Common random number simulation as the optimal coordinatewise coupling for iterated random functions

Sabrina Sixta* and Jeffrey S. Rosenthal†

Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

November 1, 2026

Abstract

In this paper, we extend the results of [19] to demonstrate that common random number (CRN) simulation applied to coordinatewise monotone iterated random functions (IRF) represents the optimal allocation of random numbers to minimize the expected distance of an n-step Markov chain (MC) in $\mathbb{R}^q$. We introduce this concept as the optimal coordinatewise Wasserstein distance (CWD) coupling. We provide specific examples of IRFs where CRN is the optimal CWD coupling, as well as cases where it is not. For IRFs that are not coordinatewise monotone, we illustrate how CRN applied to equimeasurable transformations can result in even lower expected distances compared to CRN applied to the original IRF.

Keywords: common random number, synchronous coupling, Wasserstein distance, Markov chains, MCMC, iterated random functions.

2020 Mathematics Subject Classification: Primary 60J05

Secondary 49Q22, 60J20

*email: sabrina.sixta@mail.utoronto.ca

†email: jeff@math.toronto.edu

Contents

1 Introduction **3**

2 Background **6**

2.1 Notation and definitions 6

2.2 Applying CRN to quantile functions 8

3 Coordinatewise Wasserstein distance **11**

4 Conditions when CRN is an optimal coupling for MCs **14**

4.1 Conditions when CRN is an optimal CWD coupling on the first iteration of a MC 14

4.2 Conditions when CRN is an optimal CWD coupling on the nth iteration of a MC 18

5 Examples **19**

5.1 Random logistic maps 19

5.2 Zeno's walk 20

5.3 Autoregressive normal process in $\mathbb{R}^q$ 22

5.4 Bayesian location model Gibbs sampler 23

5.5 Linear ARCH process 25

6 When monotone conditions are not satisfied **25**

6.1 Equimeasurable functions 26

6.2 Example 27

7 Appendix **32**

1 Introduction

Markov chain Monte Carlo (MCMC) is an integral part of statistical computing and Bayesian statistics is its most notable application [21, 7]. An important question when simulating MCMC is how quickly the chain converges to the stationary distribution [21]. The Wasserstein distance is a metric on probability distributions that is a popular method for evaluating whether a chain is close to stationarity [29, 17]. It scales well in high dimensions compared to drift and minorization generated total variation bounds [29] and metrizes convergence in distribution in $P_k(\mathcal{X})$ [42].

Various methods for obtaining theoretical upper bounds on the Wasserstein distance between the distribution of a Markov chain (MC) and its stationary distribution have been developed [29, 28, 39, 23]. However, such bounds can involve complex calculations, be difficult to implement, and are unavailable for many MCMC models of interest. An alternative method for estimating upper bounds on the Wasserstein distance is through common random number (CRN) simulation [19, 3, 38] alternatively called synchronous coupling [1, 8].

We consider CRN in the setting of iterated random functions (IRFs). A MC $\{X_n\}_{n\geqslant 0}$ can be represented as a sequence of IRFs [14, 39, 40, 16, 25], where

$$X_n = f(\theta_n, X_{n-1})$$

and $\{\theta_n\}_{n\geqslant 1}$ is a sequence of i.i.d. random variables. Consider two copies of the chain, $\{X_n\}_{n\geqslant 0}$ and $\{Y_n\}_{n\geqslant 0}$, initialized at x_0 and y_0, respectively. Under CRN, the same realization of θ_n is used to update both chains:

$$X_n = f(\theta_n, X_{n-1}), \qquad Y_n = f(\theta_n, Y_{n-1}).$$

Thus, CRN induces a coupling between the two copies of the chain.

The optimality of CRN as the joint distribution that minimizes the expected distance has

been established in several settings. As early as 1934, it was shown that if $\theta, \theta' \sim Unif(0,1)$ and f_X and f_Y are quantile functions, then CRN simulation minimizes the Euclidean distance [20]. Proposition 2.2 of [19] extends this result to coordinatewise increasing transforms. Related results have been established for stochastic processes and specific MCs. For example, [8] shows that, under suitable conditions, two reflected Brownian motion processes simulated using CRN converge to one another almost surely. [1] shows that CRN generates the optimal Wasserstein coupling between the solutions of stochastic differential equations. [13] shows that CRN generates a maximal coupling with respect to total variation distance for the independent Metropolis algorithm. However, the conditions under which CRN induces an optimal coupling that minimizes the expected distance between two copies of a Markov chain represented as an iterated random function have not been fully characterized.

In this paper we provide sufficient conditions for when CRN applied to an IRF minimizes the expected distance between two copies of an n-step MC in $\mathbb{R}^q$ (see Theorem 4.8). In particular, we generalize the results of [19] to quasi-antitone distance functions and extend CRN optimality results from single-step transformations to iterated random functions. We show that CRN could be the optimal coupling for a 1-step IRF, but not a 2-step IRF (See Lemma 5.2 compared to Numerical Example 6.4). When our sufficient conditions are not satisfied, we demonstrate how alternative couplings can yield smaller expected distances (Numerical Example 6.4).

In Section 2.2, we summarize the useful properties of CRN when the random transform is the quantile function (ie., the inverse distribution function). We justify why the Wasserstein distance is the natural measure for studying the convergence properties of random values simulated using CRN. We also introduce Propostion 2.1 (Theorem 2 of [9]), which lays the foundations for the rest of the theorems in the paper.

In Section 3, we first show in Theorem 3.2 that the Wasserstein distance between the laws of $\vec{X} = f(\vec{\theta}, \vec{x})$ and $\vec{Y} = f(\vec{\theta'}, \vec{y})$ is attained by minimizing with respect to the joint distribution between $\vec{\theta}$ and $\vec{\theta'}$. Next, we introduce the concept of coordinatewise Wasserstein

distance (CWD) for IRFs, Definition 3.3. If we write the MC as an IRF, the CWD is the infimum expected distance between $f(\vec{\theta}, \vec{x})$ and $f(\vec{\theta}', \vec{y})$ with respect to a product measure between $\vec{\theta}$ and $\vec{\theta}'$, $\bigotimes_{i=1}^{d} \mathcal{L}(\theta_i, \theta_i')$. We also provide an example of when the Wasserstein distance is strictly smaller than the CWD (Example 3.4).

In Section 4.1, we identify when CRN is an optimal CWD coupling on the first iteration of a MC. In Theorem 4.7 we show that if $f(\vec{\theta}, \vec{x})$ is increasing coordinatewise over $\vec{\theta}$, then CRN on the first iteration of a MC is an optimal CWD coupling.

In Section 4.2 we extend the results of Section 4.1 to provide sufficient conditions for when CRN is the optimal CWD coupling on the nth iteration of a MC. In Theorem 4.8 we show that if $f(\vec{\theta}, \vec{x})$ is increasing coordinatewise with respect to both $\vec{\theta}$ and $\vec{x}$, then CRN on the nth iteration of a MC is an optimal CWD coupling.

In Section 5, we present a range of examples illustrating when CRN is, and is not, an optimal CWD coupling. We provide examples where CRN is the optimal CWD coupling for a 1-step chain and an n-step chain (see Section 5). We show that CRN could be the optimal CWD coupling for a 1-step chain, but not an n-step chain (see Example 5.1 compared to Numerical Example 6.4).

In Section 6, we look at ways to lower the expected distance when CRN is not an optimal CWD coupling. We propose generating an equimeasurable function that is monotone in $\vec{\theta}$, which will generate the optimal CWD coupling when CRN is applied. We provide two methods of generating equimeasurable transforms (see Lemmas 6.2 and 6.3) and apply our lemmas to an example.

First, we introduce notation in the Section 2.1. This section should be referenced if the reader wants clarity on notation throughout the paper.

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

2 Background

2.1 Notation and definitions

Markov chain notation Let $\{X_n\}_{n\geqslant 1}$ be a MC such that X_n can be written as an iterative random function ie. $X_n = f(\vec{\theta}_n, X_{n-1}) \in \mathbb{R}$ and $\vec{\theta}_n$ is a random variable. Define $\Theta_n = (\vec{\theta}_1, \ldots, \vec{\theta}_n)$ where $\vec{\theta}_1, \ldots, \vec{\theta}_n$ are i.i.d. random variables. X_n can then be written as follows for initial value x_0.

$$X_n = f(\vec{\theta}_n, f(\vec{\theta}_{n-1}, \ldots, f(\vec{\theta}_1, x_0)\ldots)) = f_n(\Theta_n, x_0) \tag{1}$$

Any time-homogeneous MC can be represented as an IRF [40]. X_n is the state of the MC at iteration n.

$\mathcal{L}(X_n)$ denotes the law of X_n and $\mathcal{L}(X_n) \Rightarrow \pi$ denotes weak convergence of the law of X_n to a stationary distribution π. We write that $x \in \text{supp}(\pi)$ if x is in the support of π.

More generally, an $\mathbb{R}^q$-valued Markov chain $\{\vec{X}_n\}_{n\geqslant 1}$ can be represented as $\vec{X}_n = f(\vec{\theta}_n, \vec{X}_{n-1}) \in \mathbb{R}^q$ where, for each coordinate $l \in 1, \ldots, q$, $X_n^{(l)} = f^{(l)}(\vec{\theta}_n, \vec{X}_{n-1})$.

Pushforward measure For a fixed $\vec{x}$, let $f_{\vec{x}}(\cdot) := f(\cdot, \vec{x})$. The pushforward of μ under $f_{\vec{x}}$ is denoted by $(f_{\vec{x}})_*\mu : \mathcal{B}(\mathbb{R}^q) \to [0,1]$. For any $B \in \mathcal{B}(\mathbb{R}^q)$, $(f_{\vec{x}})_*\mu(B) := \mu[(f_{\vec{x}})^{-1}(B)]$. Since $\vec{\theta}_n \sim \mu$ and $\vec{X}_n = f(\vec{\theta}_n, \vec{X}_{n-1})$, $\mathcal{L}(\vec{X}_n \mid \vec{X}_{n-1} = \vec{x}) = (f_{\vec{x}})_*\mu$. See Section 2.4 of [6] for more details on pushforward measures.

CRN CRN simulation occurs when two Markov chains $\{X_n\}_{n\geqslant 1}$ and $\{Y_n\}_{n\geqslant 1}$, with different initial values $x_0 \neq y_0$, are simulated using the same IRF and the same realizations of the random inputs. That is,

$$X_n = f(\vec{\theta}_n, X_{n-1}) \qquad \text{and} \qquad Y_n = f(\vec{\theta}_n, Y_{n-1}), \qquad n \geqslant 1,$$

use the same $\vec{\theta}_n$ at each iteration. Equivalently,

$$X_n = f_n(\Theta_n, x_0) \qquad \text{and} \qquad Y_n = f_n(\Theta_n, y_0), \qquad n \geqslant 1,$$

where the same Θ_n is used for both chains. The CRN coupling depends on the choice of IRF representation and is therefore not uniquely determined by the Markov chain [19]. Some treatments restrict CRN to the inverse distribution function (quantile) representation [19, 22], as considered in Section 2.2.

Distance function k: Let k be a distance function.

k is quasi-antitone (in other words $-k$ is supermodular) if for all $x \leqslant x'$, $y \leqslant y'$,

$$k(x, y) + k(x', y') - k(x, y') - k(x', y) \leqslant 0$$

We will need k to be quasi-antitone to prove that CRN is an optimal coupling. We write that k is QAC if k is a quasi-antitone continuous distance function.

The distance function corresponding to the ℓ^p norm $p \geqslant 1$, $||\vec{x} - \vec{y}||_p = \left(\sum_{i=1}^{q} |x_i - y_i|^p\right)^{1/p}$ is quasi-antitone with respect to each pair of coordinates. Theorem 3 of [12] establishes that continuous functions $h(x - y)$ where h is convex and $h(0) = 0$ are QAC. Thus, $h(x_i - y_i) = |x_i - y_i|^p$ is QAC.

Wasserstein distance We define $\mathcal{L}(X, Y)$ as a joint distribution of random variables X, Y and $\mathcal{J}(\mu, \nu) = \{\mathcal{L}(X, Y) \mid \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu\}$ to be the set of all joint distribution functions with marginals μ, ν. The Wasserstein distance of order p between distributions μ, ν with distance function k is defined as $W_{k,p}(\mu, \nu) = \inf_{\gamma \in \mathcal{J}(\mu,\nu)} E_\gamma\big[k(X, Y)^p\big]^{1/p}$ for $p \in [1, \infty)$. If $p = 1$, then we denote the Wasserstein distance as $W_k(\mu, \nu)$. The joint distribution $\gamma_{opt} \in \mathcal{J}(\mu, \nu)$ is an optimal Wasserstein distance coupling if $E_{\gamma_{opt}}[k(X, Y)] = W_k(\mu, \nu)$ [42].

We define the Wasserstein space $P_k(\mathcal{X})$ as the space of Borel probability measures μ on

the metric space $(\mathcal{X}, k)$ such that $\int_{\mathcal{X}} k(x_0, x), d\mu(x) < \infty$ for some $x_0 \in \mathcal{X}$.

Define a sequence of random variables $\{X_n\}_{n \geqslant 1}$ in a Polish metric space $(\mathcal{X}, k)$ with distributions in $\mathcal{L}(X_n) \in P_k(\mathcal{X})$. Define $\mathcal{L}(X) = \pi \in P_k(\mathcal{X})$. X_n converges weakly to π in $P_k(\mathcal{X})$ (denoted as $\mathcal{L}(X_n) \Rightarrow \pi$ in $P_k(\mathcal{X})$) if (1) $\mathcal{L}(X_n) \Rightarrow \pi$ and (2) for some $x_0 \in \mathcal{X}$, $E[k(x_0, X_n)] \to E[k(x_0, X)]$ (see alternative conditions in Definition 6.8 of [42]).

2.2 Applying CRN to quantile functions

We first discuss the simplest case of CRN: when random variables $X, Y \in \mathbb{R}$ are written as the quantile transform of a uniform random variable. We define the quantile function as $F^{-1}(u) = \inf\{x \mid u \leqslant F(x)\}$, and our random variables as functions of the quantile function, $X = F_\mu^{-1}(U) \sim \mu$ and $Y = F_\nu^{-1}(V) \sim \nu$ where $U, V \sim Unif(0, 1)$. Proposition 2.1 specifies the range of possible expected distances $E[k(X, Y)]$ over all elements in $\mathcal{J}(\mu, \nu)$.

Proposition 2.1 lays the foundation for the rest of the theorems in this paper. Namely Theorem 4.2 uses Proposition 2.1 to prove that CRN is the optimal coupling for the Wasserstein distance when X is an increasing function of U.

