Bounding Wasserstein distance with couplings

Niloy Biswas
Harvard University
niloy_biswas@g.harvard.edu

Lester Mackey
Microsoft Research New England
lmackey@microsoft.com

November 3, 2023

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 approximating MCMC in a manner that improves computational speed per iteration but does not produce asymptotically consistent estimates. 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]. They are based on a transition kernel K_1 which is invariant with respect to a target distribution of interest P. MCMC methods are asymptotically unbiased in that they generate Markov chains with marginal distributions that asymptotically converge to P as the number of iterations tend to infinity. However, in modern applications with a large number of data points or high dimensions, evaluating the transition kernel K_1 at each iteration can incur high computation cost. This has catalyzed the use of asymptotically biased sampling methods such as approximate MCMC and variational inference. Approximate MCMC [e.g., Welling and Teh, 2011, Bardenet

et al., 2017, Narisetty et al., 2019, Johndrow et al., 2020] is based on a transition kernel K_2 which is an approximation of K_1 with low computation cost; these approximate Markov chains typically converge to a distribution Q that differs from the target P. Variational inference [Blei et al., 2017, e.g.,] alternatively uses optimization to inexactly approximate P with a surrogate distribution Q.

Assessing the quality of such asymptotically biased samplers is of great interest for researchers who develop new approximate inference methods. Standard MCMC diagnostic tests [e.g., Johnson, 1998, Biswas et al., 2019, Vats and Knudson, 2021, Vehtari et al., 2021] 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. However, 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. In this manuscript, we develop generic upper bound estimates of the Wasserstein distance, an appealing measure of distance between distributions discussed in Sec. 1.2. Our estimates are then applied to assess the quality of asymptotically biased samplers.

1.2 Couplings and Wasserstein distances

Consider a complete, separable metric space $(\mathcal{X}, c)$ where c is a metric. For each $p \geq 1$, let $\mathcal{P}_p(\mathcal{X})$ denote the set of all probability measures P on $(\mathcal{X}, c)$ which have finite moments of order p, i.e., for which $\int_{\mathcal{X}} c(x_0, x)^p dP(x) < \infty$ for some $x_0 \in \mathcal{X}$. Then the p-Wasserstein distance is a metric on $\mathcal{P}_p(\mathcal{X})$, defined for any probability measures P and Q in $\mathcal{P}_p(\mathcal{X})$ as

$$\mathcal{W}_p(P, Q) = (\inf_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)^p d\gamma(x, y))^{1/p} \tag{1}$$

where $\Gamma(P, Q)$ is the set of probability measures on $\mathcal{X} \times \mathcal{X}$ with marginal measures P and Q respectively. Any probability measure in $\Gamma(P, Q)$ is called a coupling of P and Q, and any coupling which attains the infimum in (1) is called p-Wasserstein optimal.

The Wasserstein distance has many advantageous properties. Here we note those most relevant for this work and refer to Villani [2008] for more details. First, it allows comparison between mutually singular distributions that may have disjoint supports, unlike common alternatives like the total variation distance, Kullback–Leibler (KL) divergence and Rényi's α-divergences [van Erven and Harremos, 2014]. Moreover, it captures geometric properties induced by the metric c and differences between moments of distributions. For example when $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|_p = (\sum_{j=1}^d |x_i - y_i|^p)^{1/p}$, Jensen's inequality and the triangle inequality imply

$$\max\left\{\|\mathbb{E}[X - Y]\|_p, |\mathbb{E}[\|X\|_p^p]^{1/p} - \mathbb{E}[\|Y\|_p^p]^{1/p}|\right\} \leq \mathbb{E}_{\gamma^*(P,Q)}[\|X - Y\|_p^p]^{1/p} = \mathcal{W}_p(P, Q) \tag{2}$$

for any $P, Q \in \mathcal{P}_p(\mathcal{X})$ and random variables (X, Y) jointly distributed according to a p-Wasserstein optimal coupling $\gamma^*(P, Q)$. Equation (2) shows that p-Wasserstein distances can control the difference between moments of order p. Indeed, Huggins et al. [2020, Thm. 3.4, Rem. 3.5, and Prop. 3.6] showed that explicit bounds on Wasserstein distances translate into explicit guarantees for a variety of downstream inferential tasks including mean estimation, covariance estimation, numerical integration

of Lipschitz functions, and prediction accuracy. Meanwhile, these guarantees are *not* implied by a small KL or α-divergence [Huggins et al., 2020].

Popular approaches to estimating $\mathcal{W}_p(P, Q)$ involve drawing independent samples from P and Q and then computing the Wasserstein distance between the corresponding empirical distributions. Such approaches produce estimates that are consistent as the number of samples tend to infinity but can suffer from the curse of dimensionality and give loose upper bounds of $\mathcal{W}_p(P, Q)$ when the number of samples does not increase exponentially with dimension [e.g., Weed and Bach, 2019]. They also incurs prohibitive computational costs which scale at a cubic rate with the number of samples [Orlin, 1988]. Entropy-regularized variants of the Wasserstein distance such as Sinkhorn distances [Cuturi, 2013] offer computational costs which scale at a quadratic rate with the number of samples but produce estimates that are not consistent [Altschuler et al., 2017].

This manuscript develops consistent upper bound estimates for 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 cover in Sec. 3.4, our work provides an appealing alternative to estimates based on empirical Wasserstein distances and Sinkhorn distances and to the upper bound 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 on challenging contraction-constant estimation.

In related work, measures of asymptotic bias based on Stein discrepancies have been developed, which do not require sampling from the target distribution of interest. For example, Gorham et al. [2019] established a near-linear relationship between Stein discrepancies and standard Wasserstein distances, 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. Our upper bound estimates of the Wasserstein distance apply to any distributions that can be targeted with Markov chains and do not require any additional distributional knowledge.

1.3 Our contributions

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

In Sec. 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, 2022]. 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.

Sec. 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 Sec. 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 Sec. 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 $\mathcal{W}_p(P, Q)$. 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 a Markovian kernel $\bar{K}$ on the joint space $\mathcal{X} \times \mathcal{X}$ such that for all $x, y \in \mathcal{X}$,

$$\bar{K}\big((x,y),(\cdot,\mathcal{X})\big) = K_1(x,\cdot) \text{ and } \bar{K}\big((x,y),(\mathcal{X},\cdot)\big) = K_2(y,\cdot). \tag{3}$$

Given the kernel $\bar{K}$, we generate a coupled Markov chain $(X_t, Y_t)_{t\geq 0}$ using Alg. 1, 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, 2022]. 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 $\bar{K}$ are covered in Sec. 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 $(X_{t+1}, Y_{t+1})|(X_t, Y_t) \sim \bar{K}\big((X_t, Y_t), \cdot\big)$
return Markov chain $(X_t, Y_t)_{t=0}^T$

For a Markov chain $(X_t, Y_t)_{t\geq 0}$ from Alg. 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 $\mathcal{W}_p(P,Q)^p \leq \liminf_{S\to\infty, T-S\to\infty} \sum_{t=S+1}^{T} \frac{\mathbb{E}[c(X_t,Y_t)^p]}{T-S}$. This motivates our coupling upper bound (CUB) estimate

$$\text{CUB}_p \triangleq \big(\tfrac{1}{I(T-S)} \textstyle\sum_{i=1}^{I} \sum_{t=S+1}^{T} c(X_t^{(i)}, Y_t^{(i)})^p\big)^{1/p}, \tag{4}$$

where $(X_t^{(i)}, Y_t^{(i)})_{t=0}^T$ are sampled using Alg. 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 Sec. 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 (4) for two Gaussian distributions on $\mathbb{R}^d$, given by

$$P = \mathcal{N}(0, \Sigma) \text{ where } \Sigma_{i,j} = 0.5^{|i-j|} \text{ for } 1 \leq i, j \leq d \text{ and } Q = \mathcal{N}(0, I_d). \tag{5}$$

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 Alg. 4 of App. F. Each chain is initialized with independent draws of $X_0^{(i)} \sim P$ and $Y_0^{(i)} \sim Q$, and the choice of initialization is covered in Sec. 3. Throughout, we will also compare to an *independent coupling* obtained by sampling the $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ chains independently using the K_1 and K_2 kernels respectively.

Fig. 1 (Left) compares several upper bound estimates of $\mathcal{W}_2(P, Q)$ for dimension $d = 100$. The solid line (—) is CUB_2 based on $I = 5$ independent chains, burn-in $S = 0$ and varying trajectory length $1 \leq T \leq 1000$, and the grey error bands represent 95% confidence intervals arising from Monte Carlo error. As the marginal chains are initialized at their respective stationary distributions, here CUB_2 produces valid upper bounds for all trajectory lengths T with zero burn-in $S = 0$. The values of T and I are chosen based on upper bound estimates and error bands of initial runs, and this choice is further discussed in Sec. 3.2. The dotted line (····) plots the independent coupling upper bound $\mathop{\mathbb{E}}_{Y\sim Q, X\sim P}[\|X - Y\|_2^2]^{1/2} = (2d)^{1/2}$ with X and Y independent. The dot-dashed line (-·-) plots an estimate based on empirical Wasserstein distances, given by $\sum_{i=1}^{I} \mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})/I$ 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 $\mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})$ is calculated exactly by solving a linear program [Orlin, 1988, see also App. A.1]. In Sec. 3.4 we examine the upper- and lower-bounding properties of this common Wasserstein distance estimate and observe that its convergence can be slow in high dimensions due to substantial bias. Finally, the dashed line (- - -) shows the true Wasserstein distance $\mathcal{W}_2(P, Q)$, which is known for this stylized example [see, e.g., Peyré and Cuturi, 2019, Rem. 2.23] and is given by the coupling $\mathop{\mathbb{E}}_{Y\sim Q, X=\Sigma^{1/2}Y\sim P}[\|X - Y\|_2^2]^{1/2}$ where $\Sigma^{1/2}$ is the positive matrix square root of Σ. At initialization ($T = 0$) CUB_2 matches the equivalent independent coupling bound. For greater trajectory lengths T, CUB_2 offers a significant improvement over the independent bound and the popular empirical Wasserstein estimate.

Fig. 1 (right) considers $\mathcal{W}_2(P, Q)$ for higher dimensions. The solid line now plots CUB_2 based on $I = 5$, $S = 0$, and $T = 1000$. Fig. 1 (right) 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 Sec. 3. Sec. 3.4 provides a further comparison of our CUB bounds with empirical Wasserstein and Sinkhorn distances, which can have

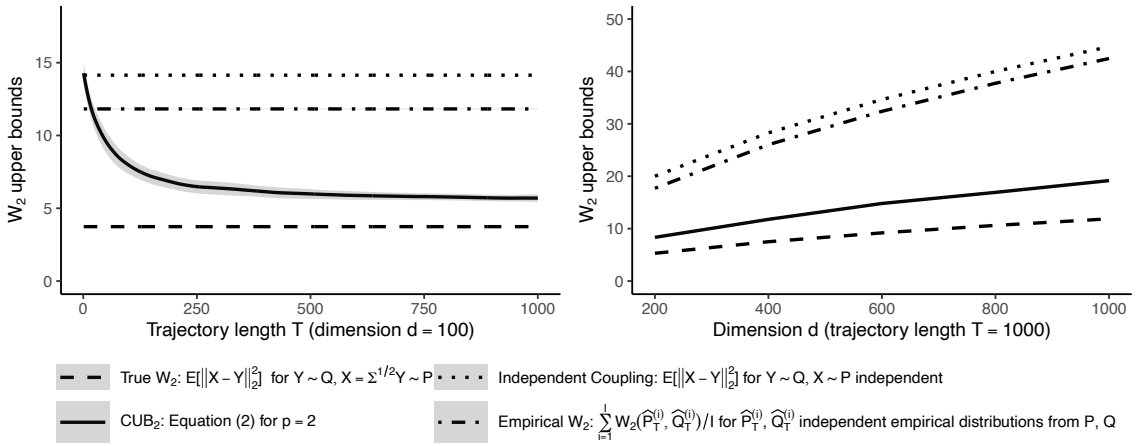


Figure 1: Upper bound estimates for $\mathcal{W}_2(P, Q)$ with $P = \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq d$, $Q = \mathcal{N}(0, I_d)$, and metric $c(x, y) = \|x - y\|_2$. See Sec. 2.1.

prohibitive computational cost for larger sample sizes and suffer from the curse of dimensionality.

2.2 Bias of approximate MCMC methods

The 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 are 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 ULA and the original target distribution of interest on a stylized example.

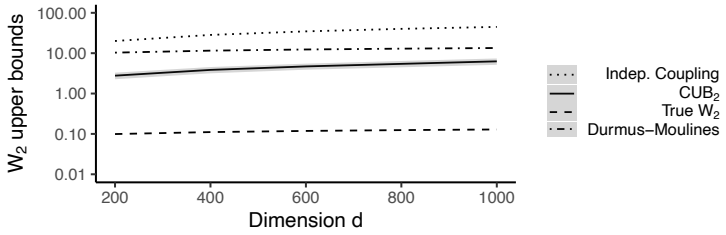


Figure 2: Upper bound estimates for the 2-Wasserstein distance with $c(x, y) = \|x - y\|_2$ between the limiting distributions of ULA and MALA targeting $P = \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq d$ on $\mathbb{R}^d$. See Sec. 2.2.

Fig. 2 shows the performance of CUB_2 (4) when the marginal kernels K_1 and K_2 are based, respectively, on the MALA and ULA Markov chains targeting the distribution $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$ defined

in (5). The MALA kernel K_1 produces an *exact* Markov chain which is $\mathcal{N}(0, \Sigma)$ invariant, and the ULA kernel K_2 produces an *approximate* Markov chain which is not $\mathcal{N}(0, \Sigma)$ invariant. The joint kernel $\bar{K}$ is based on a CRN coupling of the proposal steps of MALA and ULA, and is given in Alg. 5 of App. F. We again use a step size of $\sigma_P = \sigma_Q = 0.5d^{-1/6}$ for both marginal chains (following existing guidance for step size choice [Roberts and Rosenthal, 1998]) and initialize $X_0^{(i)} \sim \mathcal{N}(0, I_d)$ and $Y_0^{(i)} \sim \mathcal{N}(0, I_d)$ independently for each coupled chain i. Let P_t and Q_t denote the marginal distribution of $X_t^{(i)}$ and $Y_t^{(i)}$ respectively. We show in App. A.2 that $P_t \overset{t\to\infty}{\Rightarrow} P \triangleq \mathcal{N}(0, \Sigma)$, $Q_t = \mathcal{N}\big(0, \sigma_Q^2 \sum_{j=0}^{t-1} B^{2j}\big)$, and $Q_t \overset{t\to\infty}{\Rightarrow} Q \triangleq \mathcal{N}(0, \sigma_Q^2(I_d - B^2)^{-1})$, where $B \triangleq (I_d - (\sigma_Q^2/2)\Sigma^{-1})$ and the weak convergence of Q_t to Q holds for σ_Q sufficiently small.

