Bounding Wasserstein distance with couplings

Niloy Biswas
Harvard University
niloy_biswas@g.harvard.edu

Lester Mackey
Microsoft Research New England
lmackey@microsoft.com

December 30, 2021

Abstract

Markov chain Monte Carlo (MCMC) provides asymptotically consistent estimates of intractable posterior expectations as the number of iterations tends to infinity. However, in large data applications, MCMC can be computationally expensive per iteration. This has catalyzed interest in sampling methods such as approximate MCMC, which trade off asymptotic consistency for improved computational speed. In this article, we propose estimators based on couplings of Markov chains to assess the quality of such asymptotically biased sampling methods. The estimators give empirical upper bounds of the Wasserstein distance between the limiting distribution of the asymptotically biased sampling method and the original target distribution of interest. We establish theoretical guarantees for our upper bounds and show that our estimators can remain effective in high dimensions. We apply our quality measures to stochastic gradient MCMC, variational Bayes, and Laplace approximations for tall data and to approximate MCMC for Bayesian logistic regression in 4500 dimensions and Bayesian linear regression in 50000 dimensions.

1 Introduction

1.1 Quality of asymptotically biased Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods are commonly used for the approximation of intractable integrals arising in Bayesian statistics, probabilistic inference, machine learning, and other fields [Brooks et al., 2011, Liu, 2008, Robert and Casella, 2013]. MCMC methods generate Markov chains with marginal distributions that asymptotically converge to a target distribution of interest as the number of iterations tend to infinity. In modern applications with a large number of data points or high dimensions, MCMC methods can have high computation cost per iteration. This has catalyzed the use of approximate MCMC methods [e.g., Welling and Teh, 2011, Bardenet et al., 2017, Narisetty et al., 2019, Johndrow et al., 2020], which have lower computation cost per iteration but may not
converge to the target distribution of interest, and methods such as variational inference \cite[e.g.][]{blei2017}, which inexactly approximate the target distribution through optimization.

Assessing the quality of such asymptotically biased samplers is of great interest for researchers who develop new approximate inference methods. Standard MCMC diagnostic tests \cite[e.g.,][]{johnson1998, biswas2019, vats2021, vechtari2021} are not directly suitable for such settings as they do not account for asymptotic bias. Researchers often resort to comparing summary statistics or marginal univariate traceplots of samples from such methods with samples from an asymptotically unbiased Markov chain. Such marginal traceplots and summary statistics may fail to capture higher order moments and dependencies between different components. Moreover, in high-dimensional settings, visualizing all marginal traceplots may not even be feasible. These limitations of existing diagnostics and heuristics have stimulated interest in measuring the quality of sample approximations. A number of works have developed measures of sample quality based on Stein discrepancies which do not require sampling from the target distribution of interest \cite{gorham2015, chwialkowski2016, liu2016, gorham2017, huggins2018, gorham2020}. \cite{gorham2019} have further established a near-linear relationship between Stein discrepancies and standard Wasserstein distances for distributions with fast-mixing diffusions, but the constants in these results rely on specific knowledge of the gradient of the log target density that must be derived for each new target distribution. In this manuscript, we develop generic upper bound estimates of the Wasserstein distance that apply to any distributions that can be targeted with fast-mixing Markov chains and do not require any additional distributional knowledge.

### 1.2 Couplings and Wasserstein distances

Consider a complete, separable metric space \((\mathcal{X}, c)\) where \(c\) is a metric. A probability measure \(\mu\) on \((\mathcal{X}, c)\) has finite moments of order \(p\) if there exists some \(x_0 \in \mathcal{X}\) such that \(\int_{\mathcal{X}} c(x_0, x)^p d\mu(x) < \infty\). For any \(p \geq 1\), let \(\mathcal{P}_p(\mathcal{X})\) denote the set of all probability measures on \((\mathcal{X}, c)\) which have finite moments of order \(p\). Then the \(p\)-Wasserstein distance is a metric on \(\mathcal{P}_p(\mathcal{X})\), defined for any probability measures \(\mu\) and \(\nu\) in \(\mathcal{P}_p(\mathcal{X})\) as

\[
W_p(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y)^p d\gamma(x, y) \right)^{1/p}
\]

where \(\Gamma(\mu, \nu)\) is the set of probability measures on \(\mathcal{X} \times \mathcal{X}\) with marginal measures \(\mu\) and \(\nu\) respectively. Any probability measure in \(\Gamma(\mu, \nu)\) is called a coupling of \(\mu\) and \(\nu\).

The Wasserstein distance has many advantageous properties. Here we note those most relevant for this work and refer to \cite{villani2008} for more details. Firstly, it allows comparison between mutually singular distributions that may have disjoint supports, unlike common alternatives like the total variation distance, Kullback–Leibler divergence and Rényi’s \(\alpha\)-divergences \cite{van2014}. In addition, it captures geometric properties induced by the metric \(c\) and differences between moments of distributions. For example when \(\mathcal{X} = \mathbb{R}^d\) and \(c(x, y) = \|x - y\|_p = (\sum_{i=1}^d |x_i - y_i|^p)^{1/p}\),
Jensen’s inequality and the triangle inequality imply

$$\max \{ \left\| \mathbb{E}[X - Y] \right\|_p, |\mathbb{E}[\|X\|_p^{1/p}] - \mathbb{E}[\|Y\|_p^{1/p}]| \} \leq \mathbb{E}[\|X - Y\|_p^{1/p}] = \mathcal{W}_p(\mu, \nu)$$

(1)

for any $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ and random variables $X \sim \mu$ and $Y \sim \nu$ jointly distributed according to a $p$-Wasserstein optimal coupling. Equation (1) shows that $p$-Wasserstein distances can control the difference between moments of order $p$. Meanwhile, Huggins et al. [2020] present examples of pairs of distributions with large moment differences but arbitrarily small Kullback–Leibler divergence or Rényi’s $\alpha$-divergence.

This manuscript develops consistent estimates of upper bounds to Wasserstein distances. The developed algorithms and estimators are then used to assess the quality of approximate MCMC and certain variational inference methods. Specifically, we use couplings of Markov chains to estimate upper bounds on the Wasserstein distance between the limiting distribution of the asymptotically biased sampling method and the original target distribution of interest. As we discuss in Section 3.4, our work provides an appealing alternative to popular biased estimates based on empirical Wasserstein distances and Sinkhorn distances, [e.g., Orlin, 1988, Cuturi, 2013] which can suffer from curse of dimensionality, and to the upper bounds estimates of Huggins et al. [2020], which are based on worst-case divergence bounds and rely on efficient importance sampling. In addition, our upper bound estimates provably improve upon those of Dobson et al. [2021] which rely upon challenging contraction-constant estimation.

1.3 Our contributions

We introduce new tools for method developers to assess quality of their approximate inference procedures. Our primary contributions are summarized below.

In Section 2, we first introduce algorithms for coupling two Markov chains with distinct stationary distributions. Our approach generalizes recent efforts to couple Markov chains with identical transition kernels [see, e.g., Glynn and Rhee, 2014, Heng and Jacob, 2019, Middleton et al., 2019, Jacob et al., 2020, Biswas et al., 2019, 2021]. We then introduce estimators based on our coupled chains that consistently upper bound the Wasserstein distance between their stationary distributions. This enables us to assess the asymptotic bias of approximate MCMC methods and certain variational inference procedures.

Section 3 provides a theoretical analysis of our upper bound estimates. We first establish the consistency and unbiasedness of our upper bound estimates and then derive interpretable analytic upper bounds on our estimates in terms of the mixing rate of one chain and the closeness of the two transition kernels. These analytic bounds provide sufficient conditions for our upper estimates to be informative in high dimensions.

In Section 4, we demonstrate the favorable empirical performance of our upper bound estimates on modern applications. We first consider datasets with a large number of data points to assess the quality of stochastic gradient MCMC, variational Bayes, and Laplace approximations for Bayesian logistic regression. We then consider high-dimensional datasets to assess the quality of approximate MCMC for high-dimensional linear regression with continuous shrinkage priors ($d \approx 50000$) and high-dimensional logistic regression with spike-and-slab priors ($d \approx 4500$). Finally, we discuss our
results and directions for future work in Section 5. Open-source R code recreating all experiments in this paper can be found at github.com/niloyb/BoundWasserstein.

2 Bounding Wasserstein distance with couplings

Given distributions $P$ and $Q$ in $\mathcal{P}_p(\mathcal{X})$ for some $p \geq 1$, we wish to estimate upper bounds on $W_p(P,Q)$. Our estimates are based on Markov chains $(X_t)_{t \geq 0}$ and $(Y_t)_{t \geq 0}$ with marginal transition kernels $K_1$ and $K_2$ invariant for $P$ and $Q$ respectively. Specifically, we construct Markovian kernels $\bar{K}$ on the joint space $\mathcal{X} \times \mathcal{X}$ such that for all $x, y \in \mathcal{X}$,

$$\bar{K}((x,y),(\cdot,\mathcal{X})) = K_1(x,\cdot) \text{ and } \bar{K}((x,y),(\mathcal{X},\cdot)) = K_2(y,\cdot). \quad (2)$$

Given the kernel $\bar{K}$, we generate a coupled Markov chain $(X_t, Y_t)_{t \geq 0}$ using Algorithm 1. Algorithm 1 can be viewed as a generalization of existing coupling constructions [Johnson, 1998, Glynn and Rhee, 2014, Heng and Jacob, 2019, Middleton et al., 2019, Jacob et al., 2020, Biswas et al., 2019, 2021]. While prior work focused on $K_1 = K_2$ and $X_t \overset{d}{=} Y_t$ to establish convergence to a single stationary distribution $P$, our work uses distinct kernels $K_1$ and $K_2$ to bound the distance between distinct stationary distributions $P$ and $Q$. Algorithms to sample from the coupled kernel $\bar{K}$ are covered in Section 3.2.

Algorithm 1: Coupled Markov chain Monte Carlo for bounding Wasserstein distances

**Input:** Initial distribution $\bar{I}_0$ on $\mathcal{X} \times \mathcal{X}$, joint kernel $\bar{K}$, number of iterations $T$

**Initialize:** Sample $(X_0, Y_0) \sim \bar{I}_0$

for $t = 1, \ldots, T - 1$ do

Sample $(X_{t+1}, Y_{t+1}) | (X_t, Y_t) \sim \bar{K}((X_t, Y_t), \cdot)$

**return** Markov chain $(X_t, Y_t)_{t=0}^T$

For a Markov chain $(X_t, Y_t)_{t \geq 0}$ from Algorithm 1, suppose the marginal distributions of $X_t$ and $Y_t$ converge in $p$-Wasserstein distance to $P$ and $Q$ respectively as $t$ tends to infinity. Informally, the coupling representation of the Wasserstein distance implies $W_p(P,Q)^p \leq \liminf_{S \to \infty, T \to \infty} \sum_{t=S+1}^T E[c(X_t, Y_t)^p]$. This motivates our coupling upper bound (CUB) estimate

$$\text{CUB}_p \triangleq \left( \frac{1}{T-S} \sum_{i=1}^T \sum_{t=S+1}^T c(X_t^{(i)}, Y_t^{(i)})^p \right)^{1/p}, \quad (3)$$

where $(X_t^{(i)}, Y_t^{(i)})_{t=0}^T$ are sampled using Algorithm 1 independently for each $i$, with burn-in $S \geq 0$ and trajectory length $T > S$. We prove the consistency of this and related upper bound estimators in Section 3. We now consider the empirical performance of this estimator on two stylized examples, working with the Euclidean metric $c(x,y) = \|x - y\|_2$ on $\mathbb{R}^d$. 


2.1 Upper bound on Wasserstein distance

We consider the performance of CUB$_2$ (3) for two Gaussian distributions on $\mathbb{R}^d$, given by

$$P = \mathcal{N}(0, \Sigma) \text{ where } \Sigma_{i,j} = 0.5^{(i-j)} \text{ for } 1 \leq i, j \leq d \text{ and } Q = \mathcal{N}(0, I_d).$$

Here we use the marginal kernels $K_1$ and $K_2$ of the Metropolis–adjusted Langevin algorithm (MALA), with step sizes $\sigma_P = \sigma_Q = 0.5d^{-1/6}$ targeting $P$ and $Q$ respectively, following existing guidance for step size choice [Roberts and Rosenthal, 1998]. The joint kernel $\bar{K}$ is based on a common random numbers (CRN, also called “synchronous”) coupling of both the proposal step and the accept-reject step of the MALA algorithm, as detailed in Algorithm 4 of Appendix F. We initialize each chain with independent draws of $X_0^{(i)} \sim P$ and $Y_0^{(i)} \sim Q$.

Figure 1a compares several upper bound estimates for $W_2(P, Q)$ with dimension $d = 100$. The solid line plots CUB$_2$ based on $I = 5$ independent chains with burn-in $S = 0$ and varying trajectory length $1 \leq T \leq 1000$. The dotted line corresponds to the independent coupling upper bound $\mathbb{E}[\|X - Y\|_2^2]^{1/2}$ where $X \sim P$ and $Y \sim Q$ are independent and equals $(2d)^{1/2}$. The dot-dashed line represents an estimate based on empirical Wasserstein distances. Specifically, it plots $\sum_{i=1}^I W_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})/I$ where each $\hat{P}_T^{(i)}$ and $\hat{Q}_T^{(i)}$ are the empirical distributions of $T = 1000$ points sampled independently from $P$ and $Q$ respectively and $W_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})$ is calculated exactly by solving a linear program [Orlin, 1988]. In Section 3.4 we establish the consistency of this common Wasserstein distance estimate but also note that its convergence can be slow in high dimensions due to substantial bias. The dashed line corresponds to the true Wasserstein distance $W_2(P, Q)$ and equals $\|\Sigma^{1/2} - I_d\|_F$, where $\Sigma^{1/2}$ is the positive matrix square root of $\Sigma$ and $\|\cdot\|_F$ is the Frobenius norm on matrices [see, e.g., Peyré and Cuturi, 2019, Remark 2.23]. The grey error bands represent one standard error arising from Monte Carlo error. At initialization ($T = 0$) CUB$_2$ matches the equivalent independent coupling bound. For greater trajectory lengths $T$, CUB$_2$ offers a significant improvement over both the independent bound and the popular empirical Wasserstein estimate.

Figure 1b considers the 2-Wasserstein distance $W_2(P, Q)$ for higher dimensions. The solid line now plots CUB$_2$ based on $I = 5$, $S = 0$, and $T = 1000$. Figure 1b highlights that, unlike the independent and empirical Wasserstein estimates, CUB$_2$ offers bounds that remain informative even in higher dimensions. Such dimension-free properties of our upper bounds are investigated in Section 3. Section 3.4 provides a further comparison of our CUB bounds with empirical Wasserstein and Sinkhorn distances, which can have prohibitive computational cost for larger sample sizes and suffer from the curse of dimensionality.

2.2 Bias of approximate MCMC methods

Unadjusted Langevin algorithm (ULA) is a popular approximate MCMC counterpart to MALA. It has the same proposal step as MALA but now all proposed states accepted. The lack of a Metropolis–Hastings accept-reject step leads to ULA having a lower computation costs per iteration than MALA, which is beneficial for applications with large datasets [e.g. Nemeth and Fearnhead, 2021]. On the other hand, ULA is asymptotically biased [Durmus and Moulines, 2019]. In this section, we consider upper bounds of the Wasserstein distance between the limiting distribution of
Figure 1: Upper bound estimates for $W_2(P,Q)$ with $P = \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.5|i-j|$ for $1 \leq i, j \leq d$, $Q = \mathcal{N}(0, I_d)$, and metric $c(x,y) = ||x - y||_2$. See Section 2.1.

Figure 2 compares several approaches to bounding the asymptotic $W_2(P,Q)$ bias of ULA. The solid line displays our coupling upper bound estimate. For each dimension $d$, it is calculated using CUB$_2$ (3) with $I = 10$, $S = 1000$, and $T = 3000$. The dashed line shows the true asymptotic bias $W_2(P,Q)$ and the dotted line shows the independent coupling upper bound, both of which can be computed exactly in this example. The dot-dashed line plots the analytic ULA bias upper bounds of Durmus and Moulines [2019, Corollary 9] (see Appendix A.2 for more details). The tailored Durmus-Moulines bounds are significantly tighter than the convenient independent coupling bound, but CUB is tighter still, offering significantly improved estimates for all dimensions.

3 Properties and Implementation

In this section we establish the consistency of the estimators in Section 2, describe how to sample from the joint kernel $ar{K}$ in Algorithm 1, investigate the theoretical properties of our upper bounds,
and compare to alternative approaches. All proofs are in Appendix B.

3.1 Consistency of coupling upper bounds

We begin by establishing the consistency of coupling upper bound estimators. Our first result bounds the Wasserstein distance between coupled chains in terms of an instantaneous CUB estimator related to the time-averaged estimator in (3).

Proposition 3.1 (Consistency of instantaneous CUB). Let \((X_t^{(i)}, Y_t^{(i)})_{t \geq 0}\) for \(i = 1, \ldots, I\) denote coupled chains generated independently from Algorithm 1 with marginal distributions \(X_t^{(i)} \sim P_t\) and \(Y_t^{(i)} \sim Q_t\) at time \(t\). For each \(t \geq 0\), define the instantaneous CUB estimator

\[
CUB_{p,t} \triangleq \left( \frac{1}{I} \sum_{i=1}^{I} c(X_t^{(i)}, Y_t^{(i)})^p \right)^{1/p}.
\]

If \(P_s\) and \(Q_s\) have finite moments of order \(p\) for all \(s \leq t\), then \(CUB_{p,t}\) has finite moments of order \(p\), and, as \(I \to \infty\),

\[
CUB_{p,t} \overset{a.s.}{\to} \mathbb{L}^1 \mathbb{E}[CUB_{p,t}^p] \geq W_p(P_t, Q_t)^p.
\]

Our next result shows that the estimator \(CUB_p\) (3) consistently bounds the Wasserstein distance between time-averaged marginal distributions.

Corollary 3.2 (Consistency of CUB for time-averaged marginals). Under the assumptions and notation of Proposition 3.1, consider the estimator \(CUB_p\) (3) with any number of independent chains \(I \geq 0\), and trajectories with burn-in \(S \geq 1\) and length \(T \geq S\). Then \(CUB_p\) has finite moments of order \(p\), and as \(I \to \infty\),

\[
CUB_p \overset{a.s.}{\to} \mathbb{L}^1 \mathbb{E}[CUB_p^p] \geq W_p \left( \frac{1}{T-S} \sum_{t=S+1}^{T} P_t, \frac{1}{T-S} \sum_{t=S+1}^{T} Q_t \right)^p.
\]

An important implication of Corollary 3.2 is that \(CUB_p\) (3) consistently bounds the Wasserstein distance between stationary distributions whenever its chains are marginally initialized at stationarity.
Corollary 3.3 (Consistency of CUB with stationary initialization). Under the assumptions and notation of Proposition 3.1, suppose kernels $K_1$ and $K_2$ have stationary distributions $P$ and $Q$ respectively, where $P$ and $Q$ have finite moments of order $p$. Suppose we initialize $(X_0, Y_0) \sim I_0$ such that $X_0 \sim P$ and $Y_0 \sim Q$ marginally. Then for any number of independent chains $I \geq 0$, trajectories with burn-in $S \geq 1$ and length $T \geq S$, the estimator CUB$_p$ (3) has finite moments of order $p$, and as $I \to \infty$,

$$\text{CUB}_p^p \xrightarrow{a.s.} L^1} \mathbb{E}[\text{CUB}_p^p] \geq W_p(P, Q)^p.$$  

We may not always be able to initialize using the marginal stationary distributions $P$ and $Q$. To obtain upper bounds on $W_p(P, Q)$ without starting at the marginal stationary distributions $P$ and $Q$, we make an assumption related to convergence of the Markov chain marginals $(P_t)_{t \geq 0}$ and $(Q_t)_{t \geq 0}$.

Assumption 3.4 (Convergence of marginal chains). As $t \to \infty$, $P_t$ and $Q_t$ converge in $p$-Wasserstein distance respectively to $P$ and $Q$ with finite moments of order $p$.