Proposition 2.1 (Theorem 2 of [9]). *Suppose that $X, Y \in \mathbb{R}$ are two random variables with corresponding distribution functions F_μ, F_ν such that $X \sim \mu$ and $Y \sim \nu$. Suppose that k is a QAC metric function and that $\mu, \nu \in P_k(\mathbb{R})$. Then the set of possible expected distances over all couplings in $\mathcal{J}(\mu, \nu)$ is the following interval.*

$$\big\{E[k(X, Y)] \mid \mathcal{L}(X, Y) \in \mathcal{J}(\mu, \nu)\big\} = \big[E[k(F_\mu^{-1}(U), F_\nu^{-1}(U))], E[k(F_\mu^{-1}(U), F_\nu^{-1}(1 - U))]\big]$$

Proof. The result follows from Theorem 2 of [9]. Note that condition (a) of Theorem 2 of [9] is satisfied, because k is symmetric and $\mu, \nu \in P_k(\mathbb{R})$, which means that for some $x_0 \in \mathbb{R}$, $E[k(X, Y)] \leqslant E[k(X, x_0)] + E[k(x_0, Y)] < \infty$. □

Propositions 2.2 and 2.3, which are a combination of established results, summarize why we focus on Wasserstein distance to study convergence of CRN applied to quantile functions.

Namely, that CRN applied to the quantile function metrizes convergence in distribution in $P_k(\mathbb{R})$.

The importance of this observation lies in the fact that it identifies the Wasserstein distance as a natural metric for studying convergence of random variables generated via CRN. This is in contrast to the total variation distance, which is the metric used to analyze CRN convergence in [24, 13]. Total variation distance is strictly stronger than convergence in distribution (Theorem 6 of [18]). Indeed, convergence in total variation distance requires that $P(X = Y) \to 1$, which is equivalent to the existence of a coupling under which the two random variables become equal with probability tending to one [30]. In most applications of CRN for real-valued random variables, however, the coupled random variables are not equal with probability tending to 1; rather, they are constructed to be strongly positively dependent.

Proposition 2.2. *Define the random variables* $\{X_n\}_{n\geqslant 1}$ *and* X *in a Polish metric space* $(\mathbb{R}, k)$ *such that* $\sup_n E[k(X_n, x_0)] < \infty$. *The following are equivalent where* $U \sim \text{Unif}(0, 1)$.

1. $\mathcal{L}(X_n) \Rightarrow \pi$ *in* $P_k(\mathbb{R})$
2. $W_k(\mathcal{L}(X_n), \mathcal{L}(X)) \to 0$
3. $E[k(F_{X_n}^{-1}(U), F_X^{-1}(U))] \to 0$

Proof. (1) $\iff$ (2): by Theorem 6.9 of [42].

(2) $\iff$ (3): By Proposition 2.1, the Wasserstein distance is attained when CRN is applied for each n, $W_k(\mathcal{L}(X_n), \mathcal{L}(X)) = E[k(F_{X_n}^{-1}(U), F_X^{-1}(U))]$. □

Proposition 2.3 shows that CRN applied to quantile functions can provide a stronger form of convergence than convergence in the Wasserstein distance. In particular, almost sure convergence of the CRN-coupled random variables follows from weak convergence of their distributions, without requiring convergence in $P_k(\mathbb{R})$. Thus, under the quantile coupling, weak convergence of the distributions can be strengthened to almost sure convergence of the coupled random variables.

Proposition 2.3 is closely related to Lemma 21.2 of [41] and to the proof of Skorohod's representation theorem.

Proposition 2.3. *Let $\{X_n\}_{n\geqslant 1}$ be random variables in the metric space $(\mathbb{R}, |\cdot|)$ and π be a distribution. Then the following holds,*

$$\mathcal{L}(X_n) \Rightarrow \pi \iff |F_n^{-1}(U) - F^{-1}(U)| \to 0 \quad \textit{almost surely}$$

Where F_n and F are the CDF of $\mathcal{L}(X_n)$ and π, respectively, and $U \sim Unif(0,1)$.

Proof. See proof in Subsection 7.1, which closely follows the proof of Skorohod's theorem (see Theorem 25.6 in [2]). □

The following example provides a case where $\mathcal{L}(X_n) \Rightarrow \pi$ in distribution but not in $P_k(\mathbb{R})$, so Proposition 2.2 does not hold, but Proposition 2.3 holds.

Example 2.4. Let $\{X_n\}_{n\geqslant 1}$ be a sequence such that for each $n \geqslant 1$, $P(X_n = n) = \frac{1}{n}$ and $P(X_n = 0) = 1 - \frac{1}{n}$. $\mathcal{L}(X_n)$ converges weakly to the Dirac mass at point 0, $\mathcal{L}(X_n) \Rightarrow \delta_0$. In this case F_n and F_n^{-1} are defined as follows,

$$F_n(x) = \begin{cases} 0 & x < 0 \\ 1 - \frac{1}{n} & 0 \leqslant x < n \\ 1 & n \leqslant x \end{cases} \qquad F_n^{-1}(u) = \begin{cases} 0 & 0 \leqslant u \leqslant 1 - \frac{1}{n} \\ n & 1 - \frac{1}{n} < u \leqslant 1 \end{cases}$$

$\mathcal{L}(X_n)$ does not converge weakly to δ_0 in $P_k(\mathbb{R})$: Fix a point x_0, $\lim_{n\to\infty} E[|x_0 - X_n|] = \lim_{n\to\infty} |x_0 - 0|(1 - \frac{1}{n}) + |x_0 - n|\frac{1}{n} = |x_0| + 1 \neq |x_0 - 0|$.

Proposition 2.3 applies: First note that $F_{\delta_0}^{-1}(u) = 0$ for all $u \in (0, 1]$. Fix $u \in (0, 1)$, then for $n > \frac{1}{1-u}$, $F_n^{-1}(u) = 0 = F_{\delta_0}^{-1}(u)$. So F_n^{-1} converges pointwise to $F_{\delta_0}^{-1}$ for all $u \in (0, 1)$. Since $P(U \in (0, 1)) = 1$, $|F_n^{-1}(U) - F_{\delta_0}^{-1}(U)| \to 0$ w.p. 1.

Bullet (3) of Proposition 2.2 does not apply: $E[|X_n - 0|] = |n-0|\frac{1}{n} + |0-0|(1-\frac{1}{n}) = 1 \not\to 0$.

While the results related to the convergence of inverse transforms are elegant, they are generally impractical for studying the optimality of CRN applied to MCs. We usually do

not know the quantile function of X_n. Further, we can only apply inverse transforms when $X \in \mathbb{R}$ (see Section 2 of [34]). In the next section we will formally define CWD, which will be used to provide sufficient conditions under which CRN is the optimal CWD coupling between two IRFs in $\mathbb{R}^q$.

3 Coordinatewise Wasserstein distance

Let $\vec{X}, \vec{Y} \in \mathbb{R}^q$ be random vectors represented as random functions as follows:

$$\vec{X} = f(\vec{\theta}, \vec{x}), \qquad \vec{Y} = f(\vec{\theta'}, \vec{y}),$$

Where $\vec{\theta}, \vec{\theta'}$ are defined as in Definition 3.1.

Definition 3.1 $(\vec{\theta})$**.** Let $\mathcal{X}_i$, $i = 1, \ldots, d$, be topological spaces, each equipped with a total order relation (ie. for any $\theta_i, \theta_i' \in \mathcal{X}_i$ either $\theta_i \leqslant \theta_i'$ or $\theta_i' \leqslant \theta_i$). Let $\vec{\theta} = (\theta_1, \ldots, \theta_d)$ be a random vector taking values in the product space $\prod_{i=1}^d \mathcal{X}_i$, with independent coordinates $\theta_i \sim \mu_i$. Then $\mathcal{L}(\vec{\theta}) = \mu = \bigotimes_{i=1}^d \mu_i$.

We first show that the infimum expected distance between $\vec{X}$ and $\vec{Y}$ over all couplings of their distributions is equal to the infimum expected distance between $f(\vec{\theta}, \vec{x})$ and $f(\vec{\theta'}, \vec{y})$ over all couplings of $\vec{\theta}$ and $\vec{\theta'}$. Specifically,

$$\inf_{\gamma \in \mathcal{J}(\mathcal{L}(\vec{X}), \mathcal{L}(\vec{Y}))} E_{(\vec{X}, \vec{Y}) \sim \gamma}[k(\vec{X}, \vec{Y})] = \inf_{\gamma \in \mathcal{J}(\mu, \mu)} E_{(\vec{\theta}, \vec{\theta'}) \sim \gamma} \left[k\big(f(\vec{\theta}, \vec{x}), f(\vec{\theta'}, \vec{y})\big) \right]. \tag{2}$$

The equality in Equation 2 is formalized in the following theorem.

Theorem 3.2. *Suppose that $\vec{X} = f(\vec{\theta}, \vec{x})$ and $\vec{Y} = f(\vec{\theta'}, \vec{y}) \in \mathbb{R}^q$ where $\vec{\theta}, \vec{\theta'}$ are defined in Definition 3.1 and $f(\cdot, \vec{x})$ and $f(\cdot, \vec{y})$ are measurable functions. Then,*

$$W_k(\mathcal{L}(\vec{X}), \mathcal{L}(\vec{Y})) = \inf_{\gamma \in \mathcal{J}(\mu, \mu)} E_{(\vec{\theta}, \vec{\theta'}) \sim \gamma}[k(f(\vec{\theta}, \vec{x}), f(\vec{\theta'}, \vec{y}))] \tag{3}$$

The proof is in Section 7.2

However, similar to [19], the infima in the theorems of this paper are not taken over all couplings in $\mathcal{J}(\mu,\mu)$, so Theorem 3.2 does not apply. Instead, we restrict the infimum to couplings in $\mathcal{J}_{IND}(\mu,\mu) \subset \mathcal{J}(\mu,\mu)$, which allow dependence only between corresponding components of $\vec{\theta}$ and $\vec{\theta'}$.

$$\mathcal{J}_{IND}(\mu,\mu) = \left\{ \gamma \in P(\mathcal{X}^d \times \mathcal{X}^d) \mid \gamma_i \in \mathcal{J}(\mu_i,\mu_i) \text{ and } \gamma = \bigotimes_{i=1}^{d} \gamma_i \right\}$$

This class is denoted by $\mathcal{M}_0(\mu,\mu)$ in [19]. The infimum in Proposition 2.2 of [19] and in Theorems 4.6, 4.7, and 4.8 is taken over this restricted class of couplings.

Next, we introduce the coordinatewise Wasserstein distance, which is the infimum expected distance over all couplings in $\mathcal{J}_{IND}(\mu,\mu)$.

Definition 3.3 (coordinatewise Wasserstein distance (CWD))**.** The coordinatewise Wasserstein distance for IRFs is the infimum expected distance between $\vec{X} = f(\vec{\theta},\vec{x})$ and $\vec{Y} = f(\vec{\theta'},\vec{y})$ over the set of distributions $\mathcal{J}_{IND}(\mu,\mu)$.

$$\widetilde{W}_k(\mathcal{L}(\vec{X}),\mathcal{L}(\vec{Y})) = \inf_{\gamma \in \mathcal{J}_{IND}(\mu,\mu)} E_{(\vec{\theta},\vec{\theta'})\sim\gamma}[k(f(\vec{\theta},\vec{x}),f(\vec{\theta'},\vec{y}))]$$

Note that $\widetilde{W}_k(\mathcal{L}(\vec{X}),\mathcal{L}(\vec{Y})) := \widetilde{W}_k((f_{\vec{x}})_*\mu,(f_{\vec{y}})_*\mu)$, with notation being used interchangeably throughout the paper. $\gamma_{opt} \in \mathcal{J}_{IND}(\mu,\mu)$ is an optimal CWD coupling if

$$E_{(\vec{\theta},\vec{\theta'})\sim\gamma_{opt}}[k(f(\vec{\theta},\vec{x}),f(\vec{\theta'},\vec{y}))] = \widetilde{W}_k((f_{\vec{x}})_*\mu,(f_{\vec{y}})_*\mu)$$

Since the CWD minimizes over $\mathcal{J}_{IND}(\mu,\mu) \subset \mathcal{J}(\mu,\mu)$, it is greater than or equal to the corresponding Wasserstein distance. That is,

$$W_k(\mathcal{L}(\vec{X}),\mathcal{L}(\vec{Y})) \leqslant \widetilde{W}_k(\mathcal{L}(\vec{X}),\mathcal{L}(\vec{Y})) \tag{4}$$

However, equality can hold under additional structural conditions. In particular, [19] establishes equality with respect to the Euclidean distance cost for certain queueing systems that satisfy coordinatewise monotonicity and permutability.

If $d = 1$ in Definition 3.1, Equation 4 becomes an equality since $\mathcal{J}_{IND}(\mu, \mu) = \mathcal{J}(\mu, \mu)$. That is,

$$W_k(\mathcal{L}(X), \mathcal{L}(X')) = \widetilde{W}_k(\mathcal{L}(X), \mathcal{L}(X')) \tag{5}$$

Theorem 2 of [33] provides an upper bound on the Wasserstein distance between push forward measures $\mathcal{L}(\vec{X}) = (f_{\vec{x}})_*\mu$ and $\mathcal{L}(\vec{Y}) = (f_{\vec{y}})_*\mu$ based on the same initial distributions μ and two different functions $f_{\vec{x}}, f_{\vec{y}}$. The inequality is easy to interpret, but the bound does not appear to be very tight for the random logistics map, which we suspect is the case for other examples as well. Theorem 3.2 of [10] provides an upper bound on the Wasserstein distance of the pushforward measure based on the Wasserstein distance using optimal transport maps.

In Example 3.4, we provide a counterexample to show that the inequality in Equation 4 can be strict if the joint couplings are restricted to $\mathcal{J}_{IND}(\mu, \mu)$.

Example 3.4. Let $\theta_1, \theta_2 \sim \text{Bernoulli}(\frac{1}{2}) = \mu$ such that $\theta_1 \perp\!\!\!\perp \theta_2$. Denote

$$f(\vec{\theta}, x) = (\theta_1 + 2\theta_2)I_{x<2} + (\theta_2 + 2\theta_1)I_{x\geqslant 2}$$

So, $f(\vec{\theta}, 1) = \theta_1 + 2\theta_2$ and $f(\vec{\theta'}, 2) = \theta_2' + 2\theta_1'$. It is clear that $W_k((f_1)_*\mu, (f_2)_*\mu) = 0$ since $(f_1)_*\mu := \mathcal{L}(f(\vec{\theta}, 1)) = \mathcal{L}(f(\vec{\theta}, 2)) := (f_2)_*\mu$.

By Theorem 4.6 and since $f(\cdot, x)$ is coordinatewise monotone and continuous in $\vec{\theta}$ for fixed $x \in \{1, 2\}$, the CWD between the laws of $f(\vec{\theta}, 1)$ and $f(\vec{\theta}, 2)$ is attained when applying CRN,

$$\widetilde{W}_k((f_1)_*\mu, (f_2)_*\mu) = E[|f(\vec{\theta}, 1) - f(\vec{\theta}, 2)|] = E[|\theta_1 - \theta_2|] = \frac{1}{2}$$

This is an example of when the CWD ($\frac{1}{2}$) is strictly greater than the Wasserstein distance

(0).