Fig. 2 compares several approaches to bounding the asymptotic $\mathcal{W}_2(P, Q)$ bias of ULA. The solid line (—) displays our coupling upper bound estimate. For each dimension d, it is calculated using CUB_2 (4) with $I = 10$, $S = 1000$, and $T = 3000$. The dashed line (- - -) shows the true asymptotic bias $\mathcal{W}_2(P, Q)$ 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, Cor. 9] (see App. A.2 for more details). The tailored Durmus-Moulines bounds are significantly tighter than the convenient independent coupling bound, but CUB_2 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 Sec. 2, describe how to sample from the joint kernel $\bar{K}$ in Alg. 1, investigate the theoretical properties of our upper bounds, and compare to alternative approaches. All proofs are in App. 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 (4).

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*

$$\mathrm{CUB}_{p,t} \triangleq \Big(\tfrac{1}{I}\textstyle\sum_{i=1}^{I} c(X_t^{(i)}, Y_t^{(i)})^p\Big)^{1/p}. \tag{6}$$

If P_s and Q_s have finite moments of order p for all $s \leq t$, then $\mathrm{CUB}_{p,t}$ has finite moments of order p, and, as $I \to \infty$,

$$\mathrm{CUB}_{p,t}^p \overset{\text{a.s., } L^1}{\to} \mathbb{E}[\mathrm{CUB}_{p,t}^p] \geq \mathcal{W}_p(P_t, Q_t)^p.$$

Our next result shows that the estimator CUB_p (4) 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 Prop. 3.1, consider the estimator* CUB_p (4) *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* $I \to \infty$*,*

$$\mathrm{CUB}_p^p \overset{\text{a.s.},L^1}{\to} \mathbb{E}[\mathrm{CUB}_p^p] \geq \mathcal{W}_p(\tfrac{1}{T-S}\textstyle\sum_{t=S+1}^{T} P_t, \tfrac{1}{T-S}\sum_{t=S+1}^{T} Q_t)^p.$$

An important implication of Cor. 3.2 is that CUB_p (4) 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 Prop. 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* $(X_0, Y_0) \sim \bar{I}_0$ *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 (4) *has finite moments of order* p*, and as* $I \to \infty$*,*

$$\mathrm{CUB}_p^p \overset{\text{a.s.},L^1}{\to} \mathbb{E}[\mathrm{CUB}_p^p] \geq \mathcal{W}_p(P, Q)^p.$$

We may not always be able to initialize using the marginal stationary distributions P and Q. To obtain upper bounds on $\mathcal{W}_p(P, Q)$ 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 Assump. 3.4 and the assumptions and notation of Prop. 3.1, for all* $\epsilon > 0$ *there exists* $S \geq 1$ *such that for all* $T \geq S$*, the estimator* CUB_p (4) *has finite moments of order* p*, and as* $I \to \infty$*,*

$$\mathrm{CUB}_p^p \overset{\text{a.s.},L^1}{\to} \mathbb{E}\big[\mathrm{CUB}_p^p\big] \geq \mathcal{W}_p(P, Q)^p - \epsilon.$$

Prop. 3.5 establishes that CUB_p with any initialization $(X_0, Y_0) \sim \bar{I}_0$ consistently bounds $\mathcal{W}_p(P, Q)$ 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 App. A.3 for details).

We emphasize that the results of this section hold for any coupled chain sampled using Alg. 1 with joint kernel $\bar{K}$ satisfying (3). For example, this includes both the CRN coupled chains and the independently coupled chains from Sec. 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 $\bar{K}$ 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 Sec. 3.1 can produce informative upper bounds. Our construction decomposes the overall coupling into two convenient coupling steps based on a same-chain coupling kernel Γ_1 on $\mathcal{X} \times \mathcal{X}$ and a perturbative coupling kernel Γ_Δ on $\mathcal{X}$:

1. Γ_1 is a Markovian coupling of the kernel K_1 with itself: for all $x, \tilde{x} \in \mathcal{X}$, $\Gamma_1(x, \tilde{x})$ is a coupling of the distributions $K_1(x, \cdot)$ and $K_1(\tilde{x}, \cdot)$.

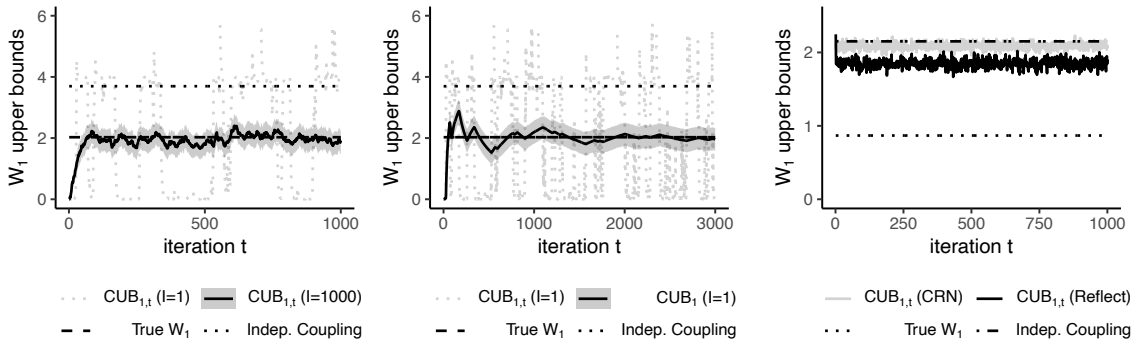
2. Γ_Δ is coupling of kernels K_1 and K_2 from the same point: for all $z \in \mathcal{X}$, $\Gamma_\Delta(z)$ is a coupling of the distributions $K_1(z, \cdot)$ and $K_2(z, \cdot)$.

This decomposition allows us to exploit the extensive and growing literature on same-chain coupling kernels and their properties (see Section 3.3) and to analyze the targeting of two distinct stationary distributions as a simple perturbation to well-studied same-chain couplings. For example, when K_1 is a Metropolis–Hastings kernel, Γ_1 can be a CRN coupling of both the proposal step and the accept-reject step. Indeed, we often make use of CRN couplings as a default choice in this work due to their broad applicability and straightforward implementation. 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 (HMC)—Γ_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 Γ_Δ 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, Γ_Δ 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 Γ_1 and Γ_Δ further in Sec. 3.3. Given Γ_1 and Γ_Δ, we sample from the joint kernel $\bar{K}$ using Alg. 2.

Algorithm 2: Joint kernel $\bar{K}$ 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 Γ_1 and Γ_Δ
Sample $(X_t, Z_t, Y_t)|X_{t-1}, Y_{t-1}$ 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)

Alg. 2 gives the conditional marginal distributions $X_t|X_{t-1}, Y_{t-1} \sim K_1(X_{t-1}, \cdot)$, $Z_t|X_{t-1}, Y_{t-1} \sim K_1(Y_{t-1}, \cdot)$, $Y_t|X_{t-1}, Y_{t-1} \sim K_2(Y_{t-1})$ so that $\bar{K}$ satisfies (3). Often Alg. 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 σ_P and σ_Q, target distributions P and Q, and Γ_1 and Γ_Δ are CRN coupled kernels. Given (X_{t-1}, Y_{t-1}), we sample $\epsilon_{CRN} \sim \mathcal{N}(0, I_d)$ and calculate the proposals $X^* = X_{t-1} + (\sigma_P^2/2)\nabla \log P(X_{t-1}) + \sigma_P \epsilon_{CRN}$, $Z^* = Y_{t-1} + (\sigma_P^2/2)\nabla \log P(Y_{t-1}) + \sigma_P \epsilon_{CRN}$, 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



(a) Impact of multiple trajectories on CRN coupling bounds.

(b) Impact of ergodic averaging on CRN coupling bounds.

(c) Impact of coupling choice on bound quality.

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 Sec. 3.2.

and ULA is included in Alg. 5 of App. F. App. 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_1 (4) and instantaneous $\mathrm{CUB}_{1,t}$ (6) estimators when bounding the 1-Wasserstein distance with $c(x, y) = \|x - y\|_2$ between $P = \frac{1}{2}\mathcal{N}(1_d, I_d) + \frac{1}{2}\mathcal{N}(-1_d, I_d)$ and $Q = \mathcal{N}(1_d, I_d)$ with $d = 4$ so that one of the marginal target distributions is bimodal with well-separated modes. We simulate the coupled chains $(X_t^{(i)}, Y_t^{(i)})_{t\geq 0}$ independently for each i using Alg. 1, where the joint kernel $\bar{K}$ is based on a CRN coupling of MALA kernels K_1 and K_2 targeting distributions P and Q respectively. The MALA kernels have a common step size $d^{-1/6}$ (following existing guidance for step size choice [Roberts and Rosenthal, 1998]), and we initialize $X_0^{(i)} = 1_d$ and $Y_0^{(i)} = 1_d$ such that both marginal chains start at the common mode. Fig. 3a isolates the impact of averaging over multiple chains when computing the $\mathrm{CUB}_{1,t}$ estimate (6). 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 $\mathrm{CUB}_{1,t}$ (6) 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 $\mathcal{W}_1(P, Q)$ 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 CUB_1 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 Fig. 3b. 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] contains related motivating discussions for unbiased estimation with couplings.

Choice of coupled kernel. Secondly, we highlight the importance of the choice of the coupled kernel $\bar{K}$. Fig. 3c examines the performance of the CUB_1 (4) estimator when bounding the 1-Wasserstein distance with $c(x, y) = \|x - y\|_2$ between $P = \frac{1}{2}\mathcal{N}(2, 1) + \frac{1}{2}\mathcal{N}(-2, 1)$ and $Q = \frac{1}{2}\mathcal{N}(1, 1) + \frac{1}{2}\mathcal{N}(-1, 1)$, 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 Fig. 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 Sec. 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, Fig. 3c highlights that our upper bounds may not always be very close to the true Wasserstein distance when the marginal Markov chains have slow mixing rates or when the coupling of the marginal transition kernels is not close to optimal. 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 (4) consistently upper bounds Wasserstein distances (Sec. 3.1) and developed algorithms to compute CUB in practice (Sec. 3.2). We next derive upper bounds on the size of CUB to provide interpretable sufficient conditions under which CUB is guaranteed to be small. We emphasize that it is possible for CUB to be significantly smaller than these interpretable bounds and for CUB to be small even when the assumptions of the interpretable bounds are not met. Hence, when bounding Wasserstein distances in practice, we would not recommend computing these intepretable bounds but rather computing the even tighter CUB Wasserstein bound directly.

Our analysis is based on Markov chain perturbation theory for $\mathcal{W}_1$ [Pillai and Smith, 2015, Johndrow and Mattingly, 2018, Rudolf and Schweizer, 2018], which we generalize to $\mathcal{W}_p$ for all $p \geq 1$. This is a useful extension, as $\mathcal{W}_2$ in particular is believed to better reflect geometric features and adapt to geometric structure than $\mathcal{W}_1$ [Villani, 2008, Rem. 6.6]. We also discuss examples where the $\mathcal{W}_p$ upper bounds do not explicitly depend on the state space dimension and are stable up to a coupling of the one-step marginal kernels.

To establish our CUB_p upper bounds, we assume that the Markovian coupling Γ_1 in Alg. 2 gives uniform contraction in Wasserstein distance. Recall that Γ_1 is a coupling of the marginal kernel K_1 with itself, so Assump. 3.6 concerns only the single kernel K_1 targeting the single stationary distribution P.

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

Assump. 3.6 is stronger than the convergence assumption of the marginal chain corresponding to kernel K_1 (Assump. 3.4 for $(P_t)_{t\geq 0}$). For many popular MCMC algorithms, Assump. 3.6 has been established under certain metrics c and coupled kernels Γ_1 to give contraction rates ρ that do not explicitly depend on the dimension of the state space $\mathcal{X}$. This includes MALA [Eberle, 2014] and HMC [Bou-Rabee et al., 2020]. When the target distributions are log-concave, these algorithms satisfy Assump. 3.6 with $c(x,y) = \|x-y\|_2$ and the coupled kernel Γ_1 based on a CRN coupling. For target distributions satisfying a weaker distant dissipativity condition [Eberle, 2016, Gorham et al., 2019] (including, for example, multimodal distributions with Gaussian tails), these algorithms satisfy Assump. 3.6 with Γ_1 based on a combination of CRN and reflection coupling and a metric $\tilde{c}$ satisfying $r\tilde{c}(x,y) \leq \|x-y\|_2 \leq R\,\tilde{c}(x,y)$ for some $0 < r \leq R < \infty$.

Furthermore, we can weaken Assump. 3.6 to a geometric ergodicity condition as in [Rudolf and Schweizer, 2018], where for some constants $C \geq 1$, $\rho \in (0,1)$, and for all $L \geq 1$, $\mathbb{E}[c(X_{t+L}, Y_{t+L})^p | X_t, Y_t]^{1/p} \leq C\rho^L c(X_t, Y_t)$ for $(X_{t+L}, Y_{t+L})|(X_t, Y_t) \sim \Gamma_P^L(X_t, Y_t)$ where $\Gamma_P^L(X_t, Y_t)$ denotes a coupling of L-steps of the kernel K_1 marginally starting from states X_t and 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 App. D for completeness.

Under Assump. 3.6, we can upper bound the distance from our coupled chains explicitly in terms of the initial distribution $\bar{I}_0$, contraction constant ρ, and coupled kernel Γ_Δ 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 Alg. 1 with initial distribution $\bar{I}_0$ and joint kernel $\bar{K}$ from Alg. 2. Suppose the coupled kernel Γ_1 satisfies Assump. 3.6 for some $\rho \in (0,1)$. Then*

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

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

For $\mathrm{CUB}_{p,t}$ based on a metric c, one obtains an analogous bound if Assump. 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 $\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq R\,\mathbb{E}[\tilde{c}(X_t, Y_t)^p]^{1/p}$. 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 Thm. 3.7, suppose that the marginal distributions Q_t converge in p-Wasserstein distance to some distribution Q as $t \to \infty$. Then for each $\epsilon > 0$, there exists $S \geq 1$ such that for all $t \geq S$,*

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

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 $Y^* \sim Q$.

