4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–9

Bounding Wasserstein distance with couplings

Anonymous Authors
Anonymous Institution

Abstract

Markov chain Monte Carlo (MCMC) methods are a powerful tool in Bayesian computation. They provide asymptotically consistent estimates as the number of iterations tends to infinity. However, in large data applications, MCMC can be computationally expensive per iteration. This has catalyzed interest in sampling methods such as approximate MCMC, which trade off asymptotic consistency for improved computational speed. In this article, we propose estimators based on couplings of Markov chains to assess the quality of such asymptotically biased sampling methods. The estimators give empirical upper bounds of the Wassertein distance between the limiting distribution of the asymptotically biased sampling method and the original target distribution of interest. We apply our sample quality measures to two stylized examples in high dimensions.

1. Introduction

1.1. Sample 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 (Gelman and Brooks, 1998; Liu, 2008; Robert and Casella, 2013). In modern applications with a large number of data points or high dimensions, MCMC methods can have high computation cost per iteration. This has catalyzed the use of approximate MCMC methods (e.g. Welling and Teh, 2011; Bardenet et al., 2017; Narisetty et al., 2019; Johndrow et al., 2020), which have lower computation cost per iteration but may not converge to the target distribution of interest, and methods such as variational inference (e.g. Wainwright and Jordan, 2008; Blei et al., 2017), which inexactly approximate the target distribution through optimization.

Measuring the sample quality of such asymptotically biased samplers is of great interest for researchers who develop new approximate inference methodology. Standard MCMC diagnostic tests (e.g., Johnson, 1998; Biswas et al., 2019; Vats and Knudson, 2020; Vehtari et al., 2020) are not directly suitable for such settings as they do not account for asymptotic bias. Researchers often resort to comparing summary statistics or marginal univariate traceplots of samples from such methods with samples from an asymptotically unbiased Markov chain. Such marginal traceplots and summary statistics may fail to capture higher order moments and dependencies between different components. Moreover, in high-dimensional settings, visualizing all marginal traceplots may not even be feasible. These limitations of existing diagnostics and heuristics have stimulated recent work in measuring the quality of sample approximations, with Gorham and Mackey (2015); Chwialkowski et al. (2016); Liu et al. (2016); Gorham and Mackey (2017); Huggins and Mackey (2018); Gorham et al. (2020) developing measures based on Stein discrepancies which do not require sampling from the target distribution of interest. In this manuscript, we develop generic upper bound estimates of the Wasserstein distance that apply to any distributions that can be targeted with fast-mixing Markov chains and do not require any additional distributional knowledge.

1.2. Couplings and Wasserstein distances

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

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

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

The Wasserstein distance has many advantageous properties compared to metrics such as Total Variation distance and divergences such as Kullback–Leibler divergence and Rényi's α-divergence (Villani, 2008; Peyré and Cuturi, 2019; Huggins et al., 2020). It allows comparison between singular distributions that may have disjoint supports, captures geometric properties characterized by the metric c and captures differences between moments of distributions. In this manuscript, 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.

2. Bounding the Wasserstein distance with couplings

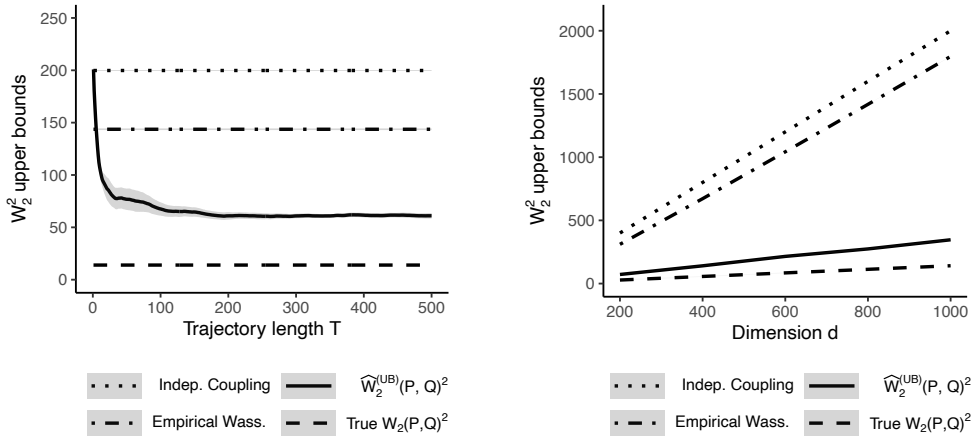
Given distributions P and Q on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and some $p \geq 1$, we wish to estimate upper bounds on $\mathcal{W}_p(P, Q)$. Our estimates are based on Markov chains $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ with marginal transition kernels K_1 and K_2 which have invariant distributions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

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

Such kernels have been used analytically to develop perturbation theory for Markov chains (Pillai and Smith, 2015; Johndrow and Mattingly, 2018; Rudolf and Schweizer, 2018). Given $\bar{K}$, we instead simulate the coupled Markov chain $(X_t, Y_t)_{t\geq 0}$ using Algorithm 1. Algorithm 1 is an extension of the joint kernels considered in (Johnson, 1998; Glynn and Rhee, 2014; Heng and Jacob, 2019; Middleton et al., 2019; Jacob et al., 2020; Biswas et al., 2019, 2021) for unbiased estimation and MCMC convergence diagnostics, where $K_1 = K_2$ and $X_0 \sim Y_0$ such that $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ have the same marginal distributions. Algorithms to sample from $\bar{K}$ are in the supplement.

Algorithm 1: Coupled Markov chain Monte Carlo for sample quality

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 $t = 1, \ldots, T$ **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 $(X_t, Y_t)_{t=0}^T$



(*a*) Dimension $d = 100$ (*b*) Varying dimension

Figure 1: Upper bounds on 2-Wasserstein distances with L_2 norm between $P = \mathcal{N}(0, \Sigma)$ and $Q = \mathcal{N}(0, I_d)$. $\widehat{\mathcal{W}}_p^{(UB)}(P, Q)$ corresponds to our upper bound estimate from (3).

Consider a coupled Markov chain $(X_t, Y_t)_{t\geq 0}$ generated using Algorithm Algorithm 1. Suppose the marginal distributions of X_t and Y_t converge in p-Wasserstein distance to P and Q respectively as t tends to infinity. Informally, the coupling representation of the Wasserstein distance implies $\mathcal{W}_p(P, Q)^p \leq \liminf_{S\to\infty} \sum_{t=S+1}^{T} \mathbb{E}[c(X_t, Y_t)^p]/(T - S)$ for all $T > S$. This motivates our upper bound

$$\widehat{\mathcal{W}}_p^{(UB)}(P, Q) := \Big(\frac{1}{I(T - S)} \sum_{t=S+1}^{T} \sum_{i=1}^{I} c(X_t^{(i)}, Y_t^{(i)})^p\Big)^{1/p}, \tag{3}$$

where $(X_t^{(i)}, Y_t^{(i)})_{t=0}^{T}$ are coupled Markov chains sampled using Algorithm Algorithm 1 independently for each chain $i = 1, \ldots, I$, with burn-in $S \geq 0$ and trajectory length $T > S$. We formally establish the consistency of this and related upper bound estimators in the supplement.

We now consider the empirical performance of this upper bound on two stylized examples. We focus on the distance metric $c(x, y) = \|x - y\|_2$ induced by the L_2 norm, which controls the difference between the first and second moments between distributions.

2.1. Upper bound on Wasserstein distance

Figure 1 highlights the performance of the upper bound in (3) for two multivariate Gaussian distributions on $\mathbb{R}^d$, given by

$$P = \mathcal{N}(0, \Sigma) \text{ where } [\Sigma]_{i,j} = 0.5^{|i-j|} \text{ for } 1 \leq i, j \leq d \text{ and } Q = \mathcal{N}(0, I_d). \tag{4}$$

The marginal kernels K_1 and K_2 are based on Metropolis–adjusted Langevin algorithm (MALA) targeting P and Q respectively. 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 MALA, and is in the supplement. We initialize each $X_0^{(i)} \sim P$ and $Y_0^{(i)} \sim Q$ independently.

Figure 1(a) shows the performance of our upper bound on the 2-Wasserstein distance $\mathcal{W}_2(P, Q)$ with dimension $d = 100$. For the coupled MALA kernel we use a step-size of $\sigma_P = \sigma_Q = 0.5$ for both the marginal chains. The solid line plots the upper bound estimates $\widehat{\mathcal{W}}_2^{(UB)}(P, Q)^2$ from (3) based on $I = 5$ independent coupling chains with burn-in $S = 0$ and varying trajectory length $1 \leq T \leq 500$. The dotted line corresponds to $\mathbb{E}[\|X - Y\|_2^2]$ for $X \sim P$ and $Y \sim Q$ independent, and equals $2d = 200$. The dot-dashed line corresponds to an upper bound estimate based on solving linear programs on the empirical distributions of P and Q. It plots $\sum_{i=1}^{I} \mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})^2/I$ where $\hat{P}_T^{(i)}$ and $\hat{Q}_T^{(i)}$ are empirical distribution of P and Q respectively based on $T = 500$ independent samples and each $\mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})$ is calculated by solving a linear program independently for $i = 1, \ldots, I = 5$. The dashed line corresponds to the true Wasserstein distance squared (e.g. Peyré and Cuturi, 2019, Remark 2.23). The grey error bands correspond to one standard deviation arising from Monte Carlo error. By our choice of initialization, the averaged trajectory has the same initial distance value as the distance under an independence coupling. For greater trajectory length T, $\widehat{\mathcal{W}}_2^{(UB)}(P, Q)^2$ gives an improved upper bound. Overall, Figure 1(a) shows that couplings give informative upper bounds to $\mathcal{W}_2(P, Q)^2$.

Figure 1(b) considers higher dimensions. The solid line plots the upper bound estimates $\widehat{\mathcal{W}}_2^{(UB)}(P, Q)^2$ from (3), based on $I = 10$, $T = 1,500$ and $S = 500$. We use a step-size of $\sigma_P = \sigma_Q = 0.5d^{-1/6}$ for both the marginal chains (Kennedy and Pendleton, 1991; Roberts and Rosenthal, 1998, 2001). The grey error bands are now too small to be visible. The dotted line plots $\mathbb{E}[\|X - Y\|_2^2]$ for $X \sim P$ and $Y \sim Q$ independent, and equals $2d$. The dot-dashed line plots an upper bound estimate based on solving linear programs on the empirical distributions of P and Q. It plots $\sum_{i=1}^{I} \mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})^2/I$ where $\hat{P}_T^{(i)}$ and $\hat{Q}_T^{(i)}$ are empirical distribution of P and Q respectively based on $T = 1,500$ independent samples and each $\mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})$ is calculated by solving a linear program for $I = 10$. The dashed line plots the true Wasserstein distance squared $\mathcal{W}_2(P, Q)^2$. Figure 1(b) highlights that couplings can give upper bounds that remain informative in higher dimensions. Theoretical analysis of such properties are in the supplement.

2.2. Bias of approximate MCMC methods

Figure 2 highlights the performance of coupling based upper bounds when the marginal kernels K_1 and K_2 are based on a MALA Markov chain and an unadjusted Langevin algorithm (ULA) Markov chain respectively, with both chains targeting the distribution $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$ as defined in (4). The MALA kernel K_1 produces an *exact* Markov chain which is $\mathcal{N}(0, \Sigma)$ invariant. 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 the MALA and the ULA algorithms, and is in the supplement. We use a step-size of $\sigma_P = \sigma_Q = 0.5d^{-1/6}$ for both chains, 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 distributions of $X_t^{(i)}$ and $Y_t^{(i)}$ respectively.

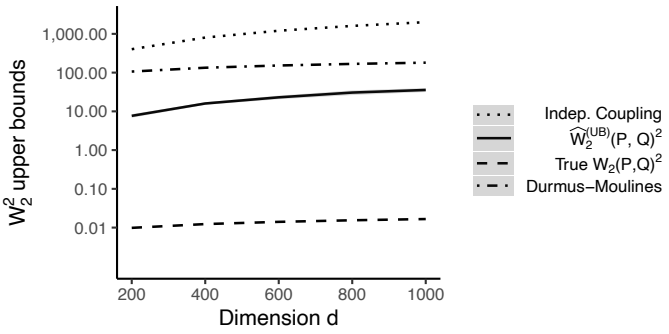


Figure 2: Upper bounds on 2-Wasserstein distances with L_2 norm between limiting distributions of ULA and MALA targeting $P = \mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$. Analytic upper bounds are from Durmus-Moulines (Durmus and Moulines, 2019). $\widehat{\mathcal{W}}_p^{(UB)}(P, Q)$ corresponds to our upper bound estimate from (3).

We obtain

$$P_t = \mathcal{N}(0, \Sigma) =: P, \; Q_t = \mathcal{N}\big(0, \sigma_Q^2 \sum_{j=0}^{t-1} B^{2j}\big), \text{ and } Q_t \overset{t\to\infty}{\Rightarrow} \mathcal{N}(0, \sigma_Q^2 (I_d - B^2)^{-1}) =: Q \quad (5)$$

where $B := I_d - (\sigma_Q^2/2)\Sigma^{-1}$, and weak convergence of Q_t to Q holds for σ_Q sufficiently small.