Proposition 3.5 (Consistency when chain marginals converge). Under Assumption 3.4 and the assumptions and notation of Proposition 3.1, for all $\epsilon > 0$ there exists some $S \geq 1$ such that for all $T \geq S$, the estimator CUB$_p$ (3) has finite moments of order $p$, and as $I \to \infty$,

$$\text{CUB}_p^p \xrightarrow{a.s.} L^1} \mathbb{E}[\text{CUB}_p^p] \geq W_p(P, Q)^p - \epsilon.$$  

Proposition 3.5 establishes that CUB$_p$ with any initialization $(X_0, Y_0) \sim I_0$ consistently bounds $W_p(P, Q)$ as $I$ and $S$ grow. In practice, we can use standard MCMC burn-in diagnostics to select an appropriate burn-in level for our marginal chains of interest [e.g., Johnson, 1998, Biswas et al., 2019, Vats and Knudson, 2021, Vehtari et al., 2021]. Alternatively, for $p = 1$, we can avoid burn-in removal and instead directly correct our bound for non-stationarity using the recent $L$-lag coupling approach of Biswas et al. [2019] (see Appendix A.3 for details).

We emphasize that the results of this section hold for any coupled chain sampled using Algorithm 1 with joint kernel $\bar{K}$ satisfying (2). For example, this includes both the CRN coupled chains and the independently coupled chains from Section 2, where the CRN coupled chains produced more informative upper bounds empirically as shown in Figures 1 and 2. We now consider how to sample from the joint kernel $\bar{K}$ and investigate when our upper bounds are informative.

3.2 Algorithms to sample from the coupled kernel $\bar{K}$

In this section, we develop algorithms to sample from the joint kernel $\bar{K}$ such that the estimators from Section 3.1 can produce informative upper bounds. Our construction makes use of the kernels $\Gamma_1$ on $\mathcal{X} \times \mathcal{X}$ and $\Gamma_\Delta$ on $\mathcal{X}$ such that:

1. $\Gamma_1$ is a Markovian coupling of the kernel of $K_1$: for all $x, y \in \mathcal{X}$, $\Gamma_1(x, y)$ is a coupling of the distributions $K_1(x, \cdot)$ and $K_1(y, \cdot)$.

2. $\Gamma_\Delta$ is coupling of kernels $K_1$ and $K_2$ from the same point: for all $z \in \mathcal{X}$, $\Gamma_\Delta(z)$ is a coupling of the distributions $K_1(z, \cdot)$ and $K_2(z, \cdot)$.

8
The kernel $\Gamma_1$ is based on the marginal chain corresponding to $K_1$. For example, when $K_1$ is a Metropolis–Hastings kernel, $\Gamma_1$ can be a CRN coupling of both the proposal step and the accept-reject step. When the Metropolis–Hastings proposal is based on a spherically symmetric distribution such as a Gaussian (as in random walk Metropolis–Hastings or the momentum component in Hamiltonian Monte Carlo), $\Gamma_1$ can be a reflection coupling of the proposal step and a CRN coupling of the accept-reject step [e.g. Bou-Rabee et al., 2020, Wang et al., 2021]. The kernel $\Gamma_\Delta$ characterizes the perturbation between the marginal kernels $K_1$ and $K_2$. For example, when $K_1$ and $K_2$ are MALA and ULA kernels respectively, $\Gamma_\Delta$ can be a CRN coupling of the proposal step. This leads to identical proposals when MALA and ULA have the same step size, but the MALA chain will have a further accept-reject step while the ULA chain will always accept the proposal. We discuss the choice of $\Gamma_1$ and $\Gamma_\Delta$ further in Section 3.3.

Given $\Gamma_1$ and $\Gamma_\Delta$, we sample from the joint kernel $\tilde{K}$ using Algorithm 2.

**Algorithm 2**: Joint kernel $\tilde{K}$ which couples the marginal kernels $K_1$ and $K_2$

**Input**: Chain states $X_{t-1}$ and $Y_{t-1}$, kernels $K_1$ and $K_2$, coupled kernels $\Gamma_1$ and $\Gamma_\Delta$

**Sample** $(X_t, Z_t, Y_t) \mid X_{t-1}, Y_{t-1}$ such that $(X_t, Z_t) \sim \Gamma_1(X_{t-1}, Y_{t-1})$, $(Z_t, Y_t) \sim \Gamma_\Delta(Y_{t-1})$

**return** $(X_t, Y_t)$

Algorithm 2 gives the conditional marginal distributions $X_t \mid X_{t-1}, Y_{t-1} \sim K_1(X_{t-1}, \cdot)$, $Z_t \mid X_{t-1}, Y_{t-1} \sim K_1(Y_{t-1}, \cdot)$, $Y_t \mid X_{t-1}, Y_{t-1} \sim K_2(Y_{t-1})$ so that $\tilde{K}$ satisfies (2). Often Algorithm 2 can be implemented without explicitly sampling $Z_t$. As an example, consider when $K_1$ and $K_2$ are MALA and ULA kernels with step sizes $\sigma_P$ and $\sigma_Q$, target distributions $P$ and $Q$, and $\Gamma_1$ and $\Gamma_\Delta$ are CRN coupled kernels. Given $(X_{t-1}, Y_{t-1})$, we sample $\epsilon_{CRN} \sim N(0, I_d)$ and calculate the proposals $X^* = X_{t-1} + (\sigma_P^2/2)\nabla \log P(X_{t-1}) + \sigma_P \epsilon_{CRN}$, $Z^* = Y_{t-1} + (\sigma_Q^2/2)\nabla \log P(Y_{t-1}) + \sigma_Q \epsilon_{CRN}$, and $Y^* = Y_{t-1} + (\sigma_Q^2/2)\nabla \log Q(Y_{t-1}) + \sigma_Q \epsilon_{CRN}$. Then we accept or reject proposals $X^*$ and $Z^*$ based on a Metropolis–Hastings correction with a common random number $U_{CRN} \sim \text{Uniform}(0,1)$ to obtain $X_t$ equal to $X^*$ or $X_{t-1}$, $Z_t$ equal to $Z^*$ or $Y_{t-1}$, and always accept $Y^*$ to obtain $Y_t = Y^*$. Notably, $Z_t$ need not be explicitly sampled to perform this update of $(X_t, Y_t)$. This CRN coupling of MALA and ULA is included in Algorithm 5 of Appendix F. Appendix F also details general CRN and reflection couplings between two Metropolis–Hastings kernels.

We now cover implementation practicalities and potential limitations.

**Number of coupled chains and chain length to simulate.** We first highlight the value of averaging over time and over independent coupled chains when producing upper bound estimates. Figures 3a and 3b examine the performance of the CUB$_p$ (3) and instantaneous CUB$_{p,t}$ (5) estimators when bounding the 1-Wasserstein distance with $\phi(x, y) = \|x - y\|_2$ between

$$P = \frac{1}{2}N(1_d, I_d) + \frac{1}{2}N(-1_d, I_d) \quad \text{and} \quad Q = N(1_d, I_d) \quad \text{with} \quad d = 4$$

so that one of the marginal target distributions is bimodal with well-separated modes. We simulate the coupled chains $(X_t^{(i)}, Y_t^{(i)})_{i \geq 0}$ independently for each $i$ using Algorithm 1, where the joint kernel $\tilde{K}$ is based on a CRN coupling of MALA kernels $K_1$ and $K_2$ targeting distributions $P$ and $Q$.
respective. The MALA kernels have a common step size $d^{-1/6}$, and we initialize $X_0^{(i)} = 1_d$ and $Y_0^{(i)} = 1_d$ such that both marginal chains start at the common mode. Figure 3a isolates the impact of averaging over multiple chains when computing the $\text{CUB}_{p,t}$ estimate (5). The grey dotted line shows the single trajectory $(c(X_t^{(1)}, Y_t^{(1)}))_{t=1}^{1000}$ and the black solid line shows the averaged trajectory $(\bar{c}(X_t, Y_t))_{t=1}^{1000}$ where $\bar{c}(X_t, Y_t) \triangleq \sum_{i=1}^I c(X_t^{(i)}, Y_t^{(i)})/I$ for $I = 100$ independent chains. The grey dotted line alternates between values close to 0 or 4, corresponding to when the marginal chains from a single trajectory are both near the common mode ($1_d$) or near different modes ($-1_d$ and $1_d$) respectively. This illustrates that instantaneous upper bound estimator $\text{CUB}_{p,t}$ (5) based on only a single trajectory of short chain length can have high variance. For multiple independent coupled chains, the averaged trajectory has lower variance and higher precision as shown by the grey confidence bands and the black solid line which remains close to the true $W_1(P,Q)$ distance (shown by black dotted line). Conveniently, these multiple chains can be simulated in parallel. Also even for upper bound estimates based on a single chain, the $\text{CUB}_p$ estimator with $I = 1$ and a sufficiently large chain length $T$ can produce estimates with low variance, as shown by the grey confidence bands and the black solid line in Figure 3b. A formal study of the optimal choice between number of independent coupled chains and chain length, given a certain coupled kernel $\bar{K}$ and a fixed number of parallel processors is an open area for further investigation. Jacob et al. [2020] contain related motivating discussions in the context of unbiased estimation with couplings.

**Choice of coupled kernel.** Secondly, we highlight the importance of the choice of the coupled kernel $\bar{K}$. Figure 3c examines the performance of the $\text{CUB}_p$ (3) estimator when bounding the 1-Wasserstein distance with $c(x,y) = \|x - y\|_2$ between

$$P = \frac{1}{2} \mathcal{N}(2,1) + \frac{1}{2} \mathcal{N}(-2,1) \quad \text{and} \quad Q = \frac{1}{2} \mathcal{N}(1,1) + \frac{1}{2} \mathcal{N}(-1,1),$$

Figure 3: Impact of multiple trajectories, ergodic averaging, and coupling choice on coupling bound quality for the 1-Wasserstein distance with $c(x,y) = \|x - y\|_2$. See Section 3.2.
so that now both the marginal target distributions are bimodal. Under this setup, we simulated coupled chains based on both a CRN coupling and a reflection coupling of MALA kernels $K_1$ and $K_2$ targeting distributions $P$ and $Q$ respectively. The MALA kernels have a common step size $2$, and we initialize such that each $X_0^{(i)} \sim P$ and $Y_0^{(i)} \sim Q$ are independent. In Figure 3c, the grey and black solid lines show averaged trajectories from $I = 1000$ independent coupled chains based on CRN and reflection coupling respectively. It highlights that reflection coupling gives tighter upper bounds compared to CRN for this example. In general, the choice of coupling can have an impact on the tightness of our upper bounds. We emphasize that any choice of such couplings still produces consistent upper bounds (as shown in Section 3.1). In practice, one can simulate different coupling algorithms to empirically assess which choice produces the tightest upper bounds and even select the smallest of multiple coupling bounds. Finally, Figure 3c highlights that our upper bounds may not always be very close to the true Wasserstein distance. Alternative coupling algorithms and tailored Wasserstein distance upper bounds between mixtures of distributions could give further improvements for this example.

### 3.3 Interpretable upper bounds for CUB

So far we have established that CUB (3) consistently upper bounds Wasserstein distances (Section 3.1) and developed algorithms to compute CUB in practice (Sections 3.2). We next derive interpretable upper bounds on the size of CUB. Our analysis is based on Markov chain perturbation theory for the 1-Wasserstein distance [Pillai and Smith, 2015, Johndrow and Mattingly, 2018, Rudolf and Schweizer, 2018], which we generalize to $p$-Wasserstein distances for all $p \geq 1$. This is a useful extension, as the 2-Wasserstein distance in particular is believed to better reflect geometric features and adapt to geometric structure than the 1-Wasserstein distance [Villani, 2008, Remark 6.6]. We also discuss examples where the upper bounds on the Wasserstein distance do not explicitly depend on the dimension of the state space and are stable up to a coupling of the one-step marginal kernels.

To establish our CUB upper bounds, we assume the Markovian coupling $\Gamma_1$ in Algorithm 2 gives uniform contraction in Wasserstein distance.

**Assumption 3.6 (Uniform contraction).** There exists a constant $\rho \in (0, 1)$ such that for all $X_t, Y_t \in \mathcal{X}$ and $(X_{t+1}, Y_{t+1})| (X_t, Y_t) \sim \Gamma_1(X_t, Y_t)$, $\mathbb{E}[c(X_{t+1}, Y_{t+1})^p|X_t, Y_t]^1/p \leq \rho c(X_t, Y_t)$.

Assumption 3.6 is stronger than the convergence assumption of the marginal chain corresponding to kernel $K_1$ (Assumption 3.4 for $(P_t)_{t \geq 0}$). For many popular MCMC algorithms, Assumption 3.6 has been established under certain metrics $c$ and coupled kernel $\Gamma_1$ to give contraction rates $\rho$ that do not explicitly depend on the dimension on the state space $\mathcal{X}$. This includes MALA [Eberle, 2014], and Hamiltonian Monte Carlo (HMC) [Bou-Rabee et al., 2020]. When the target distributions are log-concave, these algorithms satisfy Assumption 3.6 with $c(x, y) = \|x - y\|_2$ and the coupled kernel $\Gamma_1$ based on a CRN coupling. For target distributions (including, for example, multimodal distributions with Gaussian tails) satisfying a weaker distant dissipativity condition [Eberle, 2016, Gorham et al., 2019], these algorithms satisfy Assumption 3.6 with $\Gamma_1$ based on a combination of CRN and reflection coupling and the modified metric $\tilde{c}$ satisfying $r \tilde{c}(x, y) \leq \|x - y\|_2 \leq R \tilde{c}(x, y)$ for some $0 < r \leq R < \infty$. 

11
Furthermore, we can weaken Assumption 3.6 to a geometric ergodicity condition as in [Rudolf and Schweizer, 2018], where for some constants $C, \gamma \in (0, 1)$, and for all $L \geq 1$, $\mathbb{E}[c(X_{t+L}, Y_{t+L})^p | X_t, Y_t]^{1/p} \leq C p^t c(X_t, Y_t)$ for $(X_{t+L}, Y_{t+L})|(X_t, Y_t) \sim \Gamma_{\rho}^t(X_t, Y_t)$. Our analysis then is based on the construction of a multi-step coupling kernel. This may be of independent interest and is included in Appendix D for completeness.

Under Assumption 3.6, we can upper bound the distance from our coupled chains explicitly in terms of the initial distribution $\tilde{I}_0$, contraction constant $\rho$, and coupled kernel $\Gamma_{\Delta}$ corresponding to perturbations between the marginal kernels $K_1$ and $K_2$.

**Theorem 3.7 (CUB upper bound).** Let $(X_t, Y_t)_{t \geq 0}$ denote a coupled Markov chain generated using Algorithm 1 with initial distribution $I_0$ and joint kernel $K$ from Algorithm 2. Suppose the coupled kernel $\Gamma_1$ satisfies Assumption 3.6 for some $\rho \in (0, 1)$. Then

$$\mathbb{E}[\text{CUB}_{p, t}^{\rho}]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^t \rho^{t-i} \mathbb{E}[\Delta_p(Y_{i-1})]^{1/p}$$

for all $t \geq 0$, where $(X_0, Y_0) \sim \tilde{I}_0$ and $\Delta_p(z) := \mathbb{E}[c(X, Y)^p | z]$ for $(X, Y)|z \sim \Gamma_\Delta(z)$.

For CUB$_{p, t}$ based on a metric $c$, one obtains an analogous bound if Assumption 3.6 instead holds for a dominating metric $\tilde{c}$, i.e., for $\tilde{c}$ satisfying $c(x, y) \leq R \tilde{c}(x, y)$ for some constant $R \in (0, \infty)$. Then $\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq R \mathbb{E}[c(X_t, Y_t)^p]^{1/p}$. Also, when the marginal distributions $(Q_t)_{t \geq 0}$ converge, we can obtain a simpler expression for the upper bound.

**Corollary 3.8 (CUB upper bound under marginal convergence).** Under the notation and assumptions of Theorem 3.7, suppose that the marginal distributions $Q_t$ converge in $p$-Wasserstein distance to some distribution $Q$ as $t \to \infty$. Then for all $\epsilon > 0$, there exists some $S \geq 1$ such that for all $t \geq S$,

$$\mathbb{E}[\text{CUB}_{p, t}^{\rho}]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1 - \rho^t) \frac{\mathbb{E}[\Delta_p(Y^*)]^{1/p}}{1 - \rho} + \epsilon,$$

where $(X_0, Y_0) \sim \tilde{I}_0$, $\Delta_p(z) := \mathbb{E}[c(X, Y)^p | z]$ for $(X, Y) \sim \Gamma_\Delta(z)$, and $Y^* \sim Q$.

Corollary 3.8 gives $W_p(P, Q) \leq \liminf_{t \to \infty} \mathbb{E}[\text{CUB}_{p, t}^{\rho}]^{1/p} \leq \mathbb{E}[\Delta_p(Y^*)]^{1/p}/(1 - \rho)$. It implies that estimators from our coupled chains may give informative empirical upper bounds when the expected perturbation $\mathbb{E}[\Delta_p(Y^*)]$ for $Y^* \sim Q$ is small. Further if the contraction rate $\rho$ does not explicitly depend on the dimension, then our upper bounds remain informative in high dimensional settings, as observed in Figures 1b and 2.

Our next result covers the case in which the marginals $(Q_t)_{t \geq 0}$ do not converge to any limiting distribution in $p$-Wasserstein distance. In this case, our upper bound is in terms of perturbations between the marginal kernels weighted by a Lyapunov function of $K_2$.

**Proposition 3.9 (CUB upper bound weighted by a Lyapunov function).** Under the notation and assumptions of Theorem 3.7, let $V : \mathcal{X} \to [0, \infty)$ satisfy $\mathbb{E}[V(Y_{t+1})|Y_t = z] \leq \gamma V(z)^p + L$ for some fixed constants $\gamma \in (0, 1)$ and $L \in [0, \infty)$ and all $z \in \mathcal{X}$. Define $\delta \triangleq \sup_{z \in \mathcal{X}} \left( \frac{\Delta_p(z)}{1 + \gamma V(z)^p} \right)^{1/p}$ and $\kappa \triangleq (1 + \max \left\{ \mathbb{E}[V(Y_0)^p], \frac{L}{1 - \gamma} \right\})^{1/p}$, where $\Delta_p(z) \triangleq \mathbb{E}[c(X, Y)^p | z]$ for $(X, Y) \sim \Gamma_\Delta(z)$. Then for all
\[ t \geq 0, \]
\[
\mathbb{E}[\text{CUB}^p_{p,t}]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^p \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1 - \rho^p)^{\frac{\delta_x}{1 - \rho^p}}.
\]

In the case \( p = 1 \), Proposition 3.9 recovers Theorem 3.1 of Rudolf and Schweizer [2018]. For such a result to be informative, we require functions \( V \) such that \( \delta_x \) is small. An application of these results to three simple examples based on MALA, ULA, and stochastic gradient Langevin dynamics (SGLD) [Welling and Teh, 2011] chains is given in Appendix C.

### 3.4 Comparison with alternative Wasserstein bounds

In this section, we compare our coupling-based Wasserstein bounds with alternatives.

**Empirical Wasserstein and Sinkhorn distances.** A common approach to estimating \( W_p(P, Q) \) is to draw independent samples from \( P \) and \( Q \) and then exactly compute the Wasserstein distance between the corresponding empirical distributions. This is precisely the empirical Wasserstein estimate that appeared in Figure 1. As our next proposition demonstrates, this empirical Wasserstein approach consistently upper bounds \( W_p(P, Q) \).

**Proposition 3.10** (Empirical Wasserstein distance bounds). For \( P \) and \( Q \) in \( \mathcal{P}_p(\mathcal{X}) \), let \( \hat{P}_N, \tilde{P}_N, \tilde{Q}_N \) and \( Q_N \) denote empirical distributions of the samples \( (X_1, \ldots, X_N) \) and \( (\tilde{X}_1, \ldots, \tilde{X}_N) \), \( (Y_1, \ldots, Y_N) \) and \( \tilde{Y}_N \) respectively, where \( X_i \sim \tilde{X}_i \sim P \) and \( Y_i \sim \tilde{Y}_i \sim Q \) for all \( i = 1, \ldots, N \). Suppose \( (X_1, \ldots, X_N) \) and \( (\tilde{X}_1, \ldots, \tilde{X}_N) \) are independent, and \( (Y_1, \ldots, Y_N) \) and \( (\tilde{Y}_1, \ldots, \tilde{Y}_N) \) are independent. Then

\[
0 \leq \mathbb{E}[W_p(\hat{P}_N, \tilde{Q}_N)^p]^{1/p} - W_p(P, Q) \leq \mathbb{E}[W_p(\hat{P}_N, \tilde{P}_N)^p]^{1/p} + \mathbb{E}[W_p(\tilde{Q}_N, Q_N)^p]^{1/p},
\]

where \( W_p(\hat{P}_N, \tilde{Q}_N) \rightleftharpoons W_p(P, Q) \) as \( N \to \infty \).