Cor. 3.8 gives $\mathcal{W}_p(P,Q) \leq \liminf_{t\to\infty} \mathbb{E}[\mathrm{CUB}_{p,t}^p]^{1/p} \leq \mathbb{E}[\Delta_p(Y^*)]^{1/p}/(1-\rho)$, implying that CUB estimators 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 ρ does not explicitly depend on the dimension, then our upper bounds do not increase unfavorably with dimension and remain informative in high dimensional settings. Hence Cor. 3.8 provides interpretable sufficient conditions for CUB to be dimension-free, as in Figs. 1 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 Thm. 3.7, let $V : \mathcal{X} \to [0, \infty)$ satisfy $\mathbb{E}[V(Y_{t+1})^p | Y_t = z] \leq \gamma V(z)^p + L$ for some fixed constants $\gamma \in [0, 1)$ and $L \in [0, \infty)$ and all $z \in \mathcal{X}$. Define $\delta \triangleq \sup_{z\in\mathcal{X}} \left(\frac{\Delta_p(z)}{1+V(z)^p}\right)^{1/p}$ and $\kappa \triangleq \left(1 + \max\left\{\mathbb{E}[V(Y_0)^p], \frac{L}{1-\gamma}\right\}\right)^{1/p}$, where $\Delta_p(z) \triangleq \mathbb{E}[c(X, Y)^p | z]$ for $(X, Y) \sim \Gamma_\Delta(z)$. Then for all $t \geq 0$,*

$$\mathbb{E}[\mathrm{CUB}_{p,t}^p]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1 - \rho^t)\frac{\delta\kappa}{1-\rho}.$$

In the case $p = 1$, Prop. 3.9 recovers Thm. 3.1 of Rudolf and Schweizer [2018]. For such result to be informative, we require functions V such that $\delta\kappa$ 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 App. 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 $\mathcal{W}_p(P, Q)$ is to draw independent samples from P and Q and then exactly compute the $\mathcal{W}_p$ distance between the empirical distributions. This is precisely the empirical Wasserstein estimate that appeared in Fig. 1. As our next proposition, proved in App. B.3, demonstrates, this empirical Wasserstein approach consistently upper bounds $\mathcal{W}_p(P, Q)$.

Proposition 3.10 (Empirical Wasserstein distance bounds)**.** *For P and Q in $\mathcal{P}_p(\mathcal{X})$, let $\hat{P}_n$, $\tilde{P}_n$, $\hat{Q}_n$, and $\tilde{Q}_n$ denote empirical distributions of the samples $(X_i)_{i=1}^n$, $(\tilde{X}_i)_{i=1}^n$, $(Y_i)_{i=1}^n$, and $(\tilde{Y}_i)_{i=1}^n$ respectively, where $X_i, \tilde{X}_i \overset{i.i.d.}{\sim} P$ and, independently, $Y_i, \tilde{Y}_i \overset{i.i.d.}{\sim} Q$ for all $i = 1, \ldots, n$. Then, $\mathcal{W}_p(\hat{P}_n, \hat{Q}_n) \overset{\text{a.s.}}{\to} \mathcal{W}_p(P, Q)$ as $n \to \infty$, and*

$$0 \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p]^{1/p} - \mathcal{W}_p(P, Q) \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, \tilde{P}_n)^p]^{1/p} + \mathbb{E}[\mathcal{W}_p(\hat{Q}_n, \tilde{Q}_n)^p]^{1/p}.$$

However, there are two downsides to the empirical Wasserstein approach. The first is statistical. The difference between $\mathbb{E}\left[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p\right]^{1/p}$ and $\mathcal{W}_p(P, Q)$ can be quite large and decay very slowly in n. For example, for some d-dimensional target distributions, $\mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)]$ converges to $\mathcal{W}_p(P, Q)$ at rate $\Omega(n^{-1/d})$ when $d > 2p$ [Weed and Bach, 2019]. This can lead to the empirical Wasserstein distance giving loose upper bounds on $\mathcal{W}_p(P, Q)$ when the number of samples does not increase exponentially with dimension. The example in Fig. 1 illustrates this curse of dimensionality, where

the estimator CUB_p (4) with CRN coupling gives tighter upper bounds of $\mathcal{W}_p(P, Q)$ than the empirical Wasserstein estimates.

The second downside is computational. Calculating $\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)$ amounts to solving an uncapacitated minimum cost flow problem with $\mathcal{O}(n^3 \log n)$ computational cost [Orlin, 1988], prohibitive cost 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 bound accuracy. A smaller λ leads to more expensive $\mathcal{O}(n^2/(\lambda\epsilon))$ computation time for ϵ-accurate solutions [Altschuler et al., 2017] and potential instability of the Sinkhorn algorithm in practice. See App. A.4 for simulations illustrating these issues.

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 Euclidean Wasserstein distances in terms of KL or α-divergences. To estimate their upper bounds of $\mathcal{W}_p(P, Q)$ 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. Fig. 4 (left) plots the performance of the $\mathcal{W}_2$ bounds of Huggins et al. for the example in Sec. 2.1. The dot-dashed line represents the mean of $I = 20$ independent Huggins et al. importance-sampling estimators, each with $2T = 3000$ samples from Q. The CUB_2 estimator plotted for comparison uses I independent CRN coupled chains with trajectory length T and burnin $S = 500$. In this example, the Huggins et al. bounds are significantly looser than both our CRN coupling bound and the independent coupling upper bound. Furthermore, the Huggins et al. estimates 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 Sec. 2.2, the stochastic gradient-based samplers in Sec. 4.1, and the matrix approximation-based sampler in Sec. 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. Our next result, proved in App. B.4, 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,* $\mathbb{E}[\text{CUB}_1]$ (4) *lower bounds the estimated upper bound of Dobson et al. [2021].*

Fig. 4 (right) plots the 1-Wasserstein upper bounds of Dobson et al. and CUB_1 for the example

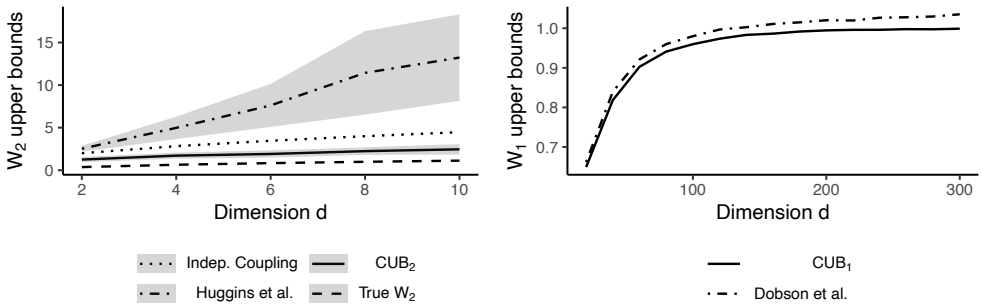


Figure 4: **(Left)** Upper bound estimates for $\mathcal{W}_2$ with $c(x,y) = \|x-y\|_2$ between $P = \mathcal{N}(0,\Sigma)$ and $Q = \mathcal{N}(0, I_d)$ for $[\Sigma]_{i,j} = 0.5^{|i-j|}$ for $1 \le i,j \le d$. The Huggins et al. [2020] bound is looser than CUB_2 and has larger variance as the dimension grows. See Sec. 3.4 for more details. **(Right)** Upper bound estimates for $\mathcal{W}_1$ with $c(x,y) = \min\{1, \|x-y\|_2\}$ between ULA and MALA chains targeting $P = \mathcal{N}(0,\Sigma)$. In line with Prop. 3.11, the CUB_1 (4) estimate is tighter than the Dobson et al. [2021] bound employing the same CRN coupling. See Sec. 3.4 for more details.

in Sec. 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 Prop. 3.11, the upper bounds of Dobson et al. are looser than CUB_1.

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 (2) 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 $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$, let P_i and Q_i denote the marginal distributions of the i^{th} component of P and Q respectively. Let $\mathcal{N}_P$ and $\mathcal{N}_Q$ denote Gaussian distributions on $\mathbb{R}^d$ with the same means and covariance matrices as P and Q respectively. Then,

$$\max\left\{\textstyle\sum_{i=1}^d \mathcal{W}_2(P_i,Q_i)^2\ ,\ \mathcal{W}_2(\mathcal{N}_P,\mathcal{N}_Q)^2\right\} \le \mathcal{W}_2(P,Q)^2. \tag{7}$$

Here, $\sum_{i=1}^d \mathcal{W}_2(P_i,Q_i)^2 \le \mathcal{W}_2(P,Q)^2$ follows from the coupling representation of $\mathcal{W}_2(P,Q)$, and $\mathcal{W}_2(\mathcal{N}_P,\mathcal{N}_Q) \le \mathcal{W}_2(P,Q)$ is the lower bound of Gelbrich [1990, Thm. 2.1]. Each one-dimensional Wasserstein distance $\mathcal{W}_2^2(P_i,Q_i)$ admits a convenient representation for estimation, given by $\int_0^1 (F_{P_i}^{-1}(u) - F_{Q_i}^{-1}(u))^2 du$ where $F_{P_i}^{-1}$ and $F_{Q_i}^{-1}$ are the inverse cumulative distribution functions of P_i and Q_i respectively, while $\mathcal{W}_2(\mathcal{N}_P,\mathcal{N}_Q)$ has the closed form $\big(\|\mu_P-\mu_Q\|_2^2 + \text{Trace}\big(\Sigma_P+\Sigma_Q - 2(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2})^{1/2}\big)\big)^{1/2}$ in terms of the means μ_P, μ_Q and covariances Σ_P, Σ_Q of P and Q [Peyré and Cuturi, 2019, Rem. 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 (7). Details of all the datasets, algorithms, and specific estimator parameters used in this section can be found in App. 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 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]. For each sampling algorithm, Fig. 5 plots CUB_2 (4) upper bounds and $\mathcal{W}_2$ lower bounds estimated using (7). 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 $\mathcal{N}(\mu_L, \Sigma_L)$ 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 $\mathcal{N}(\mu_{VB}, \Sigma_{VB})$ where $\mu_{VB} \in \mathbb{R}^d$ and $\Sigma_{VB} \in \mathrm{Diag}(\mathbb{R}^{d\times d})$ 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 $(Y_t^{(i)})_{t\geq 0}$. App. E.1 contains details about the datasets, algorithms and estimator parameters used.

Fig. 5 shows that Laplace approximation has the smallest asymptotic bias for both datasets. This promising Laplace performance 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 for the response vector $y \in \mathbb{R}^n$ is a Gaussian density with mean $X\beta$ and covariance matrix $\sigma^2 I_n$, where $X \in \mathbb{R}^{n\times d}$ 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

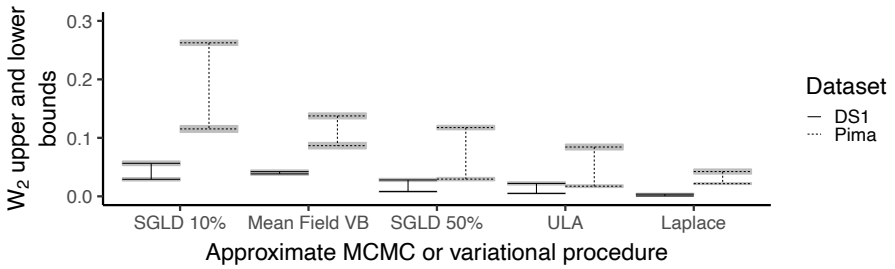


Figure 5: Bounds on the Euclidean $\mathcal{W}_2$ bias of approximate MCMC and variational inference procedures for Bayesian logistic regression. We consider the DS1 dataset ($n = 26732$ observations, $d = 10$ covariates) and the Pima dataset ($n = 768$, $d = 8$). See Sec. 4.1 for more details.

mixture priors, given by

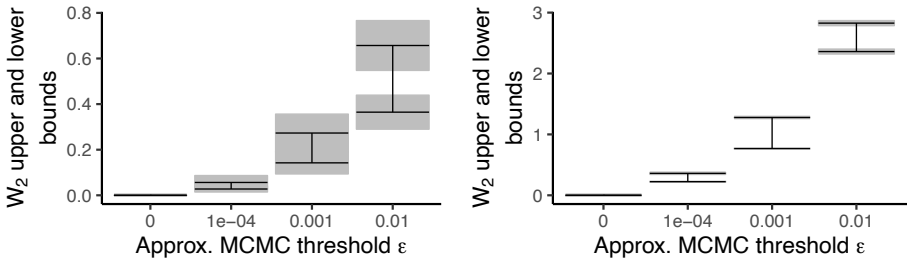
$$\xi^{-1/2} \sim \mathcal{C}_+(0,1),\ \eta_j^{-1/2} \overset{i.i.d.}{\sim} t_+(\nu),\ \sigma^{-2} \sim \text{Gamma}\Big(\tfrac{a_0}{2}, \tfrac{b_0}{2}\Big),\ \beta_j|\eta,\xi,\sigma^2 \overset{ind.}{\sim} \mathcal{N}\Big(0, \tfrac{\sigma^2}{\xi\eta_j}\Big)$$

where $\mathcal{C}_+(0,1)$ is the half-Cauchy distribution on $[0,\infty)$ and $t_+(\nu)$ is the half-t distribution on $[0,\infty)$ with ν 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 Sec. 4.1, as now the posterior distribution is multi-modal, has polynomial tails, and has infinite density about the origin [Biswas et al., 2022]. 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. [2022] extended the sampler of Johndrow et al. to all $\nu \geq 1$ and showed that using larger values of ν 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. App. E.2 contains details about the datasets, algorithms, and estimator parameters used.

Fig. 6 plots upper and lower bounds on 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 CUB_2 (4), and the lower bounds are estimated using (7). 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. In particular, the bounds enable a developer to assess the computational cost of an approximation procedure as a function of the bias introduced (and vice-versa). For example, for any maximum acceptable bias level, one can identify the largest approximation parameter ϵ with a CUB interval below the acceptable level and assess the computational savings delivered relative to an exact sampler.