Figure 2 shows the performance of our upper bound on $\mathcal{W}_2(P, Q)$ to measure the asymptotic bias of ULA. The solid line shows the asymptotic bias upper bound calculated using our coupled chains. For each dimension d, it is calculated using $\widehat{\mathcal{W}}_2^{(UB)}(P, Q)^2$ from (3) with $I = 10$ independent chains each of length $T = 3000$ with a burn-in of $S = 1000$ iterations. For such number of independent chains and chain length, the grey error bands for the the coupling based estimates are too small to be visible. The dashed line shows the true asymptotic bias $\mathcal{W}_2(P, Q)^2$, which is analytical tractable for this example. The dotted line corresponds to $\mathbb{E}[\|X - Y\|_2^2]$ for $X \sim P$ and $Y \sim Q$ independent, and is analytical tractable for this example. The dot-dashed line shows the analytic upper bounds of the asymptotic bias of ULA developed via couplings-based theoretical analysis (Durmus and Moulines, 2019, Corollary 9). Figure 2 highlights that simulating couplings can give informative empirical upper bounds to the asymptotic bias for practitioners which are much tighter compared to an independent coupling and to the analytic upper bounds derived for ULA and thus directly useful for practitioners. Overall, Figure 2 highlights that couplings can give empirically informative upper bounds to $\mathcal{W}_2(P_t, Q_t)$ in higher dimensions.

3. Discussion

We have used 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. Our supplement contains further details including:

(i) consistency of our upper bound estimates, (ii) sufficient conditions for our upper estimates to remain informative in high dimensions, (iii) comparison with alternative methods, (iii) application of our sample quality measures on stochastic gradient MCMC, variational Bayes, and Laplace approximations for tall data, and approximate MCMC for linear and logistic regression in 5,000 dimensions.

References

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *J. Mach. Learn. Res.*, 18(1):1515–1557, January 2017. ISSN 1532-4435.

Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings. In *Advances in Neural Information Processing Systems*, pages 7389–7399, 2019.

Niloy Biswas, Anirban Bhattacharya, Pierre E. Jacob, and James E. Johndrow. Coupled markov chain monte carlo for high-dimensional regression with half-t priors. *arXiv preprint arXiv:2012.04798v2*, 2021.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2606–2615. JMLR.org, 2016.

Alain Durmus and Éric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 11 2019. doi: 10.3150/18-BEJ1073. URL https://doi.org/10.3150/18-BEJ1073.

Andrew Gelman and Stephen P. Brooks. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 1998.

Peter W Glynn and Chang-han Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/698d51a19d8a121ce581499d7b701668-Paper.pdf.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1292–1301, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/gorham17a.html.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein Discrepancies. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17931–17942. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf.

J Heng and P E Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302, 02 2019. ISSN 0006-3444. doi: 10.1093/biomet/asy074. URL https://doi.org/10.1093/biomet/asy074.

Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1792–1802. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/huggins20a.html.

Jonathan H. Huggins and Lester Mackey. Random feature Stein discrepancies. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1903–1913, Red Hook, NY, USA, 2018. Curran Associates Inc.

Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020. doi: 10.1111/rssb.12336. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12336.

James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61, 2020. URL http://jmlr.org/papers/v21/19-536.html.

James E. Johndrow and Jonathan C. Mattingly. Error bounds for approximations of markov chains used in bayesian sampling. *arXiv preprint arXiv:1711.05382*, 2018.

Valen E Johnson. A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441): 238–248, 1998.

A.D. Kennedy and Brian Pendleton. Acceptances and autocorrelations in hybrid monte carlo. *Nuclear Physics B - Proceedings Supplements*, 20:118 – 121, 1991. ISSN 0920-5632. doi: https://doi.org/10.1016/0920-5632(91)90893-J. URL http://www.sciencedirect.com/science/article/pii/092056329190893J.

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

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/liub16.html.

Lawrece Middleton, George Deligiannidis, Arnaud Doucet, and Pierre E. Jacob. Unbiased smoothing using particle independent metropolis-hastings. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2378–2387. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/middleton19a.html.

Naveen N. Narisetty, Juan Shen, and Xuming He. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*, 114 (527):1205–1217, 2019. doi: 10.1080/01621459.2018.1482754. URL https://doi.org/10.1080/01621459.2018.1482754.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073.

Natesh S. Pillai and Aaron Smith. Ergodicity of approximate mcmc chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*, 2015.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Spinger New York, 2013.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 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.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statist. Sci.*, 16(4):351–367, 11 2001. doi: 10.1214/ss/1015346320. URL https://doi.org/10.1214/ss/1015346320.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 11 2018. doi: 10.3150/17-BEJ938. URL https://doi.org/10.3150/17-BEJ938.

Dootika Vats and Christina Knudson. Revisiting the gelman-rubin diagnostic. *Statistical Science (to appear)*, 2020.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Anal.*, 2020. doi: 10.1214/20-BA1221. URL https://doi.org/10.1214/20-BA1221. Advance publication.

Cédric Villani. *Optimal transport – Old and new*, volume 338. Springer, 2008. doi: 10.1007/978-3-540-71050-9.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–32

Bounding Wasserstein distance with couplings: Supplementary material

Anonymous Authors
Anonymous Institution

1. Properties and Implementation

In this section we establish the consistency of the estimators in the main text, consider how to sample from the joint kernel $\bar{K}$, and investigate theoretical properties of our upper bounds. All proofs are deferred to Appendix 4.

1.1. Consistency of upper bounds on Wasserstein distance

Let $(X_t^{(i)}, Y_t^{(i)})_{t\geq 0}$ denote coupled Markov chains, where each chain $i = 1, \ldots, I$ is independently generated with initial distribution $\bar{I}_0$ and joint kernel $\bar{K}$ on $\mathcal{X} \times \mathcal{X}$ with marginal kernels K_1 and K_2. For each $t \geq 0$, let P_t and Q_t denote the distribution of $X_t^{(i)}$ and $Y_t^{(i)}$ respectively. Fix some $p \in [1, \infty)$, and suppose P_t and Q_t have finite moments of order p for all $t \geq 0$. Under this setup, we establish the consistency of the p-Wasserstein estimators proposed.

Proposition 1.1 *Define the estimator*

$$\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t) := \Big(\frac{1}{I}\sum_{i=1}^{I} c(X_t^{(i)}, Y_t^{(i)})^p\Big)^{1/p}. \tag{1}$$

Then $\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)$ has finite moments of order p for all $I \geq 1$, and

$$\mathcal{W}_p(P_t, Q_t)^p \leq \mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p] \quad \textit{for all } t \geq 0. \tag{2}$$

Similarly, we can consider the Wasserstein distance between mixtures of the marginal distributions.

Corollary 1.2 $\mathcal{W}_p(\frac{1}{T}\sum_{t=1}^{T} P_t, \frac{1}{T}\sum_{t=1}^{T} Q_t)^p \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p]$.

When the marginal chains have stationary distributions and are initialized at stationarity marginally, we can obtain upper bound estimates of the Wasserstein distance between the two stationary distributions.

Corollary 1.3 *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 $I \geq 0$ and trajectories with burn-in $S \geq 1$ and length $T \geq S$, the estimator $\widehat{\mathcal{W}}_p^{(UB)}(P, Q)$ has finite moments of order p, and*

$$\mathcal{W}_p(P, Q)^p \leq \mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P, Q)^p]. \tag{3}$$

Under Proposition 1.1, Corollary 1.2 and Corollary 1.3, by strong law of large numbers the estimators $\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)$, $\sum_{t=1}^T \widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)/T$ and $\widehat{\mathcal{W}}_p^{(UB)}(P, Q)$ converge almost surely to upper bounds of $\mathcal{W}_p(P_t, Q_t)$, $\mathcal{W}_p(\sum_{t=1}^T P_t/T, \sum_{t=1}^T Q_t/T)$ and $\mathcal{W}_p(P, Q)$ respectively as $I \to \infty$. We may not always be able to initialize from the stationary distributions P and Q marginally. To obtain upper bounds of $\mathcal{W}_p(P, Q)$ without starting at the stationary distributions P and Q marginally, we make an assumption related to convergence of marginal distributions $(P_t)_{t\geq 0}$ and $(Q_t)_{t\geq 0}$ on $\mathcal{X}$.

Assumption 1.4 *As $t \to \infty$, P_t and Q_t converge in p-Wasserstein distance to distributions P and Q respectively, where P and Q have finite moments of order p.*

Proposition 1.5 *Under Assumption 1.4, for all $\epsilon > 0$ there exists some $S \geq 1$ such that for all $T \geq S$,*

$$\mathcal{W}_p(P, Q)^p \leq \epsilon + \frac{1}{T-S} \sum_{t=S+1}^{T} \mathbb{E}\big[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p\big] \tag{4}$$

where $\sum_{t=S+1}^T \mathbb{E}\big[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p\big]/(T-S)$ is finite.

Proposition 1.5 presents an asymptotic upper bound. In practice, we can use the estimate

$$\Big(\frac{1}{T-S} \sum_{t=S+1}^{T} \widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p\Big)^{1/p} \tag{5}$$

with a large burn-in $S \geq 1$ to obtain an upper bound to $\mathcal{W}_p(P, Q)$ under any initialization $(X_0, Y_0) \sim \bar{I}_0$.

For $p = 1$, we can also obtain non-asymptotic upper bounds using the L-lag coupling approach of (Biswas et al., 2019). Suppose $(\tilde{X}_{t-L}, X_t)_{t\geq L}$ is an L-lag coupling chain for kernel K_1 and $(\tilde{Y}_{t-L}, Y_t)_{t\geq L}$ is an L-lag coupling chain for kernel K_2. We make the following assumptions on K_1 and K_2 following (Jacob et al., 2020) and (Biswas et al., 2019).

Assumption 1.6 *For all $t \geq L$, $\mathbb{E}[c(\tilde{X}_{t-L}, X_t)^{2+\eta}] \leq D$ and $\mathbb{E}[c(\tilde{Y}_{t-L}, Y_t)^{2+\eta}] \leq D$ for some $\eta > 0, D < \infty$.*

Assumption 1.7 *The meeting times $\tau_P := \inf\{t > L : X_t = \tilde{X}_{t-L}\}$ and $\tau_Q := \inf\{t > L : Y_t = \tilde{Y}_{t-L}\}$ satisfy $\mathbb{P}(\frac{\tau_P - L}{L} > t) \leq C\delta^t$ and $\mathbb{P}(\frac{\tau_Q - L}{L} > t) \leq C\delta^t$ for all $t \geq 0$, for some constants $C < \infty$ and $\delta \in (0, 1)$.*

Assumption 1.8 *Faithfulness after meeting: $X_t = \tilde{X}_{t-L}$ for all $t \geq \tau_P$ and $Y_t = \tilde{Y}_{t-L}$ for all $t \geq \tau_Q$.*

Proposition 1.9 *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 the*

joint kernel $\bar{K}$. Under assumption 1.4 with $p = 1$, and assumptions 1.6, 1.7 and 1.8, for all $t \geq 0$

$$\mathcal{W}_1(P, Q) \leq \mathbb{E}\left[\widehat{\mathcal{W}}_1^{(UB)}(P_t, Q_t) + \sum_{j=1}^{\lceil(\tau_P - L - t)/L\rceil} c(\tilde{X}_{t+(j-1)L}, X_{t+jL}) + \sum_{j=1}^{\lceil(\tau_Q - L - t)/L\rceil} c(\tilde{Y}_{t+(j-1)L}, Y_{t+jL})\right], \tag{6}$$

where the expectation is finite and can be computed in finite time.

In Proposition 1.9, $\mathbb{E}\big[\sum_{j=1}^{\lceil(\tau_P - L - t)/L\rceil} c(\tilde{X}_{t+(j-1)L}, X_{t+jL})\big]$ and $\mathbb{E}\big[\sum_{j=1}^{\lceil(\tau_Q - L - t)/L\rceil} c(\tilde{Y}_{t+(j-1)L}, Y_{t+jL})\big]$ correspond to upper bounds of $\mathcal{W}_1(P_t, P)$ and $\mathcal{W}_1(Q_t, Q)$ respectively (Biswas et al., 2019). This captures the 1-Wasserstein convergence of the marginal distributions $(P_t)_{t\geq 0}$ and $(Q_t)_{t\geq 0}$ to give the non-asymptotic upper bound in (6).

We emphasize that the results of this section hold for any coupled chain sampled with the joint kernel $\bar{K}$. For example, this includes both the CRN coupled chains and independently coupled chains. We now consider how to sample from the joint kernel $\bar{K}$ and investigate when our upper bounds can be informative.

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

In this section, we develop algorithms to sample from the joint kernel $\bar{K}$ such that the estimators from Section 1.1 can produce informative upper bounds. Our construction makes use of the coupled kernels Γ_1 on $\mathcal{X} \times \mathcal{X}$ and Γ_Δ on $\mathcal{X}$ such that:

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

2. 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 $K_2(z, \cdot)$.