Finally, when I set $\theta_1 = \theta_2'$ and $\theta_2 = \theta_1'$, the Wasserstein distance is obtained, $E[|f((\theta_1, \theta_2), 1) - f((\theta_2, \theta_1), 2)|] = 0 = W_{|\cdot|}((f_1)_*\mu, (f_2)_*\mu)$. Note that setting $\theta_1 = \theta_2'$ and $\theta_2 = \theta_1'$ is a coupling that is not defined in $\mathcal{J}_{IND}(\mu, \mu)$, but is defined in $\mathcal{J}(\mu, \mu)$.

4 Conditions when CRN is an optimal coupling for MCs

In this section we first provide sufficient conditions for when CRN is an optimal coupling on the first iteration of a MC (Section 4.1) and then extend it to n iterations of a MC (Section 4.2).

4.1 Conditions when CRN is an optimal CWD coupling on the first iteration of a MC

In this section we provide sufficient conditions for when CRN is the optimal coupling on the first iteration. Recall the IRF notation where $\vec{X}_n = f(\vec{\theta}_n, \vec{X}_{n-1}) \in \mathbb{R}^q$ is a MC and $\theta_n \in \prod_{i=1}^d \mathcal{X}_i$ is defined in Definition 3.1. We first start with the simplest case when $d = q = 1$ (Section 4.1.1), extend this to the case when $d \geqslant 1$ and $q = 1$ (Section 4.1.2), and finally generalize this to the case when $d, q \geqslant 1$ (Section 4.1.3). In the simplest case we prove optimal Wasserstein distance coupling using Theorem 3.2 while in the second and third cases we prove optimal CWD coupling.

4.1.1 Conditions when CRN is an optimal Wasserstein distance coupling for $d = q = 1$

The following theorem provides conditions under which the expected distance using CRN between two non-decreasing random functions in $\mathbb{R}$ is equal to the Wasserstein distance.

This theorem is a generalization of Proposition 2.2 of [19], which says that the covariance between two non-decreasing random functions is maximized when generated by CRN. Before we specify the theorem we introduce the concept of monotonicity with respect to θ.

Definition 4.1 (Monotone in θ)**.** Suppose that $f(\theta, x)$ is an IRF where $\theta \sim \mu$ and $x \in \mathbb{R}$. The function f is considered to be monotone in θ, if for all $\theta \leqslant \theta'$ μ-a.e. and for all x π-a.e.

$$f(\theta, x) \leqslant f(\theta', x) \tag{6}$$

Theorem 4.2 states that if the functions $f(\cdot, x)$ and $f(\cdot, y)$ are increasing and defined on the real line, then the expected distance using CRN equals the Wasserstein distance between $\mathcal{L}(X)$ and $\mathcal{L}(Y)$.

Theorem 4.2. *Let X be a random variable such that $\mathcal{L}(X) \in P_k(\mathbb{R})$ and k is a QAC metric. Suppose that $X = f(\theta, x)$ can be written as a random function where $\theta \sim \mu$ is a random variable. Define $Y = f(\theta', y)$ where $\mathcal{L}(\theta') = \mathcal{L}(\theta) = \mu$ and $x, y \in supp(\pi)$.*

Suppose that f is monotone and right-continuous in θ. Then CRN is the optimal Wasserstein distance coupling between $\mathcal{L}(X)$ and $\mathcal{L}(Y)$,

$$W_k(\mathcal{L}(X), \mathcal{L}(Y)) = E[k(f(\theta, x), f(\theta, y))] \tag{7}$$

The proof is in Section 7.3.

Remark 4.3 (Connection to Brenier's Theorem)**.** Within the proof of Theorem 4.2, we used Proposition 2.1, however, we could have also used Brenier's theorem, with specified $k(x, y) = (x - y)^2$, which coincides with Proposition 2.2 of [19]. The following is a simplified statement of Brenier's theorem in $\mathbb{R}$ (Theorem 2.1 of [27], [5]).

> Suppose $X, Y \in \mathbb{R}$ are random variables, $k(x, y) = (x - y)^2$, and $\mathcal{L}(X), \mathcal{L}(Y) \in P_k(\mathbb{R})$. If X has a density with respect to the Lebesgue measure, then the joint distribution

that generates the optimal Wasserstein distance, $\mathcal{L}_{opt}(X, Y)$, can be defined as X and $Y = T(X)$ where T is the derivative of a convex function.

Let $X = g(\theta)$ and $Y = h(\theta')$ where $\mathcal{L}(\theta) = \mathcal{L}(\theta')$. Brenier's theorem implies that if g and h are increasing functions, then CRN is the optimal Wasserstein distance coupling: First note that $k(x, y) = (x - y)^2$ is a QAC function. Next note that if g is an increasing function, then $g^{-1}(x) = \inf\{u \mid g(u) \geqslant x\}$ is an increasing function. The coupling $\mathcal{L}_{opt}(X, Y)$ where $Y = h(g^{-1}(X))$ is optimal according to Brenier's theorem, because $T = h \cdot g^{-1}$ is an increasing function and thus is the derivative of a convex function. Now we remark that $g^{-1}(X) \sim \mathcal{L}(\theta)$ and so for $\theta \sim \mathcal{L}(\theta)$, $X = g(\theta)$ and $Y = h(\theta)$ thus implying that CRN is the optimal Wasserstein distance coupling. This is only a sketch, however, a more thorough explanation can be found in Section 2.1 of [34] and Section 2.3.1 of [22].

If our MC is in $\mathbb{R}^q, q \geqslant 1$, the state space is no longer naturally endowed with a total order and so the definition of monotonicity in $\mathbb{R}$, Definition 4.1, needs additional conditions. Brenier specifies the optimal mapping to be the gradient of a convex function, which can be applied to MCs in $\mathbb{R}^q$. In [19] and in this paper the optimal coupling occurs when CRN is applied to functions that are coordinatewise increasing.

4.1.2 Conditions when CRN is an optimal CWD coupling for $d \geqslant 1$ and $q = 1$

By Definition 3.1 each $\theta_i \in \vec{\theta}$ has an order relation. Below, we define monotonicity over $\vec{\theta}$.

Definition 4.4 (Monotone in θ_i)**.** Let $f(\vec{\theta}, \vec{x}) \in \mathbb{R}^q$ be an iterative random function where $\vec{x} \in \mathbb{R}^q$ and $\vec{\theta}$ is defined as in Definition 3.1. f is considered to be monotone in θ_i, if for all $\theta_i \leqslant \theta_i'$ μ_i-a.e. and $\vec{\theta}_{-i}$ μ_{-i}-a.e., $\vec{x}$ π-a.e.

$$f^{(l)}((\theta_1, \ldots, \theta_i, \ldots, \theta_d), \vec{x}) \leqslant f^{(l)}((\theta_1, \ldots, \theta_i', \ldots, \theta_d), \vec{x}) \tag{8}$$

for all $l \in \{1 \ldots q\}$

Definition 4.5 (Coordinatewise monotone in $\vec{\theta}$)**.** The function f is considered to be coordinatewise monotone in $\vec{\theta}$, if it is monotone in θ_i for all $i \in \{1, \ldots, d\}$.

Next we will provide sufficient conditions for when CRN generates the optimal CWD coupling on the first iteration of a MC when $\vec{\theta}$ is defined as in Definition 3.1. Similar to Theorem 4.2, Theorem 4.6 is a generalization of Proposition 2.2 of [19].

Theorem 4.6. *Let X be a random variable such that $\mathcal{L}(X) \in P_k(\mathbb{R})$ and k is a QAC metric. Suppose that $X = f(\vec{\theta}, x)$ can be written as a random function where $\vec{\theta}$ is defined in Definition 3.1. Let $Y = f(\vec{\theta}', y)$ where $\mathcal{L}(\vec{\theta}') = \mathcal{L}(\vec{\theta}) = \mu$ and $x, y \in supp(\pi)$.*

Suppose further that f is coordinatewise monotone and continuous in $\vec{\theta}$. Then the CWD between $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ is attained when CRN is applied,

$$\widetilde{W}_k(\mathcal{L}(X), \mathcal{L}(Y)) = E[k(f(\vec{\theta}, x), f(\vec{\theta}, y))] \tag{9}$$

The proof is in Section 7.4.

4.1.3 Conditions when CRN is an optimal CWD coupling for $d, q \geqslant 1$

Finally, we generalize Theorem 4.6 to apply to random variables that are defined in $\mathbb{R}^q, q \geqslant 1$.

Theorem 4.7. *Let $\vec{X}$ be a random vector such that $\mathcal{L}(\vec{X}) \in P_{||\cdot||_p^p}(\mathbb{R}^q)$. Suppose that $\vec{X} = f(\vec{\theta}, \vec{x})$ can be written as a random function where $\vec{\theta}$ is defined in Definition 3.1. Let $\vec{Y} = f(\vec{\theta}', \vec{y})$ where $\mathcal{L}(\vec{\theta}) = \mathcal{L}(\vec{\theta}') = \mu$ and $\vec{x}, \vec{y} \in supp(\pi)$.*

Suppose further that f is coordinatewise monotone and continuous in $\vec{\theta}$. Then the CWD between $\mathcal{L}(\vec{X})$ and $\mathcal{L}(\vec{Y})$ is attained when CRN is applied,

$$\widetilde{W}_{||\cdot||_p, p}(\mathcal{L}(\vec{X}), \mathcal{L}(\vec{Y})) = E[||f(\vec{\theta}, \vec{x}) - f(\vec{\theta}, \vec{y})||_p^p]^{1/p} \tag{10}$$

The proof is in Section 7.5.

4.2 Conditions when CRN is an optimal CWD coupling on the nth iteration of a MC

So far we've provided conditions for when CRN is the optimal CWD coupling on the first iteration of the MC. Next, in Theorem 4.8, we show that if a MC is coordinatewise monotone with respect to $\vec{\theta}$ and $\vec{x}$ then CRN is an optimal CWD coupling at any iteration of the MC.

Theorem 4.8. *Let $\{\vec{X}_n\}_{n\geqslant 1}$ be a MC such that $\mathcal{L}(\vec{X}_n) \in P_{||\cdot||_p^p}(\mathbb{R}^q), n \geqslant 1$. Suppose that $\vec{X}_n = f(\vec{\theta}_n, \vec{X}_{n-1}) = f_n(\Theta_n, \vec{x}_0)$ can be written as a random function where where $\Theta_n = (\vec{\theta}_1, \vec{\theta}_2, \ldots, \vec{\theta}_n)$ are i.i.d. with distribution $\vec{\theta}_j \sim \mu = \bigotimes_{i=1}^d \mu_i$ and $\vec{x}_0 \in \mathbb{R}^q$ is the initial value. Suppose that $f(\vec{\theta}, \vec{x})$ is coordinatewise monotone and continuous in $\vec{\theta}$ and $\vec{x}$.*

Define $\vec{Y}_n = f(\vec{\theta}'_n, \vec{Y}_{n-1}) = f_n(\Theta_n, \vec{y}_0)$ where $\mathcal{L}(\vec{\theta}_n) = \mathcal{L}(\vec{\theta}'_n) = \mu$ and $\vec{x}_0, \vec{y}_0 \in supp(\pi)$. Then the CWD between $\mathcal{L}(\vec{X}_n)$ and $\mathcal{L}(\vec{Y}_n)$ is attained when using CRN,

$$\widetilde{W}_{||\cdot||_p,p}(\mathcal{L}(\vec{X}_n), \mathcal{L}(\vec{Y}_n)) = E[||f_n(\Theta_n, \vec{x}_0) - f_n(\Theta_n, \vec{y}_0)||_p^p]^{1/p} \tag{11}$$

The proof is in Section 7.6

Remark 4.9. If $X_n, Y_n \in \mathbb{R}$, then Theorem 4.8 still applies for distance function k QAC and Equation 11 is replaced by the following.

$$\widetilde{W}_k(\mathcal{L}(X_n), \mathcal{L}(Y_n)) = E[k(f_n(\Theta_n, x_0), f_n(\Theta_n, y_0))] \tag{12}$$

5 Examples

The following section presents examples of MCs for which CRN yield an optimal CWD coupling. In the case of the random logistic map (Example 5.1), CRN applied to the one-step transition is optimal; however, for the n-step chain with $n \geqslant 2$, it is no longer optimal. This is demonstrated in Numerical Example 6.4, where we show that CRN fails to be the optimal CWD coupling for the two-step random logistic map. In contrast, for all other

examples considered in this section, CRN provides an optimal CWD coupling for the n-step chain.

5.1 Random logistic maps

Example 5.1 (Random logistic maps)**.** Let $\{X_n\}_{n\geqslant 1} \in [0,1]$ be a random logistic map ([39, 26]). That is for i.i.d. $\beta_n \sim \text{Beta}(\alpha + 1/2, \alpha - 1/2)$ $(\alpha > 1/2)$,

$$X_n = f(\beta_n, X_{n-1}) = 4\beta_n X_{n-1}(1 - X_{n-1})$$

In Lemma 5.2, we show that CRN is an optimal Wasserstein distance coupling on the first iteration of the random logistic map. Section 6.2 illustrates why CRN applied to the 2-step random logistic map is sometimes not the optimal CWD coupling (depending on the initial values).

Lemma 5.2. *CRN is an optimal Wasserstein distance coupling on the first iteration of the random logistic map when k is a QAC metric.*

Proof. Since $4X_{n-1}(1 - X_{n-1}) > 0$ w.p. 1, the function $4\beta_n X_{n-1}(1 - X_{n-1})$ is increasing with respect to β_n. By Theorem 4.2, CRN is an optimal Wasserstein distance coupling on the first iteration of the MC when k is QAC. □

For all $\alpha > 1/2$ the random logistic map converges in distribution to the stationary distribution, $\text{Beta}(\alpha, \alpha)$. For $\alpha \leqslant 1$ the system is not locally contractive, and for $\alpha \geqslant 2$ the system is locally contractive (Proposition 7 of [39]) and geometrically ergodic (Theorem 1 of [39]). What is unknown is whether the expected distance between two MCs simulated using CRN converges to 0 when $\alpha \leqslant 1$.

Figure 1 provides the simulated expected distance when CRN is applied for different $\alpha \in \{0.6, 1.5, 2.5\}$. The red line corresponding to $\alpha = 2.5$ converges quickly to 0, which is expected, because it is geometrically ergodic and locally contractive. The blue line corresponding to $\alpha = 1.5$ also converges quickly to 0.