Often one will choose an acceptable level of Wasserstein bias based on the direct implications for downstream inferential tasks (e.g., based on tolerable discrepancies in predictive accuracy or



(a) GWAS dataset ($n = 71, d = 4088$) (b) Synthetic dataset ($n = 500, d = 50000$)

Figure 6: Bounds on the Euclidean $\mathcal{W}_2$ bias of an approximate MCMC Gibbs sampler for high-dimensional Bayesian regression with half-t(2) prior, n observations, and d covariates. We consider both a bacteria GWAS dataset and a synthetic dataset. See Sec. 4.2 for more details.

numerical integration, as discussed in Section 1.2). When it is otherwise difficult for a user to select an acceptable level of Wasserstein bias on an absolute scale, we would recommend normalizing each CUB estimate based on the coupled chains $(X_t^{(i)}, Y_t^{(i)})_{t=0}^T$ by a second, independent-coupling CUB estimate based on the chains $(X_t^{(i)}, \tilde{X}_t^{(i)})_{t=0}^T$, where $(\tilde{X}_t^{(i)})_{t=0}^T$ is sampled independently of $(X_t^{(i)})_{t=0}^T$ using the P-invariant K_1 kernel. This enables Wasserstein bias to be assessed relative to a measure of the intrinsic variability or noise level in the target distribution P.

4.3 Approximate MCMC for high-dimensional logistic regression

We now consider high-dimensional Bayesian logistic regression with spike and slab priors, a popular choice for Bayesian variable selection [Tadesse and Vannucci, 2021]. Narisetty et al. [2019] recently developed an approximate MCMC algorithm called *Skinny Gibbs*, to sample from posteriors in this setting. Here, 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 using a CRN coupling between the exact MCMC kernel and the Skinny Gibbs kernel. App. E.3 contains further details about spike and slab priors and the datasets, algorithms, and estimator parameters used.

Fig. 7 displays CUB_2 (4) upper bounds and lower bounds estimated using (7) on the Euclidean 2-Wasserstein distance between the limiting distributions of the exact and Skinny Gibbs chains for β. We display these bounds not to draw comparisons across the datasets but rather to exemplify the level of precision provided by CUB when applied to real high-dimensional logistic regression tasks. For researchers developing approximate 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 Wasserstein distance between the limiting distribution of the approximation and the original target distribution of interest. The proposed estimators can be

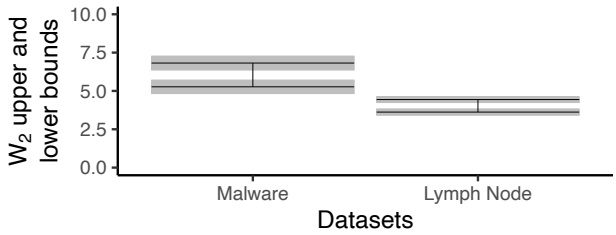


Figure 7: Bounds on the Euclidean $\mathcal{W}_2$ bias of the Skinny Gibbs sampler [Narisetty et al., 2019] for Bayesian logistic regression with a spike and slab prior; see Sec. 4.3 for details. We consider a malware dataset ($n = 373$ observations; $d = 503$ covariates) and a lymph node GWAS dataset ($n = 148$, $d = 4514$).

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 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., 2022], and, as evidenced in Sec. 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 better coupling for a given problem at hand.

Avoiding sampling from an asymptotically unbiased Markov chain. Our proposed upper bounds require sampling from a P-invariant Markov chain $(X_t)_{t\geq 0}$. 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) $\mathbb{E}[c(X_t, Y_t')^p] = \mathbb{E}[c(X_t, Y_t)^p] \leq \mathbb{E}[c(Y_t', Y_t)^p]$ 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 $\mathbb{E}[c(X_t, Y_t)^p]^{1/p}$ which is only loose by a constant factor of 2, as $\mathbb{E}[c(Y_t', Y_t)^p]^{1/p} \leq \mathbb{E}[c(X_t, Y_t)^p]^{1/p} + \mathbb{E}[c(X_t, Y_t')^p]^{1/p} = 2\mathbb{E}[c(X_t, Y_t)^p]^{1/p}$. We hope to investigate such coupling constructions in follow-up work.

Upper bounds for total variation distance. The 1-Wasserstein distance with metric $c(x, y) = \mathrm{I}\{x \neq y\}$ gives the popular total variation (TV) distance, 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.

Spot checking. Finally, an anonymous associate editor suggested the following additional application. Often one is interested in approximating an entire family of target distributions P_η with approximations Q_η indexed by a parameter η taking a large number of distinct values in $\mathbb{R}$. When it is feasible to run a P_η-invariant Markov chain only for a small number of η values but infeasible to run these exact chains for all target η values, CUB can be used to spot check Wasserstein quality at a small set of representative η values and drive decision making around the degree or type of

approximation used for the full collection of η values.

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 BayesComp workshop on "Measuring the quality of MCMC output" for helpful feedback. We also thank the anonymous reviewers and associate editor for their valuable comments and suggestions. 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. 3, 14

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. 1, 16

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. 46, 48, 49

N. Biswas, P. E. Jacob, and P. Vanetti. Estimating convergence of Markov chains with L-lag couplings. NeurIPS, pages 7389–7399, 2019. 2, 3, 4, 8, 27, 31

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. *J. R. Stat. Soc. Ser. B Methodol.*, 2022. 3, 4, 17, 19, 46, 47

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. 2, 16

N. Bou-Rabee and M. Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2012. ISSN 0272-4979. doi: 10.1093/imanum/drs003. URL https://doi.org/10.1093/imanum/drs003. 37

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for hamiltonian monte carlo. *Ann. Appl. Probab.*, 30(3):1209–1250, 2020. doi: 10.1214/19-AAP1528. URL https://doi.org/10.1214/19-AAP1528. 9, 12, 52

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011. 1

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. 17, 45

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. 17

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. 16

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. NeurIPS, pages 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf. 3, 14

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, 46

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. 6, 7, 26, 37

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. 12, 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. 12

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. 46

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, 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. 3, 12

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. 46

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, 2020. URL https://proceedings.mlr.press/v108/huggins20a.html. 2, 3, 14, 15

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. 46

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, 11, 19, 27

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. 2, 17, 45, 46

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, 45

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. 46

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. 27

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. 46

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. 2, 18, 19, 46, 50

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. 6, 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, 14

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/. 45

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. 5, 7, 10

G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 13507265. URL http://www.jstor.org/stable/3318418. 25

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. 11, 12, 13, 41

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 2000. 26

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. 16, 45

M. G. Tadesse and M. Vannucci. *Handbook of Bayesian Variable Selection*. Chapman and Hall/CRC, 2021. doi: 10.1201/9781003089018. URL https://doi.org/10.1201/9781003089018. 18

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. 16

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. 2

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. 2, 8

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ 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. 2, 8

C. Villani. *Optimal transport – Old and new*. Springer, 2008. doi: 10.1007/978-3-540-71050-9. 2, 11, 35

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. 9, 19

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. 3, 13

M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. ICML, page 681–688, 2011. ISBN 9781450306195. 1, 13, 16, 37

K. Xu, T. E. Fjelde, C. Sutton, and H. Ge. Couplings for Multinomial Hamiltonian Monte Carlo. AISTATS, pages 3646–3654, 2021. 19

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

$$\mathcal{W}_p(P,Q)^p \leq \mathbb{E}\big[\mathcal{W}_p(\hat{P}_T,\hat{Q}_T)^p\big]$$

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 corresponding estimate of this upper bound, given by

$$\Big(\frac{1}{I}\sum_{i=1}^{I}\mathcal{W}_2(\hat{P}_T^{(i)},\hat{Q}_T^{(i)})^2\Big)^{1/2} \tag{8}$$

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 $\mathcal{W}_2(\hat{P}_T^{(i)},\hat{Q}_T^{(i)})$ is calculated by solving a linear program. The error bands plot 95% confidence intervals given by $\Big[\frac{1}{I}\sum_{i=1}^{I}\mathcal{W}_2(\hat{P}_T^{(i)},\hat{Q}_T^{(i)})^2 \pm 1.96\hat{\sigma}/\sqrt{I}\Big]^{1/2}$ where $\hat{\sigma}^2$ is the empirical variance of $\big(\mathcal{W}_2(\hat{P}_T^{(i)},\hat{Q}_T^{(i)})^2\big)_{i=1}^{I}$.

Instead of (8), one could alternatively use the estimator $\mathcal{W}_2(\hat{P}_{IT},\hat{Q}_{IT})$ where $\hat{P}_{IT}$ and $\hat{Q}_{IT}$ are empirical distribution of P and Q respectively based on IT samples. Using $\mathcal{W}_2(\hat{P}_{IT},\hat{Q}_{IT})$ produces a tighter upper bound estimate compared to using (8), which is linked to consistency of empirical Wasserstein distance based estimates covered in Proposition 3.10 of Section 3.4. However, this numerical improvement is minor; for example in Figure 1 (Left) with dimension $d = 100$, a tighter empirical upper bound of 11.35 is obtained using this estimator compared to the upper bound of 11.83 using (8) and both these upper bound estimates are looser than the coupling based upper bound estimate of 5.78. Such minor numerical improvement is linked to the curse of dimensionality for empirical Wasserstein distances, as discussed in Sections 1.2 and 3.4. Furthermore, calculating $\mathcal{W}_2(\hat{P}_{IT},\hat{Q}_{IT})$ for this example requires approximately 10 times greater numerical runtimes compared to calculating (8).

A.2 Section 2.2 calculations.

As kernel K_1 is P invariant, $X_t \sim P_t \overset{t\to\infty}{\Rightarrow} P$ for all $t \geq 0$ [e.g. Roberts and Tweedie, 1996]. The ULA chain $(Y_t)_{t\geq 0}$ corresponds to an auto-regressive $AR(1)$ model, where

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

for all $t \geq 0$, where $Y_0 \sim \mathcal{N}(0, I_d)$, $Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ and $Z_0 \triangleq Y_0$, and $B = (I_d - (\sigma_Q^2/2)\Sigma^{-1})$. By induction,

$$\begin{aligned} Y_t &= B^t Z_0 + \sigma_Q \Big(B^{t-1} Z_1 + B^{t-2} Z_2 + ... + Z_t \Big) \\ &= B^t Z_0 + \sigma_Q \sum_{j=0}^{t-1} B^j Z_{t-j} \\ &\sim \mathcal{N}\big(0, B^{2t} + \sigma_Q^2 \sum_{j=0}^{t-1} B^{2j}\big) =: Q_t \end{aligned}$$

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), $\lim_{t\to\infty} \big(B^{2t} + \sum_{j=0}^{t-1} B^{2j}\big) = (I_d - B^2)^{-1}$ (see, e.g. Shumway and Stoffer [2000] for sufficient conditions for the convergence $AR(1)$ models). This gives $Q_t \overset{t\to\infty}{\Rightarrow} \mathcal{N}(0, \sigma_Q^2 (I_d - B^2)^{-1}) =: Q$.

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 π 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 $x, y \in \mathbb{R}^d$,*
$$U(x) \leq U(y) + \langle \nabla U(x), y - x\rangle + (m/2)\|x - y\|_2^2$$
3. *U is three times continuously differentiable and there exists some $\tilde{L} > 0$ such that for all $x, y \in \mathbb{R}^d$,*
$$\|\nabla^2 U(x) - \nabla^2 U(y)\|_2 \leq \tilde{L}\|x - y\|_2.$$

Let the step size σ 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 π_γ, and

$$\mathcal{W}_2(\pi, \pi_\gamma)^2 \leq 2\kappa^{-1}\gamma^2 d\Big(2L^2 + \gamma L^4\big(\frac{\gamma}{6} + \frac{1}{m}\big) + \kappa^{-1}\big(\frac{4d\tilde{L}^2}{3} + \gamma L^4 + \frac{4L^4}{3m}\big)\Big) \tag{9}$$

where $\kappa = 2mL/(m + L)$.

The dotted line in Figure 2 is plotted by applying (9) for $\pi = \mathcal{N}(0, \Sigma)$, where $L = \lambda_{min}(\Sigma)^{-1}$, $m = \lambda_{max}(\Sigma)^{-1}$ and $\tilde{L} = 0$. Here $\lambda_{max}(\Sigma)$ and $\lambda_{min}(\Sigma)$ are the largest and smallest eigenvalue of Σ 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 $\bar{K}_1$ on $\mathcal{X}\times\mathcal{X}$ such that $\bar{K}_1((x,y),(\cdot,\mathcal{X})) = K_1(x,\cdot)$ and $\bar{K}_1((x,y),(\mathcal{X},\cdot)) = K_1(y,\cdot)$ for all $x, y \in \mathcal{X}$. Then the *L-lag coupling* chain $(\tilde{X}_{t-L}, X_t)_{t\geq L}$ is generated by sampling X_0 and $\tilde{X}_0$ 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 $(\tilde{X}_{t-L}, X_t)|\tilde{X}_{t-L-1}, X_{t-1} \sim \bar{K}_1((\tilde{X}_{t-L-1}, X_{t-1}),\cdot)$ for $t > L$. Crucially, the joint kernel $\bar{K}_1$ is designed such that: (i) the marginal chains $(\tilde{X}_{t-L})_{t\geq L}$ and $(X_t)_{t\geq 0}$ exactly meet such that the random meeting time $\tau \triangleq \inf\{t > L : \tilde{X}_{t-L} = X_t\}$ is almost surely finite and (ii) the chains remain faithful after meeting such that $\tilde{X}_{t-L} = X_t$ for all $t \geq \tau$. Suppose the coupled chain $(\tilde{X}_{t-L}, X_t)_{t\geq L}$ 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 $t \geq L$, $\mathbb{E}[c(\tilde{X}_{t-L}, X_t)^{2+\eta}] \leq D$ for some constants $\eta > 0$ and $D < \infty$.*

Assumption A.3 (Sub-exponentially tailed meeting times)**.** *The meeting times $\tau \triangleq \inf\{t > L : X_t = \tilde{X}_{t-L}\}$ satisfies $\mathbb{P}(\frac{\tau-L}{L} > t) \leq C\delta^t$ 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

$$\mathcal{W}_1(P_t, P) \leq \sum_{j=1}^{\infty} \mathcal{W}_1(P_{t+jL-L}, P_{t+jL}) \tag{10}$$

$$\leq \sum_{j=1}^{\infty} \mathbb{E}[c(\tilde{X}_{t+jL-L}, X_{t+jL})] \tag{11}$$

$$= \mathbb{E}\Big[\sum_{j=1}^{\infty} c(\tilde{X}_{t+jL-L}, X_{t+jL})\Big] \tag{12}$$

$$= \mathbb{E}\Big[\sum_{j=1}^{\lceil(\tau-L-t)/L\rceil} c(\tilde{X}_{t+jL-L}, X_{t+jL})\Big], \tag{13}$$

where (10) follows from the triangle inequality using Assumption A.2, (11) follows from the coupling representation of the Wasserstein distance, and (12) follows from interchanging the summation and expectation using the dominated convergence theorem under Assumptions A.2 and A.3, and (13) follows as $c(\tilde{X}_{t+jL-L}, X_{t+jL}) = 0$ for all $j > \lceil(\tau - L - t)/L\rceil$ under Assumption A.4. Note that τ has finite expectation under Assumption A.3, which means the upper bound in (13) can be

estimated in finite time. We can estimate this upper bound by simulating multiple L-lag coupled chains $(\tilde{X}_{t-L}, X_t)_{\tau \geq t \geq L}$ independently and using the empirical average

$$\frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{\lceil (\tau^{(i)} - L - t)/L \rceil} c(\tilde{X}^{(i)}_{t+jL-L}, X^{(i)}_{t+jL})$$

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

The following Proposition employs this upper bound alongside CUB_1 (4) to obtain a non-asymptotic upper bound on $\mathcal{W}_1(P, Q)$.