However, there are two downsides to the empirical Wasserstein approach. The first is statistical. The difference between \( \mathbb{E}[W_p(\hat{P}_N, \tilde{Q}_N)^p]^{1/p} \) and \( W_p(P, Q) \) can be quite large and decay very slowly in \( N \). For example, for some \( d \)-dimensional target distributions, \( \mathbb{E}[W_p(\hat{P}_N, \tilde{Q}_N)] \) converges to \( W_p(P, Q) \) at rate \( \Omega(N^{-1/d}) \) when \( d > 2p \) [e.g., Weed and Bach, 2019]. This can lead to the empirical Wasserstein distance giving loose upper bounds of \( W_p(P, Q) \) when the number of samples does not increase exponentially with dimension. The example in Figure 1b illustrates this curse of dimensionality, where the estimator \( \text{CUB}_p \) with CRN coupling gives tighter upper bounds of \( W_p(P, Q) \) than the empirical Wasserstein estimates.

The second downside is computational. Calculating \( W_p(\hat{P}_N, \tilde{Q}_N) \) amounts to solving an uncapacitated minimum cost flow problem with \( \mathcal{O}(N^3 \log N) \) computational cost [Orlin, 1988], a cost which is prohibitive for large sample sizes. A popular alternative is to compute an entropy-regularized Wasserstein distance instead using the Sinkhorn algorithm [Cuturi, 2013]. A larger value of the regularization parameter \( \lambda > 0 \) leads to faster computation but also introduces an additional bias that can compromise the accuracy of the bound. A smaller value of \( \lambda \) leads to more expensive \( \mathcal{O}(N^2/(\lambda \epsilon)) \) computation time for \( \epsilon \)-accurate solutions [e.g., Altschuler et al., 2017] and potential
instability of the Sinkhorn algorithm in practice. Simulations illustrating such computational issues are included in Appendix A.4.

In comparison, our coupling estimators run in time linear in the sample size $N$ and do not require the solution of any expensive optimization problems. On the other hand, empirical Wasserstein estimates will eventually converge to the true Wasserstein distance given sufficiently (perhaps exponentially) large sample sizes, so the empirical Wasserstein approach can lead to tighter bounds if one has a substantial computational budget.

The approach of Huggins et al. Huggins et al. [2020] derive upper bounds for the Wasserstein distance with $c(x,y) = \|x - y\|_2$ norm on $\mathbb{R}^d$ in terms of the Kullback–Leibler divergence or Rényi’s $\alpha$-divergences. To estimate their upper bounds of $W_p(P,Q)$ for $P$ and $Q$ in $\mathcal{P}_p(\mathbb{R}^d)$ and $P$ absolutely continuous with respect to $Q$, Huggins et al. propose importance sampling based estimates which require samples from $Q$, evaluations of the normalized density of $Q$, and evaluations of the unnormalized density of $P$. Figure 4 (Left) plots the performance of the 2-Wasserstein upper bounds of Huggins et al. for the example in Section 2.1. The dot-dashed line represents the mean of 20 independent Huggins et al. importance-sampling estimators, each with $2T = 3000$ samples from $Q$. The CUB$_2$ estimator plotted for comparison uses $I = 20$ independent CRN coupled chains with trajectory length $T = 1500$ and burnin $S = 500$. In this example, the Huggins et al. bounds are significantly looser than both our common random numbers coupling bound and the independent coupling upper bound. Furthermore, the importance sampling estimates of Huggins et al. exhibit an increasing variance in higher dimensions, as shown by the large grey error bands. One advantage of the Huggins et al. estimates over CUB$_2$ is that samples from $P$ are not required. On the other hand, unlike the Huggins et al. estimates, CUB$_p$ remains applicable even when the density of $Q$ cannot be evaluated. This case arises for many approximate MCMC algorithms such as ULA in Section 2.2, the stochastic gradient-based samplers in Section 4.1, and the matrix approximation-based sampler in Section 4.2.

The approach of Dobson et al. Dobson et al. [2021] apply couplings to assess the quality of numerical approximation of stochastic differential equations. Specifically, they focus on the 1-Wasserstein distance with the capped metric $c(x,y) = \min\{1, \|x - y\|_2\}$ on $\mathbb{R}^d$ and derive upper bounds in terms of the contraction constant of one of the marginal chains which are then estimated using couplings. The following proposition shows that $\mathbb{E}[\text{CUB}_1]$ with the same coupling provides a tighter upper bound than the proposal of Dobson et al. [2021].

**Proposition 3.11 (CUB lower bounds Dobson et al.).** Consider the 1-Wasserstein distance with metric $c(x,y) = \min\{1, \|x - y\|_2\}$ on $\mathbb{R}^d$. Then, for any coupling and sufficiently large burn-in, $\mathbb{E}[\text{CUB}_1]$ (3) lower bounds the estimated upper bound of Dobson et al. [2021].

Figure 4 (Right) plots the 1-Wasserstein upper bounds of Dobson et al. and CUB$_1$ for the example in Section 2.2 with the capped metric $c(x,y) = \min\{1, \|x - y\|_2\}$ on $\mathbb{R}^d$. We use $I = 100$ independent coupled chains with trajectory length $T = 3000$ and burnin $S = 1000$ to estimate both the upper bounds of Dobson et al. and CUB$_1$. The figure shows that, in line with Proposition 3.11, the upper bounds of Dobson et al. are looser than CUB$_1$. 
Figure 4: (Left) Upper bound estimates for 2-Wasserstein distances with \( c(x, y) = \|x - y\|_2 \) between \( P = \mathcal{N}(0, \Sigma) \) and \( Q = \mathcal{N}(0, I_d) \) on \( \mathbb{R}^d \), where \( [\Sigma]_{i,j} = 0.5^{i|\cdot-j|} \) for \( 1 \leq i, j \leq d \). The Huggins et al. [2020] bound is looser than CUB\(_2\) and has larger variance as the dimension grows. See Section 3.4 for more details. (Right) Upper bound estimates for the 1-Wasserstein distance with the capped metric \( c(x, y) = \min\{1, \|x - y\|_2\} \) between ULA and MALA chains targeting \( P = \mathcal{N}(0, \Sigma) \) on \( \mathbb{R}^d \). In line with Proposition 3.11, the CUB\(_1\) (3) estimate is tighter than the Dobson et al. [2021] bound employing the same common random numbers coupling. See Section 3.4 for more details.

4 Applications

We now illustrate the value of our methods for three practical applications. We focus on the 2-Wasserstein distance with \( c(x, y) = \|x - y\|_2 \) on \( \mathbb{R}^d \), which by (1) controls first and second order moments and captures geometric features induced by the Euclidean norm \( \|\cdot\|_2 \). In this case a tractably estimated lower bound on the Wasserstein distance is also available. For any \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \), let \( \mu_i \) and \( \nu_i \) denote the distributions of the \( i \)th component of \( \mu \) and \( \nu \) respectively. Let \( \mathcal{N}_\mu \) and \( \mathcal{N}_\nu \) denote Gaussian distributions on \( \mathbb{R}^d \) with the same means and covariance matrices as \( \mu \) and \( \nu \) respectively. Then,

\[
\max \left\{ \sum_{i=1}^{d} W_2(\mu_i, \nu_i)^2, \ W_2(\mathcal{N}_\mu, \mathcal{N}_\nu)^2 \right\} \leq W_2(\mu, \nu)^2. \tag{6}
\]

Here, \( \sum_{i=1}^{d} W_2(\mu_i, \nu_i)^2 \leq W_2(\mu, \nu)^2 \) follows from the coupling representation of \( W_2(\mu, \nu) \), and \( W_2(\mathcal{N}_\mu, \mathcal{N}_\nu) \leq W_2(\mu, \nu) \) is the lower bound of Gelbrich [1990, Theorem 2.1]. Each one-dimensional Wasserstein distance \( W_2(\mu_i, \nu_i) \) admits a convenient representation for estimation, given by \( \int_{-\infty}^{\infty} |F_{\mu_i}^{-1}(u) - F_{\nu_i}^{-1}(u)| \, du \) where \( F_{\mu_i} \) and \( F_{\nu_i}^{-1} \) are the inverse cumulative distribution functions of \( \mu_i \) and \( \nu_i \) respectively, while \( W_2(\mathcal{N}_\mu, \mathcal{N}_\nu) \) has the closed form \( \left( \|m_\mu - m_\nu\|_{\Sigma}^2 + \text{Trace}(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}) \right)^{1/2} \) in terms of the means \( m_\mu, m_\nu \) and covariances \( \Sigma_\mu, \Sigma_\nu \) of \( \mu \) and \( \nu \) [Peyré and Cuturi, 2019, Remark 2.23]. Since the true Wasserstein distances are unknown in our applications to follow, we will assess the tightness of our coupling-based upper bounds by estimating the lower bound (6). Details of all the datasets, algorithms and specific estimator parameters used in this section can be found in Appendix E.
4.1 Approximate MCMC and variational inference for tall data

Our first application concerns Bayesian inference for tall datasets [Bardenet et al., 2017], where the number of observations $n$ is large compared to the dimension $d$. In such settings, exact MCMC can be computationally expensive with $\Omega(n)$ cost per iteration. This computational bottleneck and the prevalence of tall datasets has catalyzed much interest in approximate MCMC and variational approximation based algorithms. Approximate MCMC algorithms include ULA and stochastic gradient MCMC (see Nemeth and Fearnhead [2021] for a review) such as Stochastic Gradient Langevin Dynamics (SGLD) [Welling and Teh, 2011]. Popular variational approximation methods include Laplace approximation [e.g., Tierney and Kadane, 1986] and variational Bayes (VB, see Blei et al. [2017] for a review).

In this section, we assess the quality of these sampling algorithms. We consider ULA, SGLD, Laplace approximation, and mean field VB applied to Bayesian logistic regression with a Gaussian prior for the Pima diabetes dataset [Smith et al., 1988] and the DS1 life sciences dataset [Komarek and Moore, 2003]. Figure 5 plots Wasserstein distance upper and lower bounds for different sampling algorithms. The upper bounds are given by $CUB_2(3)$ and the lower bounds are estimated using (6). We simulate the coupled chains $(X_t^{(i)}, Y_t^{(i)})_{t \geq 0}$ independently for each $i$, where each $(X_t^{(i)})_{t \geq 0}$ is a MALA chain targeting the posterior $P$ and each $(Y_t^{(i)})_{t \geq 0}$ is linked to an approximate MCMC or a variational procedure. In particular, we consider $(Y_t^{(i)})_{t \geq 0}$ to be an ULA chain, SGLD chains based on sub-sampling 10% and 50% of the observations, a MALA chain targeting $N(\mu_L, \Sigma_L)$ where $\mu_L \in \mathbb{R}^d$ and $\Sigma_L \in \mathbb{R}^{d \times d}$ are from a Laplace approximation of $P$, and a MALA chain targeting $N(\mu_{VB}, \Sigma_{VB})$ where $\mu_{VB} \in \mathbb{R}^d$ and $\Sigma_{VB} \in \text{Diag}(\mathbb{R}^{d \times d})$ are from a Gaussian mean field VB approximation of $P$. In each case, we use a CRN coupling between the marginal kernels of $(X_t^{(i)})_{t \geq 0}$ and $(Y_t^{(i)})_{t \geq 0}$. Appendix E.1 contains details about the datasets, algorithms and estimator parameters used.

Figure 5 shows that all the sampling algorithms considered have smaller asymptotic bias for the taller DS1 dataset (with a larger number of observations) compared to the Pima dataset. Also, Laplace approximation has the small asymptotic bias for both datasets. This promising performance of Laplace’s approximations can be linked to posterior concentration and accuracy of the corresponding Bernstein-von Mises approximation [Bardenet et al., 2017, Chopin and Ridgway, 2017]. Our bounds also show how the Metropolis–Hastings correction and stochastic gradients affect the quality of ULA and SGLD. Overall, this application illustrates the effectiveness of our proposed quality measures for comparing approximate inference algorithms in the tall data setting.

4.2 Approximate MCMC for high-dimensional linear regression

We now consider high-dimensional Bayesian linear regression, where the dimension $d$ is larger than the number of observations $n$. The likelihood is given by $L(\beta; y, X, \sigma^2) = N(y; X\beta; \sigma^2 I_n)$, where $N(\cdot; X\beta, \sigma^2 I_n)$ denotes the probability density function of a Gaussian distribution on $\mathbb{R}^n$ with mean $X\beta$ and covariance matrix $\sigma^2 I_n$, $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times d}$ is the design matrix, $\beta \in \mathbb{R}^d$ is an unknown signal vector, and $\sigma^2 > 0$ is the unknown noise variance. We consider a class
of global-local mixture priors, given by

$$\xi^{-1/2} \sim C_+(0, 1), \quad \eta_j^{-1/2} \overset{i.i.d.}{\sim} t_+(\nu), \quad \sigma^{-2} \sim \text{Gamma}\left(\frac{a_0}{2}, \frac{b_0}{2}\right), \quad \beta_j|\eta, \xi, \sigma^2 \sim N\left(0, \frac{\sigma^2}{\xi \eta_j}\right)$$

where $C_+(0, 1)$ is the half-Cauchy distribution on $[0, \infty)$ and $t_+(\nu)$ is the half-t distribution on $[0, \infty)$ with $\nu$ degrees of freedom. When $\nu = 1$, this corresponds to the popular Horseshoe prior [Carvalho et al., 2010]. This setting differs considerably from the log-concave tall data example of Section 4.1, as now the posterior distribution is multi-modal, has polynomial tails, and has infinite density about the origin [Biswas et al., 2021]. Johndrow et al. [2020] have developed exact and approximate Gibbs samplers for the Horseshoe prior in this setting, which involves an approximation parameter $\epsilon \geq 0$. Biswas et al. [2021] extended the sampler of Johndrow et al. to all $\nu \geq 1$ and showed that using larger values of $\nu$ could improve mixing times in high dimensions.

In this section, we use couplings to assess the quality of such approximate MCMC algorithms. Following Biswas et al., we consider $\nu = 2$ applied to a genome-wide association study (GWAS) dataset [Bühlmann et al., 2014] and a synthetic dataset. We use a CRN coupling with the marginal chains corresponding to the exact and the approximate MCMC kernel. Appendix E.2 contains details about the datasets, algorithms and estimator parameters used.

Figure 6 plots upper and lower bounds of the 2-Wasserstein distance, illustrating how asymptotic bias of the approximate Gibbs sampler varies with the approximation parameter $\epsilon \geq 0$. The upper bounds are given by our estimator $\text{CUB}_2(3)$ and the lower bounds are estimated using (6). For developers of such high-dimensional approximate MCMC samplers, these bounds provide an empirical assessment of the trade-off between improved quality and higher computational cost for different datasets and posteriors.

### 4.3 Approximate MCMC for high-dimensional logistic regression

We now consider high-dimensional Bayesian logistic regression with spike and slab priors. Spike and slab priors have been commonly used for Bayesian variable selection [George and McCulloch, 1993,
Figure 6: Upper and lower bounds on the Euclidean $W_2$ bias of an approximate MCMC Gibbs sampler for high-dimensional Bayesian regression with half-t(2) prior. We consider a bacteria GWAS dataset with $n = 71$ observations and $d = 4088$ covariates and a synthetic dataset with $n = 500$ and $d = 50000$. See Section 4.2 for more details.

Figure 7: Upper and lower bounds on the Euclidean $W_2$ bias of the Skinny Gibbs sampler [Narisetty et al., 2019] for Bayesian logistic regression with the spike and slab prior. We consider a malware dataset with $n = 373$ observations and $d = 503$ covariates and a lymph node GWAS dataset with $n = 148$ and $d = 4514$. See Section 4.3 for more details.

Narisetty and He, 2014. Narisetty et al. [2019] have recently developed an approximate MCMC algorithm which does not have any tuning parameters, called Skinny Gibbs, to sample from the posteriors in this setting. In this section, we assess the quality of the Skinny Gibbs algorithm applied to a malware dataset [Dua and Graff, 2017] and a lymph node GWAS [Narisetty et al., 2019] dataset. We use a CRN coupling between one marginal chain corresponding to the exact MCMC kernel and another corresponding to the Skinny Gibbs kernel. Appendix E.3 contains further details about spike and slab priors, and the datasets, algorithms and estimator parameters used.

Figure 7 plots upper and lower bounds for the Euclidean 2-Wasserstein distance between the limiting distributions of the exact and Skinny Gibbs chains for $\beta$. The upper bounds are given by our estimator CUB$_2$ (3) and the lower bounds are estimated using (6). These bounds suggest that the asymptotic bias of the Skinny Gibbs algorithm is greater for the malware dataset compared to the lymph node GWAS dataset. For researchers developing approximate MCMC samplers, these bounds provide an empirical assessment of asymptotic bias for different datasets and posteriors under the spike and slab prior.
5 Discussion

We have introduced new estimators to assess the quality of approximate inference procedures. The estimators consistently bound the Wassertein distance between the limiting distribution of the approximation and the original target distribution of interest. The proposed estimators can be applied to approximate MCMC and certain variational inference methods in practical settings, including Bayesian regression in 50000 dimensions.

The following questions arise from our work.

**Alternative coupling algorithms.** We have chosen common random numbers (CRN) coupling as a practical default for our experiments due to its broad applicability, but a growing inventory of alternative coupling strategies is available [Heng and Jacob, 2019, Lee et al., 2020, Xu et al., 2021, Wang et al., 2021, Biswas et al., 2021], and, as evidenced in Section 3.2, alternative couplings tailored to the problem can yield tighter upper bounds. An important open question is how to best identify or construct a coupling better than the CRN default for a given problem at hand.

**Wasserstein distance upper bounds which avoid sampling from the asymptotically unbiased Markov chain.** All our upper bounds require sampling from the Markov chain \((X_t, Y_t)_{t \geq 0}\) where \((X_t)_{t \geq 0}\) and \((Y_t)_{t \geq 0}\) correspond to an asymptotically unbiased and biased chain marginally. Sampling \((X_t)_{t \geq 0}\) can be computationally expensive and thus limits the application of our quality measures to researchers developing approximate sampling methods rather than practitioners. This raises the question: can one construct a Markov chain \((Y'_t, Y_t)_{t \geq 0}\) such that (i) \((Y'_t)_{t \geq 0}\) and \((Y_t)_{t \geq 0}\) are identically distributed according to the same asymptotically biased chain marginally and (ii) 

\[
E[c(X_t, Y'_t)^p] = E[c(X_t, Y_t)^p] \leq E[c(Y'_t, Y_t)^p] \text{ for all } t \geq 0,
\]

where \((X_t)_{t \geq 0}\) is an asymptotically unbiased chain? Then we could sample from the computationally less expensive chain \((Y'_t, Y_t)_{t \geq 0}\) to obtain an upper bound of 

\[
E[c(X_t, Y_t)^p]^{1/p} \geq 2E[c(X_t, Y'_t)^p]^{1/p} = 2E[c(X_t, Y_t)^p]^{1/p}.
\]

We hope to investigate such coupling constructions in follow-up work.

**Upper bounds for total variation distance.** The 1-Wasserstein distance with metric \(c(x, y) = \mathbb{I}\{x \neq y\}\) gives the popular total variation distance (TV), which always takes values in \([0, 1]\) and is invariant to reparameterization. To obtain upper bounds of TV strictly less than 1 using our estimators, we require couplings which allow exact meetings between the two marginal chains. Our initial attempts at using maximal couplings [Johnson, 1998, Jacob et al., 2020, Wang et al., 2021] have not been effective in high dimensions and suggest a need for further methodological work.

**Acknowledgments.** We thank Juan Shen for sharing the Lymph Node dataset, and Pierre E. Jacob, Xiao-Li Meng, the participants of the International Conference on Monte Carlo Methods and Applications, and the participants of the BayesComp workshop on “Measuring the quality of MCMC output” for helpful feedback. NB was supported by the NSF grant DMS-1844695, a GSAS Merit Fellowship, and a Two Sigma Fellowship Award.
References

J. Altschuler, J. Weed, and P. Rigollet. Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. NeurIPS, page 1961–1971, 2017. ISBN 9781510860964.

R. Bardenet, A. Doucet, and C. Holmes. On Markov Chain Monte Carlo Methods for Tall Data. J. Mach. Learn. Res., 18(1):1515–1557, 2017. ISSN 1532-4435.

A. Bhattacharya, A. Chakraborty, and B. K. Mallick. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. Biometrika, 103(4):985–991, 2016. ISSN 0006-3444. doi: 10.1093/biomet/asw042. URL https://doi.org/10.1093/biomet/asw042.

N. Biswas, P. E. Jacob, and P. Vanetti. Estimating convergence of Markov chains with L-lag couplings. NeurIPS, pages 7389–7399, 2019.