The coupled kernel Γ_1 characterizes the marginal chain corresponding to K_1. 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. Alternatively when the proposal step is based on a spherically symmetric distribution such as a Gaussian (e.g. Random-Walk Metropolis–Hastings or the momentum component in Hamiltonian Monte Carlo), Γ_1 can be a reflection coupling of the proposal step and a CRN coupling of the accept-reject step ((Lindvall and Rogers, 1986; Bou-Rabee et al., 2020); see also (O'Leary et al., 2021)). The coupled 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 (when MALA and ULA have the same step-size, this gives the same proposal) with the MALA chain having a further accept-reject Metropolis–Hastings correction step and the ULA chain always accepting the proposal. Choice of coupled kernels Γ_1 and Γ_Δ is further discussed in Section 1.3.

Given coupled kernels Γ_1 and Γ_Δ, we sample from the joint kernel $\bar{K}$ using Algorithm 1.

Given X_{t-1} and Y_{t-1}, Algorithm 1 gives the conditional marginal distributions

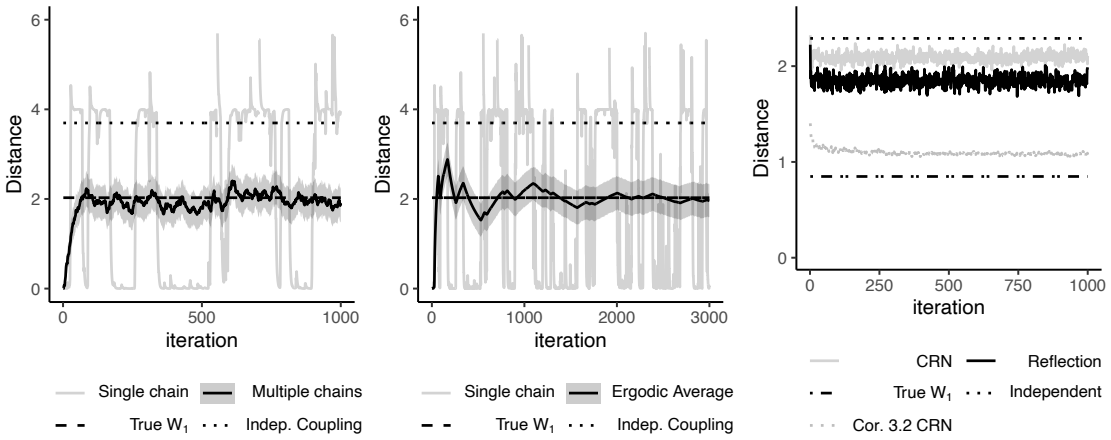
$$X_t | X_{t-1}, Y_{t-1} \sim K_1(X_{t-1}, \cdot) \qquad Z_t | X_{t-1}, Y_{t-1} \sim K_1(Y_{t-1}, \cdot) \qquad Y_t | X_{t-1}, Y_{t-1} \sim K_2(Y_{t-1}, \cdot). \tag{7}$$

Algorithm 1: Joint kernel $\bar{K}$ on $\mathcal{X} \times \mathcal{X}$, which couples 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) such that $(X_t, Z_t)|X_{t-1}, Y_{t-1} \sim \Gamma_1(X_{t-1}, Y_{t-1})$ and $(Z_t, Y_t)|X_{t-1}, Y_{t-1} \sim \Gamma_\Delta(Y_{t-1})$

return (X_t, Y_t)



(*a*) Single and averaged CRN trajectories for $\mathcal{W}_1$ with L_2 distance.

(*b*) Single and averaged CRN trajectories for $\mathcal{W}_1$ with L_2 distance.

(*c*) CRN and reflection coupling trajectories for $\mathcal{W}_1$ with L_2 distance.

Figure 1: Upper bounds based on single and multiple trajectories, and under different choices of coupling.

Often Algorithm 1 can be implemented to capture this dependency between X_t and Y_t given (X_{t-1}, Y_{t-1}) without explicitly sampling Z_t. As an example, consider Algorithm 1 for the coupled chain when K_1 and K_2 are MALA and ULA kernels with step-sizes σ_P and σ_Q and target distributions P and Q respectively. Γ_1 and Γ_Δ are chosen to be CRN coupled kernels. Given (X_{t-1}, Y_{t-1}), we sample the common random number $\epsilon_{CRN} \sim \mathcal{N}(0, I_d)$ and calculate 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 $Y_t = Y^*$. This CRN coupling of MALA and ULA is included in Algorithm 4 of Appendix 7, where each Z_t is not explicitly sampled. Appendix 7 also contains general CRN and reflection coupling between two Metropolis–Hastings kernel.

We conclude this section with a discussion on practical implementation and potential limitations.

Number of coupled chains and chain length to simulate. We first highlight the importance of simulating multiple coupled chains independently and running long chains to obtain reliable upper bound estimates. Figure 1(*a*) considers the performance of the our coupled chains to obtain 1-Wasserstein upper bounds between distributions P and Q on $\mathbb{R}^d$ for $d = 4$, given by

$$P = \frac{1}{2}\mathcal{N}(1_d, I_d) + \frac{1}{2}\mathcal{N}(-1_d, I_d) \quad \text{and} \quad Q = \mathcal{N}(1_d, I_d) \tag{8}$$

such 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 , 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 $d^{-1/6}$, and we initialize $X_0^{(i)} = 1_d$ and $Y_0^{(i)} = 1_d$ such that both marginal chains start at the common mode. Under this setup, Figure 1(*a*) of main text shows the trajectories of distance metric $c(x, y) = \|x-y\|_2$ induced by the L_2 norm. The grey solid 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 $\bar{c}(X_t, Y_t) := \sum_{i=1}^{I} c(X_t^{(i)}, Y_t^{(i)})/I$ for $I = 100$ independent chains. The grey solid 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 highlights that upper bound estimates 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 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). This highlights that upper bound estimates based on multiple chains are more reliable in the presence of multiple modes. We note that these multiple chains can be simulated in parallel, and that the memory requirements can be negligible as it suffices to only store a scalar trajectory $(c(X_t^{(i)}, Y_t^{(i)}))_{t\geq 1}^{T}$ of chain length T for each independent coupled chain i. Also even for upper bound estimates based on a single chain, the ergodic average $\sum_{t=1}^{T} c(X_t^{(1)}, Y_t^{(1)})/T$ for a sufficiently large chain length T can produce estimates with low variance, as shown by the grey confidence bands and the black solid line in Figure 1(*b*).

Choice of coupled kernel. Secondly, we highlight the importance of the choice of the coupled kernel $\bar{K}$. Figure 1(*c*) considers the performance of the our coupled chains to obtain 1-Wasserstein upper bounds between distributions P and Q on $\mathbb{R}$, given by

$$P = \frac{1}{2}\mathcal{N}(2, 1) + \frac{1}{2}\mathcal{N}(-2, 1) \quad \text{and} \quad Q = \frac{1}{2}\mathcal{N}(1, 1) + \frac{1}{2}\mathcal{N}(-1, 1), \tag{9}$$

such that now both the marginal target distributions are bimodal. Under this setup, we simulated coupled chains based on both a CRN coupling and a reflection coupling of MALA kernels K_1 and K_2 targeting distributions P and Q respectively. The MALA kernels have a common step-size 2, and we initialize such that each $X_0^{(i)} \sim P$ and $Y_0^{(i)} \sim Q$ are independent. In Figure 1(*c*), the grey and black solid lines show averaged trajectories from 100 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 valid upper bounds (as shown in Section 1.1), and in practice one can simulate different coupling algorithms to empirically assess which choice of coupling produces tighter upper bounds. Finally, Figure 1(c) highlights that our upper bounds may not always be very close to the true Wasserstein distance. Alternative coupling algorithms and Wasserstein distance upper bounds between mixtures of distributions (application of Corollary 1.2, as shown by the black dotted line) can give further improvements for this example. Details of the application of Corollary 1.2 here is included in Appendix 3.

1.3. Theoretical guarantees of upper bounds on Wasserstein distance

We have established the consistency of our estimators based on coupled chains to produce upper bounds on Wasserstein distances (Section 1.1), and have developed algorithms to sample these coupled chains (Sections 1.2). In this section, we establish theoretical guarantees of upper bounds generated from these algorithms. Our analysis is based on analytic perturbation theory for Markov chains in 1-Wasserstein distance (Pillai and Smith, 2015; Johndrow and Mattingly, 2018; Rudolf and Schweizer, 2018), and we generalize existing such results to p-Wasserstein distances for all $p \geq 1$. This is a useful extension, as distances such as 2-Wasserstein better reflect geometric features compared to the 1-Wasserstein (e.g. Villani, 2008, Remark 6.6). We also highlight examples where the upper bounds on the Wasserstein distance do not explicitly depend on the dimension of the state space, and are stable up to a coupling of the one-step marginal kernels.

To establish theoretical guarantees of our upper bounds, we assume the Markovian coupling Γ_1 in Algorithm 1 gives uniform contraction in Wasserstein distance.

Assumption 1.10 *There exists a constant $\rho \in (0,1)$ such that for all $X_t, Y_t \in \mathcal{X}$,*

$$\mathbb{E}[c(X_t, Y_t)^p | X_t, Y_t]^{1/p} \leq \rho c(X_t, Y_t) \text{ for } (X_{t+1}, Y_{t+1}) | (X_t, Y_t) \sim \Gamma_1(X_t, Y_t). \tag{10}$$

Assumption 1.10 is stronger than the convergence assumption of the marginal chain corresponding to kernel K_1 (Assumption 1.4 for the marginal distributions $(P_t)_{t\geq 0}$). For many popular MCMC algorithms, Assumption 1.10 has been established under certain metrics c and coupled kernel Γ_1 to give contraction rates ρ that do not explicitly depend on the dimension on the state space $\mathcal{X}$. This includes MALA (Eberle, 2014; Eberle and Majka, 2019), Hamiltonian Monte Carlo (HMC) (Bou-Rabee et al., 2020) and Pre-conditioned HMC (Bou-Rabee and Eberle, 2020). When the target distributions are log-concave, Assumption 1.10 can hold with metric $c(x,y) = \|x-y\|_2$ induced by L_2 norm and the coupled kernel Γ_1 based on a CRN coupling. For target distributions (including for example, multimodal distributions with Gaussian tails) which satisfy a weaker distant dissipativity condition (Eberle, 2016; Gorham et al., 2019), Assumption 1.10 can hold with transformed metrics of the form $\tilde{c}(x,y) = f(\|x-y\|_2)$ for some chosen increasing concave function f such that $\tilde{c}$ is equivalent to the original metric $c(x,y) = \|x-y\|_2$ with $r\tilde{c}(x,y) \leq c(x,y) \leq R\tilde{c}(x,y)$ for some constants $0 < r \leq R < \infty$, and the coupled kernel Γ_1 based on a combination of CRN and reflection coupling.

Further, we could directly weaken Assumption 1.10 to a geometric ergodicity condition as in (Rudolf and Schweizer, 2018), where for some $C \geq 1, \rho \in (0,1)$ 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 $(X_{t+L}, Y_{t+L})|(X_t, Y_t) \sim \Gamma_P^L(X_t, Y_t)$. Our analysis then is based on the construction of a multi-step coupling kernel. This may be of independent interest, and is included in Appendix 6 for completeness.

Under Assumption 1.10, 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 1.11 *Let $(X_t, Y_t)_{t\geq 0}$ denote a coupled Markov chain with initial distribution $\bar{I}_0$ and joint kernel $\bar{K}$ on $\mathcal{X} \times \mathcal{X}$ from Algorithm 1. Suppose the coupled kernel Γ_1 satisfies Assumption 1.10 for some $\rho \in (0, 1)$. Then for all $t \geq 0$,*

$$\mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} \mathbb{E}[\Delta_p(Y_{i-1})]^{1/p} \tag{11}$$

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

Remark 1.12 *It suffices to consider some metric $\tilde{c}$ which satisfies Assumption 1.10 and dominates the metric c of interest, such that $c(x, y) \leq R\tilde{c}(x, y)$ for some constant $R \in (0, \infty)$. Then $\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq R\mathbb{E}[\tilde{c}(X_t, Y_t)^p]^{1/p}$, and by Theorem 1.11 for $\tilde{c}$ we can obtain upper bounds with respect to metric c.*

When the marginal distributions $(Q_t)_{t\geq 0}$ converge, we obtain a simpler expression for the upper bound.

Corollary 1.13 *Under the setup and assumptions of Theorem 1.11, consider when the marginal distributions Q_t converge in p-Wasserstein distance to some distribution Q as $t \to \infty$. Then for all $\epsilon > 0$, there exists some $S \geq 1$ such that for all $t \geq S$,*

$$\mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1 - \rho^t)\frac{\mathbb{E}[\Delta_p(Y^*)]^{1/p}}{1 - \rho} + \epsilon. \tag{12}$$

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

Corollary 1.13 gives $\mathcal{W}_p(P, Q) \leq \liminf_{t\to\infty} \mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p]^{1/p} \leq \mathbb{E}[\Delta_p(Y^*)]^{1/p}/(1 - \rho)$. It implies that estimators from our coupled chains may give informative empirical upper bounds of $\mathcal{W}_p(P, Q)$ when kernels K_1 and K_2 are close such that 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 coupled chains may give upper bounds which are informative even in high dimensional settings.

The marginal distributions $(Q_t)_{t\geq 0}$ can correspond to an approximate Markov chain, and may not always converge to a limiting distribution. In such cases, we can upper bound the distance from our coupled chains in terms of perturbations between the marginal kernels weighted by a Lyapunov function of K_2.