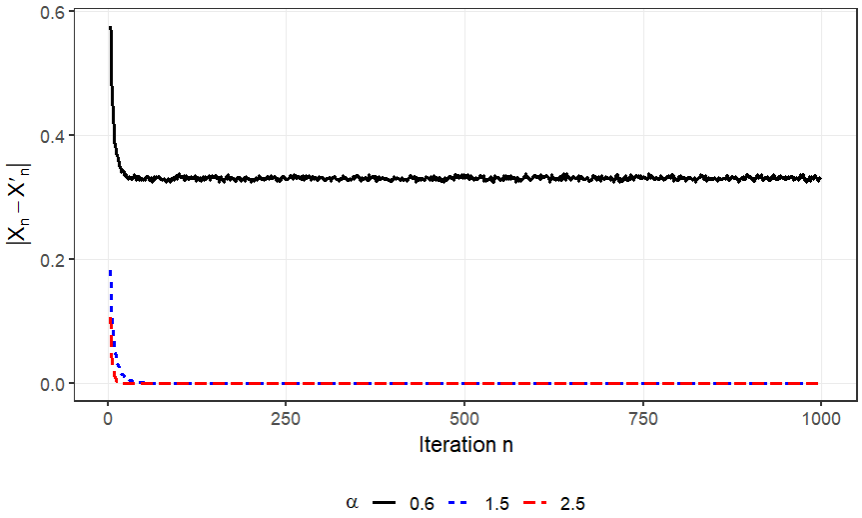


Figure 1: The simulated expected value of $|X_n - X'_n|$ where X_n, X'_n are random logistic maps and CRN is applied ($X_0 = 0.5, X'_0 = 0.125$). For $\alpha > 1/2$, the chain converges in distribution. For $\alpha \leqslant 1$ the chain is not locally contractive. For $\alpha > 2$, the chain is locally contractive. The CRN estimate uses the mean of 10000 simulations.

On the other hand, when $\alpha = 0.6$, the simulated expected distance does not appear to converge to 0 (at iteration 100 000, $E[|X_{100\,000} - X'_{100\,000}|] \approx 0.343$, which is nearly unchanged from iteration 100, $E[|X_{100} - X'_{100}|] \approx 0.331$). This indicates that when the random logistics map is not locally contractive, the two chains simulated using CRN may not converge.

5.2 Zeno's walk

Example 5.3 (Zeno's walk)**.** Let $\{X_n\}_{n\geqslant 0} \in \mathbb{R}$ be a Zeno's walk ([39]). That is for i.i.d. $B_n \sim \text{Bernoulli}(r)$, $X_n = B_n f^+(X_{n-1}) + (1 - B_n) f^-(X_{n-1})$ where,

$$f^+(x) = \begin{cases} x + 1 & 0 \leqslant x \\ \frac{x}{2} + 1 & -2 \leqslant x \leqslant 0 \\ x + 2 & x \leqslant -2 \end{cases} \qquad f^-(x) = \begin{cases} x - 1 & x \leqslant 0 \\ \frac{x}{2} - 1 & 0 \leqslant x \leqslant 2 \\ x - 2 & 2 \leqslant x \end{cases}$$

See Figure 2 for a graphical representation of $f^+(x)$ and $f^-(x)$.

Zeno's walk is locally contractive and geometrically ergodic if $\frac{1}{3} < r < \frac{2}{3}$ [39]. In Lemma 5.4, we show that CRN is the optimal CWD coupling on the nth iteration of Zeno's walk.

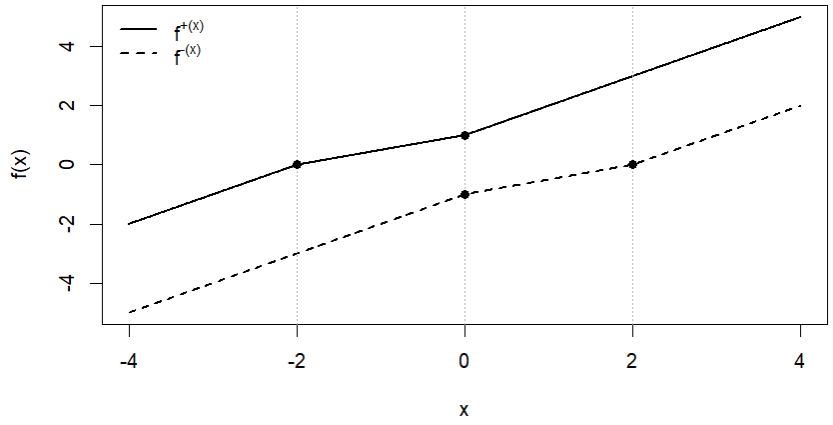


Figure 2: Graph of $f^+(x)$ and $f^-(x)$. It is clear from the graph that $f^+(x) - f^-(x) \geqslant 0$ for any $x \in \mathbb{R}$.

Lemma 5.4. *CRN is an optimal CWD coupling on the nth iteration of Zeno's walk when k is a QAC metric.*

Proof. First we show that CRN is an optimal Wasserstein distance coupling on the first iteration: We can re-write $X_n = f^-(X_{n-1}) + B_n(f^+(X_{n-1}) - f^-(X_{n-1}))$. Note that $f^+(x) - f^-(x) > 0$ for all $x \in \mathbb{R}$ (this is clear from Figure 2).This means that for any $X_{n-1} \in \mathbb{R}$, the function $X_n = f^-(X_{n-1}) + B_n(f^+(X_{n-1}) - f^-(X_{n-1}))$ is an increasing function of B_n and so by Theorem 4.2, for k QAC, CRN is an optimal Wasserstein distance coupling on the first iteration of Zeno's walk.

Next, since $f^+(x), f^-(x)$ are increasing functions of x and $B_n, 1 - B_n \geqslant 0$, the function $B_n f^+(x) + (1 - B_n)f^-(x)$ is an increasing function of x and so, by Theorem 4.8 CRN is the optimal CWD coupling for the nth iteration of the MC. □

Figure 3 displays the expected distance using CRN between two copies of the Zeno walk, X_n and X'_n, initialized at $X_0 = 0$ and $X'_0 = 100$, respectively.

Moreover, Section 7.2 of [39] asserts that the process fails to converge whenever $r \leqslant 1/3$. This theoretical result is reflected in Figure 3: the black curves (corresponding to $r = 0.1$ and $r = 1/3$) are non-convergent, whereas the red curve (corresponding to $r = 0.5$) converges.

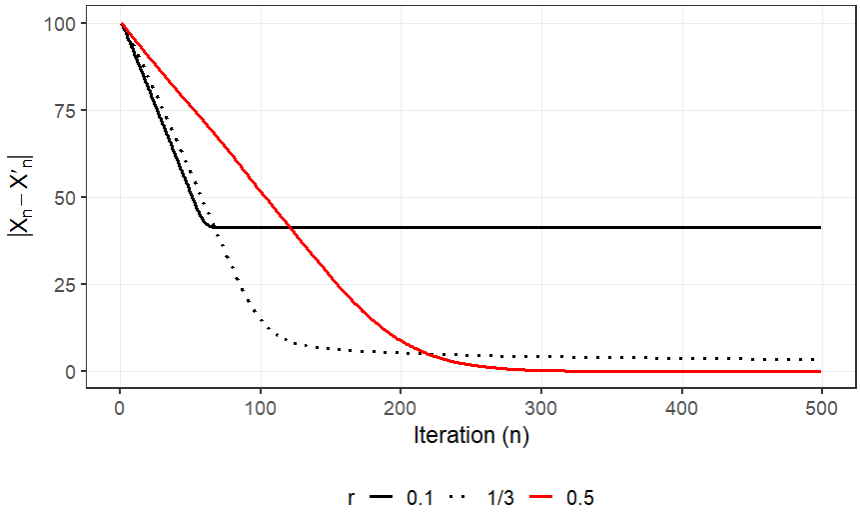


Figure 3: The simulated expected value of $|X_n - X'_n|$ where X_n, X'_n follow Zeno's walk and are simulated using CRN. The CRN estimate uses the mean of 1000 simulations. $X_0 = 0, X'_0 = 100$.

5.3 Autoregressive normal process in $\mathbb{R}^q$

Example 5.5 (Autoregressive normal process in $\mathbb{R}^q$)**.** The random vector $\{\vec{X}_n\}_{n\geqslant 1} \in \mathbb{R}^q$ is an autoregressive normal process if for matrix $A \in \mathbb{R}^{q\times q}$ and normally distributed random vector $\vec{\theta}_n \sim N(\vec{0}_q, \Sigma)$, Σ is the covariance matrix,

$$\vec{X}_n = A\vec{X}_{n-1} + \vec{\theta}_n$$

If A has only positive entries, then by Lemma 5.6 below, we show that CRN is an optimal CWD coupling for the nth iteration of the MC.

Lemma 5.6. *For the autoregressive normal process in $\mathbb{R}^q$, CRN is an optimal CWD coupling at the first iteration. If A has non-negative entries, that is, $a_{lj} \geqslant 0$ for all $l, j \in 1, \ldots, q$, then CRN is an optimal CWD coupling at every iteration $n \geqslant 1$. In particular,*

$$\widetilde{W}_{||\cdot||_p,p}\big(\mathcal{L}(\vec{X}_n), \mathcal{L}(\vec{Y}_n)\big) \;= \big(\mathbb{E}\left[||f_n(\Theta_n, \vec{x}_0) - f_n(\Theta_n, \vec{y}_0)||_p^p\right]\big)^{1/p}. \tag{13}$$

Proof. CRN is the optimal CWD coupling at the first iteration: For any fixed $i \in \{1, \ldots, q\}$

and initial value $\vec{x}_0$, the coordinate $X_{1,i} = (A\vec{x}_0)_i + \theta_{1,i}$ is an increasing function of $\theta_{1,i}$. So, by Theorem 4.7, CRN is an optimal CWD coupling for the first iteration of the autoregressive normal process in $\mathbb{R}^q$.

CRN is the optimal CWD coupling at the nth iteration: Now suppose that A has nonnegative entries. For any $\vec{x}, \vec{y} \in \mathbb{R}^q$ such that $\vec{x} \leqslant \vec{y}$ coordinatewise, we have $A\vec{x} + \vec{\theta} \leqslant A\vec{y} + \vec{\theta}$ coordinatewise, since $a_{lj} \geqslant 0$ for all $l, j \in 1, \ldots, q$. Hence, $f(\vec{\theta}, \vec{x}) = A\vec{x} + \vec{\theta}$ is coordinatewise nondecreasing in $\vec{x}$. Moreover, f is coordinatewise nondecreasing in $\vec{\theta}$, since each component of $\vec{\theta}$ enters the corresponding component of f additively.

Thus, the assumptions of Theorem 4.8 are satisfied, and CRN is an optimal CWD coupling at every iteration $n \geqslant 1$. Equation 13 follows.

□

5.4 Bayesian location model Gibbs sampler

Example 5.7 (Bayesian location model Gibbs sampler)**.** Suppose that we are given data points $Y_1, \ldots, Y_J \sim N(\mu, \tau^{-1})$ where $J \geqslant 3$ and μ, τ^{-1} are unknown with flat priors on $\mathbb{R}$ and $\mathbb{R}_+$, respectively. The Gibbs algorithm is based on the following conditional posterior distributions of μ and τ^{-1}.

$$\mu_{n+1} = \bar{y} + Z_{n+1}/\sqrt{J\tau_n} \qquad\qquad \tau_{n+1}^{-1} = \frac{S + J(\bar{y} - \mu_{n+1})^2}{2G_{n+1}}$$

Where $Z_n \sim N(0,1)$ and $G_n \sim \Gamma(\frac{J}{2} + 1, 1)$ are independent and $S = \sum_{i=1}^{J}(y_i - \bar{y})^2$. $\Gamma(\alpha, \beta)$ represents the gamma distribution, $\Gamma^{-1}(\alpha, \beta)$ represents the inverse gamma distribution where α and β represent the shape and scale, respectively.

The variable τ^{-1} can be written as an IRF: $\tau_n^{-1} = H_n G_n \tau_{n-1}^{-1} + G_n$ where $H_n \sim \Gamma\left(\frac{1}{2}, \frac{S}{2}\right)$ and $G_n \sim \Gamma^{-1}\left(\frac{J+2}{2}, \frac{S}{2}\right)$ are independent.

The following lemma states that CRN is also an optimal CWD coupling for the nth iteration of the Bayesian location model Gibbs sampler.

Lemma 5.8. *CRN is an optimal CWD coupling on the nth iteration of the Bayesian location model Gibbs sampler when k is a QAC metric. That is for i.i.d.* $\vec{\theta}_n = (H_n, G_n)$,

$$\widetilde{W}_k(\mathcal{L}(\tau_n^{-1}), \mathcal{L}(\tau_n^{-1'})) = E[k(f_n(\Theta_n, \tau_0^{-1}), f_n(\Theta_n, \tau_0^{-1'}))] \tag{14}$$

Proof. Note first that $H_n, G_n, \tau_{n-1}^{-1} \geqslant 0$. For any $H_n = h, \tau_{n-1}^{-1} = \tau^{-1}$, the function $\tau_n^{-1} = hG_n\tau^{-1} + G_n$ is increasing with respect to G_n. For any $G_n = g, \tau_{n-1}^{-1} = \tau^{-1}$, the function $\tau_n^{-1} = H_n g\tau^{-1} + g$ is increasing with respect to H_n. For any $H_n = h, G_n = g$, the function $\tau_n^{-1} = hg\tau_{n-1}^{-1} + g$ is increasing with respect to τ_{n-1}^{-1}. Thus, f is coordinatewise monotone in terms of (H_n, G_n) and τ_n^{-1}.

So by Theorem 4.8, CRN is an optimal CWD coupling on the nth iteration of the Bayesian location model Gibbs sampler when k a QAC metric and Equation 14 follows.

□

5.5 Linear ARCH process

Example 5.9 (Linear ARCH process)**.** Let $\{X_n\}_{n\geqslant 1} \in \mathbb{R}$ be a linear ARCH process. Then for i.i.d. θ_n

$$X_n = (\beta_0 + \beta_1 X_{n-1})\theta_n$$

We show that for the linear ARCH process, if $X_n \geqslant 0$ for all $n \in \mathbb{N}$, CRN is the joint coupling that generates the CWD for iteration $n \geqslant 1$.

Lemma 5.10. *Suppose that* $\beta_0, \beta_1, \theta_n > 0$ *a.s. and* $x_0, y_0 > 0$*, then CRN is an optimal CWD coupling on the nth iteration of the linear ARCH process when k is a QAC metric.*

$$\widetilde{W}_k(\mathcal{L}(X_n), \mathcal{L}(Y_n)) = E[k(f_n(\Theta_n, x_0), f_n(\Theta_n, y_0))] \tag{15}$$

Proof. Since $(\beta_0 + \beta_1 X_{n-1}) > 0$ w.p. 1, $X_n = f(\theta_n, X_{n-1})$ is always an increasing function of θ_n. Since $\beta_1\theta_n > 0$ w.p. 1, $f(\theta_n, X_{n-1})$ is always an increasing function of X_{n-1}. So by

Theorem 4.8, CRN is an optimal CWD coupling on the nth iteration of the linear ARCH process when k is a QAC metric. □

6 When monotone conditions are not satisfied

Monotonicity is important when establishing the optimal CWD coupling between two random variables, because it orders the function with regard to what is relevant: the distance between $f(\theta, x)$ and $f(\theta, y)$ as a function of $U \sim Unif(0,1)$ where $\theta = F_\theta^{-1}(U)$. However, sometimes f is not monotone. In such cases we ask ourselves, is CRN the optimal CWD coupling? Probably not. So, how could we generate a coupling that has lower expected distance? If we can propose an equimeasurable rearrangement of f w.r.t U, f^*, that is monotone, then CRN applied to the proposed f^* would generate the optimal CWD coupling.