N. Biswas, A. Bhattacharya, P. E. Jacob, and J. E. Johndrow. Coupling-based convergence assessment of some gibbs samplers for high-dimensional bayesian regression with shrinkage priors. arXiv:2012.04798v2, 2021.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc., 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

P. Bühlmann, M. Kalisch, and L. Meier. High-Dimensional Statistics with a View Toward Applications in Biology. Annu. Rev. Stat. Appl., 1(1):255–278, 2014. doi: 10.1146/annurev-statistics-022513-115545. URL https://doi.org/10.1146/annurev-statistics-022513-115545.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. Biometrika, 97(2):465–480, 2010. ISSN 00063444. URL http://www.jstor.org/stable/25734098.

N. Chopin and J. Ridgway. Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation. Statist. Sci., 32(1):64 – 87, 2017. doi: 10.1214/16-STS581. URL https://doi.org/10.1214/16-STS581.

K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. ICML, page 2606–2615, 2016.
M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. NeurIPS, pages 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572c7f59eb343d-Paper.pdf. 3, 13

M. Dobson, Y. Li, and J. Zhai. Using Coupling Methods to Estimate Sample Quality of Stochastic Differential Equations. SIAM-ASA J. Uncertain., 9(1):135–162, 2021. doi: 10.1137/20M1312009. URL https://doi.org/10.1137/20M1312009. 3, 14, 15, 35

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. 18, 45

A. Durmus and E. Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. Bernoulli, 25(4A):2854–2882, 2019. doi: 10.3150/18-BEJ1073. URL https://doi.org/10.3150/18-BEJ1073. 5, 6, 25, 36

A. Durmus, A. Eberle, A. Enfroy, A. Guillin, and P. Monmarché. Discrete sticky couplings of functional autoregressive processes. arXiv:2104.06771, 2021. 37

R. Durrett. Probability: Theory and Examples. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034. 29

A. Eberle. Error bounds for metropolis–hastings algorithms applied to perturbations of gaussian measures in high dimensions. Ann. Appl. Probab., 24(1):337–377, 2014. doi: 10.1214/13-AAP926. URL https://doi.org/10.1214/13-AAP926. 11, 37

A. Eberle. Reflection couplings and contraction rates for diffusions. Probab. Theory Relat. Fields, 166(3):851–886, 2016. doi: 10.1007/s00440-015-0673-1. URL https://doi.org/10.1007/s00440-015-0673-1. 11

M. Gelbrich. On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. Math. Nachr., 147(1):185–203, 1990. doi: https://doi.org/10.1002/mana.19901470121. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mana.19901470121. 15

E. I. George and R. E. McCulloch. Variable Selection via Gibbs Sampling. J. Am. Stat. Assoc., 88(423):881–889, 1993. doi: 10.1080/01621459.1993.10476353. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353. 17, 45

P. W. Glynn and C.-H. Rhee. Exact estimation for Markov chain equilibrium expectations. J. Appl. Probab., 51(A):377–389, 2014. 3, 4

J. Gorham and L. Mackey. Measuring Sample Quality with Stein’s Method. NeurIPS, pages 226–234, 2015. URL https://proceedings.neurips.cc/paper/2015/file/698d51a19d8a121ce581499d7b701668-Paper.pdf. 2

J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. ICML, pages 1292–1301, 2017. URL http://proceedings.mlr.press/v70/gorham17a.html. 2
J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928, 2019. doi: 10.1214/19-AAP1467. URL https://doi.org/10.1214/19-AAP1467. 2, 11

J. Gorham, A. Raj, and L. Mackey. Stochastic Stein Discrepancies. NeurIPS, pages 17931–17942, 2020. URL https://proceedings.neurips.cc/paper/2020/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf. 2

C. Hans, A. Dobra, and M. West. Shotgun Stochastic Search for “Large p” Regression. *J. Am. Stat. Assoc.*, 102(478):507–516, 2007. doi: 10.1198/016214507000000121. URL https://doi.org/10.1198/016214507000000121. 45

J. Heng and P. E. Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302, 2019. ISSN 0006-3444. doi: 10.1093/biomet/asy074. URL https://doi.org/10.1093/biomet/asy074. 3, 4, 19

J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. Validated Variational Inference via Practical Posterior Error Bounds. AISTATS, pages 1792–1802. PMLR, 2020. URL https://proceedings.mlr.press/v108/huggins20a.html. 3, 14, 15

J. H. Huggins and L. Mackey. Random Feature Stein Discrepancies. NeurIPS, page 1903–1913, 2018. 2

J. H. Huggins, T. Campbell, M. Kasprzak, and T. Broderick. Scalable Gaussian Process Inference with Finite-data Mean and Variance Guarantees. AISTATS, pages 796–805, 2019. URL http://proceedings.mlr.press/v89/huggins19a.html. 37

H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.*, 33(2):730 – 773, 2005. doi: 10.1214/009053604000001147. URL https://doi.org/10.1214/009053604000001147. 45

P. E. Jacob, J. O’Leary, and Y. F. Atchadé. Unbiased Markov chain Monte Carlo methods with couplings (with Discussion). *J. R. Stat. Soc. Ser. B Methodol.*, 82(3):543–600, 2020. doi: 10.1111/rssb.12336. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12336. 3, 4, 10, 19, 26

J. Johndrow, P. Orenstein, and A. Bhattacharya. Scalable Approximate MCMC Algorithms for the Horseshoe Prior. *J. Mach. Learn. Res.*, 21(73):1–61, 2020. URL http://jmlr.org/papers/v21/19-536.html. 1, 17, 44, 45

J. E. Johndrow and J. C. Mattingly. Error bounds for Approximations of Markov chains used in Bayesian Sampling. arXiv:1711.05382, 2018. 11

V. E. Johnson. A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *J. Am. Stat. Assoc.*, 93(441):238–248, 1998. 2, 4, 8, 19

P. Komarek and A. Moore. Fast robust logistic regression for large sparse datasets with binary outputs. AISTATS, pages 163–170, 2003. URL http://komarix.org/ac/ds/. 16, 44

22
A. Lee, S. S. Singh, and M. Vihola. Coupled conditional backward sampling particle filter. *Ann. Statist.*, 48(5):3066–3089, 2020. 19

F. Liang, Q. Song, and K. Yu. Bayesian Subset Modeling for High-Dimensional Generalized Linear Models. *J. Am. Stat. Assoc.*, 108(502):589–606, 2013. doi: 10.1080/01621459.2012.761942. URL https://doi.org/10.1080/01621459.2012.761942. 45

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Spinger New York, 2008. 1

Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. ICML, pages 276–284, 2016. URL https://proceedings.mlr.press/v48/liub16.html. 2

L. Middleton, G. Deligiannidis, A. Doucet, and P. E. Jacob. Unbiased Smoothing using Particle Independent Metropolis-Hastings. AISTATS, pages 2378–2387, 2019. URL http://proceedings.mlr.press/v89/middleton19a.html. 3, 4

L. Middleton, G. Deligiannidis, A. Doucet, and P. E. Jacob. Unbiased Markov chain Monte Carlo for intractable target distributions. *Electron. J. Statist.*, 14(2):2842–2891, 2020. doi: 10.1214/20-EJS1727. URL https://doi.org/10.1214/20-EJS1727. 26

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42(2):789 – 817, 2014. doi: 10.1214/14-AOS1207. URL https://doi.org/10.1214/14-AOS1207. 1, 18, 45

N. N. Narisetty, J. Shen, and X. He. Skinny Gibbs: A Consistent and Scalable Gibbs Sampler for Model Selection. *J. Am. Stat. Assoc.*, 114(527):1205–1217, 2019. doi: 10.1080/01621459.2018.1482754. URL https://doi.org/10.1080/01621459.2018.1482754. 1, 18, 45, 49

C. Nemeth and P. Fearnhead. Stochastic Gradient Markov Chain Monte Carlo. *J. Am. Stat. Assoc.*, 116(533):433–450, 2021. doi: 10.1080/01621459.2020.1847120. URL https://doi.org/10.1080/01621459.2020.1847120. 5, 16

J. Orlin. A Faster Strongly Polynomial Minimum Cost Flow Algorithm. STOC, page 377–387, 1988. ISBN 0897912640. doi: 10.1145/62212.62249. URL https://doi.org/10.1145/62212.62249. 3, 5, 13

G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073. 5, 15

N. S. Pillai and A. Smith. Ergodicity of Approximate MCMC Chains with Applications to Large Data Sets. *arXiv:1405.0182*, 2015. 11

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. 44

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Spinger New York, 2013. 1
G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. 
*J. R. Stat. Soc. Ser. B Methodol.*, 60(1):255–268, 1998. doi: https://doi.org/10.1111/1467-9868.00123. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00123.

D. Rudolf and N. Schweizer. Perturbation theory for markov chains via wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018. doi: 10.3150/17-BEJ938. URL https://doi.org/10.3150/17-BEJ938.

J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261 – 265, 1988.

L. Tierney and J. B. Kadane. Accurate Approximations for Posterior Moments and Marginal Densities. *J. Am. Stat. Assoc.*, 81(393):82–86, 1986. doi: 10.1080/01621459.1986.10478240. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240.

T. van Erven and P. Harremos. Rényi Divergence and Kullback–Leibler Divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

D. Vats and C. Knudson. Revisiting the Gelman–Rubin Diagnostic. *Statist. Sci.*, 36(4):518 – 529, 2021. doi: 10.1214/20-STS812. URL https://doi.org/10.1214/20-STS812.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-Normalization, Folding, and Localization: An Improved ˆR for Assessing Convergence of MCMC (with Discussion). *Bayesian Anal.*, 16(2):667 – 718, 2021. doi: 10.1214/20-BA1221. URL https://doi.org/10.1214/20-BA1221.

C. Villani. *Optimal transport – Old and new*. Springer, 2008. doi: 10.1007/978-3-540-71050-9.

G. Wang, J. O’Leary, and P. Jacob. Maximal Couplings of the Metropolis-Hastings Algorithm. AISTATS, pages 1225–1233, 2021. URL https://proceedings.mlr.press/v130/wang21d.html.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019. doi: 10.3150/18-BEJ1065. URL https://doi.org/10.3150/18-BEJ1065.

M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. ICML, page 681–688, 2011. ISBN 9781450306195.

K. Xu, T. E. Fjelde, C. Sutton, and H. Ge. Couplings for Multinomial Hamiltonian Monte Carlo. AISTATS, pages 3646–3654, 2021.
A Additional figures and discussion

A.1 Calculation of empirical Wasserstein bounds in Figure 1.

In this section we note how the empirical Wasserstein upper bounds and error bands in Figure 1 are generated. Our upper bounds are based on Proposition 3.10, which gives

$$W_p(P, Q)^p \leq E[W_p(\hat{P}_T, \hat{Q}_T)^p]$$

where $P$ and $Q$ are distributions on the metric space $(\mathcal{X}, c)$ with finite moments of order $p$, and $\hat{P}_T$ and $\hat{Q}_T$ denote empirical distributions of the samples $(X_1, ..., X_T)$ and $(Y_1, ..., Y_T)$ where $X_i \sim P$ and $Y_i \sim Q$ for all $i = 1, ..., T$. For $p = 2$ and $P \neq Q$, the dot-dashed lines in Figure 1 plots the estimate

$$\left(\frac{1}{T} \sum_{i=1}^{T} W_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})^2\right)^{1/2}$$

of this upper bound, where $\hat{P}_T^{(i)}$ and $\hat{Q}_T^{(i)}$ are empirical distribution of $P$ and $Q$ respectively based on $T$ samples. For each $i = 1, ..., I$, such empirical distributions $\hat{P}_T^{(i)}$ and $\hat{Q}_T^{(i)}$ are generated independently and then $W_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})$ is calculated by solving a linear program. The error bands plot $\hat{\sigma} / \sqrt{T}$ where $\hat{\sigma}^2$ is the empirical variance of $\left(W_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})^2\right)_{i=1}^{I}$.

A.2 Section 2.2 calculations.

As $X_0 \sim \mathcal{N}(0, \Sigma) = P$ and kernel $K_1$ is $P$ invariant, $X_t \sim P$ for all $t \geq 0$. For the ULA chain $(Y_t)_{t \geq 0}$, we have

$$Y_t = (I_d - (\sigma_Q^2/2)\Sigma^{-1})Y_{t-1} + \sigma_Q Z_t = BY_{t-1} + \sigma_Q Z_t$$

for all $t \geq 0$, where $Y_0 = 0$, and $Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ and $B = (I_d - (\sigma_Q^2/2)\Sigma^{-1})$. By induction,

$$Y_t = B^t Y_0 + \sigma_Q \left( B^{t-1} Z_1 + B^{t-2} Z_2 + ... + Z_t \right) = \sigma_Q \sum_{j=0}^{t-1} B^j Z_{t-j} \sim \mathcal{N}(0, \sigma_Q^2 \sum_{j=0}^{t-1} B^{2j}) =: Q_t$$

as required. Finally, note that for $\sigma_Q = 0.5d^{-1/6}$ sufficiently small such that $\|B\|_{\text{op}} < 1$ (where $\|\cdot\|_{\text{op}}$ is the matrix operator norm), $\lim_{t \to \infty} \sum_{j=0}^{t-1} B^{2j} = (I_d - B^2)^{-1}$. This gives $Q_t \overset{t \to \infty}{\to} \mathcal{N}(0, \sigma_Q^2 (I_d - B^2)^{-1}) =: Q$.

ULA asymptotic bias upper bound calculation for Figure 2. We recall a result of Durmus and Moulines [2019] on the asymptotic bias of ULA.

**Proposition A.1.** [Durmus and Moulines, 2019, Corollary 9] Consider an ULA Markov chain targeting the distribution $\pi$ on $\mathbb{R}^d$ with un-normalized density $\exp(-U(x))$. For $\|\cdot\|_2$ the Euclidean norm on $\mathbb{R}^d$, assume:
1. $U$ is continuously differentiable and Lipschitz: there exists some $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, 
\[ \|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|_2. \]

2. $U$ is $m$-strongly convex for some $m > 0$: there exists some $m > 0$ such that for all $x, y \in \mathbb{R}^d$, 
\[ U(x) \leq U(y) + \langle \nabla U(x), y - x \rangle + \left(\frac{m}{2}\right)\|x - y\|_2^2 \]

3. $U$ is three times continuously differentiable and there exists some $\bar{L} > 0$ such that for all $x, y \in \mathbb{R}^d$, 
\[ \|\nabla^2 U(x) - \nabla^2 U(y)\|_2 \leq \bar{L}\|x - y\|_2. \]

Let the step size $\sigma$ of the Markov chain be sufficiently small such that $\gamma \triangleq \sigma^2/2 < 1/(m + L)$. Then the ULA Markov chain converges to some distribution $\pi_\gamma$, and
\[ \mathcal{W}_2(\pi, \pi_\gamma)^2 \leq 2\kappa^{-1}\gamma^2d\left(2L^2 + \gamma L^4\left(\frac{\gamma}{6} + \frac{1}{m}\right) + \kappa^{-1}\left(\frac{4\bar{L}^2}{3} + \gamma L^4 + \frac{4L^4}{3m}\right)\right) \tag{7} \]
where $\kappa = 2mL/(m + L)$.

The dotted line in Figure 2 is plotted by applying (7) for $\pi = \mathcal{N}(0, \Sigma)$, where $L = \lambda_{\text{min}}(\Sigma)^{-1}$, $m = \lambda_{\text{max}}(\Sigma)^{-1}$ and $\bar{L} = 0$. Here $\lambda_{\text{max}}(\Sigma)$ and $\lambda_{\text{min}}(\Sigma)$ are the largest and smallest eigenvalue of $\Sigma$ respectively.

### A.3 Non-asymptotic upper bounds using L-Lag coupling

In this section, we discuss how to avoid burn-in removal and instead directly correct our bound for non-stationarity using the recent $L$-lag coupling approach of Biswas et al. [2019] in the case of the 1-Wasserstein distance.

We first informally outline the approach of Biswas et al. [2019]. Consider a Markov chain on $(\mathcal{X}, c)$ with transition kernel $K_1$, marginal distributions $(P_t)_{t \geq 0}$ and a unique stationary distribution $P$. Consider a joint kernel $K_1$ on $\mathcal{X} \times \mathcal{X}$ such that $K_1((x, y), (\cdot, \cdot)) = K_1(x, \cdot)$ and $K_1((x, y), (\cdot, \cdot)) = K_1(y, \cdot)$ for all $x, y \in \mathcal{X}$. Then the $L$-lag coupling chain $(\bar{X}_t)_{t \geq 0}$ is generated by sampling $X_0$ and $X_0$ independently from a common initial distribution $P_0$, sampling $X_t | X_{t-1} \sim K_1(X_{t-1}, \cdot)$ for $t = 1, \ldots, L$, and generating $(\bar{X}_t)_{t \geq 0}$ exactly meet such that the random meeting time $\tau \triangleq \inf\{t > L : \bar{X}_t = X_t\}$ is almost surely finite and (ii) the chains remain faithful after meeting such that $\bar{X}_t = X_t$ for all $t \geq \tau$. Suppose the coupled chain $(\bar{X}_t, X_t)_{t \geq L}$ satisfies Assumptions A.2, A.3 and A.4 [Biswas et al., 2019, Jacob et al., 2020] (see Middleton et al. [2020] for the use of polynomially-tailed meeting times).

**Assumption A.2** (Marginal convergence and uniformly bounded moments). Marginal distributions $(P_t)_{t \geq 0}$ converge to $P$ in 1-Wasserstein distance, and for all $t \geq L$, $E[c(\bar{X}_t, X_t)^{2+\eta}] \leq D$ for some constants $\eta > 0$ and $D < \infty$.  

26
Assumption A.3 (Sub-exponentially tailed meeting times). The meeting times $\tau \triangleq \inf \{ t > L : X_t = \tilde{X}_{t-L} \}$ satisfies $P(\frac{\tau - L}{L} > t) \leq C\delta^t$ for some constants $C < \infty$ and $\delta \in (0,1)$ and all $t \geq 0$.

Assumption A.4 (Faithfulness after meeting). $X_t = \tilde{X}_{t-L}$ for all $t \geq \tau$.

Under Assumptions A.2, A.3 and A.4, Biswas et al. [2019] obtain

$$W_1(P_t, P) \leq \sum_{j=1}^{\infty} W_1(P_{t+jL-L}, P_{t+jL}) \tag{8}$$

$$\leq \sum_{j=1}^{\infty} E[c(\tilde{X}_{t+jL-L}, X_{t+jL})] \tag{9}$$

$$= E\left[ \sum_{j=1}^{\infty} c(\tilde{X}_{t+jL-L}, X_{t+jL}) \right] \tag{10}$$

$$= E\left[ \sum_{j=1}^{\left[\frac{(\tau - L-t)}{L}\right]} c(\tilde{X}_{t+jL-L}, X_{t+jL}) \right], \tag{11}$$

where (8) follows from the triangle inequality using Assumption A.2, (9) follows from the coupling representation of the Wasserstein distance, and (10) follows from interchanging the summation and expectation using the dominated convergence theorem under Assumptions A.2 and A.3, and (11) follows as $c(\tilde{X}_{t+jL-L}, X_{t+jL}) = 0$ for all $j > \left[\frac{(\tau - L-t)}{L}\right]$ under Assumption A.4. Note that $\tau$ has finite expectation under Assumption A.3, which means the upper bound in (11) can be estimated in finite time. We can estimate this upper bound by simulating multiple $L$-lag coupled chains $(\tilde{X}_{t-L}, X_t)_{t \geq t \geq L}$ independently and using the empirical average

$$\frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{\left[\frac{(\tau - L-t)}{L}\right]} c(\tilde{X}_{t+jL-L}^{(i)}, X_{t+jL}^{(i)})$$

where $I \geq 1$ is the number of independent coupled chains.

The following Proposition employs this upper bound alongside CUB$_1$ (3) to obtain a non-asymptotic upper bound on $W_1(P, Q)$.