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 $(\tilde{Y}_{t-L}, Y_t)_{t \geq L}$,*

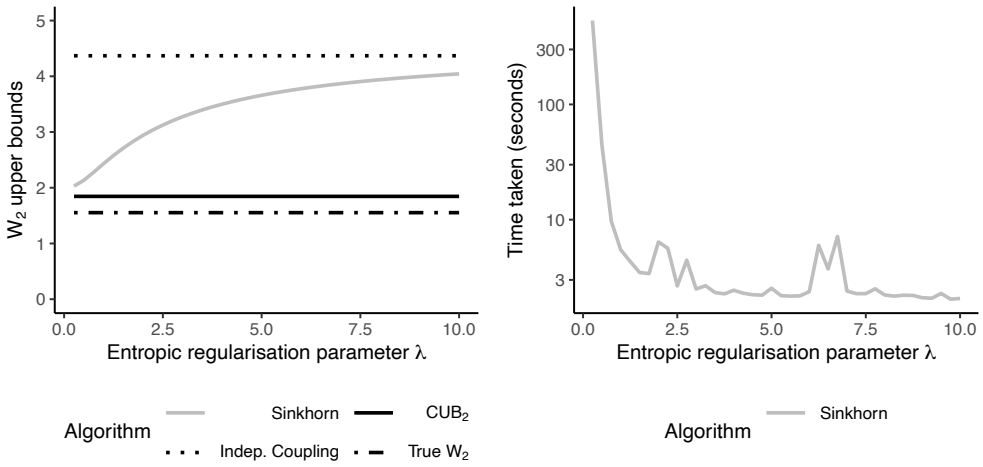
$$\mathcal{W}_1(P, Q) \leq \mathbb{E}[\mathrm{CUB}_{1,t}] + \mathbb{E}[\sum_{j=1}^{\lceil (\tau_P - L - t)/L \rceil} c(\tilde{X}_{t+(j-1)L}, X_{t+jL})] + \mathbb{E}[\sum_{j=1}^{\lceil (\tau_Q - L - t)/L \rceil} c(\tilde{Y}_{t+(j-1)L}, Y_{t+jL})] \tag{14}$$

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

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 (4) 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 indepdendent coupling (black dotted line) are analytically tractable, as given in Section 2.1. For different values of the entropic regularization parameter λ, the grey solid line plots the induced distance of the optimal matching obtained from the Sinkhorn algorithm. For each λ, 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 λ 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 λ. This example illustrates that the Sinkhorn algorithm has expensive runtime precisely for the smaller values of λ that give tighter upper bounds to the Wasserstein distance. In comparison, the CUB_2 (4) estimate does not require solving any expensive optimization problem.



(a) $\mathcal{W}_2$ upper bounds with varying λ.

(b) Sinkhorn runtime with varying λ.

Figure 8: Figure 8a plots upper bound estimates for $\mathcal{W}_2(P, Q)$ with $P = \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq d$, $Q = \mathcal{N}(0, I_d)$, metric $c(x, y) = \|x - y\|_2$ and dimension $d = 10$. Figure 8b plots the runtime of the Sinkhorn algorithm.

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 $a_j \overset{j\to\infty}{\to} 0$, and let $\rho \in (0, 1)$. Then $\sum_{j=1}^{t} \rho^{t-j} a_j \overset{t\to\infty}{\to} 0$.*

Proof of B.1. As $a_j \overset{j\to\infty}{\to} 0$, 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

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

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. □

Lemma B.2. *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 $n \to \infty$, $S_n/n \overset{\text{a.s.},L^1}{\to} \mathbb{E}[\xi_1]$ and for any $p \geq 1$, $(S_n/n)^{1/p} \overset{\text{a.s.},L^1}{\to} \mathbb{E}[\xi_1]^{1/p}$.*

Proof of B.2. As n tends to infinity, $S_n/n \overset{\text{a.s.},L^1}{\to} \mathbb{E}[\xi_1]$ 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]). $(S_n/n)^{1/p} \overset{\text{a.s.}}{\to} (\mathbb{E}[\xi_1])^{1/p}$ follows from $S_n/n \overset{\text{a.s.}}{\to} \mathbb{E}[\xi_1]$ by continuous mapping theorem on $[0,\infty)$. Finally, for $p \geq 1$,

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

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 $S_n/n \overset{L^1}{\to} \mathbb{E}[\xi_1]$. Therefore, $(S_n/n)^{1/p} \overset{L^1}{\to} \mathbb{E}[\xi_1]^{1/p}$. □

Proof of Proposition 3.1: Consistency of instantaneous CUB . Note that $\mathcal{W}_p(P_t, Q_t)$ 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

$$\mathcal{W}_p(P_t, Q_t)^p \leq \mathbb{E}[c(X_t, Y_t)^p] = \mathbb{E}[\text{CUB}^p_{p,t}],$$

where the inequality follows from the coupling representation of Wasserstein distance, and the equality follows from the definition of $\text{CUB}_{p,t}$. As $\mathbb{E}[\text{CUB}^p_{p,t}] < \infty$, by Lemma B.2, $\text{CUB}^p_{p,t} \overset{\text{a.s.}, L^1}{\to} \mathbb{E}[\text{CUB}^p_{p,t}]$ as $I \to \infty$. □

Proof of Corollary 3.2: Consistency of CUB for time-averaged marginals. We first show that

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

Let γ_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 $(X^*, Y^*)|U^* = t \sim \gamma_t$ for $U^* \sim \text{Uniform}(\{S+1, \ldots, T\})$. 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

$$\begin{aligned}
\mathcal{W}_p\Big(\frac{1}{T-S}\sum_{t=S+1}^{T} P_t, \frac{1}{T-S}\sum_{t=S+1}^{T} Q_t\Big)^p &\leq \mathbb{E}[c(X^*, Y^*)^p] \text{ by the coupling representation of } \mathcal{W}_p \\
&= \frac{1}{T-S}\sum_{t=S+1}^{T} \mathbb{E}[c(X^*, Y^*)^p | U^* = t] \\
&= \frac{1}{T-S}\sum_{t=S+1}^{T} \mathcal{W}_p(P_t, Q_t)^p.
\end{aligned}$$

Now by Proposition 3.1 and definition (4),

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

As $\mathbb{E}[\text{CUB}^p_p] < \infty$, by Lemma B.2 $\text{CUB}^p_p \overset{\text{a.s.}, L^1}{\to} \mathbb{E}[\text{CUB}^p_p]$ as $I \to \infty$. □

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,

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

where the first equality follows as $P_t = P$ and $Q_t = Q$ for all $t \ge 0$, the inequality follows Proposition 3.1, and the last equality follows from the definition of CUB_p. As $\mathbb{E}[\mathrm{CUB}_p^p] < \infty$, by Lemma B.2 $\mathrm{CUB}_p^p \overset{\text{a.s., } L^1}{\to} \mathbb{E}[\mathrm{CUB}_p^p]$ as $I \to \infty$. □

Proof of Proposition 3.5: Consistency when chain marginals converge. Let $(P_t)_{t\ge0}$ and $(Q_t)_{t\ge0}$ denote the marginal distributions of Markov chains $(X_t)_{t\ge0}$ and $(Y_t)_{t\ge0}$ respectively. By Assumption 3.4, distributions $(P_t)_{t\ge0}$, $(Q_t)_{t\ge0}$, P and Q all have finite moments of order p. Then for all $t \ge 1$,

$$\mathcal{W}_p(P,Q) \le \mathcal{W}_p(P,P_t) + \mathcal{W}_p(P_t,Q_t) + \mathcal{W}_p(Q_t,Q) \tag{15}$$
$$\le \mathcal{W}_p(P,P_t) + \mathbb{E}[c(X_t,Y_t)^p]^{1/p} + \mathcal{W}_p(Q_t,Q), \tag{16}$$

where (15) follows by the triangle inequality as $\mathcal{W}_p$ is a metric on the space of measure on $\mathcal{X}$ with finite moments of order p , and (16) follows from the coupling representation of $\mathcal{W}_p$. By Assumption 3.4, $\lim_{t\to\infty} \mathcal{W}_p(P,P_t) = 0$ and $\lim_{t\to\infty} \mathcal{W}_p(Q_t,Q) = 0$. Taking the limit infimum in (16) and raising to the p^{th} exponent gives $\mathcal{W}_p(P,Q)^p \le \liminf_{t\to\infty} \mathbb{E}[c(X_t,Y_t)^p]$. Therefore for all $\epsilon > 0$, there exists $S \ge 1$ such that for all $t \ge S$, $\mathcal{W}_p(P,Q)^p \le \epsilon + \mathbb{E}[c(X_t,Y_t)^p]$, and

$$\mathcal{W}_p(P,Q)^p \le \epsilon + \frac{1}{T-S}\sum_{t=S+1}^{T} \mathbb{E}[c(X_t,Y_t)^p] = \epsilon + \mathbb{E}[\mathrm{CUB}_p^p]$$

for all $T \ge S$. As $\mathbb{E}[\mathrm{CUB}_p^p]^p < \infty$, by Lemma B.2 $\mathrm{CUB}_p^p \overset{\text{a.s., } L^1}{\to} \mathbb{E}[\mathrm{CUB}_p^p]$ as $I \to \infty$. □

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

$$\mathcal{W}_1(P,Q) \le \mathcal{W}_1(P_t,Q_t) + \mathcal{W}_1(P_t,P) + \mathcal{W}_1(P_t,P).$$

By Proposition 3.1, $\mathcal{W}_1(P_t,Q_t) \le \mathbb{E}[\mathrm{CUB}_{1,t}]$. Under assumptions A.2, A.3 and A.4, by Biswas et al. [2019, Theorem 2.5]

$$\mathcal{W}_1(P_t,P) \le \mathbb{E}\Big[\sum_{j=1}^{\lceil(\tau_P-L-t)/L\rceil} c(\tilde{X}_{t+(j-1)L}, X_{t+jL})\Big] \text{ and}$$
$$\mathcal{W}_1(Q_t,Q) \le \mathbb{E}\Big[\sum_{j=1}^{\lceil(\tau_Q-L-t)/L\rceil} c(\tilde{Y}_{t+(j-1)L}, Y_{t+jL})\Big].$$

Equation (14) now directly follows. As the meeting times τ_P and τ_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 $\bar{K}$ 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 $(Z_t, Y_t)|X_{t-1}, Y_{t-1} \sim \Gamma_\Delta(Y_{t-1})$. This gives

$$\begin{aligned}
\mathbb{E}[c(X_t, Y_t)^p]^{1/p} &= \mathbb{E}[\mathbb{E}[c(X_t, Y_t)^p | X_{t-1}, Y_{t-1}]]^{1/p} \\
&\leq \mathbb{E}[\mathbb{E}[\big(c(X_t, Z_t) + c(Z_t, Y_t)\big)^p | X_{t-1}, Y_{t-1}]]^{1/p} && (17) \\
&\leq \mathbb{E}[\mathbb{E}[c(X_t, Z_t)^p | X_{t-1}, Y_{t-1}]]^{1/p} + \mathbb{E}[\mathbb{E}[c(Z_t, Y_t)^p | X_{t-1}, Y_{t-1}]]^{1/p} && (18) \\
&\leq \rho \mathbb{E}[c(X_{t-1}, Y_{t-1})^p]^{1/p} + \mathbb{E}[\Delta_p(Y_{t-1})]^{1/p} && (19)
\end{aligned}$$

where (17) follows as c is a metric, (18) follows by Minkowski's inequality, and (19) 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, (19) implies

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

□

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,

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

□

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

$$\mathbb{E}[V(Y_i)^p] \leq \gamma^i \mathbb{E}[V(Y_0)^p] + (1 - \gamma^i) \frac{L}{1 - \gamma} \text{ for all } i \geq 0. \tag{20}$$

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 δ, second inequality from (20), and the second inequality from the definition of κ. By Theorem 3.7, we obtain

$$\begin{aligned}
\mathbb{E}[c(X_t, Y_t)^p]^{1/p} &\leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} \mathbb{E}\Big[\Delta_p(Y_{i-1})\Big]^{1/p} \\
&\leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \delta\kappa \sum_{i=1}^{t} \rho^{t-i} \\
&= \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1-\rho^t)\frac{\delta\kappa}{1-\rho}.
\end{aligned}$$

□

B.3 Wasserstein distances of empirical distributions proofs

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

Lemma B.3. *Suppose S and T are distributions on the metric space $(\mathcal{X}, c)$ with finite moments of order p, and $n \geq 1$ is an integer. 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,*

$$\mathcal{W}_p(S, T)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{S}_n, T)^p].$$

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,

$$\begin{aligned}
\mathcal{W}_p(S, T)^p &\leq \mathbb{E}[c(U, V)^p] \text{ by the coupling representation of Wasserstein distance} \\
&= \mathbb{E}[\mathbb{E}[c(U, V)^p | U_1, \ldots U_n]] \\
&= \mathbb{E}[\mathcal{W}_p(\hat{S}_n, T)^p].
\end{aligned}$$

□

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.3 with $S = P$, $U_i = X_i$ and $T = Q$,

$$\mathcal{W}_p(P, Q)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, Q)^p].$$

As $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$ are independent, $Y_i|(X_1, ..., X_n) \sim Y_i \sim Q$ for all $i = 1, ..., n$. We can therefore apply Lemma B.3 conditional on $(X_1, ..., X_n)$ now with $S = Q$, $U_i = Y_i$ and $T = \hat{P}_n$ to obtain

$$\mathcal{W}_p(\hat{P}_n, Q)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p|X_1, ..., X_n]$$

almost surely for all $X_1, ..., X_n$. Overall, this gives

$$\mathcal{W}_p(P, Q)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, Q)^p] \leq \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p|X_1, ..., X_n]] = \mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p]$$

as required.