In the following section, we propose two equimeasurable transformations $g : [0,1] \to [0,1]$ and $h : [0,1] \times \{0,1\} \to [0,1]$ that, when used correctly, generate monotone functions and consequently lower expected distances when simulated using CRN.

6.1 Equimeasurable functions

'Equimeasurable' when used within the context of CRN was discussed as early as 1934 [20]. The equimeasurable functions, however, were limited to the quantile transformation of $U \sim \text{Unif}(0,1)$, which is described in Section 2.2. Equimeasurable rearrangements were also defined in Chapter 1 of [42], and [11], but was again only defined in terms of quantile functions.

While the quantile function is an optimal transformation, in the problem setup with IRFs we do not know the CDF of X, but only know that $X = f(\theta, x)$ and that $\theta = F_\theta^{-1}(U)$ for known F_θ. Thus we're interested in generating an equimeasurable monotone rearrangment of $f(F_\theta^{-1}(U), x)$ with respect to U by creating a mapping $g : [0,1] \to [0,1]$ such that $f(F_\theta^{-1}(g(U)), x)$ is a monotone function of U and $\mathcal{L}(f(F_\theta^{-1}(g(U)), x)) = \mathcal{L}(f(F_\theta^{-1}(U), x))$.

In this section we study equimeasurable functions as transformations on the random variable U that preserve the measure of U (see Definition 6.1). This is similar to the approach taken in [36, 35], which studied generating lower bounds on eigenvalues using rearrangements of the eigenfunctions.

Definition 6.1 (Equimeasurable function of $f(U)$)**.** Let $f : [0,1] \to \mathbb{R}$ be a measurable function and $U \sim \text{Unif}(0,1)$. f^* is an equimeasurable function of f if $\mathcal{L}(f(U)) = \mathcal{L}(f^*(U))$.

An equimeasurable function f^* is specific to the function f and the random variable U. In this paper we restrict our definition to $U \sim \text{Unif}(0,1)$.

The following are two function transformations that can be applied to U. We can then use these transformations to generate monotone equimeasurable transformations of our random function.

If the values of a function are flipped over an interval $(a,b) \in [0,1]$, Lemma 6.2 states that this is an equimeasurable function.

Lemma 6.3 states that if I partition an interval into two parts $(a,b) = (a,t) \cup (t,b)$ then I can replace the interval (a,b) with (a,t) with probability $\frac{t-a}{b-a}$ and replace it with (t,b) with probability $\frac{b-t}{b-a}$, hence 'merging' two intervals.

Lemma 6.2 (Interval flip)**.** *Let $(a,b) \subset [0,1]$, $U \sim \text{Unif}(0,1)$, and the function g to be*

$$g(U) = \begin{cases} U & \text{if } U \notin (a,b) \\ a+b-U & \text{if } U \in (a,b) \end{cases} \tag{16}$$

Then $\mathcal{L}(f(U)) = \mathcal{L}(f(g(U)))$. The proof can be found in Section 7.7.

Lemma 6.3 (Interval merge)**.** *Let $(a,b) \subset [0,1]$, $t \in (a,b)$, $U \sim \text{Unif}(0,1)$ and the function $h(U,\xi)$ be as follows for independent random variables $U \sim \text{Unif}(0,1)$ and $\xi \sim \text{Bernoulli}(\frac{t-a}{b-a})$*

$$h(U,\xi) = \begin{cases} \xi\left(\frac{t-a}{b-a}(U-a)+a\right) + (1-\xi)\left(\frac{b-t}{b-a}(U-a)+t\right) & \text{if } U \in (a,b) \\ U & \text{if } U \notin (a,b) \end{cases} \tag{17}$$

Then $\mathcal{L}(f(U)) = \mathcal{L}(f(h(U,\xi)))$. The proof can be found in Section 7.8.

6.2 Example

Numerical Example 6.4 (Random logistic maps)**.** We revisit Example 5.1, to prove by counterexample that CRN is not always the optimal CWD coupling for a two-step random logistic map, even though it is always the optimal CWD one step coupling. Suppose that the initial values of the random logistic map are $X_0 = \frac{1}{2}$ and $Y_0 = \frac{1}{8}$, then the second iteration of the functions are

$$X_2 = 4\beta_2\beta_1(1-\beta_1) \qquad\qquad Y_2 = \frac{7}{4}\beta_2'\beta_1'\left(1 - \frac{7}{16}\beta_1'\right) \tag{18}$$

Recall $\beta_1, \beta_2 \sim \text{Beta}(\alpha+1/2, \alpha-1/2)$. Denote F_β to be the CDF of $\text{Beta}(\alpha+1/2, \alpha-1/2)$ so that $F_\beta^{-1}(U) \sim \text{Beta}(\alpha+1/2, \alpha-1/2)$. We can then re-write X_2 as follows where $U_1 \sim \text{Unif}(0,1)$ and U_1, β_2 are independent,

$$X_2(U_1, \beta_2) = 4\beta_2 F_\beta^{-1}(U_1)(1 - F_\beta^{-1}(U_1))$$

Observe that for any β_2, X_2 is an increasing function of U_1 when $U_1 \in [0, F_\beta(0.5)]$ and a decreasing function of U_1 when $U_1 \in [F_\beta(0.5), 1]$. We will construct an equimeasurable transform of $X_2(U_1, \beta_2)$, $X_2''(U_1, \beta_2, \xi)$, that is monotone in U_1 for $\xi \in \{0,1\}$. To do so, we first generate an equimeasurable function, $X_2'(U_1, \beta_2) = X(g(U_1), \beta_2)$, that is increasing over $U_1 \in (F_\beta(0.5), 1]$ where $g : [0,1] \to [0,1]$ is defined as follows,

$$g(U) = \begin{cases} U & \text{if } U \in [0, F_\beta(0.5)] \\ 1 + F_\beta(0.5) - U & \text{if } U \in (F_\beta(0.5), 1] \end{cases} \tag{19}$$

By Lemma 6.2, $\mathcal{L}(X_2') = \mathcal{L}(X_2)$.

The interval flip is easiest to visualize in Figure 4. In graph (A), when $U_1 < F_\beta(0.5) \approx$

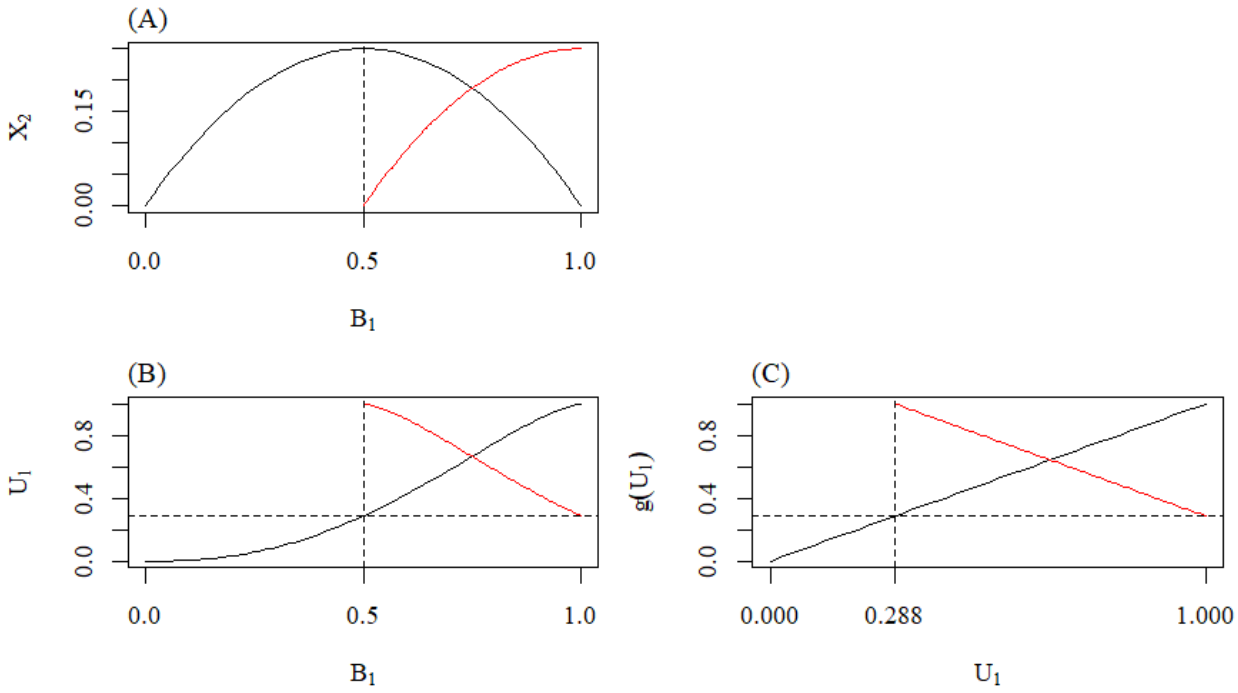


Figure 4: Set $\alpha = 2$ so that $\beta_1 \sim \text{Beta}(2.5, 1.5)$ and $F_\beta(0.5) = 0.288$. Set $\beta_2 = 1$ (its value is of little importance since it does not affect whether β_1 is increasing or not). In graph (A), the black line denotes $X_2(U_1, 1)$ and the red line denotes $X_2'(U_1, 1)$ when it is not equal to $X_2(U_1, 1)$. Graph (C) represents the transformation $g(U_1)$ in red compared to the identity function. Graph (B) then imposes the flip that was applied by g onto the values of β_1. Graph (A) shows the effect of the flip, g, through β_1 onto the value of X_2.

0.288 ($\beta_1 < 0.5$), $X_2(U_1, \beta_2) = X_2'(U_1, \beta_2)$, however when $U_1 > F_\beta(0.5)$ ($\beta_1 > 0.5$), $X_2(U_1, \beta_2)$ is represented by the black line while the transform $X_2'(U_1, \beta_2)$ is represented by the red line. Note that the transform $X_2'(U_1, \beta_2)$ is increasing when $U_1 > F_\beta(0.5)$.

The transform $X_2'(U_1, \beta_2)$ is increasing over $U_1 \in [0, F_\beta(0.5))$ and $U_1 \in (F_\beta(0.5), 1]$. However, it is not monotonically increasing. To generate an equimeasurable map that is monotonically increasing with respect to U, we use Lemma 6.3 and introduce the auxiliary random variable ξ.

Let $(a, b) = [0, 1]$ and $t = F_\beta(0.5)$. Then $X_2''(U_1, \beta_2, \xi) = X_2(g(h(U_1, \xi)), \beta_2)$ where $h(U_1, \xi)$ is generated according to Equation 17 is an equimeasurable function of U (Lemma

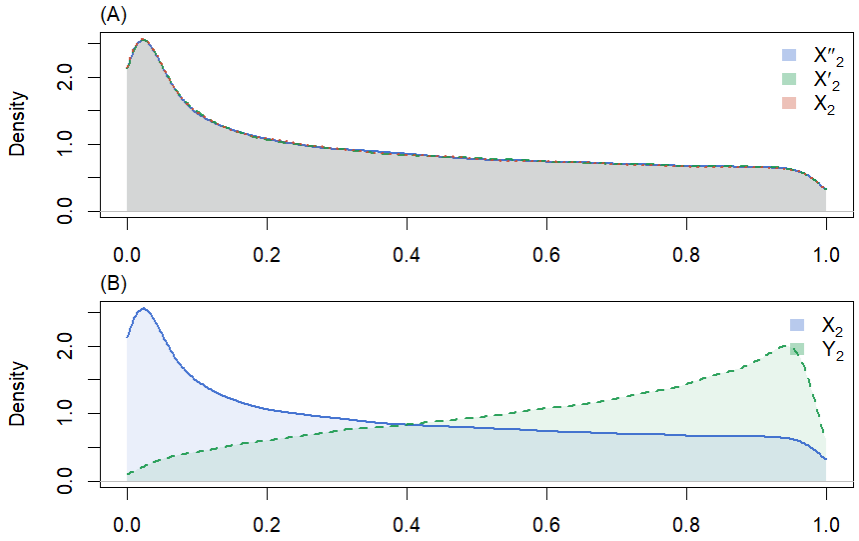


Figure 5: (A): Simulated density of $X_2 = 4\beta_2\beta_1(1-\beta_1)$, $X_2'(U_1, \beta_2)$, and $X_2''(U_1, \beta_2, \xi)$ (see Equation 20). The random variables $\beta_i \sim \text{Beta}(1.5, 0.5), i \in \{1, 2\}$, $U_1 \sim \text{Unif}(0, 1)$, $\xi \sim \text{Bernoulli}(F_\beta(0.5))$ are independent. This shows that the random variables X_2', which is based on the transformation g and X_2'', which is based on the transformations g and h are indeed equimeasurable to X_2. (B): Comparison of the density of X_2 and Y_2

6.3). To summarize, X_2'' can be defined as follows for $\xi \sim \text{Bernoulli}(F_\beta(0.5))$

$$X_2''(U_1, \beta_2, \xi) = \begin{cases} 4\beta_2 F_\beta^{-1}(1 - U_1(1 - F_\beta(0.5)))[1 - F_\beta^{-1}(1 - U_1(1 - F_\beta(0.5)))] & \text{if } \xi = 0 \\ 4\beta_2 F_\beta^{-1}(F_\beta(0.5)U_1)[1 - F_\beta^{-1}(F_\beta(0.5)U_1)] & \text{if } \xi = 1 \end{cases} \tag{20}$$

Lemmas 6.3 and 6.2 imply that $X_2''(U_1, \beta_2, \xi)$ is an equimeasurable transform of $X_2(U_1, \beta_2)$. Figure 5 (A) graphs the simulated density of $X_2(U_1, \beta_2)$, $X_2'(U_1, \beta_2)$ and $X_2''(U_1, \beta_2, \xi)$, which visually shows that the densities are equal.

