Methodology and theory for unbiased Markov chain Monte Carlo and alternatives

Tamás Péter Papp, MA (Cantab), MMath, MRes

Submitted for the degree of Doctor of Philosophy at

Lancaster University.

September 2025

STOR-i

excellence with impact

Abstract

Markov chain Monte Carlo (MCMC) is the default inference method in Bayesian statistics. However, MCMC algorithms are biased, which can limit the extent to which these algorithms can be parallelized, and which can complicate the subsequent statistical inference. Recently-proposed methods provide ways of entirely eliminating this bias, based on simulating a coupling of two Markov chains that evolve in tandem. This thesis contributes to the unbiased MCMC literature by improving the efficiency and the theoretical understanding of coupling methods. Additionally, it provides a competitive alternative to coupling methods.

The first contribution is an effective coupling for the random walk Metropolis algorithm, a widely-used practical MCMC algorithm. We design this coupling with the explicit aim of scalability to high dimensions, and analyze this coupling in a theoretical framework that can quantify the efficiency of couplings in high dimensions. This framework may be useful for designing and analyzing couplings of other MCMC algorithms.

The second contribution is a study on tuning the scalar parameters of coupling methods. We argue that the so-called time-lag parameter is crucial to the efficiency and the robustness of these methods. Even though unbiased MCMC estimators can be noisier than standard MCMC ones, we demonstrate how judicious tuning can ensure that unbiased MCMC is nearly as efficient as standard MCMC.

The final contribution is a pair of estimators of the squared Euclidean 2-Wasserstein distance, a strong measure of the discrepancy between two distributions. These es-

timators are based on approximately independent samples from the distributions of interest. We show that the estimators are often upper and lower bounds on the underlying discrepancy, and we demonstrate that they often outperform coupling methods in statistical applications.

Acknowledgements

Firstly, I would like to thank my supervisor, Chris Sherlock, for selflessly lending me his time, wisdom, and guidance throughout these years. It has been my honour and privilege to have worked with you; I am undoubtedly a better researcher for it.

I would like to thank everyone in the leadership and administrative teams of the STOR-i Centre for Doctoral Training for their tireless help and support throughout the years.

I would like to thank Paul Fearnhead for his guidance and for the opportunity to be a part of the CoSInES grant while intercalating from the PhD. The later stage of my PhD undoubtedly benefitted from numerous interactions with the community around CoSInES and its sister-grant Bayes4Health; I thank you all for the keeping me intellectually engaged.

I am grateful to my viva examiners, Pierre Jacob, Peter Neal, and Chris Nemeth, for their valuable comments, insights, and suggestions.

My PhD would have been a significantly less enjoyable experience without the welcoming environment of STOR-i. I am grateful to my cohort for fostering a collegial atmosphere, and in particular to Matt Darlington, Peter Greenstreet, Ed Mellor, Matt Randall, and Hamish Thorburn for the good times. Matt Randall deserves special thanks for housing me, for propping up my social life, and for being a reliable friend over all these years; I will forever be grateful. To all my other friends from Lancaster: I thank you for making these years that much more fun. (If perhaps less productive!)

Finally, I would have never gotten this far without the unwavering support of my family, and this PhD could not have been concluded without the invaluable encouragement of Holly Jackson. I cannot thank you all enough.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Tamás Péter Papp

Contents

List of Figures

List of Abbreviations

CDF	cumulative distribution function
CRN	common random numbers
EJC	expected jump concordance
ESJD	expected squared jump distance
ESS	effective sample size
GCRefl	gradient-corrected reflection
GCRN	gradient common random numbers
H&H	Hug and Hop
HMC	Hamiltonian Monte Carlo
IACT	integrated autocorrelation time
MALA	Metropolis-adjusted Langevin algorithm
MCMC	Markov chain Monte Carlo
ODE	ordinary differential equation
OT	optimal transport
OU	Ornstein-Uhlenbeck
PDF	probability distribution function
RWM	Random walk Metropolis
SVM	stochastic volatility model

List of Symbols

$[n]$	Set $\{1, \ldots, n\}$.
$\lfloor \cdot \rfloor$	Floor.
$\lceil \cdot \rceil$	Ceiling.
$\wedge$	Minimum.
$\vee$	Maximum.
x_+	Positive part, i.e. $\max(x, 0)$.
$\|\cdot\|$	Euclidean norm.
$\|\cdot\|_\infty$	Supremum norm.
∇	Gradient.
∇^2	Hessian.
$\mathcal{O}, \Theta, \omega, \Omega$	Bachmann-Landau asymptotic notation.
I_d	Identity matrix of dimension $d \times d$.
$\Sigma^{1/2}$	Principal square-root of symmetric positive definite matrix Σ.
$\succeq$	Loewner ordering of matrices.
$A \otimes B$	Kronecker product of matrices A and B.
$\mathrm{diag}(\cdot)$	Diagonal matrix specified according to its diagonal entries.
$\delta_x(\cdot)$	Dirac delta function.
$\mu \otimes \nu$	Product of distributions μ and ν.
$\mathbb{P}(\cdot)$	Probability.
$\mathbb{E}[\cdot]$	Expectation.

$\mathcal{P}(\cdot)$	Set of Borel probability measures.
$L^p(\mu)$	Set $\{f : \mathbb{E}_\mu[\lvert f(X)\rvert^p] < \infty\}$ of p times integrable functions under the distribution μ.
$\mathcal{N}_d(\mu, \Sigma)$	Gaussian distribution of dimension d with mean μ and covariance Σ.
$\mathcal{N}_d(x \mid \mu, \Sigma)$	Density of $\mathcal{N}_d(\mu, \Sigma)$ evaluated at x.
$\mathcal{N}(\cdot)$	One-dimensional Gaussian distribution or density.
$\phi(\cdot)$	Density function of $\mathcal{N}(0, 1)$.
$\Phi(\cdot)$	Cumulative distribution function of $\mathcal{N}(0, 1)$.
$\mathrm{BvN}(x, y \mid \rho)$	Cumulative probability $\mathcal{P}(Z_1 \leq x, Z_2 \leq y)$ under jointly Gaussian $Z_1, Z_2 \sim \mathcal{N}(0, 1)$ with correlation ρ.
$\Longrightarrow$	Weak convergence.
$\xrightarrow{\mathrm{p}}$	Convergence in probability.
$\xrightarrow{\mathrm{TV}}$	Convergence in total variation.
$\mathcal{L}(\cdot)$	Law of a random variable.

Chapter 1

Introduction

Monte Carlo algorithms (e.g. Robert and Casella, 2004) provide a way of understanding probability distributions by sampling from them, with wide-ranging uses throughout statistics, machine learning, and the applications of these disciplines. Markov chain Monte Carlo algorithms (e.g. Brooks et al., 2011) are a class of Monte Carlo algorithms that provide a flexible iterative way of sampling from probability distributions. They simulate a sequence of random samples one step at a time, drawing the next sample based on a distribution that is entirely determined by the current sample; thus, the sequence of draws forms a Markov chain (e.g. Meyn and Tweedie, 2009). In general, Markov chain Monte Carlo samples only converge towards the target distribution of interest as the number of iterations grows to infinity; at any finite number of iterations, samples from Markov chain Monte Carlo algorithms thus incur a certain bias. To account for this bias, practitioners often discard the initial portion of the Markov chain Monte Carlo run, referred to as the *burn-in*. Appropriately setting and quantifying this burn-in is, however, a notoriously challenging task.

Recent advances in Markov chain Monte Carlo provide ways of robustly assessing (Biswas et al., 2019) and even entirely eliminating (Jacob et al., 2020b) the burn-in bias using *couplings* of Markov chain Monte Carlo algorithms. These methods rely on

simulating two Markov chains in tandem, such that *marginally* the chains are simulated according to the Markov chain Monte Carlo algorithm of interest, but that *jointly* the chains evolve in such a way that their states eventually coincide after sufficiently many iterations. Eliminating the burn-in bias comes with a variety of computational and statistical advantages, chief among them that it simplifies the aggregation of multiple Markov chain Monte Carlo runs obtained by independent parallel processors. Unbiased Markov chain Monte Carlo methods, therefore, offer practitioners an appealing, principled way of exploiting parallel computing resources, contrasting with the inherently sequential nature of standard Markov chain Monte Carlo.

This thesis considers several outstanding questions related to the coupling methods of Jacob et al. (2020b); Biswas et al. (2019). Firstly, given a Markov chain Monte algorithm, how should one design an effective coupling for it, and is there an appropriate theoretical framework in which to measure the performance of such a coupling? Secondly, given that these coupling methods require various scalar tuning parameters, how should one choose these tuning parameters in practice? Finally, expecting that there may be situations where devising an effective coupling is difficult, can we assess convergence in similarly strong metrics to the method of Biswas et al. (2019), but without relying on couplings? One chapter of this thesis is devoted to each of these three questions, and each of the three chapters can be read independently of the others.

1.1 Organization of the thesis

Chapter 2: Background material and literature review

This chapter provides background for the remainder of the thesis. Section 2.1 introduces theoretical and computational aspects of optimal transport and couplings of random variables. Section 2.2 introduces Markov chain Monte Carlo. Section 2.3 introduces the unbiased Markov chain Monte Carlo method of Jacob et al. (2020b) and the convergence

bound of Biswas et al. (2019), and surveys related literature.

Chapter 3: Scalable couplings for the random walk Metropolis algorithm

This chapter is a journal contribution with co-author Chris Sherlock, reproduced in full. The manuscript has been published in the "Journal of the Royal Statistical Society, Series B (Statistical Methodology)" as Papp and Sherlock (2025b). All occurrences of "this paper" should be replaced by "this chapter." The abstract of the manuscript is given below.

There has been a recent surge of interest in coupling methods for Markov chain Monte Carlo algorithms: they facilitate convergence quantification and unbiased estimation, while exploiting embarrassingly parallel computing capabilities. Motivated by these, we consider the design and analysis of couplings of the random walk Metropolis algorithm which scale well with the dimension of the target measure. Methodologically, we introduce a low-rank modification of the synchronous coupling that is provably optimally contractive in standard high-dimensional asymptotic regimes. We expose a shortcoming of the reflection coupling, the state of the art at the time of writing, and we propose a modification which mitigates the issue. Our analysis bridges the gap to the optimal scaling literature and builds a framework of asymptotic optimality which may be of independent interest. We illustrate the applicability of our proposed couplings, and the potential for extending our ideas, with various numerical experiments.

Chapter 4: On the efficiency of lagged coupling methods

This chapter considers the problem of tuning the scalar parameters of the coupling methods of Jacob et al. (2020b); Biswas et al. (2019) for optimal efficiency. Section 4.2 highlights a robustness issue shared by these methods. Section 4.3 derives asymptotic results related to the tuning of one of the scalar parameters. Section 4.4 provides a detailed study of parameter tuning in a stylized setting. Section 4.5 provides tuning

guidelines for practitioners.

Chapter 5: Centered plug-in estimation of Wasserstein distances

This chapter is a planned journal contribution with co-author Chris Sherlock, reproduced in full. The manuscript has been submitted for publication and, at the time of writing, is undergoing peer review. All occurrences of "this paper" should be replaced by "this chapter." The abstract of the manuscript is given below.

The plug-in estimator of the squared Euclidean 2-Wasserstein distance is conservative, however due to its large positive bias it is often uninformative. We eliminate most of this bias using a simple centering procedure based on linear combinations. We construct a pair of centered plug-in estimators that decrease with the true Wasserstein distance, and are therefore guaranteed to be informative, for any finite sample size. Crucially, we demonstrate that these estimators can often be viewed as complementary upper and lower bounds on the squared Wasserstein distance. Finally, we apply the estimators to Bayesian computation, developing methods for estimating (i) the bias of approximate inference methods and (ii) the convergence of MCMC algorithms.

Chapter 6: Conclusions

This chapter summarizes the contributions of this thesis, and provides directions for further work based on the union of these contributions.

Chapter 2

Background material and literature review

2.1 Optimal transport

Optimal transport (e.g. Villani, 2003, 2009) is the study of the optimal movement of mass from one distribution $\mu \in \mathcal{P}(\mathcal{X})$ to another distribution $\nu \in \mathcal{P}(\mathcal{Y})$. Moving a unit of mass incurs a cost, specified by the *cost function* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we seek to minimize on average. Mass is moved from μ to ν according to a *transport plan* $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, which satisfies

$$\pi(A, \mathcal{Y}) = \mu(A), \;\; \pi(\mathcal{X}, B) = \nu(B), \;\; \text{for all measurable } A \subseteq \mathcal{X}, B \subseteq \mathcal{Y}. \tag{2.1.1}$$

Intuitively, in the discrete case, the transport plan π distributes the mass $\mu(x)$ at location x according to the conditional distribution $p(y \mid x) = \pi(x, y)/\mu(x)$. Condition (2.1.1) states that π is a joint distribution with marginals μ and ν, also known as a *coupling* of (μ, ν) in the probabilistic literature (Lindvall, 1992). We let $\Gamma(\mu, \nu)$ denote the set of all such couplings/transport plans π.

Our primary object of interest is the *optimal transport cost*

$$\mathcal{T}_c(\mu,\nu) = \inf_{\pi\in\Gamma(\mu,\nu)} \int c(x,y)\mathrm{d}\pi(x,y) = \inf_{(X,Y)\sim\pi\in\Gamma(\mu,\nu)} \mathbb{E}_\pi\left[c(X,Y)\right].$$

Another object of interest is the optimal coupling π^* that attains the infimum above.

As it is a linear optimization problem on a convex domain, the (primal) optimal transport problem has an equivalent dual formulation (Kantorovich, 1942)

$$\mathcal{T}_c(\mu,\nu) = \sup_{(\phi,\psi)\in\Phi(\mu,\nu)} \int \phi(x)\mathrm{d}\mu(x) + \int \psi(y)\mathrm{d}\nu(y),$$

$$\Phi(\mu,\nu) = \{(\phi,\psi)\in L^1(\mu)\times L^1(\nu) : \phi(x)+\psi(y)\le c(x,y),\ \ \forall (x,y)\in\mathcal{X}\times\mathcal{Y}\}.$$

Intuitively, the dual problem takes the perspective of a shipping company that charges, per unit of goods, the amount $\phi(x)$ at pick-up in location x and the amount $\psi(y)$ at drop-off in location y. It seeks to maximize profit subject to the maximum cost $c(x,y)$ that the customer is willing to pay for a unit of goods to be moved from x to y.

We use $(\phi_{\mu,\nu},\psi_{\mu,\nu})$ to denote an optimal solution to the Kantorovich dual problem, referred to as *Kantorovich potentials*. Kantorovich potentials are not unique: for any constant $c\in\mathbb{R}$, the pair $(\phi_{\mu,\nu}-c,\psi_{\mu,\nu}+c)$ is also optimal, and in general there may be additional degrees of freedom. When $\mu=\nu$, the optimal transport cost is zero, i.e. $\mathcal{T}_c(\mu,\mu)=0$, and an optimal solution is $\phi_{\mu,\mu}=\psi_{\mu,\mu}=0$. As we will see in Chapter 5, the Kantorovich dual formulation is particularly useful for computational purposes.

2.1.1 Monge problem

In certain cases, the optimal transport π^* does not split mass: for all locations x, all of the mass $\mathrm{d}\mu(x)$ is moved to a single location. This recovers the original formulation of the optimal transport problem of Monge (1781), where the optimal transport is given by a map $T_{\mu,\nu}$. Brenier's theorem (1991) establishes this property in the case of continuous measures and the squared-Euclidean cost $c(x,y)=\|x-y\|^2$, and provides the form of

the optimal transport map.

Theorem 2.1.1 (Brenier's theorem; e.g. Villani, 2003)**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and let μ be continuous with respect to the Lebesgue measure. Then,*

$$\mathcal{T}_{\|\cdot\|^2}(\mu, \nu) = \mathbb{E}_{X\sim\mu}[\|X - T_{\mu,\nu}(X)\|^2],$$

where $T_{\mu,\nu}$ is a deterministic transport map that pushes μ forward to ν, such that $T_{\mu,\nu}(X) \sim \nu$. Furthermore, $T_{\mu,\nu} = \nabla\varphi_{\mu,\nu}$ for some convex function $\varphi_{\mu,\nu} : \mathbb{R}^d \to \mathbb{R}$.

The relation $\varphi_{\mu,\nu}(x) = \|x\|^2/2 - \phi_{\mu,\nu}(x)$ holds between the Brenier potential $\varphi_{\mu,\nu}$ and the Kantorovich potential $\phi_{\mu,\nu}$ (Villani, 2003, Remark 2.13). Provided that ν is also continuous, the transport map $T_{\mu,\nu}$ is bijective, and the optimal transport map that pushes ν forward to μ is $T_{\nu,\mu} = T_{\mu,\nu}^{-1}$.

2.1.2 Wasserstein distance

Optimal transport can be used to define the *Wasserstein distances*, a family of distances on the space of probability measures. Let now $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a metric and let $p \geq 1$. The p-Wasserstein distance $\mathcal{W}_{p,c}(\mu, \nu)$ is defined, for $\mu, \nu \in \mathcal{P}(\mathcal{X})$, as

$$\mathcal{W}_{p,c}(\mu, \nu) = \mathcal{T}_{c^p}(\mu, \nu)^{1/p} = \inf_{(X,Y)\sim\pi\in\Gamma(\mu,\nu)} \mathbb{E}_\pi[c(X, Y)^p]^{1/p}.$$

The following properties can be found in Villani (2009, Chapter 6). The p-Wasserstein distance is a finite-valued metric on the set

$$\mathcal{P}_c^p(\mathcal{X}) = \left\{\mu \in \mathcal{P}(\mathcal{X}) : \int c(x, y)^p \mathrm{d}\mu(x) < \infty,\ \forall y \in \mathcal{X}\right\}$$

of Borel probability measures with finite p-th moment. It provides a strong control over the discrepancy between distributions, as it metrizes weak convergence and, by

the dual formulation, it controls the discrepancy between the p-th moment of Lipschitz functions. Furthermore, as the p-Wasserstein distance inherits the properties of the *ground metric* c, it induces a particularly intuitive geometry. The nice properties of Wasserstein distances make them useful tools both for analysis, as we will see in Section 2.2, as well as for methodology, as we will see in Chapter 5.

With the choice of metric $c(x, y) = \mathbb{1}\{x \neq y\}$, the 1-Wasserstein distance becomes the total variation distance

$$\mathrm{TV}(\mu, \nu) = \inf_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{P}(X \neq Y).$$

Furthermore, when (μ, ν) have densities (p_μ, p_ν) with respect to a common dominating measure λ, the total variation distance can be expressed as (Scheffé, 1947)

$$\mathrm{TV}(\mu, \nu) = \frac{1}{2}\int |p_\mu(x) - p_\nu(x)|\mathrm{d}\lambda(x) = 1 - \int p_\mu(x) \wedge p_\nu(x)\mathrm{d}\lambda(x).$$

2.1.3 Solving discrete optimal transport problems

Discrete optimal transport problems are computationally tractable. For $\mu = \sum_{i=1}^n \mu_i \delta_{x_i}$, $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, and $C_{ij} = c(x_i, y_j)$, the primal and dual optimal transport problems can be written as

$$\min_\pi \sum_{ij} C_{ij}\pi_{ij} \text{ subject to } \pi_{ij} \geq 0, \ \sum_j \pi_{ij} = \mu_i, \ \sum_i \pi_{ij} = \nu_j \text{ for all } i, j,$$

$$\max_{\phi,\psi} \sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j \text{ subject to } \phi_i + \psi_j \leq C_{ij} \text{ for all } i, j.$$

These are linear programs with a special structure, and can be solved exactly using network simplex algorithms (Dantzig, 1951). Considering the simplified case $m = n$, the best theoretical bound on the complexity of network simplex algorithms is $O(n^3 \log^2 n)$ (Orlin, 1997; Tarjan, 1997). In practice, the state-of-the art network simplex imple-

mentation of Bonneel et al. (2011) scales closer to $O(n^2)$. In Chapter 5, we consider problems of sizes n up to the low ten-thousands, which can be solved exactly in seconds.

Fast approximate solvers have also been developed. Cuturi (2013) proposes to regularize the primal objective by adding the term $\varepsilon \sum_{ij} \pi_{ij} \log(\pi_{ij})$. This results in the following "entropic" optimal transport problem:

$$\min_{\pi} \varepsilon \, \mathrm{KL}(\pi \mid \eta) \text{ subject to } \pi_{ij} \geq 0, \ \sum_j \pi_{ij} = \mu_i, \ \sum_i \pi_{ij} = \nu_j \text{ for all } i, j,$$

where η is the discrete distribution satisfying $\eta_{ij} \propto \exp(-C_{ij}/\varepsilon)$. The regularization introduces an $O(\varepsilon \log(1/\varepsilon))$ error (Chizat et al., 2020), but makes the objective convex, and so enables the use of convex optimization techniques. Cuturi (2013) proposes to use the method of alternating projections (Von Neumann, 1949), i.e. the Sinkhorn-Knopp matrix scaling algorithm (Deming and Stephan, 1940; Sinkhorn and Knopp, 1967). The Sinkhorn algorithm has complexity $O(n^2/\varepsilon^2)$ (Dvurechensky et al., 2018), but is amenable to vectorization and so can take advantage of graphics processing units (GPUs).

2.1.4 Sampling from optimal couplings

This thesis is concerned with methodologies that simulate from optimal couplings $(X, Y) \sim \pi^* \in \Gamma(\mu, \nu)$, or approximations thereof. We now describe a few settings where optimal couplings are tractable.

Maximal couplings

Couplings π^* that minimize for the total variation distance $\mathrm{TV}(\mu, \nu)$ equivalently maximize the probability that the draws from (μ, ν) are identical, i.e.

$$\pi^* = \arg\min_{\pi \in \Gamma(\mu,\nu)} \mathbb{P}(X \neq Y) = \arg\max_{\pi \in \Gamma(\mu,\nu)} \mathbb{P}(X = Y).$$

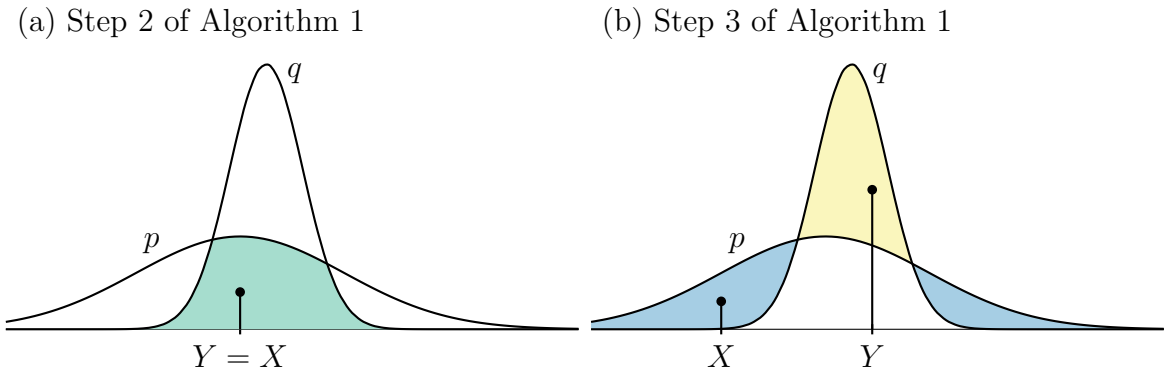

Figure 2.1.1: Illustration of a maximal coupling with independent residuals. We first draw a uniform sample under the graph of p and we retain its abscissa X. (a) If the sample falls in the area of overlap $p \wedge q$, we set $Y = X$. (b) If it does not, we sample uniformly from the residual area $q \setminus p$ (Algorithm 1 does so by rejection sampling), and we retain the abscissa Y. This provides a sample Y with the correct marginal density q.

Following Jacob et al. (2020b), we refer to such couplings as *maximal couplings*.

Algorithm 1 (Lindvall, 1992; Johnson, 1998) provides a way of simulating from a maximal coupling of (μ, ν) when these have densities (p, q) with respect to a common dominating measure. Conditional on the event $\{X \neq Y\}$, Algorithm 1 simulates the random variables (X, Y) independently. Algorithm 1 is based on the simple observation that sampling $(X, Up(X))$ uniformly under the graph of p then retaining the first coordinate provides a sample X with the correct marginal density p (Robert and Casella, 2004, Chapter 2.3.1). Figure 2.1.1 illustrates Algorithm 1.

Algorithm 1 Maximal coupling with independent residuals

1. Sample $X \sim p$ and $U \sim \text{Unif}(0, 1)$.
2. If $Up(X) \leq q(X)$, then set $Y = X$.
3. Else, sample $Y \sim q$ and $V \sim \text{Unif}(0, 1)$ until $Vq(Y) > p(Y)$.

In expectation, Algorithm 1 requires two draws from q to terminate, but the variance of its computing cost diverges as $\text{TV}(\mu, \nu) \to 0$. Gerber and Lee (2020) describes a simple modification of Algorithm 1 that does not have this issue, at the cost of being suboptimal.

(a) Step 2 of Algorithm 2 (b) Step 3 of Algorithm 2

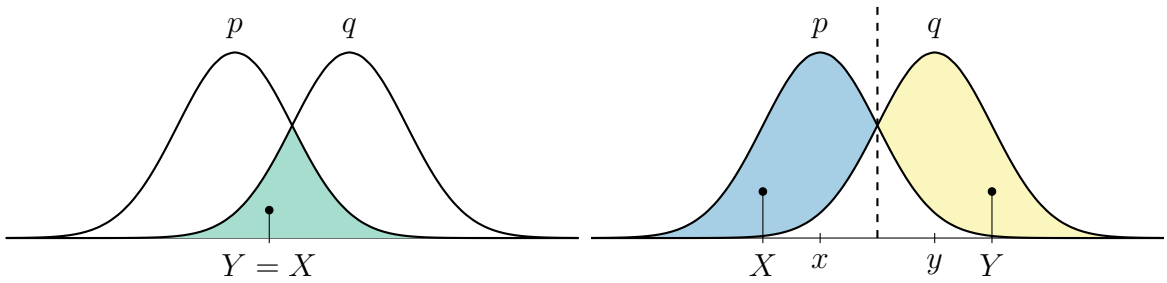

Figure 2.1.2: Illustration of a reflection-maximal coupling. We first draw a uniform sample (X, H) under the graph of p and we retain its abscissa X. (a) If the sample falls in the area of overlap $p \wedge q$, we set $Y = X$. (b) If it does not, we set Y as the reflection of X in the perpendicular bisector of (x, y); by symmetry, it holds that (Y, H) is a uniform sample from the residual area $q \setminus p$. Thus, Y has the correct marginal density q.

Algorithm 2 (Bou-Rabee et al., 2020; Jacob et al., 2020b) describes a maximal coupling (X, Y) of Gaussian measures $\mu = \mathcal{N}_d(x, \Sigma)$ and $\nu = \mathcal{N}_d(y, \Sigma)$. Conditional on the event $\{X \neq Y\}$, Algorithm 2 simulates the random variables (X, Y) based on a reflection in the perpendicular bisector of (x, y). Figure 2.1.2 provides an illustration of Algorithm 2.

Algorithm 2 Reflection-maximal coupling

Require: matrix L such that $LL^\top = \Sigma$, $z = L^{-1}(x-y)$, $e = z/\|z\|$, $s(x) = \mathcal{N}_d(x \mid 0_d, I_d)$.

1. Sample $\dot{X} \sim \mathcal{N}_d(0_d, I_d)$ and $U \sim \mathrm{Unif}(0, 1)$.
2. If $Us(\dot{X}) \leq s(\dot{X} + z)$, then set $\dot{Y} = \dot{X} + z$. ▷ Will output $Y = X$.
3. Else, set $\dot{Y} = \dot{X} - 2(e^\top \dot{X})e$.
4. Return $(X, Y) = (x + L\dot{X}, y + L\dot{Y})$.

As opposed to Algorithm 1, the computational cost of Algorithm 2 is bounded. The use of the reflection move turns out to be particularly advantageous when designing couplings of Markov kernels, as we will see in Chapter 3.

Common random numbers couplings

Couplings defined in terms of transformations of shared randomness are called *common random numbers* (CRN) or *synchronous* couplings. CRN couplings tend to be sensible strategies when we seek to bring the random variables $X \sim \mu$ and $Y \sim \nu$ as close together as possible *on average*.

For one-dimensional measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with inverse-CDFs (F_μ^{-1}, F_ν^{-1}), the coupling $(X, Y) = (F_\mu^{-1}(U), F_\nu^{-1}(U))$ with $U \sim \text{Unif}(0, 1)$ is optimal for any cost function $c(x, y) = h(x - y)$ with a convex $h : \mathbb{R} \to \mathbb{R}$ (e.g. Santambrogio, 2015, Theorem 2.9). This is because $c(x_1, y_1) + c(x_2, y_2) \leq c(x_1, y_2) + c(x_2, y_1)$ whenever $x_1 \leq x_2$ and $y_1 \leq y_2$, so the optimal transport must preserve the increasing ordering of $\mathbb{R}$.

For Gaussian $\mu = \mathcal{N}_d(m_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}_d(m_\nu, \Sigma_\nu)$ with covariance matrices (Σ_μ, Σ_ν) that commute, the CRN coupling $(X, Y) = (m_\mu + \Sigma_\mu^{1/2} Z, m_\nu + \Sigma_\nu^{1/2} Z)$ with $Z \sim \mathcal{N}_d(0_d, I_d)$ is optimal for the squared-Euclidean cost $c(x, y) = \|x - y\|^2$ (e.g. Peyré and Cuturi, 2019, Remark 2.31).

Discrete distributions

Sampling from a discrete optimal coupling π^* can be done using any method to sample from discrete distributions, such as the inverse-CDF method or the alias method (Walker, 1977).

2.2 Markov chain Monte Carlo

We consider the problem of estimating the expectation $\mathbb{E}_\pi[h(X)]$ of a test function $h : \mathcal{X} \to \mathbb{R}$ under a probability distribution $\pi \in \mathcal{P}(\mathcal{X})$. For the purposes of this exposition, we assume that all probability distributions have densities with respect to a common dominating measure λ, and we identify a distribution and its density using the same symbol.

Provided that i.i.d. draws $X_1, \ldots, X_n \sim \pi$ are available, the *Monte Carlo estimator*

$$h_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \approx \mathbb{E}_{X \sim \pi}[h(X)]$$

can be used. This has the appealing properties that it is unbiased and converges at rate $n^{-1/2}$. Furthermore, the central limit theorem

$$\sqrt{n}\,(h_n - \mathbb{E}_\pi[h(X)]) \implies \mathcal{N}(0, \mathrm{Var}_\pi(h(X))) \tag{2.2.1}$$

provides confidence intervals for $\mathbb{E}_\pi[h(X)]$ whenever $\mathrm{Var}_\pi(h(X)) < \infty$. However, i.i.d. sampling is often infeasible for probability distributions π of interest. The canonical example is the posterior distribution π in a Bayesian inference problem.

Example 2.2.1: Consider a Bayesian posterior distribution $\pi(\cdot \mid y)$ based on the prior p, data y and likelihood $L(y \mid \cdot)$. The posterior density can be evaluated up to a constant of proportionality as $\pi(x \mid y) \propto p(x)L(y \mid x)$, but i.i.d. sampling is often intractable.

Markov chain Monte Carlo (MCMC; e.g. Brooks et al., 2011) is a practical alternative to i.i.d. sampling. *Asymptotically exact* MCMC algorithms simulate a π-invariant Markov chain $(X_t)_{t \geq 0}$ with transition kernel P, generating correlated samples whose distribution $X_t \sim \pi_t$ converges towards the target π as time $t \to \infty$. Most asymptotically exact MCMC algorithms are *reversible*, with their kernel P satisfying the following *detailed balance condition* for all $(\mathrm{d}x, \mathrm{d}y)$:

$$\pi(\mathrm{d}x)P(x, \mathrm{d}y) = \pi(\mathrm{d}y)P(y, \mathrm{d}x).$$

Detailed balance implies that $\int_{x \in \mathcal{X}} \pi(\mathrm{d}x)P(x, \mathrm{d}y) = \int_{x \in \mathcal{X}} \pi(\mathrm{d}y)P(y, \mathrm{d}x) = \pi(\mathrm{d}y)$, hence π is a stationary distribution for the chain.

The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) pro-

vides a recipe for designing asymptotically exact MCMC algorithms. From the current state $X_t = x$, the Metropolis-Hastings algorithm generates a proposal $X' = y$ from a proposal kernel $q(x, \cdot)$ with density $q(x, y)$. Then, with probability

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, \tag{2.2.2}$$

it sets $X_{t+1} = X'$; otherwise, it sets $X_{t+1} = x$. The Metropolis-Hastings kernel is $P_{\mathrm{MH}}(x, \mathrm{d}y) = q(x, \mathrm{d}y)\alpha(x, y) + r(x)\delta_x(\mathrm{d}y)$, where

$$r(x) = 1 - \int q(x, \mathrm{d}y)\alpha(x, y),$$

and can be checked to satisfy the detailed balance condition with respect to π. Notably, the Metropolis-Hastings algorithm only requires evaluations of the unnormalized density of π, which makes it an appealing choice in Bayesian inference problems (Tierney, 1994).

Provided that it is aperiodic and irreducible (Roberts and Rosenthal, 2004, Theorem 4), a π-invariant Markov chain converges to π from an arbitrary initialization. Aperiodicity and irreducibility are mild technical conditions that are satisfied by all algorithms considered in this thesis. Under these conditions, the ergodic theorem (e.g. Robert and Casella, 2004, Theorem 6.63) guarantees the almost-sure convergence of long-run sample averages of $(h(X_t))_{t\geq 0}$ to $\mathbb{E}_\pi[h(X)]$.

2.2.1 Performance measures

Compared to i.i.d. sampling, MCMC algorithms produce correlated samples that only converge towards π in the long run. To use MCMC algorithms effectively, we must therefore understand how quickly the MCMC algorithm converges, as well as how the

correlation affects the performance of MCMC estimators

$$h_{k:m} = \frac{1}{m-k+1} \sum_{t=k}^{m} h(X_t),$$

where the first k states $(X_t)_{t=0}^{k-1}$ are discarded by the user as *burn-in* based on an assessment of the convergence of the MCMC algorithm. We also refer to $h_{k:m}$ as an *ergodic average*.

Long-run efficiency

The long-run efficiency of the MCMC estimator $h_{k:m}$ can be assessed using the *asymptotic variance*, defined as

$$\upsilon(P,h) = \lim_{m\to\infty} (m-k+1)\operatorname{Var}(h_{k:m}) = \operatorname{Var}_\pi(h(X_0)) + 2\sum_{t\geq 1} \operatorname{Cov}_\pi(h(X_0), h(X_t)),$$

where the subscript π indicates that $X_0 \sim \pi$ and so that expectations are taken under a stationary chain. The loss of efficiency compared to i.i.d. sampling is provided by the *integrated autocorrelation time* (IACT)

$$\operatorname{IACT}(P,h) = \frac{\upsilon(P,h)}{\operatorname{Var}_\pi(h(X_0))} = 1 + 2\sum_{t\geq 1} \operatorname{Corr}_\pi(h(X_0), h(X_t)),$$

MCMC algorithms achieving lower IACTs are said to *mix faster*. The IACT is used to define the *effective sample size* (ESS)

$$\operatorname{ESS}(h_{k:m}) = \frac{m-k+1}{\operatorname{IACT}(P,h)},$$

with an effective sample size of $\operatorname{ESS}(h_{k:m}) = m-k+1$ corresponding to i.i.d. sampling. See e.g. Geyer (2011) for further background on these quantities. Standard ways of estimating the asymptotic variance are based on spectral variance, batch means (e.g.

Flegal and Jones, 2010; Vats et al., 2019), or initial sequence (Geyer, 1992) estimators, which are consistent as the length of the simulation $m \to \infty$ but converge slower than the Monte Carlo rate $m^{-1/2}$.

MCMC output is sometimes *thinned* by only retaining the output of every r-th iteration. Thinning is useful to reduce the cost of storing the MCMC output in computer memory, but tends to produce less efficient estimators (Geyer, 1992).

Convergence

Theoretical results concerning MCMC convergence consider the discrepancy between the marginals $(\pi_t)_{t\geq 0}$ and the stationary distribution π, often measuring it in total variation distance. The chain is said to be *polynomially ergodic* if there exist $C_{\text{poly}}(\pi_0) > 0$ and $\kappa > 1$ such that $\text{TV}(\pi, \pi_t) \leq C_{\text{poly}}(\pi_0)t^{-\kappa}$ for all $t \geq 0$. Practical MCMC algorithms often satisfy a stronger form of convergence, *geometric ergodicity*, when there exist $C_{\text{geo}}(\pi_0) > 0$ and $\rho \in (0, 1)$ such that $\text{TV}(\pi, \pi_t) \leq C_{\text{geo}}(\pi_0)\rho^t$ for all $t \geq 0$.

Provided that the chain is at least polynomially ergodic, and that $h \in L^{2+\delta}(\pi)$ for some $\delta > 2/(\kappa - 1)$, the MCMC estimator $h_{k:m}$ satisfies the central limit theorem (Jones, 2004, Theorem 9)

$$\sqrt{m-k+1}\,(h_{k:m} - \mathbb{E}_\pi[h(X)]) \implies \mathcal{N}(0, v(P, h)). \tag{2.2.3}$$

Ergodicity results therefore provide some reassurance that MCMC estimators behave sensibly and converge at the usual $m^{-1/2}$ Monte Carlo rate.

However, theoretical bounds on the rate of convergence (e.g. Roberts and Rosenthal, 2004) are often too loose for practical purposes. Instead, practitioners typically assess the convergence of MCMC algorithms by analyzing the output of multiple i.i.d. replicate simulations. A popular summary statistic is the *potential scale reduction factor* $\widehat{R}$ (Gelman and Rubin, 1992; Gelman et al., 2013), a scale-free measure of convergence that compares the variance of the samples *across replicates* to the variance of the samples

within each replicate. Assuming that the initialization and all marginals $(\pi_t)_{t\geq 0}$ are appropriately overdispersed with respect to the target π, the statistic should satisfy $\widehat{R} \geq 1$, with equality for sufficiently long stationary chains. In practice, one therefore runs multiple chains from an overdispersed initialization and declares convergence once $\widehat{R}$ is close enough to 1; Vehtari et al. (2021) recommend $\widehat{R} < 1.01$. Though various extensions of $\widehat{R}$ have been proposed in the literature (e.g. Vehtari et al., 2021; Margossian et al., 2024), the basic idea remains the same.

It is apparent that there is a gap between theoretical assessments of convergence, which rely on bounding strong metrics such as the total variation distance, and practical assessments of convergence, which typically rely on tracking simple summary statistics. This thesis aims to bridge this gap. In Section 2.3, we discuss the method of Biswas et al. (2019) that directly estimates upper bounds on $\mathrm{TV}(\pi, \pi_t)$ using couplings of Markov chains, and in Chapter 5 we propose a method to estimate upper and lower bounds on $\mathcal{T}_{\|\cdot\|^2}(\pi, \pi_t)$.

2.2.2 MCMC algorithms

The random walk Metropolis (RWM) algorithm (e.g. Tierney, 1994) is one of the simplest practical Metropolis-Hastings algorithms, using proposals of the form $q(x, y) = r(x - y)$, where $r(x) = r(-x)$ for all x. Due to cancellation, the acceptance probability (2.2.2) for the RWM simplifies to

$$\alpha_{\mathrm{RWM}}(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}. \tag{2.2.4}$$

The choice of proposals is typically Gaussian, $q(x, y) = \mathcal{N}_d(y \mid x, \Sigma_d)$, for targets $\pi \in \mathcal{P}(\mathbb{R}^d)$.

The independence Metropolis-Hastings algorithm (IMH) uses position-independent

proposals $q(x, y) = q(y)$ for all x. The IMH acceptance probability is

$$\alpha_{\text{IMH}}(x, y) = 1 \wedge \frac{w(y)}{w(x)},$$

where $w(x) = \pi(x)/q(x)$ is known as the importance weight function.

Algorithms that utilize the gradient of the log-density can better adapt to the local geometry of the target than the RWM, and thus mix faster in practice. The Metropolis-adjusted Langevin algorithm (MALA; Besag, 1994) uses proposals $q(x, y) = \mathcal{N}_d(y \mid x + \frac{1}{2}\Sigma_d \nabla \log \pi(x), \Sigma_d)$, arising from a discretization of a π-invariant stochastic process called the overdamped Langevin diffusion (see e.g. Roberts and Tweedie, 1996a). Other gradient-based MCMC algorithms we will encounter in this thesis include Hamiltonian Monte Carlo (HMC; Duane et al., 1987; Neal, 2011), the method of Horowitz (1991), and Hug and Hop (Ludkin and Sherlock, 2022). Appendix A.3.3 provides a description of HMC and Hug and Hop.

Another way to adapt to the geometry of the target π is through its conditional distributions, as with Gibbs samplers (Geman and Geman, 1984; Gelfand and Smith, 1990). The popular systematic scan Gibbs sampler has a kernel of the form $P = P_1 \dots P_d$, where for each i the partial kernel $P_i(x, A) = \pi(X_i \in A_i \mid X_{-i} = x_{-i})$ updates the i-th coordinate of the state according to the conditional distribution of π, while keeping the remaining coordinates fixed. Other Gibbs samplers can be devised by updating multiple coordinates at once ("blocking"; Liu et al., 1994), by marginalizing some conditionals out ("collapsing"; Liu, 1994; Van Dyk and Park, 2008), by replacing any partial kernel with a Metropolis-Hastings kernel targeting the corresponding conditional distribution, and by combining the partial kernels using a mixture (a "random scan") instead of a systematic scan.

Although we have so far only focused on asymptotically exact MCMC algorithms, practitioners also use asymptotically inexact MCMC algorithms, whose stationary distributions are not π but only an approximation of it. Asymptotically inexact MCMC

algorithms are generally devised by approximating or omitting parts of an exact MCMC algorithm, with the goal of reducing the amount of computation needed at each iteration. In view of the Monte Carlo error in ergodic averages $h_{k:m}$, the speed-up may more than compensate for the resulting error in the stationary distribution. Examples of asymptotically inexact MCMC algorithms considered in this thesis include the unadjusted Langevin algorithm (ULA; Roberts and Tweedie, 1996a) and the OBABO discretization of underdamped Langevin dynamics (Monmarché, 2021), which essentially correspond to MALA and the method of Horowitz (1991) with the Metropolis-Hastings acceptance step omitted.

2.2.3 Optimal scaling

MCMC algorithms often involve tuning parameters. Starting from the seminal work Roberts et al. (1997), *optimal scaling* theory has been developed to provide principled parameter tuning guidelines for MCMC algorithms. Optimal scaling theory typically considers a sequence of stylized target distributions of increasing dimension d and shows that the trajectories of a summary statistic of the MCMC algorithm converge to an explicit limiting process as the dimension $d \to \infty$. Parameter tuning guidelines are then obtained by optimizing an appropriate asymptotic measure of speed. We now discuss optimal scaling results for the stationary (Roberts et al., 1997) and the non-stationary, i.e. *transient*, (Christensen et al., 2005) phases of MCMC algorithms, which have qualitatively different behaviours. We focus on tuning the proposal covariance Σ_d of a RWM algorithm with Gaussian proposals $q(x, y) = \mathcal{N}_d(y \mid x, \Sigma_d)$.

Stationary phase

Roberts et al. (1997) considers $\pi_d(x) = \prod_{i=1}^d f(x_i)$ and $\Sigma_d = \lambda^2 I_d/d$, where $(\log f)'$ is assumed to be Lipschitz and sufficiently regular, and shows that the first coordinate of a stationary process $(X_{\lfloor td \rfloor})_{t \geq 0}$ converges weakly as $d \to \infty$ in the topology of Skorokhod

(1956) to a Langevin diffusion $(Y_{h(\lambda)t})_{t\geq 0}$, where

$$h(\lambda) = \lambda^2 a(\lambda), \quad a(\lambda) = 2\Phi(-\lambda I^{1/2}/2), \quad I = \mathrm{Var}_f(\log f'(X)).$$

The quantity $h(\lambda)$ is an asymptotic version of the *expected squared jump distance* (ESJD; e.g. Sherlock and Roberts, 2009)

$$\mathrm{ESJD}(\Sigma_d) = \mathbb{E}[\|X_{t+1} - X_t\|^2];$$

both $h(\lambda)$ and the ESJD can be viewed as measures of speed. The quantity $a(\lambda)$ is the asymptotic acceptance rate of the RWM kernel. The step size $\lambda^* = 2.38/I^{1/2} = \max_\lambda h(\lambda)$ maximizes the asymptotic speed measure, and corresponds to the optimal acceptance rate of $a(\lambda^*) = 23.4\%$.

The practical guideline of Roberts et al. (1997) is thus to tune Σ_d such that approximately 23.4% of proposals are accepted, which maximizes the limiting ESJD. This guideline turns out to be remarkably robust to the assumptions imposed on the target, and generally requires scaling $\Sigma_d \propto d^{-1}$ (Roberts and Rosenthal, 2001; Sherlock and Roberts, 2009; Sherlock, 2013; Yang et al., 2020). Matching the shape of the target also leads to performance gains: $\Sigma_d \propto \mathrm{Var}_\pi(X)$ and $\Sigma_d \propto \mathrm{Var}_\pi(\nabla \log \pi(X))^{-1}$ are optimal in terms of two different metrics (Negrea, 2022, Theorems 3.5-3.6).

Transient phase

Christensen et al. (2005, Theorem 1) considers $\pi_d = \mathcal{N}_d(0_d, I_d)$ and $\Sigma_d = \lambda^2 I_d/d$, chooses the summary statistic $(W_t^{(d)})_{t\geq 0} = (\|X_{\lfloor td \rfloor}\|^2/d)_{t\geq 0}$ initialized from $W_0^{(d)} = w_0$ for all d, and shows the weak convergence $W^{(d)} \implies w$ as $d \to \infty$ in the Skorokhod

topology to the ordinary differential equation

$$\mathrm{d}w_t = a_\lambda(w_t)\mathrm{d}t,$$
$$a_\lambda(w) = \lambda^2\Phi(-\lambda w^{-1/2}/2) + \lambda^2(1-2w)e^{\lambda^2(w-1)/2}\Phi(\lambda w^{-1/2}/2 - \lambda w^{1/2}).$$

The limiting process is deterministic and continuous, so this weak convergence is equivalent to convergence in probability in the uniform topology (see e.g. the discussion in Darling and Norris, 2008):

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \left\| W_t^{(d)} - w_t \right\|_\infty > \varepsilon\right) = 0 \text{ for all } T, \varepsilon \geq 0. \tag{2.2.5}$$

Because the target π_d is spherically symmetric, the process $W^{(d)}$ completely describes how far the RWM chain is from the main target mass; by standard concentration of measure results, the main target mass is reached when $W_t^{(d)} = 1 + \Theta(d^{-1/2})$. At the same time, the convergence (2.2.5) means that the process $W^{(d)}$ concentrates in a tube of width $o(1)$ around w. This means that we can approximately track the transient phase of the RWM algorithm using the process w, that the duration of the transient phase is approximately d times the time required for w to reach $w_t = 1 + \Theta(d^{-1/2})$, and that the absolute value of the drift a_λ can be viewed as an asymptotic measure of speed for the transient phase.

On the one hand, this result indicates that the RWM requires $\Theta(d \log d)$ iterations to converge. On the other hand, since $\lim_{\lambda\to 0} a_\lambda(w) = \lim_{\lambda\to\infty} a_\lambda(w) = 0$, the correct step size scaling is $\Sigma_d \propto d^{-1}$, as larger or smaller scalings result in limiting processes that do not move. In contrast to the stationary phase, in the transient phase the optimal step size $\lambda^*(w) = \arg\max_\lambda |a_\lambda(w)|$ is position-dependent, but the choice $\lambda = 2.38$ that is optimal at stationarity also performs well during transience (Christensen et al., 2005, Figure 1).

In Chapter 3, we extend the optimal scaling results of Christensen et al. (2005) to

couplings of RWM chains. The weak convergence results we derive should be understood in the sense of equation (2.2.5).

2.3 Unbiased Markov chain Monte Carlo

Much of this thesis is motivated by recent efforts to remove the burn-in bias of MCMC estimators using couplings (Glynn and Rhee, 2014; Jacob et al., 2020b). We now introduce these, broadly following Jacob (2020); Atchadé and Jacob (2024).

2.3.1 Advantages of unbiased estimators

The default strategy to parallelize MCMC is to run multiple independent chains in parallel, each producing a copy of the standard MCMC estimator $h_{k:m}$, then to average these copies. While averaging over parallel chains reduces the variance, it cannot "average out" the burn-in bias, which remains $\mathbb{E}[h_{k:m}] - \mathbb{E}_\pi[h(X)] \neq 0$. Controlling the error thus requires a sufficiently large number of iterations m *on each processor*, which limits the potential benefits of parallelization (Rosenthal, 2000), as well as a judicious choice of the burn-in k, which can be difficult to do in practice.

Unbiased MCMC estimators (Jacob et al., 2020b) bypass the issue of burn-in, and allow for expectations to be estimated consistently by averaging over independent copies. Conveniently, these copies can be generated embarrassingly in parallel. When the unbiased MCMC estimator has a finite variance, the estimator obtained by averaging over independent copies obeys a central limit of the form (2.2.1); the average therefore converges at the usual Monte Carlo rate in the number of copies, and its uncertainty can be straightforwardly quantified with the usual Gaussian confidence intervals based on the sample variance.

2.3.2 Unbiased estimation by random truncation

Suppose that we wish to estimate the limit s_∞ of a sequence of biased approximations $(s_k)_{k\geq 0}$, where refining the approximation from s_{k-1} to s_k costs one unit of computation.

Rhee (2013), based on an idea that dates back to Forsythe and Leibler (1950), expresses the sequence as a series $s_k = \sum_{t=0}^{k} \Delta_t$, where $\Delta_0 = s_0$ and $\Delta_k = s_k - s_{k-1}$ for $k \geq 1$, introduces a random truncation variable $K \in \mathbb{N} \cup \{0\}$, and considers the estimator

$$G = \sum_{k=0}^{K} \frac{\Delta_k}{\mathbb{P}(K \geq k)}.$$

Since $\mathbb{E}[G] = \mathbb{E}[\sum_{k\geq 0} \Delta_k \mathbb{1}\{K \geq k\}/\mathbb{P}(K \geq k)]$, the estimator G would be unbiased provided that we could interchange the expectation and the sum, in particular when $\mathbb{E}[|G|] < \infty$.

Crucially, the method also works when $(\Delta_k)_{k\geq 0}$ are random and we wish to compute the limiting quantity $s_\infty = \sum_{k\geq 0} \mathbb{E}[\Delta_k]$, which suggests that it could also be used to remove the bias of MCMC estimates.

The expected cost of G is $\mathbb{E}[K] = \sum_{k\geq 0} k\mathbb{P}(K = k)$, while the variance of G involves the term $\sum_{k\geq 0} \mathbb{E}[\Delta_k^2]/\mathbb{P}(K \geq k)$. On the one hand, $(\mathbb{E}[\Delta_k^2])_{k\geq 0}$ must decrease quickly enough to ensure that the estimator has a finite variance. On the other hand, K must strike a trade-off, with lighter tails providing a smaller expected cost, at the cost of increased variance (Rhee and Glynn, 2015).

McLeish (2011, Section 3.2) attempted to perform unbiased MCMC by estimating the expectation $\mathbb{E}_\pi[h(X)]$ based on the convergent sequence of standard MCMC estimators $(h_{0:k})_{k\geq 0}$. However, this approach is problematic because $\Delta_k = h_{0:k} - h_{0:(k-1)}$ decays at rate k^{-1}, which is too slow to ensure both unbiasedness and finite expected cost. Glynn and Rhee (2014) demonstrated the feasibility of unbiased MCMC by constructing $(\Delta_k)_{k\geq 0}$ based on a *pair of coupled Markov chains*. However, the specific coupling constructions considered therein are too restrictive to be practical. Next, we

present the refinement of Jacob et al. (2020b) that makes the unbiased MCMC idea of Glynn and Rhee (2014) practical.

2.3.3 Unbiased Markov chain Monte Carlo with couplings

Let $\bar{P}((x,y),\cdot)$ be a joint Markov kernel with marginals $(P(x,\cdot),P(y,\cdot))$, which for all (x,y) and all measurable $A,B\subseteq\mathcal{X}$ satisfies

$$\bar{P}((x,y),(A,\mathcal{X}))=P(x,A),\;\;\bar{P}((x,y),(\mathcal{X},B))=P(y,B). \tag{2.3.1}$$

Fix a *time-lag* parameter $L\in\mathbb{N}$. Consider the following *lagged coupling* of Markov chains $(X_t,Y_t)_{t\geq 0}$ initialized at $X_0,Y_0\sim\pi_0$ and evolving according to

$$\begin{aligned} X_t \mid X_{t-1} &\sim P(X_{t-1},\cdot) \text{ for } t\in[L],\\ (X_{t+L},Y_t)\mid(X_{t+L-1},Y_{t-1}) &\sim \bar{P}((X_{t+L-1},Y_{t-1}),\cdot) \text{ for } t\geq 1. \end{aligned} \tag{2.3.2}$$

This coupling ensures that $X_t,Y_t\sim\pi_t$ for all $t\geq 0$.

Furthermore, suppose that we design the joint kernel $\bar{P}$ such that the chains *meet* at some time τ, after which they *coalesce*: $X_{t+L}=Y_t$ for all $t\geq\tau$. Pitman (1976) calls such couplings "successful". Coalescence is possible because the chains evolve according to the same marginal kernel P. It is often easy to ensure that coalescence happens at the first meeting time (Rosenthal, 1997), i.e. $\tau=\inf\{t:X_{t+L}=Y_t\}$, which we henceforth assume.

Jacob et al. (2020b) propose to estimate $\mathbb{E}_\pi[h(X)]$ using the lagged coupling via

$$H_k=h(X_k)+\sum_{j\geq 0}\{h(X_{k+(j+1)L})-h(Y_{k+jL})\},$$

where the extension to $L>1$ is due to Vanetti and Doucet (2020). Provided that we

could interchange the expectation and the sum, we would have

$$\mathbb{E}[H_k] = \mathbb{E}[h(X_k)] + \sum_{j\geq 0} \big\{\mathbb{E}[h(X_{k+(j+1)L})] - \mathbb{E}[h(X_{k+jL})]\big\} = \mathbb{E}_\pi[h(X)],$$

so H_k is unbiased if the differences $(h(X_{t+L}) - h(Y_t))_{t\geq 0}$ decay quickly enough. The following time-averaged estimator is also unbiased (e.g. Douc et al., 2023):

$$H_{k:m} = \frac{1}{m-k+1}\sum_{t=k}^{m} H_t = \frac{1}{m-k+1}\left[\sum_{t=k}^{m} h(X_t) + \sum_{t\geq k} c_{k,m,L}(t)\{h(X_{t+L}) - h(Y_t)\}\right],$$

where $c_{k,m,L}(t) = 1 + \lfloor (t-k)/L \rfloor - \lceil 0 \vee (t-m)/L \rceil$. Importantly, because the coupling coalesces at τ, the series in $H_{k:m}$ truncates at $t \leq \tau - 1$. The estimator $H_{k:m}$ can therefore be computed in finite time *without introducing a truncation variable* K, which can be challenging to choose appropriately in practice (Agapiou et al., 2018).

Jacob et al. (2020b); Biswas et al. (2019); Middleton et al. (2020); Douc et al. (2023) provide various formal conditions under which $H_{k:m}$ is unbiased and has a finite variance. These conditions relate to the tail decay of the meeting time τ. For example, it suffices for τ to have geometric tails and for $h \in L^{2+\delta}(\pi)$ for some $\delta > 0$ (Biswas et al., 2019). It also suffices for τ to have polynomial tails, under a stronger moment condition than the MCMC central limit theorem (2.2.3) and a mild condition on the initialization π_0 (Douc et al., 2023).

The estimator $H_{k:m}$ has two terms, one of which is the standard MCMC average $h_{k:m}$. Practical experience indicates that $H_{k:m}$ is in general less efficient than a stationary ergodic average $h_{k:m}$, because often $\mathrm{Var}(H_{k:m}) > \mathrm{Var}(h_{k:m})$, and because $H_{k:m}$ requires additional simulation compared to $h_{k:m}$. As we will see in Chapter 4, a judicious choice of the parameters (k, m, L) can nevertheless lead to unbiased estimators $H_{k:m}$ with an efficiency that is comparable to that of standard MCMC.

2.3.4 Estimating convergence

Biswas et al. (2019) use the lagged coupling construction (2.3.2) to estimate the convergence of Markov chains in 1-Wasserstein distance. As we note in Chapter 5, the idea extends straightforwardly to p-Wasserstein distances of all orders $p > 1$. The triangle inequality provides

$$\mathcal{W}_{p,c}(\pi, \pi_t) \leq \sum_{j \geq 0} \mathcal{W}_{p,c}(\pi_{t+(j+1)L}, \pi_{t+jL}) \leq \sum_{j \geq 0} \mathbb{E}[c(X_{t+(j+1)L}, Y_{t+jL})^p]^{1/p},$$

where finally we used the definition of the Wasserstein distance as the minimum over couplings. In practice, the bound is estimated using several replicates, replacing expectations by empirical averages. Since each replicate coalesces at a finite time $\tau < \infty$, the series truncates and can be computed in finite time. In the case of the total variation distance, where $p = 1$ and the ground metric is $c(x, y) = \mathbb{1}\{x \neq y\}$, the bound becomes

$$\mathrm{TV}(\pi, \pi_t) \leq \sum_{j \geq 0} \mathbb{P}(X_{t+(j+1)L} \neq Y_{t+jL}) = \sum_{j \geq 0} \mathbb{P}(\tau > t + jL) = \mathbb{E}\left[0 \vee \lceil (\tau - t)/L \rceil\right].$$

2.3.5 Couplings of Markov chains

For a Metropolis-Hastings kernel P_{MH}, a coupling that allows for the chains to meet could proceed as follows. We couple the draws from $(P_{\mathrm{MH}}(x, \cdot), P_{\mathrm{MH}}(y, \cdot))$ in two stages, first coupling the proposals drawn from $(q(x, \cdot), q(y, \cdot))$, then coupling the acceptance steps. To make the proposals identical with maximal probability, we can use Algorithm 1. Thereafter, we can maximize the probability that the proposals are accepted simultaneously by using the same uniform variate to accept both proposals. Algorithm 3 describes this coupling; coalescence is guaranteed to happen when the proposals are identical in step 1, i.e. $X' = Y'$, and both proposals are subsequently accepted in steps 3 and 4.

Algorithm 3 A coupling (X, Y) of Metropolis-Hastings kernels $(P_{\mathrm{MH}}(x, \cdot), P_{\mathrm{MH}}(y, \cdot))$

1. Sample (X', Y') from the coupling of $(q(x, \cdot), q(y, \cdot))$ given by Algorithm 1.
2. Sample $U \sim \mathrm{Unif}(0, 1)$.
3. If $U \leq \alpha(x, X')$, then return $X = X'$. Else, return $X = x$.
4. If $U \leq \alpha(y, Y')$, then return $Y = Y'$. Else, return $Y = y$.

Jacob et al. (2020b) discuss how the coupling described by Algorithm 3 does ensure that estimators $H_{k:m}$ are unbiased, under standard, but technical, geometric ergodicity assumptions (e.g. Roberts and Tweedie, 1996b). However, satisfactory performance in theory need not guarantee good performance in practice: in particular, for RWM kernels, the meeting times grow exponentially with the dimension d when using Algorithm 3, which contrasts with the $\mathcal{O}(d \log d)$ time required for the marginal chains to converge (Christensen et al., 2005; Andrieu et al., 2024). We consider the design of more effective couplings of the RWM in Chapter 3. The strategy we pursue is to design a coupling that approximately attains $\mathcal{W}_{2,\|\cdot\|}(P_{\mathrm{MH}}(x, \cdot), P_{\mathrm{MH}}(y, \cdot))$, with the goal of first bringing the chains close enough together to have a good chance of coalescing, then combining this with another coupling that allows for meetings to occur. Next, we explain how to validly combine couplings of Markov kernels.

Using condition (2.3.1), we can devise valid couplings of P based on position-dependent mixtures of any two couplings $\bar{P}((x, y), \cdot), \bar{P}'((x, y), \cdot) \in \Gamma(P(x, \cdot), P(y, \cdot))$:

$$\varepsilon(x, y)\bar{P}((x, y), \cdot) + (1 - \varepsilon(x, y))\bar{P}'((x, y), \cdot) \in \Gamma(P(x, \cdot), P(y, \cdot)),$$

for any (x, y)-indexed probability $\varepsilon(x, y) \in [0, 1]$. This allows us to combine the individual strengths of several couplings. For instance, setting $\varepsilon(x, y) = \mathbb{1}\{\|x - y\| \leq \delta\}$ results in a "two-scale" coupling strategy (e.g. Bou-Rabee et al., 2020) that uses the joint kernel $\bar{P}$ when the chains are close together, and the joint kernel $\bar{P}'$ otherwise.

2.3.6 Related work

Couplings have a long history in Markov chain theory, going back to at least Doeblin (1938), and have nowadays become a standard tool for bounding the convergence of Markov chains (e.g. Roberts and Rosenthal, 2004; Douc et al., 2018).

The use of couplings for MCMC methods is comparatively new. Reutter and Johnson (1995); Johnson (1996, 1998); Neal (1999); Sixta et al. (2025) provide various MCMC convergence diagnostics based on replicate couplings of Markov chains. Frigessi et al. (2000); Pinto and Neal (2001); Goodman and Lin (2009); Piponi et al. (2020) devise variance reductions schemes using couplings of Markov chains. Propp and Wilson (1996) propose a method termed "coupling from the past" that enables exact sampling from the target π using only π-invariant Markov chains, but requires stringent conditions to be implementable.

The unbiased estimation method of Jacob et al. (2020b) has been extended and applied in several ways. Jacob et al. (2020b, Section 5.5) use the law of total expectation to produce unbiased posterior summaries for certain modular statistical models (Jacob et al., 2017). Ruiz et al. (2021) unbiasedly estimate the gradients of certain latent variable models. Wang and Wang (2023) construct unbiased estimators of functions of expectations. Douc et al. (2023) use the unbiased estimators of Jacob et al. (2020b) and a further telescopic series argument to unbiasedly estimate the asymptotic variance of MCMC algorithms.

The problem of designing effective implementable couplings of MCMC algorithms has been explored in various works. In addition to Jacob et al. (2020b), we list a selection of these: O'Leary (2021) focuses on the RWM; Heng and Jacob (2019); Xu et al. (2021) on MALA and variants of HMC (e.g. Betancourt, 2017); Biswas et al. (2022); Atchadé and Wang (2023) on Gibbs samplers in continuous spaces; Agapiou et al. (2018) on the preconditioned Crank-Nicholson algorithm (e.g Cotter et al., 2013); Middleton et al. (2020) on pseudo-marginal MCMC algorithms (Andrieu and Roberts,

2009); Middleton et al. (2019); Jacob et al. (2020a); van den Boom et al. (2022) on particle MCMC samplers (Andrieu et al., 2010); Corenflos et al. (2023) on samplers based on piecewise-deterministic Markov processes (Fearnhead et al., 2018). Couplings of MCMC samplers operating on complex spaces have also been considered, see Nguyen et al. (2022) for Gibbs samplers operating on partitions, Kelly et al. (2023) for RWM algorithms operating on phylogenetic trees, and Bortolato (2024) for RWM algorithms operating on manifolds.

Recent theoretical works have obtained quantitative convergence bounds based on couplings that are often directly implementable, or that can guide the design of effective implementable couplings. For instance, reflection couplings are well-known to be effective in the case of diffusion processes (Lindvall and Rogers, 1986; Eberle, 2016; Eberle et al., 2019), and can even be optimal in terms of meeting times in certain cases (e.g. Connor, 2007, Chapter 3). CRN couplings have been used to quantify the convergence of various MCMC algorithms, such as variants of MALA (Eberle, 2014) and HMC (Mangoubi and Smith, 2019, 2021). Mixtures of CRN and reflection couplings have been used to quantify the convergence of HMC (Bou-Rabee et al., 2020).

Finally, various works have devised more sophisticated couplings that allow for a positive chance of meeting. By modifying two-stage constructions like Algorithm 3, which consist of conditionally maximal couplings, Wang et al. (2021) constructs maximal couplings of the entire Metropolis-Hastings transition kernel. O'Leary and Wang (2024) characterizes the form of all Metropolis-Hastings transition kernel couplings that are maximal. Corenflos and Särkkä (2022) constructs asymptotically maximal couplings of random variables based on rejection sampling from a dominating coupling. Dau and Chopin (2023, Algorithm S.9) provides a way of turning any coupling into one that can allow for exact meetings, by using an additional draw from the "overlap region" of the densities (see e.g. Figure 2.1.1).

