g(U)h(U)1_{J_k}$ and $g(U)h(U)1_{L_k}$ are no longer the product of right continuous functions, which is a result of the fact that 1_{J_k} may not necessarily be right continuous. This is because the theorem still applies by discussions in Section 4 of [7] that note that rather than requiring that $f(\theta, x)$ be right continuous it is simply necessary to assume that the discontinuous points are countable and have left and right limits. This is a necessary condition for bounded variation (assumed in the theorem). Since the function I_A has countably many discontinuities, Proposition 3.2 can be used.

Second we show that

$$\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{A^c}] = E[g(U)h(1-U)1_{A^c}] \tag{21}$$

Suppose A^c can also be written as countable intervals. The set A^c represents areas where either g is increasing and h is decreasing or vice versa. By similar reasoning to equation 20,

$$\begin{aligned}
\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{A^c}] &= -\inf_{(U,V)\sim J(U,V)} E[g(U)(-h(V))1_{A^c}] \\
&= -\inf_{(U,V)\sim J(U,V)} E[\sum_{I\in A^c} g(U)(-h(1-U))I] \\
&= -\inf_{(U,V)\sim J(U,V)} \sum_{I\in A^c} E[g(U)(-h(1-U))I] \\
&\leq -\sum_{I\in A^c} \inf_{(U,V)\sim J(U,V)} E[g(U)(-h(1-U))I] \\
&= -\sum_{I\in A^c} E[g(U)(-h(1-U))I] \\
&= -E[g(U)(-h(1-U))1_{A^c}] \\
&= E[g(U)h(1-U)1_{A^c}]
\end{aligned}$$

The third equality is by the Dominated Convergence Theorem, Theorem 1.5.8 of [14], and the third last equality is by Theorem 2 of [7].

Again, since $E[g(U)h(1-U)1_{A^c}] \leq \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{A^c}] \leq E[g(U)h(1-U)1_{A^c}]$, equality follows.

Case 1: Suppose $P(A) = 1$. We write $P(A^c) = P(U, V) \in F(A^c))$

$$\begin{aligned}
&\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)] \\
&= \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_A] + E[g(U)h(V)1_{A^c}] \\
&= \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_A] + E[g(U)h(V) \mid (U,V) \in F(A^c)]P(A^c) \\
&= \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_A] && \text{Since } P(A^c) = 0 \\
&= E[g(U)h(U)1_A] && \text{by equation 20} \\
&= E[g(U)h(U)]
\end{aligned}$$

By equation 4, equation 8 follows.

Case 2: Suppose $P(A) = 0$.

$$\begin{aligned}
&\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)] \\
&= E[g(U)h(1-U)1_{A^c}] && \text{by equation 21} \\
&= E[g(U)h(1-U)]
\end{aligned}$$

By equation 4, equation 9 follows.

Case 3: Suppose $0 < P(A) < 1$.

$$\begin{aligned}
&\sup_{(U,V)\sim J(U,V)} E[g(U)h(V)] \\
&\leq \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_A] + \sup_{(U,V)\sim J(U,V)} E[g(U)h(V)1_{A^c}] \\
&= E[g(U)h(U)1_A] + E[g(U)h(1-U)1_{A^c}] && \text{by equations 20 and 21}
\end{aligned}$$

By equation 4, equation 10 follows. □

References

[1] David Aldous and Persi Diaconis. "Strong uniform times and finite random walks". In: *Advances in Applied Mathematics* 8.1 (1987), pp. 69–97. DOI: 10.1016/0196-8858(87)90006-6.

[2] Niloy Biswas and Lester Mackey. "Bounding Wasserstein distance with couplings". In: (Dec. 2021). DOI: arXiv:2112.03152.

[3] Niloy Biswas et al. "Coupling-based Convergence Assessment of some Gibbs Samplers for High-Dimensional Bayesian Regression with Shrinkage Priors". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.3 (Mar. 2022), pp. 973–996. DOI: 10.1111/rssb.12495.

[4] Simon Breneis. "Functions of bounded variation in one and multiple dimensions". PhD thesis. Johannes Kepler University Linz, 2020.

[5] Steve Brooks et al., eds. *Handbook of Markov chain Monte Carlo*. Chapman & Hall, 2011.