Figure 6 shows how the solid line, which is the expected value of $|X_2(U_1, \beta_2) - Y_2(U_1, \beta_2)|$ simulated using CRN has a larger expected distance compared to the transformed expected value of $|X_2'(U_1, \beta_2, \xi) - Y_2(F_\beta^{-1}(U_1), \beta_2)|$ simulated using CRN. This shows that $E[|X_2(U_1, \beta_2) - Y_2(U_1, \beta_2)|]$ is strictly greater than the CWD.

$$\widetilde{W}_{|\cdot|}(\mathcal{L}(X), \mathcal{L}(Y)) \leqslant E[|X_2'(U_1, \beta_2) - Y_2(U_1, \beta_2)|] < E[|X_2(U_1, \beta_2) - Y_2(U_1, \beta_2)|]$$

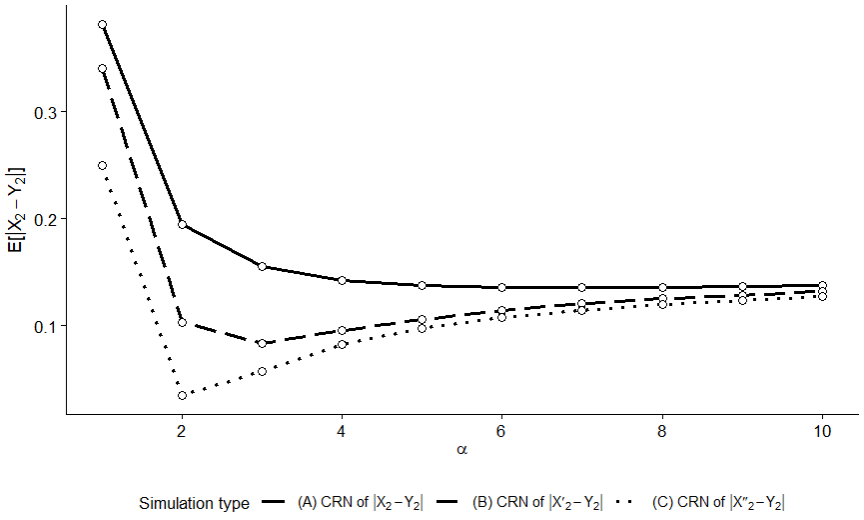


Figure 6: Line (A) represents the simulated expected distance of $|X_2(U_1, B_2) - Y_2(U_1, B_2)|$ ie. the expected distance when CRN is applied to the original function. Line (B) represents the simulated expected distance of $|X_2'(U_1, B_2) - Y_2(U_1, B_2)|$ ie. the expected distance when CRN is applied to the transformed $X_2'(U_1, B_2, \xi)$. Line (C) represents the simulated expected distance of $|X_2''(U_1, B_2, \xi) - Y_2(U_1, B_2)|$ ie. the expected distance when CRN is applied to the transformed $X_2''(U_1, B_2, \xi)$.

Figure 6 shows how the long-dashed line, which is the expected value of $|X_2'(U_1, \beta_2) - Y_2(U_1, \beta_2)|$ simulated using CRN has a larger expected distance compared to the dotted line, the transformed expected value of $|X_2''(U_1, \beta_2, \xi) - Y_2(U_1, \beta_2, \xi)|$ simulated using CRN. This shows that $E[|X_2'(U_1, \beta_2) - Y_2(U_1, \beta_2)|]$ is strictly greater than the Wasserstein distance.

$$W_{|\cdot|}(\mathcal{L}(X), \mathcal{L}(Y)) \leqslant E[|X_2''(U_1, \beta_2, \xi) - Y_2(U_1, \beta_2, \xi)|] < E[|X_2'(U_1, \beta_2) - Y_2(U_1, \beta_2)|]$$

Note that we cannot use $E[|X_2''(U_1, \beta_2, \xi) - Y_2(U_1, \beta_2, \xi)|]$ to bound the CWD from above, because the joint distribution that is generated by adding ξ may not be contained in the set of coordinate-independent joint distributions generated $\mathcal{J}_{IND}(\mu, \mu)$ where $\mu = \mathcal{L}(U_1, \beta_2)$.

7 Appendix

7.1 Proof of Proposition 2.3

Proof. Proof of Proposition 2.3

$\Longleftarrow$: $|F_n^{-1}(U) - F^{-1}(U)| \to 0$ w.p. 1 means that $F_n^{-1}(U) \to F^{-1}(U)$ almost surely, which means that $\mathcal{L}(F_n^{-1}(U)) \to \mathcal{L}(F^{-1}(U))$ in distribution. Since $F_n^{-1}(U) \sim \mathcal{L}(X_n)$ and $F^{-1}(U) \sim \mathcal{L}(X_\infty)$, this implies $\mathcal{L}(X_n) \Rightarrow \mathcal{L}(X_\infty)$.

$\Longrightarrow$ This proof greatly resembles the proof of Theorem 10.1.1 (4) of [31].

Denote $D_{F^{-1}} \in [0,1]$ as the set of discontinuous points on F^{-1} and $C_{F^{-1}} \in [0,1]$ as the set of continuous points on F^{-1}.

Let $0 < u < 1$, $\epsilon > 0$, and choose ω such that $F^{-1}(u) - \epsilon < \omega < F^{-1}(u)$ and $\pi(\omega) = 0$. Then $F(\omega) < u$ and since $F_n(\omega) \to F(\omega)$, for sufficiently large n, $F_n(\omega) < u$ and so, $F^{-1}(u) - \epsilon < \omega < F_n^{-1}(u)$. Thus,

$$\liminf F_n^{-1}(u) \geqslant F^{-1}(u) \tag{21}$$

Let $0 < u < u' < 1$ and choose ω such that $F^{-1}(u') < \omega < F^{-1}(u') + \epsilon$ and $\pi(\omega) = 0$. Then $u < u' \leqslant F(F^{-1}(u')) \leqslant F(\omega)$ and since $F_n(\omega) \to F(\omega)$, for sufficiently large n, $u \leqslant F_n(\omega)$ and $F_n^{-1}(u) \leqslant \omega < F^{-1}(u') + \epsilon$. Thus,

$$\limsup F_n^{-1}(u) \leqslant F^{-1}(u') \tag{22}$$

By equation 21 and equation 22 if F^{-1} is continuous at u (ie. $u \in C_{F^{-1}}$), then $\lim_{n\to\infty} F_n^{-1}(u) = F^{-1}(u)$. Or in other words, $P(\lim_{n\to\infty} |F_n^{-1}(u) - F^{-1}(u)| = 0 \mid U \in C_{F^{-1}}) = 1$.

Next suppose that $U \sim Unif(0,1)$. Since F^{-1} is non-decreasing on $(0,1)$, it has countably many discontinuous points and so $P(U \in D_{F^{-1}}) = 0$ (and $P(U \in C_{F^{-1}}) = 1$).

Thus, the probability of $\lim_{n\to\infty} |F_n^{-1}(u) - F^{-1}(u)| = 0$ is as follows,

$$\begin{aligned}
&P(\lim_{n\to\infty} |F_n^{-1}(U) - F^{-1}(U)| = 0) \\
&\geqslant P(\lim_{n\to\infty} |F_n^{-1}(U) - F^{-1}(U)| = 0 \mid U \in C_{F^{-1}}) P(U \in C_{F^{-1}}) \\
&= 1
\end{aligned}$$

And so, when CRN is applied to the quantile function, $|F_n^{-1}(U) - F^{-1}(U)| \to 0$ w.p. 1. □

7.2 Proof of Theorem 3.2

For simplicity, we remove the vector notation from the random variables (ex. $\vec{X}$ becomes X). Define $X = f(\theta, x_0) = g(\theta)$ and $Y = f(\theta', y_0) = h(\theta')$. Recall that $\theta, \theta' \sim \mu$. Denote $\nu \in \mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y))$. Denote $\mathcal{X}^d = \prod_{i=1}^d \mathcal{X}_i$.

By Proposition 10.4.12 of [4], since $\theta, \theta' \sim \mu$ and g, h are Borel mappings, then there exists Borel regular conditional probability measures $\mu(\cdot \mid x) = \mu(\cdot \mid g(\theta) = x), x \in X$ on θ and $\mu(\cdot \mid y) = \mu(\cdot \mid h(\theta') = y), y \in Y$ on θ'.

Definition 7.1. Define $\gamma : \mathcal{B}(\mathcal{X}^d \times \mathcal{X}^d) \to \mathbb{R}$ such that for any $C \in \mathcal{B}(\mathcal{X}^d \times \mathcal{X}^d)$,

$$\gamma(C) = \int_{\mathbb{R}^q \times \mathbb{R}^q} \int_C \mu(\mathrm{d}\theta \mid x)\mu(\mathrm{d}\theta' \mid y)\nu(\mathrm{d}x, \mathrm{d}y)$$

Note that if $C = A \times B$ where $A, B \in \mathcal{B}(\mathcal{X}^d)$

$$\gamma(C) = \int_{\mathbb{R}^q \times \mathbb{R}^q} \mu(A \mid x)\mu(B \mid y)\nu(\mathrm{d}x, \mathrm{d}y)$$

Lemma 7.2. *γ is a probability measure on $\mathcal{B}(\mathcal{X}^d \times \mathcal{X}^d)$.*

Proof. 1. $\gamma(\mathcal{X}^d \times \mathcal{X}^d) = 1$

$$\begin{aligned}
\gamma(\mathcal{X}^d \times \mathcal{X}^d) &= \int_{\mathbb{R}^q \times \mathbb{R}^q} \mu(\mathcal{X}^d \mid x)\mu(\mathcal{X}^d \mid y)\nu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} \nu(\mathrm{d}x, \mathrm{d}y) && \text{Since } \mu(\mathcal{X}^d \mid x) = \mu(\mathcal{X}^d \mid y) = 1 \\
&= 1
\end{aligned}$$

2. $\gamma(\varnothing) = 0$

$$\gamma(\varnothing) = \int_{\mathbb{R}^q \times \mathbb{R}^q} \mu(\varnothing \mid x)\mu(\varnothing \mid y)\nu(\mathrm{d}x, \mathrm{d}y) = 0 \int_{\mathbb{R}^q \times \mathbb{R}^q} \nu(\mathrm{d}x, \mathrm{d}y) = 0 \times 1 = 0$$

3. Suppose that $A_1, A_2, \ldots \in \mathcal{B}(\mathcal{X}^d \times \mathcal{X}^d)$ is a sequence of disjoint sets, then $\gamma(\dot{\cup}_i^\infty A_i) = \sum_{i=1}^\infty \gamma(A_i)$

$$\begin{aligned}
\gamma(\dot{\cup}_i^\infty A_i) &= \int_{\mathbb{R}^q \times \mathbb{R}^q} \int_{\dot{\cup}_{i=1}^\infty A_i} \mu(\mathrm{d}\theta \mid x)\mu(\mathrm{d}\theta' \mid y)\nu(\mathrm{d}x, \mathrm{d}y) \\
&= \sum_{i=1}^\infty \int_{\mathbb{R}^q \times \mathbb{R}^q} \int_{A_i} \mu(\mathrm{d}\theta \mid x)\mu(\mathrm{d}\theta' \mid y)\nu(\mathrm{d}x, \mathrm{d}y) \quad \text{by MCT (Exercise 1.5.6 of [15])} \\
&= \sum_{i=1}^\infty \gamma(A_i)
\end{aligned}$$

Thus by definition of a probability measure (Section 2.1 of [31]), γ is a probability measure. □

Lemma 7.3. *The marginals of γ are μ. That is, $\gamma(A \times \mathcal{X}^d) = \mu(A)$ for $A \in \mathcal{B}(\mathcal{X}^d)$.*

Proof.

$$\begin{aligned}
\gamma(A \times \mathcal{X}^d) &= \int_{\mathbb{R}^q \times \mathbb{R}^q} \int_{\mathcal{X}^d} \int_A \mu(d\theta \mid x)\mu(d\theta' \mid y)\nu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} \mu(A \mid x)\mu(\mathcal{X}^d \mid y)\nu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} \mu(A \mid x)\nu(\mathrm{d}x, \mathrm{d}y) \\
&= \mu(A)
\end{aligned}$$

By similar reasoning, $\gamma(\mathcal{X}^d \times A) = \mu(A)$. □

Lemma 7.4. *ν is the pushforward of γ. That is, $(g \times h)_* \gamma = \nu$.*

Proof. By definition, for any $B \in \mathcal{B}(\mathbb{R}^q \times \mathbb{R}^q)$ $(g \times h)_* \gamma(B) = \gamma((g \times h)^{-1}(B))$ (Definition 2.40 of [6]). So we must show that for any $B \in \mathcal{B}(\mathbb{R}^q \times \mathbb{R}^q)$, $\gamma((g \times h)^{-1}(B)) = \nu(B)$. Let $B \in \mathcal{B}(\mathbb{R}^q \times \mathbb{R}^q)$,

$$\begin{aligned}
(g \times h)_* \gamma(B) &= \gamma((g \times h)^{-1}(B)) && (23) \\
&= \int_{\mathcal{X}^d \times \mathcal{X}^d} \int_{\mathbb{R}^q \times \mathbb{R}^q} I\big[(\theta, \theta') \in (g \times h)^{-1}(B)\big] \mu(\mathrm{d}\theta \mid x)\mu(\mathrm{d}\theta' \mid y)\nu(\mathrm{d}x, \mathrm{d}y) && (24)
\end{aligned}$$

For fixed x, y note that for all $\theta \sim \mu(\cdot \mid x)$ and $\theta' \sim \mu(\cdot \mid y)$, either $I\big[(\theta, \theta') \in (g \times h)^{-1}(B)\big] = 1$ or $I\big[(\theta, \theta') \in (g \times h)^{-1}(B)\big] = 0$ depending on whether $(g(\theta), h(\theta')) = (x, y) \in B$, so that $I\big[(\theta, \theta') \in (g \times h)^{-1}(B)\big] = I\big[(x, y) \in B\big]$.

And so Equation 23 equals

$$
\begin{aligned}
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} \int_{\mathcal{X}^d \times \mathcal{X}^d} I[(x,y) \in B] \mu(\mathrm{d}\theta \mid x) \mu(\mathrm{d}\theta' \mid y) \nu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} I[(x,y) \in B] \int_{\mathcal{X}^d \times \mathcal{X}^d} \mu(\mathrm{d}\theta \mid x) \mu(\mathrm{d}\theta' \mid y) \nu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} I[(x,y) \in B] \mu(\mathcal{X}^d \mid x) \mu(\mathcal{X}^d \mid y) \nu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} I[(x,y) \in B] \nu(\mathrm{d}x, \mathrm{d}y) \\
&= \nu(B)
\end{aligned}
$$

Thus we have that $(g \times h)_* \gamma(B) = \nu(B)$ for arbitrary B and so, $(g \times h)_* \gamma = \nu$. □

Lemma 7.5. $E_{(\theta,\theta') \sim \gamma}[k(g(\theta), h(\theta'))] = E_{(X,Y) \sim \nu}[k(X,Y)]$

Proof.

$$
\begin{aligned}
&E_{(\theta,\theta') \sim \gamma}[k(g(\theta), h(\theta'))] \\
&= \int_{\mathcal{X}^d \times \mathcal{X}^d} k(g(\theta), h(\theta')) \gamma(d\theta, d\theta') \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} k(x,y) ((g \times h)_* \gamma)(dx, dy) && \text{By Lemma 3.12 of [6]} \\
&= \int_{\mathbb{R}^q \times \mathbb{R}^q} k[(x,y)] \nu(\mathrm{d}x, \mathrm{d}y) \\
&= E_{(X,Y) \sim \nu}[k(X,Y)]
\end{aligned}
$$

□

Proof of Theorem 3.2. Recall that $X = f(\theta, x) = g(\theta)$ and $Y = f(\theta', y) = h(\theta')$.