Proposition 1.14 *Under the setup and assumptions of Theorem 1.11, 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 $z \in \mathcal{X}$ and for some fixed constants $\gamma \in [0, 1)$ and $L \in [0, \infty)$. Define*

$$\delta := \sup_{z \in \mathcal{X}} \left(\frac{\Delta_p(z)}{1 + V(z)^p} \right)^{1/p} \qquad \kappa := \left(1 + \max \left\{ \mathbb{E}[V(Y_0)^p], \frac{L}{1-\gamma} \right\} \right)^{1/p}. \tag{13}$$

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

$$\mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p]^{1/p} = \mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1 - \rho^t) \frac{\delta \kappa}{1 - \rho}. \tag{14}$$

Remark 1.15 *For Proposition 1.14 to be informative, we wish to find functions V such that $\delta\kappa$ is finite and small. In the case $p = 1$, Proposition 1.14 is related to Theorem 3.1 of (Rudolf and Schweizer, 2018).*

An application of these results on three simple examples related to MALA, ULA and Stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011) Markov chains is given in Appendix 5.

1.4. Comparison with related methods

In this section we compare our coupling based estimators of sample quality with related methods.

Linear programs and Sinkhorn distances. Given samples from the distributions P and Q, the Wasserstein distance between the corresponding empirical distributions may be used to estimate $\mathcal{W}_p(P, Q)$. Proposition 1.16 notes a bound for the non-asymptotic bias of such estimate.

Proposition 1.16 *Suppose P and Q are distributions on the metric space $(\mathcal{X}, c)$ with finite moments of order p. Let $\hat{P}_N$, $\tilde{P}_N$, $\hat{Q}_N$ and $\tilde{Q}_N$ denote empirical distributions of the samples $(X_1, \ldots, X_N)$, $(\tilde{X}_1, \ldots, \tilde{X}_N)$, $(Y_1, \ldots, Y_N)$ and $(\tilde{Y}_1, \ldots, \tilde{Y}_N)$ respectively, where $X_i \sim P$, $\tilde{X}_i \sim P$, $Y_i \sim Q$ and $\tilde{Y}_i \sim Q$ for all $i = 1, \ldots, N$. Suppose $(X_1, \ldots, X_N)$ and $(Y_1, \ldots, Y_N)$ are independent, $(X_1, \ldots, X_N)$ and $(\tilde{X}_1, \ldots, \tilde{X}_N)$ are independent, and $(Y_1, \ldots, Y_N)$ and $(\tilde{Y}_1, \ldots, \tilde{Y}_N)$ are independent. Then,*

$$\mathbb{E}\big[\mathcal{W}_p(\hat{P}_T, \hat{Q}_T)^p\big]^{1/p} - \Big(\mathbb{E}\big[\mathcal{W}_p(\hat{P}_T, \tilde{P}_T)^p\big]^{1/p} + \mathbb{E}\big[\mathcal{W}_p(\hat{Q}_T, \tilde{Q}_T))^p\big]^{1/p}\Big) \leq \mathcal{W}_p(P, Q) \leq \mathbb{E}\big[\mathcal{W}_p(\hat{P}_T, \hat{Q}_T)^p\big]^{1/p}. \tag{15}$$

We can consider the computational cost and the statistical performance of such empirical Wasserstein's distance based estimates. Calculating $\mathcal{W}_p(\hat{P}_N, \hat{Q}_N)$ corresponds to an uncapacitated minimum cost flow problem and requires $\mathcal{O}(N^3 \log N)$ computational cost (Orlin, 1988), which can be prohibitive for large sample sizes. An alternative would be to add entropic regularization and apply Sinkhorn's algorithm (Cuturi, 2013), which involves a regularization parameter $\lambda > 0$. As λ approaches zero, the induced distance from the optimal matching of the regularized problem approaches the induced distance from an Wasserstein optimal matching. However, using smaller values of λ leads to more expensive $\mathcal{O}(N^2/(\lambda\epsilon))$

computation time for ϵ-accurate matchings (Altschuler et al., 2017; Dvurechensky et al., 2018) and potential instability of the Sinkhorn's algorithm. For empirical distributions $\hat{P}_N$ and $\hat{Q}_N$ based on N independent samples from P and Q, in the worst case $\mathcal{W}_p(\hat{P}_N, \hat{Q}_N)$ converges to $\mathcal{W}_p(P, Q)$ at rate $\Omega(N^{-1/d})$ for dimension $d > 2p$ (e.g. Dudley, 1969; Weed and Bach, 2019; Lei, 2020). This can lead to the empirical Wasserstein distance giving loose upper bounds of $\mathcal{W}_p(P, Q)$ when the number of samples does not increase exponentially with dimension.

In comparison to such estimators based on empirical distributions of independent samples, our coupling based estimators do not require solving any expensive optimization problems. Furthermore, in cases when Assumption 1.10 is satisfied with favorable dependence on the dimension, our estimates do not necessarily suffer from a curse of dimensionality. On the other hand, such linear program based estimates converge to the true Wasserstein distance as the number of samples tend to infinity. Therefore if one has access to a substantially large computational budget and solving such linear programs with much larger sample sizes is feasible, then the linear program based estimates will produce tighter upper bounds.

Comparison with the approach of Dobson et al. Related work by Dobson et al. (Dobson et al., 2019) apply couplings to assess the sample quality of discretization of stochastic differential equations. Our approach avoids the challenging problem of contraction-constant estimation required in (Dobson et al., 2019). In a future revised version of this manuscript, we intend to give a comparison of our approach and (Dobson et al., 2019).

Comparison with the approach of Huggins et al. Related work by Huggins et al. (Huggins et al., 2019) derive analytic upper bounds on Wasserstein distances in terms of divergences, and estimate such upper bounds using importance sampling. Accurate estimation using importance sampling requires a number of samples that grow exponentially with the Kullback-Leibler divergence (Agapiou et al., 2017; Chatterjee and Diaconis, 2018). Our upper bound estimators in comparison do not require importance sampling, and can remain effective even in high dimensions if the marginal Markov chains are fast-mixing. In a future revised version of this manuscript, we intend to give a comparison of our approach and (Huggins et al., 2019).

2. Applications

In this section, we illustrate our methods on three important applications.

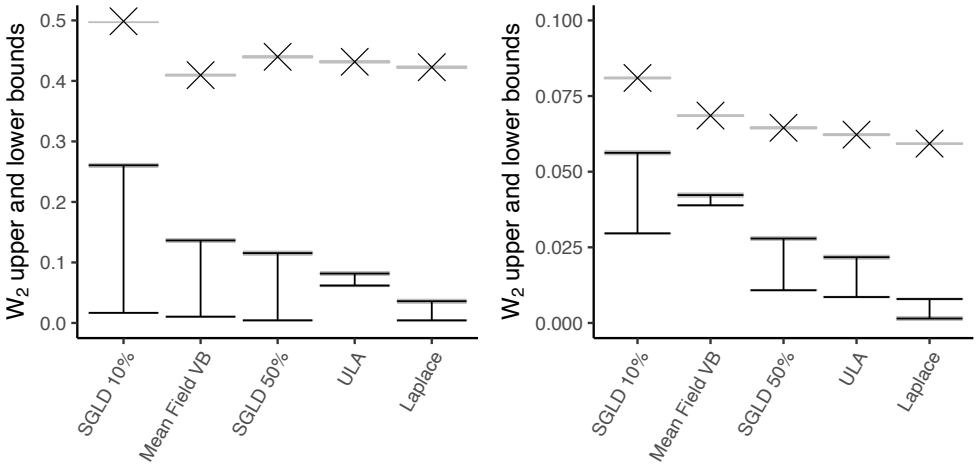
2.1. Stochastic Gradient MCMC and variational inference for tall data

Our first application concerns Bayesian inference for *tall* datasets (Bardenet et al., 2017), where the number of observations n is large compared to dimension d. In such settings, exact MCMC algorithms can be computationally expensive with $\mathcal{O}(n)$ cost per iteration. This computational bottleneck and the prevalence of tall datasets have catalysed recent interest in approximate MCMC and variational approximation based algorithms. Approximate MCMC algorithms include ULA, popular stochastic gradient based variants such as Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) (see (Nemeth and Fearnhead, 2021) for a recent review of stochastic gradient MCMC), and related algorithms based on deterministic approximations to the likelihood function (Huggins et al., 2016; Campbell and

Broderick, 2018). Popular variational approximation methods include Laplace's approximation, variational Bayes (see (Blei et al., 2017) for a review), and related variants (e.g. Hoffman et al., 2013). In this section, we use couplings to assess the quality of these sampling methods in the tall data regime. We consider ULA, SGLD, Laplace's approximation and mean field variational Bayes applied to Bayesian logistic regression for the Pima Indians Diabetes dataset (Dua and Graff, 2017) with $n = 768$ observations and $d = 8$ covariates and the DS1 life sciences dataset (Komarek and Moore, 2003) with $n = 26732$ observations and $d = 10$ covariates.

Figures 2(a) and 2(b) highlight the Wasserstein distance upper bound estimates based on coupled chains. In particular, each line corresponds to the averaged trajectory $(\bar{c}(X_t, Y_t)^2)_{t\geq 0}$, where $\bar{c}(X_t, Y_t)^2 := \sum_{i=1}^{I} c(X_t^{(i)}, Y_t^{(i)})^2/I$ for $I = 100$ independent chains, each trajectory $(X_t^{(i)})_{t\geq 0}$ is an exact MCMC MALA chain targeting the posterior P and each trajectory $(Y_t^{(i)})_{t\geq 0}$ is linked to an approximate MCMC or a variational approximation algorithm. In particular, the black solid, black dashed and black doted lines correspond to when each $(Y_t^{(i)})_{t\geq 0}$ is an ULA chain, an SGLD chain based on sub-sampling 50% of the observations and an SGLD chain based on sub-sampling 10% of the observations respectively all targeting P. The grey solid and grey dashed lines correspond to when each $(Y_t^{(i)})_{t\geq 0}$ is an MALA chain targeting $\mathcal{N}(\mu_L, \Sigma_L)$ and $\mathcal{N}(\mu_{MFVB}, \Sigma_{MFVB})$ respectively. The parameters $\mu_L \in \mathbb{R}^d, \Sigma_L \in \mathbb{R}^{d\times d}$ are obtained from a Laplace approximation of π, and parameters $\mu_{MFVB} \in \mathbb{R}^d$, $\Sigma_{MFVB} \in \mathbb{R}^{d\times d}$ (where Σ_{MFVB} is a diagonal matrix) are obtained from a mean field variational Bayes approximation of P with a Gaussian family. In all these cases, we consider a CRN coupling between the kernels corresponding to the marginal chains $(X_t^{(i)})_{t\geq 0}$ and $(Y_t^{(i)})_{t\geq 0}$ with common step-sizes. Figures 2(a) and 2(b) both highlight the promising performance of Laplace's approximation for such tall data problems. By considering the limiting distance of the trajectories, we obtain an informative upper bound estimate of approximately 10^{-3} and 10^{-7} for $\mathcal{W}_2^2\big(P, \mathcal{N}(\mu_L, \Sigma_L)\big)$ for the Pima and DS1 datasets respectively. This promising performance of Laplace's approximations can be linked to concentration of the posterior and accuracy of the corresponding Bernstein-von Mises approximation (Bardenet et al., 2017), and has been noted before for logistic regression (Chopin and Ridgway, 2017). Our bounds also highlight how the Metropolis–Hastings correction and stochastic gradients affect sample quality for ULA and SGLD for a fixed step-size.

Lastly, we note that the true Wasserstein distance may have a different ordering compared to our coupling based upper bounds. To check this, in Figure 2 we plot our 2-Wasserstein upper bound alongside the lower bound of Gelbrich (Gelbrich, 1990), based on the empirical mean and covariance estimates of the marginal chains. This allows comparison between approximate MCMC and variational approximation based methods, and highlights the promising empirical performance of our upper bounds. Furthermore, even in cases when the true Wasserstein distances are not known or the lower bounds are challenging to estimate, our upper bounds remain useful for the researcher as a geometrically faithful measure of sample quality.



(*a*) $\mathcal{W}_2$ with L_2 distance metric upper and lower bounds for Pima Indians dataset.

(*b*) Single and averaged CRN trajectories for $\mathcal{W}_1$ with L_2 distance.

Figure 2: Wasserstein distance upper bounds of Stochastic Gradient MCMC and Gaussian variational approximations for Bayesian Logistic regression.