[6] Brian S. Caffo, James G. Booth, and A. C. Davison. "Empirical Supremum Rejection Sampling". In: *Biometrika* 89.4 (2002), pp. 745–754. DOI: 10.1093/biomet/89.4.745.

[7] Stamatis Cambanis, Gordon Simons, and William Stout. "Inequalities for Ek(X, Y) when the marginals are fixed". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 36.4 (1976), pp. 285–294. DOI: 10.1007/BF00532695.

[8] Pietro Caputo, Cyril Labbé, and Hubert Lacoin. "Mixing time of the adjacent walk on the simplex". In: *The Annals of Probability* 48.5 (2020), pp. 2449–2493. DOI: 10.1214/20-AOP1428.

[9] Mary Katherine Cowles and Jeffrey S. Rosenthal. "A simulation approach to convergence rates for Markov chain Monte Carlo algorithms". In: *Statistics and Computing* 8 (1998), pp. 115–124. DOI: 10.1023/A:1008982016666.

[10] Mary Kathryn Cowles and Bradley P. Carlin. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *Journal of the American Statistical Association* 91 (1996), pp. 883–904. DOI: 10.2307/2291683.

[11] Liyi Dai. "Convergence rates of finite difference stochastic approximation algorithms part II: implementation via common random numbers". In: *Sensing and Analysis Technologies for Biomedical and Cognitive Applications*. Vol. 9871. SPIE, 2016. DOI: 10.1117/12.2228251.

[12] Persi Diaconis and David Freedman. "Iterated random functions". In: *SIAM Review* 41.1 (1999), pp. 45–76. DOI: 10.1137/S0036144598338446.

[13] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. 3rd ed. New York: Texts in Statistical Science. Chapman and Hall, 2008. DOI: 10.1201/9780367807849.

[14] Rick Durrett. *Probability: Theory and Examples*. 4th ed. New York: Cambridge Series in Statistical and Probabilistic Mathematics, 2010. DOI: 10.5555/1869916.

[15] Charles J. Geyer. "Introduction to Markov Chain Monte Carlo". In: *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall/CRC, 2011, pp. 1–46. DOI: 10.1201/b10905.

[16] Ramen Ghosh and Jakub Marecek. "Iterated Function Systems: A Comprehensive Survey". In: (Nov. 2022). DOI: arXiv.2211.14661.

[17] Alison L. Gibbs. "Convergence in the Wasserstein Metric for Markov Chain Monte Carlo Algorithms with Applications to Image Restoration". In: *Stochastic Models* 20.4 (2004), pp. 473–492. DOI: 10.1081/STM-200033117.

[18] Paul Glasserman and David D. Yao. "Some guidelines and guarantees for common random numbers". In: *Management Science* 38 (1992), pp. 884–908. DOI: 10.5555/2909967.2909978.

[19] Jeremy Heng and Pierre E. Jacob. "Unbiased Hamiltonian Monte Carlo with couplings". In: *Biometrika* 106.2 (2019), pp. 287–302. DOI: 10.1093/biomet/asy074.

[20] James P. Hobert and Galin L. Jones. "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo". In: *Statistical Science* 16.4 (2001), pp. 312–334. DOI: 10.1214/ss/1015346317.

[21] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. New York: Springer, 2009. DOI: 10.1007/978-0-387-92407-6.

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

[23] S. F. Jarner and R. L. Tweedie. "Locally Contracting Iterated Functions and Stability of Markov Chains". In: *Journal of Applied Probability* 38.2 (2001), pp. 494–507. DOI: 10.1239/jap/996986758.

[24] Zhumengmeng Jin and James P. Hobert. "Dimension free convergence rates for Gibbs samplers for Bayesian linear mixed models". In: *Stochastic Processes and their Applications* 148 (2022), pp. 25–67. DOI: 10.1016/j.spa.2022.02.003.

[25] Valen E. Johnson. "Studying convergence of Markov Chain Monte Carlo algorithms using coupled sample paths". In: *Journal of the American Statistical Association* 91 (1996), pp. 154–166. DOI: 10.1080/01621459.1996.10476672.

[26] Anthony Lee, Sumeetpal S. Singh, and Matti Vihola. "Coupled conditional backward sampling particle filter". In: *The Annals of Statistics* 48.5 (2020), pp. 3066–3089. DOI: 10.1214/19-AOS1922.