Chapter 3

Scalable couplings for the random walk Metropolis algorithm

3.1 Introduction

Couplings of Markov chain Monte Carlo (MCMC) algorithms have attracted much interest recently due to their ability to nullify the bias of MCMC estimates (Jacob et al., 2020b), conservatively estimate the rate of convergence (Biswas et al., 2019) of MCMC algorithms and the asymptotic bias of approximate inference procedures (Biswas and Mackey, 2024), as well as unbiasedly estimate the asymptotic variance of MCMC algorithms (Douc et al., 2023). One appeal of these methods is that they are able to exploit parallelism without requiring communication between processors.

In the context of unbiased MCMC and the related convergence quantification methodology, a Markovian coupling of two chains should be designed such that the chains meet after a finite number of iterations. The meeting time acts as a lower bound for the length of the simulation, and as such the efficiency of a coupling can be assessed through the distribution of the meeting time. As remarked in Jacob et al. (2020b), it is paramount to design couplings which have favourable high-dimensional properties, in the sense

that the meeting times reflect the true rate of convergence and mixing of the underlying marginal Markov chains. At the same time, the rich design space means that it is challenging to devise efficient couplings, and this design is regarded to be an art form in general (O'Leary and Wang, 2024).

We focus here on the design and analysis of scalable couplings for the random walk Metropolis (RWM) algorithm with Gaussian proposals. Methodologically (see Section 3.3), we argue for the importance of couplings which are contractive and we design couplings which attempt to optimize for contraction in squared Euclidean distance. Our *Gradient Common Random Numbers* (GCRN) coupling is provably optimally contractive in certain high-dimensional asymptotic regimes, is insensitive to the eccentricity of the target, and is consistently able to contract the chains to within a distance where coalescence in one step is achievable. Once the chains are close, GCRN or other contractive couplings can be swapped for ones which allow for exact meetings; we exemplify and advocate for such two-scale strategies in the sequel.

Our proposed couplings are designed to overcome the shortcomings of those currently available. The reflection-maximal coupling (Jacob et al., 2020b, Section 4.1), arguably the most promising candidate to date, has been seen to scale well with the dimension when the target is spherically symmetric (Jacob et al., 2020b; O'Leary, 2021). However, for high-dimensional targets which do not possess spherical symmetry this coupling has been seen to perform poorly (Papp and Sherlock, 2025a); it does not contract the chains sufficiently unless the step size of the coupled RWM algorithms is chosen to be significantly smaller than is optimal for mixing. The present work validates the favourable behaviour of the reflection coupling in the spherical case, offers an explanation for the issue when spherical symmetry is lacking, and proposes an alternative reflective coupling (Section 3.6.3) which alleviates the problem.

Our analysis bridges the gap between the coupled sampling and the optimal dimensional scaling literatures in MCMC (see the review Roberts and Rosenthal, 2001 for

optimal scaling in the stationarity phase, and Christensen et al., 2005 for the transient phase). The ODE limit in Christensen et al. (2005, Theorem 1) for high-dimensional spherical Gaussian targets, originally developed to explain the transient phase of the RWM algorithm, underpins our approach. We extend this scaling limit to coupled pairs of RWM chains in the spherical Gaussian case (Section 3.4) and further extend the salient points of this analysis to the elliptical Gaussian case (Section 3.5). Related to these scaling limits, we introduce a notion of asymptotic optimality; this extends beyond the Gaussian case (Section 3.6.2) and may guide the design of effective couplings for other MCMC algorithms.

We conclude with a series of experiments illustrating the practical appeal of our proposed couplings (Section 3.7) and a discussion of our findings and directions for further work (Section 3.8).

3.2 Background

This work is motivated by lagged coupling methodology (Jacob et al., 2020b; Biswas et al., 2019), which can be used to estimate the rate of convergence of MCMC algorithms and to obtain unbiased MCMC estimators, and which we briefly recall here. Our set-up differs slightly from the literature in that we start the index of the first chain at $-L$ rather than 0, however this will add clarity to the sequel.

Consider a π-invariant Markov kernel K, and a joint Markov kernel $\bar{K}((x, y), (\cdot, \cdot))$ with marginals $(K(x, \cdot), K(y, \cdot))$. Throughout this paper, we think of K as a Metropolis-Hastings kernel and the joint kernel $\bar{K}$ as specifying some coupling of the proposals and of the acceptance steps. We construct two Markov chains $(X_t)_{t\geq -L}$ and $(Y_t)_{t\geq 0}$. Each chain evolves marginally according to K, and after head-starting the X-chain by $L \geq 1$ iterations they evolve jointly according to $\bar{K}$:

1. Sample $X_0 \sim \pi_0 K^L$ and $Y_0 \sim \pi_0$.

2. Sample $(X_{t+1}, Y_{t+1}) \mid (X_t, Y_t) \sim \bar{K}((X_t, Y_t), \cdot)$ for $t \geq 0$.

Furthermore, we design the joint kernel $\bar{K}$ such that there is an almost surely finite meeting time $\tau = \inf\{t \geq 0 : X_t = Y_t\}$ and such that $X_t = Y_t$ for all $t \geq \tau$.

This construction can be used to estimate the rate of convergence of Markov chains (Biswas et al., 2019). Suppose that we wish to quantify the rate of convergence in a p-Wasserstein distance (Villani, 2009) of order $p \geq 1$, defined as $\mathcal{W}_p(\mu, \nu) = \inf_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}[c(X, Y)^p]^{1/p}$, where c is some real-valued ground metric and where $\Gamma(\mu, \nu)$ is the set of all couplings of the distributions (μ, ν). Let $\pi_t = \pi_0 K^t$ be the time-t marginal distribution of the Y-chain, so that $Y_t \sim \pi_t$ for all $t \geq 0$. Using firstly the triangle inequality, then the definition of the Wasserstein distance, we have that

$$\mathcal{W}_p(\pi, \pi_t) = \mathcal{W}_p(\pi_\infty, \pi_t) \leq \sum_{j\geq 0} \mathcal{W}_p(\pi_{t+(j+1)L}, \pi_{t+jL}) \leq \sum_{j\geq 0} \mathbb{E}^{1/p} \left[c(X_{t+jL}, Y_{t+jL})^p\right]. \tag{3.2.1}$$

By repeatedly simulating the pair (X, Y) and by replacing expectations with empirical averages, we obtain a consistent estimator of this upper bound. Conveniently, pairs of coupled chains can be simulated in parallel; the meeting times τ ensure that the estimator is computed in finite time. The lagged coupling framework also allows for the unbiased estimation of expectations of test functions of interest (see Jacob et al., 2020b and Appendix A.1) which enables principled parallel MCMC through averaging across multiple (pairs of) chains simulated in parallel. Such unbiased estimators also facilitate modular inference with cut posterior distributions (Jacob et al., 2017) and can be used as part of a wider multilevel Monte Carlo framework in order to unbiasedly estimate functions of expectations (Wang and Wang, 2023).

The effectiveness of the lagged coupling framework relies on designing the joint kernel $\bar{K}$ such that the meeting times τ are small, since large meeting times lengthen the duration of the simulation, loosen the upper bounds on the rate of convergence, and inflate the variance of unbiased estimators (see Jacob et al., 2020b for this final point).

In this paper, we focus on designing kernels $\bar{K}$ whose meeting times scale well with the dimension of the target π. Since the choice of lag parameter L should be guided by the choice of coupling kernel $\bar{K}$, for the remainder of the paper we focus on the joint Markov chain $(X_t, Y_t)_{t\geq 0}$, where the starting conditions (X_0, Y_0) can be arbitrary.

We illustrate the use of couplings to quantify the bias of approximate sampling algorithms (Biswas and Mackey, 2024) and to unbiasedly estimate the asymptotic variance of MCMC algorithms (Douc et al., 2023) in the experiments of Section 3.7.

3.3 Couplings of the RWM algorithm

We first restrict our attention to the random walk Metropolis (RWM) kernel K with spherical Gaussian proposals. All coupling kernels $\bar{K}$ are induced by joint updates of the form

$$X_{t+1} = X_t + hZ_xB_x, \quad Y_{t+1} = Y_t + hZ_yB_y, \tag{3.3.1}$$

where $B_x = \mathbb{1}\{\log U_x \leq \log \pi(X_t + hZ_x) - \log \pi(X_t)\}$ is the Bernoulli acceptance indicator, π is a d-dimensional target, and $Z_x \sim \mathcal{N}_d(0_d, I_d)$ and $U_x \sim \text{Unif}(0,1)$ are independent. Analogous notation applies to the Y-chain. Throughout this paper, we scale the step size as $h = \ell d^{-1/2}$, which ensures that acceptance rates remain stable as the dimension grows (Roberts et al., 1997; Christensen et al., 2005).

While the simplicity of spherical proposals is convenient for analysis, in practice better mixing may be obtained with other covariance structures. It is straightforward to extend the couplings considered in this paper to non-spherical Gaussian proposals, see Appendix A.2.2.

3.3.1 The importance of contractivity in high dimensions

To try to make the chains meet quickly, one might be tempted to use a coupling $\bar{K}$ which maximizes the chance of coalescing the chains at each iteration (see Wang et al.

(2021) for examples). However, as the dimension grows, such couplings can perform increasingly poorly for the RWM algorithm.

The issue is that the RWM proposals become increasingly local as the dimension increases, yet at the same time the distance between the coupled chains grows. To be able to coalesce the chains, a Markovian coupling of RWM chains must first propose the same state in both chains. The probability of coalescing the chains in one iteration is therefore upper bounded by the volume of overlap of the proposal densities (which is analytically tractable in terms of the standard Gaussian cumulative density function $\Phi(\cdot)$, e.g. Heng and Jacob, 2019):

$$\mathbb{P}(X_{t+1} = Y_{t+1} \mid X_t, Y_t) \leq 2\Phi\left(-\frac{1}{2h}\|X_t - Y_t\|\right) \leq 2\exp\left(-\frac{d}{8\ell^2}\|X_t - Y_t\|^2\right), \quad (3.3.2)$$

where we finally used the Chernoff bound and that $h = \ell d^{-1/2}$. Two independent chains will typically start $\|X_0 - Y_0\|^2 = \mathcal{O}(d)$ apart, yet the chance of coalescing in one iteration is infinitesimally small unless $\|X_t - Y_t\|^2 = \mathcal{O}(d^{-1})$. If the coupling cannot contract the chains to within, say, $\mathcal{O}(1)$ squared distance, then one can expect meeting times to grow exponentially with the dimension d. Clearly, this is in stark contrast with the $\mathcal{O}(d)$ time required for RWM chains to converge (Christensen et al., 2005; Andrieu et al., 2024).

In high dimensions, the coupling should therefore primarily focus on contracting the chains, as opposed to maximizing the probability of coalescing at each iteration. For the coupling to be scalable, we therefore need to design a contractive kernel $\bar{K}$.

3.3.2 Optimizing for contraction

Setting up the design of $\bar{K}$ as an optimization problem, a natural objective is the expected contraction in squared Euclidean distance. It is straightforward, following the

expansion (3.4.1) below, to show that

$$\arg\min_{\bar{K}\in\mathcal{C}} \mathbb{E}\big[\|X_{t+1} - Y_{t+1}\|^2 \mid X_t, Y_t\big] = \arg\max_{\bar{K}\in\mathcal{C}} \mathbb{E}\big[h^2 Z_x^\top Z_y B_x B_y \mid X_t, Y_t\big],$$

where $\mathcal{C}$ is any subset of $\mathcal{M}$, the class of all Markovian couplings (3.3.1). We call the quantity $\mathbb{E}\big[h^2 Z_x^\top Z_y B_x B_y \mid X_t, Y_t\big]$ the *expected jump concordance* (EJC). To optimize the expected contraction of the chains, we therefore need to maximize the EJC.

The form of the EJC suggests the following coupling strategy: correlate the acceptances B_x, B_y maximally (i.e. try to accept simultaneously in both chains) and correlate the proposal noises Z_x, Z_y maximally. The key observation is that, *as the dimension increases, these two objectives become less and less constrained by each other, so it becomes possible to satisfy both of them simultaneously.* Focusing on the acceptance steps, a Taylor expansion of the log acceptance ratio yields

$$B_x = \mathbb{1}\big\{\log U_x \le hZ_x^\top \nabla \log \pi(X_t) + (h^2/2) Z_x^\top \nabla^2 \log \pi(\bar{X}_t) Z_x\big\}, \tag{3.3.3}$$

where ∇^2 denotes the Hessian, and $\bar{X}_t$ is on the line segment from X_t to $X_t + hZ_x$. As a function of Z_x, most of the variation in B_x therefore stems from the projection of Z_x onto the logarithmic gradient $\nabla \log \pi(X_t)$. To try to make the chains accept simultaneously as often as possible, it is therefore sensible to maximally correlate $\{Z_x^\top \nabla \log \pi(X_t), Z_y^\top \nabla \log \pi(Y_t)\}$. As each projection only constrains a single coordinate, in high enough dimensions $\{Z_x, Z_y\}$ can still be correlated nearly maximally. This is the motivation behind the gradient-based GCRN coupling, which we introduce in Section 3.3.3.

The EJC is well-defined in standard high-dimensional asymptotic regimes; we will call a coupling *asymptotically optimal* if it maximizes the EJC in the limit. In such regimes, the variation from the Hessian term of (3.3.3) vanishes (e.g. Sherlock, 2013), so we expect couplings like GCRN which account for all of the variation from the

gradient term to be close to optimal. The EJC can in fact be viewed as the coupling-based analogue of the expected squared jumping distance (ESJD; Sherlock and Roberts, 2009), a widely-used measure of mixing efficiency. The ESJD has the limiting formula $\text{ESJD}(\ell) = 2\ell^2\Phi(-\ell/2)$ for a standard Gaussian target, which returns the optimal scaling of $\ell = 2.38$ and optimal acceptance rate of 23.4%. The EJC reduces to the ESJD when the the chains are stationary and are identical; it is therefore unsurprising that the same scaling $\ell = 2.38$ turns out (see Section 3.4.2) to be close to optimal for our GCRN coupling.

3.3.3 The couplings under consideration

This work focuses on the natural class of *product couplings* $\mathcal{P}$, which contains all couplings of the updates (3.3.1) such that (U_x, U_y) are independent of (Z_x, Z_y). We introduce three couplings from $\mathcal{P}$, all of which synchronize the acceptance uniforms to $U_y = U_x$, also called a common random numbers (CRN) strategy. The difference is in the coupling of the proposal increments (Z_x, Z_y):

1. **CRN:** $Z_y = Z_x$;
2. **Reflection:** $Z_y = Z_x - 2(e^\top Z_x)e$;
3. **GCRN:** $Z_x = Z - (n_x^\top Z)n_x + Z_\nabla n_x$ and $Z_y = Z - (n_y^\top Z)n_y + Z_\nabla n_y$;

where: $e = \text{Nor}(X_t - Y_t)$, $n_x = \text{Nor}(\nabla \log \pi(X_t))$ and $n_y = \text{Nor}(\nabla \log \pi(Y_t))$; $\text{Nor}(x) = x/\|x\|$ denotes the normalization operation; $Z_\nabla \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{N}_d(0_d, I_d)$ are independent. The CRN and reflection couplings serve as baselines and are already established (see e.g. O'Leary, 2021). By convention, we default to the CRN coupling when a vector to be normalized is null.

The new GCRN (*Gradient Common Random Numbers*) coupling attempts to synchronize acceptance events by ensuring that $n_x^\top Z_x = n_y^\top Z_y$; it contracts the chains due to synchronized movement towards the mode. By design and in certain high-dimensional

asymptotic regimes, GCRN is optimal for contraction within the class $\mathcal{P}$ (see Theorems 3.4.1, 3.5.2 and 3.6.2). We construct in Appendix A.2.1 an implementable modification of GCRN that is asymptotically optimal over the entire class $\mathcal{M}$; we prefer the simpler GCRN coupling as any additional gain appears very small (see Figures 3.4.1 and 3.5.2). Other variants of GCRN with similar high-dimensional properties could be devised, for instance synchronizing $n_x^\top Z_x = n_y^\top Z_y$ using an appropriate reflection or rotation. Aiming to combine favourable properties of contractive and reflective couplings, we propose a hybrid between the GCRN and reflection couplings in Section 3.6.3.

In practice, once the chains become close enough to have a reasonable chance of coalescing in one step, we propose to swap from a contractive coupling like GCRN to one that allows for exact meetings. We use such two-scale approaches in the experiments of Section 3.7; our coalescive coupling of choice is the *reflection-maximal* coupling (Jacob et al., 2020b, Section 4.1; see also Appendix A.2.2). Two-scale couplings have previously been considered in e.g. Biswas et al. (2022); Bou-Rabee et al. (2020).

As a major component to this work, we develop theory which explains the behaviour of coupled RWM chains in high dimensions. We focus on GCRN and the two baselines; a recurring quantity in our analysis is the joint distribution of $(n_x^\top Z_x, n_y^\top Z_y)$ which we characterize in Proposition 3.3.1 below. As with all our results in the main text, this is proved in Appendix A.4.

Proposition 3.3.1. *It holds that* $(n_x^\top Z_x, n_y^\top Z_y) \mid \{X_t, Y_t\} \sim \mathrm{BvN}(\rho)$*, the bivariate normal distribution with unit coordinate-wise variances and correlation* $\rho \in [-1, 1]$*, where the coupling-specific value* ρ *is*

$$\rho_{\mathrm{crn}} = n_x^\top n_y, \quad \rho_{\mathrm{refl}} = n_x^\top n_y - 2(n_x^\top e)(n_y^\top e), \quad \rho_{\mathrm{gcrn}} = 1.$$

3.4 Analysis: standard Gaussian case

We begin our analysis of the couplings of Section 3.3.3 by considering a standard Gaussian target $\pi^{(d)} = \mathcal{N}_d(0_d, I_d)$ in increasing dimension d. Clearly, this setting is very stylized, however it will allow us to obtain limit theorems that cleanly characterize the behaviour of the couplings in high dimensions. Besides, scaling limits for the RWM algorithm often rely on the target behaving asymptotically like a Gaussian (Roberts et al., 1997), and the conclusions of such analyses have been seen to hold much more widely in practice.

Firstly, we show in Theorem 3.4.1 below that the GCRN coupling is asymptotically optimal for contraction among the class $\mathcal{P}$ of product couplings, in that it optimizes a limiting form of the EJC. This coupling is therefore expected to perform well even in high dimensions.

Thereafter, our analysis centers on the three-dimensional process

$$W^{(d)} = \left(W_t^{(d)}\right)_{t\geq 0} = \frac{1}{d}\left(\|X_{\lfloor td\rfloor}\|^2, \|Y_{\lfloor td\rfloor}\|^2, X_{\lfloor td\rfloor}^\top Y_{\lfloor td\rfloor}\right)_{t\geq 0},$$

where the speed-up factor d corresponds to the natural time-scale under the step size scaling $h = \ell d^{-1/2}$. The form $B_x = \mathbb{1}\{\log U_x \leq -hZ_x^\top X_t - (h^2/2)\|Z_x\|^2\}$ of the acceptance step ensures that the process $W^{(d)}$ is Markovian under our couplings (see Appendix A.4). As the target is spherically symmetric, the first and second coordinates of this process are radial components describing the marginal behaviour of each chain. The inner product in the third coordinate captures the remaining joint behaviour of the chains under a given coupling.

We show in Theorem 3.4.4 that $W^{(d)}$ converges weakly to the solution of an ordinary differential equation (ODE) as the dimension d grows. Our approach follows Christensen et al. (2005) and extends this path-breaking work to pairs of Markov chains. By thereafter analyzing the ODE, we shed light on the high-dimensional behaviour of the

coupled chains.

3.4.1 Asymptotic optimality

It is natural to ask which coupling contracts the chains the most in the considered high-dimensional regime. As shown in Section 3.3, this is equivalent to asking which coupling maximizes the EJC, which we take as our efficiency metric. By design, the GCRN coupling optimizes an asymptotic form of the EJC over the class $\mathcal{P}$ of product couplings, and is therefore asymptotically optimal over this class. In the sequel, we quantify the gap between $\mathcal{P}$ and $\mathcal{M}$ numerically.

Theorem 3.4.1. *Conditionally on* $(\|X_t\|^2, \|Y_t\|^2, X_t^\top Y_t)/d = (x, y, v)$*, it holds that*

$$\lim_{d\to\infty} \sup_{\bar{K}\in\mathcal{P}} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\ell x^{1/2} Z - \ell^2/2} \wedge e^{\ell y^{1/2} Z - \ell^2/2}\right]$$

and this limit supremum is attained by the GCRN coupling.

3.4.2 Scaling limits

Turning to the ODE limit, we need to get a handle on the drift of the process $W^{(d)}$, as well as show that this process does not fluctuate much. In both cases, we work with conditional one-step differences in $(\|X_t\|^2, \|Y_t\|^2, X_t^\top Y_t)$.

Starting with limiting drift of the process $W^{(d)}$, its first two coordinates are coupling-invariant, are individually Markovian, and are dealt with by prior work (Christensen et al., 2005). For the third coordinate, the one-step difference is

$$X_{t+1}^\top Y_{t+1} - X_t^\top Y_t = h Y_t^\top Z_x B_x + h X_t^\top Z_y B_y + h^2 Z_x^\top Z_y B_x B_y. \tag{3.4.1}$$

The first two terms are coupling-invariant; the final term, the *jump concordance*, is coupling-dependent. We evaluate the drift in Proposition 3.4.2 below. Finally, we show

that the fluctuations of $W^{(d)}$ are negligible in Proposition 3.4.3 below.

As a technical point, we fix $\bar{x}, \bar{y} > 0$ and define the set $\mathcal{S} = \{(x, y, v) \mid x \in [0, \bar{x}], y \in [0, \bar{y}], |v| \leq \sqrt{xy}\}$. This is an arbitrarily large compact subset of $\bar{\mathcal{S}}$, the set of all feasible values of the process $W^{(d)}$. Our auxiliary results essentially cover all of $\bar{\mathcal{S}}$ as they hold uniformly over $\mathcal{S}$ for any fixed $\bar{x}, \bar{y}$. This final detail is important, but for brevity we suppress it from notation.

Proposition 3.4.2. *Under the couplings of Section 3.3.3 and uniformly over $w = (x, y, v) \in \mathcal{S}$, it holds that*

$$\lim_{d\to\infty} \mathbb{E}\left[d\big(W^{(d)}_{(t+1)/d} - W^{(d)}_{t/d}\big) \mid W^{(d)}_{t/d} = w\right] = c_\ell(w) = \big(a_\ell(x), a_\ell(y), b_\ell(x, y, v)\big),$$

where

$$\begin{aligned} a_\ell(x) &= \ell^2(1-2x)e^{\ell^2(x-1)/2}\Phi\left(\frac{\ell}{2x^{1/2}} - \ell x^{1/2}\right) + \ell^2\Phi\left(-\frac{\ell}{2x^{1/2}}\right), \\ b_\ell(x, y, v) &= \ell^2\mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}\left[1 \wedge e^{\ell x^{1/2}Z_1 - \ell^2/2} \wedge e^{\ell y^{1/2}Z_2 - \ell^2/2}\right] \\ &\quad - \ell^2 v\left[e^{\ell^2(x-1)/2}\Phi\left(\frac{\ell}{2x^{1/2}} - \ell x^{1/2}\right) + e^{\ell^2(y-1)/2}\Phi\Big(\frac{\ell}{2y^{1/2}} - \ell y^{1/2}\Big)\right], \end{aligned}$$

and where $\rho = \rho(x, y, v)$ is coupling-specific:

$$\rho_{\mathrm{crn}} = \frac{v}{(xy)^{1/2}}, \quad \rho_{\mathrm{refl}} = \frac{2xy - (x+y)v}{(xy)^{1/2}(x+y-2v)}, \quad \rho_{\mathrm{gcrn}} = 1.$$

Proposition 3.4.3. *Under the couplings of Section 3.3.3, it holds that*

$$\lim_{d\to\infty} \sup_{w\in\mathcal{S}} \mathbb{E}\left[d^2\|W^{(d)}_{(t+1)/d} - W^{(d)}_{t/d}\|^2 \mid W^{(d)}_{t/d} = w\right] < \infty.$$

Having obtained the limiting drift of the process $W^{(d)}$, as well as bounded its fluctuations, we can state our main result: the convergence of this process to a deterministic limit.

Theorem 3.4.4. *Let $W_0^{(d)} = w_0 \in \bar{\mathcal{S}}$ for all d. Then, under the couplings of Section 3.3.3, it holds that*

$$W^{(d)} \implies w \quad as \quad d \to \infty,$$

where $w : [0, \infty) \to \bar{\mathcal{S}}$ is the solution of the initial value problem

$$\mathrm{d}w(t) = c_\ell(w(t))\mathrm{d}t \quad started\ from\ w(0) = w_0, \tag{3.4.2}$$

and where the drift $c_\ell(\cdot)$ is coupling-specific as in Proposition 3.4.2.

The analysis of the process $(X_t, Y_t)_{t\geq 0}$ therefore reduces to the analysis of the ODE. We are interested in the squared distance $\|X_t - Y_t\|^2$: a change of variables leads us to an analogous ODE limit $\bar{W}^{(d)} \implies \bar{w}$ for the process

$$\left(\bar{W}_t^{(d)}\right)_{t\geq 0} = \frac{1}{d}\left(\|X_{\lfloor td\rfloor}\|^2, \|Y_{\lfloor td\rfloor}\|^2, \|X_{\lfloor td\rfloor} - Y_{\lfloor td\rfloor}\|^2\right)_{t\geq 0}$$

to the solution $\bar{w} = (x, y, s)$ of $\mathrm{d}\bar{w}(t) = \bar{c}_\ell(\bar{w}(t))\mathrm{d}t$, where $\bar{c}_\ell(\bar{w}) = (a_\ell(x), a_\ell(y), \bar{b}_\ell(x, y, s))$ and $\bar{b}_\ell(x, y, s) = a_\ell(x) + a_\ell(y) - 2b_\ell(x, y, (x + y - s)/2)$. ($a_\ell(\cdot)$ and $b_\ell(\cdot)$ are defined in Proposition 3.4.2.)

The intuition from this result is that after solving the ODE we obtain a function $s(t)$, the solution for the third component, which contracts to 0 as $t \to \infty$ for the GCRN and reflection couplings but does not contract to 0 for the CRN coupling, and is such that

$$\|X_t - Y_t\|^2 \approx s\,(t/d)\,d \quad \text{for large } d.$$

Over one step, the squared-distance changes by roughly $\bar{b}_\ell(\cdot)/d$; the smaller this value, the more a coupling contracts the chains. Since the contraction efficiency measure EJC appears in the limiting drift $\bar{b}_\ell(\cdot)$, by Theorem 3.4.1 the GCRN coupling optimizes this drift point-wise over all couplings in $\mathcal{P}$.

Numerical illustration

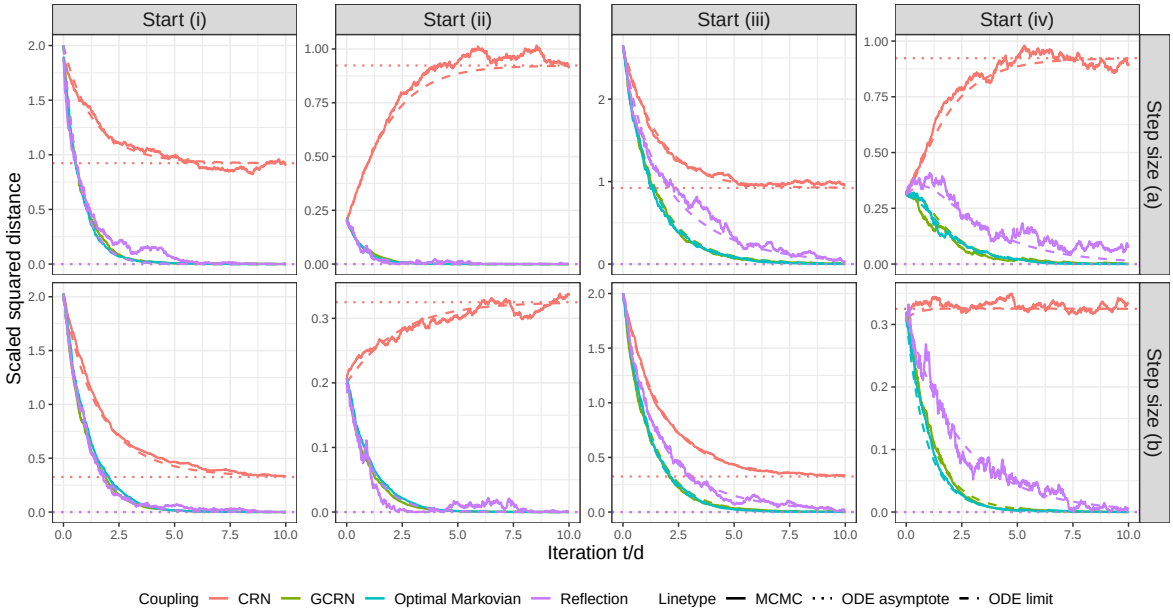

Figure 3.4.1: Trace of the scaled squared distance $\|X_t - Y_t\|^2/d$ and its ODE limit, for a target $\pi^{(d)} = \mathcal{N}_d(0_d, I_d)$ and various couplings, step sizes, and starting conditions as in Section 3.4.2.

We visualize the ODE limits in Figure 3.4.1, considering four different starting conditions (x_0, y_0, ρ_0), where (x_0, y_0) correspond to $(\|X_0\|^2/d, \|Y_0\|^2/d)$ and ρ_0 corresponds to the cosine similarity $X_0^\top Y_0/(\|X_0\|\|Y_0\|)$ between the initial states of the chains. The starting conditions are, and in a limiting sense correspond to chains initialized from:

(i) $(x_0, y_0, \rho_0) = (1, 1, 0)$, independent draws from the target.

(ii) $(x_0, y_0, \rho_0) = (1, 1, 0.9)$, positively correlated draws from the target.

(iii) $(x_0, y_0, \rho_0) = (1.5, 0.5, 0)$, independent draws from, respectively, over- and under-dispersed versions of the target.

(iv) $(x_0, y_0, \rho_0) = (0.4, 0.01, -0.5)$, negatively correlated draws from under-dispersed versions of the target.

We also compare ODE limits with MCMC traces in dimension $d = 1{,}000$. Within each scenario, the same starting value (X_0, Y_0) is used for all couplings; its coordinates

are sampled independently from $\mathrm{BvN}(x_0, y_0; (x_0 y_0)^{1/2}\rho_0)$, the bivariate normal with coordinate-wise variances (x_0, y_0) and correlation ρ_0. We use step sizes (a) $\ell = 2.38$ and (b) $\ell = \sqrt{2}$; the former is optimal at stationarity (Roberts et al., 1997) and both are close to optimal in the transient phase (see Christensen et al., 2005 and Section 3.4.2).

Figure 3.4.1 confirms that our theory consistently predicts the behaviour of the coupled RWM algorithms in high dimensions d. Although stochasticity is clearly present in the traces, this noise will vanish in the limit as $d \to \infty$. The GCRN coupling outperforms the CRN and reflection couplings. Furthermore, GCRN very closely approximates the asymptotically optimal Markovian coupling of Appendix A.2.1; for this reason, we choose to focus on GCRN, although we will revisit the optimal Markovian coupling in the numerical illustrations of Section 3.5.

Long-time behaviour

We turn to the behaviour of the chains in the joint long-time and high-dimensional limits. We shall access this through the unique stable fixed point of the limiting ODE (3.4.2); all fixed points of this process are characterized in Proposition 3.4.5 below.

The first two coordinates of the ODE dictate the marginal behaviour of the chains and are autonomous. Their fixed points correspond to the chains being marginally stationary, are stable, and are $(x^*, y^*) = (1, 1)$. The third coordinate is more involved: we require the function

$$h_\ell(\rho) = \mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}\left[1 \wedge e^{\ell Z_1 - \ell^2/2} \wedge e^{\ell Z_2 - \ell^2/2}\right],$$

which is increasing on its domain $[-1, 1]$, is bounded above by $h_\ell(1) = 2\Phi(-\ell/2)$, and has unbounded derivative as $\rho \to 1$. (See in Lemma A.4.9 in Appendix A.4.2 for properties of h_ℓ.) Proposition 3.4.2 indicates that the fixed points v^* are the solutions of $h_\ell(\rho(v)) - v h_\ell(1) = 0$, where the correlation $\rho(\cdot)$ is coupling-specific: $\rho_{\mathrm{crn}}(v) = v$

and $\rho_{\text{refl}}(v) = \rho_{\text{gcrn}}(v) = 1$. The terms of the fixed-point equation have straightforward interpretations: v is the cosine similarity between the coupled states, $h_\ell(1) = 2\Phi(-\ell/2)$ is the acceptance rate of the marginal chains, and $h_\ell(\rho(v))$ is the rate at which the chains accept their proposals simultaneously.

A scalable coupling must ensure the stability of the fixed point $v^* = 1$. However, due to the rapid growth of $h_\ell(\rho)$ near $\rho = 1$, the fixed point $v^* = 1$ is highly sensitive to the function $\rho(\cdot)$: stability can essentially only be achieved if $\rho(v) = 1$ in an interval around $v = 1$. As we discuss in Section 3.6.1, this instability is caused by the distance between the coupled chains increasing by a relatively large amount when one chain accepts its proposal while the other rejects its proposal, and points to the need to use couplings like GCRN which attempt to synchronize acceptance events. As it happens, a spherically symmetric target also allows the reflection coupling to synchronize acceptance events: intuitively, $\rho_{\text{refl}}(v) = 1$ here because, when the chains are on the same level set, the reflection is a perfect mapping between the respective logarithmic gradients.

Proposition 3.4.5. *Under the couplings of Section 3.3.3, the fixed points of* (3.4.2) *are of the form* $w^* = (1, 1, v^*)$*, where* v^* *is coupling-specific:*

- **CRN***:* $v^*_{\text{crn}} \in (0, 1)$*, stable and* $v^*_u = 1$*, unstable.*
- **Reflection***:* $v^*_{\text{refl}} = 1$*, stable.*
- **GCRN***:* $v^*_{\text{gcrn}} = 1$*, stable.*

We are interested in the limiting squared distance. Proposition 3.4.5 suggests that this equals, for the coupling-specific stable value v^*_{coup},

$$\lim_{d,t\to\infty} \|X_t - Y_t\|^2/d =: s^*_{\text{coup}} = 2(1 - v^*_{\text{coup}}).$$

For CRN, we plot $s^*_{\text{crn}}(\ell)$ in Figure 3.5.1 below and conclude that $\|X_t - Y_t\|^2 = \Theta(d)$ for large t unless $\ell = o_d(1)$. As the dimension increases, the CRN coupling becomes

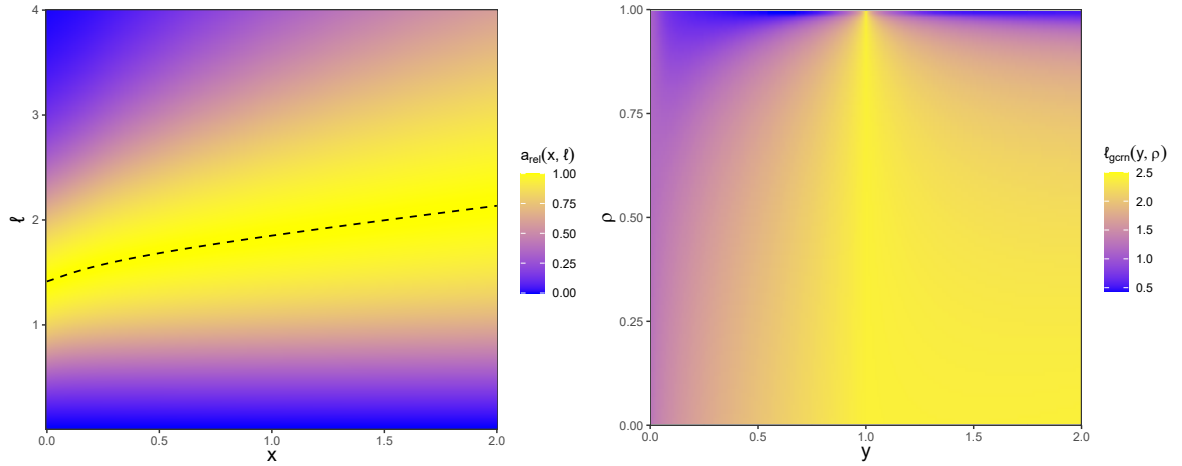

Figure 3.4.2: Optimal step size scalings for marginal rate of convergence and for contraction, as in Section 3.4.2. **Left:** Heatmap of relative drift $a_{\text{rel}}(x, \ell) = |a_\ell(x)|/\max_\ell\{|a_\ell(x)|\}$; the dashed line traces the optimum point-wise over x. **Right:** Heatmap of optimal step size $\ell_{\text{gcrn}}(y, \rho)$ for the GCRN coupling, fixing the X-chain stationary with $x = 1$.

increasingly impractical: to achieve sufficient contraction, it must sacrifice mixing by using an increasingly smaller scaling.

For GCRN and reflection, since $s^*_{\text{gcrn}} = s^*_{\text{refl}} = 0$ we conclude that $\|X_t - Y_t\|^2 = o(d)$ for large t. However, numerically testing dimensions $d \in [1, 10000]$, we found that the contraction was in fact significantly better. (Recall that, by (3.3.2), contraction to $\mathcal{O}(h^2)$ is required for a reasonable chance of coalescence.) The GCRN coupling exhibited clear $\mathcal{O}(h^2)$ behaviour in all dimensions; remarkably, for $d \geq 4$ the coupling was always able to contract the chains to within numerical precision. The reflection coupling displayed $\mathcal{O}(1)$ behaviour, with plots suggesting that it could bring the chains sufficiently close together for a reasonable chance of coalescence when $d \leq 100$, but that coalescence would be unlikely in dimensions at least an order of magnitude larger.

Optimal scaling

To optimize the rate of convergence of a single RWM chain towards the main target mass, the step size ℓ should maximize the absolute drift $|a_\ell(x)|$ point-wise over x. We plot this quantity normalized to unit scale in Figure 3.4.2 (left). Echoing Christensen

et al. (2005), the speed of convergence is insensitive to the choice of step size and no single step size is uniformly optimal, though the rough trend is that smaller step sizes should be chosen for convergence than for mixing. We single out two step sizes: (a) $\ell = 2.38$, which is optimal for mixing at stationarity, achieves at least 45% efficiency over the considered range; (b) $\ell = \sqrt{2}$ is optimal for convergence when the chain is at the mode and is at least 86% efficient over the considered range. The acceptance rate is remarkably stable at the point-wise optimal step size (the dashed line), and hovers around 35%.

We next turn to the problem of selecting a step size ℓ which optimizes the contraction of the GCRN coupling. In the ODE limit, we should optimize the drift $\bar{b}_\ell(x, y, s)$ point-wise. To obtain sensible guidelines, we fix $x = 1$: this corresponds to a stationary chain coupled with a non-stationary chain and emulates a lagged coupling with a large lag parameter. We reparametrize the state to $(x, y, s) = (1, y, 1 + y - 2y^{1/2}\rho)$, where $\rho \in [-1, 1]$ denotes cosine similarity, and we plot the optimal step size ℓ_{gcrn} as a function of (y, ρ) in Figure 3.4.2 (right).[1] Remarkably, the step size $\ell = 2.38$ is close to optimal over much of the range, and in particular when the coupled chains are marginally stationary. We also find (not pictured) that the contraction is insensitive to varying the step size near the optimum. This suggests close to optimal performance for GCRN when the step size is tuned to $\ell = 2.38$ and the acceptance rate to 23.4%, though we caution that in practice smaller scalings and higher acceptance rates may be required because it is crucial for a coupling to synchronize acceptance events between the coupled chains (see Section 3.6.1). Similar guidelines apply to the reflection coupling, but as we shall see in Section 3.5 this coupling requires a much smaller scaling $\ell = o_d(1)$ when the target is not spherically symmetric.

The different optimal scalings may seem at odds; the discrepancy can be explained as follows. The rate of contraction of two coupled chains depends on their movement

[1] The heatmap for $\rho \in [-1, 0)$ is not pictured, as it resembles the lower half of the plot.

in both the angular and the radial directions. It therefore benefits from better mixing in the angular direction, which for a single stationary chain is optimized by the scaling $\ell = 2.38$. In contrast, the rate of convergence of a single chain towards the main target mass only depends on the movement in the radial direction, which benefits from a smaller scaling.

3.5 Analysis: elliptical Gaussian case

While we do not see the asymptotic Gaussianity assumption as a major limitation to our theoretical results, the spherical symmetry assumed in Section 3.4 is certainly unrealistic: in practice, even after preconditioning, one cannot expect the RWM proposals to precisely match the structure of the target. In this section, we therefore relax this constraint and consider a more general sequence of targets $\pi^{(d)} = \mathcal{N}_d(0_d, \Sigma_d)$ of increasing dimension d. Given the form of the acceptance ratio, it will be convenient to work with the precision matrix $\Omega_d = \Sigma_d^{-1}$. To obtain a transparent asymptotic theory, we fix a positive-valued distribution μ which captures heterogeneity across the lengthscales of the target and assume that

$$\forall d: \quad \Omega_d = \mathrm{diag}(\omega_1^2, \ldots, \omega_d^2), \text{ where } (\omega_i^2)_{i \geq 1} \overset{\text{iid}}{\sim} \mu. \tag{3.5.1}$$

We impose moment conditions on the *spectral distribution* μ of the precision matrix, see Assumptions 3.5.1 and 3.5.3. We first show in Theorem 3.5.2 that the GCRN coupling is asymptotically optimal for contraction within the class $\mathcal{P}$. By distilling the argument of Section 3.4 down to its key part, we then consider a form of limiting drift in Proposition 3.5.5, which tells us about the long-time behaviour of the coupled chains and informs efficient step size scalings. To preempt our main conclusions, we find that the GCRN coupling is robust to the eccentricity of the target and that the same acceptance rate tuning guidelines apply as in the spherical case. The reflection

coupling however breaks down for eccentric targets and must sacrifice mixing in order to achieve sufficient contraction. Under stronger structural assumptions, we could arrive at ODE limits, see Remark 3.5.6.

It will be convenient to define (for all $x, y \in \mathbb{R}^d$ and $k \in \mathbb{Z}$) the inner product $\langle x, y\rangle_{[k]} = x^\top \Omega_d^k y$ and its associated norm $\|x\|_{[k]}^2 = \langle x, x\rangle_{[k]}$. The acceptance step becomes

$$B_x = \mathbb{1}\big\{\log U_x \leq -h\langle Z_x, X_t\rangle_{[1]} - (h^2/2)\|Z_x\|_{[1]}^2\big\}.$$

3.5.1 Asymptotic optimality

We first show that the GCRN coupling is asymptotically optimal for contraction among the class $\mathcal{P}$ of all product couplings, for which we require a law of large numbers assumption.

Assumption 3.5.1: The spectral distribution μ has finite first moment $z_1 = \mathbb{E}[\omega_i^2]$.

Theorem 3.5.2. *Under Assumption 3.5.1 and conditionally on* $(\|X_t\|_{[2]}^2, \|Y_t\|_{[2]}^2, \langle X_t, Y_t\rangle_{[2]})/(z_1 d) = (x_2, y_2, v_2)$*, it holds that*

$$\lim_{d\to\infty} \sup_{\bar{K}\in\mathcal{P}} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\lambda x_2^{1/2} Z - \lambda^2/2} \wedge e^{\lambda y_2^{1/2} Z - \lambda^2/2}\right],$$

where $\lambda = \ell z_1^{1/2}$*. Furthermore, the limit supremum is attained by the GCRN coupling.*

Roughly speaking, Assumption 3.5.1 states that none of the lengthscales of the target are much smaller than the average. Regularity conditions like these are among the weakest required by optimal scaling theory (Sherlock, 2013, Condition 1) and appear necessary to avoid degeneracies in the efficiency of the RWM algorithm (Beskos et al., 2018). Under Assumption 3.5.1, one should think of

$$\lambda = \ell z_1^{1/2}$$

as the natural step size parameter, with $\lambda = 2.38$ being optimal at stationarity and corresponding to a 23.4% acceptance rate. (We further relate our notation to the optimal scaling literature in Section 3.6.2.) The key consequence of Assumption 3.5.1 is that the Hessian term in the acceptance ratio tends to a constant: $\lim_{d\to\infty} h^2 \|Z_x\|^2_{[1]} = \ell^2 z_1$ in probability. It is this which enables us to prove Theorem 3.5.2. The quantities in Theorem 3.5.2 are rescaled so that e.g. $x_2 \approx 1$ when the chain is in the main target mass; we will apply a similar scaling to $\|X_t\|^2_{[k]}$ in the sequel.

3.5.2 Scaling limits

We turn to the problem of characterizing the high-dimensional behaviour of the coupled RWM chains. To be able to handle all couplings of Section 3.3.3 at once, we impose somewhat stronger assumptions.

Assumption 3.5.3: The spectral distribution μ has finite k-th moment for $k \in \{-2, 1\}$.

Assumption 3.5.3 guarantees the existence of all intermediate moments $z_k = \mathbb{E}[\omega_i^{2k}]$ for all $k \in [-2, 1]$. Intuitively, Assumption 3.5.3 asks for none of the lengthscales of the target to be much larger or smaller than the average. In practice, the target is fixed and not randomly generated as in (3.5.1), but it may still arise as a discretization of a limiting target with a given limiting spectral distribution. We exemplify a class of time series models whose limiting spectral distribution satisfies the assumption.

Example 3.5.4: Fix $p \in \mathbb{N}$. For all $d \geq p$, let $\pi^{(d)}$ be the distribution of a stationary AR(p) process of length d, with the same parameters across all dimensions d. Then, the spectrum of the precision matrix converges to a spectral distribution μ which satisfies Assumption 3.5.3.

In Example 3.5.4, the assumption is satisfied because the limiting target is well-conditioned, however we stress that a uniform control of the condition number is certainly not necessary. More broadly, we expect Assumption 3.5.3 to hold for Gaussian

Markov random fields (Rue and Held, 2005) whose dependence structure grows suitably slowly with the dimension d.

The analysis of the reflection coupling will require the following quantity, which we dub the *limiting eccentricity* of the target

$$\varepsilon = z_1 z_{-1} = \lim_{d\to\infty} \operatorname{Tr}(\Omega_d)\operatorname{Tr}(\Sigma_d)/d^2.$$

This quantity satisfies $\varepsilon \geq 1$ and is invariant to a rescaling of the precision matrix by a constant factor. The lower bound is attained when the target is spherical; the greater the imbalance between the average lengthscales of the precision and covariance matrices, the larger ε becomes.

Limiting drift

We now compute the limiting drift of triplets $W_{[k]} = (\|X_t\|^2_{[k]}, \|Y_t\|^2_{[k]}, \langle X_t, Y_t\rangle_{[k]})$ for various k. For this calculation, we let $\bar{\mathbb{E}}$ denote the expectation conditional on $W_{[j]}/(z_{j-1}d) = (x_j, y_j, v_j)$ for all $j \in \{-1, 0, 1, 2\}$, where the normalizing constants are defined subsequent to Assumption 3.5.3.

Proposition 3.5.5. *Under Assumption 3.5.3 and the couplings of Section 3.3.3, for all $k \in \{-1, 0, 1\}$ it holds that*

$$\begin{aligned}
\lim_{d\to\infty} \bar{\mathbb{E}}\left[\|X_{t+1}\|^2_{[k]} - \|X_t\|^2_{[k]}\right] &= \ell_k^2 \alpha(x_{k+1}; x_2), \\
\lim_{d\to\infty} \bar{\mathbb{E}}\left[\|Y_{t+1}\|^2_{[k]} - \|Y_t\|^2_{[k]}\right] &= \ell_k^2 \alpha(y_{k+1}; y_2), \\
\lim_{d\to\infty} \bar{\mathbb{E}}\left[\langle X_{t+1}, Y_{t+1}\rangle_{[k]} - \langle X_t, Y_t\rangle_{[k]}\right] &= \ell_k^2 \beta(v_{k+1}; x_2, y_2, \rho),
\end{aligned} \tag{3.5.2}$$

where $\ell_k = \ell z_k^{1/2}$,

$$\alpha(x_{k+1}; x_2) = (1 - 2x_{k+1}) e^{\lambda^2(x_2-1)/2} \Phi\left(\frac{\lambda}{2x_2^{1/2}} - \lambda x_2^{1/2}\right) + \Phi\left(-\frac{\lambda}{2x_2^{1/2}}\right),$$

$$\begin{aligned}\beta(v_{k+1}; x_2, y_2, \rho) = \mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)} &\left[1 \wedge e^{\lambda x_1^{1/2} Z_1 - \lambda^2/2} \wedge e^{\lambda x_2^{1/2} Z_2 - \lambda^2/2}\right] \\ &- v_{k+1}\left[e^{\lambda^2(x_2-1)/2}\Phi\left(\frac{\lambda}{2x_2^{1/2}} - \lambda x_2^{1/2}\right) + e^{\lambda^2(y_2-1)/2}\Phi\left(\frac{\lambda}{2y_2^{1/2}} - \lambda y_2^{1/2}\right)\right],\end{aligned}$$

where $\lambda = \ell z_1^{1/2}$, *and where* ρ *is coupling-specific:*

$$\rho_{\mathrm{crn}} = \frac{v_2}{(x_2 y_2)^{1/2}}, \quad \rho_{\mathrm{refl}} = \frac{v_2}{(x_2 y_2)^{1/2}} + \frac{2(x_1 - v_1)(y_1 - v_1)}{\varepsilon (x_2 y_2)^{1/2}(x_0 + y_0 - 2v_0)}, \quad \rho_{\mathrm{gcrn}} = 1.$$

Proposition 3.5.5 generalizes Proposition 3.4.2 to the elliptical case, with the notable differences that the natural step size parameter is now λ and that the correlation ρ_{refl} now depends on the eccentricity ε. By Theorem 3.5.2, the GCRN coupling optimizes point-wise the drift (3.5.2) over all couplings in $\mathcal{P}$. With more care, we could additionally bound the fluctuations of $W_{[k]}$ (as in Proposition 3.4.3) and we could make our results uniform over compact sets. However, these refinements will not provide any additional insight into the asymptotic behaviour of the coupled chains, so we avoid further technicalities.

Through a change of variables, we can understand the evolution of the squared Mahalanobis distance,

$$\lim_{d\to\infty} \bar{\mathbb{E}}\left[\|X_{t+1} - Y_{t+1}\|_{[k]}^2 - \|X_t - Y_t\|_{[k]}^2\right] = \ell_k^2 \gamma(x_{k+1}, y_{k+1}, v_{k+1}; x_2, y_2, \rho),$$

where $\gamma(x_{k+1}, y_{k+1}, v_{k+1}; x_2, y_2, \rho) := \alpha(x_{k+1}; x_2) + \alpha(y_{k+1}; y_2) - 2\beta(v_{k+1}; x_2, y_2, \rho)$.

Remark 3.5.6: Under certain structural assumptions, we can extend Proposition 3.5.5 to an ODE limit. For instance, if $\Sigma = \mathrm{diag}(1, \sigma^2, 1, \sigma^2, \dots)$, by decomposing the process $W^{(d)}$ considered in Section 3.4 into separate triplets for the odd and even coordinates,

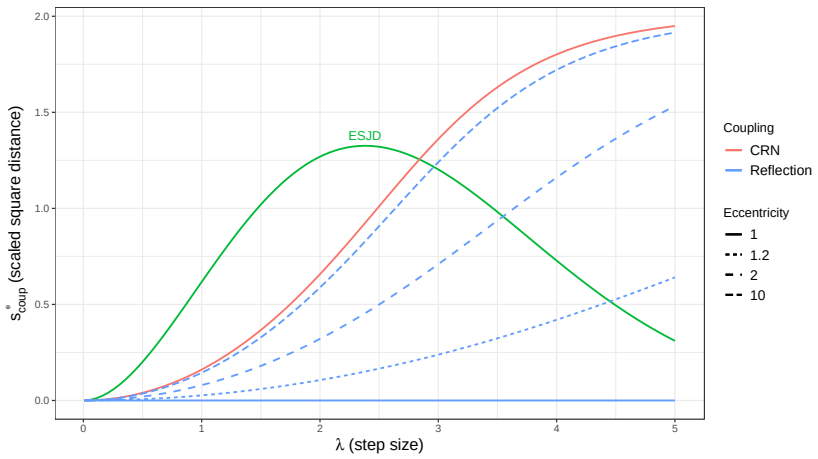

Figure 3.5.1: Scaled squared distance s^*_{coup} in the joint long-time and high-dimensional limits, as in Section 3.5.2, for targets $\pi^{(d)} = \mathcal{N}_d(0_d, \Sigma_d)$ and various couplings and values of the natural step size λ and eccentricity ε. We stress that the only desirable value is $s^*_{\text{coup}} = 0$. The efficiency measure $\text{ESJD}(\lambda)$ is overlaid for context.

we obtain a six-dimensional Markov process. Under moment conditions on σ^2, this Markov process has an ODE limit as the dimension grows.

Long-time behaviour

We infer the behaviour of the coupled chains in the joint long-time and high-dimensional limits from the stable fixed points of the system of equations (3.5.2). This is sensible, as the unconditional expected one-step change in a function of the coupled chains is null when coupled process is stationary.

This analysis mirrors, and assumes familiarity with, the discussion of Section 3.4.2. The fixed points are the solutions of $\alpha(x_k; x_2) = \alpha(y_k; y_2) = \beta(v_k; 1, 1, \rho) = 0$ for all $k \in \{0, 1, 2\}$. The solutions $x^*_k = y^*_k = 1$ correspond to marginal stationarity in each chain. This implies that the solutions $v^*_k = v^*$ must be constant in k, which collapses the fixed-point equations for the remaining joint behaviour of the chains to

$$h_\lambda(\rho(v)) - v h_\lambda(1) = 0, \tag{3.5.3}$$

where the correlation $\rho(\cdot)$ is coupling-specific: $\rho_{\text{crn}}(v) = v$, $\rho_{\text{refl}}(v) = v + \varepsilon^{-1}(1 - v)$

and $\rho_{\text{gcrn}}(v) = 1$. The terms of the fixed-point equation (3.5.3) retain their intuitive interpretations from Section 3.4.2 in terms of the cosine similarity v, the marginal acceptance rate $h_\lambda(1)$, and the synchronous acceptance rate $h_\lambda(\rho(v))$. We consider a linear stability analysis in Proposition 3.5.7; for the fixed point $v^* = 1$ to be stable, we essentially require that $\rho(v) = 1$ in an interval around $v = 1$. In other words, we require marginally stationary chains to accept their proposals simultaneously even when they are not coalesced. The reflection coupling is unable to do so in the elliptical case, which ultimately impacts its scalability.

Proposition 3.5.7. *Let the target be non-spherical Gaussian (i.e. $\varepsilon > 1$). Under the couplings of Section 3.3.3, the solutions of the fixed-point equation* (3.5.3) *are as follows:*

- **CRN***: $v^*_{\text{crn}} \in (0, 1)$, stable and $v^*_u = 1$, unstable.*
- **Reflection***: $v^*_{\text{refl}} \in (v^*_{\text{crn}}, 1)$, stable and $v^*_u = 1$, unstable. As a function of $\varepsilon \in (1, \infty)$, $v^*_{\text{refl}}(\varepsilon)$ is decreasing and tends to $\{1, v^*_{\text{crn}}\}$ at the extremes.*
- **GCRN***: $v^*_{\text{gcrn}} = 1$, stable.*

We are interested in the limiting behaviour of the squared distance. Proposition 3.5.7 suggests that, for the coupling-specific stable value v^*_{coup},

$$\lim_{d,t\to\infty} \|X_t - Y_t\|^2 / \operatorname{Tr}(\Sigma_d) =: s^*_{\text{coup}} = 2(1 - v^*_{\text{coup}}).$$

We plot this limiting quantity in Figure 3.5.1. For reflection and CRN, since $s^*_{\text{refl}}, s^*_{\text{crn}} > 0$ we conclude that $\|X_t - Y_t\|^2 = \Theta(d)$ for large t. Both couplings become increasingly impractical as the dimension increases: they must sacrifice mixing for contraction by using an increasingly smaller scaling $\lambda = o_d(1)$. This points to the need to use pre-conditioning alongside the reflection coupling, as this coupling performs adequately for spherical targets. The CRN coupling is unaffected by the eccentricity of the target but

is uniformly worse than the reflection coupling.

For GCRN, since $s^*_{\text{gcrn}} = 0$ we conclude that $\|X_t - Y_t\|^2 = o(d)$ for large t. However, experimentally (we tested dimensions $d \in [1, 10000]$ and targets similar to those of Figure 3.5.2) we found that the GCRN coupling performed significantly better. We observed $\mathcal{O}(h^2)$ behaviour in all dimensions; for $d \geq 10$, the GCRN coupling was consistently able to contract the chains to within numerical precision, even for the most eccentric target considered.

In practice, given estimates of the traces of the covariance and precision matrices, we can estimate the eccentricity ε, compute v^*_{coup} by solving (3.5.3) numerically, and then predict

$$\mathbb{E}\left[\|X_t - Y_t\|^2\right] \approx 2\,\mathrm{Tr}(\Sigma_d)(1 - v^*_{\text{coup}}) \quad \text{for large } d, t. \tag{3.5.4}$$

Optimal scaling

We turn to the problem of step size tuning for the GCRN coupling. When $k = 1$, by comparison with Proposition 3.4.2, we obtain that

$$\lim_{d\to\infty} \bar{\mathbb{E}}\left[\|X_{t+1} - Y_{t+1}\|^2_{[1]} - \|X_t - Y_t\|^2_{[1]}\right] = \bar{b}_\lambda(x_2, y_2, s_2),$$

where $\bar{b}_\lambda(\cdot)$ is the drift of the squared-distance process which we obtained for a spherical Gaussian target, which now depends on the natural step size parameter λ. To optimize the contraction of the GCRN coupling, we should therefore optimize the drift $\bar{b}_\lambda(\cdot)$ point-wise. This problem was considered in Section 3.4.2: we reiterate that (i) the contraction of the chains is insensitive to the scaling, (ii) the scaling $\lambda = 2.38$ and the acceptance rate of 23.4% are close to optimal for an elliptical Gaussian target and (iii) a somewhat smaller step size may be necessary in practice, as it is crucial for the coupling to ensure that the chains accept their proposals at the same time, see Section 3.6.1.

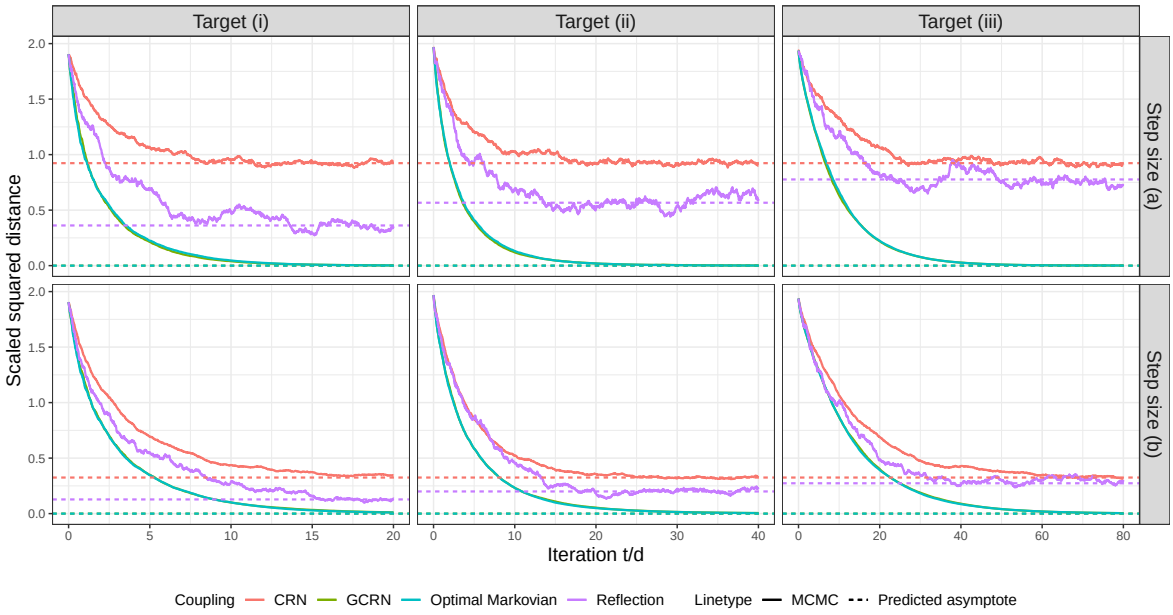

Figure 3.5.2: Trace of the scaled squared distance $\|X_t - Y_t\|^2 / \operatorname{Tr}(\Sigma_d)$ and predicted long-time asymptote, for various targets $\pi^{(d)} = \mathcal{N}_d(0_d, \Sigma_d)$, couplings, and step sizes as in Section 3.5.3.

3.5.3 Numerical illustration

We verify our findings with three targets $\pi^{(d)} = \mathcal{N}_d(0_d, \Sigma)$ of increasing limiting eccentricity ε:

(i) $\Sigma_{ij} = 0.5^{|i-j|}$ for all (i, j): $\varepsilon = 5/3$.

(ii) $\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2)$ where $(\sigma_i^2)_{i \geq 1} \overset{\text{iid}}{\sim} \chi_3^2$: $\varepsilon = 3$.

(iii) $\Sigma = \operatorname{diag}(1, 24, 1, 24, \dots)$: $\varepsilon = 25/4$.

Target (i) is an AR(1) process with autocorrelation 0.5 and unit volatility and corresponds to Example 3.5.4. Target (ii) directly corresponds to Assumption 3.5.3. Target (iii) corresponds to Remark 3.5.6, where an ODE limit can be shown. We consider the same two step size scalings in the natural parameter (a) $\lambda = 2.38$ and (b) $\lambda = \sqrt{2}$ as in Section 3.4.2, as the former is optimal for the stationary phase and we expect the latter to perform well in the transient phase. We fix $d = 2{,}000$ and we always start the coupled chains independently from the target.

Figure 3.5.2 shows that the MCMC traces resemble deterministic trajectories, hinting towards an ODE limit. For all couplings, the long-time behaviour of the chains is as predicted by our theory, with GCRN performing well and the two baselines being impractical. Remarkably, GCRN performs identically to the implementable asymptotically optimal Markovian coupling of Appendix A.2.1: here, as we show for a more general class of targets in Section 3.6.2, it is because the GCRN coupling itself is asymptotically optimal Markovian when the chains are marginally stationary. As all evidence in the main text and in Appendix A.2.1 indicates that GCRN is very nearly as efficient as the asymptotically optimal Markovian coupling, we favour parsimony and hereafter focus on GCRN.

3.6 From theory to practice

In this section, we turn our scaling limit analysis into practical recommendations. Investigating the one-step dynamics of the coupled chains, we expose the necessity for a scalable coupling to synchronize acceptance events between chains. This prompts a study into what conditions are required for the GCRN to work well in practice, as well as a hybrid between the GCRN and reflection couplings, which we expect will coalesce the chains quickly even for eccentric targets.

3.6.1 Necessity of synchronizing acceptance events

Let us interpret the one-step dynamics of coupled RWM chains with step size $h = \mathcal{O}(d^{-1/2})$. We have that

$$\|X_{t+1}-Y_{t+1}\|^2-\|X_t-Y_t\|^2 = 2h(X_t-Y_t)^\top(Z_xB_x-Z_yB_y)+h^2\|Z_xB_x-Z_yB_y\|^2. \quad (3.6.1)$$

One might think of the first term of (3.6.1) as representing the *contractive* part of the dynamics. Its expectation conditional on (X_t, Y_t) is invariant to the coupling; our

scaling limits (see e.g. the proof of Proposition 3.5.5) indicate that this expectation is roughly $h^2(X_t - Y_t)^\top(\nabla \log \pi(X_t) - \nabla \log \pi(Y_t))$ times the acceptance rate, at least when the target is sufficiently regular and each marginal chain is close enough to stationarity. We therefore expect the contraction to be roughly exponential at rate $\mathcal{O}(h^2) = \mathcal{O}(d^{-1})$, particularly if the target is log-concave.