Proposition A.5 (Non-asymptotic upper bound). For any lag $L \geq 1$, consider the coupled chain $(\tilde{X}_{t-L}, X_t, Y_t, \tilde{Y}_{t-L})_{t \geq L}$ such that $(\tilde{X}_{t-L}, X_t)_{t \geq L}$ is an $L$-lag coupling chain for the kernel $K_1$, $(\tilde{Y}_{t-L}, Y_t)_{t \geq L}$ is an $L$-lag coupling chain for the kernel $K_2$, and $(X_t, Y_t)_{t \geq L}$ is a coupled chain sampled using Algorithm 1. Under Assumption 3.4 with $p = 1$ and Assumptions A.2, A.3 and A.4 for the coupled chains $(\tilde{X}_{t-L}, X_t)_{t \geq L}$ and $(\tilde{Y}_{t-L}, Y_t)_{t \geq L}$,

$$W_1(P, Q) \leq E[CUB_{1,t}] + E\left[ \sum_{j=1}^{\left[\frac{(\tau_p - L-t)}{L}\right]} c(\tilde{X}_{t+(j-1)L}, X_{t+jL}) \right] + E\left[ \sum_{j=1}^{\left[\frac{(\tau_Q - L-t)}{L}\right]} c(\tilde{Y}_{t+(j-1)L}, Y_{t+jL}) \right] \tag{12}$$

for all $t \geq 0$, where $\tau_p \triangleq \inf \{ t > L : \tilde{X}_{t-L} = X_t \}$ and $\tau_Q \triangleq \inf \{ t > L : \tilde{Y}_{t-L} = Y_t \}$. 

27
A.4 Sinkhorn algorithm simulations for Section 3.4

In this section we consider the impact of the regularization parameter of the Sinkhorn algorithm. Figure 8a of this section plots the Wasserstein distance upper bounds for the stylized example in Section 2.1. In particular, we consider the 2-Wasserstein distance with Euclidean norm on $\mathbb{R}^d$, and the distributions $P = \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.5|i-j|$ for $1 \leq i, j \leq d$ and $Q = \mathcal{N}(0, I_d)$ in the case of dimension $d = 10$.

The CUB$_2$ (3) estimate (black line) in Figure 8a is based a CRN coupling of marginal MALA kernels, with $I = 10$ independent coupling chains and trajectories of length $T = 500$ with a burn-in of $S = 100$ for each chain. The true Wasserstein (black dot-dashed line) distance and the upper bound from independent coupling (black dotted line) are analytically tractable, as given in Section 2.1. For different values of the entropic regularization parameter $\lambda$, the grey solid line plots the induced distance of the optimal matching obtained from the Sinkhorn algorithm. For each $\lambda$, we implement the Sinkhorn algorithm on empirical distributions with $IT = 5000$ sample points from $P$ and $Q$. Figure 8a shows that we require a small entropic regularization parameter $\lambda$ to obtain informative upper bounds using the Sinkhorn algorithm. On the other hand, Figure 8b shows that the runtime for the Sinkhorn algorithm increases dramatically for smaller values of $\lambda$. This example illustrates that the Sinkhorn algorithm has expensive runtime precisely for the smaller values of $\lambda$ that give tighter upper bounds to the Wasserstein distance. In comparison, the CUB$_2$ (3) estimate does not require solving any expensive optimization problem.
B Proofs

B.1 Consistency proofs

Technical Results. We first collect some technical results for reference.

**Lemma B.1.** Let \((a_j)_{j \geq 0}\) be a real sequence with \(a_j \xrightarrow{j \to \infty} 0\), and let \(\rho \in (0, 1)\). Then \(\sum_{j=1}^{t} \rho^{t-j} a_j \xrightarrow{t \to \infty} 0\).

**Proof of B.1.** As \(a_j \xrightarrow{j \to \infty} 0\), the sequence \((a_j)_{j \geq 0}\) is bounded by some \(M \in (0, \infty)\). Also for all \(\epsilon > 0\), there exists some \(j_0 \geq 1\) such that \(|a_j| < \epsilon\) for all \(j \geq j_0\). For all \(t > j_0\), this gives

\[
\left| \sum_{j=1}^{t} \rho^{t-j} a_j \right| \leq \sum_{j=1}^{j_0} \rho^{t-j} |a_j| + \sum_{j=j_0+1}^{t} \rho^{t-j} |a_j| \leq M \rho^{j_0} \frac{1-\rho^{t-j_0}}{1-\rho} + \epsilon \frac{1-\rho^{t-j_0}}{1-\rho}.
\]

Taking \(t \to \infty\), we obtain \(\lim_{t \to \infty} \left| \sum_{j=1}^{t} \rho^{t-j} a_j \right| \leq \epsilon/(1-\rho)\), where \(\epsilon/(1-\rho)\) can be made arbitrarily small. \(\blacksquare\)

**Lemma B.2.** (Gluing lemma) [e.g. Villani, 2008, Chapter 1] Let \(\mu_i\) be probability measures on the Polish measurable spaces \((X_i, \mathcal{B}(X_i))\) for \(i = 1, \ldots, 3\). Let \(X_1, X_2, Y_2, Y_3\) be random variables such that \((X_1, X_2)\) is a coupling of \((\mu_1, \mu_2)\) and \((Y_2, Y_3)\) is a coupling of \((\mu_2, \mu_3)\). Then, there exists random variables \(Z_1, Z_2, Z_3\) such that \((Z_1, Z_2)\) has the same law as \((X_1, X_2)\) and \((Z_2, Z_3)\) has the same law as \((Y_2, Y_3)\).

**Proof of B.2.** See Villani [2008, Chapter 1]. \(\blacksquare\)

**Lemma B.3.** Let \((\xi_i)_{i \geq 0}\) be independent and identically distributed non-negative random variables with \(\mathbb{E}[\xi_1] < \infty\), and let \(S_n = \sum_{i=1}^{n} \xi_i\). Then as \(n \to \infty\), \(S_n/n \xrightarrow{a.s.} \mathbb{E}[\xi_1]\) and for any \(p \geq 1\), \((S_n/n)^{1/p} \xrightarrow{a.s.} \mathbb{E}[\xi_1]^{1/p}\).

**Proof of B.3.** As \(n\) tends to infinity, \(S_n/n \xrightarrow{a.s.} \mathbb{E}[\xi_1]\) follows from the proof of the Strong law of large numbers using backwards martingales (see, e.g., Durrett [2019, Theorem 4.7.1 and Example 4.7.4]). \((S_n/n)^{1/p} \xrightarrow{a.s.} \mathbb{E}[\xi_1]^{1/p}\) follows from \(S_n/n \xrightarrow{a.s.} \mathbb{E}[\xi_1]\) by continuous mapping theorem on \([0, \infty)\). Finally, for \(p \geq 1\),

\[
\mathbb{E}[(S_n/n)^{1/p} - (\mathbb{E}[\xi_1])^{1/p}]^p \leq \mathbb{E}[(S_n/n) - \mathbb{E}[\xi_1])^{1/p}]^p \leq \mathbb{E}[(S_n/n) - \mathbb{E}[\xi_1])^{1/p}]^p \xrightarrow{n \to \infty} 0
\]

where the first inequality follows as \(|a^{1/p} - b^{1/p}| \leq |a - b|^{1/p}\) for all \(a, b \geq 0\) and \(p \geq 1\), the second inequality follows from Jensen’s inequality and the limit follows as \(S_n/n \xrightarrow{L^1} \mathbb{E}[\xi_1]\). Therefore, \((S_n/n)^{1/p} \xrightarrow{L^1} \mathbb{E}[\xi_1]^{1/p}\). \(\blacksquare\)

**Proof of Proposition 3.1:** Consistency of instantaneous CUB. Note that \(W_p(P_t, Q_t)\) is well-defined and \(\mathbb{E}[c(X_t, Y_t)^p]\) is finite as distributions \(P\) and \(Q\) have finite moments of order \(p\). We obtain

\[
W_p(P_t, Q_t) \leq \mathbb{E}[c(X_t, Y_t)^p] = \mathbb{E}[\text{CUB}_{p,t}^p],
\]

29
where the inequality follows from the coupling representation of Wasserstein distance, and the equality follows from the definition of \( \text{CUB}_{p,t} \). As \( \mathbb{E}[\text{CUB}_{p,t}] < \infty \), by Lemma B.3, \( \text{CUB}_{p,t} \overset{a.s.}{\to}_{L^1} \mathbb{E}[\text{CUB}_{p,t}] \) as \( I \to \infty \).

\[ \text{Proof of Corollary 3.2: Consistency of CUB for time-averaged marginals.} \] We first show that

\[
\mathcal{W}_p\left(\frac{1}{T-S} \sum_{t=S+1}^{T} P_t, \frac{1}{T-S} \sum_{t=S+1}^{T} Q_t\right)^p \leq \frac{1}{T-S} \sum_{t=S+1}^{T} \mathcal{W}_p(P_t, Q_t)^p.
\]

Let \( \gamma_t \) denote the \( p \)-Wasserstein optimal coupling between distributions \( P_t \) and \( Q_t \) for \( t = S+1, \ldots, T \). Sample the coupling \((X^*, Y^*)\) such that \((X^*, Y^*)|I = t \sim \gamma_t \) for \( I \sim \text{Uniform}\{S+1, \ldots, T\} \). Then \( X^* \sim \frac{1}{T-S} \sum_{t=S+1}^{T} P_t \) and \( Y^* \sim \frac{1}{T-S} \sum_{t=S+1}^{T} Q_t \), marginally, and

\[
\mathcal{W}_p\left(\frac{1}{T-S} \sum_{t=S+1}^{T} P_t, \frac{1}{T-S} \sum_{t=S+1}^{T} Q_t\right)^p \leq \mathbb{E}[c(X^*, Y^*)^p] \text{ by the coupling representation of } \mathcal{W}_p
\]

\[
= \frac{1}{T-S} \sum_{t=S+1}^{T} \mathbb{E}[c(X^*, Y^*)^p|I = t]
\]

\[
= \frac{1}{T-S} \sum_{t=S+1}^{T} \mathcal{W}_p(P_t, Q_t)^p.
\]

Now by Proposition 3.1 and definition (3),

\[
\frac{1}{T-S} \sum_{t=S+1}^{T} \mathcal{W}_p(P_t, Q_t)^p \leq \mathbb{E}\left[\frac{1}{T-S} \sum_{t=S+1}^{T} \text{CUB}_{p,t}^p\right] = \mathbb{E}[\text{CUB}_p^p].
\]

As \( \mathbb{E}[\text{CUB}_p^p] < \infty \), by Lemma B.3 \( \text{CUB}_p \overset{a.s.}{\to}_{L^1} \mathbb{E}[\text{CUB}_p^p] \) as \( I \to \infty \).

\[ \text{Proof of Corollary 3.3: Consistency of CUB with stationary initialization.} \] Note that \( \mathcal{W}_p(P, Q) \) is well-defined and \( \sum_{t=S+1}^{T} \mathbb{E}[c(X_t, Y_t)^p]/(T-S) \) is finite as distributions \( P_t \) and \( Q_t \) have finite moments of order \( p \). We obtain,

\[
\mathcal{W}_p(P, Q)^p = \frac{1}{T-S} \sum_{t=S+1}^{T} \mathcal{W}_p(P_t, Q_t)^p \leq \frac{1}{T-S} \sum_{t=S+1}^{T} \mathbb{E}[c(X_t, Y_t)^p] = \mathbb{E}[\text{CUB}_p^p].
\]

where the first equality follows as \( P_t = P \) and \( Q_t = Q \) for all \( t \geq 0 \), the inequality follows Proposition 3.1, and the last equality follows from the definition of \( \text{CUB}_p \). As \( \mathbb{E}[\text{CUB}_p^p] < \infty \), by Lemma B.3 \( \text{CUB}_p \overset{a.s.}{\to}_{L^1} \mathbb{E}[\text{CUB}_p^p] \) as \( I \to \infty \).

\[ \text{Proof of Proposition 3.5: Consistency when chain marginals converge.} \] Let \((P_t)_{t \geq 0}\) and \((Q_t)_{t \geq 0}\) denote the marginal distributions of Markov chains \((X_t)_{t \geq 0}\) and \((Y_t)_{t \geq 0}\) respectively. By Assumption
3.4, distributions \((P_t)_{t \geq 0}, (Q_t)_{t \geq 0}\), \(P\) and \(Q\) all have finite moments of order \(p\). Then for all \(t \geq 1\),

\[
\begin{align*}
W_p(P,Q) &\leq W_p(P,P_t) + W_p(P_t,Q_t) + W_p(Q_t,Q) \\
&\leq W_p(P,P_t) + \mathbb{E}[c(X_t,Y_t)^p]^{1/p} + W_p(Q_t,Q),
\end{align*}
\]

where (13) follows by the triangle inequality as \(W_p\) is a metric on the space of measure on \(\mathcal{X}\) with finite moments of order \(p\), and (14) follows from the coupling representation of \(W_p\). By Assumption 3.4, \(\lim_{t \to \infty} W_p(P,P_t) = 0\) and \(\lim_{t \to \infty} W_p(Q_t,Q) = 0\). Taking the limit infimum in (14) and raising to the \(p^{th}\) exponent gives \(W_p(P,Q)^p \leq \lim \inf_{t \to \infty} \mathbb{E}[c(X_t,Y_t)^p]\). Therefore for all \(\epsilon > 0\), there exists \(S \geq 1\) such that for all \(t \geq S\), \(W_p(P,Q)^p \leq \epsilon + \mathbb{E}[c(X_t,Y_t)^p]\), and

\[
W_p(P,Q)^p \leq \epsilon + \frac{1}{T-S} \sum_{t=S+1}^{T} \mathbb{E}[c(X_t,Y_t)^p] = \epsilon + \mathbb{E}[\text{CUB}_p^p]
\]

for all \(T \geq S\). As \(\mathbb{E}[\text{CUB}_p^p]^p < \infty\), by Lemma B.3 \(\text{CUB}_p^p \overset{a.s.-L^1}{\rightarrow} \mathbb{E}[\text{CUB}_p^p]\) as \(I \to \infty\).

**Proof of Proposition A.5: Non-asymptotic upper bound.** By the triangle inequality,

\[
W_1(P,Q) \leq W_1(P_t,Q_t) + W_1(P_t,P) + W_1(P_t,P).
\]

By Proposition 3.1, \(W_1(P_t,Q_t) \leq \mathbb{E}[\text{CUB}_{1,t}]\). Under assumptions A.2, A.3 and A.4, by Biswas et al. [2019, Theorem 2.5]

\[
W_1(P_t,P) \leq \mathbb{E}\left[\sum_{j=1}^{[(\tau_p-L-t)/L]} c(\tilde{X}_{t+(j-1)L}, X_t+jL)\right] \quad \text{and}
\]

\[
W_1(Q_t,Q) \leq \mathbb{E}\left[\sum_{j=1}^{[(\tau_Q-L-t)/L]} c(\tilde{Y}_{t+(j-1)L}, Y_t+jL)\right].
\]

Equation (12) now directly follows. As the meeting times \(\tau_p\) and \(\tau_Q\) have sub-exponential tails by Assumption A.3, the \(L\)-lag upper bounds can be estimated in finite time.

**B.2 Wasserstein upper bound proofs**

**Proof of Theorem 3.7: CUB upper bound.** Under the coupled kernel \(\tilde{K}\) from Algorithm 2, for each \(t \geq 1\) we have the coupling \((X_t, Z_t, Y_t)\) where \((X_t, Z_t) | X_{t-1}, Y_{t-1} \sim \Gamma_1(X_{t-1}, Y_{t-1})\) and \((Z_t, Y_t) | X_{t-1}, Y_{t-1} \sim \Gamma_{\Delta}(Y_{t-1})\). This gives

\[
\mathbb{E}[c(X_t,Y_t)^p]^{1/p} = \mathbb{E}[\mathbb{E}[c(X_t,Y_t)^p|X_{t-1}, Y_{t-1}]^{1/p}
\]

\[
\leq \mathbb{E}[\mathbb{E}[(c(X_t,Z_t) + c(Z_t,Y_t))^p|X_{t-1}, Y_{t-1}]^{1/p}
\]

\[
\leq \mathbb{E}[\mathbb{E}[c(X_t,Z_t)^p|X_{t-1}, Y_{t-1}]^{1/p} + \mathbb{E}[\mathbb{E}[c(Z_t,Y_t)^p|X_{t-1}, Y_{t-1}]^{1/p}
\]

\[
\leq \rho \mathbb{E}[c(X_{t-1}, Y_{t-1})^p]^{1/p} + \mathbb{E}[\Delta_p(Y_{t-1})]^{1/p}
\]

31
where (15) follows as \( c \) is a metric, (16) follows by Minowski’s inequality, and (17) follows by Assumption 3.6 with \( \Delta_p(z) \triangleq \mathbb{E}[c(X, Y)^p|z] \) for \((X, Y) \sim \Gamma_\Delta(z)\). By induction, (17) implies

\[
\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^t \rho^{t-i} \mathbb{E}[\Delta_p(Y_{i-1})]^{1/p}.
\]

\( \square \)

**Proof of Corollary 3.8:** CUB upper bound under marginal convergence. Denote \( a \triangleq \mathbb{E}[\Delta_p(Y^*)]^{1/p} \) for \( Y^* \sim Q \) and \( a_k \triangleq \mathbb{E}[\Delta_p(Y_k)]^{1/p} \) for \( k \geq 0 \). Then \( a_k \overset{k \to \infty}{\to} a \), because \( Q_t \) converges in \( p \)-Wasserstein distance to \( Q \) as \( t \to \infty \). By Lemma B.1, this implies

\[
\sum_{i=1}^t \rho^{t-i} a_{i-1} \overset{t \to \infty}{\to} \sum_{i=1}^t \rho^{t-i} a = \frac{1 - \rho^t}{1 - \rho} a.
\]

Therefore, for all \( \epsilon > 0 \) there exists \( S \geq 1 \) such that for all \( t \geq S \), \( \sum_{i=1}^t \rho^{t-i} |a_i - a| < \epsilon \). By Theorem 3.7,

\[
\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^t \rho^{t-i} a_{i-1} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^t \rho^{t-i} a + \sum_{i=1}^t \rho^{t-i} |a_{i-1} - a| = \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \frac{1 - \rho^t}{1 - \rho} a + \epsilon.
\]

\( \square \)

**Proof of Proposition 3.9:** CUB upper bound weighted by a Lyapunov function. As \( V \) is a \( p^{th} \)-order Lyapunov function of \( K_2 \), by induction

\[
\mathbb{E}[V(Y_i)^p] \leq \gamma^i \mathbb{E}[V(Y_0)^p] + (1 - \gamma^i) \frac{L}{1 - \gamma} \text{ for all } i \geq 0.
\]

(18)

for all \( i \geq 0 \). Therefore,

\[
\mathbb{E}[\Delta_p(Y_i)] \leq \delta \mathbb{E}[1 + V(Y_{i-1})^p] \leq \delta^p \left( 1 + \gamma^{i-1} \mathbb{E}[V(Y_0)^p] + (1 - \gamma^{i-1}) \frac{L}{1 - \gamma} \right) \leq \delta^p \kappa^p
\]

for all \( i \geq 1 \), where the first inequality follows from the definition of \( \delta \), second inequality from (18),
and the second inequality from the definition of \( \kappa \). By Theorem 3.7, we obtain

\[
\mathbb{E}[c(X_t,Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0,Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} \mathbb{E} \left[ \Delta_p(Y_{i-1}) \right]^{1/p}
\]

\[
\leq \rho^t \mathbb{E}[c(X_0,Y_0)^p]^{1/p} + \delta \kappa \sum_{i=1}^{t} \rho^{t-i}
\]

\[
= \rho^t \mathbb{E}[c(X_0,Y_0)^p]^{1/p} + (1 - \rho^t) \frac{\delta \kappa}{1 - \rho}.
\]

\( \Box \)

### B.3 Wasserstein distances of empirical distributions proofs

To prove Proposition 3.10, we first record a technical result.