[27] Krzysztof Leśniak, Nina Snigireva, and Filip Strobin. "Weakly contractive iterated function systems and beyond: a manual". In: *Journal of Difference Equations and Applications* 26.8 (2020), pp. 1114–1173. DOI: 10.1080/10236198.2020.1760258.

[28] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society, 2017. DOI: 10.1090/mbk/107.

[29] Neal Madras and Denise Sezer. "Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances". In: *Bernoulli* 16.3 (2010), pp. 882–908. DOI: 10.2307/25735016.

[30] Tamás P. Papp and Chris Sherlock. "A new and asymptotically optimally contracting coupling for the random walk Metropolis". In: (Nov. 2022). DOI: arXiv:2211.12585.

[31] Alexander S. Poznyak. *Advanced Mathematical Tools for Automatic Control Engineers: Deterministic Techniques*. Oxford: Elsevier, 2008. DOI: 10.1016/B978-008044674-5.50001-8.

[32] Qian Qin and James P. Hobert. "Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 58.2 (2022), pp. 872–889. DOI: 10.1214/21-AIHP1195.

[33] Qian Qin and James P. Hobert. "Wasserstein-based methods for convergence complexity analysis of MCMC with applications". In: *The Annals of Applied Probability* 32.1 (2022), pp. 124–166. DOI: 10.1214/21-AAP1673.

[34] Svetlozar T. Rachev et al. *Bayesian Methods in Finance*. New York: Wiley and Sons, 2008. DOI: 10.1002/9781119202141.

[35] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability surveys* 1 (2004), pp. 20–71.

[36] Gareth O. Roberts and Jeffrey S. Rosenthal. "Markov Chains and De-Initializing Processes". In: *Scandinavian Journal of Statistics* 28.3 (2001), pp. 489–504.

[37] Gareth O. Roberts and Jeffrey S. Rosenthal. "One-shot coupling for certain stochastic recursive sequences". In: *Stochastic Processes and their Applications* 99 (2002), pp. 195–208. DOI: 10.1016/S0304-4149(02)00096-0.

[38] Sheldon M. Ross. *Introduction to Probability Models*. 10th ed. Academic Press, 2010.

[39] Vivekananda Roy. "Convergence Diagnostics for Markov Chain Monte Carlo". In: *Annual Review of Statistics and Its Application* 7.1 (2020), pp. 387–412. DOI: 10.1146/annurev-statistics-031219-041300.

[40] H. L. Roydon. *Real Analysis*. 2nd ed. Macmillian, 1968.

[41] Filippo Santambrogio. "Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling". In: (2015). DOI: 10.1007/978-3-319-20828-2.

[42] Sabrina Sixta and Jeffrey S. Rosenthal. "Convergence rate bounds for iterative random functions using one-shot coupling". In: *Statistics and Computing* 32.71 (2022), pp. 1573–1375. DOI: 10.1007/s11222-022-10134-x.

[43] Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton lectures in analysis. Princeton, NJ: Princeton Univ. Press, 2005. DOI: 10.2307/j.ctvd58v18.

[44] David Steinsaltz. "Locally Contractive Iterated Function Systems". In: *Annals of Probability* 27.4 (1999), pp. 1952–1979. DOI: 10.1214/aop/1022874823.

[45] Örjan Stenflo. "A Survey of Average Contractive Iterated Function Systems". In: *Journal of Difference Equations and Applications* 18.8 (2012), pp. 1355–1380. DOI: 10.1080/10236198.2011.610793.

[46] Luke Tierney. "Markov chains for exploring posterior distributions". In: *Annals of Statistics* 22 (1994), pp. 1701–1728.

[47] Guanyang Wang, John O'Leary, and Pierre E. Jacob. "Maximal Couplings of the Metropolis-Hastings Algorithm". In: *International Conference on Artificial Intelligence and Statistics*. 2021, pp. 1225–1233.

[48] Ward Whitt. "Bivariate Distributions with Given Marginals". In: *The Annals of Statistics* 4.6 (1976), pp. 1280–1289. DOI: 10.1214/aos/1176343660.

[49] James R. Wilson. "Antithetic Sampling with Multivariate Inputs". In: *American Journal of Mathematical and Management Sciences* 3.2 (1983), pp. 121–144. DOI: 10.1080/01966324.1983.10737119.

[50] Kai Xu et al. "Couplings for Multinomial Hamiltonian Monte Carlo". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* 130 (2021), pp. 3646–3654.