2.2. 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. The likelihood is given by $L(\beta; y, X, \sigma^2) = \mathcal{N}(y; X\beta; \sigma^2 I_n)$ where $\mathcal{N}(\cdot; X\beta, \sigma^2 I_n)$ denotes the probability density function of $\mathcal{N}(X\beta; \sigma^2 I_n)$ on $\mathbb{R}^n$, $y \in \mathbb{R}^n$ is the observed response vector, $X \in \mathbb{R}^{n\times d}$ is the observed design matrix, $\beta \in \mathbb{R}^d$ unknown signal vector that is assumed to be sparse, and $\sigma^2 > 0$ is the unknown noise variance. We consider a class of global-local mixture priors in this setting, given by

$$\xi^{-1/2} \sim Cauchy_+(0,1) \quad \eta_j^{-1/2} \overset{i.i.d.}{\sim} t_+(\nu) \quad \sigma^{-2} \sim \text{Gamma}\Big(\frac{a_0}{2}, \frac{b_0}{2}\Big) \quad \beta_j|\eta, \xi, \sigma^2 \overset{ind.}{\sim} \mathcal{N}\Big(0, \frac{\sigma^2}{\xi\eta_j}\Big) \tag{16}$$

where $Cauchy_+(0,1)$ is the half-Cauchy distribution on $[0,\infty)$ and $t_+(\nu)$ is the half-t distribution on $[0,\infty)$ with ν degress of freedom. When $\nu = 1$, this corresponds to the popular Horseshoe prior (Carvalho et al., 2009, 2010; Bhadra et al., 2019). This setting differs considerably from the log-concave tall data example (Section 2.1), as now the posterior distribution is multi-modal, has polynomial tails along directions in the null space of the design matrix X and has infinite density about the origin (Biswas et al., 2021). The state-of-the-art exact MCMC algorithms to sample from the posterior are Gibbs samplers which cost $\mathcal{O}(n^2 d)$ per iteration (Bhattacharya et al., 2016). This computation cost arises from a weighted matrix product calculation of the form $X\,\text{Diag}(\eta_t)^{-1}X^T$ 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 (ν=1), approximate MCMC methods have been recently developed for this problem

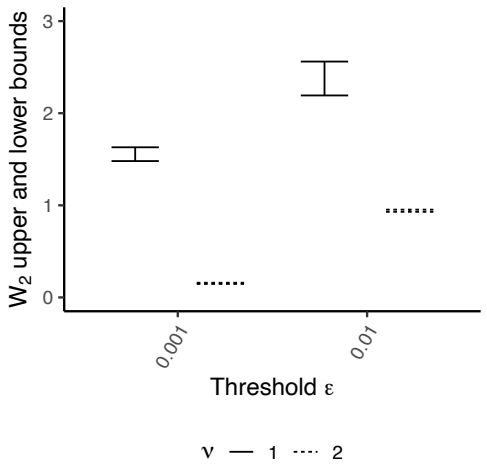


Figure 3: $\mathcal{W}_2$ with L_2 distance metric upper and lower bounds.

(Johndrow et al., 2020), which are based on the weighted matrix product $X\,\mathrm{Diag}(\tilde{\eta}_t)^{-1}X^T$, where $\tilde{\eta}_{j,t} := \eta_{j,t}\mathbb{1}_{\{\eta_{j,t}>\epsilon\}}$ is a truncated approximation of η_t for some small threshold $\epsilon > 0$.

In this section, we use couplings to measure the sample quality of the approximate MCMC algorithm. We consider a synthetic dataset with $n = 100$ and dimension $d = 5,000$. For the half-t prior in (16), 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 marginal kernels are based on the MCMC algorithms introduced in (Johndrow et al., 2020; Biswas et al., 2021)

Figure 3 highlights our results. We plot the corresponding Wasserstein upper bound and lower bounds of (Gelbrich, 1990). The upper bound and lower bounds are based on 10 independent coupled chains each of length $5,000$ with a burn-in of 500 iterations. This highlights how considering a higher value of ν (corresponding to a more concentrated prior about zero) can improve the quality of the approximate MCMC samples for the same threshold parameter ϵ. This can be linked to Theorem 1.11, and a higher degree of freedom ν giving a faster contraction under a CRN Markovian coupling of the exact MCMC kernel, as empirically observed in (Biswas et al., 2021).

2.3. Approximate MCMC for high-dimensional logistic regression

In this section, we consider high-dimensional Bayesian logistic regression with spike and slab priors. The likelihood is given by $L(\beta; y, X) = \prod_{i=1}^n (1 + \exp(-y_i x_i^T \beta))^{-1}$ where $y \in \mathbb{R}^n$ is the observed response vector with components $y_i \in \{0, 1\}$, $X \in \mathbb{R}^{n\times d}$ is the scaled design matrix with rows x_i^T, and $\beta \in \mathbb{R}^d$ is the unknown signal vector that is assumed to be sparse. We consider a spike and slab prior for Bayesian variable selection is this setting, given by

$$Z_j \overset{i.i.d.}{\sim} \text{Bernoulli}(q) \qquad \beta_j | Z_j = 0 \sim \mathcal{N}(0, \tau_0^2) \qquad \beta_j | Z_j = 1 \sim \mathcal{N}(0, \tau_1^2) \tag{17}$$

for $j = 1, \ldots, d$ where $q \in (0,1), \tau_0 > 0, \tau_1 > 0$ are hyper-parameters. We take $\tau_0 \ll \tau_1$ such that $Z_i = 0$ and $Z_i = 1$ correspond to null and a non-null components β_j respectively. Spike and slab priors have been commonly used for Bayesian variable selection (George and McCulloch, 1993; Ishwaran and Rao, 2005; Narisetty and He, 2014). By consider the posterior distributions of each variable Z_j on $\{0,1\}$, spike and slab priors provide an easily interpretable method for Bayesian variable selection. The state-of-the-art exact MCMC algorithms to sample from the posterior are Gibbs samplers which cost $\mathcal{O}(n^2 d)$ per iteration (Bhattacharya et al., 2016). Recently, (Narisetty et al., 2019) have been developed approximate MCMC methods for this logistic regression setting. For their approximate MCMC algorithm, (Narisetty et al., 2019) consider matrix approximations of the form

$$\begin{pmatrix} X_A^T X_A + \tau_1^{-2} I & X_A^T X_I \\ X_I^T X_A & X_I^T X_I + \tau_0^{-2} I \end{pmatrix} \approx \begin{pmatrix} X_A^T X_A + \tau_1^{-2} I & 0 \\ 0 & (n + \tau_0^{-2}) I \end{pmatrix} \tag{18}$$

where $A = \{j : Z_j = 1\}$, X_A is an $n \times |A|$ matrix corresponding the active columns $j \in A$ of the design matrix, and X_I is an $n \times (p - |A|)$ matrix corresponding the inactive columns $j \notin A$. The algorithm also has a corresponding correction step which ensures that the approximate chain converges to a desirable target distribution. The approximate MCMC algorithm has overall computation cost of $\mathcal{O}(n \min\{p, |A|^2\})$ per iteration.

In this section, we use couplings to measure the sample quality of the approximate MCMC algorithm. We consider a synthetic dataset with $n = 100$ and dimensions $d \in \{100, 200, 300, 400\}$. For the spike and slab prior in (17), we consider a CRN coupling on one marginal chain corresponding to the exact MCMC kernel and the other chain corresponding to the approximate MCMC kernel. The upper bound and lower bounds are then based on 10 independent coupled trajectories each of length $5,000$ with a burn-in of $5,000$ iterations.

Figure 4 highlights our results. We consider the 1-Wasserstein distance with respect to the Hamming distance on $Z \in \{0,1\}^d$. The upper bounds are based on our coupled chains. For the lower bounds, note that the Hamming distance equals square of the L_2 distance metric, so the bounds of (Gelbrich, 1990) are applicable. Figure **??** shows that the upper and lower bounds for the 1-Wasserstein distance increase with dimension on synthetic datasets, highlight the approximate chain may have worsening sample quality in higher dimensions for this example.

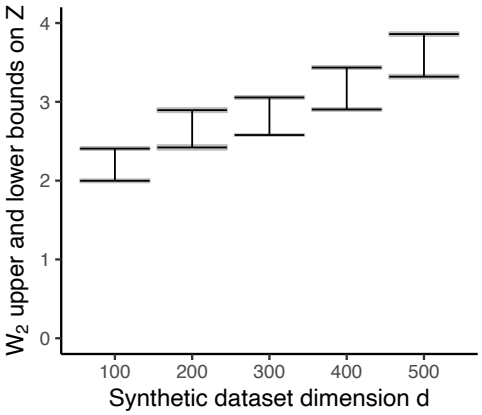


Figure 4: $\mathcal{W}_1$ upper and lower bounds with respect to the Hamming distance for Z on $\{0,1\}^d$ for Bayesian logistic regression with the spike and slab prior. $n = 100$ and varying dimension d.

References

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance Sampling: Intrinsic Dimension and Computational Cost. *Statistical Science*, 32(3):405 – 431, 2017. doi: 10.1214/17-STS611. URL https://doi.org/10.1214/17-STS611.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1961–1971, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *J. Mach. Learn. Res.*, 18(1):1515–1557, January 2017. ISSN 1532-4435.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

Anirban Bhattacharya, Antik Chakraborty, and Bani K. Mallick. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991, 10 2016. ISSN 0006-3444. doi: 10.1093/biomet/asw042. URL https://doi.org/10.1093/biomet/asw042.

Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings. In *Advances in Neural Information Processing Systems*, pages 7389–7399, 2019.

Niloy Biswas, Anirban Bhattacharya, Pierre E. Jacob, and James E. Johndrow. Coupled markov chain monte carlo for high-dimensional regression with half-t priors. *arXiv preprint arXiv:2012.04798v2*, 2021.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

N. Bou-Rabee and M. Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 03 2012. ISSN 0272-4979. doi: 10.1093/imanum/drs003. URL https://doi.org/10.1093/imanum/drs003.

Nawaf Bou-Rabee and Andreas Eberle. Two-scale coupling for preconditioned hamiltonian monte carlo in infinite dimensions. *Stochastics and Partial Differential Equations: Analysis and Computations*, 2020. doi: 10.1007/s40072-020-00175-6. URL https://doi.org/10.1007/s40072-020-00175-6.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for hamiltonian monte carlo. *Ann. Appl. Probab.*, 30(3):1209–1250, 06 2020. doi: 10.1214/19-AAP1528. URL https://doi.org/10.1214/19-AAP1528.

Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 698–706, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/campbell18a.html.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In David van Dyk and Max Welling, editors, *Proceedings of Machine Learning Research*, volume 5, pages 73–80, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/carvalho09a.html.

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

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099 – 1135, 2018. doi: 10.1214/17-AAP1326. URL https://doi.org/10.1214/17-AAP1326.

Nicolas Chopin and James Ridgway. Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation. *Statistical Science*, 32(1):64 – 87, 2017. doi: 10.1214/16-STS581. URL https://doi.org/10.1214/16-STS581.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Eustasio del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926 – 951, 2019. doi: 10.1214/18-AOP1275. URL https://doi.org/10.1214/18-AOP1275.

Matthew Dobson, Jiayu Zhai, and Yao Li. Using coupling methods to estimate sample quality for stochastic differential equations. *arXiv preprint arXiv:1912.10339*, pages 1–29, 2019. URL http://arxiv.org/abs/1912.10339.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

R. M. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40(1):40 – 50, 1969. doi: 10.1214/aoms/1177697802. URL https://doi.org/10.1214/aoms/1177697802.

Alain Durmus and Éric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 11 2019. doi: 10.3150/18-BEJ1073. URL https://doi.org/10.3150/18-BEJ1073.

Alain Durmus, Andreas Eberle, Aurélien Enfroy, Arnaud Guillin, and Pierre Monmarché. Discrete sticky couplings of functional autoregressive processes. *arXiv preprint arXiv:2104.06771*, 2021.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/dvurechensky18a.html.

Andreas Eberle. Error bounds for metropolis–hastings algorithms applied to perturbations of gaussian measures in high dimensions. *Ann. Appl. Probab.*, 24(1):337–377, 02 2014. doi: 10.1214/13-AAP926. URL https://doi.org/10.1214/13-AAP926.

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

Andreas Eberle and Mateusz B. Majka. Quantitative contraction rates for markov chains on general state spaces. *Electron. J. Probab.*, 24:36 pp., 2019. doi: 10.1214/19-EJP287. URL https://doi.org/10.1214/19-EJP287.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: https://doi.org/10.1002/mana.19901470121. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mana.19901470121.

Edward I. George and Robert E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. doi: 10.1080/01621459.

1993.10476353. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353.

Jackson Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928, 10 2019. doi: 10.1214/19-AAP1467. URL https://doi.org/10.1214/19-AAP1467.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013. URL http://jmlr.org/papers/v14/hoffman13a.html.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4087–4095, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Jonathan H. Huggins, Trevor Campbell, Mikolaj Kasprzak, and Tamara Broderick. Scalable gaussian process inference with finite-data mean and variance guarantees. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 796–805. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/huggins19a.html.

Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730 – 773, 2005. doi: 10.1214/009053604000001147. URL https://doi.org/10.1214/009053604000001147.

Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020. doi: 10.1111/rssb.12336. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12336.

James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61, 2020. URL http://jmlr.org/papers/v21/19-536.html.

James E. Johndrow and Jonathan C. Mattingly. Error bounds for approximations of markov chains used in bayesian sampling. *arXiv preprint arXiv:1711.05382*, 2018.

Paul Komarek and Andrew Moore. Fast robust logistic regression for large sparse datasets with binary outputs. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, 2003. URL http://komarix.org/ac/ds/.

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767 – 798, 2020. doi: 10.3150/19-BEJ1151. URL https://doi.org/10.3150/19-BEJ1151.

Torgny Lindvall and L. C. G. Rogers. Coupling of multidimensional diffusions by reflection. *Ann. Probab.*, 14(3):860–872, 07 1986. doi: 10.1214/aop/1176992442. URL https://doi.org/10.1214/aop/1176992442.