**Lemma B.4.** Suppose \( S \) and \( T \) are distributions on the metric space \((X,c)\) with finite moments of order \( p \). Given \( U_i \sim S \) for \( i = 1, \ldots, n \), let \( \hat{S}_N \) denote the empirical distribution of \((U_1, \ldots, U_N)\). Then,

\[
W_p(S,T)^p \leq \mathbb{E}[W_p(\hat{S}_N, T)^p].
\]

**Proof.** Our proof follows a coupling construction. Define random variables \( V \sim T \) and \( U_i \sim S \) for \( i = 1, \ldots, n \) such that \( V \) and \((U_1, \ldots, U_N)\) are independent. Then \( V|U_1, \ldots, U_N \sim V \sim T \) by independence. Let \( \hat{S}_N \) denote the empirical distribution of \((U_1, \ldots, U_N)\). Define a random variable \( U \) such that \( U|U_1, \ldots, U_N \sim \hat{S}_N \) and \((U, V)|U_1, \ldots, U_N \) is a Wasserstein optimal coupling of \( \hat{S}_N \) and \( T \). Note that unconditionally \( V \sim T \) and \( U \sim S \) as \( U_i \sim S \) for all \( i = 1, \ldots, n \). Therefore \((U, V)\) is a coupling of \( S \) and \( T \). We obtain,

\[
W_p(S,T)^p \leq \mathbb{E}[c(U,V)^p] \text{ by the coupling representation of Wasserstein distance}
\]

\[
= \mathbb{E}[\mathbb{E}[c(U,V)^p|U_1, \ldots, U_N]]
\]

\[
= \mathbb{E}[W_p(\hat{S}_N, T)^p].
\]

\( \Box \)

**Proof of Proposition 3.10: Empirical Wasserstein distance bounds.**

**Upper bound.** Let \( \hat{P}_N \) and \( \hat{Q}_N \) denote the empirical distributions of the samples \((X_1, \ldots, X_N)\) and \((Y_1, \ldots, Y_N)\) respectively, where \( X_i \sim P \), \( Y_i \sim Q \) for all \( i = 1, \ldots, n \), and \((X_1, \ldots, X_N)\) and \((Y_1, \ldots, Y_N)\) are independent. By Lemma B.4 with \( S = P \), \( U_i = X_i \) and \( T = Q \),

\[
W_p(P,Q)^p \leq \mathbb{E}[W_p(\hat{P}_N, Q)^p].
\]

As \((X_1, \ldots, X_N)\) and \((Y_1, \ldots, Y_N)\) are independent, \( Y_i|(X_1, \ldots, X_N) \sim Y_i \sim Q \) for all \( i = 1, \ldots, N \). We can therefore apply Lemma B.4 conditional on \((X_1, \ldots, X_N)\) now with \( S = Q \), \( U_i = Y_i \) and
\( T = \hat{P}_N \) to obtain

\[
W_p(\hat{P}_N, Q)^p \leq \mathbb{E}[W_p(\hat{P}_N, \hat{Q}_N)^p | X_1, \ldots, X_N]
\]

almost surely for all \( X_1, \ldots, X_N \). Overall, this gives

\[
W_p(P, Q)^p \leq \mathbb{E}[W_p(\hat{P}_N, Q)^p | X_1, \ldots, X_N] = \mathbb{E}[W_p(\hat{P}_N, \hat{Q}_N)^p]
\]

as required.

**Lower bound.** Let \( \hat{P}_N \) and \( \hat{Q}_N \) denote empirical distributions of the samples \( (X_1, \ldots, X_N) \) and \( (Y_1, \ldots, Y_N) \) respectively, where \( X_i \sim P \), \( Y_i \sim Q \) for all \( i = 1, \ldots, n \). Given \( (X_1, \ldots, X_N) \) and \( (Y_1, \ldots, Y_N) \), by the triangle inequality we obtain

\[
W_p(\hat{P}_N, \hat{Q}_N) \leq W_p(\hat{P}_N, P) + W_p(P, Q) + W_p(Q, \hat{Q}_N).
\]

By Minowski’s inequality, this gives

\[
\mathbb{E}[W_p(\hat{P}_N, \hat{Q}_N)^p]^{1/p} \leq \mathbb{E}\left[\left(W_p(\hat{P}_N, P) + W_p(P, Q) + W_p(Q, \hat{Q}_N)\right)^p\right]^{1/p}
\]

\[
\leq \mathbb{E}[W_p(\hat{P}_N, P)^p]^{1/p} + \mathbb{E}[W_p(P, Q)^p]^{1/p} + \mathbb{E}[W_p(Q, \hat{Q}_N)^p]^{1/p}
\]

\[
= \mathbb{E}[W_p(\hat{P}_N, P)^p]^{1/p} + W_p(P, Q) + \mathbb{E}[W_p(Q, \hat{Q}_N)^p]^{1/p}
\]

(19)

Let \( \hat{P}_N \) denote empirical distributions of the samples \( (\tilde{X}_1, \ldots, \tilde{X}_N) \), where \( \tilde{X}_i \sim P \) for all \( i = 1, \ldots, n \) and \( (\tilde{X}_1, \ldots, \tilde{X}_N) \) and \( (X_1, \ldots, X_N) \) are independent. Independence implies \( \tilde{X}_i | (X_1, \ldots, X_N) \sim \tilde{X}_i \sim P \) for all \( i = 1, \ldots, n \). We can therefore apply Lemma B.4 conditional on \( (X_1, \ldots, X_N) \), with \( S = P, T = \hat{P}_N \) and \( \tilde{X}_i = U_i \) to obtain

\[
W_p(\hat{P}_N, P)^p \leq \mathbb{E}[W_p(\hat{P}_N, \hat{P}_N)^p | X_1, \ldots, X_N].
\]

Similarly,

\[
W_p(Q, \hat{Q}_N)^p \leq \mathbb{E}[W_p(\hat{Q}_N, \hat{Q}_N)^p | Y_1, \ldots, Y_N]
\]

where \( \hat{Q}_N \) denotes empirical distributions of the samples \( (\tilde{Y}_1, \ldots, \tilde{Y}_N) \), where \( \tilde{Y}_i \sim Q \) for all \( i = 1, \ldots, n \) and \( (\tilde{Y}_1, \ldots, \tilde{Y}_N) \) and \( (Y_1, \ldots, Y_N) \) are independent. By (19), we obtain

\[
\mathbb{E}[W_p(\hat{P}_N, \hat{Q}_N)^p]^{1/p} \leq \mathbb{E}[W_p(\hat{P}_N, P)^p]^{1/p} + W_p(P, Q) + \mathbb{E}[W_p(Q, \hat{Q}_N)^p]^{1/p}
\]

\[
= \mathbb{E}[W_p(\hat{P}_N, P)^p | X_1, \ldots, X_N]^{1/p} + W_p(P, Q) + \mathbb{E}[W_p(Q, \hat{Q}_N)^p | Y_1, \ldots, Y_N]^{1/p}
\]

\[
\leq \mathbb{E}[W_p(\hat{P}_N, P)^p | X_1, \ldots, X_N]^{1/p} + W_p(P, Q) + \mathbb{E}[W_p(Q, \hat{Q}_N)^p | Y_1, \ldots, Y_N]^{1/p}
\]

\[
= \mathbb{E}[W_p(\hat{P}_N, P)^p]^{1/p} + W_p(P, Q) + \mathbb{E}[W_p(\hat{Q}_N, \hat{Q}_N)^p]^{1/p}
\]

as required.
Consistency. By triangle inequality,

$$|W_p(\hat{P}_N, Q_N) - W_p(P, Q)| \leq W_p(\hat{P}_N, P) + W_p(Q, \hat{Q}_N).$$

Note that $P$, $Q$, $(\hat{P}_N)_{N \geq 0}$ and $(\hat{Q}_N)_{N \geq 0}$ all have finite moments of order $p$, and that $\hat{P}_N \Rightarrow P$ and $\hat{Q}_N \Rightarrow Q$ almost surely by the Glivenko–Cantelli theorem, where the empirical distribution moments of order $p$ also converge weakly. By completeness of the $p$-Wasserstein distance on the space of probability measures with finite moments of order $p$ [Villani, 2008, Theorem 6.9], $W_p(\hat{P}_N, P) \xrightarrow{a.s.} 0$ and $W_p(Q, \hat{Q}_N) \xrightarrow{a.s.} 0$ as $N \to \infty$.

B.4 Proofs for comparison with the approach of Dobson et al.

To prove Proposition 3.11, we first outline the setup of Dobson et al. [2021]. Consider a continuous time diffusion with a unique stationary distribution $P$ on $\mathbb{R}^d$. Let $K_1$ and $K_2$ denote the Markov chain transition kernels corresponding to a discretization of this diffusion with and without an accept-reject bias correction step respectively. For example, $K_1$ and $K_2$ can be the (single or multiple step) transition kernels of an MALA and an ULA Markov chain respectively. Suppose the marginal Markov chains with kernels $K_1$ and $K_2$ converge in distribution to the unique invariant distributions $P$ and $Q$ respectively.

For some small $\epsilon > 0$, suppose there exists a compact subset $\Omega$ of $\mathbb{R}^d$ such that $P(\Omega^c) < \epsilon$ and $Q(\Omega^c) < \epsilon$. For the capped metric $c(x, y) = \min\{1, \|x - y\|_2\}$ on $\mathbb{R}^d$, suppose there exists a Markovian coupling $\Gamma_1$ of the kernel $K_1$ such that for some constant $\alpha_\Omega \in (0, 1)$ and all $X_t, X_t' \in \Omega$, $E[c(X_{t+1}, X_{t+1}')|X_t, X_t'] \leq \alpha_\Omega c(X_t, X_t')$ for $(X_{t+1}, X_{t+1}')|(X_t, X_t') \sim \Gamma_2(X_t, X_t')$. Under such assumptions, Dobson et al. [2021] show

$$W_1(P, Q) \leq \frac{E[E[c(X_1, Y_1)|Y^*]] + 2\epsilon}{1 - \alpha_\Omega}$$

(20)

where $Y^* \sim Q$ and $(X_1, Y_1)|Y^* \sim \Gamma_\Delta(Y^*)$ for some fixed coupling $\Gamma_\Delta(Y^*)$ such that $X_1|Y^* \sim K_1(Y^*, \cdot)$ and $Y_1|Y^* \sim K_2(Y^*, \cdot)$ marginally. Dobson et al. [2021] then estimate the quantities $E[E[c(X_1, Y_1)|X^*]]$ and $\alpha_\Omega$ separately using couplings to obtain a final upper bound estimate.

Given this setup, we can show that our upper bound estimator CUB1 (3) constructed using such couplings $\Gamma_1$ and $\Gamma_\Delta$ has a smaller expected value than the upper bound of (20).

Proof of Proposition 3.11: CUB lower bounds Dobson et al. We proceed as in the proofs of Theorem 3.7 and Corollary 3.8. Consider the coupling based estimator in (3) for the 1-Wasserstein distance with metric $c(x, y) = \min\{1, \|x - y\|_2\}$ on $\mathbb{R}^d$. Under the coupled kernel $\bar{K}$ from Algorithm 2, for each $t \geq 1$ we have the coupling $(X_t, Z_t, Y_t)$ where $(X_t, Z_t)|X_{t-1}, Y_{t-1} \sim \Gamma_1(X_{t-1}, Y_{t-1})$ and

$$35$$
$(Z_t, Y_t)|X_{t-1}, Y_{t-1} \sim \Gamma_\Delta(Y_{t-1})$. This gives
\[
\mathbb{E}[c(X_t, Y_t)] \leq \mathbb{E}[c(X_t, Z_t)] + \mathbb{E}[c(Z_t, Y_t)] 
\]
\[
= \mathbb{E}[c(X_t, Z_t)|\{X_{i-1}\in\Omega, Y_{i-1}\in\Omega\}] + \mathbb{E}[c(X_t, Z_t)|\{X_{i-1}\in\Omega, Y_{i-1}\in\Omega\}] + \mathbb{E}[c(Z_t, Y_t)]
\]
\[
\leq \mathbb{P}(\{X_{t-1}\in\Omega\} \cup \{Y_{t-1}\in\Omega\}) + \mathbb{E}[c(X_t, Z_t)|\{X_{i-1}\in\Omega, Y_{i-1}\in\Omega\}] + \mathbb{E}[c(Z_t, Y_t)]
\]
\[
\leq \mathbb{P}(X_{t-1} \in \Omega^c) + \mathbb{P}(Y_{t-1} \in \Omega^c) + \alpha_\Omega \mathbb{E}[c(Z_t, Y_t)];
\]
where (21) follows by the triangle inequality, (22) follows as the union bound and the definition of $\alpha_\Omega$. Denote $\Delta(z) \triangleq \mathbb{E}[c(X, Y)p|z]$ for $(X, Y) \sim \Gamma_\Delta(z)$, such that $\mathbb{E}[c(Z_t, Y_t)] = \mathbb{E}[\mathbb{E}[c(Z_t, Y_t)|Y_{t-1}]] = \mathbb{E}[\Delta(Y_{t-1})]$. Then by induction, (23) implies
\[
\mathbb{E}[c(X_t, Y_t)] \leq \alpha_\Omega^t \mathbb{E}[c(X_0, Y_0)] + \sum_{i=1}^{t} \alpha_\Omega^{t-i} \left( \mathbb{P}(X_{i-1} \in \Omega^c) + \mathbb{P}(Y_{i-1} \in \Omega^c) + \mathbb{E}[\Delta(Y_{i-1})] \right).
\]
As $X_{t-1}$ and $Y_{t-1}$ converge to $P$ and $Q$ respectively in distribution, $\mathbb{P}(X_{t-1} \in \Omega^c) \xrightarrow{t \to \infty} P(\Omega^c) < \epsilon$, $\mathbb{P}(Y_{t-1} \in \Omega^c) \xrightarrow{t \to \infty} Q(\Omega^c) < \epsilon$ and $\mathbb{E}[\Delta(Y_{t-1})] \xrightarrow{t \to \infty} \mathbb{E}[\Delta(Y^*)]$ for $Y^* \sim Q$. Following the argument in Corollary 3.8 we obtain that for all $\epsilon' > 0$, there exists some $S \geq 1$ such that for all $t \geq S$,
\[
\mathbb{E}[c(X_t, Y_t)] \leq \alpha_\Omega^t \mathbb{E}[c(X_0, Y_0)] + \sum_{i=1}^{t} \alpha_\Omega^{t-i} \left( \mathbb{E}[\Delta(Y^*)] + 2\epsilon \right) + \epsilon'.
\]
Therefore as $\alpha_\Omega \in (0, 1), \lim_{t \to \infty} \mathbb{E}[c(X_t, Y_t)] \leq \frac{\mathbb{E}[\Delta(Y^*)]+2\epsilon}{1-\alpha_\Omega}$ where $\Delta(Y^*) = \mathbb{E}[c(X_1, Y_1)|Y^*]$ and $Y^* \sim Q$ from (20) as required.

## C Example applications of theoretical results

In this section we consider the theoretical results of Section 3.3 applied to three simple examples, working with the metric $c(x, y) = \|x - y\|_2$.

**MALA and ULA.** Consider a MALA chain and an ULA chain with a common step size $\sigma$ both targeting a distribution $P$. Assume the negative log density of $P$ is gradient Lipschitz and strongly convex. In this setting, let $(X_t, Y_t)_{t \geq 0}$ be a CRN coupling of ULA and MALA simulated using Algorithm 1, such that the Markov chains $(X_t)_{t \geq 0}$ and $(Y_t)_{t \geq 0}$ marginally correspond to ULA and MALA respectively. For $\sigma$ sufficiently small, the marginal ULA chain $(X_t)_{t \geq 0}$ converges to some distribution $P_\sigma$ and satisfies Assumption 3.6 for $p = 2$ under a CRN coupling [Durmus and Moulines, 2019, Proposition 3], giving a contraction rate $\rho$ such that $1 - \rho = C\sigma^2/2$ for some constant $C$ which depends on the gradient Lipschitz constant and convexity of the negative log density of $P$ rather
than depending explicitly on the dimension of the state space. By Corollary 3.8,
\[
W_2(P_\sigma, P) \leq \liminf_{t \to \infty} \mathbb{E}[\text{CUB}_2(P_{t, 1})]^{1/2} \leq \frac{\mathbb{E}[\|Y - Y'\|^2(1 - \alpha_\sigma(Y, Y'))]}{C \sigma^2/2},
\]
where \( Y \sim P \) is the limiting distribution of the MALA chain, \( Y'|Y \sim \mathcal{N}(Y + \frac{\sigma^2}{2} \nabla \log P(Y), \sigma^2 I_d) \) corresponds to the Euler–Maruyama discretization based proposal, and \( \alpha_\sigma(Y, Y') \in [0, 1] \) is the Metropolis–Hastings acceptance probability. As the step size \( \sigma \) tends to zero, the upper bound in (24) require further analysis of the MALA acceptance probabilities [Bou-Rabee and Hairer, 2012, Eberle, 2014] and could degenerate. Recently, discrete sticky couplings [Durmus et al., 2021] have been developed for perturbed functional autoregressive processes, which produce stable upper bounds on total variation and the Wasserstein distance in such limiting regimes.

**ULA and ULA.** We can similarly consider two ULA chains with a common step size \( \sigma \) targeting different distributions \( P \) and \( Q \). As above, assume both \( \log P \) and \( \log Q \) are gradient Lipschitz and strongly convex. In this setting, let \((X_t, Y_t)_{t \geq 0}\) be a CRN coupling of two ULA chains simulated using Algorithm 1, such that the Markov chains \((X_t)_{t \geq 0}\) and \((Y_t)_{t \geq 0}\) marginally correspond to ULA targeting distributions \( P \) and \( Q \) respectively. For \( \sigma \) sufficiently small, the marginal chains \((X_t)_{t \geq 0}\) and \((Y_t)_{t \geq 0}\) converge to some distributions \( P_\sigma \) and \( Q_\sigma \) respectively. Both marginal chains also satisfy Assumption 3.6 for \( p = 2 \) under a CRN coupling, with contraction rates \( \rho_P \) and \( \rho_Q \) such that \( 1 - \rho_P = C_P \sigma^2/2 \) and \( 1 - \rho_Q = C_Q \sigma^2/2 \) respectively for some constants \( C_P \) and \( C_Q \) that do not explicitly depend on the dimension. By Corollary 3.8, this gives
\[
W_2(P_\sigma, Q_\sigma) \leq \liminf_{t \to \infty} \mathbb{E}[\text{CUB}_2(P_{t, 1})]^{1/2} \leq \frac{\mathbb{E}[\|\nabla \log P(Y_\sigma) - \nabla \log Q(Y_\sigma)\|^2]}{C_P}
\]
where \( Y \sim Q_\sigma \). By symmetry, we can obtain a similar bound in terms of some random variable \( X \sim P_\sigma \) and \( Q_\sigma \). As \( \sigma \) approaches zero, the numerator in (25) approaches the square root of the Fisher divergence between distributions \( Q \) and \( P \), given by \( F(Q, P) \triangleq \mathbb{E}[\|\nabla \log P(Y) - \nabla \log Q(Y)\|^2] \) for \( Y \sim Q \). Such link between the Fisher divergence and the Wasserstein distance has been noted previously by considering continuous-time Langevin diffusions (e.g., Huggins et al. [2019]). Finally, note that the upper bound in (25) does not explicitly depend on dimension, highlighting that estimators based on our coupled chains may give upper bounds that remain informative in high dimensions.

**ULA and SGLD.** Consider an ULA chain and a Stochastic gradient Langevin dynamics (SGLD) [Welling and Teh, 2011] chain with a common step size \( \sigma \) and both targeting a distribution \( P \). The SGLD chain is based on unbiased estimates of the gradient of the log density of \( P \), such that \( \nabla \log P_{\text{SGLD}}(z) = \nabla \log P(z) + \epsilon_{\text{SGLD}}(z) \) for all \( z \in \mathcal{X} \), where \( \epsilon_{\text{SGLD}}(z) \) is mean zero error. We assume this error is bounded such that \( \sigma^2 \triangleq \sup_{z \in \mathcal{X}} \epsilon_{\text{SGLD}}(z)/(1 + V(z)^2) < \infty \), for some \( 2^{nd} \)-order Lyapunov function \( V \) as in Proposition 3.9 and that the negative log density of \( P \) is gradient Lipschitz and strongly convex. In this setting, let \((X_t, Y_t)_{t \geq 0}\) be a CRN coupling of ULA and SGLD simulated using Algorithm 1, such that the Markov chains \((X_t)_{t \geq 0}\) and \((Y_t)_{t \geq 0}\) marginally correspond to
ULA and SGLD with marginal distributions \( (P_t^{(ULA)})_{t \geq 0} \) and \( (P_t^{(SGLD)})_{t \geq 0} \) respectively. For \( \sigma \) sufficiently small, the marginal ULA chain \( (X_t)_{t \geq 0} \) satisfies Assumption 3.6 for \( p = 2 \) under a CRN coupling, giving a contraction rate \( \rho \) such that \( 1 - \rho = C\sigma^2/2 \) for constants \( C \) that does not explicitly depend on the dimension. Then by Proposition 3.9,\[
\limsup_{t \to \infty} W_2(P_t^{(ULA)}, P_t^{(SGLD)}) \leq \liminf_{t \to \infty} E \left[ \text{CUB}_2(P_t^{(ULA)}, Q_t^{(ULA)}) \right]^{1/2} \leq \frac{\delta \kappa}{C}. \tag{26}
\]
Note that the upper bound in (26) does not explicitly depend on dimension, and approaches zero as \( \delta \) approaches zero. This shows that estimators based on our coupled chains give upper bounds which may remain informative in high dimensions and are tight with respect to the error from the stochastic gradients. This example also highlights the stability of our upper bounds even when one of marginal chains (SGLD) may not converge to a limiting distribution.

## D Multi-step couplings

In this section, we consider coupling algorithms for multi-step kernels and investigate their theoretical properties.