The second term of (3.6.1) represents the *expansive* part of the dynamics. In particular, a linear $\Theta(1)$ increase is suddenly incurred whenever there is an imbalanced acceptance event, that is acceptance in one chain at the same time as a rejection in the other. This highlights the necessity to use a coupling that synchronizes acceptance events with high probability: for instance, if imbalanced acceptance events occur with $\Theta(1)$ probability, then the equilibrium between the contractive and expansive parts of the dynamics lies at $\|X_t - Y_t\|^2 = \Theta(d)$, so the coupling cannot scale. Our fixed-point results (Propositions 3.4.5 and 3.5.7) validate this behaviour.

To us, at least, the only principled way of synchronizing acceptance events is by paying careful attention to the acceptance steps, e.g. by exploiting additional local information about the target density, such as its logarithmic gradient as in the GCRN coupling. At the same time, our optimality results concerning GCRN suggest that, in order to construct scalable couplings, judicious inclusion of gradient information is to some extent sufficient. In the next section, we therefore gauge the extent to which the GCRN coupling might perform well in practice.

3.6.2 When does GCRN work and when does it not?

To gain insight into what conditions are required for the GCRN coupling to scale well, we consider a general product-target setting with variable lengthscales as in Roberts and Rosenthal (2001). To obtain a sensible limit theory, we fix the chains to be marginally stationary; we find that GCRN is asymptotically optimal for contraction among *all Markovian couplings*. We leave extensions to non-stationary chains for further work.

Assumption 3.6.1: Let $\pi^{(d)}(x) = \prod_{i=1}^{d} \omega_i f(\omega_i x)$ with $(\omega_i^2)_{i\geq 1} \overset{\text{iid}}{\sim} \mu$, where μ has finite first moment and where $f : \mathbb{R} \to (0, \infty)$ is twice continuously differentiable with $(\log f)'$ Lipschitz continuous, $\mathbb{E}_{Y\sim f}[(\log f)'(Y)^8] < \infty$ and $\mathbb{E}_{Y\sim f}[(\log f)''(Y)^4] < \infty$.

Theorem 3.6.2. *For all $d \geq 1$, let the target $\pi^{(d)}$ be as in Assumption 3.6.1, and let the joint distribution of (X_t, Y_t) be in $\Gamma(\pi^{(d)}, \pi^{(d)})$, where $\Gamma(\mu, \nu)$ is the set of all couplings of the distributions (μ, ν). Then,*

$$\lim_{d\to\infty} \sup_{\bar{K}\in\mathcal{M}} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = 2\ell^2 \Phi\left(-\ell(bI)^{1/2}/2\right),$$

where $I = \mathbb{E}_{Y\sim f}[(\log f)'(Y)^2]$, $b = \mathbb{E}[\omega_i^2]$, and this supremum is attained by the GCRN coupling.

The optimality of GCRN is therefore not a purely Gaussian phenomenon: it occurs in Theorem 3.6.2 because the variation in the Hessian term of the log-acceptance ratio is, relative to the gradient term, asymptotically negligible. In practice and for a well-tuned RWM algorithm, this condition is approximately satisfied when the logarithmic density of the target does not have any direction in which it varies particularly rapidly (see Sherlock, 2013); equivalently, when none of the lengthscales of the target are much smaller than the average. We expect the GCRN coupling to perform well when this is the case, particularly if the target is also log-concave, with the ideal behaviour of $\|X_t - Y_t\|^2$ under GCRN being that of exponential contraction at rate $\mathcal{O}(h^2)$ to a small steady-state average.

Conversely, GCRN only explicitly accounts for first-order variation in the acceptance ratio so this coupling may perform poorly when the second-order terms in the acceptance ratio exhibit substantial variability. In practice, the issue can appear when the precision matrix of the target has a few very large eigenvalues. Preconditioning may alleviate the issue, as may reducing the step size; the latter reduces the relative variation from second-order terms in the acceptance ratio, increases the acceptance rate,

and improves the contraction by forcing the chains to accept simultaneously more often. One can also harness stochasticity to enhance the contraction of the GCRN coupling, as we show in Section 3.6.3.

Remark 3.6.3: Theorem 3.6.2 suggests good performance in finite dimensions, but it does not guarantee it. To formally establish the rate of contraction and the long-time behaviour under GCRN in any finite dimension would require a careful nonasymptotic analysis of the contractive and expansive terms of (3.6.1), which we do not perform here. Instead, we verify the behaviour of GCRN empirically, and we recall the encouraging results in the Gaussian case (Sections 3.4.2 and 3.5.2).

Remark 3.6.4: In Roberts and Rosenthal (2001), the quantity I of Theorem 3.6.2 is interpreted as a marginal roughness measure, whereas the moment b quantifies the variation across the coordinates of the target. In Assumption 3.5.1, considered in the elliptical Gaussian case, these correspond to $I = 1$ and $b = z_1$.

3.6.3 The GCRefl coupling: combining contraction with stochasticity

In the diffusion literature, it is well-known that different couplings are suited to different purposes (Chen and Li, 1989). When the goal is coalescence, reflection couplings have been seen to be highly effective: in particular, they minimize meeting times in the case of coupled Brownian motions or Ornstein-Uhlenbeck processes (e.g. Connor, 2007, Chapter 3.4). However, we have seen that the reflection coupling of the RWM does not perform well for high-dimensional eccentric targets (Section 3.5) as it is no longer able to synchronize acceptance events (Section 3.6.1).

We therefore propose a hybrid between the GCRN and reflection couplings, designed to combine the favourable properties of both of these. We call it the **GCRefl** (*Gradient-*

Corrected Reflection) coupling:

$$Z_x = Z - (e_x^\top Z)e_x + Z_\nabla e_x,$$
$$Z_y = Z - 2(e^\top Z)e - (e_y^\top Z)e_y + Z_\nabla e_y,$$

where: $e = \mathrm{Nor}(X - Y)$; $e_x = \mathrm{Nor}(n_x - (e^\top n_x)e)$ with $n_x = \mathrm{Nor}(\nabla \log \pi(X_t))$ and similarly for e_y; $Z \sim \mathcal{N}_d(0_d, I_d)$ and $Z_\nabla \sim \mathcal{N}_1(0, 1)$ are independent. We default to the reflection coupling when a vector to be normalized is null (this is always the case in dimension $d = 1$). The proposed coupling starts from the reflection coupling, then applies a GCRN-like correction in order to increase the rate of simultaneous acceptances.

One can roughly interpret the one-step dynamics of $\|X_t - Y_t\|$ under the GCRefl coupling as exponential contraction at rate $\mathcal{O}(h^2)$ together with a random walk with increments $\mathcal{O}(h)$. The exponential contraction dominates when $\|X_t - Y_t\| = \mathcal{O}(d^{1/2})$, whereas the stochasticity dominates when $\|X_t - Y_t\| = \mathcal{O}(1)$. Borrowing intuition from the case of diffusions, this added stochasticity should help drive the chains towards coalescence. Indeed, as we will see in the experiment of Section 3.7.4, the reflection move of GCRefl is particularly helpful in scenarios where GCRN is only able to contract the chains to within $\mathcal{O}(1)$ squared distance.

3.7 Numerical experiments

In this section, we illustrate practical applications of our proposed couplings, in particular the estimation of the rate of convergence and asymptotic variance of the RWM, as well as of the bias of an approximate sampling method. We also use a natural extension to the GCRN coupling to devise an effective coupling for the Hug and Hop algorithm (Ludkin and Sherlock, 2022) and to quantify the rate of convergence of this algorithm. Finally, we perform a comparative study of coupled RWM and MALA kernels, with a view towards unbiased MCMC. Our discrepancies of choice in this section are the total

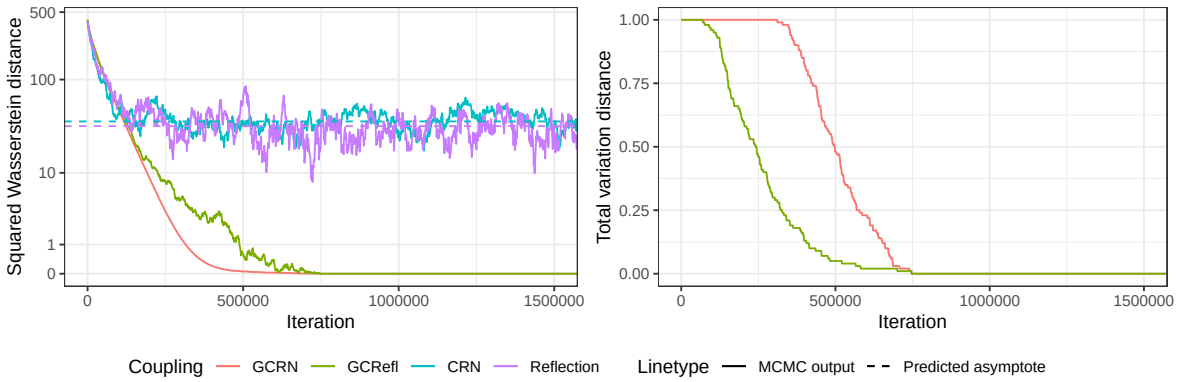

Figure 3.7.1: Rate of convergence of the RWM targeting the SVM, when started from the prior. We estimated upper bounds on $\mathcal{W}_2^2(\pi_t, \pi)$ and $\mathrm{TV}(\pi_t, \pi)$ using only the GCRN and GCRefl couplings. The reflection-maximal and CRN couplings were not sufficiently contractive for this problem, as illustrated by traces of the squared distance $\|X_t - Y_t\|^2$ in the left plot.

variation distance and the squared 2-Wasserstein distance

$$\mathrm{TV}(\mu, \nu) = \inf_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}[\mathbb{1}\{X = Y\}], \quad \mathcal{W}_2^2(\mu, \nu) = \inf_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}\left[\|X - Y\|^2\right].$$

When estimating the rate of convergence, the bound (3.2.1) for the total variation distance simplifies to $\mathrm{TV}(\pi_t, \pi) \leq \mathbb{E}[0 \vee \mathrm{Ceiling}\{(\tau - t)/L\}]$, see Biswas et al. (2019).

Throughout, we have striven to tune the considered algorithms and couplings close to optimally. We defer additional details to Appendix A.3, including: (i) further algorithmic descriptions, (ii) experiments regarding parameter choice, (iii) further discussion on the contractivity of the couplings used, (iv) alternative coupling strategies.

3.7.1 Rate of convergence of the RWM

We illustrate the effectiveness of our proposed couplings at quantifying the rate of convergence of the RWM in a challenging high-dimensional setting. We target the posterior distribution $\pi(x_{1:d} \mid y_{1:d})$ of a stochastic volatility model (SVM; Liu, 2001,

Section 9.6.2):

$$\begin{aligned} y_i \mid x_i &\sim \mathcal{N}_1(0, \beta^2 \exp(x_i)) && \text{for } i \in \{1, \dots, d\}, \\ x_{i+1} \mid x_i &\sim \mathcal{N}_1(\varphi x_i, \sigma^2) && \text{for } i \in \{1, \dots, d-1\}, \\ x_1 &\sim \mathcal{N}_1\left(0, \sigma^2/(1-\varphi^2)\right). \end{aligned}$$

We fix the dimension to $d = 360$, hyperparameters to $(\beta, \varphi, \sigma) = (0.65, 0.98, 0.15)$, and generate the data from the model. We fix the starting distribution to be the prior $\pi_0 = \pi(x_{1:d})$. Our goal is to estimate upper bounds on $\mathrm{TV}(\pi_t, \pi)$ and $\mathcal{W}_2^2(\pi_t, \pi)$ as described in Section 3.2.

Before running MCMC, we compute a Laplace approximation $\hat{\pi} = \mathcal{N}_d(\hat{\mu}, \hat{\Sigma})$. This suggests the step size $h = 2.38/\operatorname{Tr}(\hat{\Sigma}^{-1})^{1/2}$, which we employ and empirically verify as corresponding to a near-optimal acceptance rate of 23%. To predict the long-time behaviour of the chains under the CRN and reflection couplings, we plug $\hat{\pi}$ into Equation (3.5.4).

To estimate the rate of convergence, we require our couplings to coalesce the chains in finite time. The GCRN and GCRefl couplings cannot produce exact meetings on their own; instead, we opt for a two-scale approach, employing a contractive coupling when the chains are at least $\|X_t - Y_t\|^2 \geq \delta$ apart, and otherwise (when $\|X_t - Y_t\|^2 < \delta$) employing the reflection-maximal coupling. Here and in subsequent experiments with two-scale couplings, we select the switching threshold δ using a grid search on a logarithmic scale: the general trend is that the meeting times are insensitive to choosing δ smaller than optimal. We leave a formal investigation into the optimal choice of δ for further work. Our chosen thresholds are $(\delta_{\mathrm{gcrn}}, \delta_{\mathrm{gcrefl}}) = (0.1, 0.001)$ in this experiment, and we use a large lag of $L = 1.5 \times 10^6$ and 100 independent replicates to compute each bound.

The numerical results are displayed in Figure 3.7.1. Both the two-scale GCRN and GCRefl couplings effectively quantify the rate of convergence of the RWM algorithm.

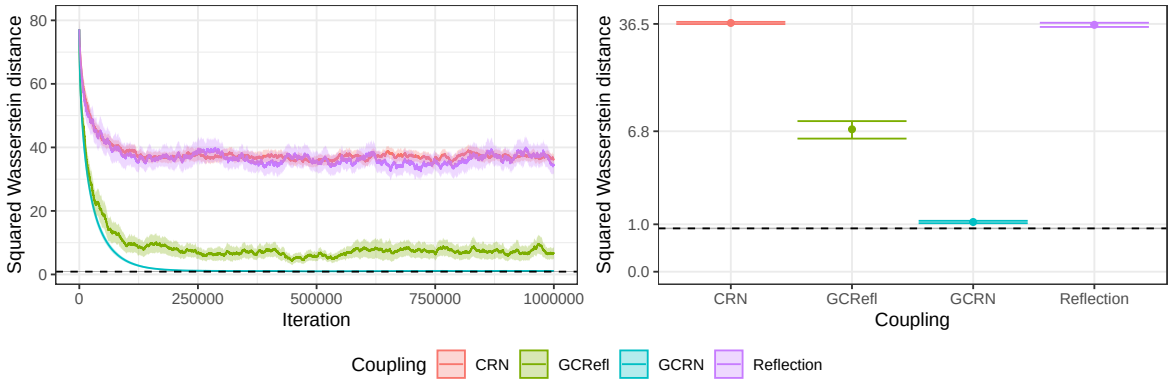

Figure 3.7.2: Bias of the Laplace approximation for the SVM. **Left**: Per-iteration sample average upper bound estimates. **Right**: Point estimates of upper bounds with $(S, T) = (3.5 \times 10^5, 10^6)$. The dashed line is the lower bound of Gelbrich (1990). All estimates are shown with ± 2 standard errors (in either shaded regions or error bars).

GCRN attains the sharper squared Wasserstein distance bound as it is focused on contraction, while GCRefl attains a sharper total variation distance bound as it is focused on coalescence. In contrast, the reflection-maximal coupling is severely hindered by the eccentricity of the target: the coupling is capable of exact meetings, but as the chains stay far apart the probability of coalescing is effectively null at all iterations. The long-time behaviour of the couplings is accurately predicted by our theory.

3.7.2 Bias of approximate sampling

Couplings can also be used to estimate the bias of approximate sampling procedures (Biswas and Mackey, 2024). Motivated by the accuracy of the quantities inferred from the Laplace approximation $\hat{\pi}$ of the SVM target π of Section 3.7.1 (such as the optimal step size h and the expected long-time behaviour under the CRN and reflection couplings), we use our proposed couplings to compute upper bounds on $\mathcal{W}_2^2(\hat{\pi}, \pi)$.

The upper bounds are computed with a small modification to the method of Section 3.2. We set the lag to be $L = 0$, target the exact distribution π with the X-chain and its approximation $\hat{\pi}$ with the Y-chain. Assuming that the chains start marginally

stationary, irrespective of their coupling it holds (Biswas and Mackey, 2024) that

$$\mathcal{W}_2^2(\hat{\pi}, \pi) \leq \sum_{t=S}^{T} \mathbb{E}\left[\|X_t - Y_t\|^2\right],$$

for any integer $T \geq S \geq 0$. In practice, we replace expectations by empirical averages and due to burn-in the bound only holds asymptotically as $T \to \infty$. In our experiment, we start the chains independently from $X_0 \sim \pi$ (using a long MCMC run) and $Y_0 \sim \hat{\pi}$. To compute upper bounds, we use natural extensions to the GCRN and GCRefl couplings, changing $n_y \leftarrow \hat{n}_y = \mathrm{Nor}(\nabla \log \hat{\pi}(Y_t))$. As coalescence is not the goal here, none of our couplings attempt to make the chains meet. We use 100 independent replicates to compute upper bounds, with the same step size h as in Section 3.7.1. We also compute (see Gelbrich, 1990) the lower bound $\mathcal{W}_2^2(\mathcal{N}_d(\hat{\mu}, \hat{\Sigma}), \mathcal{N}_d(\mu, \Sigma)) \leq \mathcal{W}_2^2(\hat{\pi}, \pi)$, where (μ, Σ) are the mean and covariance of π.

The numerical results are displayed in Figure 3.7.2. The GCRN coupling produces the most informative upper bound on the bias of the Laplace approximation in Wasserstein distance, which is also remarkably sharp compared to the lower bound. For context, the upper bound implies the relative error bound $\|\hat{\mu} - \mu\|/\|\mu\| \leq 10\%$. Compared to the Wasserstein distance, the total variation distance is a much more restrictive metric in higher dimensions, and as a consequence non-trivial bounds on $\mathrm{TV}(\hat{\pi}, \pi)$ are harder to obtain (Biswas and Mackey, 2024). Nonetheless, by adapting the coalescive two-scale GCRefl coupling, we were able to obtain the bound $\mathrm{TV}(\hat{\pi}, \pi) \leq 0.964\,(\pm 0.003)$ to two standard errors.

3.7.3 Coupling the Hug and Hop algorithm

The stochastic volatility model considered in the previous sections provided a challenging test case for the RWM and its couplings. The RWM was not necessarily a practical algorithm for the problem; one would have preferred a more scalable gradient-based

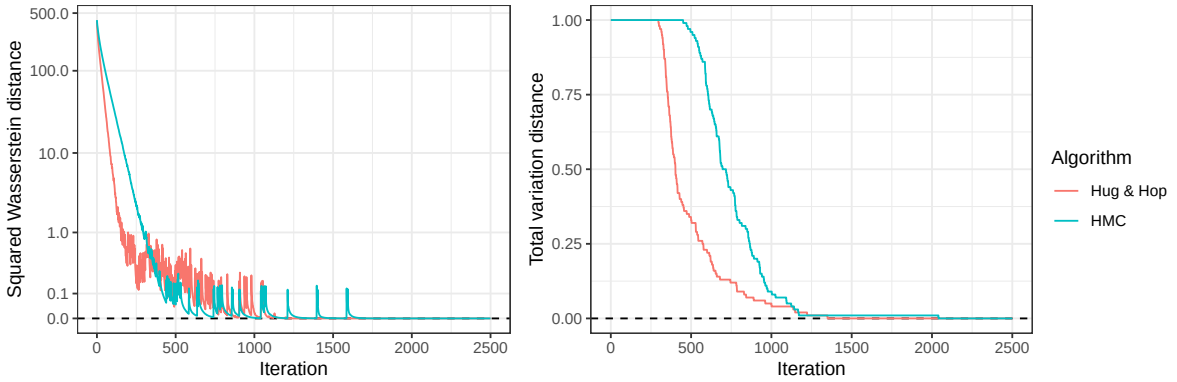

Figure 3.7.3: Rate of convergence of the H&H and HMC algorithms targeting the SVM, when started from the prior. See Section 3.7.3 of the main text for details on these upper bound estimates.

algorithm such as Hamiltonian Monte Carlo (HMC; Duane et al., 1987; Neal, 2011) or the recently-proposed Hug and Hop (H&H; Ludkin and Sherlock, 2022). Our goal in this section is two-fold: to study the rate of convergence of H&H with couplings and to demonstrate that the coupling ideas we developed for the RWM extend to other algorithms.

H&H alternates between the skew-reversible Hug kernel, which proposes large moves by approximately traversing the same level set of the log-target, and the Hop kernel, which crosses between level sets by encouraging large moves in the gradient direction. Hop uses Gaussian proposals centered at the current state; the projection of the proposal onto the subspace orthogonal to the gradient is isotropic, and the projection onto the gradient uses a much larger scaling than each coordinate of its orthogonal counterpart. Due to its similarities to the RWM kernel and its use of gradient information, GCRN-like couplings are natural for the Hop kernel; one might also expect such couplings to be contractive. To couple Hop proposals, we must couple Gaussians with different covariance matrices; whereas GCRN can be straightforwardly extended to this case, extending the reflection-maximal coupling appears significantly more challenging (Corenflos and Särkkä, 2022).

Our coupling of H&H kernels is inspired by contractive couplings of HMC and RWM

kernels and does not require additional gradient evaluations compared to, say, simulating two H&H chains independently. We synchronize the momenta in Hug proposals; such a coupling contracts HMC proposals (e.g. Heng and Jacob, 2019), and due to the parallels of Hug and HMC we expect the contractivity to carry over to Hug. For Hop, we use a two-scale coupling which aims to contract the chains when far apart and allow for exact meetings when close together. When $\|X_t - Y_t\|^2 \geq \delta$, we couple Hop proposals according to the GCRN coupling

$$X_p = X_t + \frac{\lambda}{\gamma_x} Z_\nabla n_x + \frac{(\lambda\kappa)^{1/2}}{\gamma_x}\big\{Z - (Z^\top n_x) n_x\big\},$$
$$Y_p = Y_t + \frac{\lambda}{\gamma_y} Z_\nabla n_y + \frac{(\lambda\kappa)^{1/2}}{\gamma_y}\big\{Z - (Z^\top n_y) n_y\big\},$$

where: $\gamma_x = \|\nabla \log \pi(X_t)\|$, $n_x = \nabla \log \pi(X_t)/\gamma_x$ and similarly γ_y and n_y; $Z \sim \mathcal{N}_d(0_d, I_d)$ and $Z_\nabla \sim \mathcal{N}_1(0,1)$ are independent. When $\|X_t - Y_t\|^2 < \delta$, we sample the proposals from a maximal coupling with independent residuals (Jacob et al., 2020b, Algorithm 2). We encourage simultaneous acceptance in both the Hug and Hop kernels by synchronizing the uniform acceptance variates.

We also compare H&H with HMC, following Heng and Jacob (2019): we use a synchronous coupling of HMC kernels and mix in coupled RWM kernels to allow the chains to meet. As HMC is particularly sensitive to its tuning parameters, we perform extensive experimentation to optimize the contractivity of this algorithm; to ensure a fair comparison, we also do this for Hug. We find that HMC suffers from a sharp phase transition, with long integration times adversely affecting the contractivity of the coupling, whereas Hug does not have this issue. For both HMC and Hug, the acceptance rates at the optimally contractive parameters are higher than would be optimal for mixing (Beskos et al., 2013; Ludkin and Sherlock, 2022); this is linked to contraction being adversely affected by acceptance in one chain but rejection in the other, as discussed in Section 3.6.

Figure 3.7.3 shows estimated upper bounds on the rate of convergence of H&H and HMC, computed with a lag of $L = 2500$ and with 100 replicates each. H&H coalesces quicker than HMC here, while also being more robust with respect to tuning. Both algorithms coalesce two orders of magnitude quicker than the RWM, suggesting that they would be suitable for unbiased MCMC in this setting.

3.7.4 Comparison of the RWM and MALA algorithms

We consider a logistic regression posterior π on the UCI Sonar dataset with $(n, d) = (208, 61)$ observations and regressors (covariates and intercept). Following standard practice (Gelman et al., 2008), we standardize the covariates to scale 0.5, then place a spherical Gaussian prior $\mathcal{N}_d(0_d, 25I_d)$ jointly on the regressors. The RWM and the Metropolis-adjusted Langevin algorithm (MALA; e.g. Roberts and Rosenthal, 1998) perform well in such problems of moderate dimension and with relatively few observations (Chopin and Ridgway, 2017). Our goal here is to compare these algorithms in a realistic setting, with a view towards unbiased MCMC. We are therefore interested both in their time to coalescence, as well as their performance at stationarity.

To ensure a fair comparison between algorithms, we emulate lagged couplings with large lag parameters L by always simulating couplings between a stationary X-chain and a Y-chain which is started at the posterior mean. Alongside meeting times, this set-up allows us to obtain *unbiased* estimates of the asymptotic variance of the MCMC algorithms using the "EPAVE" estimator of Douc et al. (2023), see Appendix A.3.4. As is commonly done in practice, we precondition the RWM and MALA proposals to, respectively:

$$X_t + (hP)Z_x, \quad X_t + \big(h^2 PP^\top/2\big)\nabla \log \pi(X_t) + (hP)Z_x,$$

where $Z_x \sim \mathcal{N}_d(0_d, I_d)$ and where $P \in \mathbb{R}^{d\times d}$ is estimated from a preliminary run. The

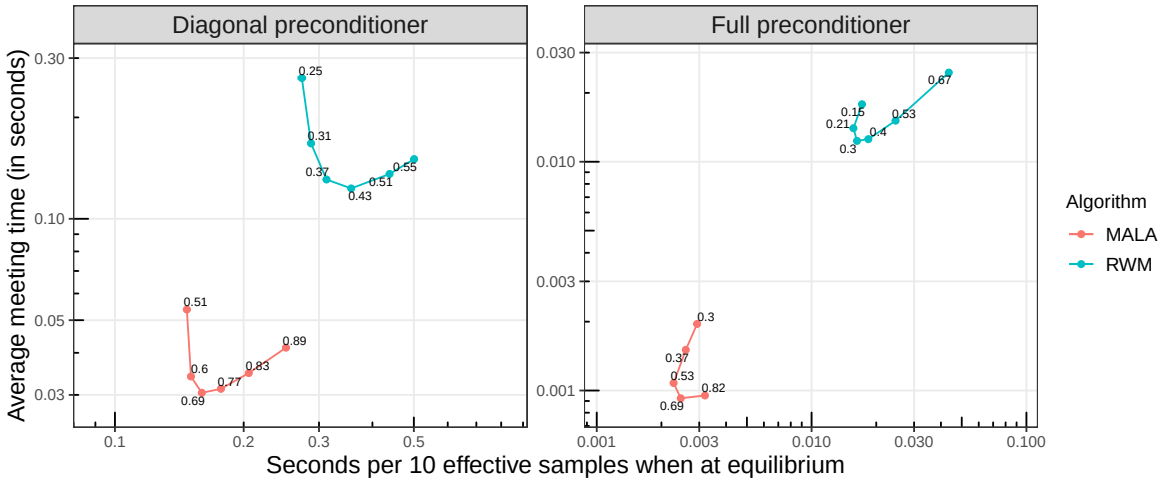

Figure 3.7.4: Comparison of RWM and MALA as in Section 3.7.4, varying the step size parameter h. All estimates are shown with two standard errors. The acceptance rate for each step size is overlaid.

preconditioning requires the change $n_x = \mathrm{Nor}(P^T \nabla \log \pi(X_t))$ in our gradient-based couplings and $e = \mathrm{Nor}(P^{-1}(X_t - Y_t))$ in our reflective couplings, see Appendix A.2.2. We consider two choices for the preconditioner P: a diagonal matrix corresponding to the standard deviations of the target marginals and the "full" Cholesky factor of the target covariance matrix.

In preliminary experiments (see Appendix A.3.4) we found reflective couplings to be particularly useful for this problem. For the RWM, we select a two-scale coupling which applies GCRefl when $\|P^{-1}(X_t - Y_t)\|^2 \geq \delta$, and otherwise applies a reflection-maximal coupling, with thresholds $\delta_{\mathrm{diag}} = 10h^2$ and $\delta_{\mathrm{full}} = 1$ depending on the preconditioner P. For the diagonal preconditioner, GCRefl was the only practical coupling; other couplings suffered disproportionately from the fact that the preconditioned target was highly skewed, with the largest eigenvalue of the preconditioned covariance being nearly three-tenths of the trace. For the full preconditioner, the reflection coupling was practical, but we found GCRefl to outperform it even after accounting for the additional gradient evaluations. For MALA, we use a reflection-maximal coupling. We emphasize that both the RWM and MALA couplings require access to the gradient oracle.

Figure 3.7.4 compares the RWM and MALA algorithms in terms of wall-time. (See Appendix A.3.4 for a figure which is not scaled by wall-time.) To measure mixing at stationarity, we consider the problem of estimating the regression coefficients and we report values corresponding to the smallest effective sample size across all coordinates. When using a diagonal preconditioner and optimizing for mixing/meeting, the RWM is $2\times/4\times$ slower than MALA. This may seem surprisingly close; it stems from the fact that MALA is more sensitive to the eccentricity of the target than the RWM (Livingstone and Zanella, 2022), so that both samplers must use similar step sizes in our setting. The efficiency gap widens when using a full preconditioner; optimizing for mixing/meeting, the RWM is $7\times/13\times$ slower than MALA. Translating these relative efficiencies to the unbiased MCMC setting, mixing efficiency is more relevant when a small number of long chains are run, and coupling efficiency is more relevant when a large number of short chains are run; we expect MALA to more clearly outperform the RWM in the latter regime. In passing, there is a trade-off between coupling and mixing efficiency when using the diagonal preconditioner, whereas using the full preconditioner mitigates this issue.

3.8 Discussion

The main takeaway from this paper is the following: in order to design effective couplings of the RWM algorithm, one should pay careful attention to the coupling of the acceptance steps. We have shown that, by making judicious use of gradient information so as to synchronize acceptance events with high probability, one can design contractive couplings which remain effective even in high-dimensional regimes. We have demonstrated the effectiveness of our proposed GCRN and GCRefl couplings at estimating the rate of convergence and the asymptotic variance of the RWM algorithm. We have also demonstrated how the contractivity of the GCRN coupling can be leveraged to

estimate the bias of an approximate sampling method.

The utility of our proposed couplings of the RWM for unbiased MCMC is less clear. They demand access to a gradient oracle; it is unclear to us how one could devise similarly scalable couplings of the RWM algorithm without such information, yet access to this oracle also enables the use of more scalable gradient-based algorithms altogether, such as MALA and HMC. We have exhibited one moderate-dimensional multi-scale setting where the gap between the RWM and MALA is not as insurmountable as one might expect a priori. Random walk proposals are also a competitive alternative in the pseudo-marginal setting (Andrieu and Roberts, 2009). Middleton et al. (2020) demonstrate the use of simple maximal couplings for unbiased pseudo-marginal MCMC; devising improved couplings will however require additional design considerations to ours, as one must also explicitly account for the noise induced by the unbiased estimator of the target density.

Our work points to several avenues of investigation. Methodologically, we expect the framework of asymptotic optimality to yield improved couplings for other MCMC algorithms. Indeed, we have demonstrated how a straightforward extension of our ideas yielded an effective coupling of the Hug and Hop algorithm that is well-suited to unbiased MCMC, as its per-iteration cost is no larger than that of two independent chains. For MALA, preliminary results (not included) suggest that, as for the RWM, higher-order derivative information than that available in the proposal should be incorporated in an asymptotically optimal coupling. Theoretically, contractive couplings have been used to obtain quantitative convergence rates for MCMC algorithms, e.g. MALA (Eberle, 2014) and HMC (Bou-Rabee et al., 2020). Our scaling limits suggest that, by adapting the couplings proposed in this paper, a sharp $\mathcal{O}(d)$ dimensional dependence for the RWM (Andrieu et al., 2024) could be recovered. Finally, extensions of our scaling limits beyond the Gaussian case may be possible (Kuntz et al., 2019), and a formal explanation of the success of the proposed couplings could be obtained by

establishing deeper connections between the Langevin diffusion and the RWM.

Chapter 4

On the efficiency of lagged coupling methods

4.1 Introduction

The unbiased MCMC estimators of Jacob et al. (2020b) and the convergence bound of Biswas et al. (2019) offer principled ways of performing statistical inference and assessing the performance of MCMC algorithms. These methods rely on a coupling construction, together with various scalar tuning parameters denoted by (k, m, L) in Atchadé and Jacob (2024). Although much research has focused on designing effective couplings for these methods, comparatively little effort has been devoted to tuning the scalar parameters (k, m, L), and to understanding how these parameters impact the efficiency of these methods. The purpose of this chapter is to help fill this gap.

4.1.1 Efficiency of unbiased estimators

Recall the L-lag coupling construction (2.3.2) from Chapter 2, generating a coupled pair of Markov chains $(X_t, Y_t)_{t\geq 0}$ based on the marginal kernel P and the joint kernel

$\bar{P}$. The chains coalesce at the meeting time

$$\tau^{(L)} = \inf\{t \geq 0 : X_{t+L} = Y_t\},$$

where here and later on we use the superscript (L) to explicitly denote a dependence on the time-lag L.

The "single-term" unbiased estimator of Jacob et al. (2020b) is

$$H_t^{(L)} = h(X_t) + \sum_{i=0}^{\lfloor(\tau^{(L)}-t-1)/L\rfloor} \{h(X_{t+iL+L}) - h(Y_{t+iL})\} =: h(X_t) + B_t^{(L)},$$

whereas the "time-averaged" unbiased estimator $H_{k:m}^{(L)} = \frac{1}{m-k+1}\sum_{t=k}^{m} H_t^{(L)}$ expands as

$$H_{k:m}^{(L)} = \frac{1}{m-k+1}\sum_{t=k}^{m} h(X_t) + \frac{1}{m-k+1}\sum_{t=k}^{\tau^{(L)}-1} c_{k:m}^{(L)}(t)\{h(X_{t+L}) - h(Y_t)\} =: h_{k:m} + B_{k:m}^{(L)},$$

where $c_{k:m}^{(L)}(t) = \lfloor (t-k)/L \rfloor - \lceil 0 \vee (t-m)/L \rceil + 1$ (e.g. Douc et al., 2023). Each of the estimators $\{H_t^{(L)}, H_{k:m}^{(L)}\}$ consists of a standard MCMC term and the respective "debiasing term" $\{B_t^{(L)}, B_{k:m}^{(L)}\}$.

Glynn and Whitt (1992) define the inefficiency of an asymptotically unbiased estimator as the product of its expected cost and its variance, in the limit as the computing budget tends to infinity. In practice, one computes expectations by averaging i.i.d. replicates of the time-averaged estimator $H_{k:m}^{(L)}$, so the inefficiency of this estimator is

$$\text{Ineff}(H_{k:m}^{(L)}) = \mathbb{E}[\text{Cost}(H_{k:m}^{(L)})] \times \text{Var}(H_{k:m}^{(L)}),$$

where $\text{Cost}(H_{k:m}^{(L)})$ denotes the computing cost of the estimator. We approximate the computing cost by counting the total number of calls to the marginal kernel P, i.e.

$$\text{Cost}(H_{k:m}^{(L)}) = m \vee (\tau^{(L)} + L) + \tau^{(L)}.$$

(This is because computing the estimator requires simulating the marginal chains exactly up to $X_{m\vee(\tau^{(L)}+L)}$ and $Y_{\tau^{(L)}}$.) This is sensible when a call to the joint kernel $\bar{P}$ costs approximately twice as much as a call to the marginal kernel P, as is often the case. As a baseline, we contrast $\mathrm{Ineff}(H^{(L)}_{k:m})$ with the inefficiency of the standard MCMC average $h_{k:m}$ as the number of iterations tends to infinity: the asymptotic variance, $\upsilon(h,P) = \lim_{m\to\infty} m \times \mathrm{Var}(h_{k:m})$.

We see that obtaining efficient estimators $H^{(L)}_{k:m}$ means balancing cost and variance. The computing cost is determined by (m, L): choosing these parameters larger, we increase the overall cost per estimator, but decrease the relative time spent on simulating the auxiliary Y-chain, whence the cost becomes comparable to standard MCMC. To understand how the parameters (k, L) relate to the variance, it is instructive to consider the single-term estimator $H^{(L)}_t = h(X_t) + B^{(L)}_t$. Intuitively, increasing (t, L) reduces the number of terms in the debiasing term $B^{(L)}_t$, which should reduce the variance of $H^{(L)}_t$. Because $B^{(L)}_t$ has roughly *a factor of L fewer terms* than $B^{(1)}_t$, increasing L can reduce the variance considerably. Irrespective of L, for large t we expect $H^{(L)}_t$ to essentially coincide with the standard MCMC term $h(X_t)$. For large (k, L), the unbiased estimator $H^{(L)}_{k:m}$ should therefore have a similar variance to the standard MCMC estimator $h_{k:m}$.

4.1.2 Efficiency of convergence bound

Chapter 5 provides the following generalization of the convergence bound of Biswas et al. (2019):

$$\mathcal{W}_{p,c}(\pi, \pi_t) \leq \sum_{j\geq 0} \mathbb{E}[c(X_{t+(j+1)L}, Y_{t+jL})^p]^{1/p} =: \mathcal{W}^{(L)}_{p,c}(\pi, \pi_t).$$

In this chapter, we mainly focus on the total variation distance bound ($p = 1$, $c(x, y) = \mathbb{1}\{x \neq y\}$)

$$\mathrm{TV}(\pi, \pi_t) \leq \sum_{i \geq 0} \mathbb{P}(\tau^{(L)} > t + iL) = \mathbb{E}\left[\left\lceil \frac{(\tau^{(L)} - t)_+}{L} \right\rceil\right] =: \mathrm{TV}^{(L)}(\pi, \pi_t),$$

which we estimate by drawing i.i.d. replicates of the meeting time $\tau^{(L)}$.

Our main efficiency consideration is the sharpness of the coupling bound $\mathrm{TV}(\pi, \pi_t) \leq \mathrm{TV}^{(L)}(\pi, \pi_t)$ point-wise over t. Since the bound consists of a series summed in increments of L, increasing L can be expected to sharpen the bound by reducing the number of terms within it.

4.1.3 Our contributions

Our contributions are structured as follows. In Section 4.2 we highlight a finite-sample robustness issue shared by the coupling bound and the unbiased estimators. In Section 4.3, we provide the asymptotic behaviour of the coupling bound and the unbiased estimators as the lag parameter L tends to infinity. These results justify taking the lag L as large as possible in practice. In Section 4.4, through explicit expressions, we provide a detailed quantitative study of the behaviour of the coupling bound and the unbiased estimators in the representative setting of an AR(1) process. In Section 4.5, based on our results and case study, we provide tuning guidelines for practitioners. We conclude with a discussion of our insight in Section 4.6.

4.2 The truncation issue

The unbiased MCMC estimators and finite-sample estimates of the coupling bound are based on truncations of infinite series. Although the truncation is crucial for estimating these quantities in finite time, it can also cause robustness issues, as we now explain.

Suppose that we estimate the coupling bound $\mathrm{TV}^{(L)}(\pi, \pi_t) = \mathbb{E}\big[\lceil (\tau^{(L)} - t)_+ / L \rceil\big]$ unbiasedly using i.i.d. replicates. Because $\lceil (\tau^{(L)} - t)_+ / L \rceil = 0$ whenever $\tau^{(L)} \leq t$, only one in every $1/\mathbb{P}(\tau^{(L)} > t)$ replicates contributes a non-zero amount to our estimate. So, unless we used a very large number of replicates, the distribution of our estimate could be heavily skewed towards zero, and with high probability we could thus underestimate the true distance to stationarity.

A similar phenomenon occurs for the unbiased estimators. Considering replicates of the single-term estimator $H_t^{(L)} = h(X_t) + B_t^{(L)}$ for simplicity, since $B_t^{(L)} \mid \{\tau^{(L)} \leq t\} = 0$ almost surely, only one in every $1/\mathbb{P}(\tau^{(L)} > t)$ replicates has a non-zero debiasing term. Yet, the debiasing term $B_t^{(L)} \mid \{\tau^{(L)} > t\}$ could be large, and thus could contribute significantly towards $H_t^{(L)}$. We might thus obtain misleading estimates of $\mathbb{E}[H_t^{(L)}] = \mathbb{E}_\pi[h(X)]$ and underestimate the variance of $\mathrm{Var}(H_t^{(L)})$ unless we used a very large number of replicates.

Considering that the truncation issue becomes more severe the larger t is, it becomes particularly problematic in situations where accuracy is needed *uniformly over* t. We highlight two such situations. The first is when the coupling bound is used to assess convergence, i.e. we estimate the entire sequence $(\mathrm{TV}^{(L)}(\pi, \pi_t))_{t \geq 0}$ using the same set of replicates. In that case, while the bound estimate may appear smooth, it will almost certainly underestimate the truth once t surpasses a large enough quantile of the empirical meeting times.

The second situation is when tuning the "burn-in" parameter k. One might think to perform several replicates with fixed (m, L), then choose whichever k minimizes the sequence of empirical variances $\big(\widehat{\mathrm{Var}}(H_{k:m}^{(L)})\big)_{k=0}^{m}$. But, due to the truncation issue, the likelihood that the empirical variance is an underestimate increases with k; therefore, we risk choosing k too small and using estimators that are substantially less efficient than we might have anticipated.

Prior work has illustrated that at lag $L = 1$ the truncation issue can be quite severe,

using diffusive MCMC samplers targeting multimodal distributions where the chains tend to "stick" to one mode, see Biswas et al. (2019, Section 2.2.2) for the coupling bound and Jacob et al. (2020b, Section 5.1) for the unbiased estimators. In Section 4.4, we demonstrate that the marginal chains need not exhibit pathological convergence for the truncation issue to appear: it suffices for the lag L to be significantly smaller than the mixing time of the chain. Conversely, Biswas et al. (2019) observed that choosing L large enough can mitigate the truncation issue. We confirm that this is a general phenomenon in Section 4.3.

4.3 Large L asymptotics

Our intuition from Section 4.1 and numerical evidence from prior work (Biswas et al., 2019; Atchadé and Jacob, 2024, e.g.) indicates that choosing the lag parameter L large leads to good performance for both the coupling bound and the unbiased estimators. The purpose of this section is to formalize this intuition, by deriving the behaviour of the meeting times, the coupling bounds and the unbiased estimators in the limit as L grows large. These limits require an appropriate asymptotic framework, which we next introduce.

Definitions. Starting from the construction (2.3.2) with a given lag $L \in \mathbb{N}$, we define the coupled process $(X_t^{(L)}, Y_t^{(L)})_{t\geq 0} := (X_{t+L}, Y_t)_{t\geq 0}$. Because the larger L is, the closer to stationarity the $X^{(L)}$-component is initialized, under weak conditions the coupled process has a well-defined limit as $L \to \infty$, which we shall represent using $(X_t^{(\infty)}, Y_t^{(\infty)})_{t\geq 0}$, initialized at a suitable $(X_0^{(\infty)}, Y_0^{(\infty)}) \in \Gamma(\pi, \pi_0)$ and evolving according to the joint kernel $\bar{P}$. Finally, we define the meeting time $\tau^{(L)} = \inf\{t \geq 0 : X_t^{(L)} = Y_t^{(L)}\}$ for $L \in \mathbb{N} \cup \{\infty\}$.

4.3.1 Meeting times and coupling bounds

Theorem 4.3.2 shows that the distribution of the meeting times stabilizes as $L \to \infty$. To ensure that this and subsequent limits are well-posed, we impose the very weak regularity assumption on how the initializations jointly converge with L.

Assumption 4.3.1: $\mathcal{L}(X_0^{(L)}, Y_0^{(L)}) \overset{\text{TV}}{\to} \mathcal{L}(X_0^{(\infty)}, Y_0^{(\infty)})$ as $L \to \infty$.

Theorem 4.3.2. *Under Assumption 4.3.1, we have* $\tau^{(L)} \implies \tau^{(\infty)}$ *as* $L \to \infty$.

Theorem 4.3.4 shows that the total variation distance coupling bound $\text{TV}^{(L)}(\pi, \pi_t)$ reduces to the usual coupling inequality of Doeblin (1938) as $L \to \infty$. We assume that the expected meeting time (eventually) grows sub-linearly with L, which is a very weak condition since, reasonably, $\mathbb{E}[\tau^{(\infty)}] < \infty$.

Assumption 4.3.3: $\lim_{L\to\infty} \mathbb{E}[\tau^{(L)}/L] = 0$.

Theorem 4.3.4. *Under Assumptions 4.3.1 and 4.3.3, we have* $\lim_{L\to\infty} \text{TV}^{(L)}(\pi, \pi_t) = \mathbb{P}\big(\tau^{(\infty)} > t\big)$ *for all* $t \geq 0$.

There are two main consequences of Theorem 4.3.4. The first is that the coupling bound is non-trivial when L is large, which suggests that it is also more informative. The second is that the bound is more robust when L is large: there is a limit to the severity of the truncation issue (Section 4.2), because now the survivor function of the meeting times is bounded below by the total variation distance to stationarity, i.e. $\mathbb{P}\big(\tau^{(\infty)} > t\big) \geq \text{TV}(\pi, \pi_t)$.

Remark 4.3.5: We can expect the analogue $\lim_{L\to\infty} \mathcal{W}_{p,c}^{(L)}(\pi, \pi_t) = \mathbb{E}[c(X_t^{(\infty)}, Y_t^{(\infty)})^p]$ to hold for the Wasserstein distance coupling bound, but deriving such a result under minimal conditions appears to be more difficult. A simple sufficient condition is $\lim_{L\to\infty} \mathbb{E}[c(X_t^{(L)}, Y_t^{(L)})^p] = \mathbb{E}[c(X_t^{(\infty)}, Y_t^{(\infty)})^p]$, together with $\mathbb{E}[c(X_t^{(L)}, Y_t^{(L)})^p] \lesssim t^{-\kappa}$ uniformly in L for some $\kappa > 1$. We avoid further technical details.

Remark 4.3.6: Because the coupling bound recovers the coupling inequality of Doeblin (1938) for large L, we expect any substantial methodological improvements to the bound to also require substantial changes to the coupling construction (2.3.2). Indeed, while Craiu and Meng (2022) obtains a sharper total variation distance bound based on the same coupling construction, the new bound can only be an improvement when the old bound is trivial, i.e. when $\mathrm{TV}^{(L)}(\pi, \pi_t) > 1$ (Craiu and Meng, 2022, Eqn. 2.22).

4.3.2 Unbiased estimators

In this section, we show that the unbiased estimator $H_{k:m}^{(L)}$ behaves similarly to a standard MCMC estimator, when the lag L is chosen large enough and when m is scaled appropriately with L.

Firstly, Theorem 4.3.8 states that the bias-correction term $B_{k:m}^{(L)}$ vanishes at essentially the rate $\Theta(m^{-1})$, provided that the limiting meeting time is almost surely finite. This assumption is strictly weaker than Assumption 4.3.3.

Assumption 4.3.7: $\mathbb{P}(\tau^{(\infty)} < \infty) = 1$.

Theorem 4.3.8. *Under Assumptions 4.3.1 and 4.3.7, for $m = \Theta(L)$ and for any $f : \mathbb{N} \to (0, \infty)$ such that $f(m) = o(m)$, it holds that $f(m)B_{k:m}^{(L)} \xrightarrow{p} 0$ as $L \to \infty$.*

As a consequence, Theorem 4.3.10 states that the unbiased estimator $H_{k:m}^{(L)} = h_{k:m} + B_{k:m}^{(L)}$ satisfies the same central limit theorem as the standard MCMC ergodic average $h_{k:m}$, under identical regularity assumptions (Jones, 2004). (See Section 2.2 for an overview of polynomial ergodicity.)

Assumption 4.3.9: The Markov chain $(X_t)_{t\geq 0}$ is polynomially ergodic, converging in total variation distance polynomially at rate $\kappa > 1$. Furthermore, the test function h satisfies $\mathbb{E}_\pi[|h(X)|^{2+\delta}] < \infty$ where $\delta > 2/(\kappa - 1)$.

Theorem 4.3.10. *Under Assumptions 4.3.1, 4.3.7 and 4.3.9, for $m = \Theta(L)$, as $L \to \infty$ it holds that*

$$\sqrt{m-k+1}\left(H_{k:m}^{(L)} - \mathbb{E}_\pi[h(X)]\right) \implies \mathcal{N}(0, v(P,h)).$$

Theorem 4.3.10 is similar to Douc et al. (2023, Proposition 19), but requires weaker assumptions. When choosing (m, L) to be large, we expect the coupling to be a relatively small part of the computing cost, whence Theorem 4.3.10 indicates that the unbiased estimator $H_{k:m}^{(L)}$ is nearly as efficient as the standard MCMC estimator.

Next, we complement the qualitative insight provided by Theorems 4.3.8 and 4.3.10 with numerical case studies that provide the precise quantitative behaviour of the unbiased estimator $H_{k:m}^{(L)}$. While quantitative insight could also be provided by theoretical bounds on the variance of the unbiased estimator $H_{k:m}^{(L)}$ (Jacob et al., 2020b, Proposition 3; Middleton et al., 2020, Proposition 1), in contrast to our case studies, these bounds can hide large implicit constants.

4.4 Case study: AR(1) process

We provide a detailed quantitative study of the coupling bound and the unbiased MCMC estimators in a stylized setting. We consider an L-lag reflection-maximal coupling of Gaussian AR(1) processes, where the marginal transition kernel is $P(x, \cdot) = \mathcal{N}(\rho x, 1-\rho^2) \in \mathcal{P}(\mathbb{R})$ and the stationary distribution is $\pi = \mathcal{N}(0,1)$. We focus on the unbiased estimators with the test function $h(x) = x$, and on total variation distance coupling bound.

Douc et al. (2023) considered a similar setting, and bounded the survivor function $\mathbb{P}(\tau^{(L)} > t)$ of the meeting times based on drift and minorization techniques. Here, we go substantially further: we obtain tractable expressions for $\mathbb{P}(\tau^{(L)} > t)$, the coupling bound, and the variance of the single-term estimators $H_t^{(L)}$. This enables us to relate our performance measures of interest directly to the mean-regression parameter ρ, and

furthermore to illustrate the truncation issue (Section 4.2) without falling victim to it.

The theoretical results from this section are proved in Appendix B.2.

4.4.1 Explicit results

By recognizing a reflection-maximal coupling of AR(1) processes as an exact time-discretization of reflection-coupled Ornstein-Uhlenbeck processes, we obtain an explicit expression for the conditional survivor function $\mathbb{P}\big(\tau^{(L)} \geq t \mid X_0^{(L)}, Y_0^{(L)}\big)$.

Theorem 4.4.1. *Consider a reflection-maximal coupling $(X_t^{(L)}, Y_t^{(L)})_{t\geq 0}$ of AR(1) processes with marginal transition kernels $P(x,\cdot) = \mathcal{N}(\rho x, 1-\rho^2) \in \mathcal{P}(\mathbb{R})$. Then, the meeting time $\tau^{(L)}$ has survivor function*

$$\mathbb{P}\left(\tau^{(L)} > t \mid X_0^{(L)}, Y_0^{(L)}\right) = 2\Phi\left(\frac{|X_0^{(L)} - Y_0^{(L)}|\rho^t}{2\sqrt{1-\rho^{2t}}}\right) - 1.$$

The unconditional survivor $\mathbb{P}(\tau^{(L)} > t)$ can be computed by quadrature. Alternatively, when $X_0^{(L)} - Y_0^{(L)}$ is Gaussian, we can express the unconditional survivor in terms of bivariate normal probabilities, see Appendix B.2. Appendix B.2 also provides

$$\mathbb{P}\left(\tau^{(L)} > t\right) = \frac{1}{\sqrt{2\pi}} \cdot \mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] \cdot \rho^t + \Theta(\rho^{2t}). \tag{4.4.1}$$

We estimate the coupling bound $\mathrm{TV}^{(L)}(\pi, \pi_t) = \sum_{j\geq 0} \mathbb{P}\left(\tau^{(L)} > t + jL\right)$ by truncating the series, then correcting for this truncation using the asymptotics (4.4.1):

$$\mathrm{TV}^{(L)}(\pi, \pi_t) \approx \sum_{j=0}^{J} \mathbb{P}\left(\tau^{(L)} > t + jL\right) + \frac{\rho^L}{1-\rho^L} \mathbb{P}\left(\tau^{(L)} > t + JL\right). \tag{4.4.2}$$

This approximation is accurate at least when $t + JL$ is large.

For the variance of the single-term estimator $H_k^{(L)}$, we have the following result.

Theorem 4.4.2. *Consider an L-lag reflection-maximal coupling of AR(1) processes*

$(X_t, Y_t)_{t\geq 0}$ with marginal transition kernels $P(x, \cdot) = \mathcal{N}(\rho x, 1-\rho^2) \in \mathcal{P}(\mathbb{R})$. Then, for the test function $h(x) = x$, we have that

$$\begin{aligned}\mathrm{Var}(H_t^{(L)}) &= \mathbb{E}[X_t^2] + 2\mathbb{E}[X_t B_t^{(L)}] + \mathbb{E}[(B_t^{(L)})^2] \\ &= \mathbb{E}[X_t^2] + \frac{2}{1-\rho^L}\mathbb{E}[X_t(X_{t+L} - Y_t)] + \frac{1+\rho^L}{1-\rho^L}\sum_{j\geq 0}\mathbb{E}[(X_{t+jL}^{(L)} - Y_{t+jL}^{(L)})^2].\end{aligned}$$

We compute the expectations in Theorem 4.4.2 using explicit expressions and/or quadrature, see Appendix B.2. Appendix B.2 also provides

$$\mathbb{E}[(X_t^{(L)} - Y_t^{(L)})^2] = \frac{8}{\sqrt{2\pi}} \cdot \mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] \cdot \rho^t + \Theta(\rho^{2t}), \tag{4.4.3}$$

hence we estimate the infinite series in Theorem 4.4.2 by a corrected truncation that is similar to (4.4.2).

While similar derivations can be carried out for the variance of the time-averaged estimator, $\mathrm{Var}(H_{k:m}^{(L)})$, the resulting expression is considerably more complex and thus its behaviour is less readily apparent. In our experiments, we therefore simply estimate $\mathrm{Var}(H_{k:m}^{(L)})$ using a number of Monte Carlo replicates (10^6) that is large enough to overcome the truncation issue.

Quantitative insight

Parameter tuning is most important when the mixing of the chain is slow, i.e. $\rho \approx 1$. We provide quantitative insight in this regime.

For the meeting times, we recall the asymptotics (4.4.1),

$$\mathbb{P}\left(\tau^{(L)} > t\right) = \frac{1}{\sqrt{2\pi}} \cdot \mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] \cdot \rho^t \text{ to leading order.}$$

Suppose that the initializations $(X_0^{(L)}, Y_0^{(L)})$ are coupled optimally; then, by taking $\rho \to 1$ with a fixed L, the meeting times concentrate at $\tau^{(L)} \leq 1$ with probability

one. So, paradoxically, *the slower the chain converges, the earlier we can force the meetings to occur*. In particular, we can always force early meetings by choosing π_0 as a point-mass, but this is inadvisable for small L because it worsens the truncation issue.

For the coupling bound, the asymptotics (4.4.1) suggest that

$$\mathrm{TV}^{(L)}(\pi, \pi_t) \approx \frac{1}{\sqrt{2\pi}} \cdot \mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] \cdot \frac{\rho^t}{1-\rho^L} \text{ to leading order.}$$

When $\mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] = \Theta(1)$ with L, we have that $\mathrm{TV}^{(L)}(\pi, \pi_t)$ goes as L^{-1} initially, so increasing L can sharpen the coupling bound considerably. Conversely, when the initializations $(X_0^{(L)}, Y_0^{(L)})$ are tightly coupled, the numerator $\mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|]$ might balance out the denominator $1-\rho^L$. Tightly coupling the initializations, we could thus sharpen the coupling bound for small L by encouraging early meetings, at the cost of worsening the truncation issue.

For the debiasing term of the single-term estimator $H_t^{(L)}$, Theorem 4.4.2 and the asymptotics (4.4.3) suggest that

$$\mathbb{E}[(B_t^{(L)})^2] \approx \frac{8}{\sqrt{2\pi}} \cdot \mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] \cdot \frac{\rho^t}{(1-\rho^L)^2} \text{ to leading order.}$$

When $\mathbb{E}[|X_0^{(L)} - Y_0^{(L)}|] = \Theta(1)$ with L, we have that $\mathbb{E}[(B_t^{(L)})^2]$ goes as L^{-2} initially. For slowly mixing chains, increasing L can thus reduce the variance dramatically, and can initially be more effective than increasing t. Although it can reduce it, tightly coupling the initializations $(X_0^{(L)}, Y_0^{(L)})$ cannot fully control the variance for small L, as we will see in the next section.

4.4.2 Behaviour of coupling bound and single-term estimators

We fix $\rho = 0.95$, vary $L \in \{1, 100\}$, and consider three settings: (a) $\pi_0 = \delta_3$; (b) $\pi_0 = \mathcal{N}(3, 9)$ with $(X_0^{(L)}, Y_0^{(L)})$ perfectly correlated; (c) $\pi_0 = \mathcal{N}(3, 9)$ with $(X_0^{(L)}, Y_0^{(L)})$ independent. Figure 4.4.1 displays the coupling bound $\mathrm{TV}^{(L)}(\pi, \pi_t)$, the survivor func-

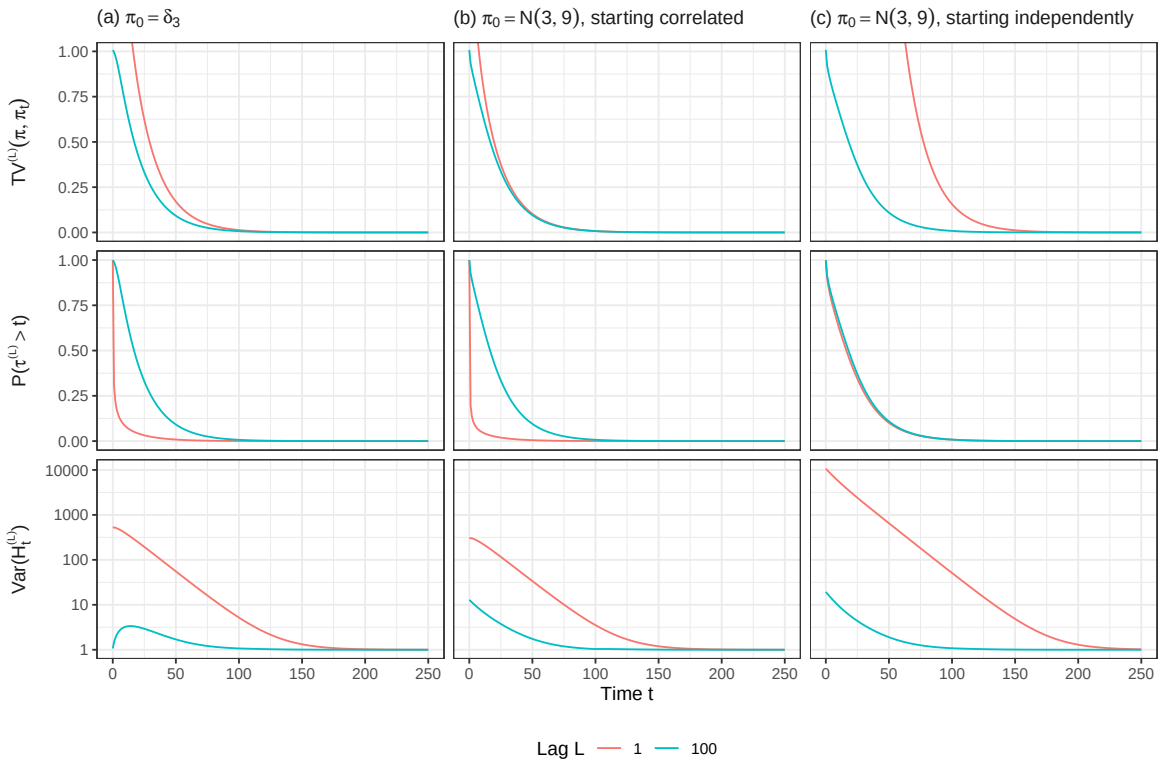

Figure 4.4.1: Behaviour of total variation distance coupling bound, survivor function of meeting times, and variance of single-term estimators, for a reflection coupling of AR(1) processes. See Section 4.4 for details.

tion of the meeting times $\mathbb{P}(\tau^{(L)} > t)$, and the variance of the single-term unbiased estimator $\mathrm{Var}(H_t^{(L)})$, against the time t.

Coupling bound

First, we focus on the coupling bound at lag $L = 1$. Due to the concentration of the meeting times near $t = 0$, in settings (a) and (b) the bound is relatively sharp. However, this concentration near zero also causes Monte Carlo estimates of the bound to be less robust to estimation error: unless a large number of replicates are used, the distribution of the Monte Carlo estimate of the bound has a large point-mass at zero, so it is likely to underestimate the true distance to stationarity. For example, in setting (b) we have that $\mathrm{TV}^{(1)}(\pi, \pi_{50}) \approx 0.1$ while $\mathbb{P}(\tau^{(1)} > 50) \approx 0.00514$. To contribute a non-zero amount to the bound estimate, a replicate must meet at $\tau^{(L)} > 50$, but on average only one in

every 194 replicates does so. Therefore, thousands of replicates would be required to obtain a sensible estimate of the bound. In setting (c), the independent overdispersed initializations ensure that the chains only meet when the marginal distributions are close to stationarity. While this circumvents the robustness issue, it also makes the bound relatively loose. Overall, we see that the coupling bound at $L = 1$ is either loose, or it suffers from the early truncation issue.

Increasing the lag to $L = 100$ makes the coupling bound uniformly sharper in all three settings, and overcomes the truncation issue affecting settings (a) and (b).

Variance of single-term estimators

Next, we turn to the variance of the single-term estimators at lag $L = 1$. We see that $\mathrm{Var}(H_t^{(1)})$ is extremely large for $t = 0$, and that it decreases geometrically at rate ρ towards the limiting value $\lim_{t\to\infty} \mathrm{Var}(H_t^{(1)}) = \mathrm{Var}_\pi(X) = 1$ as t increases.

Increasing the lag to $L = 100$ decreases the variance of $\mathrm{Var}(H_t^{(L)})$ uniformly for all t. In particular, for small t, *the variance is reduced by over an order of magnitude*. By increasing L, we can thus retain considerably more of the initial simulation output while controlling the variance, resulting in time-averaged estimators that are more efficient for the same amount of computation, as we will see in Figure 4.4.3.

Figure 4.4.1 provides the exact variance $\mathrm{Var}(H_t^{(L)})$, but does not illustrate the behaviour of Monte Carlo estimates of this variance. To assess the extent of the truncation issue for the unbiased estimator $H_t^{(L)}$, we next consider its entire density.

Density of single-term estimators

The density of the single-term estimator $H_t^{(L)}$ is a mixture of two components: the primary component $H_t^{(L)} \mid \tau^{(L)} \leq t$ i.e. $X_t \mid \tau^{(L)} \leq t$ with weight $\mathbb{P}(\tau^{(L)} \leq t)$, and the residual component $H_t^{(L)} \mid \tau^{(L)} > t$ with weight $\mathbb{P}(\tau^{(L)} > t)$. As t increases, the primary component tends to the stationary distribution π, whereas the residual component tends

to a non-trivial limit that we describe in Appendix B.2. Figure 4.4.2 illustrates these limits and contrasts them against the density of $H_0^{(L)}$.

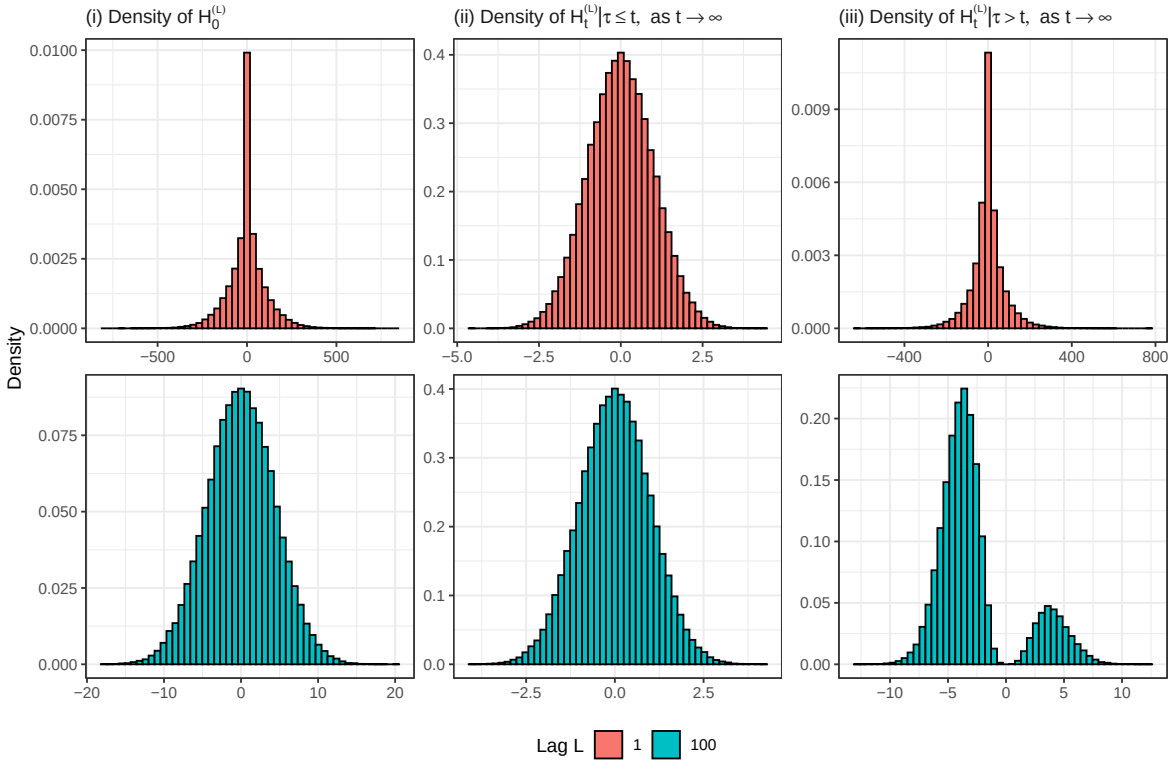

Figure 4.4.2: Behaviour of single-term estimator $H_t^{(L)}$ in setting (c) of Figure 4.4.1. The plots are based on 10^5 replicate simulations each; the behaviour in settings (a) and (b) is similar. See Section 4.4 for details.