Naveen N. Narisetty, Juan Shen, and Xuming He. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*, 114 (527):1205–1217, 2019. doi: 10.1080/01621459.2018.1482754. URL https://doi.org/10.1080/01621459.2018.1482754.

Naveen Naidu Narisetty and Xuming He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789 – 817, 2014. doi: 10.1214/14-AOS1207. URL https://doi.org/10.1214/14-AOS1207.

Christopher Nemeth and Paul Fearnhead. Stochastic gradient markov chain monte carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021. doi: 10.1080/01621459.2020.1847120. URL https://doi.org/10.1080/01621459.2020.1847120.

James Orlin. A faster strongly polynomial minimum cost flow algorithm. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, page 377–387, New York, NY, USA, 1988. Association for Computing Machinery. ISBN 0897912640. doi: 10.1145/62212.62249. URL https://doi.org/10.1145/62212.62249.

John O'Leary, Guanyang Wang, and Pierre E. Jacob. Maximal Couplings of the Metropolis–Hastings algorithm. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS) (to appear)*, Proceedings of Machine Learning Research. PMLR, 2021.

Natesh S. Pillai and Aaron Smith. Ergodicity of approximate mcmc chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*, 2015.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 11 2018. doi: 10.3150/17-BEJ938. URL https://doi.org/10.3150/17-BEJ938.

Cédric Villani. *Optimal transport – Old and new*, volume 338. Springer, 2008. doi: 10.1007/978-3-540-71050-9.

Jonathan Weed and Francis 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.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

3. Additional calculations

Calculation of linear program upper bounds in Figure ?? of main text. In this section we note how the linear program upper bounds and error bands in Figure **??** of main text are generated. Our upper bounds are based on Proposition 1.16, which gives

$$\mathcal{W}_p(P,Q)^p \leq \mathbb{E}\big[\mathcal{W}_p(\hat{P}_T,\hat{Q}_T)^p\big] \tag{19}$$

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 $(X_1, \ldots, X_T)$ and $(Y_1, \ldots, Y_T)$ where $X_i \sim P$ and $Y_i \sim Q$ for all $i = 1, \ldots, T$. For $p = 2$ and $P \neq Q$, the dot-dashed lines in Figure **??** of main text plots the estimate

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

of this upper bound, where $\hat{P}_T^{(i)}$ and $\hat{Q}_T^{(i)}$ are empirical distribution of P and Q respectively based on T samples. For each $i = 1, \ldots, I$, 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 $\hat{\sigma}/\sqrt{I}$ for $\hat{\sigma}^2$ defined as the variance of $\big(\mathcal{W}_2(\hat{P}_T^{(i)}, \hat{Q}_T^{(i)})^2\big)_{i=1}^{I}$, corresponding to one standard deviation of the upper bound estimate.

For large T and $P \neq Q$, we could also employ a central limit theorem of (del Barrio and Loubes, 2019) for the 2-Wasserstein distance between empirical distributions of independent samples on $\mathbb{R}^d$ under the L_2 distance metric.

Theorem 3.1 *(del Barrio and Loubes, 2019, Theorem 4.1) Consider the 2-Wasserstein distance on $\mathbb{R}^d$ with metric c as the L_2 distance. Suppose P and Q are distributions on $\mathbb{R}^d$ each with finite moments of order $4 + \delta$ for some $\delta > 0$ and positive density in the interior of its convex support. Let $\hat{P}_T$ and $\hat{Q}_T$ denote empirical distributions corresponding to two independent sets of independent samples of size T from P and Q respectively. Then,*

$$\sqrt{T}\Big(\mathcal{W}_2(\hat{P}_T, \hat{Q}_T)^2 - \mathbb{E}\big[\mathcal{W}_2(\hat{P}_T, \hat{Q}_T)^2\big]\Big) \overset{T \to \infty}{\Rightarrow} \mathcal{N}\big(0, \sigma^2(P, Q) + \sigma^2(Q, P)\big), \tag{20}$$

where $\sigma^2(P, Q) := \mathrm{Var}\big(\|X\|_2^2 - 2\phi_0(X)\big)$ for $X \sim P$ and ϕ_0 is the optimal transport potential from P to Q (such that $\nabla\phi_0$ is an optimal transportation map from P to Q), and $\sigma^2(Q, P) := \mathrm{Var}\big(\|Y\|_2^2 - 2\phi_1(Y)\big)$ for $Y \sim Q$ and ϕ_1 is the optimal transport potential from Q to P.

3.1. Stylized example calculations

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

$$Y_t = (I_d - (\sigma_Q^2/2)\Sigma^{-1})Y_{t-1} + \sigma_Q Z_t = BY_{t-1} + \sigma_Q Z_t \tag{21}$$

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

$$Y_t = B^t Y_0 + \sigma_Q\Big(B^{t-1}Z_1 + B^{t-2}Z_2 + \cdots + BZ_{t-1} + Z_t\Big) = \sigma_Q \sum_{j=0}^{t-1} B^j Z_{t-j} \sim \mathcal{N}\big(0, \sigma_Q^2 \sum_{j=0}^{t-1} B^{2j}\big) =: Q_t \tag{22}$$

as required. Finally, note that for $\sigma_Q = 0.5d^{-1/6}$ sufficiently small such that $\|B\|_2 < 1$ (where $\|\cdot\|_2$ is the matrix operator norm), $\lim_{t\to\infty} \sum_{j=0}^{t-1} B^{2j} = (I_d - B^2)^{-1}$. This gives $Q_t \overset{t \to \infty}{\Rightarrow} \mathcal{N}(0, \sigma_Q^2(I_d - B^2)^{-1}) =: Q$.

ULA asymptotic bias upper bound calculation. We recall a result on the asymptotic bias of ULA.

Proposition 3.2 *(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 L_2 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 $\gamma := \sigma^2/2 < 1/(m+L)$. Then the ULA Markov chain converges to some distribution π_γ, and

$$\mathcal{W}_2(\pi, \pi_\gamma)^2 \leq 2\kappa^{-1}\gamma^2 d\Big(2L^2 + \gamma L^4\big(\frac{\gamma}{6} + \frac{1}{m}\big) + \kappa^{-1}\big(\frac{4d\tilde{L}^2}{3} + \gamma L^4 + \frac{4L^4}{3m}\big)\Big) \tag{23}$$

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

The dotted line in Figure **??** of main text is plotted by applying (23) 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.

Application of Corollary 1.2 in Figure 1(*c*). By Corrollary 1.2 applied to the targets in Equation 9, we obtain

$$\mathcal{W}_p(P, Q)^p = \mathcal{W}_p\Big(\frac{1}{2}\mathcal{N}(2, 1) + \frac{1}{2}\mathcal{N}(-2, 1), \frac{1}{2}\mathcal{N}(1, 1) + \frac{1}{2}\mathcal{N}(-1, 1)\Big)^p \tag{24}$$
$$\leq \frac{1}{2}\Big(\mathcal{W}_p\big(\mathcal{N}(2, 1), \mathcal{N}(1, 1)\big)^p + \mathcal{W}_p\big(\mathcal{N}(-2, 1), \mathcal{N}(-1, 1)\big)^p\Big) \tag{25}$$
$$\leq \frac{1}{2}\Big(\mathbb{E}\big[\widehat{\mathcal{W}}_p^{(UB)}(\mathcal{N}(2, 1), \mathcal{N}(1, 1))^p\big] + \mathbb{E}\big[\widehat{\mathcal{W}}_p^{(UB)}(\mathcal{N}(-2, 1), \mathcal{N}(-1, 1))^p\big]\Big). \tag{26}$$

The black solid line in Figure 1(*c*) was obtained by calculating $\widehat{\mathcal{W}}_p^{(UB)}(\mathcal{N}(-2, 1), \mathcal{N}(-1, 1))$ based on a CRN coupling of MALA kernels targeting $\mathcal{N}(-2, 1)$ and $\mathcal{N}(-1, 1)$ marginally with a common step-size 2, and by calculating $\widehat{\mathcal{W}}_p^{(UB)}(\mathcal{N}(2, 1), \mathcal{N}(1, 1))$ based on a CRN coupling of MALA kernels targeting $\mathcal{N}(2, 1)$ and $\mathcal{N}(1, 1)$ marginally also with a common step-size 2.

4. Proofs

4.1. Consistency proofs

Technical Results. We first record some technical results.

Lemma 4.1 *Let $(a_j)_{j\geq 0}$ be a sequence on $\mathbb{R}$ such that $a_j \overset{j\to\infty}{\to} 0$, and let $\rho \in (0,1)$. Then $\sum_{j=1}^{t} \rho^{t-j} a_j \overset{t\to\infty}{\to} 0$.*

Proof 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

$$\Big|\sum_{j=1}^{t} \rho^{t-j} a_j\Big| \leq \sum_{j=1}^{j_0} \rho^{t-j}|a_j| + \sum_{j=j_0+1}^{t} \rho^{t-j}|a_j| \leq M\rho^{t-j_0}\frac{1-\rho^{j_0}}{1-\rho} + \epsilon\frac{1-\rho^{t-j_0}}{1-\rho}. \tag{27}$$

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

Lemma 4.2 *(Gluing lemma) (Villani, 2008, Chapter 1) Let μ_i be probability measures on the Polish measurable spaces $(\mathcal{X}_i, \mathcal{B}(\mathcal{X}_i))$ for $i = 1,\ldots,3$. Let X_1, X_2, Y_2, Y_3 be random variables such that (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3). Then, there exists random variables Z_1, Z_2, Z_3 such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3).*

Proof [Proof of Proposition 1.1] 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

$$\mathcal{W}_p(P_t, Q_t)^p \leq \mathbb{E}[c(X_t, Y_t)^p] = \mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)^p], \tag{28}$$

where the inequality follows from the coupling representation of Wasserstein distance, and the equality follows from the definition of $\widehat{\mathcal{W}}_p^{(UB)}(P_t, Q_t)$. ■

Proof [Proof of Corollary 1.2] It suffices to check that $\mathcal{W}_p(\frac{1}{T}\sum_{t=1}^{T} P_t, \frac{1}{T}\sum_{t=1}^{T} Q_t)^p \leq \frac{1}{T}\sum_{t=1}^{T} \mathcal{W}_p(P_t, Q_t)^p$ and apply Proposition 1.1. Let γ_t denote the p-Wasserstein optimal coupling between distributions P_t and Q_t for $t = 1,\ldots,T$. Sample the coupling (X, Y) such that $(X, Y)|I = t \sim \gamma_t$ for $I \sim \text{Uniform}(\{1,\ldots,T\})$. Then $X \sim \frac{1}{T}\sum_{t=1}^{T} P_t$ and $Y \sim \frac{1}{T}\sum_{t=1}^{T} Q_t$ marginally, and

$$\mathcal{W}_p(\frac{1}{T}\sum_{t=1}^{T} P_t, \frac{1}{T}\sum_{t=1}^{T} Q_t)^p \leq \mathbb{E}[c(X,Y)^p] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[c(X,Y)^p|I=t] = \frac{1}{T}\sum_{t=1}^{T} \mathcal{W}_p(P_t, Q_t)^p \tag{29}$$

as required. ■

Proof [Proof of Corollary 1.3] 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,

$$\mathcal{W}_p(P,Q)^p = \frac{1}{T-S}\sum_{t=S+1}^{T}\mathcal{W}_p(P_t,Q_t)^p \le \frac{1}{T-S}\sum_{t=S+1}^{T}\mathbb{E}[c(X_t,Y_t)^p] = \mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P,Q)^p]. \tag{30}$$

where the first equality follows as $P_t = P$ and $Q_t = Q$ for all $t \ge 0$, the inequality follows Proposition 1.1, and the last equality follows from the definition of $\widehat{\mathcal{W}}_p^{(UB)}(P,Q)$. ■

Proof [Proof of Proposition 1.5] 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 1.4, distributions $(P_t)_{t\ge0}$, $(Q_t)_{t\ge0}$, P and Q all have finite moments of order p. Then for all $t \ge 1$,

$$\mathcal{W}_p(P,Q) \le \mathcal{W}_p(P,P_t) + \mathcal{W}_p(P_t,Q_t) + \mathcal{W}_p(Q_t,Q) \tag{31}$$

$$\le \mathcal{W}_p(P,P_t) + \mathbb{E}[c(X_t,Y_t)^p]^{1/p} + \mathcal{W}_p(Q_t,Q), \tag{32}$$

where (31) 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 (32) follows from the coupling representation of $\mathcal{W}_p$. By Assumption 1.4, $\lim_{t\to\infty}\mathcal{W}_p(P,P_t) = 0$ and $\lim_{t\to\infty}\mathcal{W}_p(Q_t,Q) = 0$. Taking the limit infimum in (32) 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

$$\mathcal{W}_p(P,Q)^p \le \epsilon + \frac{1}{T-S}\sum_{t=S+1}^{T}\mathbb{E}[c(X_t,Y_t)^p] = \epsilon + \frac{1}{T-S}\sum_{t=S+1}^{T}\mathbb{E}\big[\widehat{\mathcal{W}}^{(UB)}(P_t,Q_t)^p\big] \tag{33}$$

for all $T \ge S$. ■

Proof [Proof of Proposition 1.9] By the triangle inequality,

$$\mathcal{W}_1(P,Q) \le \mathcal{W}_1(P_t,Q_t) + \mathcal{W}_1(P_t,P) + \mathcal{W}_1(P_t,P). \tag{34}$$