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

$$\mathcal{W}_p(\hat{P}_n, \hat{Q}_n) \leq \mathcal{W}_p(\hat{P}_n, P) + \mathcal{W}_p(P, Q) + \mathcal{W}_p(Q, \hat{Q}_n).$$

By Minkowski's inequality, this gives

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p]^{1/p} &\leq \mathbb{E}\Big[\Big(\mathcal{W}_p(\hat{P}_n, P) + \mathcal{W}_p(P, Q) + \mathcal{W}_p(Q, \hat{Q}_n)\Big)^p\Big]^{1/p} \\
&\leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, P)^p]^{1/p} + \mathbb{E}[\mathcal{W}_p(P, Q)^p]^{1/p} + \mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_n)^p]^{1/p} \\
&= \mathbb{E}[\mathcal{W}_p(\hat{P}_n, P)^p]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_n)^p]^{1/p} \qquad (21)
\end{aligned}$$

Let $\tilde{P}_n$ denote empirical distributions of the samples $(\tilde{X}_1, ..., \tilde{X}_n)$, where $\tilde{X}_i \sim P$ for all $i = 1, ..., n$ and $(\tilde{X}_1, ..., \tilde{X}_n)$ and $(X_1, ..., X_n)$ are independent. Independence implies $\tilde{X}_i|(X_1, ..., X_n) \sim \tilde{X}_i \sim P$ for all $i = 1, ..., n$. We can therefore apply Lemma B.3 conditional on $(X_1, ..., X_n)$, with $S = P$, $T = \hat{P}_n$ and $\tilde{X}_i = U_i$ to obtain

$$\mathcal{W}_p(\hat{P}_n, P)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_n, \tilde{P}_n)^p|X_1, ..., X_n].$$

Similarly,

$$\mathcal{W}_p(Q, \hat{Q}_n)^p \leq \mathbb{E}[\mathcal{W}_p(\tilde{Q}_n, \hat{Q}_n)^p|Y_1, ..., Y_n]$$

where $\tilde{Q}_n$ denotes empirical distributions of the samples $(\tilde{Y}_1, ..., \tilde{Y}_n)$, where $\tilde{Y}_i \sim Q$ for all $i = 1, ..., n$

and $(\tilde{Y}_1, ..., \tilde{Y}_n)$ and $(Y_1, ..., Y_n)$ are independent. By (21), we obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_p(\hat{P}_n, \hat{Q}_n)^p]^{1/p} \leq & \mathbb{E}[\mathcal{W}_p(\hat{P}_n, P)^p]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_n)^p]^{1/p} \\
= & \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\hat{P}_n, P)^p | X_1, ..., X_n]]^{1/p} + \mathcal{W}_p(P, Q) + \\
& \mathbb{E}[\mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_n)^p | Y_1, ..., Y_n]]^{1/p} \\
\leq & \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\hat{P}_n, \tilde{P}_n)^p | X_1, ..., X_n]]^{1/p} + \mathcal{W}_p(P, Q) + \\
& \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\tilde{Q}_n, \hat{Q}_n)^p | Y_1, ..., Y_n]]^{1/p} \\
= & \mathbb{E}[\mathcal{W}_p(\hat{P}_n, P_n)^p]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathcal{W}_p(\tilde{Q}_n, \hat{Q}_n)^p]^{1/p}
\end{aligned}$$

as required.

Consistency. By triangle inequality,

$$|\mathcal{W}_p(\hat{P}_n, \hat{Q}_n) - \mathcal{W}_p(P, Q)| \leq \mathcal{W}_p(\hat{P}_n, P) + \mathcal{W}_p(Q, \hat{Q}_n).$$

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], $\mathcal{W}_p(\hat{P}_n, P) \overset{\text{a.s.}}{\to} 0$ and $\mathcal{W}_p(Q, \hat{Q}_n) \overset{\text{a.s.}}{\to} 0$ 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 is a compact subset Ω 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 Γ_1 of the kernel K_1 such that for some constant $\alpha_\Omega \in (0, 1)$ and all $X_t, X'_t \in \Omega$, $\mathbb{E}[c(X_{t+1}, X'_{t+1})|X_t, X'_t] \leq \alpha_\Omega c(X_t, X'_t)$ 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

$$\mathcal{W}_1(P, Q) \leq \frac{\mathbb{E}[\mathbb{E}[c(X_1, Y_1)|Y^*]] + 2\epsilon}{1 - \alpha_\Omega} \tag{22}$$

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 $\mathbb{E}[\mathbb{E}[c(X_1, Y_1)|X^*]]$ and α_Ω separately using couplings to obtain a final upper bound estimate.

Given this setup, we can show that our upper bound estimator CUB_1 (4) constructed using such couplings Γ_1 and Γ_Δ has a smaller expected value than the upper bound of (22).

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 (4) 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 $(Z_t, Y_t)|X_{t-1}, Y_{t-1} \sim \Gamma_\Delta(Y_{t-1})$. This gives

$$\begin{aligned}
\mathbb{E}[c(X_t,Y_t)] \leq& \mathbb{E}[c(X_t,Z_t)] + \mathbb{E}[c(Z_t,Y_t)] && (23)\\
=& \mathbb{E}[c(X_t,Z_t)\mathrm{I}_{\{X_{t-1}\in\Omega, Y_{t-1}\in\Omega\}^c}] + \mathbb{E}[c(X_t,Z_t)\mathrm{I}_{\{X_{t-1}\in\Omega, Y_{t-1}\in\Omega\}}] + \mathbb{E}[c(Z_t,Y_t)]\\
\leq& \mathbb{P}(\{X_{t-1}\in\Omega^c\}\cup\{Y_{t-1}\in\Omega^c\}) + \mathbb{E}[c(X_t,Z_t)\mathrm{I}_{\{X_{t-1}\in\Omega, Y_{t-1}\in\Omega\}}] +\\
& \mathbb{E}[c(Z_t,Y_t)] && (24)\\
\leq& \mathbb{P}(X_{t-1}\in\Omega^c) + \mathbb{P}(Y_{t-1}\in\Omega^c) + \alpha_\Omega \mathbb{E}[c(X_{t-1},Y_{t-1})] + \mathbb{E}[c(Z_t,Y_t)] && (25)
\end{aligned}$$

where (23) follows by the triangle inequality, (24) follows as c is bounded by 1, and (25) follows by the union bound and the definition of α_Ω. 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, (25) implies

$$\mathbb{E}[c(X_t,Y_t)] \leq \alpha_\Omega^t \mathbb{E}[c(X_0,Y_0)] + \sum_{i=1}^{t} \alpha_\Omega^{t-i}\Big(\mathbb{P}(X_{t-1}\in\Omega^c) + \mathbb{P}(Y_{t-1}\in\Omega^c) + \mathbb{E}[\Delta(Y_{i-1})]\Big).$$

As X_{t-1} and Y_{t-1} converges to P and Q respectively in distribution, $\mathbb{P}(X_{t-1}\in\Omega^c) \overset{t\to\infty}{\to} P(\Omega^c) < \epsilon$, $\mathbb{P}(Y_{t-1}\in\Omega^c) \overset{t\to\infty}{\to} Q(\Omega^c) < \epsilon$ and $\mathbb{E}[\Delta(Y_t)] \overset{t\to\infty}{\to} \mathbb{E}[\Delta(Y^*)]$ 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 $t \geq S$,

$$\mathbb{E}[c(X_t,Y_t)] \leq \alpha_\Omega^t \mathbb{E}[c(X_0,Y_0)] + \sum_{i=1}^{t} \alpha_\Omega^{t-i}\Big(\mathbb{E}[\Delta(Y^*)] + 2\epsilon\Big) + \epsilon'.$$

Therefore as $\alpha_\Omega \in (0,1)$, $\liminf_{t\to\infty} \mathbb{E}[c(X_t,Y_t)] \leq \frac{\mathbb{E}[\Delta(Y^*)]+2\epsilon}{1-\alpha_\Omega}$ where $\Delta(Y^*) = \mathbb{E}[c(X_1,Y_1)|Y^*]$ and $Y^* \sim Q$ from (22) 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 σ 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 σ sufficiently small, the marginal ULA chain $(X_t)_{t\geq 0}$ converges to some

distribution P_σ and satisfies Assumption 3.6 for $p = 2$ under a CRN coupling [Durmus and Moulines, 2019, Proposition 3], giving a contraction rate ρ 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,

$$\mathcal{W}_2(P_\sigma, P) \leq \liminf_{t\to\infty} \mathbb{E}[\mathrm{CUB}_{2,t}^2]^{1/2} \leq \frac{\mathbb{E}\big[\|Y - Y'\|^2\big(1 - \alpha_\sigma\big(Y, Y'\big)\big)\big]^{1/2}}{C\sigma^2/2}, \tag{26}$$

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 $\alpha_\sigma\big(Y, Y'\big) \in [0, 1]$ is the Metropolis–Hastings acceptance probability. As the step size σ tends to zero, the upper bound in (26) 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 σ 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 σ sufficiently small, the marginal chains $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ converge to some distributions P_σ and Q_σ respectively. Both marginal chains also satisfy Assumption 3.6 for $p = 2$ under a CRN coupling, with contraction rates ρ_P and ρ_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

$$\mathcal{W}_2(P_\sigma, Q_\sigma) \leq \liminf_{t\to\infty} \mathbb{E}[\mathrm{CUB}_2(P_t, Q_t)^2]^{1/2} \leq \frac{\mathbb{E}\big[\|\nabla \log P(Y_\sigma) - \nabla \log Q(Y_\sigma)\|^2\big]^{1/2}}{C_P} \tag{27}$$

where $Y \sim Q_\sigma$. By symmetry, we can obtain a similar bound in terms of some random variable $X \sim P_\sigma$ and C_Q. As σ approaches zero, the numerator in (27) 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 (27) 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 σ 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 $\widehat{\nabla \log P}_{SGLD}(z) = \nabla \log P(z) + e_{SGLD}(z)$ for all $z \in \mathcal{X}$, where $e_{SGLD}(z)$ is mean zero error. We assume this error is bounded such that $\delta^2 \triangleq \sup_{z\in\mathcal{X}} e_{SGLD}(z)/(1 + V(z)^2) < \infty$, 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 σ 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 ρ such that $1-\rho = C\sigma^2/2$ for constants C that does not explicitly depend on the dimension. Then by Proposition 3.9,

$$\limsup_{t\to\infty} \mathcal{W}_2\big(P_t^{(ULA)}, P_t^{(SGLD)}\big) \leq \liminf_{t\to\infty} \mathbb{E}\Big[\mathrm{CUB}_2\big(P_t^{(ULA)}, Q_t^{(ULA)}\big)^2\Big]^{1/2} \leq \frac{\delta\kappa}{C}. \tag{28}$$

Note that the upper bound in (28) does not explicitly depend on dimension, and approaches zero as δ 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 the 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_P^L and K_Q^L respectively. Following (3) 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 $A \in \mathcal{B}(\mathcal{X})$,

$$\bar{K}_{L-step}\big((x,y),(A,\mathcal{X})\big) = K_P^L(x,A) \text{ and } \bar{K}_{L-step}\big((x,y),(\mathcal{X},A)\big) = K_Q^L(y,A). \tag{29}$$

Given coupled kernels Γ_1 and Γ_Δ, Figure 9 illustrates how to sample from the joint kernel $\bar{K}_{L-step}$. By construction, this gives the marginal distributions $X_s|X_0, Y_0 \sim K_P^s(X_0,\cdot)$ and $Y_s|X_0,Y_0 \sim K_Q^s(Y_0,\cdot)$ for all $s = 1,\ldots,L$, such that Equation (29) 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

$$\begin{aligned} X_{Lt}|X_{L(t-1)}, Y_{L(t-1)} &\sim K_P^L(X_{L(t-1)},\cdot) \\ Z_L^{(j)}|Y_{L(t-1)+(j-1)} &\sim K_P^{L-(j-1)}(Y_{L(t-1)+(j-1)},\cdot) \\ Y_{Lt}|X_{L(t-1)}, Y_{L(t-1)} &\sim K_Q^L(Y_{L(t-1)},\cdot) \end{aligned}$$

for $s = 1,\ldots,L-1$. When $L = 1$, we obtain $\bar{K}_{L-step} = \bar{K}$ 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.

We give concrete implementations of Algorithm 3 for the ULA and MALA Markov chain with common random numbers and reflection couplings. These are based on common random numbers

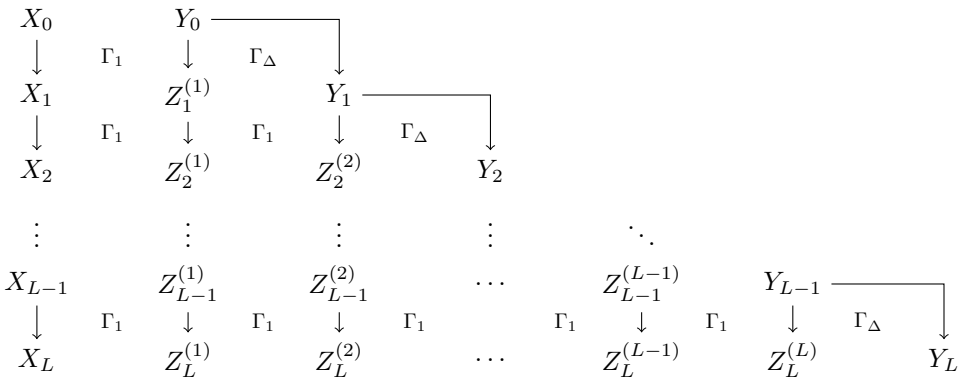


Figure 9: Joint kernel $\bar{K}_{L-step}$ on $\mathcal{X} \times \mathcal{X}$, which couples the marginal kernels K_P^L and K_Q^L

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

Input: chain states X_0 and Y_0, kernels K_1 and K_2, coupled kernels Γ_1 and Γ_Δ

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

$$(X_s, Z_s^{(1)}) \sim \Gamma_1(X_{s-1}, Z_{s-1}^{(1)}) \tag{30}$$

$$(Z_s^{(j)}, Z_s^{(j+1)}) \sim \Gamma_1(Z_{s-1}^{(j)}, Z_{s-1}^{(j+1)}) \text{ for } j = 1, ..., s-1 \tag{31}$$

$$(Z_s^{(s)}, Y_s) \sim \Gamma_\Delta(Y_{s-1}) \tag{32}$$

end

return $(X_{L(t-1)+s}, Y_{L(t-1)+s})$ for $s = 1, ..., L$.

and reflection couplings for the single-step coupling kernels included in Appendices F.