At lag $L = 1$, the residual component $H_t^{(1)} \mid \tau^{(1)} > t$ is two orders of magnitude more dispersed than the primary component. This inflates the variance of $H_t^{(1)}$ (recall Figure 4.4.1); worse, because the large residual component does not shrink with t, as t increases it becomes increasingly challenging to obtain accurate estimates of the variance $\mathrm{Var}(H_t^{(1)})$ from Monte Carlo simulations, as a manifestation of the truncation issue. For example, at $t = 150$ we have that $\mathrm{Var}(H_{150}^{(1)}) \approx 4.85$, most of which comes from the residual component that has the very small weight $\mathbb{P}(\tau^{(1)} > 150) \approx 0.0006$. Because, on average, only one in every 1660 replicates is sampled from the residual component, our Monte Carlo estimate of $\mathrm{Var}(H_{150}^{(1)})$ is likely to underestimate the truth unless tens of thousands of replicates are drawn.

Increasing the lag to $L = 100$ shrinks the size of the residual component, bringing it to the same scale as the primary component. As a consequence, the variance $\mathrm{Var}(H_t^{(L)})$ is greatly reduced; furthermore, because considerably less of the variance comes from the small-weight residual component, we are considerably less likely to underestimate this variance from Monte Carlo simulations.

4.4.3 Efficiency of time-averaged estimators

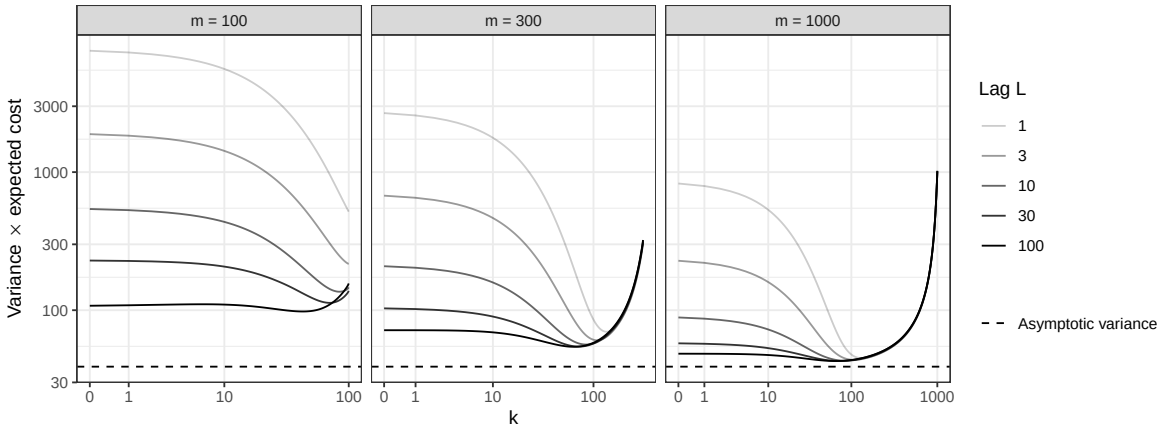

Figure 4.4.3: Inefficiency of time-averaged unbiased estimators $H_{k:m}^{(L)}$ in setting (a) of Figure 4.4.1; the behaviour in settings (b) and (c) is similar. The baseline inefficiency is the asymptotic variance of a standard MCMC estimator. See Section 4.4 for details.

We investigate the efficiency of the time-averaged estimators $H_{k:m}^{(L)}$ as (k, m, L) are varied. Recall that the inefficiency of an estimator is defined as its variance times its expected cost. Figure 4.4.3 plots this quantity for several values of (k, m, L), in setting (a) of the previous experiment.

First, consider the case of $L = 1$. We see that the efficiency of the estimator $H_{k:m}^{(1)}$ is extremely sensitive to choosing k sufficiently large. Yet at the same time, Monte Carlo estimates of the variance $\mathrm{Var}(H_{k:m}^{(1)})$ are increasingly likely to underestimate the truth as k increases, due to the truncation issue. For small L, practitioners setting k based on preliminary runs therefore risk using estimators $H_{k:m}^{(L)}$ that are considerably less efficient than they might have anticipated. Interestingly, we see that for relatively

small (m, L) time-averaging need not help: setting $k = m$ can be optimal.

Increasing L improves the efficiency of the unbiased estimator $H^{(L)}_{k:m}$, reduces its sensitivity to the tuning parameter k, and increases the robustness of its variance estimates. Practitioners choosing L large can therefore not only obtain more efficient estimators, but can also be more confident that they have accurately assessed the efficiency of their estimators. Consistent with Figure 4.4.1, the benefits of increasing L are most apparent for smaller values of m: when $m = 100$, increasing the lag from $L = 1$ to $L = 100$ improves the efficiency by up to two orders of magnitude, depending on the value of k.

Finally, for large enough m, and appropriately large L and/or k, we see that the estimator $H^{(L)}_{k:m}$ is almost as efficient as a stationary MCMC ergodic average. Intuitively, this is the regime where the debiasing term of $H^{(L)}_{k:m}$ only contributes a small amount to the variance, and where the overhead of the coupling is small compared to simulating a single MCMC chain for m iterations.

4.5 Tuning advice

Our main advice to practitioners is the following:

Choose the lag $L \gg 1$.

A simple rule of thumb is that L should be on the order of the meeting times, or larger.

To understand our recommendation, it is helpful to recapitulate all of the problems that we have seen to occur when L is small. We have seen that the coupling bound can be quite loose. For the unbiased estimators, we have seen them to be extremely variable, and therefore inefficient. Furthermore, we have seen that finite-sample estimates of the coupling bound and of the variability of the unbiased estimators have a tendency to underestimate the truth. Perhaps paradoxically, this tendency to underestimate is caused by the same random truncation that enables the computation to occur in finite

time. Altogether, practitioners using small L not only risk using inefficient bounds and estimators, but also risk being unaware of this fact.

Choosing L large enough can mitigate all of these issues. The theoretical results of Section 4.3 provide reassurance that, when L is large, the coupling bound is informative, that the unbiased estimators are as efficient as the underlying standard MCMC estimators, and that finite-sample estimates of both the coupling bound and of the variability of the unbiased estimators are representative. Indeed, robustness is achieved because there is a limit to how early the random truncation can occur, since for large L the survivor function of the meeting times is bounded below by the distance of the time-marginals to stationarity. Our case studies confirm that an appropriate choice of tuning parameters does indeed result in efficient and robust coupling bounds and unbiased estimators.

Practical tuning guidelines

We now provide practical guidelines for tuning the parameters (k, m, L), based on preliminary runs.

The lag parameter L should be tuned first. We recommend tuning L similarly for both the coupling bound and the unbiased estimators. Our guideline is to set L to a large quantile (e.g. 90%) of the distribution of the meeting time $\tau^{(L)}$ *once this distribution has stabilized with increasing* L. To assess this stability, one could compare histograms or empirical cumulative distributions of the meeting times as e.g. L is successively doubled, and stop when the visual change is negligible compared to Monte Carlo error. Our guideline requires L to be larger than the mixing time of the marginal MCMC algorithm. We thus recommend starting with a value of L set to an initial guess of the mixing time; this could be obtained using the method of Chapter 5, a running plot of the $\widehat{R}$ convergence diagnostic of Gelman and Rubin (1992) as in e.g. Vats and Knudson (2021, Figure 2) (declaring that approximate convergence has occurred once a

small enough threshold has been reached), or algorithm- and target-specific theoretical bounds (e.g. Durmus and Moulines, 2019; Bou-Rabee et al., 2020).

For the unbiased estimators, we suggest reparametrizing $(k, m) = (k, T + k - 1)$ so that T represents the effective number of single-term estimators within the time-averaged estimator $H^{(L)}_{k:m} = H^{(L)}_{k:(T+k-1)}$. We recommend setting the "burn-in" k such that the variance of the single-term estimators $(H^{(L)}_t)_{t\geq k}$ is similar to that of the MCMC terms $(h(X_t))_{t\geq k}$. This ensures that the unbiased time-averaged estimator has a similar variance to a standard MCMC estimator. With an appropriately large L, a conservative choice of burn-in would be $k = L$. Finally, it makes sense to set T as a multiple of L, to reduce the computational overhead of the coupling compared to standard MCMC. Remark 4.5.1 indicates that $T = 5L$ should ensure that $H^{(L)}_{k:(T+k-1)}$ is at worst 60% more inefficient than a stationary MCMC average with no burn-in.

Remark 4.5.1: Concretely, assuming that $k \leq L$, that $T+k \geq L$ and that $\mathbb{E}[\tau^{(L)}] \leq L$, we have that $\mathbb{E}[\text{Cost}(H^{(L)}_{k:(T+k-1)})] \leq T+k+2\mathbb{E}[\tau^{(L)}] \leq T+3L$. Compared to a length-$T$ stationary MCMC average with no burn-in, the relative inefficiency of the unbiased estimator $H^{(L)}_{k:(T+k-1)}$ is therefore at worst $1 + 3L/T$, and significantly less than this whenever $T \gg k \vee \mathbb{E}[\tau^{(L)}]$.

Comparison with previous tuning guidelines

For the coupling bound, Biswas et al. (2019) suggests tuning the lag L by starting with $L = 1$ and increasing it until $\text{TV}^{(L)}(\pi, \pi_0) \lessapprox 1$. For the unbiased estimators, Douc et al. (2023); Atchadé and Jacob (2024) suggest tuning (k, m, L) by starting with $L = 1$, then redefining L as a large quantile of the meeting time $\tau^{(1)}$, setting $k = L$, and setting m as a multiple of k. Since these guidelines start the search at $L = 1$, they are vulnerable to underestimating L due to the early meeting issue.

Vanetti and Doucet (2020) suggest tuning the unbiased MCMC by setting $T = L$ very large, which simplifies the unbiased estimator to

$$H^{(T)}_{k:(T+k-1)} = \frac{1}{T}\sum_{t=k}^{T+k-1} h(X_t) + \frac{1}{T}\sum_{t=k}^{\tau^{(T)}-1} \{h(X_{t+T}) - h(Y_t)\}\,.$$

Intuitively, this estimator corrects a long MCMC run with a short auxiliary chain *post hoc*. When tuning (T, k) according to our guidelines, we expect $H^{(T)}_{k:(T+k-1)}$ to be nearly as efficient as a standard MCMC estimator, whence the proposal of Vanetti and Doucet (2020) is a viable practical alternative with fewer tuning parameters.

Alternative perspective on tuning parameters

The better the coupling of the chains, the better we expect our unbiased estimator $H^{(L)}_k$ to perform. This suggests an alternative perspective on the relative merits of tuning the parameters (k, L). The parameter L indicates that the estimator is computed using the L-step coupling kernel $\bar{P}^L$. As $L \to \infty$, we expect the L-step coupling kernel $\bar{P}^L$ to coalesce the chains in one step from any initialization: the benefits of increasing L are therefore uniform across the entire state-space and across all k. Meanwhile, the parameter k dictates the coupling of the "initializations" $(X_{k+L}, Y_k) \in \Gamma(\pi_{k+L}, \pi_k)$ in the debiasing term $B^{(L)}_k$. Increasing k improves this coupling; however, when the lag L is small, the unbiased estimator still relies on a potentially inefficient coupling kernel, so we expect the worst-case performance of the estimator to remain the same. In order of importance, it thus makes sense to tune L first, and k second.

4.6 Discussion

Our study confirms unbiased MCMC as a principled alternative to standard MCMC. We have demonstrated that unbiased MCMC estimators can be as efficient as standard MCMC estimators when the computational budget per estimator is sufficiently large,

while having several additional advantages: a "self-terminating" property, the ability to circumvent the need for traditional burn-in (and to quantify the burn-in via the bound of Biswas et al., 2019), and ease of use when averaging independent replicates obtained in parallel. That the estimators are truly unbiased also facilitates various inferential tasks, see the applications Jacob et al. (2017); Ruiz et al. (2021); Wang and Wang (2023); Douc et al. (2023).

Whether unbiased MCMC can be a faster alternative to well-tuned standard MCMC remains to be seen. In the highly-parallel regime where the priority is to return acceptable results as quickly as possible, empirical evidence (Wang et al., 2024, Section 5) points to the contrary. This concurs with our case study, which suggests that unbiased MCMC requires a large enough budget per estimator to perform well.

Chapter 5

Centered plug-in estimation of Wasserstein distances

5.1 Introduction

Wasserstein distances are a class of probability metrics rooted in the theory of optimal transport (Villani, 2003, 2009) that increasingly underpin methodological developments in statistics (Panaretos and Zemel, 2019) and machine learning (Peyré and Cuturi, 2019).

We are motivated by two important problems from Bayesian computation: (i) assessing the quality of approximate inference methods, and (ii) assessing the convergence of Markov chain Monte Carlo (MCMC) algorithms to their limiting distributions. The former is one of the present *grand challenges* in Bayesian computation (Bhattacharya et al., 2024), whereas the latter has been challenging practitioners for over thirty years (Gelman and Rubin, 1992). Assessing the accuracy in terms of Wasserstein distances is particularly appealing in these contexts, because bounds on Wasserstein distances guarantee the accuracy of various downstream inferential tasks (Huggins et al., 2020). At the same time, because we want to recognize when an approximation is accurate,

one key requirement for Wasserstein distance estimators in these contexts is that they decrease with the Wasserstein distance itself.

Standard plug-in estimators of the Wasserstein distance have substantial positive biases that are particularly apparent when the Wasserstein distance is small. Furthermore, due to fundamental statistical challenges related to estimating Wasserstein distances (Hütter and Rigollet, 2021), this bias can often not be meaningfully reduced by increasing the sample size. To obtain informative estimators of the Wasserstein distance, we must therefore resolve the issue of bias by a different approach.

We eliminate most of the bias using a simple centering procedure based on linear combinations. Because this centering ensures that the bias decreases with the true Wasserstein distance for any finite sample size, it allows us to circumvent statistical challenges and obtain informative estimates, at moderate sample sizes, even in high dimensions. In a nutshell, we construct a pair of complementary estimators: U, which is often an approximate *upper bound* on the squared Wasserstein distance, and L, which is always an approximate *lower bound*. Formal sufficient conditions for U to be conservative may be interpreted as a form of overdispersion between the two distributions, which aligns naturally with our motivating problems from Bayesian computation.

The paper is organized as follows. In Section 5.2 we review key aspects of Wasserstein distances and their estimation. In Section 5.3 we introduce the new centered estimators, analyze their finite-sample statistical properties, and discuss how to efficiently quantify their uncertainties. In Section 5.4 we develop a methodology for assessing the quality of approximate inference methods, and in Section 5.5 we develop a methodology for assessing the convergence of MCMC algorithms; both of these are based on post-processing the output of multiple replicate Markov chains using the centered estimators. We summarize our findings and outline directions for further research in Section 5.6. R (R Core Team, 2025) code is available on GitHub.

5.2 Plug-in estimation of Wasserstein distances

We review here selected aspects of Wasserstein distances and their estimation. We refer the reader to the works Villani (2009); Peyré and Cuturi (2019); Panaretos and Zemel (2019) for further theoretical, computational, and statistical details, respectively.

Let $(\mathcal{X}, c)$ be a metric space and let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be probability distributions on $\mathcal{X}$. The p-Wasserstein distance is defined, through its p-th power, as the solution to the optimal transportation problem

$$\mathcal{W}_p^p(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \int c(x,y)^p \mathrm{d}\pi(x,y) = \inf_{X \sim \mu, Y \sim \nu} \mathbb{E}\left[c(X,Y)^p\right], \tag{5.2.1}$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ with marginals (μ, ν). The primal problem (5.2.1) admits the Kantorovich dual formulation

$$\mathcal{W}_p^p(\mu, \nu) = \sup_{(\phi,\psi) \in \Phi(\mu,\nu)} \int \phi(x) \mathrm{d}\mu(x) + \int \psi(y) \mathrm{d}\nu(y),$$

$$\Phi(\mu, \nu) = \{(\phi, \psi) \in L_1(\mu) \times L_1(\nu) \mid \phi(x) + \psi(y) \leq c(x,y)^p, \ \ \forall x, y\}.$$

We use $(\phi_{\mu,\nu}, \psi_{\mu,\nu})$ to denote a pair of optimal potentials for the Kantorovich dual. Properties of Wasserstein distances include (Villani, 2009): $\mathcal{W}_p$ defines a metric on the set of distributions with finite p-th moments, it induces an intuitive geometry and controls weak convergence on this set, and it controls the discrepancy between certain moments of Lipschitz functions.

In this paper, we are interested in estimating Wasserstein distances in practice. Since the behavior of Wasserstein distance estimators is extremely rich from a statistical perspective, and depends on the features of the distributions of interest as well as of the distance itself, we must make some assumptions. In this paper, we specialize to continuous distributions in $\mathcal{X} = \mathbb{R}^d$, we fix the ground metric to be Euclidean $c(x, y) = \|x - y\|$, and we fix the exponent to $p = 2$. Throughout the entire sequel, we

impose the regularity assumption:

(A0) The distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and satisfy $\mathbb{E}_\mu\left[\|X\|^2\right], \mathbb{E}_\nu\left[\|Y\|^2\right] < \infty$.

Brenier's theorem (1991) then provides the unique solution $\mathcal{W}_2^2(\mu,\nu) = \mathbb{E}_\nu[\|T_{\nu,\mu}(Y) - Y\|^2]$ in terms of an optimal transport map $T_{\nu,\mu}$ that pushes ν forward to μ.

We focus on the case where independent samples $X_{1:n} \sim \mu$ and $Y_{1:n} \sim \nu$ are available from each distribution. We define the empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, and we call $\mathcal{W}_2^2(\mu_n,\nu_n)$ the *plug-in estimator* of the squared Wasserstein distance $\mathcal{W}_2^2(\mu,\nu)$, which we now review from a computational and statistical perspective.

5.2.1 Computational aspects

Exact computational methods treat the plug-in estimator $\mathcal{W}_2^2(\mu_n,\nu_n)$ as the solution to a linear assignment problem. Although the worst-case theoretical complexity of exact assignment problem solvers is $O(n^3)$, particularly efficient solvers (Bonneel et al., 2011; Guthe and Thuerck, 2021) have complexities closer to $O(n^2)$ in practice, see the benchmark of Appendix C.6.1.

Among approximate methods, the most popular is that of Cuturi (2013), which solves for an entropy-regularized version of $\mathcal{W}_2^2(\mu_n,\nu_n)$ using Sinkhorn's algorithm. This has complexity $O(n^2/\varepsilon^2)$ (Dvurechensky et al., 2018) depending on the size of the regularization parameter ε, but is well-suited to vectorized hardware such as GPUs.

In this paper, we use the exact solver of Guthe and Thuerck (2021). This allows us to compute plug-in estimators at relatively large sample sizes $n = \Theta(10^4)$ and dimensions $d = \Theta(10^3)$ in a matter of seconds, *even while only using a single CPU core*. These sample sizes suffice for our all applications. Scaling to larger n would require caching (Guthe and Thuerck, 2021) or batching (Charlier et al., 2021) to overcome memory limitations, and would benefit from parallelism to reduce the computing time.

5.2.2 Statistical aspects

We turn to the statistical properties of Wasserstein distance estimators. The plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ is consistent (Villani, 2009) and has a positive bias which decreases with the sample size n (see Appendix C.1.1):

$$\lim_{n\to\infty} \mathcal{W}_2^2(\mu_n, \nu_n) = \mathcal{W}_2^2(\mu, \nu) \text{ almost surely,}$$
$$\forall n: \ \mathbb{E}\left[\mathcal{W}_2^2(\mu_n, \nu_n)\right] \geq \mathbb{E}\left[\mathcal{W}_2^2(\mu_{n+1}, \nu_{n+1})\right] \geq \mathcal{W}_2^2(\mu, \nu).$$

To make further progress, we separately impose two standard assumptions from the literature:

(A1) The distributions μ, ν are supported in the same compact set of diameter at most 1.

(A2) The distributions μ, ν have connected support with negligible boundary. Additionally, there exists a $\delta > 0$ such that $\mathbb{E}_\mu\left[\|X\|^{4+\delta}\right] < \infty$ and $\mathbb{E}_\nu\left[\|Y\|^{4+\delta}\right] < \infty$.

Under Assumption **(A1)**, the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ concentrates around its mean exponentially, and has an L_1 rate of convergence that decays with the dimension d (Fournier and Guillin, 2015; Weed and Bach, 2019; Chizat et al., 2020):

$$\forall \varepsilon \geq 0: \ \mathbb{P}\left(\left|\mathcal{W}_2^2(\mu_n, \nu_n) - \mathbb{E}\left[\mathcal{W}_2^2(\mu_n, \nu_n)\right]\right| \geq \varepsilon\right) \leq 2\exp(-n\varepsilon^2),$$
$$\forall d \geq 5: \ \mathbb{E}\left[\left|\mathcal{W}_2^2(\mu_n, \nu_n) - \mathcal{W}_2^2(\mu, \nu)\right|\right] \lesssim n^{-2/d},$$

where $\lesssim$ hides constants that do not depend on n. The rate of convergence also holds in the unbounded setting (Staudt and Hundrieser, 2024), and is furthermore minimax optimal (Hütter and Rigollet, 2021). Although smoother estimators can achieve better rates under stronger assumptions, they also require much greater computational expense (Hütter and Rigollet, 2021; Deb et al., 2021).

Under Assumption **(A2)**, the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ satisfies a central limit

theorem (CLT; del Barrio and Loubes, 2019; del Barrio et al., 2024). As $n \to \infty$,

$$\sqrt{n}\left\{\mathcal{W}_2^2(\mu_n, \nu_n) - \mathbb{E}\left[\mathcal{W}_2^2(\mu_n, \nu_n)\right]\right\} \Longrightarrow \mathcal{N}_1\left(0, \mathrm{Var}\left\{\phi_{\mu,\nu}(X) + \psi_{\mu,\nu}(Y)\right\}\right), \quad (5.2.2)$$

where $X \sim \mu$ and $Y \sim \nu$ are independent. We can therefore view $\mathcal{W}_2^2(\mu_n, \nu_n)$ as estimating $\mathbb{E}[\mathcal{W}_2^2(\mu_n, \nu_n)]$ up to Gaussian error. We now benchmark several variance estimators that could be used to construct Gaussian confidence intervals for this quantity.

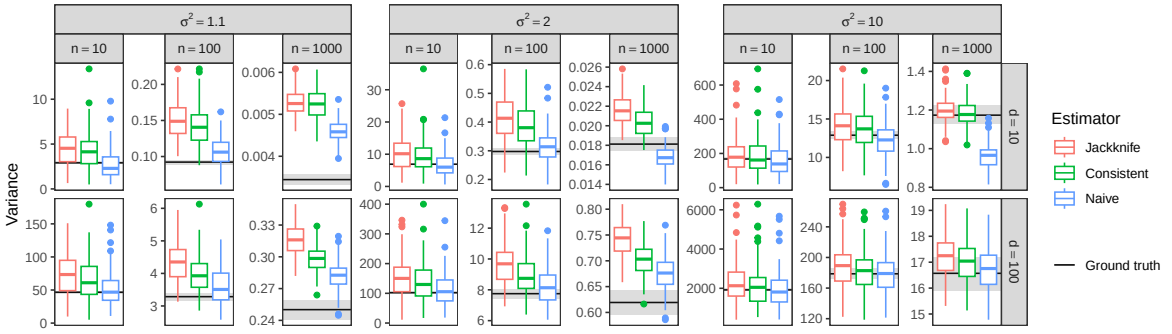

Figure 5.2.1: Variance estimation for $\mathcal{W}_2^2(\mu_n, \nu_n)$ with $\mu = \mathcal{N}_d(0_d, I_d)$, $\nu = \mathcal{N}_d(0_d, \sigma^2 I_d)$ and various methods and values of (σ^2, n, d). Unbiased estimates of the ground truth from 5000 replicates are shown with 95% bootstrap confidence intervals.

Variance estimation. We consider several ways of estimating the variance of the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$: (i) the jackknife (Efron and Stein, 1981), (ii) a consistent estimator based on the Kantorovich potentials (del Barrio et al., 2024) and (iii) a naive estimator.

Firstly, jackknife variance estimates are known to be conservative; in our context due to algorithmic considerations, the jackknife can be computed in $O(n^3)$ operations. (See Appendix C.2.1 for our procedure "Flapjack" based on Mills-Tettey et al., 2007.) Secondly, the central limit theorem (5.2.2) suggests the consistent variance estimator

$$\mathrm{Var}\left(\mathcal{W}_2^2(\mu_n, \nu_n)\right) \approx \frac{1}{n}\mathrm{Var}\left(\{\phi_{\mu_n,\nu_n}(X_i) + \psi_{\mu_n,\nu_n}(Y_i)\}_{i=1}^n\right),$$

where $\mathrm{Var}(\{x_i\}_{i=1}^n) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \frac{1}{n}\sum_{i=1}^n x_i)^2$ is the sample variance. This estimator

is appealing as optimal potentials $(\phi_{\mu_n,\nu_n}, \psi_{\mu_n,\nu_n})$ are available without additional computation with many solvers, including that of Guthe and Thuerck (2021). Finally, since $\mathcal{W}_2^2(\mu_n, \nu_n) = \frac{1}{n}\sum_{i=1}^n \|X_i - Y_{\sigma(i)}\|^2$ for an optimal permutation σ, one might naively consider the sample variance of the preceding average, implicitly assuming that all terms are independent. This estimator is also available for little added cost, but is inconsistent.

We compare the three methods in Figure 5.2.1: we prefer the consistent estimator (ii) as it is slightly conservative and quick to compute. All variance estimators have substantial positive biases as $\nu \Rightarrow \mu$, because in this regime $\phi_{\mu,\nu}, \psi_{\mu,\nu} \to 0$ and therefore the asymptotics (5.2.2) break down to a point mass δ_0.

5.2.3 Tractable scenarios

Certain structural conditions ease the computational and statistical challenges in estimating Wasserstein distances. For Gaussians, it holds that

$$\mathcal{W}_2^2(\mathcal{N}_d(m_\mu, \Sigma_\mu), \mathcal{N}_d(m_\nu, \Sigma_\nu)) = \|m_\mu - m_\nu\|^2 + \mathrm{Tr}\left(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}\right),$$

where $\Sigma^{1/2}$ denotes the principal square-root of Σ. An estimator of $\mathcal{W}_2^2$ with favorable statistical properties (Rippl et al., 2016) can be obtained by plugging in estimated means and covariances, for $\Theta(n^2 d + d^3)$ overall cost. Similar considerations apply to compatible elliptical distributions, see Peyré and Cuturi (2019, Remarks 2.31-32).

For one-dimensional measures, it holds that $\mathcal{W}_2^2(\mu, \nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^2 \mathrm{d}u$ where (F_μ^{-1}, F_ν^{-1}) are the inverse-CDFs of (μ, ν) which need not be continuous. In this case, the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ has favorable statistical properties (Bobkov and Ledoux, 2019). It is also fast to compute, requiring the $O(n \log n)$ sorting of the two samples; the Kantorovich potentials can be recovered in $\Theta(n)$ operations (Sejourne et al., 2022, Algorithm 3). Similar considerations apply to product measures, due to

tensorization: $\mathcal{W}_2^2(\otimes_{i=1}^d \mu^i, \otimes_{i=1}^d \nu^i) = \sum_{i=1}^d \mathcal{W}_2^2(\mu^i, \nu^i)$.

Gelbrich (1990) and the tensorization of the squared Euclidean metric provide the tractable lower bound

$$\mathcal{W}_2^2\left(\mathcal{N}_d(m_\mu, \Sigma_\mu), \mathcal{N}_d(m_\nu, \Sigma_\nu)\right) \vee \mathcal{W}_2^2(\otimes_{i=1}^d \mu^i, \otimes_{i=1}^d \nu^i) \leq \mathcal{W}_2^2(\mu, \nu), \tag{5.2.3}$$

where now (m, Σ) denote expectations and covariances, and where superscripts denote coordinate-wise marginals. In Section 5.4, we make use of this lower bound; since its finite-sample estimators are positively biased and noisy, we use the jackknife to correct the bias (Miller, 1974) and to quantify the additional noise (Efron and Stein, 1981).

5.3 Centered plug-in estimators

In applications, it is important for estimators of $\mathcal{W}_2^2(\mu, \nu)$ to be informative in the regime $\nu \Rightarrow \mu$: in addition to distinguishing between measures, we want to be able to recognize when they are similar. Even in low-dimensional scenarios, the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ does not satisfy this criterion, because it has a large bias that decays slowly with n and becomes particularly apparent as $\nu \Rightarrow \mu$. Since the bias cannot be meaningfully reduced by increasing the sample size, we must obtain informative estimators by different means.

We propose to render plug-in estimators of $\mathcal{W}_2^2(\mu, \nu)$ informative by centering them. Formally, we assume that empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}_i}$, $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, $\bar{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{Y}_i}$ are available, based on independent samples $X_{1:n}, \bar{X}_{1:n} \sim \mu$ and $Y_{1:n}, \bar{Y}_{1:n} \sim \nu$. The new centered estimators are:

$$\begin{aligned} U(\bar{\mu}_n, \mu_n, \nu_n) &= \mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\bar{\mu}_n, \mu_n), \\ L(\bar{\mu}_n, \mu_n, \nu_n) &= \left[\mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}_n, \mu_n)\right]_+^2, \end{aligned}$$

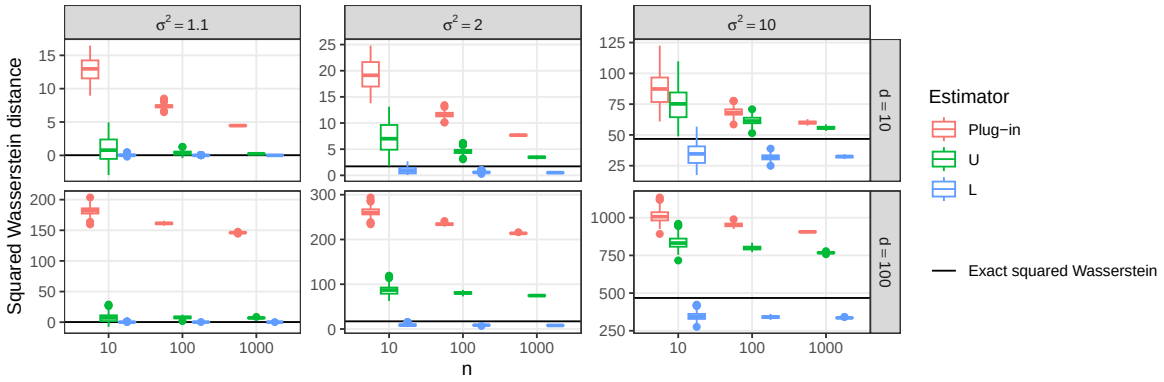

Figure 5.3.1: Comparison of plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ and proposed estimators $U(\bar{\mu}_n, \mu_n, \nu_n)$ and $L(\bar{\mu}_n, \mu_n, \nu_n)$, with $\mu = \mathcal{N}_d(0_d, I_d)$, $\nu = \mathcal{N}_d(0_d, \sigma^2 I_d)$ and various values of (σ^2, n, d).

where $[x]_{\pm}^2 = \text{sgn}(x)x^2$, i.e. L is the signed square of $\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) = \mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}_n, \mu_n)$.

The centering ensures that the proposed estimators are informative, as their expectations decrease to zero with $\mathcal{W}_2^2(\mu, \nu)$ for any finite sample size. More importantly, the proposed estimators can be viewed as complementary bounds on $\mathcal{W}_2^2(\mu, \nu)$: $U(\bar{\mu}_n, \mu_n, \nu_n)$ is an approximate upper bound when ν is overdispersed with respect to μ; $L(\bar{\mu}_n, \mu_n, \nu_n)$ is an approximate lower bound in general. We establish these properties, and we discuss suitable notions of overdispersion, in Section 5.3.1. Figure 5.3.1 illustrates these properties: notably, centering reduces the bias without increasing the variance.

Increasing the sample size n benefits the proposed estimators by decreasing the variance, reducing the bias and, as we shall see, further relaxing the conditions required for U to be conservative. Trading some of these benefits off for faster computation, $\{U, L\}$ could be replaced by sample averages computed at a lower sample size. We establish the statistical properties of the proposed estimators in Section 5.3.2, and we discuss uncertainty quantification in Section 5.3.3.

We conclude this introduction with two practical refinements of our methodology.

Hedging. Taking the maximum of two estimators, we can obtain more generally applicable upper bounds and tighter lower bounds, as with the pair

$$V(\mu_n, \nu_n, \bar{\mu}_n, \bar{\nu}_n) = U(\bar{\mu}_n, \mu_n, \nu_n) \vee U(\bar{\nu}_n, \nu_n, \mu_n) \text{ and } L(\bar{\mu}_n, \mu_n, \nu_n) \vee L(\bar{\nu}_n, \nu_n, \mu_n).$$

The first hedging strategy is particularly useful when *a priori* it is unclear which one of $\{\mu, \nu\}$ is more dispersed. Our experiments indicate that V is often conservative, even when it is used naively.

Variance reduction using couplings. When the sample generation can be controlled, positively correlating (μ_n, ν_n) can reduce the variances of $U(\bar{\mu}_n, \mu_n, \nu_n)$, $L(\bar{\mu}_n, \mu_n, \nu_n)$ and $V(\mu_n, \nu_n, \bar{\mu}_n, \bar{\nu}_n)$ with little effect to their biases. This technique can reduce the variance substantially, particularly when $\mathcal{W}_2^2(\mu, \nu)$ is small, see Section 5.4.

5.3.1 Analysis of the bias

We analyze the biases of the proposed estimators, showing that they are informative and providing conditions under which they can be viewed as approximate bounds. We recall that the minimal regularity Assumption **(A0)** applies.

Proposition 5.3.1 establishes that $\bar{L}$ is not conservative.

Proposition 5.3.1. *It holds that* $\mathbb{E}\left[\bar{L}(\bar{\mu}_n, \mu_n, \nu_n)\right]^2 = \mathbb{E}\left[\mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}_n, \mu_n)\right]^2 \leq \mathcal{W}_2^2(\mu, \nu)$.

Theorem 5.3.3 establishes properties of U. We show that an appropriate condition on the optimal transport map $T_{\nu,\mu}$ ensures that U is conservative, that U remains informative as $\nu \Rightarrow \mu$, and that U is location-free.

Definition 5.3.2 (Contractive optimal transport)**.** *We write* $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$*, and say that* ν *is contractively optimally transported to* μ*, if the optimal transport map* $T_{\nu,\mu}$ *is a contraction, that is it has Lipschitz constant* $\|T_{\nu,\mu}\|_{\text{Lip}} \leq 1$.

Theorem 5.3.3. *The following assertions hold:*

(i) *If* $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$, *then* $\mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] = \mathbb{E}\left[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\bar{\mu}_n, \mu_n)\right] \geq \mathcal{W}_2^2(\mu, \nu)$.

(ii) $\mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] \leq K(\mu, \nu)\, \mathcal{W}_2(\mu, \nu)$, *where* $K(\mu, \nu) = 3\mathbb{E}_\mu[\|X\|^2]^{1/2} + \mathbb{E}_\nu[\|Y\|^2]^{1/2}$.

(iii) $\mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] - \mathcal{W}_2^2(\mu, \nu)$ *is invariant to shifting the expectation of either* μ *or* ν.

We emphasize that the condition $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ of Theorem 5.3.3(i) is purely sufficient: it is what we use to formulate an otherwise generic result, which holds for all sample sizes n, all dimensions d, and does not impose structural assumptions on either measure μ or ν.

We interpret the condition $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ in Section 5.3.1; in Section 5.3.1, we demonstrate that the estimator U is in fact conservative much more generally. Since the inequality $\mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] \geq \mathcal{W}_2^2(\mu, \nu)$ is location-free, its validity clearly only depends on how the dispersions of μ and ν are related. For U to be an overestimate, the correct relation turns out to be that of overdispersion.

Remark 5.3.4: Brenier's theorem states that $T_{\nu,\mu} = \nabla\varphi_{\nu,\mu}$ for a convex $\varphi_{\nu,\mu}$. The condition of Theorem 5.3.3(i) is the global Hessian bound $\nabla^2\varphi_{\nu,\mu} \preceq I_d$ and resembles conditions used by recent computational (Paty et al., 2020) and theoretical (Hütter and Rigollet, 2021; Deb et al., 2021; Manole et al., 2024) work. After finalizing a preliminary version of this manuscript, we became aware of an independently derived result from a preprint version of Manole et al. (2024) that is similar to Theorem 5.3.3(i). We use our result for different purposes.

Interpreting contractive optimal transport

The condition $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ is location-free. This hints at a connection between $\overset{\text{COT}}{\rightsquigarrow}$ and stochastic orderings (Shaked and Shanthikumar, 2007), which we now discuss.

For one-dimensional measures, the univariate dispersive ordering $\nu \geq_{\text{disp}} \mu$ (Shaked, 1982) requires the quantiles of ν to lie further apart than the corresponding quantiles of μ. The condition $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ coincides with $\nu \geq_{\text{disp}} \mu$, because the optimal transport map $T_{\nu,\mu} = F_\mu^{-1} \circ F_\nu$ maps between the corresponding quantiles of ν and μ. In general, $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ implies the SD-ordering of Giovagnoli and Wynn (1995), which requires the existence of a contractive map transporting ν to μ. However, the SD-ordering does not provide a meaningful way of distinguishing between measures: for instance, μ and ν are equal under this ordering whenever they differ by a rotation, yet $\mathcal{W}_2^2(\mu, \nu)$ could be arbitrarily large.

We draw further connections between $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ and stochastic orderings under structural assumptions.

Proposition 5.3.5. *The following assertions hold:*

(i) For Gaussians, $\mathcal{N}_d(m_\nu, \Sigma_\nu) \overset{\text{COT}}{\rightsquigarrow} \mathcal{N}_d(m_\mu, \Sigma_\mu)$ if and only if $\Sigma_\nu \succeq \Sigma_\mu$, where $\succeq$ is the Loewner order.

(ii) For spherically symmetric measures, $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ if and only if the same relation holds between the distributions of their radial components.

(iii) For product measures, $(\otimes_{i=1}^d \nu^i) \overset{\text{COT}}{\rightsquigarrow} (\otimes_{i=1}^d \mu^i)$ if and only if $\nu^i \overset{\text{COT}}{\rightsquigarrow} \mu^i$ for all i.

(iv) If $\nu(x) \propto \exp(-N(x))$ and $\mu(x) \propto \exp(-M(x))$ with twice differentiable N, M with convex support, and if $\nabla^2 N \preceq A \preceq \nabla^2 M$ holds point-wise for a fixed positive definite matrix A, then $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$.

Overall, we view $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ as a global overdispersion condition: ν must be a shifted version of μ that is more spread-out in all directions. In addition to providing key intuition, this condition suggests that the estimators could be useful to assess the quality of Bayesian computation methods, where over- and underdispersion is pervasive, see Sections 5.4 and 5.5.

We conjecture that $\overset{\text{COT}}{\rightsquigarrow}$ does not define a partial order in general, and leave this as an open problem.

When is U conservative in practice?

We investigate the conditions required for U to be conservative in practice. We begin with a sharp characterization of the small-n case, see Appendix C.1.2.

Example 5.3.6 ($n = 1$)**:** The inequality $\mathbb{E}\left[U(\bar{\mu}_1, \mu_1, \nu_1)\right] \geq \mathcal{W}_2^2(\mu, \nu)$ is equivalent to

$$\sup_{(X,Y)\sim(\mu,\nu)} \mathrm{Tr}(\mathrm{Cov}(X, Y)) \geq \mathrm{Tr}(\mathrm{Var}(X)), \text{ denoted by } \nu \overset{\text{PCA}}{\rightsquigarrow} \mu.$$

Intuitively, *ν is more dispersed than μ, averaged along the principal components of μ.*

In particular, $\overset{\text{PCA}}{\rightsquigarrow}$ is partially closed under mixtures. Furthermore, $\nu \overset{\text{PCA}}{\rightsquigarrow} \mu$ holds under the convex order (Strassen, 1965), which provides the intuition that it suffices for ν to be a diffuse version of μ.

For large n, one might expect consistency to weaken the conditions under which U is conservative. However, the challenge in obtaining the exact expression of the bias to first order in n precludes a general analysis. Instead, we derive a sharp result in the one-dimensional case, see Appendix C.1.2.

Example 5.3.7 ($d = 1$)**:** Under regularity conditions, in dimension $d = 1$ it holds that

$$\lim_{n\to\infty} n\left(\mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] - \mathcal{W}_2^2(\mu, \nu)\right) \geq 0 \text{ if and only if } J(\mu, \nu) \geq J(\mu, \mu),$$

where $J(\mu, \nu) = \int_0^1 u(1-u)(F_\mu^{-1})'(u)(F_\nu^{-1})'(u)\mathrm{d}u$. This condition is significantly milder than $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$, which asks for $(F_\nu^{-1})' \geq (F_\mu^{-1})'$ uniformly.

Examples 5.3.6 and 5.3.7 indicate that a partial overdispersion can also ensure that U is conservative. This recommends the estimator V for general use. Whether V is conservative depends on the compatibility of the measures: the centering term of V may

over-correct when most of the masses of μ and ν lie in directions orthogonal to each other. In practice, the compatibility of the measures can be checked using a principal component analysis.

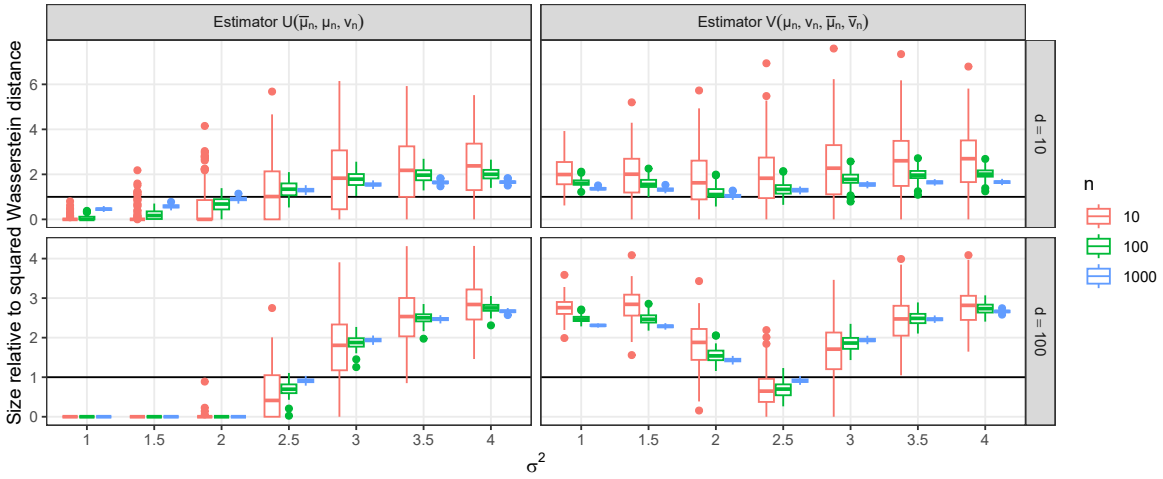

Figure 5.3.2: Robustness of proposed estimators $\{U, V\}$ to the degree of overdispersion, with $\mu = \mathcal{N}_d(0_d, \mathrm{diag}(1,4) \otimes I_{d/2})$ and $\nu = \mathcal{N}_d(0_d, \sigma^2 I_d)$ and various (σ^2, n, d). The relation $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$ holds for $\sigma^2 = 4$ and $\nu \overset{\text{PCA}}{\rightsquigarrow} \mu$ holds for $\sigma^2 \geq 2.87$ (resp. $\mu \overset{\text{PCA}}{\rightsquigarrow} \nu$ for $\sigma^2 \leq 2.25$ and $\mu \overset{\text{COT}}{\rightsquigarrow} \nu$ for $\sigma^2 = 1$). Negative estimates are set to zero.

Figure 5.3.2 illustrates that the proposed estimators $\{U, V\}$ are robust: they are conservative under relatively weak forms of overdispersion. We see that $U(\bar{\mu}_n, \mu_n, \nu_n)$ is conservative as long as ν is more dispersed than μ on average. That it is not conservative when ν is significantly less dispersed than μ should not be concerning to the reader: in practice, one would have swapped the roles of the measures and used the estimator $U(\bar{\nu}_n, \nu_n, \mu_n)$ instead. This effectively amounts to using the estimator V, which at the largest sample size is sensible throughout the considered scenario.

5.3.2 Statistical properties

We study the statistical properties of the proposed estimators. It is clear that they are consistent; they additionally inherit the concentration and near-minimax rate of convergence of the plug-in estimators they are composed of.

Theorem 5.3.8. *Under Assumption **(A1)**, it holds that*

$$\forall \varepsilon \geq 0: \ \mathbb{P}\left(|U(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right]| \geq \varepsilon\right) \leq 2\exp\left(-n\varepsilon^2/3\right),$$
$$\forall \varepsilon \geq 0: \ \mathbb{P}\left(\left|\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[\bar{L}(\bar{\mu}_n, \mu_n, \nu_n)\right]\right| \geq \varepsilon\right) \leq 2\exp\left(-n\varepsilon^4/32\right).$$

Theorem 5.3.9. *Let $\mu \neq \nu$. Under Assumption **(A1)**, it holds that*

$$\forall d \geq 5: \ \mathbb{E}\left[\left|U(\bar{\mu}_n, \mu_n, \nu_n) - \mathcal{W}_2^2(\mu, \nu)\right|\right] \lesssim n^{-2/d}, \quad \mathbb{E}\left[\left|\mathcal{W}_2(\mu, \nu) - \bar{L}(\bar{\mu}_n, \mu_n, \nu_n)\right|\right] \asymp n^{-1/d},$$

where $\asymp$ denotes decay at the exact rate.

As a consequence, the proposed estimators are high-probability bounds as soon as they are bounds in expectation. We emphasize that this does not require the sufficient condition $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$: in Corollary 5.3.10, we ask for $U(\bar{\mu}_n, \mu_n, \nu_n)$ to be positively biased by an amount which decays in n at the rate of Theorem 5.3.9.

Corollary 5.3.10. *Let $\mu \neq \nu$. Under Assumption **(A1)**, it holds that*

$$\forall d \geq 5: \ \mathbb{P}\left(L(\bar{\mu}_n, \mu_n, \nu_n) \leq \mathcal{W}_2^2(\mu, \nu)\right) = \mathbb{P}\left(\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) \leq \mathcal{W}_2(\mu, \nu)\right) \geq 1 - \exp\left(-C_1 n^{1-4/d}\right).$$

If additionally $\mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] - \mathcal{W}_2^2(\mu, \nu) \gtrsim n^{-2/d}$, it holds that

$$\forall d \geq 5: \ \mathbb{P}\left(U(\bar{\mu}_n, \mu_n, \nu_n) \geq \mathcal{W}_2^2(\mu, \nu)\right) \geq 1 - \exp\left(-C_2 n^{1-4/d}\right).$$

The constants $C_1, C_2 > 0$ only depend on the measures μ, ν and on the dimension d.

Similar properties can be shown for V, with the hedging ensuring that this estimator is a high-probability bound on $\mathcal{W}_2^2(\mu, \nu)$ under weaker conditions. We avoid further technical details.

5.3.3 Uncertainty quantification

We describe how to quantify the uncertainty of the proposed estimators based on their asymptotic distributions. The estimator U obeys a Gaussian CLT, as a direct consequence of del Barrio et al. (2024, Theorem 4.10) and Slutsky's theorem, which we state without proof.

Theorem 5.3.11. *Under Assumption **(A2)** it holds that, as $n \to \infty$,*

$$\sqrt{n}\left(U(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right]\right) \Longrightarrow \mathcal{N}_1\left(0, \sigma^2\right) \quad \textit{and} \quad \lim_{n\to\infty} n \operatorname{Var}\left\{U(\bar{\mu}_n, \mu_n, \nu_n)\right\} = \sigma^2,$$

where $\sigma^2 = \operatorname{Var}\{\phi_{\mu,\nu}(X) + \psi_{\mu,\nu}(Y)\}$ under independent $X \sim \mu$ and $Y \sim \nu$.

Formal results for $\bar{L}$ are more challenging because $\mathcal{W}_2(\bar{\mu}_n, \mu_n)$ lacks a satisfactory limiting theory (del Barrio et al., 2024), but experiments indicate that $\bar{L}$ is approximately Gaussian even for small n.

To quantify the variability of $\{U, \bar{L}\}$, we therefore use Gaussian confidence intervals. For $L = [\bar{L}]_{\pm}^2$, we transform by $[\cdot]_{\pm}^2$ the confidence interval for $\bar{L}$. For estimators like V that are formed as the maximum of two components, we use the confidence interval corresponding to the active component; the slight underestimation balances out with our conservative variance estimates, which we next describe.

The confidence intervals require variance estimates, we propose to use

$$\begin{aligned}
\operatorname{Var}(U) &\approx \frac{1}{n} \operatorname{Var}\left(\left\{\phi_{\bar{\mu}_n,\nu_n}(\bar{X}_i) + \psi_{\bar{\mu}_n,\nu_n}(Y_i) - \phi_{\bar{\mu}_n,\mu_n}(\bar{X}_i) - \psi_{\bar{\mu}_n,\mu_n}(X_i)\right\}_{i=1}^n\right), \\
\operatorname{Var}(\bar{L}) &\approx \frac{1}{n} \operatorname{Var}\left(\left\{\frac{\phi_{\bar{\mu}_n,\nu_n}(\bar{X}_i) + \psi_{\bar{\mu}_n,\nu_n}(Y_i)}{2\,\mathcal{W}_2(\bar{\mu}_n, \nu_n)} - \frac{\phi_{\bar{\mu}_n,\mu_n}(\bar{X}_i) + \psi_{\bar{\mu}_n,\mu_n}(X_i)}{2\,\mathcal{W}_2(\bar{\mu}_n, \mu_n)}\right\}_{i=1}^n\right),
\end{aligned}$$

justified in turn by Theorem 5.3.11 and an approximate delta method for $\bar{L}$ (detailed in Appendix C.2.2). Figure 5.3.3 compares the proposed consistent variance estimator of U with the jackknife, which is conservative and available with an additional $O(n^3)$ computation using the Flapjack algorithm (Appendix C.2.1). We prefer the consistent

estimator for its smaller positive bias and lower computing cost. Similar considerations hold for the variance estimator of $\bar{L}$.

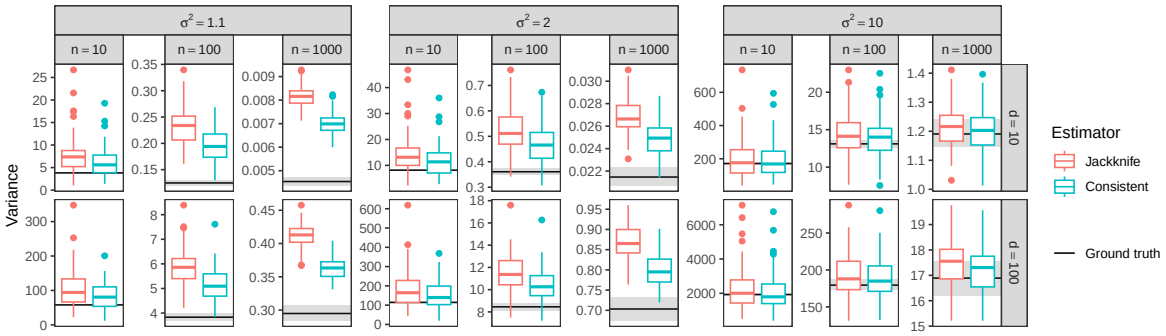

Figure 5.3.3: Variance estimates for $U(\bar{\mu}_n, \mu_n, \nu_n)$ with $\mu = \mathcal{N}_d(0_d, I_d)$, $\nu = \mathcal{N}_d(0_d, \sigma^2 I_d)$ and various methods and values of (σ^2, n, d). Unbiased estimates of the ground truth from 5000 replicates are shown with 95% bootstrap confidence intervals.

The variance estimates also remain valid when the pairs (X_i, Y_i) are sampled i.i.d. from any coupling of (μ, ν), so the confidence intervals correctly account for the variance reduction afforded by positively correlating (μ_n, ν_n). As we explain in Appendix C.2.3, we can estimate this reduction in variance without additional simulation. Finally, these uncertainty quantification methods can be generalized to correlated samples and to averages of plug-in estimators, see Appendices C.2.3 and C.2.4. These generalizations will prove useful in the applications of Sections 5.4 and 5.5.

5.4 Assessing the quality of approximate inference methods

Reliably assessing the quality of approximate Bayesian inference methods is one of the grand challenges of Bayesian computation (Bhattacharya et al., 2024), a problem that is of interest both to the researchers developing such methods, as well as to the practitioners using them. We propose here to estimate the squared Wasserstein distance $\mathcal{W}_2^2(\mu, \nu)$ of approximations ν to exact models μ with the centered estimators of Section 5.3.

5.4.1 Methodology

We advocate using MCMC to sample from the model μ and the approximation ν, in the following way. We sample i.i.d. μ-invariant Markov chains $(X_k^{(t)})_{t=0}^{B_\mu+T_\mu(I-1)}$ and ν-invariant chains $(Y_k^{(t)})_{t=0}^{B_\nu+T_\nu(I-1)}$ for $k \in [2K]$. We discard, respectively, $\{B_\mu, B_\nu\}$ iterations as burn in, and thin the remainder of each chain by factors of $\{T_\mu, T_\nu\}$ to provide the empirical measures

$$\mu_n = \frac{1}{KI}\sum_{k=1}^{K}\sum_{i=0}^{I-1}\delta_{X_k^{(B_\mu+T_\mu i)}}, \quad \bar{\mu}_n = \frac{1}{KI}\sum_{k=K+1}^{2K}\sum_{i=0}^{I-1}\delta_{X_k^{(B_\mu+T_\mu i)}},$$
$$\nu_n = \frac{1}{KI}\sum_{k=1}^{K}\sum_{i=0}^{I-1}\delta_{Y_k^{(B_\nu+T_\nu i)}}, \quad \bar{\nu}_n = \frac{1}{KI}\sum_{k=K+1}^{2K}\sum_{i=0}^{I-1}\delta_{Y_k^{(B_\nu+T_\nu i)}},$$

each with $n = KI$ samples. We modify the confidence intervals to account for within-chain sample dependence in Appendix C.2.3.

Our parameter guidelines are motivated by the insight that the biases of the proposed estimators primarily depend on the smallest of the effective sample sizes (ESSes; e.g. Vats et al., 2019) within the empirical measures $\{\mu_n, \nu_n\}$. We therefore recommend setting the thinning factors $\{T_\mu, T_\nu\}$ such that the ESSes are roughly equal,[1] and increasing $\{K, I\}$ until a target ESS is attained. The burn-ins $\{B_\mu, B_\nu\}$ should be set based on estimates of the rate of convergence, see Section 5.5. Our experience is that the estimators are robust to small $\{T_\mu, T_\nu\}$.

To reduce the variance of estimators, we can induce a positive correlation between (μ_n, ν_n) by coupling the pairs $(X_k^{(t)}, Y_k^{(t)})$ and setting $(B_\mu, T_\mu) = (B_\nu, T_\nu)$. We consider various practical coupling strategies based on common random numbers (CRNs) in Section 5.4.4.

[1] That is, we recommend performing more iterations with the slowest-mixing chain.

5.4.2 Approximate inference methods

We discuss several common types of approximate inference methods ν, focusing on how their variabilities relate to that of the exact model μ. The general trend is that approximate inference methods tend to be over- or underdispersed versions of the exact model, and so we typically expect the estimators of Section 5.3 to reliably bound the squared Wasserstein distance $\mathcal{W}_2^2(\mu, \nu)$.

Laplace approximations. A Laplace approximation ν is the best Gaussian fit around a mode of the density of the true model μ. Since Laplace approximations are local, they typically underestimate the variability, particularly if μ has heavier-than-Gaussian tails, or it has multiple modes. Other types of localized approximations can similarly be expected to underestimate the variability in the true model.

Variational approximations. Variational inference (VI; Blei et al., 2017) uses optimization to fit the approximation ν. The approximating family is often Gaussian. The objective is typically the *exclusive* (or *reverse*) Kullback-Leibler (KL) divergence $\mathrm{KL}(\cdot\|\mu)$, which tends to produce local approximations which underestimate the true variability (Wang and Titterington, 2005). Conversely, expectation propagation (EP; Minka, 2001) is an algorithm that optimizes for the *inclusive* (or *forward*) KL divergence $\mathrm{KL}(\mu\|\cdot)$. EP appears to have two regimes, either globally overestimating the true variability or globally underestimating it (Dehaene and Barthelmé, 2018).

Approximate MCMC algorithms. Certain gradient-based unadjusted MCMC algorithms, such as the unadjusted Langevin algorithm (ULA; Roberts and Tweedie, 1996a) and the OBABO discretization of the *underdamped* (or *kinetic*) Langevin diffusion (e.g. Monmarché, 2021), tend to have stationary distributions ν that are overdispersed versions of the exact target μ. We verify this analytically for Gaussian targets μ.

Proposition 5.4.1. *The stationary distribution ν of an ULA or OBABO chain targeting a Gaussian μ satisfies $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$.*

Stochastic gradient MCMC algorithms (Ma et al., 2015) are gradient-based unadjusted MCMC algorithms where the gradient is replaced by an unbiased estimate; they are popular in tall-data applications. The additional noise typically causes the stationary distribution ν of a stochastic gradient MCMC algorithm to be an overdispersed version of the target μ (Nemeth and Fearnhead, 2021).

Exact and approximate Gibbs samplers for high-dimensional linear regression models with horseshoe priors (Carvalho et al., 2010) were developed in Johndrow et al. (2020); these samplers were later extended to more general half-t priors in Biswas et al. (2022); Biswas and Mackey (2024). We explain why these approximate Gibbs samplers generate overdispersed versions ν of the exact target μ in Appendix C.4.2.

Approximate Bayesian computation. Approximate Bayesian computation (ABC) methods perform Bayesian inference using noisy surrogate versions of the likelihood. Due to this noise, the ABC posterior is typically more dispersed than the true posterior (Sisson et al., 2018).

5.4.3 Related methods

Biswas and Mackey (2024) use couplings to assess the quality of approximate sampling methods. The idea is to sample a pair of coupled Markov chains $(X^{(t)}, Y^{(t)})_{t\geq 0}$ with kernels (P, Q) and stationary distributions (μ, ν). In the idealized setting where the chains are stationary, for all (B, I) it holds that

$$\mathcal{W}_2^2(\mu, \nu) \leq \mathbb{E}\left[\frac{1}{I}\sum_{t=B}^{B+I-1} \left\|X^{(t)} - Y^{(t)}\right\|^2\right]. \tag{5.4.1}$$

In practice, we discard the first B iterations as burn-in, and we estimate the coupling bound by averaging over K replicates.

The method of Biswas and Mackey (2024) can only perform well if (P, Q) are similar in a uniform sense. It additionally requires the user to carefully design a contractive coupling of (P, Q). As we demonstrate in Section 5.4.4, sensible couplings of (P, Q) can still produce loose bounds, whereas any coupling that positively correlates the chains can reduce the variance of our proposed estimators.

Huggins et al. (2020) derive computable upper bounds on $\mathcal{W}_2^2(\mu, \nu)$ based on a series of worst-case theoretical bounds and importance sampling using ν as a proposal. Dobson et al. (2021) propose a coupling-based upper bound that is similar to (5.4.1), but incurs an additional term related to the rate of contraction of the kernel Q. Because Biswas and Mackey (2024, Section 3.4) demonstrates that the method of Huggins et al. (2020) deteriorates rapidly with increasing dimension and that the method of Dobson et al. (2021) produces a looser bound than (5.4.1), we do not compare with these methods in the sequel.

5.4.4 Numerical illustrations

We illustrate the proposed methodology with various applications, comparing our method with the coupling-based bound of Biswas and Mackey (2024) and assessing the sharpness of all estimates against the tractable lower bound (5.2.3). Because the squared Wasserstein distance does not have a global upper bound, we instead provide the trace of the covariance $\mathrm{Tr}(\mathrm{Cov}_\mu(X))$ as a measure of scale, that intuitively indicates a poor approximation ν. (Since $\mathcal{W}_2^2(\mu, \nu) = \mathrm{Tr}(\mathrm{Cov}_\mu(X))$ when ν is a Dirac mass centered at the mean of μ.) We defer additional experimental details to Appendix C.6.2.

Asymptotic bias of unadjusted MCMC algorithms

We estimate the asymptotic bias of two unadjusted MCMC algorithms, ULA and the OBABO discretization of the underdamped Langevin diffusion. The algorithms target $\mu = \mathcal{N}_d(0_d, \Sigma_d)$ with $(\Sigma_d)_{ij} = 0.5^{|i-j|}$ and use spherical Gaussian proposals with standard deviation $h = d^{-1/6}$ in various dimensions d. The underdamped algorithm uses critical damping. This synthetic Gaussian setting presents us with a dual advantage: it allows us to compare estimators against the true squared Wasserstein distance, as well as to assess the sensitivity of estimators to the dynamics of each approximate MCMC algorithm, since both algorithms have identical Gaussian stationary distributions μ_h at identical step sizes h, see Appendix C.4.1.

We follow Biswas and Mackey (2024, Section 2.2) and couple each unadjusted algorithm with its Metropolis-adjusted counterpart by CRNs, ULA with the Metropolis-adjusted Langevin algorithm (MALA; Besag, 1994) and OBABO with the method of Horowitz (1991). We do not use the couplings to reduce the variance of the proposed estimators.

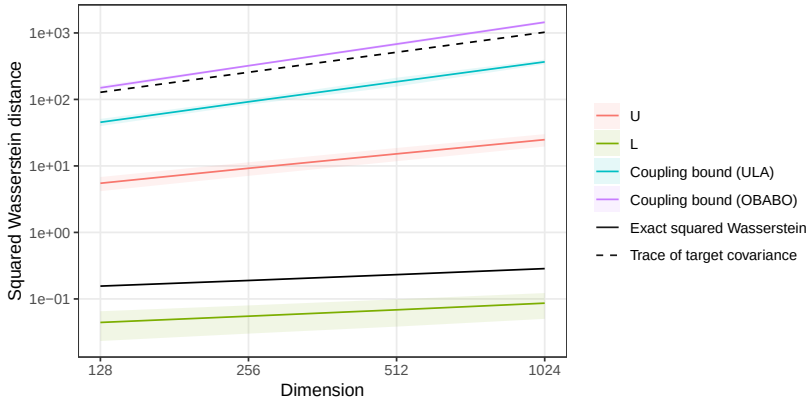

Figure 5.4.1: Asymptotic bias of unadjusted MCMC algorithms in increasing dimension, see Section 5.4.4 for details. The considered algorithms (ULA and OBABO) have identical stationary distributions. Solid lines represent empirical means, shaded areas represent two standard deviations.

Figure 5.4.1 displays estimates of the asymptotic bias $\mathcal{W}_2^2(\mu, \mu_h)$. The proposed estimators $\{U, L\}$ reveal that the asymptotic bias is small even in high dimensions and

provide identical results for both approximate algorithms. In contrast, the coupling bound is at least an order of magnitude looser and performs significantly worse for OBABO than it does for ULA. We estimate that the coupling of ULA (resp. OBABO) could have reduced the variance of U by a factor of $2\times$ (resp. $1.1\times$).