Note that by discussion below Proposition 2.2 of [19],

$$\mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y)) \supseteq \{\mathcal{L}(g(\theta), h(\theta')) \mid \theta, \theta' \sim \mu\}$$

and so by Equation 4,

$$W_k(\mathcal{L}(X), \mathcal{L}(Y)) \leqslant \inf_{\gamma \in \mathcal{J}(\mu,\mu)} E_{(\theta,\theta')\sim\gamma}[k(g(\theta), h(\theta'))] \tag{25}$$

We want to show that the reverse is also true. Ie.

$$W_k(\mathcal{L}(X), \mathcal{L}(Y)) \geqslant \inf_{\gamma \in \mathcal{J}(\mu,\mu)} E_{(\theta,\theta')\sim\gamma}[k(g(\theta), h(\theta'))]$$

Let $\nu \in \mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y))$ and propose $\gamma(\theta, \theta')$ as defined in Definition 7.1.

$\gamma \in \mathcal{J}(\mu, \mu)$ since by Lemma 7.2, γ is a probability measure on $\mathcal{B}(\mathcal{X}^d \times \mathcal{X}^d)$ and by Lemma 7.3, the marginals of γ are μ. Further, by Lemma 7.4, $(g \times h)_* \gamma = \nu$ and so, for every $\nu \in \mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y))$, there exists a $\gamma \in \mathcal{J}(\mu, \mu)$ such that $\mathcal{L}(g(\theta), h(\theta')) \in \mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y))$. This means that

$$\mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y)) \subseteq \{\mathcal{L}(g(\theta), h(\theta')) \mid \theta, \theta' \sim \mu\}$$

Furthermore for every $\nu \in \mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y))$, the corresponding $\gamma \in \mathcal{J}(\mu, \mu)$ generates the same expected distance between X and Y. That is, by Lemma 7.5,

$$E_{(\theta,\theta')\sim\gamma}[k(g(\theta), h(\theta'))] = E_{(X,Y)\sim\nu}[k(X, Y)]$$

And so we get the following inequality,

$$\inf_{\nu \in \mathcal{J}(\mathcal{L}(X),\mathcal{L}(Y))} E_{(X,Y)\sim\nu}[k(X, Y)] \geqslant \inf_{\gamma \in \mathcal{J}(\mu,\mu)} E_{(\theta,\theta')\sim\gamma}[k(g(\theta), h(\theta'))]$$

Thus the Wasserstein distance is bounded below by the infimum expected distance with

respect to the joint probability between θ, θ'

$$W_k(\mathcal{L}(X), \mathcal{L}(Y)) = \inf_{\nu \in \mathcal{J}(\mathcal{L}(X), \mathcal{L}(Y))} E_{(X,Y)\sim\nu}[k(X, Y)] \tag{26}$$

$$\geqslant \inf_{\gamma \in \mathcal{J}(\mu,\mu)} E_{(\theta,\theta')\sim\gamma}[k(g(\theta), h(\theta'))] \tag{27}$$

Combining equations 25 and 26, we get equality.

$$W_k(\mathcal{L}(X), \mathcal{L}(Y)) = \inf_{\gamma \in \mathcal{J}(\mu,\mu)} E_{(\theta,\theta')\sim\gamma}[k(g(\theta), h(\theta'))] \tag{28}$$

□

7.3 Proof of Theorem 4.2

Proof of Theorem 4.2. Denote $g(\theta_1) = f(\theta_1, x)$ and $h(\theta_2) = f(\theta_2, y)$ for fixed x and y.

Since g and h are monotone in θ by assumption, if $\theta_1 \leqslant \theta_1'$ and $\theta_2 \leqslant \theta_2'$, then $g(\theta_1) \leqslant g(\theta_1')$ and $h(\theta_2) \leqslant h(\theta_2')$. By the quasi-antitone property of k,

$$k\left(g(\theta_1), h(\theta_2)\right) + k\left(g(\theta_1'), h(\theta_2')\right) - k\left(g(\theta_1'), h(\theta_2)\right) - k\left(g(\theta_1), h(\theta_2')\right) \leqslant 0$$

And so, $k(g(\theta_1), h(\theta_2))$ is quasi-antitone with respect to θ_1 and θ_2.

Further, since g and h are right-continuous functions, $k(g(\theta_1), h(\theta_2))$ is right-continuous with respect to θ_1, θ_2 and so $k(g(\theta_1), h(\theta_2))$ is QAC with respect to θ_1, θ_2. We can then apply Proposition 2.1.

$$\begin{aligned}
\inf_{\mathcal{J}(\mu,\mu)} E[k(g(\theta_1), h(\theta_2))] &= E[k(g(F_\theta^{-1}(U)), h(F_\theta^{-1}(U)))] \qquad \text{by Prop 2.1} \\
&= E[k(g(\theta), h(\theta))] \\
&= E[k(f(\theta, x), f(\theta, y))]
\end{aligned}$$

Combining the above equation with Theorem 3.2, $W_k(\mathcal{L}(X), \mathcal{L}(Y)) = E[k(f(\theta, x), f(\theta, y))]$.

□

7.4 Proof of Theorem 4.6

Proof. CRN is the coordinatewise optimal CWD coupling First note that by assumption, for all $i \in \{1, \ldots, d\}$ the function $f(\vec{\theta}, x)$ with fixed $\vec{\theta}_{-i}$ μ_{-i}- a.e. is an increasing continuous function with respect to θ_i. Next since k is QAC, then by Theorem 4.2, for almost every pair $(\vec{\theta}_{-i}, \vec{\theta}'_{-i})$ and fixed x, y, CRN is an optimal coupling with respect to the ith coordinate,

$$\inf_{\gamma_i \in \mathcal{J}(\mu_i, \mu_i)} E_{(\theta_i, \theta'_i) \sim \gamma_i}[k(f((\theta_i, \vec{\theta}_{-i}), x), f((\theta'_i, \vec{\theta}'_{-i}), y))] \tag{29}$$

$$= E_{\theta_i \sim \mu_i}[k(f((\theta_i, \vec{\theta}_{-i}), x), f((\theta_i, \vec{\theta}'_{-i}), y))] \tag{30}$$

We apply a proof by induction to show that CRN is an optimal CWD coupling with respect to the unconditional expected value.

Base step:

The expected value of $k(f(\vec{\theta}, x), f(\vec{\theta}', y))$ conditioned on coordinates $(\theta_{2:d}, \theta'_{2:d})$ is bounded below by the expected value when CRN is applied to coordinate 1 ($\theta_1 = \theta'_1$). This is true by equation 29 for arbitrary joint distribution $\gamma_1 \in \mathcal{J}(\mu_1, \mu_1)$ and almost every pair $(\theta_{2:d}, \theta'_{2:d})$.

$$E_{(\theta_1, \theta'_1) \sim \gamma_1}[k(f((\theta_1, \vec{\theta}_{2:d}), x), f((\theta'_1, \vec{\theta}'_{2:d}), y))] \geqslant E_{\theta_1 \sim \mu_1}[k(f((\theta_1, \vec{\theta}_{2:d}), x), f((\theta_1, \vec{\theta}'_{2:d}), y))]$$

Induction step: Suppose that for $0 \leqslant N \leqslant d-1$, the joint expected distance conditioned on fixed coordinates $(\vec{\theta}_{N+1:d}, \vec{\theta}'_{N+1:d})$ is bounded below by the expected distance when CRN

is applied to coordinates 1 to N ($\vec{\theta}_{1:N} = \vec{\theta}'_{1:N}$). Denote $\gamma_n \in \mathcal{J}(\mu_n, \mu_n)$.

$$E_{(\vec{\theta}_{1:N}, \vec{\theta}'_{1:N}) \sim \bigotimes_{n=1}^{N} \gamma_n}[k(f((\vec{\theta}_{1:N}, \vec{\theta}_{N+1:d}), x), f((\vec{\theta}'_{1:N}, \vec{\theta}'_{N+1:d}), y))] \tag{31}$$

$$\geqslant E_{\vec{\theta}_{1:N} \sim \bigotimes_{n=1}^{N} \mu_n}[k(f((\vec{\theta}_{1:N}, \vec{\theta}_{N+1:d}), x), f((\vec{\theta}_{1:N}, \vec{\theta}'_{N+1:d}), y))] \tag{32}$$

We will then show that for $N+1$, the expected value of k conditioned on fixed coordinates $(\vec{\theta}_{N+2:d}, \vec{\theta}'_{N+2:d})$ is also attained when CRN is applied to coordinates 1 to $N+1$.

$$\begin{aligned}
&E_{(\vec{\theta}_{1:N+1}, \vec{\theta}'_{1:N+1}) \sim \bigotimes_{n=1}^{N+1} \gamma_n}[k(f((\vec{\theta}_{1:N+1}, \vec{\theta}_{N+2:d}), x), f((\vec{\theta}'_{1:N+1}, \vec{\theta}'_{N+2:d}), y))] \\
&= E_{(\vec{\theta}_{1:N}, \vec{\theta}'_{1:N}) \sim \bigotimes_{n=1}^{N} \gamma_n}[E_{(\theta_{N+1}, \theta'_{N+1}) \sim \gamma_{N+1}}[k(f((\vec{\theta}_{1:N}, \theta_{N+1}, \vec{\theta}_{N+2:d}), x), f((\vec{\theta}'_{1:N}, \theta'_{N+1}, \vec{\theta}'_{N+2:d}), y))]] \\
&\geqslant E_{(\vec{\theta}_{1:N}, \vec{\theta}'_{1:N}) \sim \bigotimes_{n=1}^{N} \gamma_n}[E_{\theta_{N+1} \sim \mu_{N+1}}[k(f((\vec{\theta}_{1:N}, \theta_{N+1}, \vec{\theta}_{N+2:d}), x), f((\vec{\theta}'_{1:N}, \theta_{N+1}, \vec{\theta}'_{N+2:d}), y))]] && \text{by eq. 29} \\
&= E_{\theta_{N+1} \sim \mu_{N+1}}[E_{(\vec{\theta}_{1:N}, \vec{\theta}'_{1:N}) \sim \bigotimes_{n=1}^{N} \gamma_n}[k(f((\vec{\theta}_{1:N}, \theta_{N+1}\vec{\theta}_{N+2:d}), x), f((\vec{\theta}'_{1:N}, \theta_{N+1}\vec{\theta}'_{N+2:d}), y))]] \\
&\geqslant E_{\theta_{N+1} \sim \mu_{N+1}}[E_{\vec{\theta}_{1:N} \sim \bigotimes_{n=1}^{N} \mu_n}[k(f((\vec{\theta}_{1:N}, \theta_{N+1}\vec{\theta}_{N+2:d}), x), f((\vec{\theta}_{1:N}, \theta_{N+1}\vec{\theta}'_{N+2:d}), y))]] && \text{by eq. 31} \\
&= E_{\vec{\theta}_{1:N+1} \sim \bigotimes_{n=1}^{N+1} \mu_n}[k(f((\vec{\theta}_{1:N}, \theta_{N+1}\vec{\theta}_{N+2:d}), x), f((\vec{\theta}_{1:N}, \theta_{N+1}\vec{\theta}'_{N+2:d}), y))]
\end{aligned}$$

The above implies that the expected value of k for joint distribution $\gamma \in \mathcal{J}_{IND}(\mu, \mu)$ is bounded below by the expected value of k when CRN is applied.

$$E_{(\vec{\theta}, \vec{\theta}') \sim \gamma}[k(f(\vec{\theta}, x), f(\vec{\theta}', y))] \geqslant E_{\vec{\theta} \sim \mu}[k(f(\vec{\theta}, x), f(\vec{\theta}, y))] \tag{33}$$

Since the joint distribution $\gamma \in \mathcal{J}_{IND}(\mu, \mu)$ was arbitrary,

$$\inf_{\gamma \in \mathcal{J}_{IND}(\mu, \mu)} E_{(\vec{\theta}, \vec{\theta}') \sim \gamma}[k(f(\vec{\theta}, x), f(\vec{\theta}', y))] \geqslant E_{\vec{\theta} \sim \mu}[k(f(\vec{\theta}, x), f(\vec{\theta}, y))]$$

Further, since $\bigotimes_{n=1}^{d} \mathcal{L}(\theta_n, \theta_n) \in \mathcal{J}_{IND}(\mu, \mu)$, the inequality goes the other way,

$$\inf_{\gamma \in \mathcal{J}_{IND}(\mu, \mu)} E_{(\vec{\theta}, \vec{\theta}') \sim \gamma}[k(f(\vec{\theta}, x), f(\vec{\theta}', y))] \leqslant E_{\vec{\theta} \sim \mu}[k(f(\vec{\theta}, x), f(\vec{\theta}, y))]$$

And so, equality ensues, $\inf_{\gamma \in \mathcal{J}_{IND}(\mu,\mu)} E_{(\vec{\theta},\vec{\theta'})\sim\gamma}[k(f(\vec{\theta},x), f(\vec{\theta'},y))] = E_{\vec{\theta}\sim\mu}[k(f(\vec{\theta},x), f(\vec{\theta},y))]$. Combining the above equations, we get that $\widetilde{W}_k(\mathcal{L}(X), \mathcal{L}(Y)) = E[k(f(\vec{\theta},x), f(\vec{\theta},y))]$. □

7.5 Proof of Theorem 4.7

Proof. Denote the ith coordinate of $\vec{X}$ as $\vec{X}^{(i)} = f^{(i)}(\vec{\theta},\vec{x})$.

$$\inf_{\gamma \in \mathcal{J}_{IND}(\mu,\mu)} E_{(\vec{\theta},\vec{\theta'})\sim\gamma}[||f(\vec{\theta},\vec{x}) - f(\vec{\theta'},\vec{y})||_p^p] \tag{34}$$

$$= \inf_{\gamma \in \mathcal{J}_{IND}(\mu,\mu)} \sum_{i=1}^{q} E_{(\vec{\theta},\vec{\theta'})\sim\gamma}[|f^{(i)}(\vec{\theta},\vec{x}) - f^{(i)}(\vec{\theta'},\vec{y})|^p] \tag{35}$$

$$\geqslant \sum_{i=1}^{q} \inf_{\gamma \in \mathcal{J}_{IND}(\mu,\mu)} E_{(\vec{\theta},\vec{\theta'})\sim\gamma}[|f^{(i)}(\vec{\theta},\vec{x}) - f^{(i)}(\vec{\theta'},\vec{y})|^p] \tag{36}$$

$$= \sum_{i=1}^{q} E_{\vec{\theta}\sim\mu}[|f^{(i)}(\vec{\theta},\vec{x}) - f^{(i)}(\vec{\theta},\vec{y})|^p] \qquad \text{by Equation 9} \tag{37}$$

The last equality is a result of Equation 9 and since $h(x-y) = |x-y|^p$ is a quasi-antitone function. Since $\bigotimes_{n=1}^{d} \mathcal{L}(\theta_n,\theta_n) \in \mathcal{J}_{IND}(\mu,\mu)$, equality follows. Next,

$$\widetilde{W}_{||\cdot||_p,p}(\mathcal{L}(\vec{X}),\mathcal{L}(\vec{Y})) = \left(\inf_{\gamma \in \mathcal{J}_{IND}(\mu,\mu)} E_{(\vec{\theta},\vec{\theta'})\sim\gamma}[||f(\vec{\theta},\vec{x}) - f(\vec{\theta'},\vec{y})||_p^p]\right)^{1/p} \quad \text{By Definition 3.3 \& Section 2}$$

$$= \left(\sum_{i=1}^{q} E_{\vec{\theta}\sim\mu}[|f^{(i)}(\vec{\theta},\vec{x}) - f^{(i)}(\vec{\theta},\vec{y})|^p]\right)^{1/p} \quad \text{By eq. 34}$$

$$= \left(E_{\vec{\theta}\sim\mu}[||f(\vec{\theta},\vec{x}) - f(\vec{\theta},\vec{y})||_p^p]\right)^{1/p}$$

□

7.6 Proof of Theorem 4.8

Proof. Part 1: coordinatewise monotone in Θ_n: Recall $\vec{X} \in \mathbb{R}^q$ and $\vec{X}^{(l)} = f^{(l)}(\vec{\theta}, \vec{x})$. Fix $j \in \{1, \ldots, d\}$. Define $\vec{\theta}_{1,-j} = \vec{\theta_1} \backslash \theta_{1,j}$.