### D.1 Coupling algorithms for multi-step kernels

Consider the \( L \)-step Markov chains \((X_{Lt})_{t \geq 0}\) and \((Y_{Lt})_{t \geq 0}\) for \( L \geq 1 \), corresponding to marginal multi-step Markov kernels \( K^L_P \) and \( K^L_Q \) respectively. Following (2) and Section 3.2, we now construct a kernel \( \bar{K}_{L-step} \) on the joint space \( \mathcal{X} \times \mathcal{X} \) such that for all \( x,y \in \mathcal{X} \) and all \( A \in B(\mathcal{X}) \),\[
\bar{K}_{L-step}((x,y),(A,\mathcal{X})) = K^L_P(x,A) \text{ and } \bar{K}_{L-step}((x,y),(\mathcal{X},A)) = K^L_Q(y,A). \tag{27}
\]

Given coupled kernels \( \Gamma_1 \) and \( \Gamma_\Delta \), Figure 9 illustrates how to sample from the joint kernel \( \bar{K}_{L-step} \). By construction, this gives the marginal distributions \( X_s | X_0, Y_0 \sim K^s_P(X_0, \cdot) \) and \( Y_s | X_0, Y_0 \sim K^s_Q(Y_0, \cdot) \) for all \( s = 1, \ldots, L \), such that Equation (27) is satisfied. Algorithm 3 samples from this coupled kernel \( \bar{K}_{L-step} \). It characterizes the dependency between \( X_{Lt} \) and \( Y_{Lt} \) such that\[
X_{Lt} | X_{L(t-1)}, Y_{L(t-1)} \sim K^L_P(X_{L(t-1)}, \cdot)
\]
\[
Z^{(j)}_L | Y_{L(t-1)+(j-1)} \sim K^{L-(j-1)}_P(Y_{L(t-1)+(j-1)}, \cdot)
\]
\[
Y_{Lt} | X_{L(t-1)}, Y_{L(t-1)} \sim K^L_Q(Y_{L(t-1)}, \cdot)
\]
for \( s = 1, \ldots, L-1 \). When \( L = 1 \), we obtain \( \bar{K}_{1-step} = \bar{K} \) from Algorithm 2. Note that \( \bar{K}_{1-step} \) is the single-step kernel \( \bar{K} \) from Algorithm 2, but \( \bar{K}_{L-step} \) and \( \bar{K}^L \) are not equivalent in general.

Having developed algorithms to sample from the coupled kernels \( \bar{K} \) and \( \bar{K}_{L-step} \), we now investigate theoretical properties our upper bounds.
Figure 9: Joint kernel $\bar{K}_{L-step}$ on $\mathcal{X} \times \mathcal{X}$, which couples the marginal kernels $K^L_P$ and $K^L_Q$

Algorithm 3: Joint kernel $\bar{K}_{L-step}$ on $\mathcal{X} \times \mathcal{X}$, which couples the marginal kernels $K^L_P$ and $K^L_Q$

**Input:** chain states $X_0$ and $Y_0$, kernels $K_1$ and $K_2$, coupled kernels $\Gamma_1$ and $\Gamma_\Delta$

```plaintext
for s=1,...,L do
    Sample
    $$(X_s, Z_s^{(1)}, ..., Z_s^{(s)}, Y_s) | (X_{s-1}, Z_{s-1}^{(1)}, ..., Z_{s-1}^{(s-1)}, Y_{s-1})$$
    jointly such that
    $$(X_s, Z_s^{(1)}) \sim \Gamma_1(X_{s-1}, Z_{s-1}^{(1)})$$
    $$(Z_s^{(j)}, Z_s^{(j+1)}) \sim \Gamma_1(Z_{s-1}^{(j)}, Z_{s-1}^{(j+1)}) \text{ for } j = 1, ..., s - 1$$
    $$(Z_s^{(s)}, Y_s) \sim \Gamma_{\Delta}(Y_{s-1})$$
end
return $(X_{L(t-1)+s}, Y_{L(t-1)+s})$ for $s = 1, ..., L$.
```
D.2 Theoretical properties of couplings of multi-step kernels

To establish theoretical guarantees of coupled Markov chains based on the coupled kernel \( \tilde{K}_{L-step} \), we assume the Markovian coupling \( \Gamma_1 \) in Algorithm 3 satisfies a geometric ergodicity condition.

**Assumption D.1.** There exists constants \( C \in [1, \infty) \) and \( \rho \in (0, 1) \) such that for all \( L \geq 1 \),

\[
\mathbb{E}[c(X_{t+L}, Y_{t+L})^p|X_t, Y_t]^{1/p} \leq C \rho^{L} c(X_t, Y_t) \text{ for } (X_{t+L}, Y_{t+L})|(X_t, Y_t) \sim \Gamma_1^p(X_t, Y_t).
\]

Assumption D.1 is weaker than uniform contraction in Wasserstein’s distance as in Assumption 3.6. Under Assumption D.1, we now characterize the distance from our coupled chains based on the coupled kernel \( \tilde{K}_{L-step} \) explicitly in terms of the initial distribution \( \bar{I}_0 \) and the coupled kernel \( \Gamma_\Delta \) corresponding to perturbations between the marginal kernels \( K_1 \) and \( K_2 \). At the heart of our analysis is the construction of the coupled kernel \( \tilde{K}_{L-step} \) given in Figure 9 and Algorithm 3. When the coupled kernel \( \Gamma_\Delta \) characterizing the perturbation between the marginal kernels \( K_1 \) and \( K_2 \) is Wasserstein optimal, our analysis is linked to Rudolf and Schweizer [2018], which only considers the 1-Wasserstein distance and establishes similar results using analytic rather than probabilistic arguments.

**Theorem D.2.** Let \((X_t, Y_t)_{t \geq 0}\) denote a coupled Markov chain generated using Algorithm 1 with initial distribution \( \bar{I}_0 \) and joint kernel \( \bar{K} \) on \( \mathcal{X} \times \mathcal{X} \) from Algorithm 2. Suppose the coupled kernel \( \Gamma_1 \) satisfies Assumption D.1 for some \( C \geq 1 \) and \( \rho < 1 \). Fix some \( L \geq 1 \) such that \( \bar{\rho} = C \rho^L < 1 \), and consider the coupled chain \((X_t, Y_t)_{t \geq 0}\) generated using Algorithm 3 with the \( L \)-step coupled kernel \( \tilde{K}_{L-step} \). Then for all \( t \geq 0 \),

\[
\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \bar{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \bar{\rho}^{t-i} \left( \sum_{j=1}^{L} C \rho^{L-j} \mathbb{E} \left[ \Delta_p(Y_{L(i-1)+j}) \right]^{1/p} \right)
\]

where \((X_0, Y_0) \sim \bar{I}_0\) and \( \Delta_p(z) := \mathbb{E}[c(X, Y)^p]\) for \((X, Y)|z \sim \Gamma_\Delta(z)\).

**Corollary D.3.** Under the setup and assumptions of Theorem D.2, consider when the marginal distributions \( Q_t \) converge in \( p \)-Wasserstein distance to some distribution \( Q \) with finite moments of order \( p \) as \( t \to \infty \). Then for all \( \epsilon > 0 \), there exists some \( S \geq 1 \) such that for all \( t \geq S \),

\[
\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq (C \rho^L)^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + C \left( \frac{1 - (C \rho^L)^t}{1 - C \rho^L} \right) (1 - \rho^L) \mathbb{E}[\Delta_p(Y^*)]^{1/p} + \epsilon.
\]

where \((X_0, Y_0) \sim \bar{I}_0\), \( \Delta_p(z) \triangleq \mathbb{E}[c(X, Y)^p|z] \) for \((X, Y) \sim \Gamma_\Delta(z)\) and \( Y^* \sim Q \).

As in Section 3.3, we can also upper bound the limiting distance from our coupled chains in terms of the perturbations between the marginal kernels weighted by a Lyapunov function of \( K_2 \).

**Proposition D.4.** Under the setup and assumptions of Theorem D.2, let \( V : \mathcal{X} \to [0, \infty) \) be a \( p^{th} \)-order Lyapunov function of \( K_2 \) such that

\[
\mathbb{E}[V(Y_{t+1})^p|Y_t = z] \leq \gamma V(z)^p + L
\]

40
where \( \Delta_p(z) \triangleq \mathbb{E}[c(X,Y)^p|z] \) for \((X,Y) \sim \Gamma_{\Delta}(z)\). Then for all \( t \geq 0 \),

\[
\mathbb{E}[\text{CUB}_{p,t}^p]^{1/p} = \mathbb{E}[c(X_t,Y_t)^p]^{1/p} \leq (C\rho^L)^t \mathbb{E}[c(X_0,Y_0)^p]^{1/p} + C \left( \frac{1 - (C\rho^L)^t}{1 - C\rho^L} \right) \left( \frac{1}{1 - \rho} \right) \delta_K.
\]

### D.3 Proofs

**Proof of Theorem D.2.** Under the coupled kernel \( \tilde{K}_{L-step} \) from Algorithm 2, for each \( t \geq 1 \) we obtain

\[
(X_{Lt}, Z^{(1)}_L, \ldots, Z^{(L)}_L, Y_{Lt})
\]

where

\[
(X_{Lt}, Z^{(1)}_L)|X_{L(t-1)}, Y_{L(t-1)} \sim \Gamma^L_t(X_{L(t-1)}, Y_{L(t-1)})
\]

\[
(Z^{(j)}_L, Z^{(j+1)}_L)|Y_{L(t-1)+j-1} \sim \Gamma_{\Delta}(Y_{L(t-1)+j-1}) \Gamma_{L-1}^{L-j} \text{ for } j = 1, \ldots, L - 1
\]

\[
(Z_L^{(L)}, Y_{Lt})|Y_{L(t-1)+L-1} \sim \Gamma_{\Delta}(Y_{L(t-1)+L-1}).
\]

As \((X_{Lt}, Z^{(0)}_L)|X_{L(t-1)}, Y_{L(t-1)} \sim \Gamma^L_t(X_{L(t-1)}, Y_{L(t-1)})\), we obtain

\[
\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} = \mathbb{E}[\mathbb{E}[c(X_{Lt}, Y_{Lt})^p|X_{L(t-1)}, Y_{L(t-1)}]]^{1/p}
\]

\[
\leq \mathbb{E}[\mathbb{E}[(c(X_{Lt}, Z^{(1)}_L) + c(Z^{(1)}_L, Y_{Lt}))^p|X_{L(t-1)}, Y_{L(t-1)}]]^{1/p}
\]

\[
\leq \mathbb{E}[\mathbb{E}[c(X_{Lt}, Z^{(L)}_L)^p|X_{L(t-1)}, Y_{L(t-1)}]]^{1/p} + \mathbb{E}[\mathbb{E}[c(Z^{(1)}_L, Y_{Lt})^p|X_{L(t-1)}, Y_{L(t-1)}]]^{1/p}
\]

\[
\leq \tilde{\rho} \mathbb{E}[c(X_{L(t-1)}, Y_{L(t-1)})^p]^{1/p} + \mathbb{E}[c(Z^{(1)}_L, Y_{Lt})^p]^{1/p}
\]

where \( \tilde{\rho} \) is a constant.
where (28) follows as $c$ is a metric, (29) follows by Minowski’s inequality, and (30) follows by Assumption D.1. Denote $\Delta_p(z) \triangleq \mathbb{E}[c(X, Y)^p|z]$ for $(X, Y) \sim \Gamma_\Delta(z)$. Then,

$$
\mathbb{E}[c(Z^{(1)}_L, Y_{L_t})^{1/p}]^{1/p} \leq \mathbb{E}\left[\left(c(Z^{(L)}_L, Y_{L_t}) + \sum_{j=1}^{L-1} c(Z^{(j)}_L, Z^{(j+1)}_L)\right)^{1/p}\right]
$$

as $c$ is a metric

$$
\leq \mathbb{E}\left[c(Z^{(L)}_L, Y_{L_t})^{1/p}\right]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\left[c(Z^{(j)}_L, Z^{(j+1)}_L)^{1/p}\right]^{1/p}
$$

by Minowski’s inequality

$$
= \mathbb{E}\left[\mathbb{E}\left[c(Z^{(L)}_L, Y_{L_t})^{1/p}\right]|Y_{L_{(t-1)+L-1}}\right]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\left[\mathbb{E}\left[c(Z^{(j)}_L, Z^{(j+1)}_L)^{1/p}\right]|Y_{L_{(t-1)+j-1}}\right]^{1/p}
$$

$$
= \mathbb{E}[\Delta_p(Y_{L_{(t-1)+L-1}})]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}[\Delta_p(Y_{L_{(t-1)+j-1}})]^{1/p}
$$

by Assumption D.1

$$
\leq \sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}[\Delta_p(Y_{L_{(t-1)+j}})]^{1/p}
$$

as $C \geq 1$.

Equation (30) now gives

$$
\mathbb{E}[c(X_{L_t}, Y_{L_t})^{1/p}]^{1/p} \leq \bar{\rho}\mathbb{E}[c(X_{L_{(t-1)}}, Y_{L_{(t-1)}})^{1/p}]^{1/p} + \sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}[\Delta_p(Y_{L_{(t-1)+j}})]^{1/p}
$$

(31)

By induction, (31) implies

$$
\mathbb{E}[c(X_{L_t}, Y_{L_t})^{1/p}]^{1/p} \leq \bar{\rho}^{t}\mathbb{E}[c(X_{0}, Y_{0})^{1/p}]^{1/p} + \sum_{i=1}^{t} \bar{\rho}^{t-i}\left(\sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}[\Delta_p(Y_{L_{(i-1)+j}})]^{1/p}\right)
$$

as required.

Proof of Corollary D.3. Denote $a \triangleq \mathbb{E}[\Delta_p(Y^*)]^{1/p}$ for $Y^* \sim Q$ and $a_k \triangleq \mathbb{E}[\Delta_p(Y_k)]^{1/p}$ for $k \geq 0$. Then $a_k \xrightarrow{k \to \infty} a$, because $Q_t$ converges in $p$-Wasserstein distance to $Q$ as $t \to \infty$. This implies

$$
\sum_{i=1}^{t} \bar{\rho}^{t-i}\left(\sum_{j=1}^{L} C\rho^{L-j}a_{L_{(i-1)+j}}\right) \xrightarrow{t \to \infty} \sum_{i=1}^{t} \bar{\rho}^{t-i}\left(\sum_{j=1}^{L} C\rho^{L-j}a\right).
$$

Therefore, for all $\epsilon > 0$ there exists $S \geq 1$ such that for all $t \geq S$, $\sum_{i=1}^{t} \bar{\rho}^{t-i}\sum_{j=1}^{L} C\rho^{L-j}a_{L_{(i-1)+j}}$
\( a |< \epsilon \). By Theorem D.2,

\[
\mathbb{E}[c(X_{Lt}, Y_{Lt})^{1/p}] \leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + \sum_{i=1}^{t} \tilde{\rho}^i \left( \sum_{j=1}^{L} C\rho^{L-j} a L(i-1+j) \right)
\]

\[
\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + \sum_{i=1}^{t} \tilde{\rho}^i \sum_{j=1}^{L} C\rho^{L-j} a + \sum_{i=1}^{t} \tilde{\rho}^i \left( \sum_{j=1}^{L} C\rho^{L-j} |a L(i-1+j) - a| \right)
\]

\[
\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + \sum_{i=1}^{t} \tilde{\rho}^i \sum_{j=1}^{L} C\rho^{L-j} a + \epsilon
\]

\[
= (C\rho^L)^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + C \left( \frac{1 - (C\rho^L)^t}{1 - C\rho^L} \right) \left( \frac{1 - \rho^L}{1 - \rho} \right) a + \epsilon
\]

as required. \( \Box \)

**Proof of Proposition D.4.** As \( V \) is a a \( p \)-th-order Lyapunov function of \( K_2 \), by induction

\[
\mathbb{E}[V(Y_j)^p] \leq \gamma^j \mathbb{E}[V(Y_0)^p] + (1 - \gamma^j) \frac{L}{1 - \gamma}
\]

for all \( j \geq 0 \). This gives

\[
\mathbb{E}\left[ \Delta_p(Y_j) \right]^{1/p} \leq \delta \mathbb{E}[1 + V(Y_{j-1})^{1/p}]
\]

\[
\leq \delta \left( 1 + \mathbb{E}[V(Y_{j-1})^{1/p}] \right)
\]

\[
\leq \delta \left( 1 + \left( \gamma^{t-1} \mathbb{E}[V(Y_0)^p] + (1 - \gamma^{t-1}) \frac{L}{1 - \gamma} \right)^{1/p} \right)
\]

\[
\leq \delta \left( 1 + \max \left\{ \mathbb{E}[V(Y_0)^p]^{1/p}, \left( \frac{L}{1 - \gamma} \right)^{1/p} \right\} \right)
\]

\[
= \delta \kappa
\]

for all \( j \geq 0 \). By Theorem D.2, we obtain

\[
\mathbb{E}[c(X_{Lt}, Y_{Lt})^{1/p}] \leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + \sum_{i=1}^{t} \tilde{\rho}^i \left( \sum_{j=1}^{L} C\rho^{L-j} \mathbb{E}\left[ \Delta_p(Y_{L(i-1)+j}) \right]^{1/p} \right)
\]

\[
\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + \sum_{i=1}^{t} \tilde{\rho}^i \left( \sum_{j=1}^{L} C\rho^{L-j} \delta \kappa \right)
\]

\[
\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^{1/p}] + C \left( \frac{1 - (C\rho^L)^t}{1 - C\rho^L} \right) \left( \frac{1 - \rho^L}{1 - \rho} \right) \delta \kappa
\]

\( \Box \)
Details for the practical applications in Section 4

In this section, we provide details of the datasets, algorithms and parameters used for the three practical applications in Section 4. Open-source R code [R Core Team, 2013] recreating all experiments in this paper can be found at github.com/niloyb/BoundWasserstein.

E.1 Approximate MCMC and variational inference for tall data

Section 4.1 considers Bayesian logistic regression with a Gaussian prior applied to the Pima Diabetes dataset [Smith et al., 1988] and the DS1 life sciences dataset [Komarek and Moore, 2003]. The Pima Diabetes dataset has $n = 768$ binary observations (corresponding to the presence of diabetes), and $d = 8$ covariates (containing information such as body mass index, insulin level and age), and is publicly available on kaggle.com/uciml/pima-indians-diabetes-database. The DS1 life sciences dataset has $n = 26732$ binary observations (corresponding to reactivity of the compound observed in a life sciences experiment), and $d = 10$ covariates (containing information about the inputs to the life sciences experiment), and is publicly available on komarix.org/ac/ds/ (ds1.10 file).

In Figure 5, the upper bounds are given by our estimator $CUB_2(3)$ with $S = 1000, T = 2000$, and $I = 100$ for the Pima dataset and $S = 500, T = 100$, and $I = 40$ for the DS1 dataset, where these values were chosen based on initial runs. The lower bounds are estimated using (6) based on the same samples from the coupled chains used to calculate the upper bound estimate. For all the cases considered in Figure 5, we use a CRN coupling of the marginal kernels with a common step-size of $0.05$ for the Pima dataset and a common step-size of $0.05$ for the DS1 dataset. We also considered switching between CRN and reflection couplings based on the Euclidean norm between the two chains. This did not produce tighter upper bounds than CRN in our experiments, but it may be effective in other examples, so we have included this option in our released code.

E.2 Approximate MCMC for high-dimensional linear regression

Section 4.2 considers Bayesian linear regression with the half-t global-local shrinkage prior applied to a bacteria genome-wide association study (GWAS) dataset [Bühlmann et al., 2014] and a synthetically generated dataset. The GWAS dataset has $n = 71$ observations (corresponding to production of the vitamin riboflavin) and $d = 4088$ covariates (corresponding to single nucleotide polymorphisms (SNPs) in the genome) and is publicly available. The synthetically generated dataset has $n = 500$ observations and $d = 50000$ covariates. For the synthetic dataset, we generate $[X]_{i,j} \sim i.i.d. N(0,1)$ and $y \sim N(X\beta^*, \sigma^2I_n)$, where $\sigma_* = 2$ and $\beta^* \in \mathbb{R}^d$ is chosen to be sparse such that $\beta_{*,j} = 2^{(9-j)/4}$ for $1 \leq j \leq 20$ and $\beta_{*,j} = 0$ for all $j > 20$.