This experiment highlights a limitation of the coupling bound. Although seemingly a reasonable default, coupling unadjusted MCMC algorithms with their Metropolized counterparts turns out to only be effective when the acceptance rate of the Metropolized algorithm is extremely high, i.e. the mixing is poor. For ULA coupled with MALA, we observe that the squared-distance between the chains increases by $\Theta(h^2 d)$ upon rejection in MALA, whereas the chains contract exponentially at rate $\Theta(h^2)$ upon acceptance. The equilibrium therefore lies at $\Theta(d)$ times the rejection rate, which is typically much larger than $\mathcal{W}_2^2(\mu, \mu_h) = \Theta(h^2 d)$ (Durmus and Moulines, 2019). For OBABO coupled with the Horowitz method, the situation worsens because the Horowitz method reverses direction upon rejection; the persistent momentum then causes the chains to move away from each other for several iterations. In this experiment, the step size $h = d^{-1/6}$ ensures a small asymptotic bias and a relatively high acceptance rate of $\approx 70\%$, yet the coupling bound is still loose.

Approximate inference for tall data

We assess the quality of various approximate inference methods for tall datasets (Bardenet et al., 2017), where the number of observations is much larger than the number of covariates. We consider stochastic gradient Langevin dynamics (SGLD; Welling and Teh, 2011) subsampling 10% of the data per iteration, SGLD with control variates (SGLD-cv; Baker et al., 2019) subsampling 1% of the data per iteration, the Laplace approximation, and full-rank Gaussian VI (Kucukelbir et al., 2017). We compare these methods on Bayesian logistic regression models with the following datasets: Pima Indians (Smith et al., 1988; 768 observations, 8 covariates) and DS1 life sciences (Komarek

and Moore, 2003; 26733 observations, 10 covariates).

For parity across methods, and to reduce the variance, we compute all Wasserstein distance estimators based on the same coupled pairs of Markov chains targeting (μ, ν). We target μ and optimization-based approximations ν with MALA and use CRN couplings, as in Biswas and Mackey (2024, Section 4.1). To make the implementation generic across different approximations, we use the proposed estimator V based on splitting the available sample.

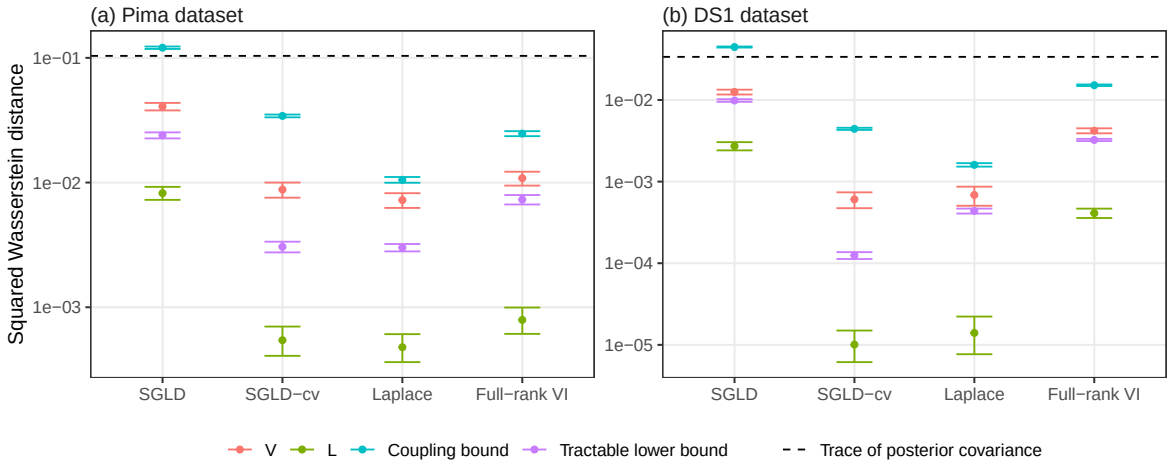

Figure 5.4.2: Quality of various approximate inference methods applied to Bayesian logistic regression models with various datasets, see Section 5.4.4 for details. Error bars represent approximate 95% confidence intervals.

Figure 5.4.2 displays estimates of the asymptotic bias of each approximate inference method. Consistent with the concentration of the posterior due to Bernstein-von-Mises limit, SGLD-cv and the Laplace approximation have the smallest biases. In contrast, SGLD overestimates the posterior variance due to noisy gradient estimates, whereas VI underestimates the posterior variance.

The proposed estimators accurately quantify the asymptotic bias: V is often remarkably close to the tractable lower bound (5.2.3), which we expect to be tight due to the proximity of the model to its Bernstein-von-Mises limit. The coupling bound is uniformly looser: similarly to Section 5.4.4, the issue is partly caused by the challenge in

coupling MCMC algorithms that involve accept-reject decisions. We estimate that the coupling reduced the variance of $\{V, L\}$ by factors of up to 1.6× for the Pima dataset and 2.2× for the DS1 dataset.

Finally, sampling from the exact model μ with MALA becomes a significant bottleneck for datasets larger than the ones considered here. The proposed estimators can scale to larger datasets by amortizing the cost of sampling from μ using recent advances in exact MCMC algorithms based on subsampling (e.g. Fearnhead et al., 2018; Prado et al., 2024). However, because these algorithms have complex dynamics or parametrizations, it is less clear how one can couple them effectively.

Approximate sampling for high-dimensional Bayesian linear regression

We consider a high-dimensional Bayesian linear regression model with half-t(η) priors. Johndrow et al. (2020) developed exact and approximate Gibbs samplers for the case $\eta = 1$, corresponding to the horseshoe prior; Biswas et al. (2022) and Biswas and Mackey (2024) extended these samplers to degrees of freedom $\eta > 1$. We assess the asymptotic bias of such approximate Gibbs samplers with $\eta = 2$ on the Riboflavin dataset (Bühlmann et al., 2014; 71 observations, 4088 covariates).

This is a challenging scenario: the distributions we compare are high-dimensional, multimodal and heavy-tailed. This setting is also ideal for the coupling bound, since considerable effort has been spent on devising effective couplings for these samplers (Biswas et al., 2022; Biswas and Mackey, 2024). We follow Biswas and Mackey (2024) and use CRN couplings between the approximate and exact Gibbs samplers. We also use the couplings to reduce the variance of our proposed estimators. Since we know that the exact model is the less dispersed distribution, we draw an additional set of samples from it to use throughout the experiment, and we use the estimator U.

Figure 5.4.3 displays estimates of the asymptotic bias $\mathcal{W}_2^2(\mu, \mu_\varepsilon)$ against the parameter $\varepsilon \geq 0$ that controls the quality of the approximation, where μ is the exact and μ_ε

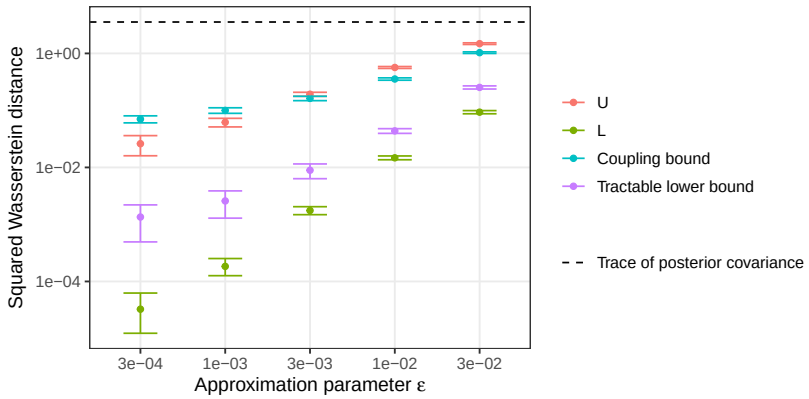

Figure 5.4.3: Asymptotic bias of approximate Gibbs sampler for high-dimensional linear regression model with half-t(2) prior, see Section 5.4.4. Error bars represent approximate 95% confidence intervals. The estimate of the tractable lower bound (5.2.3) has a considerable positive bias for small ε.

the approximate posterior marginal of the regression coefficients. The figure suggests that $\mathcal{W}_2^2(\mu, \mu_\varepsilon) \approx \Theta(\varepsilon)$, which is consistent with the results of Johndrow et al. (2020) for the case $\eta = 1$ and confirms that their recommended default of setting ε as the reciprocal of the number of covariates ($\approx 2.5 \times 10^{-4}$ here) achieves a small asymptotic bias.

The experiment illustrates that the proposed estimators can be effective in complex problems of very high dimensionality. The estimator U is competitive with the coupling bound and outperforms it for smaller values of ε. In fact, whereas our proposed estimators are guaranteed to be informative for all ε, the coupling bound becomes uninformative as $\varepsilon \to 0$ because the CRN coupling is not uniformly contractive when $\varepsilon = 0$ (Biswas et al., 2022, Appendix B). Nevertheless, because it reduces the variance of $\{U, L\}$ by a factor of 22$\times$ for the smallest ε, the coupling appears crucial for controlling the variance of the proposed estimators when the true Wasserstein distance is small.

5.5 Assessing the convergence of MCMC algorithms

MCMC algorithms undergo an initial warm-up phase wherein the time-marginals $(\pi^{(t)})_{t\geq 0}$ converge towards the stationary distribution $\pi^{(\infty)}$. Assessing how quickly MCMC algorithms converge is of great importance to the researchers developing such methods, as well as to the practitioners using them. We propose here to estimate the convergence in squared Wasserstein distance $\mathcal{W}_2^2(\pi^{(\infty)}, \pi^{(t)})$, by post-processing the output of several parallel Markov chains using the estimators of Section 5.3.

5.5.1 Methodology

We simulate $2n$ replicate Markov chains up to a large time $T \gg 1$. We split the samples from $\pi^{(t)}$ into equally weighted empirical measures $\{\pi_n^{(t)}, \bar{\pi}_n^{(t)}\}$ for all $t \geq 0$. When $\pi^{(t)}$ is more dispersed than $\pi^{(T)}$, we estimate

$$L(\bar{\pi}_n^{(T)}, \pi_n^{(T)}, \pi_n^{(t)}) \lessapprox \mathcal{W}_2^2(\pi^{(T)}, \pi^{(t)}) \lessapprox U(\bar{\pi}_n^{(T)}, \pi_n^{(T)}, \pi_n^{(t)}). \tag{5.5.1}$$

Conversely, we estimate $L(\bar{\pi}_n^{(t)}, \pi_n^{(t)}, \pi_n^{(T)}) \lessapprox \mathcal{W}_2^2(\pi^{(T)}, \pi^{(t)}) \lessapprox U(\bar{\pi}_n^{(t)}, \pi_n^{(t)}, \pi_n^{(T)})$ when $\pi^{(t)}$ is less dispersed than $\pi^{(T)}$.

The standard practice in MCMC is to use overdispersed initializations. Because the time-marginals $\pi^{(t)}$ tend to gradually concentrate towards the stationary distribution $\pi^{(\infty)}$ when the initialization is overdispersed, see Section 5.5.2, in this setting we expect our estimators (5.5.1) to reliably bound the convergence. We describe in Appendix C.5.1 a reduced-variance methodology based on time-averaging that is tailored to overdispersed initializations.

We highlight three reasons why the proposed methodology is appealing. Firstly, the method closely approximates the true convergence rate, since by Theorem 5.3.3(ii) rates estimated by U are loose by at most a factor of two. Secondly, the method is plug-in, so its performance is unaffected by how complex the implementation or dynamics of

the MCMC kernel are. Finally, the method also applies to non-Markovian processes, so it can estimate the convergence of adaptive MCMC algorithms (Andrieu and Thoms, 2008). Competing methods lack one or more of these properties, see Section 5.5.3.

The method can however be vulnerable to issues of pseudo-convergence, since it assumes that the replicate MCMC runs have converged and become stationary within the computing budget, so that $\mathcal{W}_2^2(\pi^{(T)}, \pi^{(t)}) \approx \mathcal{W}_2^2(\pi^{(\infty)}, \pi^{(t)})$. Convergence diagnostics (e.g. Gorham and Mackey, 2017; Margossian et al., 2024) can help check stationarity in practice.

5.5.2 On MCMC with an overdispersed initialization

The guideline of choosing overdispersed initializations dates back to the early days of parallel MCMC (Gelman and Rubin, 1992). Decades of experience suggest that overdispersed initializations facilitate both exploration and convergence diagnosis, with the intuition being that such initializations cause the time-marginals $\pi^{(t)}$ to concentrate towards $\pi^{(\infty)}$ over time. We verify this intuition in a stylized setting that is prototypical for many popular MCMC samplers.

Proposition 5.5.1. *Let $(\pi^{(t)})_{t\geq 0}$ be the time-marginals of a Gaussian AR(1) process with a Gaussian initialization $\pi^{(0)}$. If $\pi^{(0)} \overset{\text{COT}}{\rightsquigarrow} \pi^{(\infty)}$, then $\pi^{(t)} \overset{\text{COT}}{\rightsquigarrow} \pi^{(\infty)}$ for all $t \geq 0$.*

Remark 5.5.2: Proposition 5.5.1 directly applies to discretizations of the overdamped Langevin diffusion. An extension of Proposition 5.5.1 holds for the position component of discretizations of the underdamped Langevin diffusion. In a small step-size asymptotic limit (Bou-Rabee and Vanden-Eijnden, 2010), Proposition 5.5.1 applies to MALA and the method of Horowitz (1991), and similar insight (Roberts et al., 1997) can be expected to hold for the random walk Metropolis (RWM; Tierney, 1994) algorithm. Finally, Proposition 5.5.1 applies to deterministic scan Gibbs samplers (Roberts and Sahu, 1997), and overdispersion persists in the sense of $\pi^{(t)} \overset{\text{PCA}}{\rightsquigarrow} \pi^{(\infty)}$ for random scan Gibbs samplers. We provide verification in Appendix C.4.4.

For unimodal targets, Proposition 5.5.1 suggests that samplers initialized overdispersed should gradually concentrate towards their stationary distributions. Simulations with non-Gaussian unimodal targets in Appendix C.6.3 support this insight. For multimodal targets, simulations in Appendix C.6.3 suggest that the convergence happens in a similar way provided that the initialization is dispersed across all modes.

The choice of an appropriately overdispersed initialization should be guided by the target at hand. In Bayesian inference problems (e.g. Gelman et al., 2013), the prior is often a suitable initialization, because it tends to be less concentrated than the (target) posterior distribution. More generally, initializing from an overdispersed version of an approximation to the target is a sensible strategy: Gelman and Rubin (1992) use heavy-tailed mixtures centered at the target modes; Carpenter et al. (2017) use uniform distributions adapted to the length-scales of the target parameters.

5.5.3 Related methods

Biswas et al. (2019) use couplings to bound the convergence MCMC algorithms. Originally devised for 1-Wasserstein distances, we extend this method to p-Wasserstein distances of all orders $p \geq 1$ in Appendix C.5.2. With an appropriate choice of parameters, the method effectively amounts to repeatedly sampling coupled Markov chains $(\bar{X}^{(t)}, X^{(t)})_{t\geq 0}$ with initializations $(\bar{X}^{(0)}, X^{(0)}) \in \Gamma(\pi^{(\infty)}, \pi^{(0)})$ and marginal evolutions according to the Markov kernel of interest, then estimating the coupling inequality

$$\mathcal{W}_2^2(\pi^{(\infty)}, \pi^{(t)}) \leq \mathbb{E}\big[\|\bar{X}^{(t)} - X^{(t)}\|^2\big]. \tag{5.5.2}$$

using empirical averages. In our experiments, we estimate the idealized bound (5.5.2) based on independent initializations $(\bar{X}^{(0)}, X^{(0)})$.

Johnson (1996); Sixta et al. (2025) propose to estimate a looser version of the idealized bound (5.5.2) based on a rejection-sampling construction. Since this suffers

from the curse of dimensionality, we do not compare with it in the sequel.

Coupling-based methods require the user to design and implement couplings that contract the chains $(\bar{X}^{(t)}, X^{(t)})$ quickly over time. As we demonstrate in Section 5.5.4, the availability of effective couplings is case-specific, and couplings can be sensitive to the dynamics and the parametrization of the MCMC algorithm at hand. In particular, we will see that Metropolis accept-reject steps, which are ubiquitously used to devise asymptotically exact MCMC algorithms, complicate the design of effective couplings in high dimensions (see also Papp and Sherlock, 2025b).

5.5.4 Numerical illustrations

We illustrate the proposed methodology with various moderate- to high-dimensional applications. We focus on the case of overdispersed initializations and use the reduced-variance method of Appendix C.5.1. We compare our method against the coupling bound of Biswas et al. (2019), using state-of-the art couplings (e.g. Heng and Jacob, 2019; Jacob et al., 2020b; Monmarché, 2021) based on CRNs unless stated otherwise. As a default, we compute estimators based on $n = 1024$ replicates. We defer additional experimental details to Appendix C.6.3.

Synthetic examples

We consider synthetic examples with Gaussian target distributions. These allow us to directly assess the sharpness of our estimators against the exact squared Wasserstein distance $\mathcal{W}_2^2(\pi^{(\infty)}, \pi^{(t)})$.

Gibbs sampler. We target a periodic-boundary AR(1) process $\pi^{(\infty)} = \mathcal{N}_d(0_d, \Sigma_d)$ with autocorrelation $\rho = 0.95$ in dimension $d = 50$ with a systematic scan Gibbs sampler. We consider two initializations: (a) a fully overdispersed start $\pi^{(0)} = \mathcal{N}_d(0_d, 4\Sigma_d) \overset{\text{COT}}{\rightsquigarrow} \pi^{(\infty)}$; (b) a naive start $\pi^{(0)} = \mathcal{N}_d(0_d, \text{diag}(\Sigma_d)) \overset{\text{PCA}}{\not\rightsquigarrow} \pi^{(\infty)}$ representing a mean-field approximation

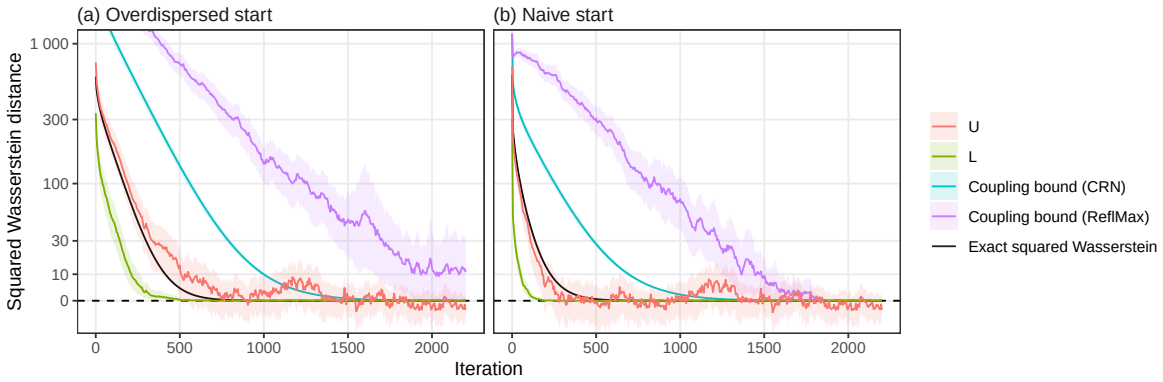

Figure 5.5.1: Convergence of a Gibbs sampler with various initializations, see Section 5.5.4 for details. Shaded areas represent approximate 95% confidence intervals. Recall that we estimate the idealized coupling bound (5.5.2) with infinite time-lag parameter.

to $\pi^{(\infty)}$.

Figure 5.5.1 displays estimates of the convergence of the Gibbs sampler with various methods. We see that the estimator U is conservative when the initialization is overdispersed and is robust to using naive initializations. Remarkably, in both cases, the true squared Wasserstein distance consistently falls within the confidence interval for U; we speculate that this relates to the target having a few very large principal components which dominate the overall contribution to the Wasserstein distance. The proposed estimator L provides a sensible companion lower bound to U. The sharpness of the coupling bound is highly dependent on the coupling used (coordinate-wise CRN or reflection-maximal, Jacob et al., 2020b), but even with the optimal Markovian CRN coupling this bound is relatively loose compared to the estimator U.

Mixing time of Langevin algorithms. We study the mixing time of MCMC algorithms based on the over- and underdamped Langevin diffusions. For each diffusion, we consider a discretization and its Metropolis-adjusted version: ULA and MALA in the overdamped case, the OBABO discretization and the Horowitz (1991) method in the underdamped case.

We revisit the setting of Section 5.4.4, targeting $\pi = \mathcal{N}_d(0_d, \Sigma_d)$ with $(\Sigma_d)_{ij} = 0.5^{|i-j|}$

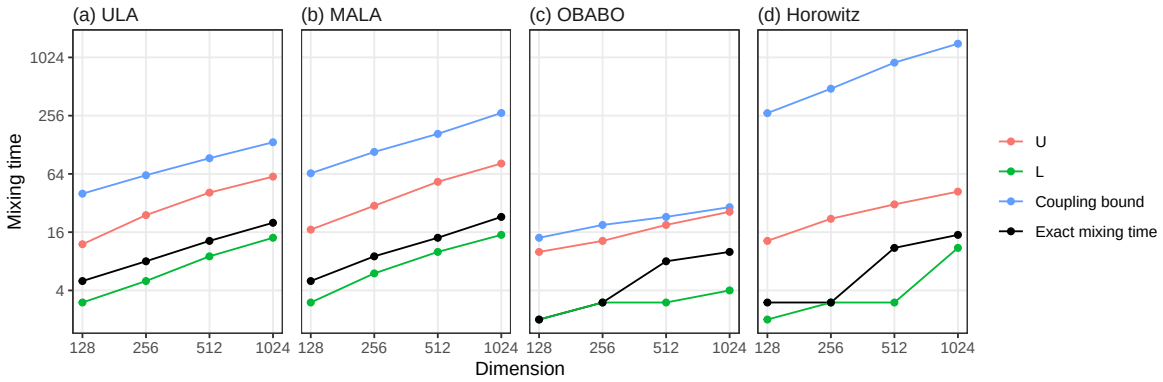

Figure 5.5.2: Mixing time of various adjusted and unadjusted MCMC algorithms, see Section 5.5.4 for details.

and using spherical Gaussian proposals with standard deviation $h = d^{-1/6}$ in various dimensions d. The target condition number is $\kappa \approx 9$ in all dimensions. The initialization $\pi^{(0)} = \mathcal{N}_d(0_d, 3I_d)$ satisfies $\pi^{(0)} \overset{\text{COT}}{\rightsquigarrow} \pi$.

While theoretical bounds must consider worst-case scenarios, the proposed estimators allow for comparisons to be drawn in the operational regime. Our scaling $h \sim d^{-1/6}$ is larger than ones suggested by non-asymptotic analyses (e.g. Wu et al., 2022), but it is consistent with asymptotic analyses at stationarity (Roberts and Rosenthal, 1998) and at transience when converging "inward" from the tails of the target (Christensen et al., 2005). The initialization ensures that we are in the latter regime.

Figure 5.5.2 displays estimates of the mixing time $\tau_6 = \inf\{t : \mathcal{W}_2^2(\pi^{(\infty)}, \pi^{(t)}) \leq 6\}$. The proposed estimators $\{U, L\}$ allow for meaningful comparisons to be drawn between algorithms: our findings are in line with the better scaling of the underdamped diffusion with the condition number of the target, as well as with the common belief that Metropolization slows down mixing. For the Horowitz method, the slow-down is due to the momentum reversals that occur whenever proposals are rejected, which cause the sampler to back-track. These momentum reversals are particularly problematic for the coupling bound, because they cause the coupled chains to drift apart when acceptances (resp. rejections) do not occur simultaneously. The coupling bound therefore wrongly suggests that the Horowitz method converges significantly slower than MALA.

Stochastic volatility model

We consider the posterior distribution of a stochastic volatility model (e.g. Liu, 2001) of dimension $d = 360$, a popular benchmark for MCMC algorithms. We target this model with various MCMC algorithms: the RWM algorithm with spherical Gaussian proposals and either (a) the optimal step size scaling (24% acceptance rate; Roberts et al., 1997) or (b) a smaller step size scaling (64% acceptance rate); (c) MALA with spherical proposals and the optimal step size scaling (57% acceptance rate; Roberts and Rosenthal, 1998); (d) Fisher-MALA (Titsias, 2023), an adaptive MCMC algorithm that learns the proposal covariance structure together with the global scale parameter. The algorithms are initialized from the prior, which we verified to be substantially more dispersed than the target posterior distribution.

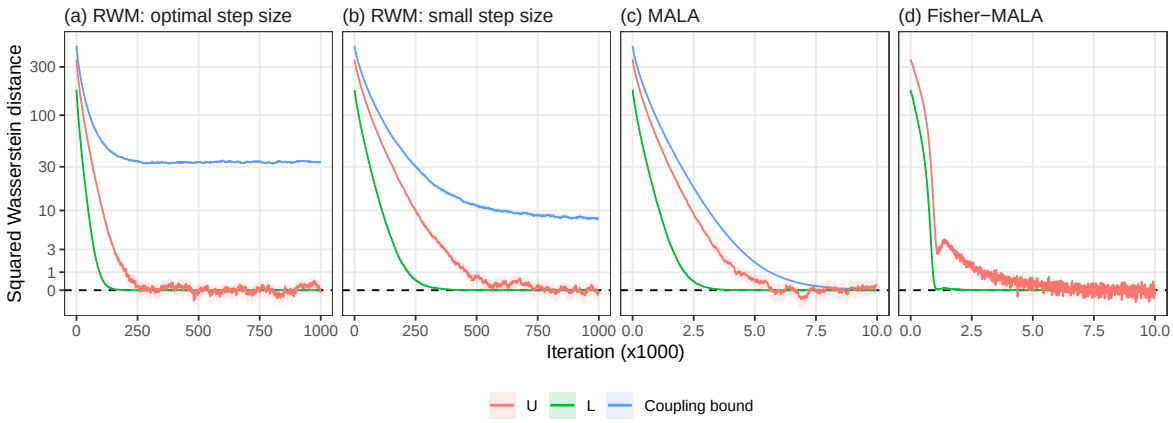

Figure 5.5.3: Convergence of various MCMC algorithms targeting a stochastic volatility model, see Section 5.5.4 for details. Shaded areas represent approximate 95% confidence intervals. Recall that we estimate the idealized coupling bound (5.5.2) with infinite time-lag parameter; samples from the target are obtained by very long MCMC runs. No coupling bound is computed for Fisher-MALA.

Figure 5.5.3 displays estimates of the convergence rates of the considered algorithms. Based on the proposed estimators $\{U, L\}$, we see that the RWM converges faster with the larger step size, that MALA converges an order of magnitude faster than the RWM due to its use of more informative proposals, and that Fisher-MALA converges faster than MALA due to the adaptation. Notably, the initial convergence rate of Fisher-

MALA is super-exponential due to rapid initial adaptation, which slows down to approximately exponential as the adaptation stabilizes. Consistent with the findings of Titsias (2023, Appendix E), Fisher-MALA appears not to converge monotonically in Wasserstein distance.

The proposed estimators provide such insights without requiring specific step sizes or that the underlying algorithm be Markovian. In contrast, the effectiveness of coupling-based estimators depends on the considered MCMC algorithm and its tuning parameters, with the considered reflection-maximal coupling of the RWM (Jacob et al., 2020b) failing to produce informative bounds in this experiment. Furthermore, the coupling-based methodology of Biswas et al. (2019) is not yet applicable to non-Markovian adaptive algorithms.

In Appendix C.6.3, we perform simulations with a more contractive but considerably more compute-intensive RWM coupling from Papp and Sherlock (2025b), follow-up work from the initial version of this manuscript, as well as with a Gaussian approximation to the model where the exact squared Wasserstein distance is available.

5.6 Discussion

Centering is a simple and effective strategy for obtaining informative estimates of the squared Euclidean 2-Wasserstein distance. We have demonstrated that our proposed centered estimators can often be viewed as approximate bounds on the squared Wasserstein distance, and have developed them into methodologies for assessing the quality approximate inference methods and the convergence of MCMC algorithms. The proposed methodologies compare favorably with coupling-based methods (Biswas et al., 2019; Biswas and Mackey, 2024), while requiring considerably less expertise from the user.

We highlight a few methodological extensions that could be explored by further

work.

Fast approximations. Practitioners with access to GPUs could speed up the computation of the proposed estimators, at the cost of introducing a small degree of approximation, by using regularized versions of $\mathcal{W}_2^2$ with a small regularization parameter (e.g. Cuturi, 2013; Genevay et al., 2018). In settings like Section 5.5 where multiple related optimal transport problems must be solved, progressive solvers based on successive warm starts (e.g. Kassraie et al., 2024) could speed up the computation further.

Importance-weighted empirical measures. Importance sampling schemes (e.g. Chopin and Papaspiliopoulos, 2020), which approximate distributions by unequally weighted empirical measures, can provide an appealing alternative to MCMC in Bayesian computation applications such as those in Section 5.4. Exploring the use of importance-weighted empirical measures within our centered estimators is thus a promising direction for further work. We speculate that, as in Section 5.4, the behavior of the proposed estimators would primarily depend on the effective sample sizes (Kong, 1992) of the importance-weighted empirical measures.

Chapter 6

Conclusions

Markov chain Monte Carlo algorithms suffer from burn-in bias; recently-proposed coupling methods promise ways of assessing (Biswas et al., 2019) and eliminating (Jacob et al., 2020b) this bias. In Chapter 3, we showed how to design and analyze efficient couplings of Markov chain Monte Carlo algorithms based on the framework of high-dimensional limiting processes and optimal scaling (Roberts et al., 1997; Christensen et al., 2005). In Chapter 4, we explored how the scalar parameters of these coupling methods should be tuned for maximum efficiency. Finally, in Chapter 5, we provided an alternative method, that does not rely on couplings, of assessing the burn-in bias of Markov chain Monte Carlo algorithms and of assessing the bias of approximate sampling methods.

This thesis, thus, primarily addresses aspects of the problem of making Markov chain Monte Carlo more principled. For researchers of Monte Carlo methods, it provides design principles for couplings and contributes towards the understanding coupling methods. For practitioners using Monte Carlo methods, it enhances the practicality of coupling methods, and provides a competitive alternative to these methods that does not rely on couplings.

We now provide directions for further work, in addition to those outlined in the

conclusions of Chapters 3, 4 and 5.

The work carried out for the RWM in Chapter 3 could provide a template for the design and analysis of couplings for other Metropolis-Hastings algorithms. Both MALA and HMC would be natural candidates, as they have well-developed optimal scaling theories (Roberts and Rosenthal, 1998; Christensen et al., 2005; Beskos et al., 2013). Two potential technical difficulties are that these algorithms can require smaller step size scalings at transience than at stationarity, and that the marginal dynamics of HMC may not be diffusive at stationarity. Nevertheless, our preliminary theoretical results suggest that near-optimal couplings for these algorithms may require less computational overhead than those of the RWM.

The framework of Chapter 3 explicitly optimizes for contraction when the coupled chains are relatively far apart. Could this theory be extended to the setting where the chains are closer together? A related methodological question is how to best combine couplings that contract with couplings that coalesce, beyond the simple two-scale strategy used in Chapter 3. Dau and Chopin (2023, Algorithm S.9), which can turn any coupling into one that coalesces based on a preliminary draw from the "overlap zone," seems to be a promising step in this direction.

A deeper understanding of the effectiveness of the unbiased estimators of Jacob et al. (2020b) could be obtained through results that explicitly connect their efficiency (Chapter 4) to the contractivity of the underlying coupling kernel (Chapters 3). We are currently pursuing non-asymptotic results in this direction.

The Wasserstein distance estimators of Chapter 5 are quite generic, and so potentially lend themselves to a variety of statistical applications. For example, consider the problem of fitting a parametric model μ_θ against samples from an unknown data-generating distribution μ by minimizing an appropriate discrepancy (Basu et al., 2011; Bernton et al., 2019). We could fit the parametric model by sampling from it and

directly minimizing the plug-in Wasserstein distance estimator

$$\theta^* = \arg\min_\theta \mathcal{W}_2(\mu_n, \mu_{\theta,n}),$$

however, the large bias of the plug-in estimator may affect the accuracy of the parameter estimate θ^*. We speculate that fitting models using the debiased Wasserstein distance estimators of Chapter 5 could improve the accuracy of the parameter estimates substantially in moderate to high dimensions.

Appendix A

Appendix for Chapter 3

A.1 Unbiased MCMC with couplings

The lagged coupling framework recalled in Section 3.2 can also be used for the unbiased estimation of expectations of test functions $h(\cdot)$ with respect to the target π (Jacob et al., 2020b; Biswas et al., 2019; Douc et al., 2023). We require additional conditions on the tail decay of the meeting time τ and the moments of the function of interest $h(\cdot)$, namely:

1. There exist $\delta \in (0,1)$ and $C > 0$ such that $\mathbb{P}(\tau > t) \leq C\delta^t$ for all $t \geq 0$.
2. It holds that $\lim_{t\to\infty} \mathbb{E}[h(Y_t)] = \mathbb{E}_\pi[h(Y)]$. Additionally, there exist $\eta, C > 0$ such that $\mathbb{E}[h(Y_t)^{2+\eta}] \leq C$ for all $t \geq 0$.

Then, for any integer $m \geq k \geq 0$, the following is an unbiased estimator of $\mathbb{E}_\pi[h(X)]$:

$$H_{k:m} = \frac{1}{m+k-1}\sum_{t=k}^{m} h(X_{t-L}) + \sum_{t=k}^{\tau-1} \frac{\gamma(t;k,m,L)}{m-k+1}\{h(X_t) - h(Y_t)\}$$

where $\gamma(t;k,m,L) = 1 + \lfloor (t-k)/L \rfloor - \lceil 0 \vee (t-m)/L \rceil$. The assumptions also ensure that $H_{k:m}$ has finite variance, so that an average of i.i.d. copies of $H_{k:m}$ is guaranteed to converge at the Monte Carlo rate. In turn, this enables principled parallel MCMC,

through averages of estimators computed through pairs (X, Y) simulated in parallel.

It is clear that the meeting time τ imposes a lower bound on the length of the simulation. Numerical evidence also indicates that the variance of the estimator grows with the meeting time. It is therefore important to design couplings which ensure that the meeting times stay small, say, on average.

See Douc et al. (2023, Appendix B) for a derivation of the estimator $H_{k:m}$; the form originally given in the rejoinder of Jacob et al. (2020b) is unfortunately incorrect. The geometric tail decay condition for the meeting time τ was relaxed to polynomial in Middleton et al. (2020), however the moment condition on $h(\cdot)$ appears crucial. We briefly discuss the use of couplings for the unbiased estimation of the asymptotic variance of an MCMC algorithm (Douc et al., 2023) in Appendix A.3.4.

A.2 Additional discussion on couplings of the RWM

A.2.1 Asymptotically optimal Markovian coupling

We recall the approximations that motivated the GCRN coupling. Let $V(x) = \log \pi(x)$. Using firstly a Taylor expansion, the acceptance indicator at $X_t = x$ is

$$\begin{aligned} B_x &\approx \mathbb{1}\big\{ \log U_x \leq h Z_x^\top \nabla V(x) + h^2 Z_x^\top \nabla^2 V(x) Z_x \big\}, \\ &\approx \mathbb{1}\big\{ \log U_x \leq g(x) Z_{\nabla x} + c(x) \big\}, \end{aligned}$$

where: $Z_{\nabla x} = Z_x^\top n_x \sim \mathcal{N}_1(0,1)$ with $n_x = \nabla V(x)/\|\nabla V(x)\|$; $g(x) = h\|\nabla V(x)\|$ and $c(x) = h^2 \operatorname{Tr}(\nabla^2 V(x))$ are constants. Scaling the step size as $h = \ell d^{-1/2}$ with the dimension d and assuming standard asymptotics as $d \to \infty$ (as in e.g. Roberts et al., 1997), the above approximations are sharp in the limit.

These observations motivated the GCRN coupling and were used to prove that it was asymptotically optimal for contraction over the class of product couplings $\mathcal{P}$. However,

by considering the random variables $\xi_x = g(x)Z_{\nabla x} - \log U_x$ and $\xi_y = g(y)Z_{\nabla y} - \log U_y$ directly, it is possible to construct a coupling that is asymptotically optimal over the entire class of Markovian couplings $\mathcal{M}$.

An implementable optimal coupling

We seek a coupling of RWM kernels which is optimally contractive with respect to the squared Euclidean distance, so equivalently we seek to maximize $\mathbb{E}[h^2 Z_x^\top Z_y B_x B_y \mid (X_t, Y_t) = (x, y)]$. To derive an implementable optimal coupling, we use the following upper bound on the objective:

$$\begin{aligned}\mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y \mid (X_t, Y_t) = (x, y)\right] &\lessapprox \ell^2 \mathbb{E}[\mathbb{1}\{0 \le \xi_x + c(x)\}\mathbb{1}\{0 \le \xi_y + c(y)\}], \\ &\le \ell^2 \mathbb{P}(0 \le \xi_x + c(x)) \wedge \mathbb{P}(0 \le \xi_y + c(y)),\end{aligned} \tag{A.2.1}$$

where the first approximate inequality is asymptotically sharp as $d \to \infty$, and the second inequality is trivial. The first inequality is satisfied if $Z_x \approx Z_y$ up to a low-rank perturbation. Using a standard optimal transport argument (Villani, 2003, Remark 2.19), one can show that an optimal coupling of (ξ_x, ξ_y) which attains the second inequality is $\xi_x = F_x^{-1}(U)$ and $\xi_y = F_y^{-1}(U)$, where $U \sim \mathrm{Unif}(0, 1)$ and $F_{x,y}$ are the respective CDFs of $\xi_{x,y}$. Since ξ_x only constrains one coordinate of Z_x (and since ξ_y similarly constrains Z_y), we can construct a coupling which renders both inequalities of Equation (A.2.1) asymptotically tight.

Algorithm 4 describes our proposed modification to the GCRN coupling.[1] The algorithm induces a coupling of RWM kernels which is asymptotically optimally contractive coupling over $\mathcal{M}$ in the regimes considered in this paper.[2]

To construct Algorithm 4, we exploited that $\xi_x \sim \mathrm{EMG}(0, g^2(x), 1)$, where $\mathrm{EMG}(\mu, \sigma^2, \lambda)$ denotes the *exponentially modified Gaussian* distribution (EMG; Grushka, 1972), the

[1] The coupling in Line 3 of Algorithm 4 is arbitrary.

[2] We omit the proof, which follows similar lines to that of Theorem 3.5.2.

Algorithm 4 Asymptotically optimal Markovian coupling

Require: Target density $\pi : \mathbb{R}^d \to \mathbb{R}$, score $\nabla \log \pi : \mathbb{R}^d \to \mathbb{R}^d$, step size $h > 0$, $g(\cdot) = h\|\nabla \log \pi(\cdot)\|$ and $n(\cdot) = h\nabla \log \pi(\cdot)/g(\cdot)$, current state $x \in \mathbb{R}^d$.
Require: Inverse CDF function $F^{-1}(\cdot \mid \mu, \sigma^2, \lambda)$ of $\mathrm{EMG}(\mu, \sigma^2, \lambda)$ distribution.

Sample $U \sim \mathrm{Unif}(0, 1)$.
Set $\xi_x = F^{-1}(U \mid 0, g(x)^2, 1)$ and $\xi_y = F^{-1}(U \mid 0, g(y)^2, 1)$. ▷ Optimal coupling

Sample $Z_{\nabla x} \sim \overline{\mathcal{N}}(g(x), 1 \mid \xi_x/g(x))$ and $Z_{\nabla y} \sim \overline{\mathcal{N}}(g(y), 1 \mid \xi_x/g(y))$. ▷ Used Lemma A.2.1
Set $\log U_{\nabla x} = g(x)Z_{\nabla x} - \xi_x$ and $\log U_{\nabla y} = g(y)Z_{\nabla y} - \xi_y$.

Sample $Z \sim \mathcal{N}_d(0_d, I_d)$.
Set $Z_x = Z + \{Z_{\nabla_x} - n(x)^\top Z\}n(x)$ and $Z_y = Z + \{Z_{\nabla y} - n(y)^\top Z\}n(y)$. ▷ As in GCRN

return (Z_x, Z_y) and (U_x, U_y)

distribution of the convolution of a Gaussian $\mathcal{N}_1(\mu, \sigma^2)$ variate and an exponential variate with rate λ. The correctness of Algorithm 4 stems from the conditional distributions of EMG random variables, see Lemma A.2.1 below. Algorithm 4 is implementable using numerical inversion, as the cumulative distribution function $F(\cdot \mid \mu, \sigma^2, \lambda)$ of an $\mathrm{EMG}(\mu, \sigma^2, \lambda)$ variable has the tractable expression,

$$F(x \mid \mu, \sigma^2, \lambda) = \Phi(x \mid \mu, \sigma^2) - \frac{1}{2}\exp\left(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)\right)\mathrm{erfc}\left(\frac{\mu + \lambda\sigma^2 - x}{\sqrt{2}\sigma}\right),$$

where $\Phi(\cdot \mid \mu, \sigma^2)$ is the CDF of a $\mathcal{N}_1(\mu, \sigma^2)$ variate and $\mathrm{erfc}(\cdot)$ is the complementary error function.

Lemma A.2.1. *Let $\xi = Z + E \sim \mathrm{EMG}(\mu, \sigma^2, \lambda)$, that is independently $Z \sim \mathcal{N}_1(\mu, \sigma^2)$ and $E \sim \mathrm{Exp}(\lambda)$. Then,*

$$Z \mid \xi \sim \overline{\mathcal{N}}(\mu + \lambda\sigma^2, \sigma^2 \mid \xi),$$
$$E \mid \xi \sim \underline{\mathcal{N}}(\xi - \mu - \lambda\sigma^2, \sigma^2 \mid 0),$$

where $\overline{\mathcal{N}}(\mu, \sigma^2 \mid u)$ *denotes the normal distribution* $\mathcal{N}_1(\mu, \sigma^2)$ *truncated above at* u*, and analogously* $\underline{\mathcal{N}}(\mu, \sigma^2 \mid l)$ *denotes* $\mathcal{N}_1(\mu, \sigma^2)$ *truncated below at* l*.*

Proof. The relevant density functions are

$$p_Z(x) \propto \exp\{-(x-\mu)^2/(2\sigma^2)\}, \quad p_E(x) \propto \exp(-\lambda x)\mathbb{1}\{0 \le x\}.$$

Since $E + Z$ is a convolution, we have that

$$p(Z = x \mid E + Z = \xi) \propto p_Z(x)p_E(\xi - x) \propto \exp\{-(x - \mu - \lambda\sigma^2)^2/(2\sigma^2)\}\mathbb{1}\{x \le \xi\},$$
$$p(E = x \mid E + Z = \xi) \propto p_Z(\xi - x)p_E(x) \propto \exp\{-(x - \xi + \mu + \lambda\sigma^2)^2/(2\sigma^2)\}\mathbb{1}\{x \ge 0\}.$$

These are truncated normal distributions, which concludes the proof. □

ODE limit

When the target is $\pi^{(d)} = \mathcal{N}_d(0_d, I_d)$ the proposed coupling satisfies a high-dimensional ODE limit as in Theorem 3.4.4. The drift in Proposition 3.4.2 requires the change

$$\begin{aligned} b_{\mathrm{opt}}(x, y, v) &= \ell^2\mathbb{E}\big[1 \wedge e^{\ell x^{1/2}Z - \ell^2/2}\big] \wedge \mathbb{E}\big[1 \wedge e^{\ell y^{1/2}Z - \ell^2/2}\big] - \ell^2 v\left\{q_\ell(x) + q_\ell(y)\right\} \\ &=: p_\ell(x) \wedge p_\ell(x) - v\left\{q_\ell(x) + q_\ell(y)\right\}, \end{aligned}$$

where $Z \sim \mathcal{N}_1(0, 1)$,

$$q_\ell(x) = \ell^2 e^{\ell^2(x-1)/2}\Phi\left(\frac{\ell}{2x^{1/2}} - \ell x^{1/2}\right), \qquad \text{(Lemma A.4.8)}$$
$$p_\ell(x) = q_\ell(x) + \ell^2\Phi\left(-\frac{\ell}{2x^{1/2}}\right). \qquad \text{(Proposition 3.4.2)}$$

By construction, the drift $b_{\mathrm{opt}}(\cdot)$ is point-wise optimal among all couplings in $\mathcal{M}$. This drift is for the inner-product process; the corresponding drift for the squared-distance process is $\bar{b}_{\mathrm{opt}}(x, y, s) = a(x) + a(y) - 2b_{\mathrm{opt}}(x, y, (x + y - s)/2)$.

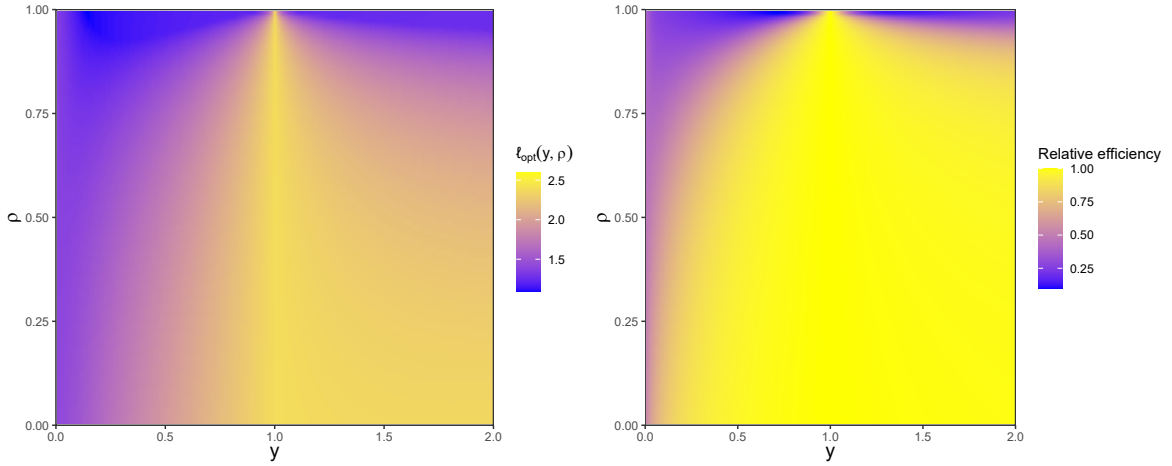

Figure A.2.1: **Left:** Heatmap of optimal step size $\ell_{\text{opt}}(y, \rho)$ for the optimal Markovian coupling. **Right:** Efficiency of the GCRN coupling relative to that of the optimal Markovian coupling, at the point-wise optimal step sizes.

Optimal scaling and relative efficiency

We consider the optimal scaling of the optimal Markovian coupling. As in Section 3.4.2, we optimize for contraction in the drift $\bar{b}(\cdot)$ corresponding to the squared Euclidean distance between the chains, parametrizing the state as (x, y, ρ) where $\rho \in [-1, 1]$ denotes cosine similarity. Figure A.2.1 (left) shows the point-wise optimal scaling $\ell_{\text{opt}}(y, \rho)$, having fixed $x = 1$ so as to emulate a lagged coupling with a large lag parameter L.

Figure A.2.1 (right) shows the relative efficiency of the GCRN coupling against that of the optimal Markovian one. For each coupling, we chose the step size optimally at each (y, ρ); the relative efficiency was then computed as the ratio of the drifts $\bar{b}_{\text{gcrn}}(\cdot)/\bar{b}_{\text{opt}}(\cdot)$ at these optimal step sizes. We see that the GCRN coupling loses very little efficiency over most of the range.

A.2.2 Preconditioning

Linear preconditioning can often speed up computations in practical MCMC. For a RWM algorithm with Gaussian proposals, preconditioning is equivalent to letting the proposals be of the form $x + LZ_x \sim \mathcal{N}(x, \Sigma)$ with $Z_x \sim \mathcal{N}_d(0_d, I_d)$ and $LL^\top = \Sigma$. (For

instance, L could be the Cholesky factor of Σ.) We describe here how couplings of proposals $(x + LZ_x, y + LZ_y)$ should be implemented.

GCRN coupling. The natural extension to the GCRN coupling of the RWM should account for first-order variation in the log acceptance ratio of, whose Taylor expansion is

$$V(x + LZ) - V(x) = Z^\top \{L^\top \nabla V(x)\} + Z^\top \{L^\top \nabla^2 V(\bar{x}) L\} Z,$$

where $V = \log \pi$, and $\bar{x}$ is on the segment from x to $x + LZ$. It follows that the GCRN coupling should be

$$Z_x = Z - (n_x^\top Z) n_x + Z_1 n_x, \quad Z_y = Z - (n_y^\top Z) n_y + Z_1 n_y,$$

when preconditioning is used, where: $n_x = \mathrm{Nor}\{L^\top \nabla V(x)\}$ and $n_y = \mathrm{Nor}\{L^\top \nabla V(x)\}$; $\mathrm{Nor}(x) = x/\|x\|$; $Z_1 \sim \mathcal{N}_1(0, 1)$ and $Z \sim \mathcal{N}_d(0_d, I_d)$ are independent. In effect, this amounts to preconditioning the logarithmic gradient by $L^\top$.

One appeal of the GCRN coupling is that it can be straightforwardly adapted to chains which use position-dependent preconditioning $L(x)$, for instance by changing $L \leftarrow L(x)$ in the coupling above. This can lead to effective coupling strategies, as we have demonstrated in the case of the Hug kernel in Section 3.7.3.

Reflection (-maximal) coupling. The reflection coupling should maximize the variation of $L(Z_x - Z_y)$ in the direction of $(x - y)$, while minimizing the variation in all other directions. This can be achieved through the coupling

$$Z_y = Z_x - 2(e^\top Z_x) e, \tag{A.2.2}$$

where $e = \mathrm{Nor}(L^{-1}(x - y))$. Following Jacob et al. (2020b, Section 4.1), one appeal of this coupling is that it can be modified to allow for the proposals to be identical with

maximal probability. This, of course, allows the chains to coalesce. Let $s(\cdot)$ be the density function of a $\mathcal{N}_d(0_d, I_d)$ variate and let $z = L^{-1}(x-y)$. The resulting *reflection-maximal* coupling sets $Z_y = Z_x - z$ with probability $s(x-z)/s(x)$, and otherwise employs the reflection move (A.2.2).

The reflection-maximal coupling furthermore applies to any pair of Gaussians $\mathcal{N}(x, \Sigma)$ and $\mathcal{N}(y, \Sigma)$ with the same covariance matrix. We use it to couple MALA proposals in the experiment of Section 3.7.4.

A.3 Further details on the numerical experiments

All computations were carried out in R (R Core Team, 2025); runtime-critical components were written in C++. Code to reproduce the numerical experiments can be found at https://github.com/tamaspapp/rwmcouplings. Part of the binary regression experiments were run on up to 112 processors of a computing cluster. All other experiments were run on a 2019-era Lenovo T490s laptop with 8 processors. (The processor counts include hyper-threading.)

A.3.1 Experiments with standard Gaussian targets

Solving the ODEs We solve the limiting ODEs numerically using `deSolve::ode` (Soetaert et al., 2010). To calculate the drifts in Proposition 3.4.2, we must compute the expectation

$$g(x, y, \rho) = \mathbb{E}_{(Z_1, Z_2) \sim \mathrm{BvN}(\rho)} \left[1 \wedge e^{\ell x^{1/2} Z_1 - \ell^2/2} \wedge e^{\ell y^{1/2} Z_2 - \ell^2/2} \right].$$

For $\rho \in (0,1)$, we evaluate $g(\cdot)$ in terms of numerically tractable quantities in Lemma A.3.1 below. We require the bivariate normal probabilities

$$\overline{\mathrm{BvN}}(p,q \mid \rho) := \mathbb{P}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}(Z_1 \le p, Z_2 \le q),$$
$$\underline{\mathrm{BvN}}(p,q \mid \rho) := \mathbb{P}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}(Z_1 \ge p, Z_2 \ge q),$$

which we compute using `mvtnorm::pmvnorm` (Genz et al., 2025). For $\rho = 1$, we evaluate $g(\cdot)$ using the expression in Lemma A.4.8 below instead.

Lemma A.3.1. *When $\rho < 1$ it holds that*

$$g(x,y,\rho) = \underline{\mathrm{BvN}}\left(\frac{\ell}{2x^{1/2}}, \frac{\ell}{2y^{1/2}} \;\middle|\; \rho\right) + f(x,y,\rho) + f(y,x,\rho),$$

where:

$$f(x,y,\rho) = e^{\ell^2(x-1)/2}\overline{\mathrm{BvN}}\left(\frac{b\ell x^{1/2}}{\sqrt{1+b^2}}, U \;\middle|\; -\frac{b}{\sqrt{1+b^2}}\right),$$
$$b = -\frac{(x/y)^{1/2}-\rho}{\sqrt{1-\rho^2}}, \quad U = \frac{\ell}{2x^{1/2}} - \ell x^{1/2}.$$

Faster simulation of high-dimensional chains In our experiments (Figure 3.4.1), we simulated the full coupled chains $(X_t, Y_t)_{t\ge 0}$ as this was fast even in our considered dimension $d = 2{,}000$. We however note that it is possible to reduce the computation time by a factor $\mathcal{O}(d)$ by simulating the Markov process $(\|X_t\|^2, \|Y_t\|^2, \|X_t - Y_t\|^2)_{t\ge 0}$ directly.

Proof of Lemma A.3.1

Write $g(x,y,\rho) = \mathbb{E}\left[\exp(0 \wedge A \wedge B)\right]$, where $A = \ell x^{1/2} Z_1 - \ell^2/2$ and $B = \ell y^{1/2} Z_2 - \ell^2/2$. Partition the expectation according to the events:

$$\{A \ge 0, B \ge 0\}, \quad \{A < 0, A < B\}, \quad \{A < 0, A = B\}, \quad \{A < 0, A > B\}.$$

Using that $\mathbb{P}(A = B) = 0$, we have that

$$\begin{aligned} g(x, y, \rho) &= \mathbb{P}(A \geq 0, B \geq 0) + \mathbb{E}\left[\mathbb{1}_{\{A<0\}}\mathbb{1}_{\{A<B\}}\exp(A)\right] + \mathbb{E}\left[\mathbb{1}_{\{B<0\}}\mathbb{1}_{\{B<A\}}\exp(B)\right] \\ &= \mathbb{P}\left(Z_1 \geq \frac{\ell}{2x^{1/2}}, Z_2 \geq \frac{\ell}{2y^{1/2}}\right) + f(x, y, \rho) + f(y, x, \rho), \end{aligned}$$

where $f(\cdot)$ is defined as

$$f(x, y, \rho) := \mathbb{E}\left[\mathbb{1}_{\{\ell x^{1/2} Z_1 < \ell^2/2\}}\mathbb{1}_{\{\ell x^{1/2} Z_1 < \ell y^{1/2} Z_2\}} e^{\ell x^{1/2} Z_1 - \ell^2/2}\right].$$

To express $f(\cdot)$ in terms of tractable quantities, we expand $Z_2 = \rho Z_1 + \sqrt{1-\rho^2} Z_*$ where $Z_* \sim \mathcal{N}_1(0, 1)$ is independent of Z_1. Collecting all terms in the integrand of $f(\cdot)$ that depend on Z_* and integrating them out, we have the expression

$$\mathbb{E}_{Z_*}\left[\mathbb{1}_{\{\ell x^{1/2} Z_1 < \ell y^{1/2} Z_2\}}\right] = \mathbb{P}_{Z_*}\left((x/y)^{1/2} Z_1 < \rho Z_1 + \sqrt{1-\rho^2} Z_*\right) =: \Phi(b Z_1),$$

where $b := -\{(x/y)^{1/2} - \rho\}/\sqrt{1-\rho^2}$. It immediately follows that

$$\begin{aligned} f(x, y, \rho) &= \mathbb{E}\left[\mathbb{1}_{\{\ell x^{1/2} Z_1 < \ell^2/2\}} \Phi(b Z_1) e^{\ell x^{1/2} Z_1 - \ell^2/2}\right] \\ &= e^{\ell^2(x-1)/2} \mathbb{E}\left[\mathbb{1}_{\{Z_1 < \ell/(2x^{1/2})\}} \Phi(b Z_1) e^{\ell x^{1/2} Z_1 - \ell^2 x/2}\right] \\ &= e^{\ell^2(x-1)/2} \int_{-\infty}^{\ell/(2x^{1/2})} \Phi(bz) e^{\ell x^{1/2} z - \ell^2 x/2} \phi(z) \mathrm{d}z \\ &= e^{\ell^2(x-1)/2} \int_{-\infty}^{\ell/(2x^{1/2})} \Phi(bz) \phi\left(z - \ell x^{1/2}\right) \mathrm{d}z \\ &= e^{\ell^2(x-1)/2} \int_{-\infty}^{\ell/(2x^{1/2}) - \ell x^{1/2}} \Phi\left(b\left(z + \ell x^{1/2}\right)\right) \phi(z) \mathrm{d}z \\ &=: e^{\ell^2(x-1)/2} \int_{-\infty}^{U} \Phi(a + bz) \phi(z) \mathrm{d}z \\ &= e^{\ell^2(x-1)/2} \overline{\mathrm{BvN}}\left(\frac{a}{\sqrt{1+b^2}}, U \,\middle|\, -\frac{b}{\sqrt{1+b^2}}\right), \text{ (Owen, 1980, Eqn. (10,010.1))} \end{aligned}$$

where: (i) we have used the identity $e^{\ell x^{1/2} z - \ell^2 x/2} \phi(z) = \phi(z - \ell x^{1/2})$ to obtain the fourth

line; (ii) we have defined

$$a := b\ell x^{1/2}, \quad U := \frac{\ell}{2x^{1/2}} - \ell x^{1/2}.$$

Substituing this expression into $g(\cdot)$ completes the proof.

A.3.2 Experiments with elliptical Gaussian targets

We evaluate the eccentricties of targets used in Section 3.5.

AR(1) process An AR(1) process with unit noise increments and correlation ρ has a covariance with entries $\Sigma_{ij}^{(d)} = \rho^{|i-j|}$ for all (i, j). This is a Kac-Murdock-Szegő matrix (see e.g. Trench, 1999). It holds that

$$\frac{1}{d}\operatorname{Tr}\left(\Sigma^{(d)}\right) = 1, \quad \lim_{d\to\infty}\frac{1}{d}\operatorname{Tr}\left(\left(\Sigma^{(d)}\right)^{-1}\right) = \frac{1+\rho^2}{1-\rho^2},$$

so that the limiting eccentricity is $\varepsilon = (1+\rho^2)/(1-\rho^2)$.

Chi-square eigenvalues Let $\lambda \sim \chi^2_\nu$. It holds that $\mathbb{E}[\lambda] = \nu$ and (if $\nu > 2$) that $\mathbb{E}\left[\lambda^{-1}\right] = 1/(\nu-2)$. If the eigenvalues $\lambda_1, \ldots, \lambda_d$ of the covariance matrix $\Sigma^{(d)}$ are sampled i.i.d from χ^2_ν, it holds that

$$\lim_{d\to\infty}\frac{1}{d}\operatorname{Tr}\left(\Sigma^{(d)}\right) = \mathbb{E}[\lambda] = \nu, \quad \lim_{d\to\infty}\frac{1}{d}\operatorname{Tr}\left(\left(\Sigma^{(d)}\right)^{-1}\right) = \mathbb{E}\left[\lambda^{-1}\right] = \frac{1}{\nu-2},$$

so that the limiting eccentricity is $\varepsilon = \nu/(\nu-2)$.

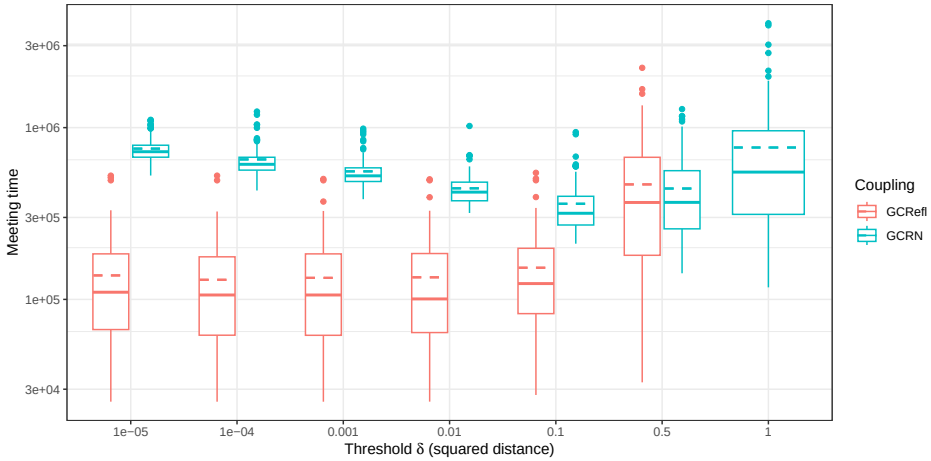

Figure A.3.1: **Stochastic volatility model.** Box plots of $R = 100$ meeting times for various thresholds δ and two-scale RWM couplings. The dashed lines denote the sample means.

A.3.3 Experiments with stochastic volatility model

Posterior log-density and score The posterior log-density of the stochastic volatility model is

$$\log \pi(x_{1:d} \mid y_{1:d}) = -\frac{1}{2}\left(\sum_{t=1}^{d} x_t + \frac{1}{\beta^2}\sum_{t=1}^{d} y_t^2 \exp(-x_t) + \frac{1}{\sigma^2}\sum_{t=1}^{d-1}(\varphi x_t - x_{t+1})^2 + \frac{1-\varphi^2}{\sigma^2}x_1^2\right) + \text{const},$$

where "const" is an offset constant in $x_{1:d}$. The score has entries

$$\frac{\partial \log \pi}{\partial x_t} = -\frac{1}{2} + \frac{1}{2\beta^2} y_t^2 \exp(-x_t) - \frac{\varphi}{\sigma^2}(\varphi x_t - x_{t+1})\mathbb{1}_{\{t \neq d\}} - \frac{1}{\sigma^2}(x_t - \varphi x_{t-1})\mathbb{1}_{\{t \neq 1\}} - \frac{1-\varphi^2}{\sigma^2} x_1 \mathbb{1}_{\{t=1\}}$$

for all $t \in \{1, 2, \ldots, d\}$.

Laplace approximation We use the LBFGS optimizer of the R function `optim` to compute the Laplace approximation. The optimization is initialized at a single draw from the prior.

Rate of convergence of the RWM

Choice of threshold δ in two-scale couplings We search for a sensible threshold $\|X_t - Y_t\|^2 = \delta$ over a coarse a grid $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1\}$. For each coupling and choice of threshold, we measured the meeting time between chains initialized at independent draws from the Laplace approximation (with no lag, i.e. $L = 0$) and repeated this 100 times. The results are displayed in Figure A.3.1. Compared to GCRN, GCRefl prefers a smaller δ, is less sensitive to values smaller than optimal but is much more sensitive to values larger than optimal. (We omitted $\delta = 1$ for the GCRefl coupling as we found the meeting times in preliminary runs to be exceedingly large.)

With the above grid of values and when using a maximal coupling, the one-step probability of coalescing the proposals at $\|X_t - Y_t\|^2 = \delta$ would be $\{0.81, 0.45, 0.02, \dots\}$, where dots denote probabilities under 10^{-13}. The fact that $\delta = 0.1$ performs best for the GCRN coupling, even though the chance of meeting is extremely small at this threshold, indicates that the extra variability induced by the reflection coupling contracts the chains, on average, faster than exponential rate $\mathcal{O}(h^2)$ of GCRN.

Bias of Laplace approximation

Lower bound of Gelbrich (1990) We have the explicit expression

$$\mathcal{W}_2^2(\mathcal{N}_d(\hat{\mu}, \hat{\Sigma}), \mathcal{N}_d(\mu, \Sigma)) = \|\hat{\mu} - \mu\|^2 + \operatorname{Tr}(\Sigma) + \operatorname{Tr}(\hat{\Sigma}) - 2 \operatorname{Tr}\left(\left(\Sigma^{1/2}\hat{\Sigma}\Sigma^{1/2}\right)^{1/2}\right).$$

We estimated the posterior mean and covariance (μ, Σ) by averaging over $R = 50$ independent Hug and Hop chains (see Appendix A.3.3); each chain was run for 50,000 iterations and was warm-started from an independent draw from the Laplace approximation. We followed the guidelines in Ludkin and Sherlock (2022) and tuned Hug to $(T, B) = (0.5, 10)$ integration time and bounce count (for an acceptance rate of 79%) and Hop to $(\lambda, \kappa) = (20, 1)$ for an acceptance rate of 40%.