ULA with common random numbers coupling. For each $s = 1, \ldots, L$ in Algorithm 3, sample $\epsilon_s \sim \mathcal{N}(0, I_d)$ and

- Sample $(X_s, Z_s^{(1)}) \sim \Gamma_1(X_{s-1}, Z_{s-1}^{(1)})$ in (30) such that $X_s = X_{s-1} + \frac{1}{2}\sigma_P^2 \nabla \log p(X_{s-1}) + \sigma_P \epsilon_s$ and $Z_s^{(1)} = Z_{s-1}^{(1)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(1)}) + \sigma_P \epsilon_s$.
- Sample $(Z_s^{(j)}, Z_s^{(j+1)}) \sim \Gamma_1(Z_{s-1}^{(j)}, Z_{s-1}^{(j+1)})$ for each $j = 1, ..., s-1$ in (31) such that $Z_s^{(j)} = Z_{s-1}^{(j)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(j)}) + \sigma_P \epsilon_s$ and $Z_s^{(j+1)} = Z_{s-1}^{(j+1)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(j+1)}) + \sigma_P \epsilon_s$.
- Sample $(Z_s^{(s)}, Y_s) \sim \Gamma_\Delta(Y_{s-1})$ in (32) such that $Z_s^{(s)} = Z_{s-1}^{(s)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(s)}) + \sigma_P \epsilon_s$ and $Y_s = Y_{s-1} + \frac{1}{2}\sigma_Q^2 \nabla \log p(Y_{s-1}) + \sigma_Q \epsilon_s$.

MALA with common random numbers coupling. For $s = 1, \ldots, L$ in Algorithm 3, sample $\epsilon_s \sim \mathcal{N}(0, I_d)$ and generate proposals $X_s^*, Z_s^{(1),*}, \ldots, Z_s^{(s),*}, Y_s^*$ using the steps for ULA with common random numbers coupling given above. Then sample $U^{(s)} \sim \text{Uniform}([0,1])$ and accept each of these proposals if $U^{(s)}$ is less than the respective Metropolis-Hastings acceptance probabilities.

ULA with reflection coupling. For each $s = 1, \ldots, L$ in Algorithm 3, sample $\epsilon_s \sim \mathcal{N}(0, I_d)$ and

- Sample $(X_s, Z_s^{(1)}) \sim \Gamma_1(X_{s-1}, Z_{s-1}^{(1)})$ in (30) such that $X_s = X_{s-1} + \frac{1}{2}\sigma_P^2 \nabla \log p(X_{s-1}) + \sigma_P \epsilon_s$ and $Z_s^{(1)} = Z_{s-1}^{(1)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(1)}) + \sigma_P (I_d - e^{(1)} e^{(1)\top}) \epsilon_s$ for $e^{(1)} = \frac{X_{s-1} - Z_{s-1}^{(1)}}{\|X_{s-1} - Z_{s-1}^{(1)}\|}$.
- Sample $(Z_s^{(j)}, Z_s^{(j+1)}) \sim \Gamma_1(Z_{s-1}^{(j)}, Z_{s-1}^{(j+1)})$ for each $j = 1, ..., s-1$ in (31) such that $Z_s^{(j)} = Z_{s-1}^{(j)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(j)}) + \sigma_P \epsilon_s$ and $Z_s^{(j+1)} = Z_{s-1}^{(j+1)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(j+1)}) + \sigma_P (I_d - e^{(j+1)} e^{(j+1)\top}) \epsilon_s$ for $e^{(j+1)} = \frac{Z_{s-1}^{(j)} - Z_{s-1}^{(j+1)}}{\|Z_{s-1}^{(j)} - Z_{s-1}^{(j+1)}\|}$.
- Sample $(Z_s^{(s)}, Y_s) \sim \Gamma_\Delta(Y_{s-1})$ in (32) such that $Z_s^{(s)} = Z_{s-1}^{(s)} + \frac{1}{2}\sigma_P^2 \nabla \log q(Z_{s-1}^{(s)}) + \sigma_P \epsilon_s$ and $Y_s = Y_{s-1} + \frac{1}{2}\sigma_Q^2 \nabla \log p(Y_{s-1}) + \sigma_Q (I_d - e^{(s+1)} e^{(s+1)\top}) \epsilon_s$ for $e^{(s+1)} = \frac{Z_{s-1}^{(s)} - Y_{s-1}}{\|Z_{s-1}^{(s)} - Y_{s-1}\|}$.

MALA with reflection coupling. For $s = 1, \ldots, L$ in Algorithm 3, sample sample $\epsilon_s \sim \mathcal{N}(0, I_d)$ and generate proposals $X_s^*, Z_s^{(1),*}, \ldots, Z_s^{(s),*}, Y_s^*$ using the steps for ULA with reflection coupling given above. Then sample $U^{(s)} \sim \text{Uniform}([0,1])$ and accept each of these proposals if $U^{(s)}$ is less than the respective Metropolis-Hastings acceptance probabilities.

Having developed algorithms to sample from the coupled kernels $\bar{K}$ and $\bar{K}_{L-step}$, we now investigate theoretical properties our upper bounds.

D.2 Theoretical properties of couplings of multi-step kernels

To establish theoretical guarantees of coupled Markov chains based on the coupled kernel $\bar{K}_{L-step}$, we assume the Markovian coupling Γ_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 $L \geq 1$,*

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

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 $\bar{K}_{L-step}$ explicitly in terms of the initial distribution $\bar{I}_0$ and the coupled kernel Γ_Δ 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 $\bar{K}_{L-step}$ given in Figure 9 and Algorithm 3. When the coupled kernel Γ_Δ 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 Γ_1 satisfies Assumption D.1 for some $C \geq 1$ and $\rho < 1$. Fix some $L \geq 1$ such that $\tilde{\rho} = C\rho^L < 1$, and consider the coupled chain $(X_t, Y_t)_{t \geq 0}$ generated using Algorithm 3 with the L-step coupled kernel $\bar{K}_{L-step}$. Then for all $t \geq 0$,*

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

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 $t \geq S$,*

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

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*

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

for all $z \in \mathcal{X}$, where $\gamma \in [0,1)$ and $L \in [0,\infty)$ are constants. Define

$$\delta \triangleq \sup_{z\in\mathcal{X}} \left(\frac{\Delta_p(z)}{1+V(z)^p}\right)^{1/p} \qquad \kappa \triangleq 1 + \max\left\{\mathbb{E}[V(Y_0)^p]^{1/p}, \left(\frac{L}{1-\gamma}\right)^{1/p}\right\}\Big).$$

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

$$\mathbb{E}[\mathrm{CUB}_{p,t}^p]^{1/p} = \mathbb{E}[c(X_t,Y_t)^p]^{1/p} \leq (C\rho^L)^t \mathbb{E}[c(X_0,Y_0)^p]^{1/p} + C\Big(\frac{1-(C\rho^L)^t}{1-C\rho^L}\Big)\Big(\frac{1-\rho^L}{1-\rho}\Big)\delta\kappa.$$

D.3 Proofs

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

$$(X_{Lt}, Z_L^{(1)}, ..., Z_L^{(L)}, Y_{Lt})$$

where

$$\begin{aligned}
(X_{Lt}, Z_L^{(1)})|X_{L(t-1)}, Y_{L(t-1)} &\sim \Gamma_P^L(X_{L(t-1)}, Y_{L(t-1)}) \\
(Z_L^{(j)}, Z_L^{(j+1)})|Y_{L(t-1)+j-1} &\sim \Gamma_\Delta(Y_{L(t-1)+j-1})\Gamma_1^{L-j} \text{ for } j = 1, ..., L-1 \\
(Z_L^{(L)}, Y_{Lt})|Y_{L(t-1)+L-1} &\sim \Gamma_\Delta(Y_{L(t-1)+L-1}).
\end{aligned}$$

As $(X_{Lt}, Z_t^{(0)})|X_{L(t-1)}, Y_{L(t-1)} \sim \Gamma_1^L(X_{L(t-1)}, Y_{L(t-1)})$, we obtain

$$\begin{aligned}
\mathbb{E}[c(X_{Lt},Y_{Lt})^p]^{1/p} =& \mathbb{E}[\mathbb{E}[c(X_{Lt},Y_{Lt})^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} \\
\leq& \mathbb{E}[\mathbb{E}[\big(c(X_{Lt},Z_L^{(1)}) + c(Z_L^{(1)},Y_{Lt})\big)^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} \qquad (33)\\
\leq& \mathbb{E}[\mathbb{E}[c(X_{Lt},Z_L^{(1)})^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} + \\
& \mathbb{E}[\mathbb{E}[c(Z_L^{(1)},Y_{Lt})^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} \qquad (34)\\
\leq& \tilde{\rho}\mathbb{E}[c(X_{L(t-1)},Y_{L(t-1)})^p]^{1/p} + \mathbb{E}[c(Z_L^{(1)},Y_{Lt})^p]^{1/p} \qquad (35)
\end{aligned}$$

where (33) follows as c is a metric, (34) follows by Minkowski's inequality, and (35) 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,

$$\begin{aligned}
\mathbb{E}[c(Z_L^{(1)}, Y_{Lt})^p]^{1/p} &\leq \mathbb{E}\Big[\Big(c(Z_L^{(L)}, Y_{Lt}) + \sum_{j=1}^{L-1} c(Z_L^{(j)}, Z_L^{(j+1)})\Big)^p\Big]^{1/p} \text{ as } c \text{ is a metric} \\
&\leq \mathbb{E}\Big[c(Z_L^{(L)}, Y_{Lt})^p\Big]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\Big[c(Z_L^{(j)}, Z_L^{(j+1)})^p\Big]^{1/p} \text{ by Minkowski's inequality} \\
&= \mathbb{E}\Big[\mathbb{E}\Big[c(Z_L^{(L)}, Y_{Lt})^p | Y_{L(t-1)+L-1}\Big]\Big]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\Big[\mathbb{E}\Big[c(Z_L^{(j)}, Z_L^{(j+1)})^p | Y_{L(t-1)+j-1}\Big]\Big]^{1/p} \\
&= \mathbb{E}[\Delta_p(Y_{L(t-1)+(L-1)})]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\Big[\mathbb{E}\Big[c(Z_L^{(j)}, Z_L^{(j+1)})^p | Y_{L(t-1)+j-1}\Big]\Big]^{1/p} \\
&\leq \mathbb{E}[\Delta_p(Y_{L(t-1)+(L-1)})]^{1/p} + \sum_{j=1}^{L-1} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(t-1)+j-1})\Big]^{1/p} \text{ by Assumption D.1} \\
&\leq \sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(t-1)+j})\Big]^{1/p} \text{as } C \geq 1.
\end{aligned}$$

Equation (35) now gives

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \tilde{\rho}\mathbb{E}[c(X_{L(t-1)}, Y_{L(t-1)})^p]^{1/p} + \sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(t-1)+j})\Big]^{1/p} \tag{36}$$

By induction, (36) implies

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

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 $a_k \overset{k\to\infty}{\to} a$, because Q_t converges in p-Wasserstein distance to Q as $t \to \infty$. This implies

$$\sum_{i=1}^{t} \tilde{\rho}^{t-i}\Big(\sum_{j=1}^{L} C\rho^{L-j} a_{L(i-1)+j}\Big) \overset{t\to\infty}{\to} \sum_{i=1}^{t} \tilde{\rho}^{t-i}\Big(\sum_{j=1}^{L} C\rho^{L-j} a\Big).$$

Therefore, for all $\epsilon > 0$ there exists $S \geq 1$ such that for all $t \geq S$, $\sum_{i=1}^{t} \tilde{\rho}^{t-i} \sum_{j=1}^{L} C\rho^{L-j}|a_{L(i-1)+j} -$

$a| < \epsilon$. By Theorem D.2,

$$\begin{aligned}
\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} &\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} a_{L(i-1)+j} \Big) \\
&\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \sum_{j=1}^{L} C\rho^{L-j} a + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} |a_{L(i-1)+j} - a| \Big) \\
&\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \sum_{j=1}^{L} C\rho^{L-j} a + \epsilon \\
&= (C\rho^L)^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + C\Big(\frac{1 - (C\rho^L)^t}{1 - C\rho^L} \Big) \Big(\frac{1 - \rho^L}{1 - \rho} \Big) a + \epsilon
\end{aligned}$$

as required. □

Proof of Proposition D.4. As V is a a p^{th}-order Lyapunov function of K_2, by induction

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

for all $j \geq 0$. This gives

$$\begin{aligned}
\mathbb{E}\Big[\Delta_p(Y_j)\Big]^{1/p} &\leq \delta \mathbb{E}[1 + V(Y_{j-1})^p]^{1/p} \\
&\leq \delta(1 + \mathbb{E}[V(Y_{j-1})^p]^{1/p}) \\
&\leq \delta\left(1 + \left(\gamma^{t-1} \mathbb{E}[V(Y_0)^p] + (1 - \gamma^{t-1}) \frac{L}{1 - \gamma}\right)^{1/p}\right) \\
&\leq \delta\left(1 + \max\left\{\mathbb{E}[V(Y_0)^p]^{1/p}, \left(\frac{L}{1 - \gamma}\right)^{1/p}\right\}\right) \\
&= \delta\kappa
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} &\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} \mathbb{E}\Big[\Delta_p(Y_{L(i-1)+j})\Big]^{1/p} \Big) \\
&\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} \delta\kappa \Big) \\
&\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + C\Big(\frac{1 - (C\rho^L)^t}{1 - C\rho^L} \Big) \Big(\frac{1 - \rho^L}{1 - \rho} \Big) \delta\kappa
\end{aligned}$$

□

E 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 (4) 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 (7) 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 $[X]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $y \sim \mathcal{N}(X\beta_*, \sigma_*^2 I_n)$, 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 $\mathcal{O}(n^2 d)$ 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 ν=1), approximate MCMC methods have been developed by Johndrow et al.