Base step: If $\theta_{1,j} \leqslant \theta'_{1,j}$ μ_j-a.e., then for each $k \in \{1 \ldots q\}$ and $\vec{\theta}_{1,-j} \sim \mu_{-j}$, $f^{(k)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}) \leqslant f^{(k)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x})$ by monotonicity of $\vec{\theta}$. Thus for arbitrary $l \in \{1 \ldots q\}$,

$$
\begin{aligned}
& f^{(l)}(\theta_2, f((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x})) \\
& = f^{(l)}[\theta_2, (f^{(1)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}), f^{(2)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}), \ldots f^{(q)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}))] \\
& \leqslant f^{(l)}[\theta_2, (f^{(1)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x}), f^{(2)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}), \ldots f^{(q)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}))] \quad \text{Since } f^{(l)} \text{ is monotone in } x_1 \\
& \leqslant f^{(l)}[\theta_2, (f^{(1)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x}), f^{(2)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x}), \ldots f^{(q)}((\vec{\theta}_{1,-j}, \theta_{1,j}), \vec{x}))] \quad \text{Since } f^{(l)} \text{ is monotone in } x_2 \\
& \ldots \\
& \leqslant f^{(l)}[\theta_2, (f^{(1)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x}), f^{(2)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x}), \ldots f^{(q)}((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x}))] \quad \text{Since } f^{(l)} \text{ is monotone in } x_q \\
& = f^{(l)}[\theta_2, f((\vec{\theta}_{1,-j}, \theta'_{1,j}), \vec{x})]
\end{aligned}
$$

Since j was arbitrary, $f(\vec{\theta_2}, f(\vec{\theta_1}, \vec{x}))$, is coordinatewise monotone in $\vec{\theta_1}$. In other words, $f_2(\Theta_2, \vec{x})$ is monotone in Θ_2 where $\Theta_2 = (\vec{\theta_1}, \vec{\theta_2})$.

Induction step: Next, suppose that $f_n(\Theta_n, \vec{x})$ is monotone in Θ_n. Show that $f_{n+1}(\Theta_{n+1}, \vec{x})$ is monotone in Θ_{n+1}

Fix $i \in \{1 \ldots, n\}$ and $j \in \{1 \ldots, d\}$. Define θ_j as the jth coordinate of the Θ_n. Define $\Theta_{n,-j} = \Theta_n \backslash \theta_j$. If $\theta_j \leqslant \theta'_j$ μ_i-a.e., then for each $l \in \{1 \ldots q\}$, $\Theta_{n,-j} \sim \mu_{-j} \times \mu^{n-1}$-a.e. and $\vec{x}$ π-a.e. the following inequality holds, because f_n is coordinatewise monotone in Θ_n and f is coordinatewise monotone in $\vec{\theta}$.

$$
\begin{aligned}
f^{(l)}_{n+1}(\Theta_{n+1}, \vec{x}) &= f^{(l)}_{n+1}((\Theta_{n+1,-j}, \theta_j), \vec{x})) \\
&= f^{(l)}(\vec{\theta}_{n+1}, f_n((\Theta_{n,-j}, \theta_j), \vec{x})) \\
&\leqslant f^{(l)}(\vec{\theta}_{n+1}, f_n((\Theta_{n,-j}, \theta'_j), \vec{x}))
\end{aligned}
$$

Since j was arbitrary, $f_{n+1}(\Theta_{n+1}, \vec{x})$, is coordinatewise monotone in Θ_{n+1} if $f_n(\Theta_n, \vec{x})$, is coordinatewise monotone in Θ_n.

By induction and base step, $f_n(\Theta_n, \vec{x})$ is coordinatewise monotone in Θ_n, $n \geqslant 1$.

Theorem 4.7 of [32] states that the composition of two continuous functions is continuous. Since $f(\vec{\theta}, \vec{x})$ is continuous with respect to $\vec{\theta}$ and $\vec{x}$, $f_n(\Theta_n, \vec{x})$ is continuous with respect to Θ_n. Thus we apply Theorem 4.6 if $x \in \mathbb{R}$ or Theorem 4.7 if $\vec{x} \in \mathbb{R}^q$ to obtain Equation 11. □

7.7 Proof of Lemma 6.2

Proof of Lemma 6.2. Let $t \in \mathbb{R}$. We want to show that $P(f(g(U)) \leqslant t) = P(f(U) \leqslant t)$

$$\begin{aligned}
P(f(g(U)) \leqslant t) &= \int_{[0,1]} I(f(g(u)) \leqslant t)\mathrm{d}u \\
&= \int_a^b I(f(a+b-u) \leqslant t)\mathrm{d}u + \int_{(a,b)^C} I(f(u) \leqslant t)\mathrm{d}u \\
&= -\int_b^a I(f(v) \leqslant t)\mathrm{d}v + \int_{(a,b)^C} I(f(u) \leqslant t)\mathrm{d}u \\
&= \int_a^b I(f(u) \leqslant t)\mathrm{d}u + \int_{(a,b)^C} I(f(u) \leqslant t)\mathrm{d}u \\
&= \int_{[0,1]} I(f(u) \leqslant t)\mathrm{d}u \\
&= P(f(U) \leqslant t)
\end{aligned}$$

□

7.8 Proof of Lemma 6.3

Proof of Lemma 6.3. Let $x \in \mathbb{R}$. We want to show that $P(f(h(U, \xi)) \leqslant x) = P(f(U) \leqslant x)$.

First note the following integrals.

$$\int_{(a,b)} I(f(h(u,1)) \leqslant x)\mathrm{d}u \tag{38}$$

$$= \int_{(a,b)} I\left(f\left(\frac{u-a}{b-a}(t-a)+a\right) \leqslant x\right)\mathrm{d}u \tag{39}$$

$$= \int_{(a,t)} I\left(f\left(v\right) \leqslant x\right)\frac{b-a}{t-a}\mathrm{d}v \qquad \text{by a change of variable } v = \frac{u-a}{b-a}(t-a)+a \tag{40}$$

$$\int_{(a,b)} I(f(h(u,0)) \leqslant x)\mathrm{d}u \tag{41}$$

$$= \int_{(a,b)} I\left(f\left(\frac{u-a}{b-a}(b-t)+t\right) \leqslant x\right)\mathrm{d}u \tag{42}$$

$$= \int_{(t,b)} I\left(f\left(v\right) \leqslant x\right)\frac{b-a}{b-t}\mathrm{d}v \qquad \text{by a change of variable } v = \frac{u-a}{b-a}(b-t)+t \tag{43}$$

We show that $P(f(h(U,\xi)) \leqslant x) = P(f(U) \leqslant x)$ as follows,

$$
\begin{aligned}
&P(f(h(U,\xi)) \leqslant x) \\
&= P(f(h(U,\xi)) \leqslant x \mid \xi = 1)P(\xi = 1) + P(f(g(U,\xi)) \leqslant x \mid \xi = 0)P(\xi = 0) \\
&= P(f(h(U,1)) \leqslant x)\frac{t-a}{b-a} + P(f(h(U,0)) \leqslant x)\frac{b-t}{b-a} \\
&= \left(\int_{(a,b)} I(f(h(u,1)) \leqslant x)\mathrm{d}u + \int_{(a,b)^c} I(f(h(u,1)) \leqslant x)\mathrm{d}u\right)\frac{t-a}{b-a} + \\
&\qquad \left(\int_{(a,b)} I(f(h(u,0)) \leqslant x)\mathrm{d}u + \int_{(a,b)^c} I(f(h(u,0)) \leqslant x)\mathrm{d}u\right)\frac{b-t}{b-a} \\
&= \int_{(a,b)} I(f(h(u,1)) \leqslant x)\mathrm{d}u\frac{t-a}{b-a} + \int_{(a,b)} I(f(h(u,0)) \leqslant x)\mathrm{d}u\frac{b-t}{b-a} + \\
&\qquad \int_{(a,b)^c} I(f(u) \leqslant x)\mathrm{d}u \\
&= \int_{(a,t)} I\left(f\left(v\right) \leqslant x\right)\frac{b-a}{t-a}\mathrm{d}v\frac{t-a}{b-a} + \int_{(t,b)} I\left(f\left(v\right) \leqslant x\right)\frac{b-a}{b-t}\mathrm{d}v\frac{b-t}{b-a} + \\
&\qquad \int_{(a,b)^c} I(f(u) \leqslant x)\mathrm{d}u \qquad \text{By eq. 38, 41} \\
&= \int_{(a,t)} I\left(f\left(v\right) \leqslant x\right)\mathrm{d}v + \int_{(t,b)} I\left(f\left(v\right) \leqslant x\right)\mathrm{d}v + \int_{(a,b)^c} I(f(u) \leqslant x)\mathrm{d}u \\
&= \int_{[0,1]} I(f(u) \leqslant x)\mathrm{d}u \\
&= P(f(U) \leqslant x)
\end{aligned}
$$

□

References

[1] Julio Backhoff-Veraguas, Sigrid Källblad, and Benjamin A. Robinson. "Adapted Wasserstein distance between the laws of SDEs". In: *Stochastic Processes and their Applications* 189 (Nov. 2025), p. 104689. DOI: 10.1016/j.spa.2025.104689.

[2] Patrick Billingsley. *Probability and Measure: Anniversary Edition*. New York: Wiley Series in Probability and Statistics, 2012. ISBN: 978-1-118-34191-9.

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

[4] Vladimir I. Bogachev. *Measure theory. Vol. I, II*. Berlin: Springer-Verlag, 2007. DOI: 10.1007/978-3-540-34514-5.

[5] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417. DOI: 10.1002/cpa.3160440402.

[6] Eric W. Bridgeford. *Measure Theoretic Probability Theory: A Hands-On, Proof-Based Approach*. 2025. URL: https://ericwb.me/prob_book/coverpage.html.

[7] Steve Brooks et al., eds. *Handbook of Markov chain Monte Carlo*. Chapman & Hall, 2011. DOI: 10.1201/b10905.

[8] Krzysztof Burdzy, Zhen-Qing Chen, and Peter Jones. "Synchronous couplings of reflected Brownian motions in smooth domains". In: *Illinois Journal of Mathematics* 50.1-4 (2006), pp. 189–268. DOI: 10.1215/ijm/1258059475.

[9] 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.

[10] Guillaume Carlier, Alex Delalande, and Quentin Mérigot. "Quantitative Stability of the Pushforward Operation by an Optimal Transport Map". In: *Foundations of Computational Mathematics* 25.4 (2025), pp. 1259–1286. DOI: 10.1007/s10208-024-09669-4.

[11] Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. "Quantile and Probability Curves Without Crossing". In: *Econometrica* 78.3 (2010), pp. 1093–1125. DOI: 10.3982/ecta7880.

[12] J.A. Crowe, J.A. Zweibel, and P.C. Rosenbloom. "Rearrangements of functions". In: *Journal of Functional Analysis* 66.3 (1986), pp. 432–438. DOI: 10.1016/0022-1236(86)90067-4.

[13] George Deligiannidis et al. *On importance sampling and independent Metropolis-Hastings with an unbounded weight function*. 2024. arXiv: 2411.09514 [math.ST].

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

[15] Rick Durrett. *Probability: Theory and Examples*. 4th ed. New York: Cambridge Series in Statistical and Probabilistic Mathematics, 2010. DOI: 10.1017/9781108591034.

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

[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] Alison L. Gibbs and Francis Edward Su. "On Choosing and Bounding Probability Metrics". In: *International Statistical Review / Revue Internationale de Statistique* 70.3 (2002), pp. 419–435. DOI: 10.2307/1403865.

[19] Paul Glasserman and David D. Yao. "Some guidelines and guarantees for common random numbers". In: *Management Science* 38 (1992), pp. 884–908. DOI: 10.1287/mnsc.38.6.884.

[20] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. 1st ed. Cambridge: At the University Press, 1934.

[21] 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.

[22] Pierre E. Jacob. *Lecture notes for couplings and Monte Carlo*. URL: https://sites.google.com/site/pierrejacob/cmclectures?authuser=0.

[23] 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.

[24] 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.

[25] 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.

[26] 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: 10.3150/09-bej238.

[27] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport: With Applications to Data Science". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607. DOI: 10.1561/2200000073.

[28] 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.

[29] 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.

[30] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability surveys* 1 (2004), pp. 20–71. DOI: 10.1214/154957804100000024.

[31] Jeffrey S. Rosenthal. *A First Look at Rigorous Probability Theory*. 2nd ed. World Scientific, 2016. DOI: 10.1142/6300.

[32] Walter Rudin. *Principles of Mathematical Analysis*. 3rd. New York: McGraw-Hill, 1976. ISBN: 978-0070542358.

[33] Amir Sagiv. "The Wasserstein Distances Between Pushed-Forward Measures with Applications to Uncertainty Quantification". In: *Communications in Mathematical Sciences* 18.3 (2020), pp. 707–724. DOI: 10.4310/CMS.2020.v18.n3.a6.

[34] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling*. 1st ed. Birkauser, 2015. DOI: 10.1007/978-3-319-20828-2.

[35] Philip Savoye. "Equimeasurable Rearragnements of Functions and Fourth Order Boundary Value Problems". In: *The Rocky Mountain Journal of Mathematics* 26.1 (1996), pp. 281–293. DOI: 10.1216/rmjm/1181072116.

[36] Binyamin Schwarz. "On the Extrema of the Frequencies of Nonhomogeneous Strings with Equimeasurable Density". In: *Journal of Mathematics and Mechanics* 10.3 (1961), pp. 401–422.

[37] 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.

[38] Sabrina Sixta, Jeffrey S. Rosenthal, and Austin Brown. "Estimating MCMC convergence rates using common random number simulations". In: *Stochastic Models* (2026), pp. 1–42. DOI: 10.1080/15326349.2026.2688991.

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

[40] 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.

[41] A. W. van der Vaart. *Asymptotic Statistics*. 1st ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

[42] Cédric Villani. *Optimal Transport: Old and New*. 1st ed. Springer Berlin Heidelberg, 2008. DOI: 10.1007/978-3-540-71050-9.