Jackknife (Efron and Stein, 1981) bias and standard error estimates suggested that the mean-squared error of our estimate of $\mathcal{W}_2^2(\mathcal{N}_d(\hat{\mu}, \hat{\Sigma}), \mathcal{N}_d(\mu, \Sigma))$ was small. We note that the bootstrap is known to be consistent in our case, see Rippl et al. (2016, Section 2.3).

Total variation distance bound The total variation distance bound was computed with a two-scale GCRefl coupling with the same parameter settings (h, δ) as in the experiment of Section 3.7.1.

Convergence of Hug and Hop

Algorithm 5 Hug kernel

Require: Target density $\pi : \mathbb{R}^d \to \mathbb{R}$, score $\nabla \log \pi : \mathbb{R}^d \to \mathbb{R}^d$, current state $x \in \mathbb{R}^d$.
Require: Step size $\delta > 0$ and bounce count $B \in \mathbb{N}$.

function BOUNCE(x, z, δ, B)
  for $i = 1, \ldots, B$ **do**
    $x \leftarrow x + (\delta/2)z$
    $z \leftarrow z - 2(g(x)^\top z)z$
    $x \leftarrow x + (\delta/2)z$
  end for
  return (x, z)
end function

Sample $z \sim \mathcal{N}_d(0_d, I_d)$. ▷ Generate proposal
Propose $(X, Z) \leftarrow$ BOUNCE(x, z, δ, B).
Sample $U \sim \text{Unif}(0, 1)$. ▷ Metropolis filter
if $\log U < (\log \pi(X) - \log \pi(x))$ **then**
  $x \leftarrow X$
end if
return x

Additional algorithmic descriptions The Hug kernel is described in Algorithm 5 (see also Ludkin and Sherlock, 2022, Algorithm 1) and the Hamiltonian Monte Carlo (HMC) kernel is described in Algorithm 6. The synchronous coupling of Hug and HMC

Algorithm 6 Hamiltonian Monte Carlo kernel

Require: Target density $\pi : \mathbb{R}^d \to \mathbb{R}$ and score $\nabla \log \pi : \mathbb{R}^d \to \mathbb{R}^d$, current state $x \in \mathbb{R}^d$.
Require: Leapfrog step size $\delta > 0$ and iteration count $B \in \mathbb{N}$.

function LEAPFROG(x, z, δ, B)
  for $i = 1, \ldots, B$ **do**
    $z \leftarrow z + (\delta/2)\nabla \log \pi(x)$
    $x \leftarrow z + \delta z$
    $z \leftarrow z + (\delta/2)\nabla \log \pi(x)$
  end for
  return (x, z)
end function

Sample $z \sim \mathcal{N}_d(0_d, I_d)$. ▷ Generate proposal
Propose $(X, Z) \leftarrow$ LEAPFROG(x, z, δ, B).
Sample $U \sim \text{Unif}(0, 1)$. ▷ Metropolis filter
if $\log U < (\log \pi(X) - \log \pi(x) - \|Z\|^2/2 + \|z\|^2/2)$ **then**
  $x \leftarrow X$
end if
return x

proceeds by using identical initial momenta (Line 9) and identical acceptance uniforms (Line 11).

Tuning Hop We follow the guidelines of Ludkin and Sherlock (2022) and first tune Hop independently of Hug, choosing $(\lambda, \kappa) = (20, 1)$ for an acceptance rate of 40%.

Tuning HMC and Hug for contractivity We assess the contractivity of coupled HMC and coupled Hug and Hop (H&H) as follows: we independently initialize a pair of coupled chains from the Laplace approximation and track the squared distance between the chains for a fixed number of iterations. We make no attempt to coalesce the chains: HMC is not mixed with the RWM, and for Hop we fix $\delta_{\text{hop}} = 0$ so that only the GCRN coupling is applied.

HMC and Hug require the tuning of two parameters (T, B), where $T = \varepsilon B$ represents the integration time, B is either the number of leafrog steps (for HMC) or the

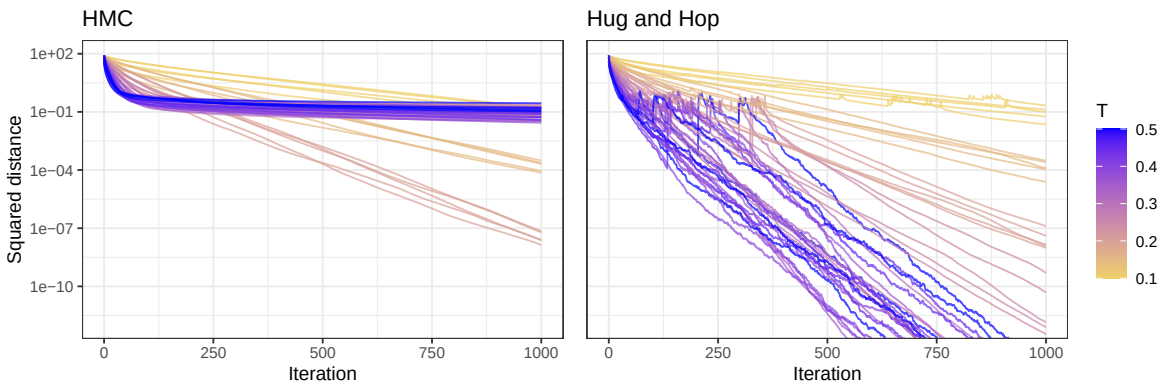

Figure A.3.2: **Stochastic volatility model.** Contractivity of HMC, and Hug and Hop, varying the integration time T and using a fine integration grid.

number of "bounces" (for Hug), and ε represents the step size used to generate the respective discretized proposals. First, we look at the impact of the integration time on the contractivity, fixing a small $\varepsilon = 10^{-3}$ and varying $T \in \{0.1, 0.15, \ldots, 0.5\}$. Figure A.3.2 displays trace plots of 5 replicates for each algorithm and parameter setting. HMC requires a short integration time in order to be contractive, suffering a sharp phase transition from $T = 0.2$ to $T = 0.25$; this is in line with the empirical observations of Heng and Jacob (2019) and the theory of Bou-Rabee et al. (2020) that imposes an upper bound on T to achieve contractivity. In contrast, Hug is more robust with respect to this tuning parameter and benefits from longer integration times than HMC.

Next, we assess which configurations are contractive among the grid of parameters

$$T_{\text{hug}} \in \{0.2, 0.25, \ldots, 0.5\}, \quad T_{\text{hmc}} \in \{0.15, 0.175, \ldots, 0.25\}, \quad B \in \{10, 20, 30\}.$$

Figure A.3.3 displays box plots of the squared distance between chains after 1,000 and 2,000 iterations, from 20 replicates each. We select from among the more efficient contractive configurations $(T_{\text{hug}}, B_{\text{hug}}) = (0.35, 10)$ and $(T_{\text{hmc}}, B_{\text{hmc}}) = (0.225, 10)$, for approximately equal cost per iteration for both H&H and HMC. The configurations correspond to acceptance rates of $\alpha_{\text{hug}} = 90\%$ and $\alpha_{\text{hmc}} = 78\%$.

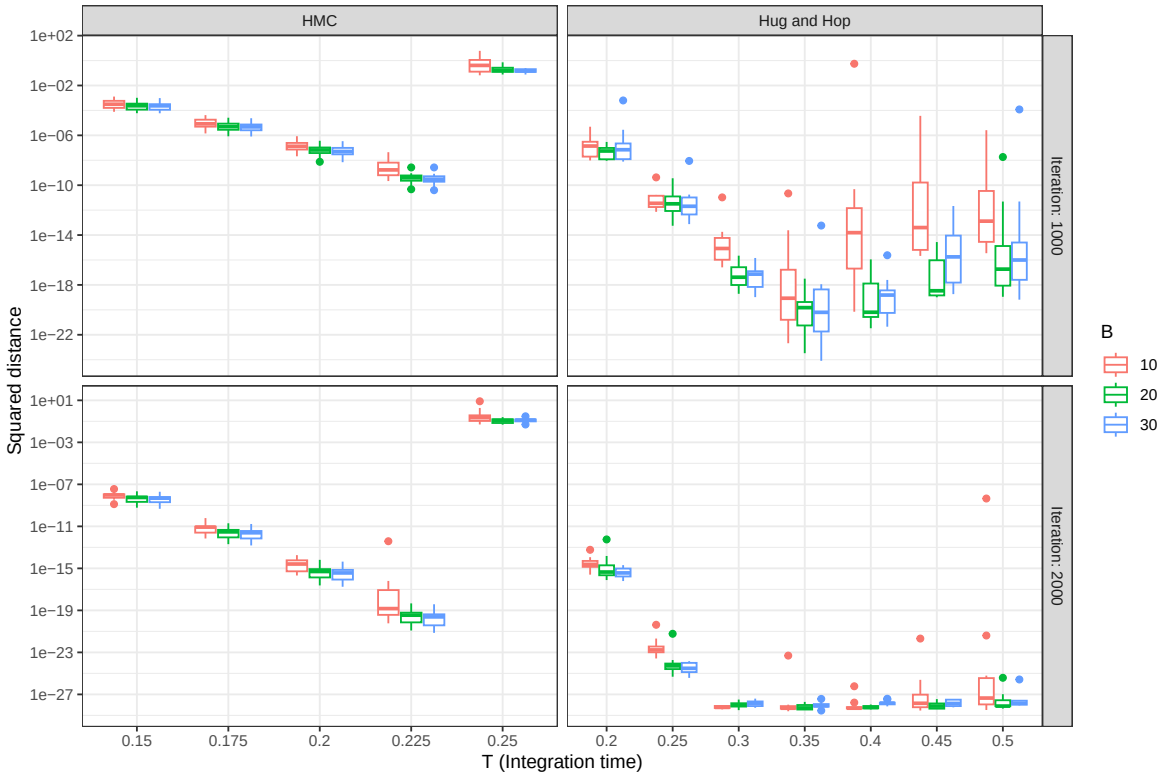

Figure A.3.3: **Stochastic volatility model.** Contractivity of HMC, and Hug and Hop, varying the integration time T and leapfrog steps/bounces B.

RWM and HMC mixture With probability $\gamma = 0.05$ (the default in Heng and Jacob, 2019) we switch from coupled HMC to coupled RWM kernels in order to allow the chains to meet. For the RWM, we use the parameter tuning and two-scale GCRN coupling of Section 3.7.1, which we know perform well. We expect these settings to be close to optimal as Heng and Jacob (2019) demonstrate that the performance of the overall algorithm is insensitive to the mixture probability γ and to the tuning of the RWM.

Choosing δ_{hop} for two-scale Hop coupling We sweep over a grid $\delta \in [10^{-9}, 5 \times 10^{-3}]$ in search of a sensible threshold for the two-scale Hop coupling. Figure A.3.4 displays box plots of 100 replicates for each δ. The meeting times are insensitive to the theshold as long as the chains have a high probability of meeting when $\|X_t - Y_t\|^2 < \delta$; we select $\delta_{\text{hop}} = 10^{-5}$.

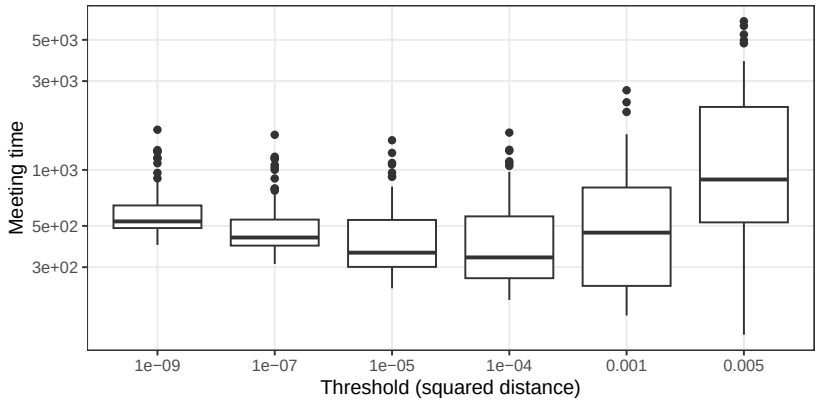

Figure A.3.4: **Stochastic volatility model.** Box plots of the meeting times for various choices of the threshold δ in the two-scale Hug and Hop coupling.

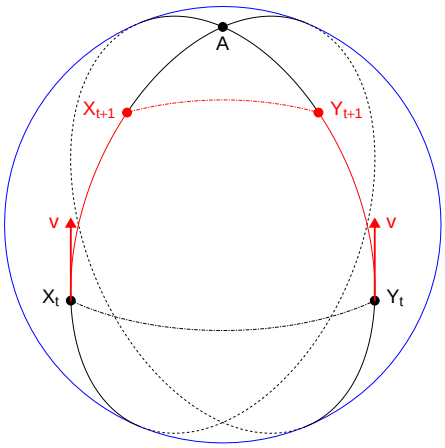

Figure A.3.5: Visualisation of the contractive behaviour of the synchronous Hug coupling. For a spherical target of increasing dimension, the pictured scenario occurs with probability approaching 1.

Contractivity of synchonous Hug coupling in high dimensions We illustrate our intuition as to why the CRN coupling of Hug chains is contractive in high dimensions in Figure A.3.5. For clarity of exposition, we assume that the target is spherically symmetric (so its level sets are hyperspheres), that the Hug dynamics are exact (so they traverse great circles with constant speed), and that the two CRN-coupled Hug chains start at (X_t, Y_t) which lie on the same level set. In high dimensions, the velocity v is essentially always orthogonal to the great circle traversing (X_t, Y_t). As a consequence, essentially almost sure contraction is achieved by synchonous movement towards one of

the two intersections of the two great circles which support the coupled Hug paths (in Figure A.3.5, movement is towards the point A or its antipode).

Intriguingly, as the coupled paths are farthest away at (X_t, Y_t), the chains contract irrespectively of the length of the integration time. This is consistent with our empirical observation that the synchonous Hug coupling is contractive even if the integration time is long (see Figure A.3.2).

Alternative coupling strategies for Hop We have also experimented with replacing the GCRN coupling of Hop with a CRN coupling. The CRN coupling of Hop was significantly less contractive, and produced significantly larger meeting times. This was primarily due to its inability to synchronize acceptance events with the same frequency as GCRN.

A.3.4 Experiments with binary regression

Posterior log-density and score Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the design matrix with rows $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$ and let $\mathbf{y} \in \{0,1\}^n$ be the response. For all $x \in \mathbb{R}^d$, the posterior log-density of a binary regression model with a $\mathcal{N}_d(0_d, \lambda^2 I_d)$ prior is

$$\log \pi(x \mid \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \log F\left((2\mathbf{y}_i - 1)\mathbf{x}_i^\top x\right) - \frac{1}{2\lambda^2}\|x\|^2 + \text{const},$$

where: $F : \mathbb{R} \to (0,1)$ is a cumulative distribution function; "const" is an offset constant in x. The score is

$$\nabla \log \pi(x \mid \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} (\log F)'\left((2\mathbf{y}_i - 1)\mathbf{x}_i^\top x\right)(2\mathbf{y}_i - 1)\mathbf{x}_i - \frac{1}{\lambda^2}x.$$

We consider the logistic case, in which case $\log F(z) = -\log(1 + e^{-z})$ is the logistic log-CDF.

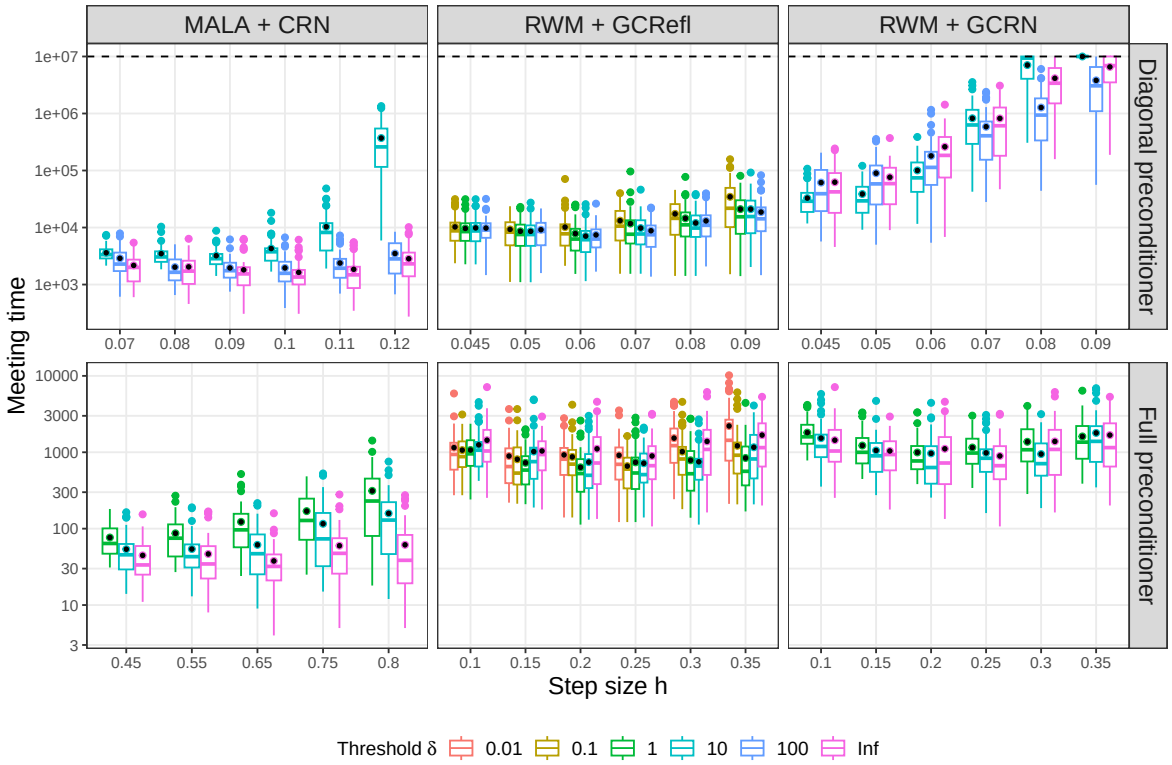

Figure A.3.6: **Binary regression.** Box plots of meeting times for various algorithms and two-scale couplings, see Appendix A.3.4 for details. Black dots indicate sample means. In the top row, the meeting times were truncated at $T = 10^7$ as indicated by the black dashed line. Note that the definition of δ may vary between plots.

Choice of two-scale couplings

We experimented with two-scale couplings that swapped from a contractive coupling to a reflection-maximal coupling when the chains were close enough. As a default, when coupling proposals $\mathcal{N}(x, PP^\top)$ and $\mathcal{N}(y, PP^\top)$ we specified the swapping rule as $\|P^{-1}(x-y)\|^2 \leq \delta^2$ and we varied the value of $\delta > 0$. However, for the RWM using a diagonal preconditioner and the GCRefl coupling, we instead specified the swapping rule as $\|P^{-1}(X-Y)\|^2 \leq \delta^2 h^2$, which is adapted to the local scale of the proposal, as we found the performance of the resulting two-scale coupling to be less sensitive to the choice of δ.

Figure A.3.6 displays box plots of $R = 50$ replicates for several configurations; we measured meeting times between coupled chains started independently from a Gaus-

sian approximation of the target. For the RWM, the two-scale GCRefl coupling was consistently the most effective: (1) when using the diagonal preconditioner, this was the only practical coupling out of the ones considered; (2) when using the full preconditioner, GCRefl outperformed both GCRN and the reflection coupling, particularly for larger step sizes. For MALA, the reflection-maximal coupling on its own consistently performed the best, across both types of preconditioning.

Comparison of RWM and MALA

Parameters for the main experiment We first approximately sampled from the target using $R = 112{,}000$ optimally-scaled MALA chains, warm-started from a Gaussian approximation to the posterior π. Discarding burn-in, we used these chains to estimate the posterior mean and covariance to a very low degree of error; these quantities are relevant for the initialization of our experiment and for the asymptotic variance estimator below.

For the diagonal preconditioner, we used the grid of step sizes

$$h_{\text{rwm}} = \{0.045, 0.05, 0.06, 0.07, 0.08, 0.09\},$$
$$h_{\text{mala}} = \{0.07, 0.08, 0.09, 0.1, 0.11, 0.12\},$$

and we ran $R = 112{,}000$ coupled chains to estimate the relevant efficiency metrics for the RWM and MALA algorithms. For the full preconditioner, we used the grid of step sizes

$$h_{\text{rwm}} = \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35\},$$
$$h_{\text{mala}} = \{0.45, 0.55, 0.65, 0.75, 0.8\},$$

and we ran $R = 11{,}200$ coupled chains to estimate the relevant efficiency metrics for the RWM and MALA algorithms.

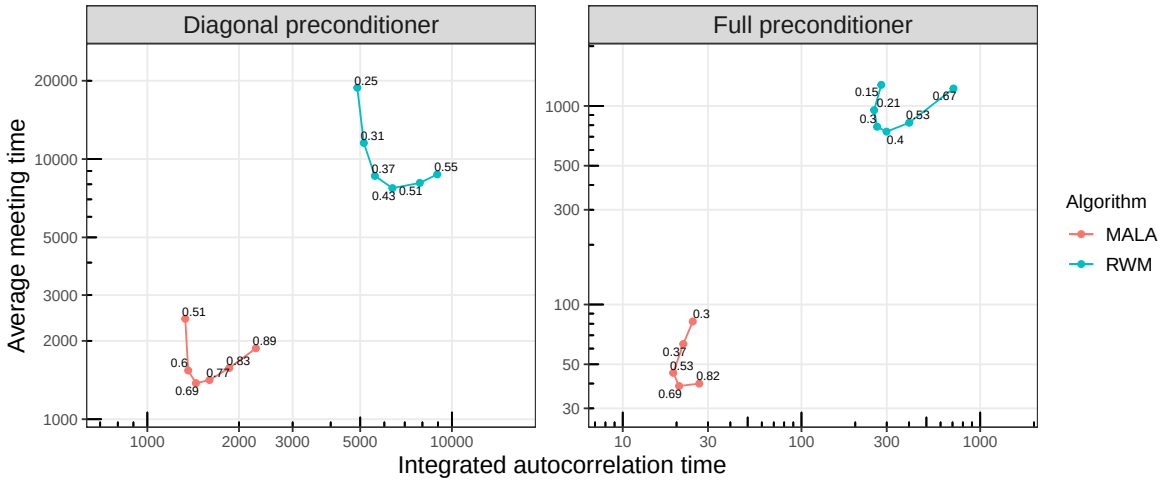

Figure A.3.7: **Binary regression.** Efficiency comparison of RWM and MALA, varying the step sizes. All estimates are shown with two standard errors. The acceptance rate for each step size is overlaid.

Measuring wall-time The experiment in the main text was performed on a shared computing cluster; as a consequence, even when using identical random seeds, we observed large variations in wall-time between different runs. To obtain more reliable measurements, in Figure 3.7.4 we instead scaled the meeting times according to the per-iteration cost of a marginal chain. To account for the gradient evaluations performed by the GCRefl coupling of the RWM, we used the fact that the GCRefl coupling only updates the gradient when a marginal chain accepts, and scaled values according to the the acceptance rate of a single stationary RWM chain. Note that this is a slight overestimate of the actual computing cost for the RWM, since (1) we combined GCRefl with a reflection-maximal coupling, which does not need gradient evaluations, and (2) one of the coupled chains was initialized near the target mode, so its acceptance rate was somewhat smaller than at stationarity.

Efficiency metrics We measure the stationary efficiency of the Markov kernel K by the integrated autocorrelation time (IACT) of a test function $f(\cdot)$,

$$\mathrm{IACT}(K, f) = \frac{\mathrm{Var}(K, f)}{\mathrm{Var}_\pi(f(X))},$$

where the denominator is the target variance and where the numerator is the asymptotic variance,

$$\mathrm{Var}(K, f) = \mathrm{Var}_\pi(f(X_0)) + 2\sum_{t=1}^{\infty} \mathrm{Cov}_\pi(f(X_0), f(X_t)),$$

with the subscript π indicating a stationary chain ($X_0 \sim \pi$ and $X_{t+1} \sim K(X_t, \cdot)$ for all $t \geq 0$).

Figure A.3.7 shows the efficiency comparison of the RWM and MALA in terms of worst IACT (when estimating the regression coefficients; we report the highest coordinate-wise sample average) and average meeting time. We account for wall-time in the main text: we define the time per effective sample as the worst IACT multiplied by the wall-time of one iteration.

On the computing cost of the RWM and MALA With our hardware and hand-optimized C++ implementation, a gradient evaluation was as expensive as a log-density evaluation, so one iteration of MALA took roughly twice as much as one of the RWM. We found the log-density and gradient computations to dominate the cost of the MCMC iteration, even when dense preconditioners were used.

Unbiased estimation of the asymptotic variance Consider the setup of Section 3.2 at no lag ($L = 0$): we run two coupled chains (X, Y) which marginally evolve under the same Markov kernel K, and which meet at some time τ. Suppose furthermore that the chains are initialized *independently* at $X_0 \sim \pi$ and $Y_0 \sim \pi_0$. Douc et al. (2023) show that

$$V(X, Y) = -\mathrm{Var}_\pi(f(X)) + 2\left\{f(X_0) - \mathbb{E}_\pi[f(X)]\right\} \sum_{t=0}^{\tau-1} \{f(X_t) - f(Y_t)\}$$

is an unbiased estimator of the asymptotic variance $\mathrm{Var}(K, f)$. This estimator coincides with "EPAVE" (Douc et al., 2023, Equation 3.2) with $t = 0$, and is unbiased in our case as the X-chain is stationary.

In our experiments, we estimate $\mathrm{IACT}(K, f)$ near-unbiasedly by $V(X, Y)/\mathrm{Var}_\pi(f(X))$, up to negligible MCMC errors in computing $\mathbb{E}_\pi[f(X)]$ and $\mathrm{Var}_\pi(f(X))$. The near-unbiased estimator allows us to exploit parallelism and to investigate the efficiency of the marginal kernel K without running long chains. Furthermore, it does not suffer from the underestimation bias of standard (spectral variance or batch means) estimators of the asymptotic variance (see Vats and Flegal (2021) and references therein), a bias which we have observed in preliminary experiments with the packages `coda` (Plummer et al., 2006) and `mcmcse` (Flegal et al., 2021).

A.4 Proofs

A.4.1 Notation

We make the following conventions in the proofs below. For simplicity of notation, we omit most subscripts and superscripts relating to the dimension d. The Euclidean norm is denoted by $\|\cdot\|$. The standard normal probability density function is denoted by ϕ, the cumulative density function is Φ. $\mathrm{BvN}(\sigma_1^2, \sigma_2^2; \eta)$ denotes the bivariate normal distribution with coordinate-wise variances (σ_1^2, σ_2^2) and covariance $\eta \in [-\sigma_1\sigma_2, \sigma_1\sigma_2]$. We write $\mathrm{BvN}(\eta)$ instead when both variances are unit ($\sigma_1^2 = \sigma_2^2 = 1$). We use $\Longrightarrow$ to denote the weak convergence of random variables and stochastic processes.

A.4.2 Auxiliary results

We shall first require a few auxiliary results. Lemma A.4.1 below recalls the Lipschitz property of the function $f(x) = 1 \wedge \exp(x)$ which appears in the Metropolis-Hastings acceptance ratio. Lemma A.4.2 below extends Lemma A.4.1.

Lemma A.4.1 (Roberts et al., 1997, Proposition 2.2)**.** *Let $a \in \mathbb{R}$. For all $x, y \in \mathbb{R}$, it holds that*

$$|1 \wedge \exp(x) - 1 \wedge \exp(y)| \leq 1 \wedge |x - y|.$$

Lemma A.4.2. *For all $a, b, x, y, u, v \in \mathbb{R}$, it holds that*

$$|a\{1 \wedge \exp(x) \wedge \exp(u)\} - b\{1 \wedge \exp(y) \wedge \exp(v)\}| \leq |a-b| + |b|(1 \wedge |x-y| + 1 \wedge |u-v|).$$

Proof. Firstly, $a(1 \wedge e^{x \wedge u}) - b(1 \wedge e^{y \wedge v}) = (a-b)(1 \wedge e^{x \wedge u}) + b(1 \wedge e^{x \wedge u} - 1 \wedge e^{y \wedge v})$. By the triangle inequality, it follows that

$$|a(1 \wedge e^{x \wedge u}) - b(1 \wedge e^{y \wedge v})| \leq |a-b| + b|1 \wedge e^{x \wedge u} - 1 \wedge e^{y \wedge v}|.$$

Now, by adding and subtracting the cross-term $1 \wedge e^{y \wedge u}$, then using the triangle inequality, we have that

$$\begin{aligned} |1 \wedge e^{x \wedge u} - 1 \wedge e^{y \wedge v}| &= |1 \wedge e^{x \wedge u} - 1 \wedge e^{y \wedge u}| + |1 \wedge e^{y \wedge u} - 1 \wedge e^{y \wedge v}| \\ &\leq 1 \wedge |x \wedge u - y \wedge u| + 1 \wedge |y \wedge u - y \wedge v|, \qquad \text{(Lemma A.4.1)} \\ &\leq 1 \wedge |x-y| + 1 \wedge |u-v|. \end{aligned}$$

where in the final line we used that the function $f_z(x) = z \wedge x$ is 1-Lipschitz, for all $z \in \mathbb{R}$. This concludes the proof. □

Lemmas A.4.3 and A.4.4 below are used to express the limiting drifts in Propositions 3.4.2 and 3.5.5.

Lemma A.4.3. *Let $(Z_1, Z_2) \sim \mathrm{BvN}(\sigma_1^2, \sigma_2^2, \eta)$, let A be a constant and let $A_d \to A$ in probability as $d \to \infty$. Then, it holds that*

$$\lim_{d \to \infty} \mathbb{E}\left[Z_2 1 \wedge \exp(Z_1 - A_d)\right] = \mathbb{E}\left[Z_2 1 \wedge \exp(Z_1 - A)\right].$$

Furthermore, the convergence is uniform over $(\sigma_1^2, \sigma_2^2, \eta)$ in a compact set provided that $A_d \to A$ uniformly over the same set.

Proof. We shall actually show uniform convergence in L_2. Firstly by the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
\mathbb{E}^2\left[Z_2\left(1\wedge e^{Z_1-A_d}-1\wedge e^{Z_1-A}\right)\right] &\leq \mathbb{E}\left[Z_2^2\right]\mathbb{E}\left[\left(1\wedge e^{Z_1-A_d}-1\wedge e^{Z_1-A}\right)^2\right] \\
&\leq \mathbb{E}\left[Z_2^2\right]\mathbb{E}\left[1\wedge(A_d-A)^2\right] \qquad \text{(Lemma A.4.1)} \\
&= \sigma_2^2\mathbb{E}\left[1\wedge(A_d-A)^2\right] \\
&\to 0 \quad \text{as } d\to\infty,
\end{aligned}$$

by the weak convergence $(A_d - A) \implies 0$ as $d \to \infty$, because the function $f(x) = 1 \wedge x^2$ is continuous and bounded and $f(0) = 0$. The squeeze / sandwich theorem takes us to the claimed convergence. The unformity of the convergence follows from the bound. □

Lemma A.4.4. *Let $(Z_1, Z_2) \sim \mathrm{BvN}(\sigma_1^2, \sigma_2^2, \eta)$, let (A, B, C) be constants and let $\{A_d, B_d\} \to \{A, B\}$ in probability and $C_d \to C$ in L_1 as $d \to \infty$. Then, it holds that*

$$\lim_{d\to\infty}\mathbb{E}\left[C_d 1\wedge\exp(Z_1-A_d)\wedge\exp(Z_2-B_d)\right] = C\mathbb{E}\left[1\wedge\exp(Z_1-A)\wedge\exp(Z_2-B)\right].$$

Furthermore, the convergence is uniform over $(\sigma_1^2, \sigma_2^2, \eta)$ in a compact set provided that $\{A_d, B_d, C_d\} \to \{A, B, C\}$ uniformly over the same set.

Proof. We shall actually show uniform convergence in L_1. By Lemma A.4.2, we have that

$$\begin{aligned}
\mathbb{E}\left[\left|C_d 1\wedge e^{(Z_1-A_d)\wedge(Z_2-B_d)}-C1\wedge e^{(Z_1-A)\wedge(Z_2-B)}\right|\right] &\leq \mathbb{E}\left[|C_d-C|\right]+C\mathbb{E}\left[1\wedge|A_d-A|\right]+ \\
&\quad + C\mathbb{E}\left[1\wedge|B_d-B|\right] \\
&\to 0 \quad \text{as } d\to\infty,
\end{aligned}$$

since $C_d \to C$ in L_1 and $\{A_d, B_d\} \to \{A, B\}$ in probability as $d \to \infty$. The squeeze theorem takes us to the claimed convergence; the uniformity of the convergence follows

from the bound. □

Lemma A.4.5 below is used to prove Proposition 3.3.1; it recalls a standard fact concerning the joint distribution of projections a multivariate Gaussian.

Lemma A.4.5. *Let $Z \sim \mathcal{N}_d(0_d, I_d)$ be independent of $u, v \in \mathbb{R}^d$. Then,*

$$\left(u^\top Z, v^\top Z\right) \mid \left\{\|u\|^2, \|v\|^2, u^\top v\right\} \sim \mathrm{BvN}\left(\|u\|^2, \|v\|^2; u^\top v\right).$$

Lemma A.4.6 below is used to prove the optimality of GCRN in Theorems 3.4.1 and 3.5.2; it characterizes the optimal coupling under the utility function $u(x, y) = x \wedge y$.

Lemma A.4.6. *Let μ, ν be real-valued distributions with finite first moments and let $\Gamma(\mu, \nu)$ be the set of all couplings of (μ, ν). Then,*

$$\arg\max_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}[X \wedge Y] = \arg\min_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}[|X - Y|].$$

Furthermore, the optimal coupling is $X = F_\mu^{-1}(U)$, $Y = F_\nu^{-1}(U)$, where $U \sim \mathrm{Unif}(0, 1)$ and where (F_μ, F_ν) denote the cumulative distribution functions of (μ, ν) respectively.

Proof. We have that $|X - Y| = X + Y - 2X \wedge Y$. Since $\mathbb{E}[X + Y]$ is independent of the coupling, the first claim immediately follows. The second claim follows by Villani (2003, Remark 2.19(iii)). □

Lemma A.4.7 states that the expected maximum of positive i.i.d. random variables grows strictly sub-linearly; it is used to show the limiting drift in Proposition 3.5.5. See Correa and Romero (2021) for a simple proof.

Lemma A.4.7 (Downey, 1990, Theorem 6)**.** *Let μ be a positive real-valued distribution such that $\mathbb{E}_\mu[X] < \infty$ and let $(X_i)_{i=1}^d \overset{\text{iid}}{\sim} \mu$. Then,*

$$\lim_{d\to\infty} \mathbb{E}\left[\max\left\{X_1, \ldots, X_d\right\}\right]/d = 0.$$

Lemma A.4.8 evaluates some Gaussian integrals that appear in the limiting drifts of Propositions 3.4.2 and 3.5.5.

Lemma A.4.8. *Let $\alpha, \beta, \ell > 0$ and let $Z \sim \mathcal{N}_1(0,1)$. Then, it holds that*

$$\mathbb{E}\left[Z1 \wedge e^{-\ell\alpha Z - \ell^2/2}\right] = -\ell\alpha e^{\ell^2(\alpha^2-1)/2}\Phi\left(\frac{\ell}{2\alpha} - \ell\alpha\right),$$

$$\mathbb{E}\left[1 \wedge e^{-\ell\alpha Z - \ell^2/2} \wedge e^{-\ell\beta Z - \ell^2/2}\right] = \Phi\left(-\frac{\ell}{2m}\right) + e^{\ell^2(m^2-1)/2}\left\{\Phi\left(\frac{\ell}{2m} - lm\right) - \Phi(-\ell m)\right\} + e^{\ell^2(M^2-1)/2}\Phi(-\ell M),$$

where $m = \alpha \wedge \beta$ and $M = \alpha \vee \beta$.

Lemma A.4.9 below establishes some properties of the function $h : [-1,1]\times(0,\infty) \to (0,\infty)$,

$$h(\rho;\ell) = \mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}\left[1 \wedge e^{\ell Z_1 - \ell^2/2} \wedge e^{\ell Z_2 - \ell^2/2}\right],$$

and is the key to our fixed-point analyses in Propositions 3.4.5 and 3.5.7.

Lemma A.4.9 (Properties of h)**.** *The function $h(\cdot)$ has the following properties:*

1. *$h(1;\ell) = 2\Phi(-\ell/2)$.*
2. *$\partial_\rho h(\rho;\ell) > 0$ for all $\rho \in (-1,1)$ and $\lim_{\rho\nearrow 1}\partial_\rho h(\rho;\ell) = \infty$.*
3. *$\partial_\rho^2 h(\rho;\ell) > 0$ for all $\rho \in [0,1)$.*

The proofs of Lemmas A.4.8 and A.4.9 rely on repeated applications of elementary calculus. As they are not instructive, we postpone them to Appendix A.4.14 at the end of this section.

A.4.3 Proof of Proposition 3.3.1

This is a consequence of Lemma A.4.5. It is clear that $(n_x^\top Z_x, n_y^\top Z_y)$ is bivariate Gaussian under each coupling, and that its coordinates have unit variance. The covariance is coupling-specific:

- For CRN, $Z_y = Z_x$ and therefore $\mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_y) = n_x^\top n_y$.
- For reflection, $Z_y = Z_x - 2(e^\top Z_x)e$ and so

$$\mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_y) = \mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_x - 2(n_y^\top e)e^\top Z_x) = n_x^\top n_y - 2(n_y^\top e)(n_x^\top e).$$

- For GCRN, $n_x^\top Z_x = n_y^\top Z_y$ and so trivially $\mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_y) = 1$.

This concludes the proof.

A.4.4 The process W is Markov in the standard Gaussian case

Suppose that the target is standard Gaussian $\pi = \mathcal{N}_d(0_d, I_d)$. We assume that $\|X_t\|, \|Y_t\| \neq 0$; dealing with the case when one or both are null is straightforward and the proof is omitted. Let $\widehat{X}_t = X_t/\|X_t\|$ and $\widehat{Y}_t = Y_t/\|Y_t\|$.

In order to show that W is Markov, it is sufficient to show that $\{(\|X_t\|^2, \|Y_t\|^2, X_t^\top Y_t)\}_{t\geq 0}$ is Markov. We have the following expressions, shared by all three couplings:

$$\begin{aligned}
\|X_{t+1}\|^2 &= \|X_t\|^2 + \big(2h\|X_t\|\widehat{X}_t^\top Z_x + h^2\|Z_x\|^2\big)B_x,\\
\|Y_{t+1}\|^2 &= \|Y_t\|^2 + \big(2h\|Y_t\|\widehat{Y}_t^\top Z_y + h^2\|Z_y\|^2\big)B_y,\\
X_{t+1}^\top Y_{t+1} &= X_t^\top Y_t + h\|Y_t\|\widehat{Y}_t^\top Z_x B_x + h\|X_t\|\widehat{X}_t^\top Z_y B_y + h^2 Z_x^\top Z_y B_x B_y,
\end{aligned}$$

where $B_x = \mathbb{1}\big\{\log U \leq -h\|X_t\|\widehat{X}_t^\top Z_x - h^2\|Z_x\|^2/2\big\}$ and analogously for B_y. Since U is independent of the remaining randomness, $\big(\widehat{X}_t^\top Z_x, \widehat{Y}_t^\top Z_x, \widehat{X}_t^\top Z_y, \widehat{Y}_t^\top Z_y, \|Z_x\|^2, \|Z_y\|^2\big)$, it suffices to show that the joint distribution of these six random variables is uniquely determined by the triplet $(\|X_t\|^2, \|Y_t\|^2, X_t^\top Y_t)$.

Consider the projections of Z_x, Z_y onto $\mathrm{span}\{X_t, Y_t\}$ and its orthogonal complement separately. Under the couplings of Section 3.3.3 (CRN, reflection, or GCRN), the joint randomness reduces to a χ^2_{d-2} random variable and an independent 6-dimensional multivariate normal with zero mean and a covariance matrix that is uniquely determined

by $(\|X_t\|^2, \|Y_t\|^2, X_t^\top Y_t)$. (The calculations themselves are straightforward, but tedious, and are omitted.) It follows that the process of interest W is Markov, which concludes the proof.

A.4.5 Proof of Proposition 3.4.2

Throughout this proof, we condition on $W_{t/d} = \left(\|X_t\|^2, \|Y_t\|^2, X_t^\top Y_t\right)/d = (x, y, v) \in \mathcal{S}$.

Firstly, we expand the drift to

$$d(W_{(t+1)/d} - W_{t/d}) = \left(\|X_{t+1}\|^2 - \|X_t\|^2, \|Y_{t+1}\|^2 - \|Y_t\|^2, X_{t+1}^\top Y_{t+1} - X_t^\top Y_t\right).$$

Further expanding the first and last terms, we have that

$$\begin{aligned} \|X_{t+1}\|^2 - \|X_t\|^2 &= (2hX_t^\top Z_x + h^2\|Z_x\|^2)B_x, \\ X_{t+1}^\top Y_{t+1} - X_t^\top Y_t &= hY_t^\top Z_x B_x + hX_t^\top Z_y B_y + h^2 Z_x^\top Z_y B_x B_y, \end{aligned} \tag{A.4.1}$$

where $B_x = \mathbb{1}\{U \le \exp(-hX_t^\top Z_x - h^2\|Z_x\|^2/2)\}$ and similarly for B_y using the same $U \sim \text{Unif}(0,1)$. Except for the jump concordance $h^2 Z_x^\top Z_y B_x B_y$, all terms are coupling-independent, so we deal with these first.

Coupling-independent terms

Integrating over $U \sim \text{Unif}(0,1)$, we have that

$$\begin{aligned} \mathbb{E}\left[\|X_{t+1}\|^2 - \|X_t\|^2\right] &= \mathbb{E}\left[(-2Z_1 + h^2\|Z_x\|^2)1 \wedge \exp\left(Z_1 - h^2\|Z_x\|^2/2\right)\right], \\ \mathbb{E}\left[hY_t^\top Z_x B_x\right] &= -\mathbb{E}\left[Z_2 1 \wedge \exp\left(Z_1 - h^2\|Z_x\|^2/2\right)\right], \end{aligned}$$

where $(Z_1, Z_2) = (-hX_t^\top Z_x, -hY_t^\top Z_x)$. By Lemma A.4.5, we have that

$$(Z_1, Z_2) \sim \text{BvN}(h^2\|X_t\|^2, h^2\|X_t\|^2; h^2 X_t^\top Y_t) = \text{BvN}(\ell^2 x, \ell^2 y; \ell^2 v).$$

We have that $\lim_{d\to\infty} h^2\|Z_x\|^2 = \ell^2$ in L_1 uniformly over $(x, y, v) \in \mathcal{S}$, for any compact $\mathcal{S}$. By Lemmas A.4.3 and A.4.4 it follows that

$$\lim_{d\to\infty} \mathbb{E}\left[\|X_{t+1}\|^2 - \|X_t\|^2\right] = \mathbb{E}\left[(-2Z_1 + \ell^2)1 \wedge \exp\left(Z_1 - \ell^2/2\right)\right],$$
$$\lim_{d\to\infty} \mathbb{E}\left[hY_t^\top Z_x B_x\right] = -\mathbb{E}\left[Z_2 1 \wedge \exp\left(Z_1 - \ell^2/2\right)\right],$$

uniformly over the same set. The desired formulae follow by Lemma A.4.8; the first limit is also derived in Christensen et al. (2005). The quantities $\lim_{d\to\infty} \mathbb{E}\left[\|Y_{t+1}\|^2 - \|Y_t\|^2\right]$ and $\lim_{d\to\infty} \mathbb{E}\left[hX_t^\top Z_y B_y\right]$ follow by symmetry.

This completes the calculations relating to the coupling-independent terms. We now turn to the coupling-dependent term, the jump concordance.

Correlation of projections

To evaluate the limiting expected jump concordance, we must express $\rho = \mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_y)$ as a function of (x, y, v). Recall that $e = (X_t - Y_t)/\|X_t - Y_t\|$; the normalized gradient at X_t is $n_x = -X_t/\|X_t\|$. We turn to Proposition 3.3.1 and compute

$$n_x^\top n_y = \frac{X_t^\top Y_t}{\|X_t\|\|Y_t\|} = \frac{v}{xy^{1/2}}, \quad n_x^\top e = \frac{X_t^\top (Y_t - X_t)}{\|X_t\|\|X_t - Y_t\|} = \frac{v - x}{x^{1/2}(x + y - 2v)^{1/2}},$$

and also $n_y^\top e = (v - y)\{y(x + y - 2v)\}^{-1/2}$ by symmetry. Plugging these into Proposition 3.3.1, we have that

$$\rho_{\mathrm{crn}} = \frac{v}{(xy)^{1/2}}, \quad \rho_{\mathrm{refl}} = \frac{2xy - (x + y)v}{(xy)^{1/2}(x + y - 2v)}, \quad \rho_{\mathrm{gcrn}} = 1.$$

Coupling-dependent term

Integrating over $U \sim \mathrm{Unif}(0, 1)$, we have that

$$\mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \mathbb{E}\left[h^2 Z_x^\top Z_y 1 \wedge \exp\left(Z_3 - h^2\|Z_x\|^2/2\right) \wedge \exp\left(Z_4 - h^2\|Z_y\|^2/2\right)\right],$$

where $(Z_3, Z_4) = (-hX_t^\top Z_x, -hY_t^\top Z_y) = (\ell x^{1/2} n_x^\top Z_x, \ell y^{1/2} n_y^\top Z_y)$. By Proposition 3.3.1, we have that

$$(Z_3, Z_4) \sim \mathrm{BvN}(\ell^2 x, \ell^2 y; \ell^2 (xy)^{1/2} \rho),$$

where ρ is coupling-specific and evaluated above.

We have the following limits in L_1 as $d \to \infty$: $h^2 \|Z_x\|^2 \to \ell^2$, $h^2 \|Z_y\|^2 \to \ell^2$ and $h^2 Z_x^\top Z_y \to \ell^2$. The latter limit holds for all considered couplings, as they are low-rank perturbations of the CRN coupling. Moreover, the limits hold uniformly over $(x, y, v) \in \mathcal{S}$ for any compact $\mathcal{S}$. By Lemma A.4.4 it follows that

$$\lim_{d\to\infty} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \ell^2 \mathbb{E}\left[1 \wedge \exp\left(Z_3 - \ell^2/2\right) \wedge \exp\left(Z_4 - \ell^2/2\right)\right],$$

and this limit is uniform the same set. Putting the limits together concludes the proof.

A.4.6 Proof of Proposition 3.4.3

Repeat the proof of Proposition 3.4.2 up to and including the expansions (A.4.1). We bound the second moment of each term in these expansions; all bounds hold due to $B_{x,y} \in [0, 1]$:

$$\begin{aligned}
\mathbb{E}[(2hX_t^\top Z_x B_x)^2] &\le \mathbb{E}[(2hX_t^\top Z_x)^2] = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[4h^2 \|X_t\|^2 Z^2] = 4\ell^2 x, \\
\mathbb{E}[(h^2 \|Z_x\|^2 B_x)^2] &\le \mathbb{E}[h^4 \|Z_x\|^4] = \ell^4 (d+2)/d \le 3\ell^4, \\
\mathbb{E}[(h^2 Z_x^\top Z_y B_x B_y)^2] &\le \mathbb{E}[h^4 (Z_x^\top Z_y)^2] \le \mathbb{E}[h^4 (\|Z_x\|^4 + \|Z_y\|^4)/2] = \ell^4 (d+2)/d \le 3\ell^4,
\end{aligned}$$

where for the last two bounds we used the moments of $\|Z_x\|^2 \sim \chi^2_d$ and that $d \ge 1$. By symmetry, $\mathbb{E}[(2hY_t^\top Z_y B_y)^2] \le 4\ell^2 y$, $\mathbb{E}[(hY_t^\top Z_x B_x)^2] \le \ell^2 y$ and $\mathbb{E}[(hX_t^\top Z_y B_y)^2] \le \ell^2 x$. These bounds are independent of the coupling and the dimension d.

Now, by Cauchy-Schwarz, $(x + y + z)^2 \le 3(x^2 + y^2 + z^2)$ for all $x, y, z \in \mathbb{R}$. Using this inequality twice, we obtain that $\mathbb{E}[d^2 \|W_{(t+1)/d} - W_{t/d}\|^2] \le f_\ell(x, y, v)$ for all $d \ge 1$

and for all couplings, where $f_\ell(x, y, v)$ is a linear combination of the moment bounds obtained above; $f_\ell(\cdot)$ is therefore linear in all of its arguments and is independent of d. It follows that, for any compact set $\mathcal{S} \subset [0, \infty)^3$,

$$\lim_{d\to\infty} \sup_{(x,y,v)\in\mathcal{S}} \mathbb{E}\left[d^2 \|W_{(t+1)/d} - W_{t/d}\|^2\right] \leq \lim_{d\to\infty} \sup_{(x,y,v)\in\mathcal{S}} f_\ell(x, y, v) < \infty.$$

This concludes the proof.

A.4.7 Proof of Theorem 3.4.4

We show here that the infinitesimal generator of the process $W^{(d)}$ converges to that of the ODE $\dot{w(t)} = c_\ell(w(t))$ as $d \to \infty$. Textbook results concerning the convergence of stochastic processes then allow us to conclude. This proof mirrors that of Christensen et al. (2005, Theorem 1) and is identical for each coupling of Section 3.3.3.

Technical preliminaries

We first recall some standard technical results. Let $\bar{\mathcal{S}} = \{(x, y, v) : x \geq 0, y \geq 0, v \leq \sqrt{xy}\}$ and let $\mathcal{C}^\infty$ be the set of all infinitely differentiable functions $h : \bar{\mathcal{S}} \to \mathbb{R}^3$ with compact support. Let $c : \bar{\mathcal{S}} \to \mathbb{R}^3$ be as in Proposition 3.4.2 and let $w : [0, \infty) \to \bar{\mathcal{S}}$ be the solution to the ordinary differential equation (ODE) $\dot{w}(t) = c(w(t))$. Let G be the infinitesimal generator of $\dot{w}(t) = c(w(t))$, which we recall satisfies $Gh(x) = \nabla h(x)^\top c(x)$ for all $h \in \mathcal{C}^\infty$ and $x \in \mathcal{S}$ (e.g. Øksendal, 1998, Theorem 7.3.3). Moreover, the set $\mathcal{C}^\infty$ is a core for the infinitesimal generator of the ODE $\dot{w}(t) = c(w(t))$ (e.g. Sato, 1999, Theorem 31.5, as an ODE is a Lévy process).

Convergence of generator

We now proceed to the main body of the proof, showing the convergence of the discrete time generator to the continuous time generator.

Let $G^{(d)}$ be the discrete time infinitesimal generator of the process $W = W^{(d)}$ and let $h \in \mathcal{C}^\infty$. For $w \in \bar{\mathcal{S}}$, a Taylor expansion gives

$$\begin{aligned} G^{(d)}h(w) &= \mathbb{E}\left[h(W_{(t+1)/d}) - h(W_{t/d}) \mid W_{t/d} = w\right] d \\ &= \mathbb{E}\left[\nabla h(w)^\top (W_{(t+1)/d} - w) \mid W_{t/d} = w\right] d \\ &\quad + \frac{1}{2}\mathbb{E}\left[(W_{(t+1)/d} - w)^\top \nabla^2 h(w^*)(W_{(t+1)/d} - w) \mid W_{t/d} = w\right] d, \end{aligned}$$

for some w^* on the segment from $W_{(t+1)/d}$ to w, where $\nabla^2 h$ denotes the Hessian of h.

Recall that the support of h is compact. Proposition 3.4.2 therefore implies that the first term above converges to $Gh(w) = \nabla h(w)^\top c(w)$, uniformly over $w \in \bar{\mathcal{S}}$. We claim that the second term converges to zero uniformly over the same set. To see this, since $h \in \mathcal{C}^\infty$, it follows that there exists an $M < \infty$ such that $\sup_{x \in \bar{\mathcal{S}}} \|\nabla^2 h(x)\|_\infty \leq M$, where $\|\cdot\|_\infty$ is the sup-norm. It follows that

$$(W_{(t+1)/d} - w)^\top \nabla^2 h(w^*)(W_{(t+1)/d} - w) \leq M \|W_{(t+1)/d} - w\|^2.$$

Proposition 3.4.3 then takes us to the claimed convergence.

Altogether, we have that

$$\lim_{d \to \infty} \sup_{w \in \bar{\mathcal{S}}} \left| G^{(d)}h(w) - Gh(w) \right| = 0, \tag{A.4.2}$$

that is the convergence of the infinitesimal generators. The limit (A.4.2) is analogous to Christensen et al. (2005, Eqn. 7).

Convergence of stochastic process

The final part of the proof promotes the convergence of the infintesimals $G^{(d)} \to G$ to the weak convergence of the stochastic processes $W^{(d)} \implies w$. Since $\mathcal{C}^\infty$ is a core for the generator G, the limit (A.4.2) is equivalent to point (i) of Kallenberg (2021,

Theorem 17.28). Point (iv) of the same theorem concludes the proof.

A.4.8 Proof of Proposition 3.4.5

Recall that the drift is $c_\ell(x,y,v) = (a_\ell(x), a_\ell(y), b_\ell(x,y,v))$, where

$$\begin{aligned} a_\ell(x) &= \ell^2(1-2x)e^{\ell^2(x-1)/2}\Phi\Big(\frac{\ell}{2x^{1/2}} - \ell x^{1/2}\Big) + \ell^2\Phi\Big(-\frac{\ell}{2x^{1/2}}\Big), \\ b_\ell(x,y,v) &= \ell^2\mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho(x,y,v))}\Big[1 \wedge e^{\ell x^{1/2}Z_1 - \ell^2/2} \wedge e^{\ell y^{1/2}Z_2 - \ell^2/2}\Big] \\ &\quad - \ell^2 v\Big[e^{\ell^2(x-1)/2}\Phi\Big(\frac{\ell}{2x^{1/2}} - \ell x^{1/2}\Big) + e^{\ell^2(y-1)/2}\Phi\Big(\frac{\ell}{2y^{1/2}} - \ell y^{1/2}\Big)\Big], \end{aligned}$$

and where $\rho(\cdot)$ is coupling-specific:

$$\rho_{\mathrm{crn}}(x,y,v) = \frac{v}{(xy)^{1/2}}, \quad \rho_{\mathrm{refl}}(x,y,v) = \frac{2xy - (x+y)v}{(xy)^{1/2}(x+y-2v)}, \quad \rho_{\mathrm{gcrn}}(x,y,v) = 1.$$

The fixed points are the solutions of $a_\ell(x) = a_\ell(y) = b_\ell(x,y,v) = 0$. Since all but one of the coordinates of the ODE are autonomous, the fixed points are stable if and only if $\partial_x a_\ell(x) < 0$, $\partial_y a_\ell(y) < 0$ and $\partial_v b_\ell(x,y,v) < 0$. The remainder of the proof is entirely elementary calculus.

Fixed points and their stability

We start with the fixed points and their linear stability analysis. Firstly, by Kuntz et al. (2019, Lemma 4.1): $a_\ell(x) > 0$ for $x \in [0,1)$; $a_\ell(1) = 0$; $a_\ell(x) < 0$ for $x \in (1,\infty)$. It follows that the unique solution of $a_\ell(x) = 0$ is $x^* = 1$, and furthermore this solution must be stable.

The fixed points are therefore of the form $(1,1,v^*)$, with stability in the first two coordinates and where $v^* \in [-1,1]$ is a root of $b_\ell(1,1,v) = 0$. By Lemma A.4.9, we can re-write this equation to

$$h_\ell(\rho(v)) - v h_\ell(1) = 0,$$

where $h_\ell(\rho) = \mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}\left[1 \wedge e^{\ell Z_1 - \ell^2/2} \wedge e^{\ell Z_2 - \ell^2/2}\right]$ and $\rho(\cdot)$ is coupling-specific and takes the values

$$\rho_{\mathrm{crn}}(v) = v, \quad \rho_{\mathrm{refl}}(v) = \rho_{\mathrm{gcrn}}(v) = 1.$$

For both the GCRN and reflection couplings, since $h_\ell(1) > 0$ it follows that $v^* = 1$ is the unique fixed point, and it is stable since $\partial_v g_\ell(1) = -h_\ell(1) < 0$. For the CRN coupling, let $g_\ell(v) = h_\ell(v) - v h_\ell(1)$. By Lemma A.4.9, g_ℓ is convex on $[0,1]$ and satisfies $g_\ell(0) > 0$, $g_\ell(1) = 0$ and $\lim_{v\to 1} \partial_v g_\ell(v) = \infty$. It follows that there are two fixed points: $v^*_{\mathrm{u}} = 1$, which is unstable and $v^*_{\mathrm{crn}} \in (0,1)$, which is stable as we must have $g_\ell(v^*_{\mathrm{crn}}) < 0$ due to the convexity of g_ℓ. This concludes the proof.

A.4.9 Proof of Theorem 3.4.1

Condition on the same quantities as in the proof of Proposition 3.4.2.

GCRN attains the claimed upper bound

From the proof of Proposition 3.4.2, under the GCRN coupling it holds that

$$\lim_{d\to\infty} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\ell x^{1/2} Z - \ell^2/2} \wedge e^{\ell y^{1/2} Z - \ell^2/2}\right].$$

This coincides with the claimed limit supremum.

To complete the proof, it is sufficient to show that the quantity on the right-hand side is an upper upper bound on the limit supremum over $\mathcal{P}$ of the left-hand side. We now show this by an argument similar to the proof of Lemma A.4.4.

Claimed upper bound on the limit supremum

We have the sequence of bounds

$$\begin{aligned}
\mathbb{E}[h^2 Z_x^\top Z_y B_x B_y] &= \ell^2 \mathbb{E}[B_x B_y] + \ell^2 \mathbb{E}\left[\left(\frac{1}{d} Z_x^\top Z_y - 1\right) B_x B_y\right] \\
&\leq \ell^2 \mathbb{E}[B_x B_y] + \frac{\ell^2}{2} \mathbb{E}\left[\Big(\frac{1}{d}\|Z_x\|^2 + \frac{1}{d}\|Z_y\|^2 - 2\Big) B_x B_y\right] \\
&\leq \ell^2 \mathbb{E}[B_x B_y] + \frac{\ell^2}{2} \mathbb{E}\left[\Big|\frac{1}{d}\|Z_x\|^2 + \frac{1}{d}\|Z_y\|^2 - 2\Big|\right] \\
&\leq \ell^2 \mathbb{E}[B_x B_y] + \ell^2 \mathbb{E}\left[\Big|\frac{1}{d}\|Z_x\|^2 - 1\Big|\right],
\end{aligned}$$

where in the second line we have used that $B_x B_y \geq 0$ and that $x^\top y \leq (\|x\|^2 + \|y\|^2)/2$ for all $x, y \in \mathbb{R}^d$; in the third line that $B_x B_y \leq 1$; in the final line the triangle inequality and that $\{Z_x, Z_y\}$ are equal in distribution. Since $\lim_{d\to\infty} \|Z_x\|^2/d = 1$ in L_1, the second term in the bound converges to zero; moreover, the convergence is uniform over the kernel coupling $\bar{K}$ used.

We now bound the first term above. Let $(Z_1, Z_2) = (n_x^\top Z_x, n_y^\top Z_y)$ so that

$$\mathbb{E}[B_x B_y] = \mathbb{E}\left[\mathbb{1}\{U_x \leq e^{\ell x^{1/2} Z_1 - (\ell^2/2)\|Z_x\|^2/d}\}\mathbb{1}\{U_y \leq e^{\ell y^{1/2} Z_2 - (\ell^2/2)\|Z_y\|^2/d}\}\right].$$

Since we are in the class $\mathcal{P}$ of product couplings, we have that (U_x, U_y) are independent of $\{Z_x, Z_y, Z_1, Z_2\}$. Taking expectations with respect to the uniforms first, it follows that

$$\begin{aligned}
\mathbb{E}[B_x B_y] &\leq \mathbb{E}\left[1 \wedge e^{\ell x^{1/2} Z_1 - (\ell^2/2)\|Z_x\|^2/d} \wedge e^{\ell y^{1/2} Z_2 - (\ell^2/2)\|Z_y\|^2/d}\right] \\
&\leq \mathbb{E}\left[1 \wedge e^{\ell x^{1/2} Z_1 - \ell^2/2} \wedge e^{\ell y^{1/2} Z_2 - \ell^2/2}\right] + \ell^2 \mathbb{E}\left[1 \wedge \Big|\frac{1}{d}\|Z_x\|^2 - 1\Big|\right],
\end{aligned}$$

where finally we have used Lemma A.4.2 and that $\{Z_x, Z_y\}$ are equal in distribution. The second term converges to zero as $d \to \infty$, uniformly over the kernel coupling $\bar{K}$

used. For the first term, Lemma A.4.6 implies that

$$\sup_{\bar{K}\in\mathcal{P}} \mathbb{E}\left[1 \wedge e^{\ell x^{1/2} Z_1 - \ell^2/2} \wedge e^{\ell y^{1/2} Z_2 - \ell^2/2}\right] = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\ell x^{1/2} Z - \ell^2/2} \wedge e^{\ell y^{1/2} Z - \ell^2/2}\right],$$

since $Z_1, Z_2 \sim \mathcal{N}(0,1)$ and since $f_a(z) = 1 \wedge \exp(az - \ell^2/2)$ is increasing for all $a \geq 0$.

Putting all bounds together, we have that

$$\lim_{d\to\infty} \sup_{\bar{K}\in\mathcal{P}} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] \leq \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\ell x^{1/2} Z - \ell^2/2} \wedge e^{\ell y^{1/2} Z - \ell^2/2}\right].$$

This concludes the proof.

A.4.10 Proof of Theorem 3.5.2

This proof is virtually identical to that of Theorem 3.4.1. Recall that the target is $\pi = \mathcal{N}(0, \Omega^{-1})$ and that we have defined the inner product $\langle x, y\rangle_{[k]} = x^\top \Omega^k y$ and the squared norm $\|x\|_{[k]}^2 = x^\top \Omega^k x$. Throughout the proof, we condition on

$$\left(\|X_t\|_{[2]}^2, \|Y_t\|_{[2]}^2, \langle X_t, Y_t\rangle_{[2]}\right)/(z_1 d) = (x_2, y_2, v_2).$$

We have that $B_x = \mathbb{1}\{U \leq \exp(-h\langle X_t, Z_x\rangle_{[1]} - h^2\|Z_x\|_{[1]}^2/2)\}$ is the acceptance step, with the analogous expression for B_y with the same $U \sim \text{Unif}(0,1)$.

GCRN attains the claimed upper bound

Fix the coupling to be GCRN. By Proposition 3.3.1, we have that

$$(-h\langle X_t, Z_x\rangle_{[1]}, -h\langle Y_t, Z_y\rangle_{[1]}) = (\ell^2 z_1 x_2 Z, \ell^2 z_1 y_2 Z) = (\lambda x_2 Z, \lambda y_2 Z)$$

where $Z \sim \mathcal{N}(0,1)$. The following limits hold as $d \to \infty$: $h^2\|Z_x\|_{[1]}^2 \to \ell^2 z_1 = \lambda^2$ and $h^2\|Z_y\|_{[1]}^2 \to \lambda^2$ in probability (by Assumption 3.5.1 and the law of large numbers);

$h^2 Z_x^\top Z_y \to \ell^2$ in L_1. By Lemma A.4.4, it follows that

$$\lim_{d\to\infty} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\lambda x_2^{1/2} Z - \lambda^2/2} \wedge e^{\lambda y_2^{1/2} Z - \lambda^2/2}\right],$$

so the GCRN coupling attains the upper bound claimed in Theorem 3.5.2.

To complete the proof, we show that the above quantity is indeed an upper upper bound on the limit supremum over $\mathcal{P}$ of the left-hand side.

Upper bound on limit supremum

As in the proof of Theorem 3.4.1, we have the coupling-independent bound

$$\begin{aligned}\sup_{\bar{K}\in\mathcal{P}} \mathbb{E}[h^2 Z_x^\top Z_y B_x B_y] \le \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\lambda x_1^{1/2} Z - \lambda^2/2} \wedge e^{\lambda y^{1/2} Z - \lambda^2/2}\right] \\ + \ell^2 \mathbb{E}\left[\left|\frac{1}{d}\|Z_x\|^2 - 1\right|\right] + \ell^2 \mathbb{E}\left[1 \wedge \left|h^2 \|Z_x\|_{[1]}^2 - \lambda^2\right|\right].\end{aligned}$$

The second term tends to zero as $d \to \infty$, and by Assumption 3.5.1 so does the third. It follows that

$$\lim_{d\to\infty} \sup_{\bar{K}\in\mathcal{P}} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] \le \ell^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[1 \wedge e^{\lambda x_2^{1/2} Z - \lambda^2/2} \wedge e^{\lambda y_2^{1/2} Z - \lambda^2/2}\right],$$

which concludes the proof.