[2020] based on approximations of the form

$$X \operatorname{Diag}(\xi\eta_t)^{-1}X^\top \approx X \operatorname{Diag}((\xi^{-1}\eta_j^{-1}\mathrm{I}_{\{\xi^{-1}\eta_j^{-1}>\epsilon\}})_{j=1}^p)\, X^\top \tag{37}$$

for some small threshold $\epsilon > 0$. Biswas et al. [2022] 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 (37) for $\nu = 2$. The upper bounds in Figure 6 are given by our estimator CUB_2 (4). 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. [2022]. The lower bounds in Figure 6 are estimated using (7) 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 $L(\beta; y, X) = \prod_{i=1}^n (1 + \exp(-y_i x_i^\top \beta))^{-1}$ 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

$$Z_j \overset{i.i.d.}{\sim} \text{Bernoulli}(q), \qquad \beta_j | Z_j = 0 \sim \mathcal{N}(0, \tau_0^2), \qquad \beta_j | Z_j = 1 \sim \mathcal{N}(0, \tau_1^2) \tag{38}$$

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 $Z_i = 0$ and $Z_i = 1$ correspond to null and non-null components of β_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 (38) are Gibbs samplers which cost $\mathcal{O}(n^2 d)$ 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 $C_t \triangleq (\beta_t, \eta_t, \sigma_t^2, \xi_t) \in \mathbb{R}^d \times \mathbb{R}^d_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, approximate chain current state $\tilde{C}_t \triangleq (\tilde{\beta}_t, \tilde{\eta}_t, \tilde{\sigma}_t^2, \tilde{\xi}_t) \in \mathbb{R}^d \times \mathbb{R}^d_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and approximation threshold $\epsilon > 0$.

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

$$\pi(\eta_{t+1,j}|...) \propto \frac{e^{-m_{t,j}\eta_{t+1,j}}}{\eta_{t+1,j}^{\frac{1-\nu}{2}}(1+\nu\eta_{t+1,j})^{\frac{\nu+1}{2}}} \text{ and } \pi(\tilde{\eta}_{t+1,j}|...) \propto \frac{e^{-\tilde{m}_{t,j}\eta_{t+1,j}}}{\eta_{t+1,j}^{\frac{1-\nu}{2}}(1+\nu\eta_{t+1,j})^{\frac{\nu+1}{2}}}$$

   for $m_{t,j} \triangleq \left(\xi_t \beta_{t,j}^2\right)/\left(2\sigma_t^2\right)$ and $\tilde{m}_{t,j} \triangleq \left(\tilde{\xi}_t \tilde{\beta}_{t,j}^2\right)/\left(2\tilde{\sigma}_t^2\right)$ respectively using common random numbers. This can be done using the slice sampler of Biswas et al. [2022].

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

   (a) Sample $(\xi_{t+1}, \tilde{\xi}_{t+1})$ via Metropolis-Hastings with step size $\sigma_{\text{MH}} = 0.8$:
   Propose $\log(\xi^*) = \log(\xi_t) + \sigma_{\text{MH}} Z^*$ and $\log(\tilde{\xi}^*) = \log(\tilde{\xi}_t) + \sigma_{\text{MH}} Z^*$ for $Z^* \sim \mathcal{N}(0,1)$.
   Calculate acceptance probabilities

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

   where $\pi_\xi(\cdot)$ is the prior density of ξ, $M \triangleq I_n + \xi_t^{-1} X \operatorname{Diag}(\eta_{j,t}^{-1}) X^\top$,
   $\tilde{M} \triangleq I_n + X \operatorname{Diag}((\tilde{\xi}_t^{-1}\tilde{\eta}_{j,t}^{-1} \mathrm{I}_{\{\tilde{\xi}_{\max}^{-1}\tilde{\eta}_{j,t}^{-1} > \epsilon\}})_{j=1}^p) X^\top$ for $\tilde{\xi}_{\max} = \max\{\tilde{\xi}_t, \tilde{\xi}^*\}$,

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

   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 $\tilde{\xi}_{t+1} \triangleq \tilde{\xi}^*$ if $U^* \leq \min(1, \tilde{q})$, else set $\tilde{\xi}_{t+1} \triangleq \tilde{\xi}_t$.

Algorithm 2: continued

2. [(a)]

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

$$\sigma^2_{t+1}|\xi_{t+1}, \eta_{t+1} \sim \text{InvGamma}\Big(\frac{a_0+n}{2}, \frac{y^\top M^{-1}_{\xi_{t+1},\eta_{t+1}} y + b_0}{2}\Big) \text{ and}$$
$$\tilde{\sigma}^2_{t+1}|\tilde{\xi}_{t+1}, \tilde{\eta}_{t+1} \sim \text{InvGamma}\Big(\frac{a_0+n}{2}, \frac{y^\top M^{-1}_{\tilde{\xi}_{t+1},\tilde{\eta}_{t+1}} y + b_0}{2}\Big).$$

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

$$\beta_{t+1}|\sigma^2_{t+1}, \xi_{t+1}, \eta_{t+1} \sim \mathcal{N}\big(\Sigma^{-1}X^\top y, \sigma^2_{t+1}\Sigma^{-1}\big) \text{ for } \Sigma = X^\top X + \xi_{t+1}\text{Diag}(\eta_{t+1})$$
$$\tilde{\beta}_{t+1}|\tilde{\sigma}^2_{t+1}, \tilde{\xi}_{t+1}, \tilde{\eta}_{t+1} \sim \mathcal{N}\big(\tilde{\Sigma}^{-1}X^\top y, \tilde{\sigma}^2_{t+1}\tilde{\Sigma}^{-1}\big) \text{ for } \tilde{\Sigma} = X^\top X + \tilde{\xi}_{t+1}\text{Diag}(\tilde{\eta}_{t+1})$$

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

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

$$\begin{pmatrix} X_A^\top X_A + \tau_1^{-2} I & X_A^\top X_{A^c} \\ X_{A^c}^\top X_A & X_{A^c}^\top X_{A^c} + \tau_0^{-2} I \end{pmatrix} \approx \begin{pmatrix} X_A^\top X_A + \tau_1^{-2} I & 0 \\ 0 & ((n-1) + \tau_0^{-2}) I \end{pmatrix}$$

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^c} 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 CUB_2 (4) 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 (7) 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

$$(\beta_{A,t+1}, \beta_{A^c,t+1})|z_t, e_t, w_t \sim \mathcal{N}\left(\Sigma^{-1}X^\top W y, \Sigma^{-1}\right) \text{ for } \Sigma = \begin{pmatrix} X_A^\top W X_A + \tau_1^{-2} I & X_A^\top W X_{A^c} \\ X_{A^c}^\top W X_A & X_{A^c}^\top W X_{A^c} + \tau_0^{-2} I \end{pmatrix},$$

$$(\tilde{\beta}_{\tilde{A},t+1}, \tilde{\beta}_{\tilde{I},t+1})|\tilde{z}_t, \tilde{e}_t, \tilde{w}_t \sim \mathcal{N}\left(\tilde{\Sigma}^{-1}X^\top \tilde{W} y, \tilde{\Sigma}^{-1}\right) \text{ for } \tilde{\Sigma} = \begin{pmatrix} X_{\tilde{A}}^\top \tilde{W} X_{\tilde{A}} + \tau_1^{-2} I & 0 \\ 0 & ((n-1) + \tau_0^{-2}) I \end{pmatrix}$$

   where $W = \mathrm{Diag}(w_t)$ and $\tilde{W} = \mathrm{Diag}(\tilde{w}_t)$, $A = \{j : z_{j,t} = 1\}$ and $\tilde{A} = \{j : \tilde{z}_{j,t} = 1\}$ are the index sets of active components, X_A and $X_{\tilde{A}}$ are matrices corresponding to the active (or inactive) columns of X with columns $j \in A$ and $j \in \tilde{A}$ respectively, $\beta_{A,t+1}$ and $\tilde{\beta}_{\tilde{A},t+1}$ are vectors of active components of β_{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 $j = 1, \ldots, p$ such that each $z_{j,t+1}$ and $\tilde{z}_{j,t+1}$ are Bernoulli random variables with odds

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

   respectively where $\mathcal{N}(\cdot; \mu, \Sigma)$ is the probability density of the normal distribution with mean μ and variance Σ, $C_j \triangleq \{k : \tilde{z}_{k,t+1} = 1 \text{ for } k < j \text{ or } \tilde{z}_{k,t} = 1 \text{ for } k > j\}$ is the index set of active components in $\{1, \ldots, p\} \backslash \{j\}$, 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.

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 $i = 1, ..., n$

$$e_{i,t+1} \sim \begin{cases} \mathcal{N}(x_i^\top \beta_{t+1}, w_{i,t}^{-1})\mathrm{I}_{[0,\infty)} & \text{if } y_i = 1 \\ \mathcal{N}(x_i^\top \beta_{t+1}, w_{i,t}^{-1})\mathrm{I}_{(-\infty,0)} & \text{if } y_i = 0 \end{cases} \text{ and}$$

$$\tilde{e}_{i,t+1} \sim \begin{cases} \mathcal{N}(x_{\tilde{A},i}^\top \tilde{\beta}_{\tilde{A},t+1}, \tilde{w}_{i,t}^{-1})\mathrm{I}_{[0,\infty)} & \text{if } y_i = 1 \\ \mathcal{N}(x_{\tilde{A},i}^\top \tilde{\beta}_{\tilde{A},t+1}, \tilde{w}_{i,t}^{-1})\mathrm{I}_{(-\infty,0)} & \text{if } y_i = 0 \end{cases}$$

where $x_i^\top$ and $x_{\tilde{A},i}^\top$ 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 $w_{i,t} = \tilde{w}_{i,t} = 3/\pi^2$ for all $i = 1, ..., n$ and $t \geq 0$, where $3/\pi^2$ 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 $i = 1, ..., n$,

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

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 $C_t \triangleq (\tilde{\beta}_t, \tilde{z}_t, \tilde{e}_t, \tilde{w}_t)$.

F Additional Algorithms

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

Input: (X_t, Y_t), unnormalized densities p and q of P and Q respectively, step sizes σ_P and σ_Q
Sample $\epsilon_{CRN} \sim \mathcal{N}(0, I_d)$. Calculate proposals

$$X^* \triangleq X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t) + \sigma_P \epsilon_{CRN} \text{ and } Y^* \triangleq Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t) + \sigma_Q \epsilon_{CRN}$$

Sample $U_{CRN} \sim \text{Uniform}([0,1])$
if $U_{CRN} \leq \frac{p(X^*)\mathcal{N}(X^*; X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t), \sigma_P^2 I_d)}{p(X_t)\mathcal{N}(X_t; X^* + \frac{1}{2}\sigma_P^2 \nabla \log p(X^*), \sigma_P^2 I_d)}$, **then** set $X_{t+1} = X^*$; **else** set $X_{t+1} = X_t$
if $U_{CRN} \leq \frac{q(Y^*)\mathcal{N}(Y^*; Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t), \sigma_Q^2 I_d)}{q(Y_t)\mathcal{N}(Y_t; Y^* + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y^*), \sigma_Q^2 I_d)}$, **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 σ_P and σ_Q
Sample $\epsilon_{CRN} \sim \mathcal{N}(0, I_d)$. Calculate proposals

$$X^* \triangleq X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t) + \sigma_P \epsilon_{CRN} \text{ and } Y^* \triangleq Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t) + \sigma_Q \epsilon_{CRN}.$$

Sample $U \sim \text{Uniform}([0,1])$
if $U \leq \frac{p(X^*)\mathcal{N}(X^*; X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t), \sigma_P^2 I_d)}{p(X_t)\mathcal{N}(X_t; X^* + \frac{1}{2}\sigma_P^2 \nabla \log p(X^*), \sigma_P^2 I_d)}$, **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})

Algorithm 6: Reflection coupling of two MALA kernels marginally targetting distributions P and Q respectively [see, e.g. Bou-Rabee et al., 2020].

Input: (X_t, Y_t), unnormalized densities p and q of P and Q respectively, step sizes σ_P and σ_Q
Sample $\epsilon \sim \mathcal{N}(0, I_d)$. Calculate proposals

$$X^* \triangleq X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t) + \sigma_P \epsilon$$
$$Y^* \triangleq Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t) + \sigma_Q (I_d - ee^\top)\epsilon \text{ for } e = \frac{X_t - Y_t}{\|X_t - Y_t\|_2}.$$

Sample $U_{CRN} \sim \text{Uniform}([0,1])$.
if $U_{CRN} \leq \frac{p(X^*)\mathcal{N}(X^*; X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t), \sigma_P^2 I_d)}{p(X_t)\mathcal{N}(X_t; X^* + \frac{1}{2}\sigma_P^2 \nabla \log p(X^*), \sigma_P^2 I_d)}$, **then** set $X_{t+1} = X^*$; **else** set $X_{t+1} = X_t$.
if $U_{CRN} \leq \frac{q(Y^*)\mathcal{N}(Y^*; Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t), \sigma_Q^2 I_d)}{q(Y_t)\mathcal{N}(Y_t; Y^* + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y^*), \sigma_Q^2 I_d)}$, **then** set $Y_{t+1} = Y^*$; **else** set $Y_{t+1} = Y_t$.
return (X_{t+1}, Y_{t+1})

Algorithm 7: Reflection maximal coupling of two MALA kernels marginally targetting distributions P and Q respectively [see, e.g. Bou-Rabee et al., 2020].

Input: (X_t, Y_t), unnormalized densities p and q of P and Q respectively, step sizes σ_P and σ_Q
Sample $\epsilon \sim \mathcal{N}(0, I_d)$, $U^* \sim \text{Uniform}([0,1])$. Calculate proposals

$$X^* \triangleq X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t) + \sigma_P \epsilon$$
$$Y^* \triangleq \begin{cases} X^* & \text{if } U^* \leq \frac{\mathcal{N}(X^*; Y_t + \frac{1}{2}\sigma_P^2 \nabla \log p(Y_t), \sigma_P^2 I_d)}{\mathcal{N}(X^*; X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t), \sigma_P^2 I_d)} \\ Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t) + \sigma_Q (I_d - ee^\top)\epsilon & \text{otherwise for } e = \frac{X_t - Y_t}{\|X_t - Y_t\|_2}. \end{cases}$$

Sample $U_{CRN} \sim \text{Uniform}([0,1])$.
if $U_{CRN} \leq \frac{p(X^*)\mathcal{N}(X^*; X_t + \frac{1}{2}\sigma_P^2 \nabla \log p(X_t), \sigma_P^2 I_d)}{p(X_t)\mathcal{N}(X_t; X^* + \frac{1}{2}\sigma_P^2 \nabla \log p(X^*), \sigma_P^2 I_d)}$, **then** set $X_{t+1} = X^*$; **else** set $X_{t+1} = X_t$.
if $U_{CRN} \leq \frac{q(Y^*)\mathcal{N}(Y^*; Y_t + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y_t), \sigma_Q^2 I_d)}{q(Y_t)\mathcal{N}(Y_t; Y^* + \frac{1}{2}\sigma_Q^2 \nabla \log q(Y^*), \sigma_Q^2 I_d)}$, **then** set $Y_{t+1} = Y^*$; **else** set $Y_{t+1} = Y_t$.
return (X_{t+1}, Y_{t+1})