By Proposition 1.1, $\mathcal{W}_1(P_t,Q_t) \le \mathbb{E}[\widehat{\mathcal{W}}_1^{(UB)}(P_t,Q_t)]$. Under assumptions 1.6, 1.7 and 1.8, by (Biswas et al., 2019, Theorem 2.5) we obtain

$$\mathcal{W}_1(P_t,P) \le \mathbb{E}\Big[\sum_{j=1}^{\lceil(\tau_Q-L-t)/L\rceil} c(\tilde{Y}_{t+(j-1)L}, Y_{t+jL})\Big] \text{ and} \tag{35}$$

$$\mathcal{W}_1(Q_t,Q) \le \mathbb{E}\Big[\sum_{j=1}^{\lceil(\tau_P-L-t)/L\rceil} c(\tilde{X}_{t+(j-1)L}, X_{t+jL})\Big]. \tag{36}$$

Equation (6) now directly follows. ■

4.2. Wasserstein upper bound proofs

Proof [Proof of Theorem 1.11] Under the coupled kernel $\bar{K}$ from Algorithm 1, 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

$$\begin{aligned}
\mathbb{E}[c(X_t, Y_t)^p]^{1/p} &= \mathbb{E}[\mathbb{E}[c(X_t, Y_t)^p | X_{t-1}, Y_{t-1}]]^{1/p} && (37)\\
&\leq \mathbb{E}[\mathbb{E}[\big(c(X_t, Z_t) + c(Z_t, Y_t)\big)^p | X_{t-1}, Y_{t-1}]]^{1/p} && (38)\\
&\leq \mathbb{E}[\mathbb{E}[c(X_t, Z_t)^p | X_{t-1}, Y_{t-1}]]^{1/p} + \mathbb{E}[\mathbb{E}[c(Z_t, Y_t)^p | X_{t-1}, Y_{t-1}]]^{1/p} && (39)\\
&\leq \rho \mathbb{E}[c(X_{t-1}, Y_{t-1})^p]^{1/p} + \mathbb{E}[\Delta_p(Y_{t-1})]^{1/p} && (40)
\end{aligned}$$

where (38) follows as c is a metric, (39) follows by Minowski's inequality, and (40) follows by Assumption 1.10 with $\Delta_p(z) := \mathbb{E}[c(X, Y)^p | z]$ for $(X, Y) \sim \Gamma_\Delta(z)$. By induction, (40) implies

$$\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} \mathbb{E}[\Delta_p(Y_{i-1})]^{1/p}. \tag{41}$$

■

Proof [Proof of Corollary 1.13] Denote $a := \mathbb{E}[\Delta_p(Y^*)]^{1/p}$ for $Y^* \sim Q$ and $a_k := \mathbb{E}[\Delta_p(Y_k)]^{1/p}$ 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 4.1, this implies

$$\sum_{i=1}^{t} \rho^{t-i} a_{i-1} \overset{t\to\infty}{\to} \sum_{i=1}^{t} \rho^{t-i} a = \frac{1-\rho^t}{1-\rho} a. \tag{42}$$

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 1.11,

$$\begin{aligned}
\mathbb{E}[c(X_t, Y_t)^p]^{1/p} &\leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} a_{i-1} && (43)\\
&\leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} a + \sum_{i=1}^{t} \rho^{t-i} |a_{i-1} - a| && (44)\\
&= \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \frac{1-\rho^t}{1-\rho} a + \epsilon. && (45)
\end{aligned}$$

■

Proof [Proof of Proposition 1.14] As V is a a p^{th}-order Lyapunov function of K_2, by induction

$$\mathbb{E}[V(Y_i)^p] \leq \gamma^i \mathbb{E}[V(Y_0)^p] + (1-\gamma^i) \frac{L}{1-\gamma} \text{ for all } i \geq 0. \tag{46}$$

for all $i \geq 0$. Therefore,

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

for all $i \geq 1$, where the first inequality follows from the definition of δ, second inequality from (46), and the second inequality from the definition of κ. By Theorem 1.11, we obtain

$$\mathbb{E}[c(X_t, Y_t)^p]^{1/p} \leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \rho^{t-i} \mathbb{E}\Big[\Delta_p(Y_{i-1})\Big]^{1/p} \tag{47}$$

$$\leq \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \delta\kappa \sum_{i=1}^{t} \rho^{t-i} \tag{48}$$

$$= \rho^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + (1-\rho^t)\frac{\delta\kappa}{1-\rho}. \tag{49}$$

■

4.3. Wasserstein distances of empirical distributions proofs

To prove Proposition 1.16, we first record a technical result.

Lemma 4.3 *Suppose S and T are distributions on the metric space $(\mathcal{X}, c)$ with finite moments of order p. Given $U_i \sim S$ for $i = 1, \ldots, n$, let $\hat{S}_N$ denote the empirical distribution of $(U_1, \ldots, U_N)$. Then,*

$$\mathcal{W}_p(S, T)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{S}_N, T)^p]. \tag{50}$$

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,

$$\mathcal{W}_p(S, T)^p \leq \mathbb{E}[c(U, V)^p] \text{ by the coupling representation of Wasserstein distance} \tag{51}$$

$$= \mathbb{E}[\mathbb{E}[c(U, V)^p | U_1, \ldots U_N]] \tag{52}$$

$$= \mathbb{E}[\mathcal{W}_p(\hat{S}_N, T)^p]. \tag{53}$$

■

Proof [Proof of Proposition 1.16]

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 4.3 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 $(X_1, \ldots, X_N)$ and $(Y_1, \ldots, Y_N)$ are independent, $Y_i|(X_1, \ldots, X_N) \sim Y_i \sim Q$ for all $i = 1, \ldots, N$. We can therefore apply Lemma 4.3 conditional on $(X_1, \ldots, X_N)$ now with $S = Q$, $U_i = Y_i$ and $T = \hat{P}_N$ to obtain

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

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

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

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

$$\mathcal{W}_p(\hat{P}_N, \hat{Q}_N) \leq \mathcal{W}_p(\hat{P}_N, P) + \mathcal{W}_p(P, Q) + \mathcal{W}_p(Q, \hat{Q}_N). \tag{54}$$

By Minowski's inequality, this gives

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_p(\hat{P}_N, \hat{Q}_N)^p]^{1/p} &\leq \mathbb{E}\Big[\Big(\mathcal{W}_p(\hat{P}_N, P) + \mathcal{W}_p(P, Q) + \mathcal{W}_p(Q, \hat{Q}_N)\Big)^p\Big]^{1/p} && (55)\\
&\leq \mathbb{E}[\mathcal{W}_p(\hat{P}_N, P)^p]^{1/p} + \mathbb{E}[\mathcal{W}_p(P, Q)^p]^{1/p} + \mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_N)^p]^{1/p} && (56)\\
&= \mathbb{E}[\mathcal{W}_p(\hat{P}_N, P)^p]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_N)^p]^{1/p} && (57)
\end{aligned}$$

Let $\tilde{P}_N$ denote empirical distributions of the samples $(\tilde{X}_1, \ldots, \tilde{X}_N)$, where $\tilde{X}_i \sim P$ for all $i = 1, \ldots, n$ and $(\tilde{X}_1, \ldots, \tilde{X}_N)$ and $(X_1, \ldots, X_N)$ are independent. Independence implies $\tilde{X}_i|(X_1, \ldots, X_N) \sim \tilde{X}_i \sim P$ for all $i = 1, \ldots, n$. We can therefore apply Lemma 4.3 conditional on $(X_1, \ldots, X_N)$, with $S = P$, $T = \hat{P}_N$ and $\tilde{X}_i = U_i$ to obtain

$$\mathcal{W}_p(\hat{P}_N, P)^p \leq \mathbb{E}[\mathcal{W}_p(\hat{P}_N, \tilde{P}_N)^p | X_1, \ldots, X_N]. \tag{58}$$

Similarly,

$$\mathcal{W}_p(Q, \hat{Q}_N)^p \leq \mathbb{E}[\mathcal{W}_p(\tilde{Q}_N, \hat{Q}_N)^p | Y_1, \ldots, Y_N] \tag{59}$$

where $\tilde{Q}_N$ denotes empirical distributions of the samples $(\tilde{Y}_1, \ldots, \tilde{Y}_N)$, where $\tilde{Y}_i \sim Q$ for all $i = 1, \ldots, n$ and $(\tilde{Y}_1, \ldots, \tilde{Y}_N)$ and $(Y_1, \ldots, Y_N)$ are independent. By (57), we now obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_p(\hat{P}_N, \hat{Q}_N)^p]^{1/p} &\leq \mathbb{E}[\mathcal{W}_p(\hat{P}_N, P)^p]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_N)^p]^{1/p} && (60)\\
&= \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\hat{P}_N, P)^p | X_1, \ldots, X_N]]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathbb{E}[\mathcal{W}_p(Q, \hat{Q}_N)^p | Y_1, \ldots, Y_N]]^{1/p} && (61)\\
&\leq \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\hat{P}_N, \tilde{P}_N)^p | X_1, \ldots, X_N]]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathbb{E}[\mathcal{W}_p(\tilde{Q}_N, \hat{Q}_N)^p | Y_1, \ldots, Y_N]]^{1/p} && (62)\\
&= \mathbb{E}[\mathcal{W}_p(\hat{P}_N, P_N)^p]^{1/p} + \mathcal{W}_p(P, Q) + \mathbb{E}[\mathcal{W}_p(\tilde{Q}_N, \hat{Q}_N)^p]^{1/p} && (63)
\end{aligned}$$

as required. ■

5. Example applications of theoretical results

In this section we consider the theoretical results of Section 1.3 applied to three simple examples, working with the metric $c(x, y) = \|x - y\|_2$ induced by the L_2 norm.

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 **??**, 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 1.10 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 1.13,

$$\mathcal{W}_2(P_\sigma, P) \leq \liminf_{t\to\infty} \mathbb{E}[\widehat{\mathcal{W}}_2^{(UB)}(P_t, Q_t)^2]^{1/2} \leq \frac{\mathbb{E}\big[\|Y - Y'\|^2\big(1 - \alpha_\sigma\big(Y, Y'\big)\big)\big]^{1/2}}{C\sigma^2/2}, \tag{64}$$

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 (64) 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 **??**, 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 1.10 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 1.13, this gives

$$\mathcal{W}_2(P_\sigma, Q_\sigma) \leq \liminf_{t\to\infty} \mathbb{E}[\widehat{\mathcal{W}}_2^{(UB)}(P_t, Q_t)^2]^{1/2} \leq \frac{\mathbb{E}\big[\|\nabla \log P(Y_\sigma) - \nabla \log Q(Y_\sigma)\|^2\big]^{1/2}}{C_P} \tag{65}$$

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 (65) approaches the square root of the Fisher divergence between distributions Q and P, given by $F(Q, P) := \mathbb{E}[\|\nabla \log P(Y) - \nabla \log Q(Y)\|^2]$ 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 (65) 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 $\delta^2 := \sup_{z\in\mathcal{X}} e_{SGLD}(z)/(1+V(z)^2) < \infty$, for some 2^{nd}-order Lyapunov function V as in Proposition 1.14 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 **??**, 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 1.10 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 1.14,

$$\limsup_{t\to\infty} \mathcal{W}_2\big(P_t^{(ULA)}, P_t^{(SGLD)}\big) \leq \liminf_{t\to\infty} \mathbb{E}\Big[\widehat{\mathcal{W}}_2^{(UB)}\big(P_t^{(ULA)}, Q_t^{(ULA)}\big)^2\Big]^{1/2} \leq \frac{\delta\kappa}{C}. \tag{66}$$

Note that the upper bound in (66) 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 marginal chains (SGLD) may not converge to a limiting distribution.

6. Multi-step couplings

We can consider coupling algorithms for multi-step kernels and investigate their theoretical properties.