A.4.11 Proof of Proposition 3.5.5

This proof almost is identical to that of Proposition 3.4.2, though some additional bookkeeping is required. Recall that the target is $\pi = \mathcal{N}(0, \Omega^{-1})$ and that we have defined the inner product $\langle x, y\rangle_{[k]} = x^\top \Omega^k y$ and the squared norm $\|x\|_{[k]}^2 = x^\top \Omega^k x$. Recall that we condition on

$$\left(\|X_t\|_{[j]}^2, \|Y_t\|_{[j]}^2, \langle X_t, Y_t\rangle_{[j]}\right) / (z_{j-1} d) = (x_j, y_j, v_j) \quad \text{for all } j \in \{-1, 0, 1, 2\}$$

in all expectations below.

Unless specified otherwise, the claims that follow hold over all $k \in \{-1, 0, 1\}$. The relevant one-step differences are

$$\begin{aligned} \|X_{t+1}\|_{[k]}^2 - \|X_t\|_{[k]}^2 &= \left\{2h\langle X_t, Z_x\rangle_{[k]} + h^2\|Z_x\|_{[k]}^2\right\} B_x, \\ \langle X_{t+1}, Y_{t+1}\rangle_{[k]} - \langle X_t, Y_t\rangle_{[k]} &= h\langle Y_t, Z_x\rangle_{[k]} B_x + h\langle X_t, Z_y\rangle_{[k]} B_y + h^2\langle Z_x, Z_y\rangle_{[k]} B_x B_y. \end{aligned} \tag{A.4.3}$$

where $B_x = \mathbb{1}\{U \leq \exp(-h\langle X_t, Z_x\rangle_{[1]} - h^2\|Z_x\|_{[1]}^2/2)\}$, with the analogous expression for B_y with the same $U \sim \text{Unif}(0,1)$. All terms except for $h^2\langle Z_x, Z_y\rangle_{[k]} B_x B_y$ are coupling-independent, so we deal with them first.

Coupling-independent terms

Integrating over $U \sim \text{Unif}(0,1)$, we have that

$$\begin{aligned} \mathbb{E}\left[\|X_{t+1}\|_{[k]}^2 - \|X_t\|_{[k]}^2\right] &= \mathbb{E}\left[(-2Z_2 + h^2\|Z_x\|_{[k]})1 \wedge \exp\left(Z_1 - h^2\|Z_x\|_{[1]}^2/2\right)\right], \\ \mathbb{E}\left[h\langle Y_t, Z_x\rangle_{[k]} B_x\right] &= -\mathbb{E}\left[Z_3 1 \wedge \exp\left(Z_1 - h^2\|Z_x\|_{[1]}^2/2\right)\right], \end{aligned}$$

where $(Z_1, Z_2, Z_3) = (-h\langle X_t, Z_x\rangle_{[1]}, -h\langle X_t, Z_x\rangle_{[k]}, -h\langle Y_t, Z_x\rangle_{[k]})$. By Lemma A.4.5, we have that

$$\begin{aligned} (Z_1, Z_2) &\sim \text{BvN}(\ell^2 z_1 x_2, \ell^2 z_{2k-1} x_{2k}; \ell^2 z_k x_{k+1}) \\ (Z_1, Z_3) &\sim \text{BvN}(\ell^2 z_1 x_2, \ell^2 z_{2k-1} y_{2k}; \ell^2 z_k v_{k+1}). \end{aligned}$$

By Assumption 3.5.3 and the law of large numbers, we have the following limits in L_1 as $d \to \infty$: $h^2\|Z_x\|_{[1]}^2 \to \ell^2 z_1$ and $h^2\|Z_x\|_{[k]}^2 \to \ell^2 z_k$. By Lemmas A.4.3 and A.4.4, and using the short-hand $\lambda^2 = \ell^2 z_1$, it follows that

$$\begin{aligned} \lim_{d\to\infty} \mathbb{E}\left[\|X_{t+1}\|_{[k]}^2 - \|X_t\|_{[k]}^2\right] &= \mathbb{E}\left[\left(-2Z_2 + \ell^2 z_k\right) 1 \wedge \exp\left(Z_1 - \lambda^2/2\right)\right], \\ \lim_{d\to\infty} \mathbb{E}\left[h\langle Y_t, Z_x\rangle_{[k]} B_x\right] &= -\mathbb{E}\left[Z_3 1 \wedge \exp\left(Z_1 - \lambda^2/2\right)\right]. \end{aligned}$$

The analytical formulae for these quantities follow from Lemma A.4.8. The expressions for

$$\lim_{d\to\infty} \mathbb{E}\left[\|Y_{t+1}\|_{[k]}^2 - \|Y_t\|_{[k]}^2\right], \quad \lim_{d\to\infty} \mathbb{E}\left[h\langle X_t, Z_y\rangle_{[k]} B_y\right]$$

follow by symmetry.

Correlation of projections

To evaluate the term $h^2\langle Z_x, Z_y\rangle_{[k]} B_x B_y$, we must express $\rho = \mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_y)$ in terms of the quantities we have conditioned on. Recall that $e = (X_t - Y_t)/\|X_t - Y_t\|$; the normalized gradient at X_t is $n_x = -\Omega X_t/\|\Omega X_t\|$. We turn to Proposition 3.3.1 and compute

$$n_x^\top n_y = \frac{\langle X_t, Y_t\rangle_{[2]}}{\|X_t\|_{[2]}\|Y_t\|_{[2]}} = \frac{v_2}{(x_2y_2)^{1/2}}, \quad n_x^\top e = \frac{\langle X_t, Y_t\rangle_{[1]} - \|X_t\|_{[1]}^2}{\|X_t\|_{[2]}\|X_t - Y_t\|_{[0]}} = \frac{v_1 - x_1}{(z_1x_2)^{1/2}\{z_{-1}(x_0 + y_0 - 2v_0)\}^{1/2}},$$

and also $n_y^\top e = (v_1 - y_1)(z_1y_2)^{-1/2}\{z_{-1}(x_0 + y_0 - 2v_0)\}^{-1/2}$ by symmetry. Plugging these into Proposition 3.3.1, and using that $\varepsilon = z_1z_{-1}$, we have that

$$\rho_{\mathrm{crn}} = \frac{v_2}{(x_2y_2)^{1/2}}, \quad \rho_{\mathrm{refl}} = \frac{v_2}{(x_2y_2)^{1/2}} + \frac{2(x_1 - v_1)(y_1 - v_1)}{\varepsilon(x_2y_2)^{1/2}(x_0 + y_0 - 2v_0)}, \quad \rho_{\mathrm{gcrn}} = 1,$$

as claimed.

Coupling-dependent term

Integrating over $U \sim \mathrm{Unif}(0, 1)$, we have that

$$\mathbb{E}\left[h^2\langle Z_x, Z_y\rangle_{[k]} B_x B_y\right] = \mathbb{E}\left[h^2\langle Z_x, Z_y\rangle_{[k]} 1 \wedge e^{Z_4 - h^2\|Z_x\|_{[1]}^2/2} \wedge e^{Z_5 - h^2\|Z_y\|_{[1]}^2/2}\right],$$

where $(Z_4, Z_5) = (-h\langle X_t, Z_x\rangle_{[1]}, -h\langle Y_t, Z_y\rangle_{[1]}) = (\lambda x_2^{1/2} n_x^\top Z_x, \lambda y_2^{1/2} n_y^\top Z_y)$. By Proposition 3.3.1, we have that

$$(Z_4, Z_5) \sim \mathrm{BvN}(\lambda^2 x_2, \lambda^2 y_2; \lambda^2 (x_2 y_2)^{1/2} \rho),$$

where $\rho = \mathrm{Cov}(n_x^\top Z_x, n_y^\top Z_y)$ is coupling-specific and evaluated above.

By Assumption 3.5.3, the following limits hold in L_1 as $d \to \infty$: $h^2 \|Z_x\|_{[1]}^2 \to \lambda^2$, $h^2 \|Z_y\|_{[1]}^2 \to \lambda^2$ and $h^2 \langle Z_x, Z_y\rangle_{[k]} \to \ell^2 z_k$. We prove the final limit at the end; it holds since all couplings are low-rank perturbations of the CRN coupling. By Lemma A.4.4 it follows that

$$\lim_{d\to\infty} \mathbb{E}\left[h^2 Z_x^\top Z_y B_x B_y\right] = \ell^2 z_k \mathbb{E}\left[1 \wedge e^{Z_4 - \lambda^2/2} \wedge e^{Z_5 - \lambda^2/2}\right].$$

Putting the limits together concludes the proof, up to showing that $h^2 \langle Z_x, Z_y\rangle_{[k]} \to \ell^2 z_k$ in L_1.

Convergence of limiting inner product

We finally show that, for all $k \in \{-1, 0, 1\}$ and under all considered couplings, it holds that: $\lim_{d\to\infty} \langle Z_x, Z_y\rangle_{[k]}/d \to z_k$ in L_1.

For the CRN coupling, $\langle Z_x, Z_y\rangle_{[k]}/d = \|Z_x\|_{[k]}^2/d$. The claimed limit follows by Assumption 3.5.3.

For the reflection coupling, $\langle Z_x, Z_y\rangle_{[k]} = \|Z_x\|_{[k]}^2 - 2(e^\top Z_x)\langle e, Z_x\rangle_{[k]}$, so it suffices to show that $(e^\top Z_x)\langle e, Z_x\rangle_{[k]}/d \to 0$ in L_1. Now, using the representation (3.5.1) for the precision matrix, $\langle e, Z_x\rangle_{[k]}$ is mean-zero Gaussian with standard deviation at most $\max\{\omega_1^{2k}, \ldots \omega_d^{2k}\}$, so by Cauchy-Schwarz we have that

$$\mathbb{E}\left[|(e^\top Z_x)\langle e, Z_x\rangle_{[k]}|\right]/d \le \mathbb{E}\left[\max\{\omega_1^{2k}, \ldots \omega_d^{2k}\}\right]/d.$$

By Assumption 3.5.3, Lemma A.4.7 applies and the right-hand side tends to zero as $d \to \infty$. This leads us to the claimed convergence for the reflection coupling.

For the GCRN coupling, we have that

$$\begin{aligned}\langle Z_x, Z_y\rangle_{[k]} &= \langle Z - (n_x^\top Z_x + Z_1)n_x, Z - (n_y^\top Z_y + Z_1)n_y\rangle_{[k]} \\ &= \|Z\|_{[k]}^2 - (n_x^\top Z_x + Z_1)\langle n_x, Z\rangle_{[k]} - (n_y^\top Z_y + Z_1)\langle n_y, Z\rangle_{[k]} \\ &\quad + (n_x^\top Z_x + Z_1)(n_y^\top Z_y + Z_1)\langle n_x, n_y\rangle_{[k]}.\end{aligned}$$

As for the reflection coupling, when scaled by d^{-1}, each of the final three "residual" terms tends to zero in L_1. This is since we can bound the expected absolute value of each residual by $4\mathbb{E}\left[\max\{\omega_1^{2k}, \dots \omega_d^{2k}\}\right]$. The claimed convergence follows, which completes the proof.

A.4.12 Proof of Proposition 3.5.7

We have dealt with the CRN and GCRN couplings in the proof of Proposition 3.4.5.

For the reflection coupling, we seek the roots v^* of

$$h_\ell(v + \varepsilon^{-1}(1-v)) - v h_\ell(1) = 0, \quad v \in [0,1],$$

where $h_\ell(\rho) = \mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}\left[1 \wedge e^{\ell Z_1 - \ell^2/2} \wedge e^{\ell Z_2 - \ell^2/2}\right]$. An equivalent formulation is obtained by reparametrizing with $v = (w - \varepsilon^{-1})/(1-\varepsilon^{-1})$:

$$g_\ell(w) := h_\ell(w) - (w - \varepsilon^{-1})/(1-\varepsilon^{-1}) h_\ell(1) = 0.$$

By Lemma A.4.9, g_ℓ is convex and satisfies $g_\ell(0) > 0$, $g_\ell(w) = 0$ and $\lim_{w\to 1} \partial_w g_\ell(w) = \infty$. It follows that there are two fixed points: $w_{\mathrm{u}}^* = 1$, which is unstable and $w_{\mathrm{refl}}^* \in (0,1)$, which is stable as we must have $g_\ell(w_{\mathrm{refl}}^*) < 0$ due to the convexity of g_ℓ.

Returning to the original parametrization, $v_{\mathrm{u}}^* = 1$ is unstable and $v_{\mathrm{refl}}^* = (w_{\mathrm{refl}}^* -$

$\varepsilon^{-1})/(1-\varepsilon^{-1}) \in (0,1)$ is stable. To show that $v^*_{\text{refl}} \geq v^*_{\text{crn}}$, recall that

$$0 = h_\ell(v^*_{\text{refl}} + \varepsilon^{-1}(1 - v^*_{\text{refl}})) - v^*_{\text{refl}} h_\ell(1) \geq h_\ell(v^*_{\text{refl}}) - v^*_{\text{refl}} h_\ell(1).$$

Given that v^*_{crn} is the unique solution of $h_\ell(v) - v h_\ell(1) = 0$ over $v \in (0,1)$, Lemma A.4.9 and the above inequality imply that $v^*_{\text{refl}} \geq v^*_{\text{crn}}$.

Behaviour with eccentricity ε

Let $f_\ell(\varepsilon, v) = h_\ell(v + \varepsilon^{-1}(1-v)) - v h_\ell(1)$, so that the fixed-point equation is $f_\ell(\varepsilon, v^*_{\text{refl}}) = 0$. By Lemma A.4.9, $f_\ell(\varepsilon, v)$ is decreasing in ε, and furthermore (due to the convexity of f_ℓ in v) it holds that $f_\ell(\varepsilon, v)$ is decreasing in v near v^*_{refl}. Since, $f_\ell(\varepsilon, v^*_{\text{refl}}) = 0$, it follows that v^*_{refl} must be decreasing in ε.

We conclude by showing the limits. Since $f_\ell(1, v) = h_\ell(1) - v h_\ell(1)$, we have that $\lim_{\varepsilon \to 1} v^*_{\text{refl}}(\varepsilon) = 1$. Since $f_\ell(\infty, v) = h_\ell(v) - v h_\ell(1)$, we have that $\lim_{\varepsilon \to \infty} v^*_{\text{refl}}(\varepsilon) = v^*_{\text{crn}}$. This concludes the proof.

A.4.13 Proof of Theorem 3.6.2

We proceed as in the proof of Theorem 3.4.1, first proving the upper bound, then showing that the GCRN coupling attains it.

Showing the upper bound

We have that

$$\begin{aligned}\mathbb{E}[h^2 Z_x^\top Z_y B_x B_y] &= \ell^2 \mathbb{E}[B_x B_y] + \ell^2 \mathbb{E}\left[\left(\frac{1}{d} Z_x^\top Z_y - 1\right) B_x B_y\right] \\ &\leq \ell^2 \mathbb{E}[B_x] + \frac{\ell^2}{2}\left\{\mathbb{E}[(\frac{1}{d}\|Z_x\|^2 + \frac{1}{d}\|Z_y\|^2 - 2) B_x B_y]\right\} \\ &\leq \ell^2 \mathbb{E}[B_x] + \frac{\ell^2}{2}\left\{\mathbb{E}[(\frac{1}{d}\|Z_x\|^2 - 1)_+] + \mathbb{E}[(\frac{1}{d}\|Z_y\|^2 - 1)_+]\right\}, \\ &= \ell^2 \mathbb{E}[B_x] + \ell^2 \mathbb{E}\left[\left(\frac{1}{d}\|Z_x\|^2 - 1\right)_+\right],\end{aligned} \tag{A.4.4}$$

where $(x)_+ = 0 \vee x$ is the positive part of x; we have used that $B_y \in [0,1]$ and $Z_x^\top Z_y \leq (\|Z_x\|^2 + \|Z_y\|^2)/2$ to obtain the second line, and that $B_{x,y} \in [0,1]$ to obtain the third line.

The bound obtained above is invariant to the coupling used. The first term converges to $2\Phi(-\ell(bI)^{1/2}/2)$ by Roberts and Rosenthal (2001, Theorem 5) (this calculation is also performed below), while the second term converges to zero since $\|Z_x\|^2/d \to 1$ in L_2. It follows that

$$\lim_{d\to\infty} \sup_{\bar{K}\in\mathcal{M}} \mathbb{E}[h^2 Z_x^\top Z_y B_x B_y] \leq 2\ell^2 \Phi(-\ell(bI)^{1/2}/2),$$

which is the claimed bound.

Showing that GCRN attains the bound

To show that the GCRN coupling attains the claimed bound, we hereafter restrict to this coupling and we return to the decomposition (A.4.4).

To get a handle on the first term of (A.4.4), we Taylor-expand the logarithm of the acceptance ratio for the X-chain (as in Roberts and Rosenthal, 2001, proof of

Theorem 5),

$$\begin{aligned}
&\log \pi(X + hZ_x) - \log \pi(X) = \\
&\quad = h\nabla \log \pi(X)^\top Z_x + \frac{h^2}{2} Z_x^\top \nabla^2 \log \pi(X) Z_x + R_x \\
&\quad = \left\{ \frac{1}{d} \sum_{i=1}^{d} (\ell\omega_i)^2 \left[(\log f)'(\omega_i X_i) \right]^2 \right\}^{1/2} Z_\nabla + \frac{1}{2d} \sum_{i=1}^{d} (\ell\omega_i Z_{x,i})^2 (\log f)''(\omega_i X_i) + R_x \\
&\quad =: G_x^{1/2} Z_\nabla + H_x,
\end{aligned}$$

where $Z_\nabla \sim \mathcal{N}_1(0,1)$ corresponds to the gradient direction, and R_x is the third-order remainder term.

By the law of large numbers, the following limit holds in probability:

$$\begin{aligned}
\lim_{d\to\infty} G_x &= \ell^2 \mathbb{E}_{X\sim\pi^{(1)}} \left[\omega_1^2 (\log f)'(\omega_i X)^2 \right] \\
&= \ell^2 \mathbb{E}_{Y\sim f} \left[\omega_1^2 (\log f)'(Y)^2 \right] \\
&= \ell^2 b I.
\end{aligned}$$

Our assumptions ensure that the remainder term satisfies $\lim_{d\to\infty} R_x = 0$ in probability (see e.g. Sherlock, 2013, Lemma 6). By this and the law of large numbers, the following limit holds in probability:

$$\begin{aligned}
\lim_{d\to\infty} H_x &= \frac{1}{2} \mathbb{E}_{X\sim\pi^{(1)}} \left[(\ell\omega_1 Z_{x,1})^2 (\log f)''(\omega_1 X) \right] \\
&= \frac{\ell^2 b}{2} \mathbb{E}_{Y\sim f} \left[(\log f)''(Y) \right] \\
&= -\frac{\ell^2 b I}{2},
\end{aligned}$$

where in the last equality we have used the identity $\mathbb{E}_{Y\sim f}[(\log f)''(Y)] = -\mathbb{E}_{Y\sim f}[(\log f)'(Y)^2] = -I$, which follows by integration by parts.

The analogous Taylor expansion for the Y-chain is: $\log \pi(Y + hZ_y) - \log \pi(Y) =$

$G_y^{1/2} Z_\nabla + H_y$, where, by the definition of the GCRN coupling, the random variable $Z_\nabla \sim \mathcal{N}_1(0,1)$ corresponding to the gradient direction is identical to the analogous one for the X-chain. As before, the following limits hold in probability: $\lim_{d\to\infty} G_y = \ell^2 bI$, $\lim_{d\to\infty} H_y = (-\ell^2 bI)/2$.

With the above limits in probability in hand, the first term in (A.4.4) writes as

$$\begin{aligned}\ell^2 \mathbb{E}[B_x B_y] &= \ell^2 \mathbb{E}\left[1 \wedge \exp(G_x^{1/2} Z_\nabla + H_x) \wedge \exp(G_y^{1/2} Z_\nabla + H_y)\right] \\ &= \ell^2 \mathbb{E}\left[1 \wedge \left[\exp(G_x^{1/2} Z_\nabla + H_x)\left\{1 \wedge \exp(\partial G \cdot Z_\nabla + \partial H)\right\}\right]\right],\end{aligned}$$

where both $\partial G := G_y^{1/2} - G_x^{1/2}$ and $\partial H := H_y - H_x$ go to 0 in probability as $d \to \infty$. Several applications of Slutsky's Theorem (ST) and the Continuous Mapping Theorem (CMT) lead us to

$$\lim_{d\to\infty} \ell^2 \mathbb{E}[B_x B_y] = 2\ell^2 \Phi(-\ell (bI)^{1/2}/2).$$

- By ST and CMT, $\lim_{d\to\infty} 1 \wedge \exp(\partial G \cdot Z_\nabla + \partial H) = 1$ in probability.
- Again, by ST and CMT, $\exp(G_x^{1/2} Z_\nabla + H_x)$ converges weakly to a log-normal $L \sim \log \mathcal{N}(-\ell^2 bI/2, \ell^2 bI)$ as $d \to \infty$. A further application of ST, shows that the sequence of non-negative random variables $A_d = \exp(G_x^{1/2} Z_\nabla + H_x)\left[1 \wedge \exp(\partial G \cdot Z_\nabla + \partial H)\right]$ converges weakly to the same limit.
- The function $g(x) = 1 \wedge x$ is bounded for $x \in [0, \infty)$. It follows that $\lim_{d\to\infty} \mathbb{E}[B_x B_y] = \lim_{d\to\infty} \mathbb{E}[1 \wedge A_d] = \mathbb{E}[1 \wedge L]$, by the definition of weak convergence.
- Lemma A.4.8 evaluates $\mathbb{E}[1 \wedge L]$ and completes the calculation.

We therefore have a limit for the first term of (A.4.4). The second term of (A.4.4) satisfies $\lim_{d\to\infty} \ell^2 \mathbb{E}[(Z_x^\top Z_y/d - 1) B_x B_y] = 0$, since $Z_x^\top Z_y/d \to 1$ in L_1 and since $B_x B_y \in [0,1]$. Putting it all together, under the GCRN coupling we have that

$$\lim_{d\to\infty} \mathbb{E}[h^2 Z_x^\top Z_y B_x B_y] = 2\ell^2 \Phi(-\ell (bI)^{1/2}/2),$$

which coincides with the upper bound and therefore concludes the proof of Theorem 3.6.2.

A.4.14 Postponed proofs

Proof of Lemma A.4.8

Let $E_1 = \mathbb{E}\left[Z1 \wedge e^{-\ell\alpha Z - \ell^2/2}\right]$. Using that $\mathrm{d}\phi(z) = -z\phi(z)\mathrm{d}z$, it holds that $\int_a^b z\phi(z)\mathrm{d}z = -\int_a^b \mathrm{d}\phi(z) = \phi(a) - \phi(b)$. It thus follows that

$$\begin{aligned}
E_1 &= \int_{-\infty}^{-\ell/(2\alpha)} z\phi(z)\mathrm{d}z + \int_{-\ell/(2\alpha)}^{\infty} ze^{-\ell\alpha z - \ell^2/2}\phi(z)\mathrm{d}z \\
&= -\phi\left(-\frac{\ell}{2\alpha}\right) + e^{\ell^2(\alpha^2-1)/2}\int_{-\ell/(2\alpha)}^{\infty} z\phi(z+\ell\alpha)\mathrm{d}z \\
&= -\phi\left(-\frac{\ell}{2\alpha}\right) + e^{\ell^2(\alpha^2-1)/2}\left(\int_{-\ell/(2\alpha)}^{\infty} (z+\ell\alpha)\phi(z+\ell\alpha)\mathrm{d}z - \ell\alpha\int_{-\ell/(2\alpha)}^{\infty}\phi(z+\ell\alpha)\mathrm{d}z\right) \\
&= -\phi\left(-\frac{\ell}{2\alpha}\right) + e^{\ell^2(\alpha^2-1)/2}\left(\int_{-\ell/(2\alpha)+\ell\alpha}^{\infty} z\phi(z)\mathrm{d}z - \ell\alpha\int_{-\ell/(2\alpha)+\ell\alpha}^{\infty}\phi(z)\mathrm{d}z\right) \\
&= -\phi\left(-\frac{\ell}{2\alpha}\right) + e^{\ell^2(\alpha^2-1)/2}\phi\left(-\frac{\ell}{2\alpha}+\ell\alpha\right) - \ell\alpha e^{\ell^2(\alpha^2-1)/2}\Phi\left(\frac{\ell}{2\alpha}-\ell\alpha\right),
\end{aligned}$$

where we used the identity $\phi(z) = e^{\ell^2\alpha^2/2+\ell\alpha z}\phi(z+\ell\alpha)$ in the second line. The desired formula follows by applying the same identity with $z = -\ell/(2\alpha)$; the first two terms in the final expression cancel.

The second expectation equivalently writes as $E_2 = \mathbb{E}\left[1 \wedge e^{-\ell m Z - \ell^2/2} \wedge e^{-\ell M Z - \ell^2/2}\right]$, where $m = \alpha \wedge \beta$ and $M = \alpha \vee \beta$. The form of the integrand depends on the sign of Z.

Therefore,

$$\begin{aligned}
E_2 &= \int_{-\infty}^{-\ell/(2m)} \phi(z)\mathrm{d}z + \int_{-\ell/(2m)}^{0} e^{-\ell m z - \ell^2/2}\phi(z)\mathrm{d}z + \int_{0}^{\infty} e^{-\ell M z - \ell^2/2}\phi(z)\mathrm{d}z \\
&= \Phi\left(-\frac{\ell}{2m}\right) + e^{\ell^2(m^2-1)/2}\int_{-\ell/(2m)}^{0} \phi(z+\ell m)\mathrm{d}z + e^{\ell^2(M^2-1)/2}\int_{0}^{\infty} \phi(z+\ell M)\mathrm{d}z \\
&= \Phi\left(-\frac{\ell}{2m}\right) + e^{\ell^2(m^2-1)/2}\int_{-\ell/(2m)+\ell m}^{\ell m} \phi(z)\mathrm{d}z + e^{\ell^2(M^2-1)/2}\int_{\ell M}^{\infty} \phi(z)\mathrm{d}z \\
&= \Phi\left(-\frac{\ell}{2m}\right) + e^{\ell^2(m^2-1)/2}\left\{\Phi\left(\frac{\ell}{2m} - \ell m\right) - \Phi(-\ell m)\right\} + e^{\ell^2(M^2-1)/2}\Phi(-\ell M),
\end{aligned}$$

where we used identity $e^{-\ell\alpha z}\phi(z) = e^{\ell^2\alpha^2/2}\phi(z+\ell\alpha)$ with $\alpha \in \{m, M\}$ in the second line. This completes the proof.

Proof of Lemma A.4.9

Proof of claim 1. This is an immediate consequence of Lemma A.4.8.

Proof of claim 2. Firstly, the integral re-writes as

$$h(\rho;\ell) = \mathbb{E}_{(Z_1,Z_2)\sim \mathrm{BvN}(\rho)}\left[\exp\left\{0 \wedge (\ell Z_1 - \ell^2/2) \wedge (\ell Z_2 - \ell^2/2)\right\}\right].$$

We use the reparametrization trick $Z_2 = \rho Z_1 + \sqrt{1-\rho^2}Z_*$, where $Z_* \sim \mathcal{N}_1(0,1)$ is independent of Z_1. This expresses $h(\cdot)$ as an integral over randomness (Z_1, Z_*) which does not depend on ρ; thereafter, only Z_2 in the integrand depends on ρ. Differentiating and bringing the derivative inside the integral, we obtain

$$\partial_\rho h(\rho;\ell) = \mathbb{E}\left[\mathbb{1}\{Z_2 \le \ell/2\}\mathbb{1}\{Z_2 \le Z_1\}\partial_\rho(\ell Z_2 - \ell^2/2)e^{\ell Z_2 - \ell^2/2}\right].$$

We use a second reparametrization trick: $Z_1 = \rho Z_2 + \sqrt{1-\rho^2}Z_{**}$, where $Z_{**} \sim \mathcal{N}_1(0,1)$

is independent of Z_2. We can now evaluate:

$$\partial_\rho Z_2 = Z_1 - \frac{\rho}{\sqrt{1-\rho^2}} Z_* = Z_1 - \frac{\rho}{\sqrt{1-\rho^2}} \frac{Z_2 - \rho Z_1}{\sqrt{1-\rho^2}} = \frac{Z_1 - \rho Z_2}{1-\rho^2} = \frac{1}{\sqrt{1-\rho^2}} Z_{**},$$
$$\mathbb{1}\{Z_2 \le Z_1\} = \mathbb{1}\left\{Z_2 \le \rho Z_2 + \sqrt{1-\rho^2} Z_{**}\right\} = \mathbb{1}\left\{\sqrt{\frac{1-\rho}{1+\rho}} Z_2 \le Z_{**}\right\}.$$

Therefore,

$$\begin{aligned}\partial_\rho h(\rho;\ell) &= \frac{\ell}{\sqrt{1-\rho^2}} \mathbb{E}\left[\mathbb{1}\{Z_2 \le \ell/2\} \mathbb{1}\left\{\sqrt{\frac{1-\rho}{1+\rho}} Z_2 \le Z_{**}\right\} Z_{**} e^{\ell Z_2 - \ell^2/2}\right] \\ &= \frac{\ell}{\sqrt{1-\rho^2}} \mathbb{E}\left[\mathbb{1}\{Z_2 \le \ell/2\} \phi\left(\sqrt{\frac{1-\rho}{1+\rho}} Z_2\right) e^{\ell Z_2 - \ell^2/2}\right],\end{aligned}$$

where in the last line we have used that $\mathbb{E}[Z_{**} \mathbb{1}\{x \le Z_{**}\}] = \phi(x)$ for all x. It follows that $\partial_\rho h(\rho) > 0$ for all $\rho \in (-1, 1)$, and that $\lim_{\rho \nearrow 1} \partial_\rho h(\rho) = \infty$, as claimed.

Proof of claim 3. Firstly, repeated applications of the chain rule yield:

$$\partial_\rho \left\{\phi\left(\sqrt{\frac{1-\rho}{1+\rho}} Z_2\right)\right\} = \frac{1}{(1+\rho)^2} z^2 \phi\left(\sqrt{\frac{1-\rho}{1+\rho}} z\right).$$

(We have used that $\partial_\rho \phi(\sqrt{f(\rho)} z) = -\partial_\rho f(\rho) z^2 \phi(\sqrt{f(\rho)} z)/2$ for any non-negative differentiable $f(\cdot)$, and then that $\partial_\rho f(\rho) = -2/(1+\rho)^2$ when $f(\rho) = (1-\rho)/(1+\rho)$.)

Now, differentiating twice, we have that

$$\begin{aligned}\partial_\rho^2 h(\rho;\ell) &= \partial_\rho \left\{\partial_\rho h(\rho;\ell)\right\} \\ &= \partial_\rho \left\{\frac{\ell}{\sqrt{1-\rho^2}} \mathbb{E}\left[\mathbb{1}\{Z_2 \le \ell/2\} \phi\left(\sqrt{\frac{1-\rho}{1+\rho}} Z_2\right) e^{\ell Z_2 - \ell^2/2}\right]\right\} \\ &= \frac{\rho}{1-\rho^2} \partial_\rho h(\rho;\ell) + \frac{\ell}{\sqrt{1-\rho^2}} \mathbb{E}\left[\mathbb{1}\{Z_2 \le \ell/2\} \frac{1}{(1+\rho)^2} Z_2^2 \phi\left(\sqrt{\frac{1-\rho}{1+\rho}} Z_2\right) e^{\ell Z_2 - \ell^2/2}\right],\end{aligned}$$

where we have used the product rule of differentiation for the final line. By Claim 2, the first term is non-negative for $\rho \in [0, 1)$; the second term is strictly positive over the same range. It follows that $\partial_\rho^2 h(\rho;\ell) > 0$ for $\rho \in [0, 1)$, as claimed. This concludes the

proof of Lemma A.4.9.

Appendix B

Appendix for Chapter 4

B.1 Proofs for Section 4.3

Let $\gamma_L := \mathcal{L}(X_0^{(L)}, Y_0^{(L)})$ for all $L \in \mathbb{N} \cup \{\infty\}$.

Proof of Theorem 4.3.2. Since the function $\mathbb{1}\{x \neq y\}$ is TV-class, for all $t \geq 0$ holds that

$$\left|\mathbb{P}(\tau^{(L)} > t) - \mathbb{P}(\tau^{(\infty)} > t)\right| \leq \mathrm{TV}(\gamma_L \bar{P}^t, \gamma_\infty \bar{P}^t) \leq \mathrm{TV}(\gamma_L, \gamma_\infty),$$

where finally we used the data-processing inequality. Assumption 4.3.1 concludes the proof. □

Proof of Theorem 4.3.4. Using the elementary identity $0 \vee \lceil x \rceil = \mathbb{1}\{x \geq 1\} + 0 \vee \lceil x - 1 \rceil$, we have that

$$\mathrm{TV}^{(L)}(\pi, \pi_t) = \mathbb{E}\left[0 \vee \lceil (\tau^{(L)} - t)/L \rceil\right] = \mathbb{P}\left(\tau^{(L)} > t\right) + \mathbb{E}\left[0 \vee \lceil (\tau^{(L)} - L - t)/L \rceil\right].$$

By Theorem 4.3.2, the first term satisfies $\lim_{L\to\infty} \mathbb{P}(\tau^{(L)} > t) = \mathbb{P}(\tau^{(\infty)} > t)$ for all $t \geq 0$. For the remainder, for all $t \geq 0$ and all $L \geq 1$ we have that

$$0 \leq \mathbb{E}\left[0 \vee \lceil (\tau^{(L)} - L - t)/L \rceil\right] \leq \mathbb{E}\left[0 \vee \lceil (\tau^{(L)} - L)/L \rceil\right] \leq \mathbb{E}\left[\tau^{(L)}/L\right],$$

which decays to zero as $L \to \infty$ by Assumption 4.3.3. This concludes the proof. $\square$

Proof of Theorem 4.3.8. We define

$$B^{(\cdot)}_{k,m,L} := \frac{1}{m-k+1} \sum_{t=k}^{\tau^{(\cdot)}-1} c_{k,m,L}(t) \left\{ h(X^{(\cdot)}_t) - h(Y^{(\cdot)}_t) \right\},$$

so that $B^{(L)}_{k:m} = B^{(L)}_{k,m,L}$. We work with the definition of convergence in probability. Fix $\varepsilon > 0$. Since the function $\mathbb{1}\{|x| > \varepsilon\}$ is TV-class, we have that

$$\mathbb{P}\left(\left| f(m) B^{(L)}_{k,m,L} \right| > \varepsilon \right) \leq \mathbb{P}\left(\left| f(m) B^{(\infty)}_{k,m,L} \right| > \varepsilon \right) + \mathrm{TV}\left(\mathcal{L}(f(m)\mathrm{BC}^{(L)}_{k,m,L}), \mathcal{L}(f(m) B^{(\infty)}_{k,m,L}) \right).$$

For the first term, since we can bound the coefficients as $\sup_{t,k} |c_{k,m,L}(t)| \leq m/L + 1$ and since $m = \Theta(L)$, there exists a constant $C > 0$ so that $\sup_{t,k} |c_{k,m,L}(t)| \leq C < \infty$. It follows that

$$(m-k+1)|B^{(\infty)}_{k,m,L}| \leq C \sum_{t=0}^{\tau^{(\infty)}-1} \left\{ h(X^{(\infty)}_t) - h(Y^{(\infty)}_t) \right\}.$$

The right-hand-side does not depend on L, and furthermore it is finite since $\tau^{(\infty)} < \infty$ almost surely. Since $m = \Theta(L)$ and $f(m) = o(m)$, it follows that $f(m) B^{(\infty)}_{k,m,L} \xrightarrow{p} 0$ as $L \to \infty$.

For the second term, we construct an explicit coupling between $B^{(L)}_{k,m,L}$ and $B^{(\infty)}_{k,m,L}$. We synchronize the transitions $\bar{P}$ of the processes $(X^{(L)}, Y^{(L)})$ and $(X^{(\infty)}, Y^{(\infty)})$, which ensures that $B^{(L)}_{k,m,L} = B^{(\infty)}_{k,m,L}$ whenever $(X^{(L)}_0, Y^{(L)}_0) = (X^{(\infty)}_0, Y^{(\infty)}_0)$. Now, by maximally coupling the joint initializations we have that

$$\mathrm{TV}\left(\mathcal{L}(f(m) B^{(L)}_{k,m,L}), \mathcal{L}(f(m) B^{(\infty)}_{k,m,L}) \right) \leq \mathrm{TV}(\gamma_L, \gamma_\infty) \to 0 \text{ as } L \to \infty.$$

Putting it all together, we have that $f(m) B^{(L)}_{k,m,L} \xrightarrow{p} 0$ as $L \to \infty$, as desired. $\square$

Proof of Theorem 4.3.10. Assumption 4.3.9 ensures that the MCMC CLT

$$\sqrt{m-k+1}\,(h_{k:m} - \mathbb{E}_\pi[h(X)]) \implies \mathcal{N}(0, v(P,h)) \text{ as } m \to \infty$$

holds (Jones, 2004, Theorem 9). Since $m = \Theta(L)$, the MCMC CLT equivalently holds as $L \to \infty$. By Theorem 4.3.8, Assumptions 4.3.1 and 4.3.7 and $m = \Theta(L)$ ensure that $\sqrt{m-k+1}B_{k:m}^{(L)} \xrightarrow{p} 0$ as $L \to \infty$. Since $H_{k:m}^{(L)} = h_{k:m} + B_{k:m}^{(L)}$, Slutsky's theorem concludes the proof. □

B.2 On the AR(1) case study

B.2.1 Auxiliary results

Our first auxiliary result establishes that the reflection-maximal coupling of two AR(1) processes is an exact time-discretization of a reflection coupling of Ornstein-Uhlenbeck (OU) processes.

Proposition B.2.1 (Equivalence of reflection couplings)**.** *Let $(X_t, Y_t)_{t\geq 0}$ be a reflection-maximal coupling of one-dimensional AR(1) processes with marginal transition kernels $P(x,\cdot) = \mathcal{N}(\rho x, 1-\rho^2)$. Let $(\tilde{X}_t, \tilde{Y}_t)$ be a reflection coupling of two Ornstein-Uhlencheck processes, with marginal dynamics $\mathrm{d}\tilde{X}_t = -\frac{1}{2}\tilde{X}_t\mathrm{d}t + \mathrm{d}W_t$, where W is Brownian motion, such that $(\tilde{X}, \tilde{Y})$ coalesce at their first meeting time.*

Then, for all (x,y) and all $\mathrm{d}z$ it holds that

$$\mathbb{P}((X_t, Y_t)_{t\geq 0} \in \mathrm{d}z \mid (X_0, Y_0) = (x,y)) = \mathbb{P}((\tilde{X}_{t\delta}, \tilde{X}_{t\delta})_{t\geq 0} \in \mathrm{d}z \mid (\tilde{X}_0, \tilde{Y}_0) = (x,y)),$$

where $\delta = -2\log\rho$.

Proof. Firstly, it is a standard result that an AR(1) process is an exact time-discretization

of an OU process, in that for all integer $t \geq 0$, all x, and all $\mathrm{d}z$, it holds that

$$\mathbb{P}\left(\tilde{X}_{(t+1)\delta} \in \mathrm{d}z \mid \tilde{X}_{t\delta} = x\right) = \mathbb{P}\left(X_{t+1} \in \mathrm{d}z \mid X_t = x\right),$$

where $\delta = -2\log\rho$.

To see that the equivalence also holds for the joint processes, by Connor (2007, Lemma 3.14), for all $t \geq 0$ the coupling of the transition $(\tilde{X}_{(t+1)\delta}, \tilde{Y}_{(t+1)\delta}) \mid (\tilde{X}_{t\delta}, \tilde{Y}_{t\delta})$ is maximal. Since the coupling of $(\tilde{X}_t, \tilde{Y}_t)_{t\geq 0}$ is by reflection, it follows that the coupling of $(\tilde{X}_{(t+1)\delta}, \tilde{Y}_{(t+1)\delta}) \mid (\tilde{X}_{t\delta}, \tilde{Y}_{t\delta})$ must be reflection-maximal. We thus obtain the following equivalence between the joint transition kernels,

$$\mathbb{P}\left((\tilde{X}_{(t+1)\delta}, \tilde{Y}_{(t+1)\delta}) \in \mathrm{d}z \mid (\tilde{X}_{t\delta}, \tilde{Y}_{t\delta}) = (x, y)\right) = \mathbb{P}\left((X_{t+1}, Y_{t+1}) \in \mathrm{d}z \mid (X_t, Y_t) = (x, y)\right),$$

for all integer $t \geq 0$, all (x, y), and all $\mathrm{d}z$. The claimed result follows from the Markov property of the joint processes. □

Our second auxiliary result provides the survivor function of the meeting time, for a reflection coupling of Ornstein-Uhlenbeck processes.

Proposition B.2.2 (Meeting time under reflection coupling)**.** *Let $(\tilde{X}_t, \tilde{Y}_t) \mid (\tilde{X}_0, \tilde{Y}_0) = (x, y)$ be a reflection coupling of two Ornstein-Uhlencheck processes, with joint dynamics*

$$\mathrm{d}\tilde{X}_t = -\frac{1}{2}\tilde{X}_t\mathrm{d}t + \mathrm{d}W_t, \quad \mathrm{d}\tilde{Y}_t = -\frac{1}{2}\tilde{Y}_t\mathrm{d}t - \mathrm{d}W_t$$

where W is Brownian motion. Let $\tau = \inf\{t : \tilde{X}_t = \tilde{Y}_t\}$ be the first meeting time of the processes. Then, it holds that

$$\mathbb{P}(\tau > t \mid \tilde{X}_0, \tilde{Y}_0) = 2\Phi\left(\frac{|\tilde{X}_0 - \tilde{Y}_0|}{2\sqrt{e^t - 1}}\right) - 1.$$

Proof. These derivations are entirely standard and can be found in e.g. Connor (2007,

Section 3.4.2).

Firstly, we consider the difference process $D_t := \tilde{X}_t - \tilde{Y}_t$, which has dynamics

$$\mathrm{d}D_t = -\frac{1}{2}D_t\mathrm{d}t + 2\mathrm{d}W_t.$$

The first zero-hitting time of the difference process is precisely the meeting time

$$\tau = \inf\{t : \tilde{X}_t = \tilde{Y}_t\} = \inf\{t : D_t = 0\}.$$

Since the difference process is OU, by Doob's transform, it can be rewritten in terms of a time-changed and rescaled Brownian motion $\tilde{W}$ as $D_t = e^{-t/2}(D_0 + \tilde{W}_{f(t)})$, where $f(t) = 4(e^t - 1)$. Defining $\tau_{\mathrm{BM}} = \inf\{t : \tilde{W}_t = 0, \tilde{W}_0 = D_0\}$, it follows that

$$\tau = \inf\{t : D_t = 0\} = \inf\{t : \tilde{W}_{f(t)} = 0, \tilde{W}_{f(0)} = D_0\} = f^{-1}\left(\tau_{\mathrm{BM}}\right).$$

Finally, it is well-known (e.g. Connor, 2007, Eqn. 3.40) that the first zero-hitting time of Brownian motion has survivor function

$$\mathbb{P}(\tau_{\mathrm{BM}} > t \mid \tilde{W}_0 = D_0) = 2\Phi\left(\frac{|D_0|}{\sqrt{t}}\right) - 1.$$

Therefore,

$$\mathbb{P}(\tau > t \mid \tilde{X}_0, \tilde{Y}_0) = \mathbb{P}(\tau_{\mathrm{BM}} > f(t) \mid \tilde{W}_0 = \tilde{X}_0 - \tilde{Y}_0) = 2\Phi\left(\frac{|\tilde{X}_0 - \tilde{Y}_0|}{2\sqrt{e^t - 1}}\right) - 1,$$

which concludes the proof. □

B.2.2 Proofs of main results

Proof of Theorem 4.4.1. By Proposition B.2.1, the reflection-maximal coupling $(X_t^{(L)}, Y_t^{(L)})_{t\geq 0}$ is an exact time-discretization of a reflection coupling of OU processes with time sped

up by a factor of $\delta = -2\log\rho$. By Proposition B.2.2, it follows that the meeting time has survivor function

$$\mathbb{P}(\tau^{(L)} > t \mid X_0^{(L)}, Y_0^{(L)}) = 2\Phi\left(\frac{|X_0^{(L)} - Y_0^{(L)}|}{2\sqrt{e^{t\delta}-1}}\right) - 1 = 2\Phi\left(\frac{|X_0^{(L)} - Y_0^{(L)}|}{2\sqrt{1-\rho^{2t}}}\rho^t\right) - 1,$$

as claimed. □

Proof of Theorem 4.4.2. Using that, for $s \geq t$,

$$\begin{aligned}\mathbb{E}[X_t(X_{s+L} - Y_s)] &= \rho^{s-t}\mathbb{E}[X_t(X_{t+L} - Y_t)],\\ \mathbb{E}[(X_{t+L} - Y_t)(X_{s+L} - Y_s)] &= \rho^{s-t}\mathbb{E}[(X_{t+L} - Y_t)^2],\end{aligned}$$

and carefully summing up several geometric series, we have that

$$\begin{aligned}\mathbb{E}[(H_k^{(L)})^2] &= \mathbb{E}[X_k^2] + 2\sum_{i\geq 0}\mathbb{E}[X_k(X_{k+iL+L} - Y_{k+iL})] + \mathbb{E}\left[\left\{\sum_{i\geq 0}(X_{k+iL+L} - Y_{k+iL})\right\}^2\right]\\
&= \mathbb{E}[X_k^2] + \frac{2}{1-\rho^L}\mathbb{E}[X_k(X_{k+L} - Y_k)] + \sum_{i\geq 0}\mathbb{E}\left[(X_{k+iL+L} - Y_{k+iL})^2\right]\\
&\quad + 2\sum_{j>i}\mathbb{E}\left[(X_{k+iL+L} - Y_{k+iL})(X_{k+jL+L} - Y_{k+jL})\right]\\
&= \mathbb{E}[X_k^2] + \frac{2}{1-\rho^L}\mathbb{E}[X_k(X_{k+L} - Y_k)] + \frac{1+\rho^L}{1-\rho^L}\sum_{i\geq 0}\mathbb{E}\left[\left(X_{k+iL}^{(L)} - Y_{k+iL}^{(L)}\right)^2\right],\end{aligned}$$

where we recalled that $(X_t^{(L)}, Y_t^{(L)}) := (X_{t+L}, Y_t)$ in the last line. This concludes the proof. □

B.2.3 Further calculations

We let $\phi(x \mid \mu, \sigma^2)$ denote the probability distribution function and $\Phi(x \mid \mu, \sigma^2)$ denote the cumulative distribution function of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

Unconditional survivor function. When $X_0^{(L)} - Y_0^{(L)} \sim \mathcal{N}(\mu, \sigma^2)$ is Gaussian, by Owen (1980, Eqn. 10,010.2) we can express $\mathbb{P}(\tau^{(L)} > t)$ in terms of bivariate normal probabilities:

$$\mathbb{P}(\tau^{(L)} > t) = 2\,\mathrm{BvN}\left(\frac{\mu}{\sigma}, \frac{\mu}{\sqrt{c(t)+\sigma^2}} \mid \frac{\sqrt{c(t)}}{\sqrt{c(t)+\sigma^2}}\right) + 2\,\mathrm{BvN}\left(-\frac{\mu}{\sigma}, -\frac{\mu}{\sqrt{c(t)+\sigma^2}} \mid \frac{\sqrt{c(t)}}{\sqrt{c(t)+\sigma^2}}\right) - 1,$$

where $c(t) := 4(1-\rho^{2t})/\rho^{2t}$. We use this expression in all our experiments, evaluating the bivariate normal probabilities using the package `mvtnorm` (Genz et al., 2025).

Expected squared differences. We have the following formula for the conditional expected squared-difference $\mathbb{E}[(X_t^{(L)} - Y_t^{(L)})^2 \mid X_0^{(L)}, Y_0^{(L)}]$.

Lemma B.2.3. *It holds that*

$$\mathbb{E}[(X_t^{(L)} - Y_t^{(L)})^2 \mid X_0^{(L)}, Y_0^{(L)}] = 4(1-\rho^{2t}) \cdot \left\{(\alpha_t^2+1)(2\Phi(\alpha_t)-1) + 2\alpha_t\phi(\alpha_t)\right\},$$

where $\alpha_t = \frac{(X_0^{(L)} - Y_0^{(L)})\rho^t}{2\sqrt{1-\rho^{2t}}}$.

Proof. For simplicity, we drop the superscripts (L).

By Proposition B.2.1 and Connor (2007, Lemma 3.14), we have that $(X_t, Y_t) \mid (X_0, Y_0)$ is a reflection-maximal coupling of the marginals $(\mathcal{N}(\rho^t X_0, 1-\rho^{2t}), \mathcal{N}(\rho^t Y_0, 1-\rho^{2t}))$.

Now, let $\mathbb{E}_{\alpha,1}$ denote the expectation of under the reflection-maximal coupling of $(X, Y) \in \Gamma(\mathcal{N}(-\alpha, 1), \mathcal{N}(-\alpha, 1))$. By symmetry and a simple change of variables, we have that

$$\mathbb{E}[(X_t - Y_t)^2 \mid X_0, Y_0] = (1-\rho^{2t})\mathbb{E}_{\alpha,1}[(X-Y)^2],$$

where $\alpha = \frac{(X_0 - Y_0)\rho^t}{2\sqrt{1-\rho^{2t}}}$.

To evaluate the final expectation, we calculate

$$\int_0^\infty x^2\phi(x \mid -\alpha, 1)\mathrm{d}x = \int_0^\infty x^2\phi(x+\alpha)\mathrm{d}x = (\alpha^2+1)\Phi(-\alpha) - \alpha\phi(\alpha),$$

using e.g. Owen (1980, Eqn. 102). Finally, by symmetry (see Figure 2.1.2), we have that

$$\begin{aligned}
\mathbb{E}_{\alpha,1}[(X-Y)^2] &= 4\left\{\int_0^\infty x^2\phi(x \mid \alpha, 1)\mathrm{d}x - \int_0^\infty x^2\phi(x \mid -\alpha, 1)\mathrm{d}x\right\} \\
&= 4\left\{\int_{-\infty}^\infty x^2\phi(x \mid \alpha, 1)\mathrm{d}x - 2\int_0^\infty x^2\phi(x \mid -\alpha, 1)\mathrm{d}x\right\} \\
&= 4\left\{\mathbb{E}_{\alpha,1}[X^2] - 2\left[(\alpha^2+1)\Phi(-\alpha) - \alpha\phi(\alpha)\right]\right\} \\
&= 4\left\{(\alpha^2+1)(2\Phi(\alpha)-1) + 2\alpha\phi(\alpha)\right\},
\end{aligned}$$

where in the third line we substituted the previous calculation, and in the fourth that $\mathbb{E}_{\alpha,1}[X^2] = \alpha^2+1$. This concludes the proof. □

To evaluate the unconditional expected squared-difference $\mathbb{E}[(X_t^{(L)} - Y_t^{(L)})^2]$, we use the default quadrature routine in the `stats` package of the language `R` (R Core Team, 2025).

Expected inner products. Recall that $(X_t^{(L)}, Y_t^{(L)}) = (X_{t+L}, Y_t)$.

We evaluate $\mathbb{E}[X_tX_s] = \rho^{s-t}\mathbb{E}[X_t^2]$ for all $s \geq t$.

For $\mathbb{E}[X_tY_t]$, we must consider several cases. For $t \geq L$, we have that

$$\mathbb{E}[X_tY_t] = \rho^L\mathbb{E}[X_tY_{t-L}] = \frac{\rho^L}{2}\left\{\mathbb{E}[X_t^2] + \mathbb{E}[X_{t-L}^2] - \mathbb{E}\left[(X_{t-L}^{(L)} - Y_{t-L}^{(L)})^2\right]\right\}.$$

For $t < L$, we consider two sub-cases related to how $(X_0^{(L)}, Y_0^{(L)}) = (X_L, Y_0)$ are correlated:

- When (X_L, Y_0) are independent, we have that

$$\mathbb{E}[X_t Y_t] = \mathbb{E}[X_t]\mathbb{E}[Y_t] = \mathbb{E}[X_t]^2.$$

- When (X_L, Y_0) are perfectly correlated and Gaussian, then $Y_0 = aX_L + b$ with

$$a = \sqrt{\frac{\mathrm{Var}(Y_0)}{\mathrm{Var}(X_L)}} = \sqrt{\frac{\mathrm{Var}(X_0)}{\mathrm{Var}(X_L)}}, \quad b = \mathbb{E}[Y_0] - a\mathbb{E}[X_L] = \mathbb{E}[X_0] - a\mathbb{E}[X_L],$$

  and so

$$\mathbb{E}[X_t Y_t] = \rho^t \mathbb{E}[X_t Y_0] = a\rho^L \mathbb{E}[X_t^2] + b\rho^t \mathbb{E}[X_t].$$

Marginal moments. Finally, the marginal moments of the process are

$$\mathbb{E}[X_t] = \rho^t \mathbb{E}[X_0], \quad \mathbb{E}[X_t^2] = \rho^{2t}\mathbb{E}[X_0^2] + 1 - \rho^{2t}.$$

B.2.4 Asymptotics

Conditional survivor function

Using the Taylor expansion $2\Phi(x) - 1 = 2\phi(0)x + \Theta(x^3)$ about zero, we have that

$$\mathbb{P}(\tau^{(L)} > t \mid X_0^{(L)}, Y_0^{(L)}) = 2\Phi\left(\frac{|X_0^{(L)} - Y_0^{(L)}|}{2\sqrt{1-\rho^{2t}}}\rho^t\right) - 1 = \frac{1}{\sqrt{2\pi}}|X_0^{(L)} - Y_0^{(L)}|\rho^t + \Theta(\rho^{3t}).$$

Expected squared difference

Similarly to the conditional survivor function, by Lemma B.2.3 it holds that

$$\mathbb{E}[(X_t^{(L)} - Y_t^{(L)})^2 \mid X_0^{(L)}, Y_0^{(L)}] = \frac{8}{\sqrt{2\pi}}|X_0^{(L)} - Y_0^{(L)}|\rho^t + \Theta(\rho^{2t}).$$

Single-term estimator

We provide a generative description of the conditioned single-term estimator $H_t^{(L)} \mid \tau^{(L)} > t$, in the limit as $t \to \infty$.

Firstly, for $t \geq L$ we can write

$$H_t^{(L)} = h(X_t) + \sum_{i \geq 0} \{h(X_{t+iL+L}) - h(Y_{t+iL})\} = X_{t-L}^{(L)} + \sum_{i \geq 0} \{h(X_{t+iL}^{(L)}) - h(Y_{t+iL}^{(L)})\}.$$

Notice that the expression for $L > 1$ is analogous to the one for $L = 1$, up to setting $\rho \leftarrow \rho^L$ and changing the joint distribution of $(X_{t-L}^{(L)}, X_t^{(L)}, Y_t^{(L)})$. We thus focus on $L = 1$, and we provide the required changes for general L below.

Case $L = 1$. For simplicity, we drop all superscripts (L). The estimator of interest is

$$H_t = h(X_{t-1}) + \sum_{s \geq t} \{h(X_s) - h(Y_s)\}. \tag{B.2.1}$$

Since the joint process $(X_t, Y_t)_{t \geq 0}$ is Markov, we can simulate $H_t \mid \tau > t$ by proceeding as follows:

1. Sample $(X_0, Y_0) \mid \tau > t$.
2. Sample $(X_{t-1}, Y_{t-1}) \mid \big\{(X_0, Y_0),\, \tau > t\big\}$.
3. Sample $(X_t, Y_t) \mid \big\{(X_{t-1}, Y_{t-1}),\, \tau > t\big\}$.
4. Sample $(X_s - Y_s)_{s \geq t+1} \mid (X_t, Y_t)$ as usual.
5. Form H_t with the expression (B.2.1).

We now outline how each of these steps must be altered as $t \to \infty$, and how to sample from their limiting distributions.

Step 1. By Bayes' rule,

$$\mathbb{P}((X_0, Y_0) \in \mathrm{d}z \mid \tau > t) = \mathbb{P}((X_0, Y_0) \in \mathrm{d}z) \frac{\mathbb{P}(\tau > t \mid (X_0, Y_0) = z)}{\mathbb{P}(\tau > t)},$$

which tends to $\mathbb{P}((X_0, Y_0) \in \mathrm{d}z)$ as $t \to \infty$. In the limit, the initializations are thus independent of the meeting time.

Step 2. By Bayes' rule,

$$\begin{aligned}\mathbb{P}((X_{t-1}, Y_{t-1}) \in \mathrm{d}(x, y) \mid X_0, Y_0, \tau > t) \propto \mathbb{P}((X_{t-1}, Y_{t-1}) \in \mathrm{d}(x, y) \mid X_0, Y_0, \tau > t-1) \\ \times \mathbb{P}(\tau > t \mid (X_{t-1}, Y_{t-1}) = (x, y)).\end{aligned}$$

For the second term, by Theorem 4.4.1 and the Markov property, we have that

$$\mathbb{P}(\tau > t \mid (X_{t-1}, Y_{t-1}) = (x, y)) = 2\Phi\left(\frac{|x-y|}{2\sqrt{1-\rho^2}}\rho\right) - 1,$$

which does not depend on t.

For the first term, $(X_{t-1}, Y_{t-1}) \mid \big\{(X_0, Y_0),\, \tau > t-1\big\}$ is a draw from the reflection coupling of the residual densities of the transition kernels $\mathcal{N}(\rho^{t-1}X_0, 1-\rho^{2(t-1)})$ and $\mathcal{N}(\rho^{t-1}Y_0, 1-\rho^{2(t-1)})$; refer to Figure 2.1.2(b) for an illustration. Asymptotically as $t \to \infty$, we must therefore output $Y_{t-1} = -X_{t-1}$. Now, the conditional density of $X_{t-1} = x$ is

$$\begin{aligned}p_t(x) \propto \{\phi(x \mid \rho^{t-1}X_0, 1-\rho^{2(t-1)}) - \phi(x \mid \rho^{t-1}Y_0, 1-\rho^{2(t-1)})\} \\ \times \mathbb{1}\{(x - \rho^{t-1}\frac{X_0+Y_0}{2})\,\mathrm{sgn}(X_0 - Y_0) > 0\},\end{aligned}$$

which converges to $p_\infty(x) = x\exp(-x^2/2)\mathbb{1}\{x \cdot \mathrm{sgn}(X_0 - Y_0) > 0\}$ in the limit as $t \to \infty$. We recognize this as the density of a χ_2 random variable multiplied by $\mathrm{sgn}(X_0 - Y_0)$.

We can thus sample from $(X_{t-1}, Y_{t-1}) \mid \big\{(X_0, Y_0),\, \tau > t\big\}$ in the limit as $t \to \infty$ by

drawing

$$V \sim f, \quad X_{t-1} = \mathrm{sgn}(X_0 - Y_0) \cdot V, \quad Y_{t-1} = -X_{t-1},$$

where f has density

$$f(x) = \frac{x \exp(-x^2/2) \left\{ 2\Phi\left(x \cdot \frac{\rho}{\sqrt{1-\rho^2}} \right) - 1 \right\}}{\rho} \mathbb{1}\{x \geq 0\},$$

and where we used Owen (1980, Eqn. 1,011.2) to obtain the normalizing constant. Below, we propose two complementary rejection samplers for f, based on its $\rho \to 0$ and $\rho \to 1$ asymptotics.

Step 3. We need to sample $(X_t, Y_t) \mid \big\{(X_{t-1}, Y_{t-1}), \tau > t\big\}$ from the reflection coupling of the residual densities of the transition kernels $\mathcal{N}(\rho X_{t-1}, 1-\rho^2)$ and $\mathcal{N}(\rho Y_{t-1}, 1-\rho^2)$; refer to Figure 2.1.2(b) for an illustration.

Equivalently, this is the same as sampling $(\varepsilon, -\varepsilon)$ from the residual densities of $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(-\mu, 1)$, and setting

$$X_t = c + \sqrt{1-\rho^2}\varepsilon, \quad Y_t = c - \sqrt{1-\rho^2}\varepsilon,$$

where $\mu = \rho(X_{t-1} - Y_{t-1})/\{2\sqrt{1-\rho^2}\}$, and $c = \rho(X_{t-1} + Y_{t-1})/\{2\sqrt{1-\rho^2}\}$.

Now, considering $\mu > 0$ for simplicity, the density of ε is

$$q_\mu(x) = \frac{\phi(x \mid \mu, 1) - \phi(x \mid -\mu, 1)}{\Phi(\mu) - \Phi(-\mu)} \mathbb{1}\{x > 0\}.$$

We sample $\varepsilon \sim q_\mu$ by rejection: (i) when $\mu > 2$, we simply rejection-sample using $\mathcal{N}(\mu, 1)$ as the proposal; (ii) when $\mu \leq 2$, guided by the fact that $\lim_{\mu \to 0} q_\mu(x) = x \exp(-x^2/2) \mathbb{1}\{x > 0\}$ is the density of a random variable $W \sim \chi_2$, we rejection-sample using $\sigma \cdot W$ as a proposal, where we found $\sigma = 1.39$ and a conservative rejection constant by a grid search.

Steps 4 and 5. These steps are straightforward, and require no modification as $t \to \infty$.

General L. To adapt the above procedure to a general L, in step 1 we replace $(X_0, Y_0) \leftarrow (X_0^{(L)}, Y_0^{(L)}) \in \Gamma(\pi_L, \pi_0)$, and in steps 2-5 we replace $\rho \leftarrow \rho^L$.

The case of the test function $h(x) = x$. Since the test function $h(x) = x$ is odd, the distribution of $H_t^{(L)} \mid \tau^{(L)} > t$ as $t \to \infty$ is formed by two mixture components, one taking positive values and one taking negative values. Furthermore, the two mixture components are mirror images of each other about the origin, and the positive component has mixture weight $\mathbb{P}(X_0^{(L)} > Y_0^{(L)})$.

We can thus obtain the correct output by first drawing $H_t^{(L)} \mid \{H_t^{(L)} > 0, \tau^{(L)} > t\}$ as $t \to \infty$, then negating the result with probability $\mathbb{P}(X_0^{(L)} < Y_0^{(L)})$.

Rejection samplers for f. We propose two rejection samplers for the density f. Below, $U \sim \text{Unif}(0, 1)$ is used to accept the proposals.

- The first sampler is based on the $\rho \to 1$ asymptote of f. It uses the proposal $p_1(x) = x \exp(-x^2/2)$ of a χ_2 distribution, and accepts the proposal x if $U \leq f(x)/\{M_1 \cdot p_1(x)\}$, where

$$M_1 = \sup_x f(x)/p_1(x) = 1/\rho.$$

- The second sampler is based on the $\rho \to 0$ asymptote of f. It uses the proposal density $p_0(x) = \sqrt{2/\pi} x^2 \exp(-x^2/2)$ of a χ_3 distribution, and accepts the proposal x if $U \leq f(x)/\{M_0 \cdot p_0(x)\}$, where

$$M_0 = \sup_x f(x)/p_0(x) = \lim_{x \to 0} f(x)/p_0(x) = 1/\sqrt{1 - \rho^2},$$

and where the supremum is the limit because $f'(x) < p_0'(x)$ for all $x \geq 0$, and so $f(x)/p_0(x)$ is decreasing in $x \geq 0$.

To minimize the worst-case rejection rate, we use the first rejection sampler if $\rho \geq 1/(1 + \pi/2)$, and otherwise the second.

Appendix C

Appendix for Chapter 5

C.1 Analysis for Sections 5.2 and 5.3

It will be convenient to consider the Wasserstein distance of general order $p \geq 1$, defined through its p-th power as

$$\mathcal{W}_p^p(\mu,\nu) = \inf_{\pi\in\Gamma(\mu,\nu)} \int \|x-y\|^p \mathrm{d}\pi(x,y) = \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}\left[\|X-Y\|^p\right],$$

where $\Gamma(\mu,\nu)$ is the set of all joint distributions π with marginals (μ,ν). This has the Kantorovich dual

$$\mathcal{W}_p^p(\mu,\nu) = \sup_{(\phi,\psi)\in\Phi(\mu,\nu)} \int \phi(x)\mathrm{d}\mu(x) + \int \psi(y)\mathrm{d}\nu(y),$$

$$\Phi(\mu,\nu) = \{(\phi,\psi) \in L_1(\mu)\times L_1(\nu) \mid \phi(x)+\psi(y) \leq \|x-y\|^p, \ \ \forall x,y\},$$

with an optimal solution $(\phi_{\mu,\nu}, \psi_{\mu,\nu})$. We implicitly assume that $\mathbb{E}_\mu[\|X\|^p] < \infty$ and $\mathbb{E}_\nu[\|Y\|^p] < \infty$ whenever this distance is in use.