The state-of-the-art exact MCMC algorithms to sample from posteriors corresponding to the half-t prior are Gibbs samplers which cost $O(n^2d)$ per iteration. This computation cost arises from a weighted matrix product calculation of the form $X \text{Diag}(\eta_t)^{-1}X^\top$, where $\eta_t \in [0, \infty)^p$ corresponds to the local scale parameters which take different values at each iteration $t$. For the Horseshoe prior (degrees of freedom $\nu=1$), approximate MCMC methods have been developed by Johndrow et al.
based on approximations of the form
\[ X \operatorname{Diag}(\xi \eta)^{-1} X^\top \approx X \operatorname{Diag}(\xi^{-1} \eta_j^{-1} I_{(\xi^{-1} \eta_j^{-1} > \epsilon)} \rho_j = 1) X^\top \] (32)
for some small threshold \( \epsilon > 0 \). Biswas et al. [2021] extended the exact marginal chain of [Johndrow et al., 2020] to all degrees of freedom \( \nu \geq 1 \).

In Section 4.2, we use couplings to assess the quality of the approximate MCMC algorithm characterized by the approximation in (32) for \( \nu = 2 \). The upper bounds in Figure 6 are given by our estimator CUB_2 (3). We take \( S = 1000, T = 3000, \) and \( I = 100 \) for both datasets, where these values were chosen based on initial runs and the coupling-based convergence assessment of the exact chain from Biswas et al. [2021]. The lower bounds in Figure 6 are estimated using (6) based on same samples from the coupled chains used to calculate the upper bound estimate. We consider a CRN coupling with one marginal chain corresponding to the exact MCMC kernel and the other chain corresponding to the approximate MCMC kernel. The CRN coupled kernel is given in Algorithm 4.

E.3 Approximate MCMC for high-dimensional logistic regression
Section 4.3 considers Bayesian logistic regression with spike and slab priors applied to a malware detection dataset and a lymph node GWAS dataset. The Malware detection dataset from the UCI machine learning repository [Dua and Graff, 2017] has \( n = 373 \) observations (corresponding to a binary response vector indicating whether a file is malicious or non-malicious) and \( d = 503 \) covariates (corresponding to features of the files), and is publicly available on kaggle.com/piyushrumao/malware-executable-detection. The lymph node GWAS dataset [Hans et al., 2007, Liang et al., 2013, Narisetty et al., 2019] has \( n = 148 \) observations (corresponding to a binary response vector indicating high or low risk status of the lymph node that is related to breast cancer) and \( d = 4514 \) covariates (corresponding to SNPs in the genome) is not publicly available.

The logistic regression likelihood is given by
\[ L(\beta; y, X) = \prod_{i=1}^{n} \left( 1 + \exp(-y_i x_i^\top \beta) \right)^{-1} \]
where \( y \in \{-1, 1\}^n \) is the response vector, \( X \in \mathbb{R}^{n \times d} \) is the scaled design matrix with rows \( x_i^\top \), and \( \beta \in \mathbb{R}^d \) is an unknown signal vector. The spike and slab prior is given by
\[ Z_j \sim \text{Bernoulli}(q), \quad \beta_j | Z_j = 0 \sim \mathcal{N}(0, \tau_0^2), \quad \beta_j | Z_j = 1 \sim \mathcal{N}(0, \tau_1^2) \] (33)
for \( j = 1, \ldots, d \) where \( q \in (0, 1), \tau_0 > 0, \) and \( \tau_1 > 0 \) are hyper-parameters with \( \tau_0 \ll \tau_1 \) such that \( Z_j = 0 \) and \( Z_j = 1 \) correspond to null and non-null components of \( \beta_j \) respectively. By considering the posterior distribution of each variable \( Z_j \) on \( \{0, 1\} \), spike and slab priors provide an interpretable method for Bayesian variable selection [e.g. George and McCulloch, 1993, Ishwaran and Rao, 2005, Narisetty and He, 2014].

The state-of-the-art exact MCMC algorithms to sample from posteriors corresponding to the prior in (33) are Gibbs samplers which cost \( \mathcal{O}(n^2d) \) per iteration [Bhattacharya et al., 2016]. Narisetty et al. [2019] have recently developed approximate MCMC methods for this setting. Their approximate
Algorithm 4: Common random numbers coupling of an exact and an approximate Markov chain for Bayesian regression with half-t priors.

**Input:** exact chain current state $C_t \triangleq (\beta_t, \eta_t, \sigma^2_t, \xi_t) \in \mathbb{R}^d \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, approximate chain current state $\hat{C}_t \triangleq (\hat{\beta}_t, \hat{\eta}_t, \hat{\sigma}^2_t, \hat{\xi}_t) \in \mathbb{R}^d \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and approximation threshold $\epsilon > 0$.

1. Sample $(\eta_{t+1}, \hat{\eta}_{t+1})|\xi_t, \hat{\xi}_t, \sigma^2_t, \hat{\sigma}^2_t, \beta_t, \hat{\beta}_t$ component-wise, for each component $j$ targeting

$$\pi(\eta_{t+1}, j | \ldots) \propto \frac{e^{-m_{t,j} \eta_{t+1,j}}}{\eta_{t+1,j}^{1 + \nu \eta_{t+1,j}}} and \pi(\hat{\eta}_{t+1}, j | \ldots) \propto \frac{e^{-\hat{m}_{t,j} \eta_{t+1,j}}}{\eta_{t+1,j}^{1 + \nu \eta_{t+1,j}}}$$

for $m_{t,j} \triangleq (\xi_t \beta^2_{t,j})/(2\sigma^2_t)$ and $\hat{m}_{t,j} \triangleq (\hat{\xi}_t \hat{\beta}^2_{t,j})/(2\hat{\sigma}^2_t)$ respectively using common random numbers. This can be done using the slice sampler of Biswas et al. [2021].

2. Sample $(\xi_{t+1}, \hat{\xi}_{t+1}, \sigma^2_{t+1}, \hat{\sigma}^2_{t+1}, \beta_{t+1}, \hat{\beta}_{t+1})$ given $\eta_{t+1}$ and $\hat{\eta}_{t+1}$ as follows:

(a) Sample $(\xi_{t+1}, \hat{\xi}_{t+1})$ via Metropolis-Hastings with step size $\sigma_{\text{MH}} = 0.8$:

Propose $\log(\xi^*) = \log(\xi_t) + \sigma_{\text{MH}} Z^*$ and $\log(\hat{\xi}^*) = \log(\hat{\xi}_t) + \sigma_{\text{MH}} Z^*$ for $Z^* \sim \mathcal{N}(0, 1)$.

Calculate acceptance probabilities

$$q = \frac{L(y|\xi, \eta_{t+1})\pi_{\xi}(\xi^*)}{L(y|\xi_t, \eta_{t+1})\pi_{\xi}(\xi_t)} and \tilde{q} = \frac{L(y|\xi^*, \hat{\eta}_{t+1})\pi_{\xi}(\xi^*)}{L(y|\hat{\xi}_t, \hat{\eta}_{t+1})\pi_{\hat{\xi}}(\hat{\xi}_t)}$$

where $\pi_{\xi}(\cdot)$ is the prior density of $\xi$, $M \triangleq I_n + \xi_t^{-1}X \text{Diag}(\eta_{j_t}^{-1})X^\top$, $\hat{M} \triangleq I_n + X \text{Diag}(\hat{\xi}_t^{-1}\hat{\eta}_{j_t}^{-1}1_{\hat{\xi}_t^{-1}\hat{\eta}_{j_t}^{-1}>1})X^\top$ for $\hat{\xi}_{\text{max}} = \max\{\xi_t, \hat{\xi}_t\}$,

$$\log(L(y|\xi, \eta)) = -\frac{1}{2} \log(|M|) - \frac{a_0 + n}{2} \log(b_0 + y^\top M^{-1}y) and$$

$$\log(L(y|\xi, \eta)) = -\frac{1}{2} \log(|\hat{M}|) - \frac{a_0 + n}{2} \log(b_0 + y^\top \hat{M}^{-1}y).$$

Sample $U^* \sim \text{Uniform}(0, 1)$. Set $\xi_{t+1} \triangleq \xi^*$ if $U^* \leq \min(1, q)$, else set $\xi_{t+1} \triangleq \xi_t$. Set $\hat{\xi}_{t+1} \triangleq \hat{\xi}^*$ if $U^* \leq \min(1, \tilde{q})$, else set $\hat{\xi}_{t+1} \triangleq \hat{\xi}_t$. 

46
Algorithm 2: continued

2. [a)]

Sample \((\sigma_{t+1}^2, \tilde{\sigma}_{t+1}^2)|\xi_{t+1}, \tilde{\xi}_{t+1}, \eta_{t+1}, \tilde{\eta}_{t+1}\) using common random numbers, marginally targeting

\[
\sigma_{t+1}^2 | \xi_{t+1}, \eta_{t+1} \sim \text{InvGamma} \left( \frac{a_0 + n}{2}, \frac{y^T M_{\xi_{t+1}, \eta_{t+1}} y + b_0}{2} \right) \quad \text{and} \\
\tilde{\sigma}_{t+1}^2 | \tilde{\xi}_{t+1}, \tilde{\eta}_{t+1} \sim \text{InvGamma} \left( \frac{a_0 + n}{2}, \frac{y^T M_{\tilde{\xi}_{t+1}, \tilde{\eta}_{t+1}} y + b_0}{2} \right).
\]

(b) Sample \((\beta_{t+1}, \tilde{\beta}_{t+1}) | \sigma_{t+1}^2, \tilde{\sigma}_{t+1}^2, \xi_{t+1}, \tilde{\xi}_{t+1}, \eta_{t+1}, \tilde{\eta}_{t+1}\) with common random numbers and the fast sampling algorithms of Bhattacharya et al. [2016], marginally targeting

\[
\beta_{t+1} | \sigma_{t+1}^2, \xi_{t+1}, \eta_{t+1} \sim \mathcal{N} \left( \Sigma^{-1} X^T y, \sigma_{t+1}^2 \Sigma^{-1} \right) \quad \text{for} \quad \Sigma = X^T X + \xi_{t+1} \text{Diag}(\eta_{t+1}) \\
\tilde{\beta}_{t+1} | \tilde{\sigma}_{t+1}^2, \tilde{\xi}_{t+1}, \tilde{\eta}_{t+1} \sim \mathcal{N} \left( \tilde{\Sigma}^{-1} X^T y, \tilde{\sigma}_{t+1}^2 \tilde{\Sigma}^{-1} \right) \quad \text{for} \quad \tilde{\Sigma} = X^T X + \tilde{\xi}_{t+1} \text{Diag}(\tilde{\eta}_{t+1})
\]

return \(C_{t+1} \triangleq (\beta_{t+1}, \eta_{t+1}, \sigma_{t+1}^2, \xi_{t+1})\) and \(\tilde{C}_{t+1} \triangleq (\tilde{\beta}_{t+1}, \tilde{\eta}_{t+1}, \tilde{\sigma}_{t+1}^2, \tilde{\xi}_{t+1})\).

MCMC algorithm, called **Skinny Gibbs**, is based on matrix approximations of the form

\[
\begin{pmatrix}
X_A^T X_A + \tau_i^{-2}I & X_A^T X_A' \\
X_A' X_A & X_A' X_A' + \tau_0^{-2}I
\end{pmatrix} \approx \begin{pmatrix}
X_A^T X_A + \tau_i^{-2}I & 0 \\
0 & (n-1) + \tau_0^{-2}I
\end{pmatrix}
\]

where \(A = \{j : Z_j = 1\}, X_A\) is an \(n \times |A|\) matrix corresponding to the active columns \(j \in A\) of the design matrix, and \(X_A'\) is an \(n \times (d - |A|)\) matrix corresponding to the inactive columns \(j \notin A\). This gives an overall computation cost of \(\mathcal{O}(n \min\{d, |A|^2\})\) per iteration.

In Section 4.3, we use couplings to assess the quality of the Skinny Gibbs algorithm. The upper bounds in Figure 6 are given by our estimator CUB2 (3) with \(S = 1000, T = 3000,\) and \(I = 100\) for both the malware and lymph node GWAS datasets, where these values were chosen based on initial runs. The lower bounds in Figure 6 are estimated using (6) based on the same samples from the coupled chains used to calculate the upper bound estimate. We consider a CRN coupling between one marginal chain corresponding to the exact MCMC kernel and another corresponding to the Skinny Gibbs kernel. The CRN coupled kernel is given in Algorithm 3.
Algorithm 3: Common random numbers coupling of an exact and an approximate Markov chain for Bayesian logistic regression with spike and slab priors.

**Input:** exact chain current state $C_t \triangleq (\beta_t, z_t, e_t, w_t) \in \mathbb{R}^d \times \{0,1\}^d \times \mathbb{R}^n \times \mathbb{R}^n$ and approximate chain current state $\tilde{C}_t \triangleq (\tilde{\beta}_t, \tilde{z}_t, \tilde{e}_t, \tilde{w}_t) \in \mathbb{R}^d \times \{0,1\}^d \times \mathbb{R}^n \times \mathbb{R}^n$.

1. Sample $(\beta_{t+1}, \tilde{\beta}_{t+1})|z_t, e_t, w_t, \tilde{z}_t, \tilde{e}_t, \tilde{w}_t$ with common random numbers and the fast sampling algorithms of Bhattacharya et al. [2016], marginally targeting $W$ vectors of active components of $\tilde{C}_t$ in odds algorithms of Bhattacharya et al. [2016], marginally targeting indices in active components in $\mu$ and variance respectively where $\beta_{t+1}$ and $\tilde{\beta}_{t+1}$ are vectors of active components of $\beta_{t+1}$ and $\tilde{\beta}_{t+1}$ respectively.

2. Sample $(z_{t+1}, \tilde{z}_{t+1})$ given $\beta_{t+1}, \tilde{\beta}_{t+1}, e_t, \tilde{e}_t, w_t, \tilde{w}_t$ with common random numbers sequentially in order for $j = 1, ..., p$ such that each $z_{j,t+1}$ and $\tilde{z}_{j,t+1}$ are Bernoulli random variables with odds

$$\frac{q\mathcal{N}(\beta_{j,t+1},0,\tau_2^2)}{(1-q)\mathcal{N}(\beta_{j,t+1},0,\tau_0^2)}$$

and

$$\frac{q\mathcal{N}(\tilde{\beta}_{j,t+1},0,\tau_1^2)}{(1-q)\mathcal{N}(\tilde{\beta}_{j,t+1},0,\tau_0^2)} \exp \left( \tilde{\beta}_{j,t+1} X_j^T \tilde{W} (Y - X_{C_j} \tilde{\beta}_{C_j,t+1}) + \frac{1}{2} X_j^T (I - \tilde{W}) X_j \beta_{j,t+1}^2 \right)$$

respectively where $\mathcal{N}(:, \mu, \Sigma)$ is the probability density of the normal distribution with mean $\mu$ and variance $\Sigma$, $C_j \triangleq \{ k : \tilde{z}_{k,t+1} = 1 \text{ for } k < j \text{ or } \tilde{z}_{k,t} = 1 \text{ for } k > j \}$ is the index set of active components in $\{1, ..., p\}\{j\}$, $X_{C_j}$ is a matrix of the columns of $X$ which correspond to indices in $C_j$, and $\tilde{\beta}_{C_j,t+1}$ is a vector of the components of $\tilde{\beta}_{t+1}$ which correspond to indices in $C_j$. 

48
Algorithm 3: continued

3. Sample \((e_{t+1}, \tilde{e}_{t+1})|\beta_{t+1}, \tilde{\beta}_{t+1}, z_{t+1}, \tilde{z}_{t+1}, w_t, \tilde{w}_t\) with common random numbers component-wise independently such that for each \(i = 1, \ldots, n\)

\[
e_{i,t+1} \sim \begin{cases} 
N(x_i^\top \beta_{t+1}, w_{i,t}^{-1})I_{[0,\infty)} & \text{if } y_i = 1 \\
N(x_i^\top \beta_{t+1}, w_{i,t}^{-1})I_{(-\infty,0)} & \text{if } y_i = 0
\end{cases}
\]

\[
\tilde{e}_{i,t+1} \sim \begin{cases} 
N(x_i^\top \tilde{\beta}_{t+1}, \tilde{w}_{i,t}^{-1})I_{[0,\infty)} & \text{if } y_i = 1 \\
N(x_i^\top \tilde{\beta}_{t+1}, \tilde{w}_{i,t}^{-1})I_{(-\infty,0)} & \text{if } y_i = 0
\end{cases}
\]

where \(x_i^\top\) and \(x_i^{\top \tilde{A}}\) are the \(i^{th}\) row of the \(X\) and \(X_{\tilde{A}}\) respectively.

4. Sample \((w_{t+1}, \tilde{w}_{t+1})|\beta_{t+1}, \tilde{\beta}_{t+1}, z_{t+1}, \tilde{z}_{t+1}, e_{t+1}, \tilde{e}_{t+1}\). We take this variable to be fixed, and set

\[
w_{i,t} = \tilde{w}_{i,t} = 3/\pi^2 \text{ for all } i = 1, \ldots, n \text{ and } t \geq 0, \text{ where } 3/\pi^2 \text{ is the precision of the logistic distribution. In the case this variable can vary, they can be sampled using common random numbers such that for each } i = 1, \ldots, n,
\]

\[
w_{i,t+1} \sim \Gamma\left(\frac{\nu + 1}{2}, \frac{K(y_i - x_i^\top \beta_{t+1})^2}{2}\right) \text{ and } \tilde{w}_{i,t+1} \sim \Gamma\left(\frac{\nu + 1}{2}, \frac{K(y_i - x_i^{\top \tilde{A}} \tilde{\beta}_{t+1})^2}{2}\right)
\]

where \(\nu = 7.3, K \triangleq (\pi^2(\nu - 2)/3)\) are fixed constants as given in Narisetty et al. [2019].

return \(C_t \triangleq (\beta_t, z_t, e_t, w_t)\) and \(\tilde{C}_t \triangleq (\tilde{\beta}_t, \tilde{z}_t, \tilde{e}_t, \tilde{w}_t)\).
F Additional Algorithms

Algorithm 4: Common random numbers coupling of two MALA kernels marginally targeting distributions $P$ and $Q$ respectively

**Input:** $(X_t, Y_t)$, unnormalized densities $p$ and $q$ of $P$ and $Q$ respectively, step sizes $\sigma_P$ and $\sigma_Q$

Sample $\epsilon_{CRN} \sim N(0, I_d)$. Calculate proposals

$$X^* \triangleq X_t + \frac{1}{2} \sigma_P^2 \nabla \log p(X_t) + \sigma_P \epsilon_{CRN}$$

$$Y^* \triangleq Y_t + \frac{1}{2} \sigma_Q^2 \nabla \log q(Y_t) + \sigma_Q \epsilon_{CRN}$$

Sample $U_{CRN} \sim \text{Uniform}(0, 1)$

if $U_{CRN} \leq \frac{p(X^*) N(X^*; X_t + \frac{1}{2} \sigma_P^2 \nabla \log p(X_t))}{p(X_t) N(X_t; X_t + \frac{1}{2} \sigma_P^2 \nabla \log p(X_t))}$, then set $X_{t+1} = X^*$; else set $X_{t+1} = X_t$

if $U_{CRN} \leq \frac{q(Y^*) N(Y^*; Y_t + \frac{1}{2} \sigma_Q^2 \nabla \log q(Y_t))}{q(Y_t) N(Y_t; Y_t + \frac{1}{2} \sigma_Q^2 \nabla \log q(Y_t))}$, then set $Y_{t+1} = Y^*$; else set $Y_{t+1} = Y_t$

return $(X_{t+1}, Y_{t+1})$

Algorithm 5: Common random numbers coupling of a MALA kernel and an ULA kernel marginally targeting distributions $P$ and $Q$ respectively

**Input:** $(X_t, Y_t)$, unnormalized densities $p$ and $q$ of $P$ and $Q$ respectively, step sizes $\sigma_P$ and $\sigma_Q$

Sample $\epsilon_{CRN} \sim N(0, I_d)$. Calculate proposals

$$X^* \triangleq X_t + \frac{1}{2} \sigma_P^2 \nabla \log p(X_t) + \sigma_P \epsilon_{CRN}$$

$$Y^* \triangleq Y_t + \frac{1}{2} \sigma_Q^2 \nabla \log q(Y_t) + \sigma_Q \epsilon_{CRN}$$

Sample $U \sim \text{Uniform}(0, 1)$

if $U \leq \frac{p(X^*) N(X^*; X_t + \frac{1}{2} \sigma_P^2 \nabla \log p(X_t))}{p(X_t) N(X_t; X_t + \frac{1}{2} \sigma_P^2 \nabla \log p(X_t))}$, then set $X_{t+1} = X^*$; else set $X_{t+1} = X_t$

Set $Y_{t+1} = Y^*$

return $(X_{t+1}, Y_{t+1})$