6.1. 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 (**??**) of main text and Section 1.2, 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 $A \in \mathcal{B}(\mathcal{X})$,

$$\bar{K}_{L-step}\big((x,y),(A,\mathcal{X})\big) = K_P^L(x,A) \text{ and } \bar{K}_{L-step}\big((x,y),(\mathcal{X},A)\big) = K_Q^L(y,A). \tag{67}$$

Given coupled kernels Γ_1 and Γ_Δ, Figure 5 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 (67) is satisfied. Algorithm 2 samples from this coupled kernel $\bar{K}_{L-step}$. It characterizes the dependency between X_{Lt} and

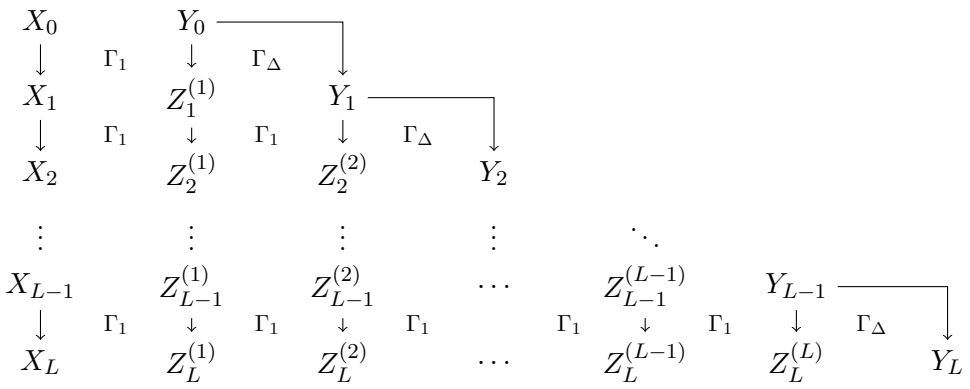


Figure 5: Joint kernel $\bar{K}_{L-step}$ on $\mathcal{X} \times \mathcal{X}$, which couples marginal kernels K_P^L and K_Q^L

Y_{Lt} such that

$$X_{Lt}|X_{L(t-1)}, Y_{L(t-1)} \sim K_P^L(X_{L(t-1)}, \cdot) \tag{68}$$

$$Z_L^{(j)}|Y_{L(t-1)+(j-1)} \sim K_P^{L-(j-1)}(Y_{L(t-1)+(j-1)}, \cdot) \tag{69}$$

$$Y_{Lt}|X_{L(t-1)}, Y_{L(t-1)} \sim K_Q^L(Y_{L(t-1)}, \cdot) \tag{70}$$

for $s = 1, \ldots, L-1$. When $L = 1$, we obtain $\bar{K}_{L-step} = \bar{K}$ from Algorithm 1. Note that $\bar{K}_{1-step}$ is the single-step kernel $\bar{K}$ from Algorithm 1, but $\bar{K}_{L-step}$ and $\bar{K}^L$ are not equivalent in general.

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

Algorithm 2: Joint kernel $\bar{K}_{L-step}$ on $\mathcal{X} \times \mathcal{X}$, which couples 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

$$(X_s, Z_s^{(1)}, \ldots, Z_s^{(s)}, Y_s)|(X_{s-1}, Z_{s-1}^{(1)}, \ldots, Z_{s-1}^{(s-1)}, Y_{s-1}) \tag{71}$$

jointly such that

$$(X_s, Z_s^{(1)}) \sim \Gamma_1(X_{s-1}, Z_{s-1}^{(1)}) \tag{72}$$

$$(Z_s^{(j)}, Z_s^{(j+1)}) \sim \Gamma_1(Z_{s-1}^{(j)}, Z_{s-1}^{(j+1)}) \text{ for } j = 1, \ldots, s-1 \tag{73}$$

$$(Z_s^{(s)}, Y_s) \sim \Gamma_\Delta(Y_{s-1}) \tag{74}$$

end

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

6.2. Theoretical properties of coupling 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 2 satisfies a geometric ergodicity condition.

Assumption 6.1 *There exists constants $C \in [1, \infty)$ and $\rho \in (0, 1)$ such that for all $L \geq 1$,*

$$\mathbb{E}[c(X_{t+L}, Y_{t+L})^p | X_t, Y_t]^{1/p} \leq C\rho^L c(X_t, Y_t) \text{ for } (X_{t+L}, Y_{t+L}) | (X_t, Y_t) \sim \Gamma_P^L(X_t, Y_t). \quad (75)$$

Assumption 6.1 is weaker than uniform contraction in Wasserstein's distance as in Assumption 1.10. Under Assumption 6.1, 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 5 and Algorithm 2. 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 6.2 *Let $(X_t, Y_t)_{t \geq 0}$ denote a coupled Markov chain generated using Algorithm ?? with initial distribution $\bar{I}_0$ and joint kernel $\bar{K}$ on $\mathcal{X} \times \mathcal{X}$ from Algorithm 1. Suppose the coupled kernel Γ_1 satisfies Assumption 6.1 for some $C = 1$ 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 2 with the L-step coupled kernel $\bar{K}_{L-step}$. Then for all $t \geq 0$,*

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} \mathbb{E}\Big[\Delta_p(Y_{L(i-1)+j})\Big]^{1/p} \Big) \quad (76)$$

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 6.3 *Under the setup and assumptions of Theorem 6.2, 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 $t \geq S$,*

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq (C\rho^L)^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + C\Big(\frac{1 - (C\rho^L)^t}{1 - C\rho^L}\Big)\Big(\frac{1 - \rho^L}{1 - \rho}\Big)\mathbb{E}[\Delta_p(Y^*)]^{1/p} + \epsilon. \quad (77)$$

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

As in Section 1.3, 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 6.4 *Under the setup and assumptions of Theorem 6.2, let $V : \mathcal{X} \to [0, \infty)$ be a p^{th}-order Lyapunov function of K_2 such that*

$$\mathbb{E}[V(Y_{t+1})^p | Y_t = z] \leq \gamma V(z)^p + L \quad (78)$$

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

$$\delta := \sup_{z\in\mathcal{X}} \left(\frac{\Delta_p(z)}{1+V(z)^p}\right)^{1/p} \qquad \kappa := 1 + \max\left\{\mathbb{E}[V(Y_0)^p]^{1/p}, \left(\frac{L}{1-\gamma}\right)^{1/p}\right\}. \tag{79}$$

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

$$\mathbb{E}[\widehat{\mathcal{W}}_p^{(UB)}(P_t,Q_t)^p]^{1/p} = \mathbb{E}[c(X_t,Y_t)^p]^{1/p} \leq (C\rho^L)^t \mathbb{E}[c(X_0,Y_0)^p]^{1/p} + C\Big(\frac{1-(C\rho^L)^t}{1-C\rho^L}\Big)\Big(\frac{1-\rho^L}{1-\rho}\Big)\delta\kappa. \tag{80}$$

6.3. Proofs

Proof [Proof of Theorem 6.2] Under the coupled kernel $\bar{K}_{L-step}$ from Algorithm 1, for each $t \geq 1$ we obtain

$$(X_{Lt}, Z_L^{(1)}, \ldots, Z_L^{(L)}, Y_{Lt}) \tag{81}$$

where

$$(X_{Lt}, Z_L^{(1)})|X_{L(t-1)}, Y_{L(t-1)} \sim \Gamma_P^L(X_{L(t-1)}, Y_{L(t-1)}) \tag{82}$$

$$(Z_L^{(j)}, Z_L^{(j+1)})|Y_{L(t-1)+j-1} \sim \Gamma_\Delta(Y_{L(t-1)+j-1})\Gamma_1^{L-j} \text{ for } j = 1,\ldots,L-1 \tag{83}$$

$$(Z_L^{(L)}, Y_{Lt})|Y_{L(t-1)+L-1} \sim \Gamma_\Delta(Y_{L(t-1)+L-1}). \tag{84}$$

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

$$\mathbb{E}[c(X_{Lt},Y_{Lt})^p]^{1/p} = \mathbb{E}[\mathbb{E}[c(X_{Lt},Y_{Lt})^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} \tag{85}$$

$$\leq \mathbb{E}[\mathbb{E}[\big(c(X_{Lt},Z_L^{(1)}) + c(Z_L^{(1)},Y_{Lt})\big)^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} \tag{86}$$

$$\leq \mathbb{E}[\mathbb{E}[c(X_{Lt},Z_L^{(1)})^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} + \mathbb{E}[\mathbb{E}[c(Z_L^{(1)},Y_{Lt})^p|X_{L(t-1)},Y_{L(t-1)}]]^{1/p} \tag{87}$$

$$\leq \tilde{\rho}\mathbb{E}[c(X_{L(t-1)},Y_{L(t-1)})^p]^{1/p} + \mathbb{E}[c(Z_L^{(1)},Y_{Lt})^p]^{1/p} \tag{88}$$

where (86) follows as c is a metric, (87) follows by Minowski's inequality, and (88) follows by Assumption 6.1. Denote $\Delta_p(z) := \mathbb{E}[c(X,Y)^p|z]$ for $(X,Y) \sim \Gamma_\Delta(z)$. Then,

$$\mathbb{E}[c(Z_L^{(1)}, Y_{Lt})^p]^{1/p} \leq \mathbb{E}\Big[\Big((Z_L^{(L)}, Y_{Lt}) + \sum_{j=1}^{L-1} c(Z_L^{(j)}, Z_L^{(j+1)})\Big)^p\Big]^{1/p} \tag{89}$$

$$\leq \mathbb{E}\Big[(Z_L^{(L)}, Y_{Lt})^p\Big]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\Big[c(Z_L^{(j)}, Z_L^{(j+1)})^p\Big]^{1/p} \tag{90}$$

$$= \mathbb{E}\Big[\mathbb{E}\Big[(Z_L^{(L)}, Y_{Lt})^p | Y_{L(t-1)+L-1}\Big]\Big]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\Big[\mathbb{E}\Big[c(Z_L^{(j)}, Z_L^{(j+1)})^p | Y_{L(t-1)+j-1}\Big]\Big]^{1/p} \tag{91}$$

$$= \mathbb{E}[\Delta_p(Y_{L(t-1)+(L-1)})]^{1/p} + \sum_{j=1}^{L-1} \mathbb{E}\Big[\mathbb{E}\Big[c(Z_L^{(j)}, Z_L^{(j+1)})^p | Y_{L(t-1)+j-1}\Big]\Big]^{1/p} \tag{92}$$

$$\leq \mathbb{E}[\Delta_p(Y_{L(t-1)+(L-1)})]^{1/p} + \sum_{j=1}^{L-1} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(t-1)+j-1})\Big]^{1/p} \tag{93}$$

$$\leq \sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(t-1)+j})\Big]^{1/p} \tag{94}$$

(88) now gives

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \tilde{\rho}\mathbb{E}[c(X_{L(t-1)}, Y_{L(t-1)})^p]^{1/p} + \sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(t-1)+j})\Big]^{1/p} \tag{95}$$

By induction, (95) implies

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \tilde{\rho}^t\mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i}\Big(\sum_{j=1}^{L} C\rho^{L-j}\mathbb{E}\Big[\Delta_p(Y_{L(i-1)+j})\Big]^{1/p}\Big) \tag{96}$$

as required. ∎

Proof [Proof of Corollary 6.3] Denote $a := \mathbb{E}[\Delta_p(Y^*)]^{1/p}$ for $Y^* \sim Q$ and $a_k := \mathbb{E}[\Delta_p(Y_k)]^{1/p}$ 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

$$\sum_{i=1}^{t} \tilde{\rho}^{t-i}\Big(\sum_{j=1}^{L} C\rho^{L-j} a_{L(i-1)+j}\Big) \overset{t\to\infty}{\to} \sum_{i=1}^{t} \tilde{\rho}^{t-i}\Big(\sum_{j=1}^{L} C\rho^{L-j} a\Big). \tag{97}$$

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 6.2,

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} a_{L(i-1)+j} \Big) \tag{98}$$

$$\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \sum_{j=1}^{L} C\rho^{L-j} a + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} |a_{L(i-1)+j} - a| \Big) \tag{99}$$

$$\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \sum_{j=1}^{L} C\rho^{L-j} a + \epsilon \tag{100}$$

$$= (C\rho^L)^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + C\Big(\frac{1 - (C\rho^L)^t}{1 - C\rho^L} \Big) \Big(\frac{1 - \rho^L}{1 - \rho} \Big) a + \epsilon \tag{101}$$

as required. ■

Proof [Proof of Proposition 6.4] As V is a a p^{th}-order Lyapunov function of K_2, by induction

$$\mathbb{E}[V(Y_j)^p] \leq \gamma^j \mathbb{E}[V(Y_0)^p] + (1 - \gamma^t) \frac{L}{1 - \gamma} \tag{102}$$

for all $j \geq 0$. This gives

$$\mathbb{E}\Big[\Delta_p(Y_j)\Big]^{1/p} \leq \delta \mathbb{E}[1 + V(Y_{j-1})^p]^{1/p} \tag{103}$$

$$\leq \delta(1 + \mathbb{E}[V(Y_{j-1})^p]^{1/p}) \tag{104}$$

$$\leq \delta\left(1 + \left(\gamma^{t-1} \mathbb{E}[V(Y_0)^p] + (1 - \gamma^{t-1}) \frac{L}{1 - \gamma} \right)^{1/p} \right) \tag{105}$$

$$\leq \delta\left(1 + \max\left\{ \mathbb{E}[V(Y_0)^p]^{1/p}, \left(\frac{L}{1 - \gamma} \right)^{1/p} \right\} \right) \tag{106}$$

$$= \delta\kappa \tag{107}$$

for all $j \geq 0$. By Theorem 6.2, we obtain

$$\mathbb{E}[c(X_{Lt}, Y_{Lt})^p]^{1/p} \leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} \mathbb{E}\Big[\Delta_p(Y_{L(i-1)+j})\Big]^{1/p} \Big) \tag{108}$$

$$\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + \sum_{i=1}^{t} \tilde{\rho}^{t-i} \Big(\sum_{j=1}^{L} C\rho^{L-j} \delta\kappa \Big) \tag{109}$$

$$\leq \tilde{\rho}^t \mathbb{E}[c(X_0, Y_0)^p]^{1/p} + C\Big(\frac{1 - (C\rho^L)^t}{1 - C\rho^L} \Big) \Big(\frac{1 - \rho^L}{1 - \rho} \Big) \delta\kappa \tag{110}$$

■

7. Additional Algorithms

Algorithm 3: Common random numbers coupling of two MALA kernels marginally targetting distributions P and Q

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

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

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 4: Common random numbers coupling of a MALA kernel and an ULA kernel 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

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

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})