Recall that we have drawn independent samples $X_{1:n}, \bar{X}_{1:n} \overset{\text{iid}}{\sim} \mu$ and $Y_{1:n}, \bar{Y}_{1:n} \overset{\text{iid}}{\sim} \nu$ and

defined the empirical measures

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i},\ \ \bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{X}_i},\ \ \nu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i},\ \ \bar{\nu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{Y}_i}.$$

C.1.1 Bias of estimators

Plug-in estimator

Lemma C.1.1 shows that the plug-in estimator of $\mathcal{W}_p^p$ has a non-negative bias.

Lemma C.1.1. *It holds that* $\mathbb{E}\left[\mathcal{W}_p^p(\mu_n,\nu_n)\right] \geq \mathbb{E}\left[\mathcal{W}_p^p(\mu,\nu_n)\right] \geq \mathcal{W}_p^p(\mu,\nu)$.

Proof. We prove that $\mathbb{E}\left[\mathcal{W}_p^p(\mu_n,\nu_n)\right] \geq \mathcal{W}_p^p(\mu,\nu)$. Since $L_1(\nu) \subset L_1(\nu_n)$ it holds that $\Phi(\mu,\nu) \subset \Phi(\mu_n,\nu_n)$, therefore

$$\begin{aligned}\mathcal{W}_p^p(\mu_n,\nu_n) &= \sup_{(\phi,\psi)\in\Phi(\mu_n,\nu_n)} \int \phi \mathrm{d}\mu_n + \int \psi \mathrm{d}\nu_n \\ &\geq \sup_{(\phi,\psi)\in\Phi(\mu,\nu)} \int \phi \mathrm{d}\mu_n + \int \psi \mathrm{d}\nu_n \geq \int \phi_{\mu,\nu}\mathrm{d}\mu_n + \int \psi_{\mu,\nu}\mathrm{d}\nu_n.\end{aligned}$$

It follows that

$$\mathbb{E}[\mathcal{W}_p^p(\mu_n,\nu_n)] \geq \mathbb{E}\left[\int \phi_{\mu,\nu}\mathrm{d}\mu_n + \int \psi_{\mu,\nu}\mathrm{d}\nu_n\right] = \int \phi_{\mu,\nu}\mathrm{d}\mu + \int \psi_{\mu,\nu}\mathrm{d}\nu = \mathcal{W}_p^p(\mu,\nu),$$

as claimed. The other inequalities follow by similar arguments, using in turn that $\mathbb{E}_{\mu_n}[\int \phi_{\mu,\nu_n}\mathrm{d}\mu_n] = \int \phi_{\mu,\nu_n}\mathrm{d}\mu$ and that $\mathbb{E}_{\nu_n}[\int \phi_{\mu,\nu}\mathrm{d}\nu_n] = \int \phi_{\mu,\nu}\mathrm{d}\nu$. □

Lemma C.1.2 shows that the bias of the plug-in estimator of $\mathcal{W}_p^p$ decreases with the sample size.

Lemma C.1.2. *It holds that* $\mathbb{E}\left[\mathcal{W}_p^p(\mu_{n-1},\nu_{n-1})\right] \geq \mathbb{E}\left[\mathcal{W}_p^p(\mu_n,\nu_n)\right]$.

Proof. We define the leave-one-out empirical measures $\mu_{-i} = \frac{1}{n-1}\sum_{j\in[n]\setminus i}\delta_{X_j}$ and $\nu_{-i} =$

$\frac{1}{n-1}\sum_{j\in[n]\setminus i}\delta_{Y_j}$. Using Kantorovich duality,

$$\begin{aligned}\mathcal{W}_p^p(\mu_n,\nu_n) &= \int \phi_{\mu_n,\nu_n}\mathrm{d}\mu_n + \int \psi_{\mu_n,\nu_n}\mathrm{d}\nu_n \\ &= \int \phi_{\mu_n,\nu_n}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{d}\mu_{-i}\right) + \int \psi_{\mu_n,\nu_n}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{d}\nu_{-i}\right) \\ &= \frac{1}{n}\sum_{i=1}^n\left(\int \phi_{\mu_n,\nu_n}\mathrm{d}\mu_{-i} + \int \psi_{\mu_n,\nu_n}\mathrm{d}\nu_{-i}\right) \\ &\leq \frac{1}{n}\sum_{i=1}^n \sup_{(\phi,\psi)\in\Phi(\mu_{-i},\nu_{-i})}\int \phi\mathrm{d}\mu_{-i} + \int \psi\mathrm{d}\nu_{-i} = \frac{1}{n}\sum_{i=1}^n \mathcal{W}_p^p(\mu_{-i},\nu_{-i}),\end{aligned}$$

where finally we used that $(\phi_{\mu_n,\nu_n},\psi_{\mu_n,\nu_n})\in\Phi(\mu_n,\nu_n)\subset\Phi(\mu_{-i},\nu_{-i})$, then Kantorovich duality. The claimed result follows by taking expectations and using that $\mathbb{E}[\mathcal{W}_p^p(\mu_{-i},\nu_{-i})] = \mathbb{E}[\mathcal{W}_p^p(\mu_{n-1},\nu_{n-1})]$ for all i. □

Proof of Theorem 5.3.3(i)

The proof relies on a few standard results, which we recall without proof. For a convex $\varphi:\mathbb{R}^d\to\mathbb{R}$, we let $\varphi^*(x) = \sup_y\{x^\top y - \varphi(y)\}$ be its Legendre transform, which is convex and satisfies $\varphi^{**}=\varphi$. We say that φ is m-strongly convex for $m>0$ if and only if $f(x)=\varphi(x)-m\|x\|^2/2$ is convex.

Lemma C.1.3 (Duality between smoothness and strong convexity; Zhou, 2018)**.** *Let $\varphi:\mathbb{R}^d\to\mathbb{R}$ be convex and let $L>0$. Then, $\|\nabla\varphi\|_{\mathrm{Lip}}\leq L$ if and only if φ^* is $(1/L)$-strongly convex.*

Lemma C.1.4 (Brenier's theorem; McCann, 1995)**.** *Let $\mu,\nu\in\mathcal{P}(\mathbb{R}^d)$ satisfy Assumption **(A0)**. Then,*

$$\mathcal{W}_2^2(\mu,\nu) = \mathbb{E}_\mu[\|X - T_{\mu,\nu}(X)\|^2] = \mathbb{E}_\nu[\|T_{\nu,\mu}(Y) - Y\|^2],$$

where $T_{\mu,\nu}$ and $T_{\nu,\mu}$ are push-forward maps ($T_{\mu,\nu}\#\mu=\nu$, $T_{\nu,\mu}\#\nu=\mu$). Furthermore, the maps are uniquely determined by $T_{\mu,\nu}=\nabla\varphi_{\mu,\nu}$ and $T_{\nu,\mu}=\nabla\varphi_{\nu,\mu}$ where $\varphi_{\mu,\nu},\varphi_{\nu,\mu}$:

$\mathbb{R}^d \to \mathbb{R}$ *are convex and conjugate* ($\varphi_{\nu,\mu} = \varphi^*_{\mu,\nu}$).

Lemma C.1.5. $\mathcal{W}_2^2(\mu_n, \nu_n) = \min_\sigma \frac{1}{n}\sum_{i=1}^n \|X_i - Y_{\sigma(i)}\|^2$ *over all permutations* σ.

We proceed with the proof of Theorem 5.3.3(i). Since we have assumed that $\|T_{\nu,\mu}\|_{\mathrm{Lip}} = \|\nabla\varphi_{\nu,\mu}\|_{\mathrm{Lip}} \leq 1$, it follows that $\varphi^*_{\nu,\mu} = \varphi_{\mu,\nu}$ is 1-strongly convex. Therefore, $T_{\mu,\nu} = \mathrm{id} + \nabla f$, where $f(x) = \varphi_{\mu,\nu}(x) - \|x\|^2/2$ is convex. For all $(x, \bar{x})$, we therefore have that

$$\begin{aligned} \|T_{\mu,\nu}(x) - \bar{x}\|^2 &= \|T_{\mu,\nu}(x) - x\|^2 + 2\nabla f(x)^\top (x - \bar{x}) + \|x - \bar{x}\|^2 \\ &\geq \|T_{\mu,\nu}(x) - x\|^2 + 2\{f(x) - f(\bar{x})\} + \|x - \bar{x}\|^2, \end{aligned} \tag{C.1.1}$$

where finally we used the convexity of f.

Now, without loss of generality, we set $Y_i = T_{\mu,\nu}(X_i)$. By the primal formulation,

$$\begin{aligned} \mathbb{E}\left[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n)\right] &= \mathbb{E}\left[\min_\sigma \frac{1}{n}\sum_{i=1}^n \|\bar{X}_i - T_{\mu,\nu}(X_{\sigma(i)})\|^2\right] \\ &\geq \mathbb{E}\left[\min_\sigma \frac{1}{n}\sum_{i=1}^n \left(\|\bar{X}_i - X_{\sigma(i)}\|^2 + 2\left\{f(\bar{X}_i) - f(X_{\sigma(i)})\right\} + \|X_{\sigma(i)} - T_{\mu,\nu}(X_{\sigma(i)})\|^2\right)\right] \\ &= \mathbb{E}\left[\min_\sigma \frac{1}{n}\sum_{i=1}^n \|\bar{X}_i - X_{\sigma(i)}\|^2 + \frac{1}{n}\sum_{i=1}^n \|X_i - T_{\mu,\nu}(X_i)\|^2\right] \\ &= \mathbb{E}\left[\mathcal{W}_2^2(\bar{\mu}_n, \mu_n)\right] + \mathcal{W}_2^2(\mu, \nu), \end{aligned}$$

where we used (C.1.1) for the second line, that σ is a permutation and that $X_i, \bar{X}_i \sim \mu$ for the third, and the primal formulation for the last. This concludes the proof.

Proof of Theorem 5.3.3(ii)

Using the primal formulation, we have that

$$\begin{aligned} \mathbb{E}\left[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\bar{\mu}_n, \mu_n)\right] &= \mathbb{E}\left[\|Y\|^2 - \|X\|^2\right] - 2\mathbb{E}\left[\max_\sigma \frac{1}{n}\sum_{i=1}^n Y_i^\top \bar{X}_{\sigma(i)} - \max_\sigma \frac{1}{n}\sum_{i=1}^n X_i^\top \bar{X}_{\sigma(i)}\right] \\ &=: ① - ②, \end{aligned}$$

where $(X, Y) \sim (\mu, \nu)$ and where the maxima are over all permutations σ.

Term ①. By the Minkowski inequality, $\left|\mathbb{E}[\|Y\|^2]^{1/2} - \mathbb{E}[\|X\|^2]^{1/2}\right| \leq \inf_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}[\|Y - X\|^2]^{1/2} = \mathcal{W}_2(\mu, \nu)$. It follows that

$$\left|\mathbb{E}\left[\|Y\|^2 - \|X\|^2\right]\right| \leq \mathcal{W}_2(\mu, \nu) \left(\mathbb{E}\left[\|X\|^2\right]^{1/2} + \mathbb{E}\left[\|Y\|^2\right]^{1/2}\right).$$

Term ②. Without loss of generality, we choose to sample the pairs $(X_i, Y_i) \sim (\mu, \nu)$ i.i.d. from the optimal coupling. We have that

$$\begin{aligned}
\frac{1}{2}\left|②\right| &\leq \left|\mathbb{E}\left[\max_\sigma \frac{1}{n}\sum_{i=1}^n (Y_i - X_i)^\top \bar{X}_{\sigma(i)}\right]\right| && \text{(max is convex)} \\
&\leq \mathbb{E}\left[\max_\sigma \left(\frac{1}{n}\sum_{i=1}^n \|Y_i - X_i\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|\bar{X}_{\sigma(i)}\|^2\right)^{1/2}\right] && \text{(Cauchy-Schwarz)} \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \|Y_i - X_i\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|\bar{X}_i\|^2\right)^{1/2}\right] && \left(\textstyle\sum_i \|\bar{X}_{\sigma(i)}\|^2 = \sum_i \|\bar{X}_i\|^2\right) \\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|Y_i - X_i\|^2\right]^{1/2} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\bar{X}_i\|^2\right]^{1/2} && \text{(Cauchy-Schwarz)} \\
&= \mathcal{W}_2(\mu, \nu)\mathbb{E}\left[\|X\|^2\right]^{1/2}. && \text{(couplings } (X_i, Y_i) \text{ are optimal)}
\end{aligned}$$

Therefore,

$$\left|\mathbb{E}\left[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\bar{\mu}_n, \mu_n)\right]\right| \leq \left|①\right| + \left|②\right| \leq \mathcal{W}_2(\mu, \nu) \left(3\mathbb{E}\left[\|X\|^2\right]^{1/2} + \mathbb{E}\left[\|Y\|^2\right]^{1/2}\right),$$

which concludes the proof.

Proof of Theorem 5.3.3(iii)

Let $\{\mu^c, \nu^c\}$ be versions of $\{\mu, \nu\}$ with expectations 0, and let $\{\bar{\mu}_n^c, \mu_n^c, \nu_n^c\}$ be the analogous transformations of $\{\bar{\mu}_n, \mu_n, \nu_n\}$. From Panaretos and Zemel (2019, Section 2), it

holds that

$$\begin{aligned}\mathcal{W}_2^2(\mu,\nu) &= \|\mathbb{E}_\mu[X] - \mathbb{E}_\nu[Y]\|^2 + \mathcal{W}_2^2(\mu^c,\nu^c),\\ \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n,\nu_n)] &= \|\mathbb{E}_\mu[X] - \mathbb{E}_\nu[Y]\|^2 + \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n^c,\nu_n^c)],\\ \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n,\mu_n)] &= \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n^c,\mu_n^c)].\end{aligned}$$

It follows that

$$\mathbb{E}[U(\bar{\mu}_n,\mu_n,\nu_n)] - \mathcal{W}_2^2(\mu,\nu) = \mathbb{E}[U(\bar{\mu}_n^c,\mu_n^c,\nu_n^c)] - \mathcal{W}_2^2(\mu^c,\nu^c),$$

hence the difference is location-free, as claimed.

Proof of Proposition 5.3.1

By the Jensen and triangle inequalities,

$$\left|\mathbb{E}\left[\mathcal{W}_2(\bar{\mu}_n,\nu_n) - \mathcal{W}_2(\bar{\mu}_n,\mu_n)\right]\right| \le \mathbb{E}\left[\left|\mathcal{W}_2(\bar{\mu}_n,\nu_n) - \mathcal{W}_2(\bar{\mu}_n,\mu_n)\right|\right] \le \mathbb{E}\left[\mathcal{W}_2(\mu_n,\nu_n)\right]. \tag{C.1.2}$$

Now, using the linearity of the expectation, without loss of generality (without changing the left-hand-side of (C.1.2)) we choose to instead sample $(X_i, Y_i) \sim (\mu,\nu)$ i.i.d. from the optimal coupling. By Jensen's inequality and the primal formulation,

$$\mathbb{E}[\mathcal{W}_2(\mu_n,\nu_n)] \le \mathbb{E}\left[\mathcal{W}_2^2(\mu_n,\nu_n)\right]^{1/2} \le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|X_i - Y_i\|^2\right]^{1/2} = \mathcal{W}_2(\mu,\nu). \tag{C.1.3}$$

Combining inequalities (C.1.2) and (C.1.3) completes the proof.

C.1.2 Overdispersion conditions

Proof of Proposition 5.3.5

We first require the following characterization of $\overset{\text{COT}}{\rightsquigarrow}$ and we recall an auxiliary result.

Lemma C.1.6. *The following claims are equivalent:* (i) $\nu \overset{\text{COT}}{\rightsquigarrow} \mu$*;* (ii) $\|T_{\nu,\mu}\|_{\text{Lip}} \leq 1$*;* (iii) $\varphi_{\mu,\nu}$ *is 1-strongly convex;* (iv) $\nabla^2 \varphi_{\nu,\mu} \preceq I_d$ *uniformly;* (v) $\nabla^2 \varphi_{\mu,\nu} \succeq I_d$ *uniformly.*

Proof. Since the Brenier potentials $(\varphi_{\mu,\nu}, \varphi_{\nu,\mu})$ are convex, by Alexandroff's theorem their gradients and Hessians exist almost-everywhere.

The equivalence $(i) \iff (ii)$ follows by definition. The equivalence $(ii) \iff (iii)$ follows from the duality of smoothness and strong convexity. The equivalence $(ii) \iff (iv)$ is shown in Nesterov (2004, Theorem 2.1.6). The equivalence $(iii) \iff (v)$ is shown in Nesterov (2004, Theorem 2.1.11). Therefore, all claims are equivalent. □

Lemma C.1.7 (Lawson and Lim, 2001, Corollary 3.5)**.** *Let* $M, N \in \mathbb{R}^{d\times d}$ *be positive definite matrices. Define* $M^{-1}\#N := M^{-1/2}(M^{1/2}NM^{1/2})^{1/2}M^{-1/2}$*. Then, it holds that* $I \preceq M^{-1}\#N$ *if and only if* $M \preceq N$*.*

We proceed to the main proof. Since $\overset{\text{COT}}{\rightsquigarrow}$ is location-free, without loss of generality we let $\mathbb{E}_\mu[X] = \mathbb{E}_\nu[Y] = 0$.

Claim (i). By Peyré and Cuturi (2019, Remark 2.31), the Brenier potential from $\mu = \mathcal{N}(0, \Sigma_\mu)$ to $\nu = \mathcal{N}(0, \Sigma_\nu)$ is $\varphi_{\mu,\nu}(x) = x^\top(\Sigma_\mu^{-1}\#\Sigma_\nu)x/2$. By Lemma C.1.6, we have that

$$\nu \overset{\text{COT}}{\rightsquigarrow} \mu \iff I \preceq \nabla^2\varphi_{\mu,\nu} \text{ uniformly} \iff I \preceq \Sigma_\mu^{-1}\#\Sigma_\nu \iff \Sigma_\mu \preceq \Sigma_\nu,$$

where finally we used Lemma C.1.7. This concludes the proof of the claim.

Claim (ii). Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be spherically symmetric and let $\mathcal{S}^{d-1}$ be the unit sphere. Any $X \sim \mu$ can be written as $X = R_\mu U_\mu$ in terms of an angular component $U_\mu \sim$

$\text{Unif}(\mathcal{S}^{d-1})$ and an independent radial component $R_\mu \sim r_\mu \in \mathcal{P}((0,\infty))$. Similarly, so can $Y = R_\nu U_\nu \sim \nu$. Now, $\mathbb{E}[\|R_\mu U_\mu - R_\nu U_\nu\|^2] \geq \mathbb{E}[(R_\mu - R_\nu)^2]$. Since the lower bound is attained by the coupling

$$(X, Y) = \big(F^{-1}_{r_\mu}(U_1)U, F^{-1}_{r_\nu}(U_1)U\big) \sim (\mu, \nu),$$

where $U_1 \sim \text{Unif}([0,1])$ and $U \sim \text{Unif}(\mathcal{S}^{d-1})$, this coupling must be optimal. The optimal transport map is therefore

$$T_{\nu,\mu}(x) = \big(F^{-1}_{r_\mu} \circ F_{r_\nu}\big)(\|x\|) \cdot \frac{x}{\|x\|},$$

and so $\|T_{\nu,\mu}\|_{\text{Lip}} \leq 1$ if and only if $\|F^{-1}_{r_\mu} \circ F_{r_\nu}\|_{\text{Lip}} \leq 1$, as claimed.

Claim (iii). Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be product measures, say $\mu = \otimes_{i=1}^d \mu^i$ and $\nu = \otimes_{i=1}^d \nu^i$. By the tensorization property of the squared Wasserstein distance, the optimal transport map is

$$T_{\nu,\mu}(x) = \big(T_{\nu^1,\mu^1}(x_1), \ldots, T_{\nu^d,\mu^d}(x_d)\big)^\top,$$

where $T_{\nu^i,\mu^i} = F^{-1}_{\mu^i} \circ F_{\nu^i}$. Therefore, $\|T_{\nu,\mu}\|_{\text{Lip}} \leq 1$ if and only if $\|T_{\nu^i,\mu^i}\|_{\text{Lip}} \leq 1$ for all i, as claimed.

Claim (iv). This is lifted from Chewi and Pooladian (2023, Theorem 13).

On Example 5.3.6

Deriving the result. The inequality $\mathbb{E}\left[U(\bar{\mu}_1, \mu_1, \nu_1)\right] \geq \mathcal{W}_2^2(\mu, \nu)$ is equivalent to

$$\mathbb{E}[\|\bar{X} - Y\|^2 - \|\bar{X} - X\|^2] \geq \inf_{(X,Y)\sim(\mu,\nu)} \mathbb{E}[\|Y - X\|^2],$$

where $\bar{X} \sim \mu$ is independent of $(X, Y) \sim (\mu, \nu)$. Rearranging, this is equivalent to

$$\sup_{(X,Y)\sim(\mu,\nu)} 2\mathbb{E}\left[X^\top Y - \mathbb{E}[X]^\top \mathbb{E}[Y]\right] \geq 2\mathbb{E}\left[\|X\|^2 - \mathbb{E}[X]^\top \mathbb{E}[X]\right].$$

Recognizing the outer expectations as $\mathrm{Tr}(\mathrm{Cov}(X, Y))$ and $\mathrm{Tr}(\mathrm{Var}(X))$ provides the result of Example 5.3.6.

Partial closure under mixtures. Let $\nu = \sum_k p_k \nu^k$ be a mixture. By Jensen's inequality and the linearity of the expectation, it holds that

$$\sup_{(X,Y)\sim(\mu,\nu)} \mathrm{Tr}(\mathrm{Cov}(X, Y)) \geq \sum_k p_k \sup_{(X,Y_k)\sim(\mu,\nu^k)} \mathrm{Tr}(\mathrm{Cov}(X, Y_k)).$$

So, if $\sup \mathrm{Tr}(\mathrm{Cov}(X, Y_k)) \geq \mathrm{Tr}(\mathrm{Var}(X))$ for all k, then $\sup \mathrm{Tr}(\mathrm{Cov}(X, Y)) \geq \mathrm{Tr}(\mathrm{Var}(X))$. In other words, the relation of Example 5.3.6 is partially closed under mixtures.

Relation to convex ordering. The convex ordering $\nu \geq_{\mathrm{cvx}} \mu$ states that $\mathbb{E}_\nu[f(Y)] \geq \mathbb{E}_\mu[f(X)]$ for any convex f for which the expectations are well-defined. Strassen's martingale coupling theorem (Strassen, 1965) states that this is equivalent to the existence of coupling $(X, Y) \sim (\mu, \nu)$ such that $\mathbb{E}[Y \mid X] = X$.

Now, suppose that that a convex ordering holds between versions of μ and ν which are centered at 0, i.e. that there exists a coupling of $(X, Y) \sim (\mu, \nu)$ such that $\mathbb{E}[Y - \mathbb{E}[Y] \mid X] = X - \mathbb{E}[X]$. Under this coupling, $\mathrm{Tr}(\mathrm{Cov}(X, Y)) = \mathrm{Tr}(\mathrm{Cov}(X, X)) = \mathrm{Tr}(\mathrm{Var}(X))$, so the condition of Example 5.3.6 is satisfied.

On Example 5.3.7

The asymptotic result of Example 5.3.7 is a consequence of Proposition C.1.9. We require Lemma C.1.8, which provides a tractable formula for the bias of the plug-in estimator in the one-dimensional setting.

Lemma C.1.8. *Let (μ, ν) be one-dimensional measures with inverse-CDFs (G, H), and let $U_{(1:n)}$ be the order statistics of $U_{1:n} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. Then,*

$$\mathbb{E}[\mathcal{W}_2^2(\mu_n, \nu_n)] - \mathcal{W}_2^2(\mu, \nu) = \frac{2}{n} \sum_{i=1}^{n} \text{Cov}\left(G(U_{(i)}), H(U_{(i)})\right).$$

Proof. Since $\mathbb{E}[\mathcal{W}_2^2(\mu_n, \nu_n)] = \frac{1}{n}\mathbb{E}\big[\sum_{i=1}^n (X_{(i)} - Y_{(i)})^2\big]$ and since $X_{1:n}$ is independent of $Y_{1:n}$, it holds that

$$\begin{aligned}\mathbb{E}[\mathcal{W}_2^2(\mu_n, \nu_n)] &= \mathbb{E}\left[X_1^2 + Y_1^2\right] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_{(i)}Y_{(i)}\right] = \mathbb{E}\left[X_1^2 + Y_1^2\right] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_{(i)}\right]\mathbb{E}\left[Y_{(i)}\right] \\ &= \mathbb{E}\left[X_1^2 + Y_1^2\right] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\left[G(U_{(i)})\right]\mathbb{E}\left[H(U_{(i)})\right].\end{aligned}$$

Now, it holds that

$$\begin{aligned}\mathcal{W}_2^2(\mu, \nu) &= \mathbb{E}[G(U)^2 + H(U)^2] - 2\mathbb{E}[G(U)H(U)] = \mathbb{E}[X_1^2 + Y_1^2] - 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} G(U_i)H(U_i)\right] \\ &= \mathbb{E}[X_1^2 + Y_1^2] - 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} G(U_{(i)})H(U_{(i)})\right].\end{aligned}$$

The claimed result follows by subtracting off the previous identities. □

Proposition C.1.9. *Let (μ, ν) be one-dimensional measures with inverse-CDFs (G, H) that are twice differentiable with uniformly bounded second derivatives. Then,*

$$\mathbb{E}\left[\mathcal{W}_2^2(\mu_n, \nu_n) - \mathcal{W}_2^2(\mu, \nu)\right] = 2J(\mu, \nu)n^{-1} + o(n^{-1}),$$

where $J(\mu, \nu) = \int_0^1 u(1-u)G'(u)H'(u)\mathrm{d}u$.

Proof. Let $U_{(1):(n)}$ be the order statistics of $U_{1:n} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. By Lemma C.1.8, we have that

$$n\mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\mu, \nu)] = 2\sum_{i=1}^{n} \text{Cov}\left(G(U_{(i)}), H(U_{(i)})\right).$$

We will estimate $\mathrm{Cov}\left(G(U_{(i)}), H(U_{(i)})\right)$ using Taylor expansions. We require the first two moments of $U_{(i)} \sim \mathrm{Beta}(i, n+1-i)$,

$$a_{(i)} := \mathbb{E}[U_{(i)}] = \frac{i}{n+1} \text{ and } \sigma^2_{(i)} := \mathrm{Var}(U_{(i)}) = \frac{i(n+1-i)}{(n+1)^2(n+2)} = \frac{\frac{i}{n+1}\left(1-\frac{i}{n+1}\right)}{(n+1)} + O(n^{-2}). \tag{C.1.4}$$

Recall that $\sup_u |G''(u)| \le G''_{\max}$ and $\sup_u |H''(u)| \le H''_{\max}$ by assumption.

By the usual Taylor expansion,

$$G(U_{(i)}) = G(a_{(i)}) + (U_{(i)} - a_{(i)})G'(a_{(i)}) + r_G(U_{(i)}), \text{ where } |r_G(U_{(i)})| \le G''_{\max}(U_{(i)} - a_{(i)})^2.$$

Taking expectations on both sides, $\left|\mathbb{E}[G(U_{(i)})] - G(a_i)\right| \le G''_{\max}\,\mathrm{Var}(U_{(i)}) = G''_{\max}\sigma^2_{(i)}$. So, the triangle inequality gives

$$\left|G(U_{(i)}) - \mathbb{E}[G(U_{(i)})] - (U_{(i)} - a_i)G'(a_{(i)})\right| \le G''_{\max}\sigma^2_{(i)} + G''_{\max}(U_{(i)} - a_{(i)})^2\sigma^2_{(i)},$$

with a similar result for H. Combining these results with the elementary inequality $|g_1h_1 - g_2h_2| \le |g_1 - g_2||h_1 - h_2| + |g_2||h_1 - h_2| + |h_2||g_1 - g_2|$, we obtain that

$$\begin{aligned}&\left|\left\{G(U_{(i)}) - \mathbb{E}[G(U_{(i)})]\right\}\left\{H(U_{(i)}) - \mathbb{E}[H(U_{(i)})]\right\} - (U_{(i)} - a_i)^2 G'(a_{(i)})H'(a_{(i)})\right| \le \\ &\le G''_{\max}H''_{\max}\left(\sigma^2_{(i)} + (U_{(i)} - a_{(i)})^2\right)^2 + \left(G'(a_{(i)})G''_{\max} + H'(a_{(i)})H''_{\max}\right)|U_{(i)} - a_i|\left(\sigma^2_{(i)} + (U_{(i)} - a_{(i)})^2\right).\end{aligned}$$

The expectation of the right-hand side is $O(n^{-3/2})$. Therefore,

$$\mathrm{Cov}\left(G(U_{(i)}), H(U_{(i)})\right) = G'(a_{(i)})H'(a_{(i)})\,\mathrm{Var}(U_{(i)}) + O(n^{-3/2}).$$

Given the definition of $a_{(i)}$ and approximation of $\mathrm{Var}(U_{(i)})$ in equation (C.1.4), the result follows from the definition of the Riemann integral and the size of the remainder when it is approximated by a Riemann sum. □

Proposition C.1.9 requires lighter-than-Gaussian tails (Bobkov and Ledoux, 2019, Section 5.1) and generalizes Solomon et al. (2022, Proposition 5.5) and Bobkov and Ledoux (2019, Theorem 5.1).

C.1.3 Statistical properties

Proof of Theorem 5.3.8

Estimator U. The estimator $U(\bar{\mu}_n, \mu_n, \nu_n) = \mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\bar{\mu}_n, \mu_n)$ satisfies the bounded difference property under the compact space Assumption **(A1)**. Following Weed and Bach (2019); Chizat et al. (2020), since the space has diameter 1 by Assumption **(A1)**, changing any of the samples within ν_n or μ_n can only change U by at most $\pm n^{-1}$, and changing any one of the samples within $\bar{\mu}_n$ can only change U by at most $\pm 2n^{-1}$. By the bounded difference inequalities (McDiarmid, 1989), it follows that

$$\begin{aligned}
&\mathbb{P}\left(U(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] \geq t\right) \leq \exp\left(-2t^2 / \left\{2n(n^{-1})^2 + n(2n^{-1})^2\right\}\right) = \exp(-nt^2/3), \\
&\mathbb{P}\left(U(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[U(\bar{\mu}_n, \mu_n, \nu_n)\right] \leq -t\right) \leq \exp(-nt^2/3),
\end{aligned} \tag{C.1.5}$$

for any $t \geq 0$. A union bound concludes the proof.

Estimator $\bar{L}$. The proof for the estimator $\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) = \mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}_n, \mu_n)$ is more involved. Following Boissard and Le Gouic (2014, Appendix A) we use the transportation method, which provides concentration bounds for Lipschitz functionals. Technical details are postponed to Lemma C.1.11.

The key step is to establish that, when viewed as a function of its constituent samples, the estimator $\bar{L} : \mathbb{R}^{3nd} \to \mathbb{R}$ is Lipschitz. We show that $\|\bar{L}\|_{\text{Lip}} \leq 2n^{-1/2}$ in Lemma C.1.13. The compact support Assumption **(A1)** puts us in the setting of

Corollary C.1.12, hence

$$\begin{aligned}\mathbb{P}\left(\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[L(\bar{\mu}_n, \mu_n, \nu_n)\right] \geq t\right) &\leq \exp(-nt^4/32),\\ \mathbb{P}\left(\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) - \mathbb{E}\left[L(\bar{\mu}_n, \mu_n, \nu_n)\right] \leq -t\right) &\leq \exp(-nt^4/32),\end{aligned} \tag{C.1.6}$$

for any $t \geq 0$. A union bound concludes the proof.

Proof of Theorem 5.3.9

We first require Lemma C.1.10, which recalls the exact convergence rates of $\mathcal{W}_2(\bar{\mu}_n, \mu_n)$ and $\mathcal{W}_2^2(\bar{\mu}_n, \mu_n)$.

Lemma C.1.10. *Let $d \geq 5$ and consider Assumption **(A1)**. Then,*

$$\mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \mu_n)]^2 \asymp \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \mu_n)] \asymp n^{-2/d}.$$

Proof. By Jensen's inequality, we have that

$$\mathbb{E}[\mathcal{W}_1(\bar{\mu}_n, \mu_n)]^2 \leq \mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \mu_n)]^2 \leq \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \mu_n)].$$

Now, Chizat et al. (2020, Theorem 2) provides $\mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \mu_n)] \lesssim n^{-2/d}$. To see the lower asymptote, by Lemma C.1.1 and Panaretos and Zemel (2019, Section 3.3) it holds that $\mathbb{E}[\mathcal{W}_1(\bar{\mu}_n, \mu_n)] \geq \mathbb{E}[\mathcal{W}_1(\mu, \mu_n)] \gtrsim n^{-1/d}$. The claimed result follows. □

We turn to the proof of the main result.

Estimator U. By the triangle inequality,

$$\mathbb{E}\left[|U - \mathcal{W}_2^2(\mu, \nu)|\right] \leq \mathbb{E}\left[|\,\mathcal{W}_2^2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2^2(\mu, \nu)|\right] + \mathbb{E}\left[\mathcal{W}_2^2(\bar{\mu}_n, \mu_n)\right] \lesssim n^{-2/d},$$

where we finally used Chizat et al. (2020, Theorem 2) and Lemma C.1.10.

Estimator $\bar{L}$. By the triangle inequality,

$$\left|\mathbb{E}\left[\mathcal{W}_2(\mu,\nu)-\bar{L}\right]-\mathbb{E}[\mathcal{W}_2(\bar{\mu}_n,\mu_n)]\right| \leq \mathbb{E}\left[|\mathcal{W}_2(\bar{\mu}_n,\nu_n)-\mathcal{W}_2(\mu,\nu)|\right] \lesssim n^{-2/d},$$

where we finally used Chizat et al. (2020, Corollary 1). Since $\mathbb{E}[\mathcal{W}_2(\bar{\mu}_n,\mu_n)] \asymp n^{-1/d}$ by Lemma C.1.10 and since $\mathbb{E}[\mathcal{W}_2(\mu,\nu)-\bar{L}] \geq 0$ by Proposition 5.3.1, it follows that $\mathbb{E}[\mathcal{W}_2(\mu,\nu)-\bar{L}] \asymp \mathbb{E}[\mathcal{W}_2(\bar{\mu}_n,\mu_n)] \asymp n^{-1/d}$, as claimed.

Proof of Corollary 5.3.10

Estimator U. Using the lower deviation bound (C.1.5) from the proof of Theorem 5.3.8,

$$\mathbb{P}(U \geq \mathcal{W}_2^2(\mu,\nu)) \geq 1-\exp\left(-\frac{n}{3}\left(\mathbb{E}[U]-\mathcal{W}_2^2(\mu,\nu)\right)^2\right) \geq 1-\exp\left(C_1 n^{1-4/d}\right)$$

for some constant $C_1 > 0$, since we have assumed that $\mathbb{E}[U]-\mathcal{W}_2^2(\mu,\nu) \gtrsim n^{-2/d}$.

Estimator L. Using the upper deviation bound (C.1.6) from the proof of Theorem 5.3.8,

$$\mathbb{P}(L \leq \mathcal{W}_2^2(\mu,\nu)) = \mathbb{P}(\bar{L} \leq \mathcal{W}_2(\mu,\nu)) \geq 1-\exp\left(-\frac{n}{32}\left(\mathcal{W}_2(\mu,\nu)-\mathbb{E}[\bar{L}]\right)^4\right) \geq 1-\exp\left(C_2 n^{1-4/d}\right)$$

for some constant $C_2 > 0$, since $\mathcal{W}_2(\mu,\nu)-\mathbb{E}[\bar{L}] \gtrsim n^{-2/d}$ by Theorem 5.3.9.

Postponed auxiliary results

Lemma C.1.11 details the key ingredients of the transportation method of obtaining concentration inequalities. The idea is the following: if the fluctuations of μ, measured in some function of the Wasserstein distance, can be controlled by the Kullback-Leibler divergence $\mathrm{KL}(Q \mid \mu) = \int (\mathrm{d}Q/\mathrm{d}\mu)\log(\mathrm{d}Q/\mathrm{d}\mu)\mathrm{d}\mu$, then Lipschitz functions of $X \sim \mu$ concentrate. We refer to Boucheron et al. (2013, Chapter 8) for a pedagogical treatment.

Lemma C.1.11. *Let $\alpha, \beta : \mathbb{R} \to [0, \infty)$ be increasing with $\alpha(0) = \beta(0) = 0$. Let $\Omega \subseteq \mathbb{R}^d$ and let $\mathcal{P}(\Omega)$ be the set of all Ω-valued distributions. Let the ground metric be Euclidean throughout. For $\mu \in \mathcal{P}(\Omega)$ we define the following conditions*

$$\mathbf{T}_1(\alpha): \ \forall Q \in \mathcal{P}(\Omega) \text{ it holds that } \alpha(\mathcal{W}_1(Q, \mu)) \leq \mathrm{KL}(Q \mid \mu),$$
$$\mathbf{T}_2^2(\beta): \ \forall Q \in \mathcal{P}(\Omega) \text{ it holds that } \beta(\mathcal{W}_2^2(Q, \mu)) \leq \mathrm{KL}(Q \mid \mu).$$

The following claims hold:

(i) Suppose that $\mu \in \mathcal{P}(\Omega)$ satisfies condition $\mathbf{T}_1(\alpha)$. Then, for all $f : \Omega \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \leq L$ it holds that

$$\forall t \geq 0: \ \mathbb{P}_{X \sim \mu}\left(f(X) - \mathbb{E}\left[f(X)\right] \geq t\right) \leq \exp\left(-\alpha\left(t/L\right)\right).$$

(ii) Suppose that Ω has diameter at most D and let $\mu \in \mathcal{P}(\Omega)$. Then, μ satisfies condition $\mathbf{T}_2^2(\beta)$ with $\beta(t) = t^2/(2D^4)$.

(iii) Suppose that $\mu_1, \ldots, \mu_m \in \mathcal{P}(\Omega)$ all satisfy condition $\mathbf{T}_2^2(\beta)$. Then, $\mu = \otimes_{i=1}^m \mu_i \in \mathcal{P}(\Omega^m)$ satisfies condition $\mathbf{T}_2^2(\beta)$.

(iv) Suppose that $\mu \in \mathcal{P}(\Omega)$ satisfies condition $\mathbf{T}_2^2(\beta)$. Then, μ satisfies condition $\mathbf{T}_1(\alpha)$ with $\alpha(t) = \beta(t^2)$.

Proof. Claim (i) is equivalent to Gozlan and Léonard (2007, Lemma 5). Claim (ii) is a particular case of Bolley and Villani (2005, Particular case 2.5). Claim (iii) is a particular case of Gozlan and Léonard (2007, Theorem 5): as the squared Euclidean metric tensorizes, so must $\mathbf{T}_2^2(\beta)$. Claim (iv) uses the following argument: as $\beta \geq 0$ is an increasing function, by Jensen's inequality it holds that $\beta(\mathcal{W}_2^2(Q, \mu)) \geq \beta(\mathcal{W}_1^2(Q, \mu))$. Therefore, if μ satisfies $\mathbf{T}_2^2(\beta)$, we have that

$$\forall Q \in \mathcal{P}(\Omega): \ \beta(\{\mathcal{W}_1(Q, \mu)\}^2) \leq \mathrm{KL}(Q \mid \mu),$$

which is precisely condition $\mathbf{T}_1(\alpha)$ with $\alpha(t) = \beta(t^2)$. □

Corollary C.1.12 uses Lemma C.1.11 to derive a concentration bound for Lipschitz functions of compactly-supported product measures, and is used in the proof of Theorem 5.3.8.

Corollary C.1.12. *Let $\mu_1, \ldots, \mu_m \in \mathcal{P}(\Omega)$, where $\Omega \in \mathbb{R}^d$ has diameter at most 1. Then, $\mu = \otimes_{i=1}^m \mu_i$ satisfies inequality $\mathbf{T}_1(\alpha)$ with $\alpha = t^4/2$. Therefore, for all $t \geq 0$,*

$$\mathbb{P}_{X\sim\mu}\left(f(X) - \mathbb{E}\left[f(X)\right] \geq t\right) \leq \exp\left(-t^4/(2\|f\|_{\mathrm{Lip}}^4)\right),$$
$$\mathbb{P}_{X\sim\mu}\left(f(X) - \mathbb{E}\left[f(X)\right] \leq -t\right) \leq \exp\left(-t^4/(2\|f\|_{\mathrm{Lip}}^4)\right).$$

Proof. Lemma C.1.11(ii)-(iii) implies that $\mu = \otimes_{i=1}^m \mu_i$ satisfies $\mathbf{T}_2^2(\beta)$ with $\beta(t) = t^2/2$. Lemma C.1.11(iv) implies that μ also satisfies $\mathbf{T}_1(\alpha)$ with $\alpha(t) = \beta(t^2) = t^4/2$. Lemma C.1.11(i) concludes, noting that both f and $-f$ have Lipschitz constant $\|f\|_{\mathrm{Lip}}$. □

Lemma C.1.13 establishes that $\bar{L}$ is Lipschitz, and is used in the proof of Theorem 5.3.8.

Lemma C.1.13. *Viewing $\bar{L}(\bar{\mu}_n, \mu_n, \nu_n)$ as a function of its constituent samples, it holds that $\|\bar{L}\|_{\mathrm{Lip}} \leq 2n^{-1/2}$.*

Proof. Let $Z = [\bar{X}_{1:n}, X_{1:n}, Y_{1:n}] \in \mathbb{R}^{3nd}$ denote a concatenation. We define $Z' = [\bar{X}'_{1:n}, X'_{1:n}, Y'_{1:n}]$ and $\bar{\mu}'_n = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}'_i}$, $\mu'_n = \frac{1}{n}\sum_{i=1}^n \delta_{X'_i}$, $\nu'_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y'_i}$. We consider a minor abuse of notation and we equivalently define $\bar{L}(Z) = \bar{L}(\bar{\mu}_n, \mu_n, \nu_n) =$

$\mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}_n, \nu_n)$. The function $\bar{L}$ is Lipschitz because

$$\begin{aligned}
\left|\bar{L}\left(Z\right) - \bar{L}\left(Z'\right)\right| &= |\mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}'_n, \nu'_n) - \mathcal{W}_2(\bar{\mu}_n, \mu_n) + \mathcal{W}_2(\bar{\mu}'_n, \mu'_n)| \\
&\leq \mathcal{W}_2(\bar{\mu}_n, \bar{\mu}'_n) + \mathcal{W}_2(\nu_n, \nu'_n) + \mathcal{W}_2(\bar{\mu}_n, \bar{\mu}'_n) + \mathcal{W}_2(\mu_n, \mu'_n) \\
&\leq n^{-1/2}\left(\|\bar{X}_{1:n} - \bar{X}'_{1:n}\| + \|Y_{1:n} - Y'_{1:n}\| + \|\bar{X}_{1:n} - \bar{X}'_{1:n}\| + \|X_{1:n} - X'_{1:n}\|\right) \\
&\leq 2n^{-1/2}\left\|Z - Z'\right\|,
\end{aligned}$$

where we firstly used several applications of the triangle inequality, secondly the definition of the primal formulation (5.2.1), and finally the sharp inequality $x^{1/2} + y^{1/2} \leq \{2(x+y)\}^{1/2}$ twice. □

C.2 Uncertainty quantification

C.2.1 Jackknife variance estimation

The jackknife estimator of variance (Efron and Stein, 1981) for the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$, based on leave-one-out empirical measures of the form $\mu_{-i} = \frac{1}{n-1}\sum_{j\in[n]\setminus i}\delta_{X_j}$, reads

$$\mathrm{Var}(\mathcal{W}_2^2(\mu_n, \nu_n)) \approx \frac{n-1}{n}\sum_{i=1}^{n}\Big(\mathcal{W}_2^2(\mu_{-i}, \nu_{-i}) - \frac{1}{n}\sum_{j=1}^{n}\mathcal{W}_2^2(\mu_{-j}, \nu_{-j})\Big)^2.$$

Analogous jackknife estimators can be derived for e.g. $U(\bar{\mu}_n, \mu_n, \nu_n)$ using leave-one-out versions $U(\bar{\mu}_{-i}, \mu_{-i}, \nu_{-i})$.

Naively computing all $i \in [n]$ leave-one-out estimators would have complexity $O(n^4)$. Below, we present the Flapjack algorithm, which takes advantage of warm starts (Mills-Tettey et al., 2007) to reduce the complexity to $O(n^3)$. Understanding how this saving is obtained requires some background on linear assignment problem solvers, which we next recall.

Solving assignment problems

The primal and dual formulations of the linear assignment problem are

$$\min_{\sigma \in \mathbb{S}_n} \sum_{i=1}^{n} C_{i\sigma(i)} = \max_{u,v \in \mathbb{R}^n} \sum_{i=1}^{n} (u_i + v_i) \text{ subject to } \forall (i,j) : u_i + v_j \leq C_{ij}, \tag{C.2.1}$$

where $\mathbb{S}_n$ is the set of permutations of $[n]$, and $C \in \mathbb{R}^{n \times n}$ is a cost matrix.

Primal-dual assignment problem solvers (e.g. Kuhn, 1955; Munkres, 1957; Jonker and Volgenant, 1987) have the following general structure. We initialize with a set of feasible duals (u, v) and an empty partial assignment σ, where we write $\sigma(i) = *$ if a row i has not been assigned to any column j. Each iteration, we apply a procedure $\texttt{stage}(C, u, v, \sigma)$ that returns a new triple (u, v, σ) and: (i) increases the number of columns in the assignment by one; (ii) maintains feasibility across all duals, i.e. $\forall (i,j) : u_i + v_j \leq C_{ij}$; (iii) ensures that there is no dual slack across the matched pairs, i.e. $\forall i : u_i + v_{\sigma(i)} = C_{i\sigma(i)}$ if $\sigma(i) \neq *$. The complementary slackness conditions ensure that we terminate correctly after n iterations of `stage`.

Efficient implementations (e.g. Jonker and Volgenant, 1987) of `stage` have worst-case complexities $O(n^2)$, so the worst-case complexity of assignment problem solvers is $O(n^3)$.

Solving leave-one-out assignment problems

Suppose that we wish to solve the "leave-one-out" assignment problem, where row i and column i of the cost matrix C are removed. A naive solution would require solving this modified assignment problem from scratch, and thus $O(n^3)$ operations. However, by starting from the solution (u, v, σ) to the full-data assignment problem (C.2.1), it turns out that we can reduce this complexity by an order of magnitude.

Algorithm 7 uses the method of Mills-Tettey et al. (2007) to solve for the leave-one-out assignment cost. It solves an equivalent problem: $C_{ii} = \varepsilon$ is set small enough

Algorithm 7 Leave-one-out assignment cost, row i and column i of cost matrix removed

Input: Cost matrix C, optimal solution (u, v, σ) to primal-dual pair (C.2.1).

1. Remove row i from assignment: $\sigma(i) = *$.
2. Set small cost $C_{ii} = \varepsilon$ to guarantee assignment of pair (i, i), e.g. $\varepsilon < \min_{ij} C_{ij} - 2 \max_{ij} C_{ij}$.
3. Restore feasibility: if $u_i + v_i > C_{ii}$ set $u_i = C_{ii} - v_i$.
4. Solve for assignment: $(u, v, \sigma) \leftarrow \texttt{stage}(C, u, v, \sigma)$.
5. Return $\sum_{j=1, j\neq i}^{n} C_{j\sigma(j)}$ and reset C_{ii}.

so that row i is guaranteed to be assigned to column i; C_{ii} is then discarded in the final calculation. By removing row i from the assignment (line 1) and then restoring feasibility (line 3), we still obey complementary slackness with $(n-1)$ assigned rows, so one iteration of `stage` (line 4) suffices to obtain the correct solution.

Efficient implementations of Algorithm 7 have $O(n^2)$ complexities, a significant saving compared to the $O(n^3)$ cost of solving the leave-one-out problem without a warm start.

Flapjack algorithm

The procedure we call "Flapjack" starts from an optimal solution to the assignment problem (C.2.1), then applies Algorithm 7 for $i \in [n]$ to return all leave-one-out assignment costs.

Our implementation of Flapjack uses `stage` from Jonker and Volgenant (1987), so has a worst-case complexity of $O(n^3)$. We also observe this scaling in practice (see Figure C.6.1), which relates to a tendency of the algorithm of Jonker and Volgenant (1987) to perform many scans when most of the partial assignment σ has been filled (see also Guthe and Thuerck, 2021, Section 3.1).

Flapjack can be used to compute the jackknife estimate of variance for the plug-in estimator $\mathcal{W}_2^2(\mu_n, \nu_n)$ by fixing $C_{ij} = \|X_i - Y_j\|^2$, in which case the full-data assignment cost is $n\,\mathcal{W}_2^2(\mu_n, \nu_n)$ and the i-th leave-one-out assignment cost is $(n-1)\,\mathcal{W}_2^2(\mu_{-i}, \nu_{-i})$.

C.2.2 Approximate delta method for $\bar{L}$

We detail our approximate delta method for $\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) = \mathcal{W}_2(\bar{\mu}_n, \nu_n) - \mathcal{W}_2(\bar{\mu}_n, \nu_n)$.

Let $\Delta(\alpha, \beta) := \mathcal{W}_2^2(\alpha, \beta) - \mathbb{E}[(\alpha, \beta)]$. Taylor's theorem and a further approximation provide

$$\mathcal{W}_2(\bar{\mu}_n, \nu_n) \approx \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n)]^{1/2} + \frac{\Delta(\bar{\mu}_n, \nu_n)}{2\mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n)]^{1/2}} \approx \mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \nu_n)] + \frac{\Delta(\bar{\mu}_n, \nu_n)}{2\mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \nu_n)]},$$

the former of which is accurate when $\mathrm{Var}(\mathcal{W}_2^2(\bar{\mu}_n, \nu_n)) \ll \mathbb{E}[\mathcal{W}_2^2(\bar{\mu}_n, \nu_n)]$, whereas the latter when $\mathrm{Var}(\mathcal{W}_2(\bar{\mu}_n, \nu_n)) \ll \mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \nu_n)]^2$. Both conditions hold as $n \to \infty$. This suggests the approximation

$$\bar{L} - \mathbb{E}[\bar{L}] \approx \frac{\Delta(\bar{\mu}_n, \nu_n)}{2\mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \nu_n)]} - \frac{\Delta(\bar{\mu}_n, \mu_n)}{2\mathbb{E}[\mathcal{W}_2(\bar{\mu}_n, \mu_n)]}. \tag{C.2.2}$$

Using that $\Delta(\bar{\mu}_n, \nu_n) = \frac{1}{n}\sum_{i\in[n]}[\phi_{\bar{\mu}_n,\nu_n}(\bar{X}_i) + \psi_{\bar{\mu}_n,\nu_n}(Y_i)] + \text{const}$, we derive the variance estimate

$$\mathrm{Var}(\bar{L}) \approx \frac{1}{n}\,\mathrm{Var}\left(\left\{\frac{\phi_{\bar{\mu}_n,\nu_n}(\bar{X}_i) + \psi_{\bar{\mu}_n,\nu_n}(Y_i)}{2\,\mathcal{W}_2(\bar{\mu}_n, \nu_n)} - \frac{\phi_{\bar{\mu}_n,\mu_n}(\bar{X}_i) + \psi_{\bar{\mu}_n,\mu_n}(X_i)}{2\,\mathcal{W}_2(\bar{\mu}_n, \mu_n)}\right\}_{i=1}^{n}\right),$$

based on (C.2.2) and the insight that the empirical Kantorovich potentials are asymptotically i.i.d. (implicit in the results of del Barrio et al., 2024). Although this is only a heuristic, experiments in a setting similar to Figure 5.3.3 reveal that the variance estimate is more accurate than the jackknife, while being slightly conservative.

C.2.3 Estimators that use independent blocks of correlated samples

We describe how to quantify uncertainty in the setting of Section 5.4.1.

For simplicity, let $B_\mu = B_\nu = 0$ and $T_\mu = T_\nu = 1$. Define the sum of the Kantorovich

potentials $f_{\mu,\nu}(x,y) := \phi_{\mu,\nu}(x) + \psi_{\mu,\nu}(y)$. The proposed estimators are

$$U(\bar{\mu}_n, \mu_n, \nu_n) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{I}\sum_{i=0}^{I-1}\Big[f_{\bar{\mu}_n,\nu_n}(X_{k+K}^{(i)}, Y_k^{(i)}) - f_{\bar{\mu}_n,\mu_n}(X_{k+K}^{(i)}, X_k^{(i)})\Big],$$
$$\bar{L}(\bar{\mu}_n, \mu_n, \nu_n) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{I}\sum_{i=0}^{I-1}\left[\frac{f_{\bar{\mu}_n,\nu_n}(X_{k+K}^{(i)}, Y_k^{(i)})}{\mathcal{W}_2(\bar{\mu}_n,\nu_n)} - \frac{f_{\bar{\mu}_n,\mu_n}(X_{k+K}^{(i)}, X_k^{(i)})}{\mathcal{W}_2(\bar{\mu}_n,\mu_n)}\right].$$

To quantify the uncertainty in $\{U, \bar{L}\}$, we use Gaussian confidence intervals, based on the empirical variances

$$\mathrm{Var}(U) \approx \frac{1}{K}\mathrm{Var}\left(\left\{\frac{1}{I}\sum_{i=0}^{I-1}\Big[f_{\bar{\mu}_n,\nu_n}(X_{k+K}^{(i)}, Y_k^{(i)}) - f_{\bar{\mu}_n,\mu_n}(X_{k+K}^{(i)}, X_k^{(i)})\Big]\right\}_{k=1}^{K}\right),$$
$$\mathrm{Var}(\bar{L}) \approx \frac{1}{K}\mathrm{Var}\left(\left\{\frac{1}{I}\sum_{i=0}^{I-1}\left[\frac{f_{\bar{\mu}_n,\nu_n}(X_{k+K}^{(i)}, Y_k^{(i)})}{\mathcal{W}_2(\bar{\mu}_n,\nu_n)} - \frac{f_{\bar{\mu}_n,\mu_n}(X_{k+K}^{(i)}, X_k^{(i)})}{\mathcal{W}_2(\bar{\mu}_n,\mu_n)}\right]\right\}_{k=1}^{K}\right),$$

with consistency as $K \to \infty$. These can be justified using an extension of del Barrio et al. (2024, Theorem 4.10) and the approximate delta method of Appendix C.2.2.

Quantifying the variance reduction due the coupling. When instead (μ_n, ν_n) are correlated, we can use the estimator

$$\mathrm{Var}(U_{\mathrm{indep}}) \approx \frac{1}{K}\mathrm{Var}\left(\left\{\frac{1}{I}\sum_{i=0}^{I-1}\left[\phi_{\bar{\mu}_n,\nu_n}(X_{k+K}^{(i)}) - \phi_{\bar{\mu}_n,\mu_n}(X_{k+K}^{(i)})\right]\right\}_{k=1}^{K}\right)$$
$$+ \frac{1}{K}\mathrm{Var}\left(\left\{\frac{1}{I}\sum_{i=0}^{I-1}\psi_{\bar{\mu}_n,\nu_n}(Y_k^{(i)})\right\}_{k=1}^{K}\right) + \frac{1}{K}\mathrm{Var}\left(\left\{\frac{1}{I}\sum_{i=0}^{I-1}\psi_{\bar{\mu}_n,\mu_n}(X_k^{(i)})\right\}_{k=1}^{K}\right)$$

to *estimate the variance of* U *as if* (μ_n, ν_n) *were independent*, without actually requiring us to draw independent versions of these empirical measures. A similar estimator can be considered for $\bar{L}$.

When $\mathrm{Var}(U_{\mathrm{indep}}) \geq \mathrm{Var}(U)$, since $\mathrm{Var}(U_{\mathrm{indep}})$ and $\mathrm{Var}(U)$ are noisy overestimates of the actual variances, we expect to obtain a noisy underestimate of the factor of

variance reduction $\mathrm{Var}(U_{\mathrm{indep}})/\,\mathrm{Var}(U)$.

C.2.4 Time-averaged estimators

We describe how to quantify uncertainty in the setting of Appendix C.5.1.

Let $\pi_n^{(t)} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^{(t)}}$, based on replicates $(X_i^{(t)})_{t\geq 0}$ of a stochastic process for $i \in [n]$. Define the sum of the Kantorovich potentials $f_{\mu,\nu}(x,y) := \phi_{\mu,\nu}(x) + \psi_{\mu,\nu}(y)$. The estimators of Appendix C.5.1 are

$$U_{T,t} = \frac{1}{n}\sum_{i=1}^n \Big[f_{\pi_n^{(T)},\pi_n^{(t)}}(X_i^{(T)}, X_i^{(t)}) - \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}} f_{\pi_n^{(T)},\pi_n^{(S)}}(X_i^{(T)}, X_i^{(S)}) \Big],$$
$$\bar{L}_{T,t} = \frac{1}{n}\sum_{i=1}^n \left[\frac{f_{\pi_n^{(T)},\pi_n^{(t)}}(X_i^{(T)}, X_i^{(t)})}{\mathcal{W}_2(\pi_n^{(T)}, \pi_n^{(t)})} - \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}} \frac{f_{\pi_n^{(T)},\pi_n^{(S)}}(X_i^{(T)}, X_i^{(S)})}{\mathcal{W}_2(\pi_n^{(T)}, \pi_n^{(S)})} \right].$$

To quantify the uncertainty in $\{U_{T,t}, \bar{L}_{T,t}\}$, we use Gaussian confidence intervals, based on the empirical variances

$$\mathrm{Var}(U_{T,t}) \approx \frac{1}{n}\,\mathrm{Var}\left(\left\{ f_{\pi_n^{(T)},\pi_n^{(t)}}(X_i^{(T)}, X_i^{(t)}) - \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}} f_{\pi_n^{(T)},\pi_n^{(S)}}(X_i^{(T)}, X_i^{(S)}) \right\}_{i=1}^n \right),$$
$$\mathrm{Var}(\bar{L}_{T,t}) \approx \frac{1}{n}\,\mathrm{Var}\left(\left\{ \frac{f_{\pi_n^{(T)},\pi_n^{(t)}}(X_i^{(T)}, X_i^{(t)})}{2\,\mathcal{W}_2(\pi_n^{(T)}, \pi_n^{(t)})} - \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}} \frac{f_{\pi_n^{(T)},\pi_n^{(S)}}(X_i^{(T)}, X_i^{(S)})}{2\,\mathcal{W}_2(\pi_n^{(T)}, \pi_n^{(S)})} \right\}_{i=1}^n \right),$$

where consistency is as $n \to \infty$. These can be justified using extensions of del Barrio et al. (2024, Theorem 4.10) and the approximate delta method of Appendix C.2.2. For $L_{T,t} = \left[\bar{L}_{T,t}\right]_{\pm}^2$, we scale up the confidence interval for $\bar{L}_{T,t}$ accordingly.

C.3 Description of MCMC Algorithms

We describe the MCMC algorithms that are used in the analysis of Appendix C.4.

C.3.1 ULA

The unadjusted Langevin algorithm (ULA) targeting π generates a Markov chain $(X^{(t)})_{t\geq 0}$ based on the recursion

$$X^{(t+1)} = X^{(t)} + \frac{h^2}{2} A \nabla \log \pi(X^{(t)}) + \varepsilon^{(t)}, \;\; \varepsilon^{(t)} \sim \mathcal{N}_d(0_d, h^2 A),$$

where the user sets the step size $h > 0$ and the preconditioner $A \succ 0$.

C.3.2 OBABO

The OBABO discretization of the underdamped Langevin diffusion, targeting π in the X-component, generates a Markov chain $(X^{(t)}, Z^{(t)})_{t\geq 0}$ based on the recursion[1]

$$\begin{aligned}
&\textbf{O:} \;\; Z_\eta^{(t)} = \eta Z^{(t)} + \sqrt{1-\eta^2}\varepsilon^{(t)}, \;\; \varepsilon^{(t)} \sim \mathcal{N}_d(0_d, A), \\
&\textbf{B:} \;\; Z^{(t+1/2)} = Z_\eta^{(t)} + \frac{h}{2} A \nabla \log \pi(X^{(t)}), \\
&\textbf{A:} \;\; X^{(t+1)} = X^{(t)} + h Z^{(t+1/2)}, \\
&\textbf{B:} \;\; Z^{(t+1)} = Z^{(t+1/2)} + \frac{h}{2} A \nabla \log \pi(X^{(t+1)}),
\end{aligned}$$

where the user sets the step size $h > 0$, the preconditioner $A \succ 0$, and the momentum persistence parameter $\eta \in [0, 1)$. When $\eta = 0$, the process $(X^{(t)})_{t\geq 0}$ is an ULA chain.

[1] For simplicity, we collapsed the two partial O-steps into a full step.

C.3.3 Gibbs sampler for half-t regression

We consider a linear regression model with half-t(ν) priors,

$$\begin{aligned} y \mid X, \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I_n) \\ \beta_j | \eta, \xi, \sigma^2 \sim \mathcal{N}\Big(0, \frac{\sigma^2}{\xi\eta_j}\Big), \;\; \eta_j^{-1/2} &\sim t_+(\nu), \;\; \text{independently for } j \in [d], \\ \xi^{-1/2} \sim \mathcal{C}_+(0,1), \;\; \sigma^{-2} &\sim \text{Gamma}\Big(\frac{a_0}{2}, \frac{b_0}{2}\Big), \end{aligned} \tag{C.3.1}$$

where $\mathcal{C}_+(0,1)$ is the half-Cauchy distribution with density $\pi_\xi(x) \propto 1/(1+x^2)$ and $t_+(\nu)$ is the half-t distribution with ν degrees of freedom.

Algorithm 8 describes the approximate Gibbs sampler[2] of Biswas and Mackey (2024, Section 4.2) targeting the posterior distribution $\pi(\eta, \xi, \sigma^2, \beta \mid X, y)$ of the regression model (C.3.1), where

$$\begin{aligned} M(\xi, \eta, X) &= I_n + \xi^{-1} X \operatorname{diag}(\eta^{-1}) X^\top, \\ \log L(y, M) &= -\frac{1}{2}\log\det(M) - \frac{a_0 + n}{2}\log(b_0 + y^\top M^{-1} y). \end{aligned}$$

The exact algorithm of Biswas et al. (2022) corresponds to setting $\varepsilon = 0$ in Algorithm 8.

C.4 Analysis for Sections 5.4 and 5.5

C.4.1 Proof of Proposition 5.4.1

Proposition 5.4.1 is an immediate consequence of the following result.

Proposition C.4.1. *Let $\pi = \mathcal{N}_d(\mu, \Sigma)$ and let the spectral radius $\rho\big(\frac{h^2}{4} A^{1/2}\Sigma^{-1}A^{1/2}\big) < 1$. The following claims hold:*

(i) The invariant distribution of the OBABO chain of Appendix C.3.2 is $\pi^{(\infty)} \otimes \mathcal{N}_d(0_d, A)$, where $\pi^{(\infty)} = \mathcal{N}_d\big(\mu, (I_d - \frac{h^2}{4}A^{1/2}\Sigma^{-1}A^{1/2})^{-1}\Sigma\big)$.

[2]The selection of the active set $\mathbb{I}_\varepsilon$ mirrors the implementations of Johndrow et al. and Biswas and Mackey.

Algorithm 8 Approximate Gibbs sampler for regression model (C.3.1)

Input: current state $(\eta, \xi, \sigma^2, \beta)$, approximation parameter $\varepsilon \geq 0$, step size σ_{MH}.

1. Sample $\eta \mid \xi, \sigma^2, \beta$ component-wise. For each component j, target

$$\pi(\eta_j | \dots) \propto \eta_j^{\frac{\nu-1}{2}} (1 + \nu\eta_j)^{-\frac{\nu+1}{2}} \exp(-m_j\eta_j),$$

   with $m_j = \xi\beta_j^2/(2\sigma^2)$, using the slice sampler of Biswas et al. (2022, Algorithm 4).

2. Sample $\xi, \sigma^2, \beta \mid \eta$ as follows:

   (a) Sample $\xi \mid \eta$ with approximate Metropolis-Hastings.
   Propose $\log \xi^* \sim \mathcal{N}_1(\log \xi, \sigma^2_{\mathrm{MH}})$ and fix $\mathbb{I}_\varepsilon = \operatorname{diag}\big(\mathbb{1}\{\min(\xi^*, \xi)^{-1}\eta^{-1} > \varepsilon\}\big)$.
   Calculate acceptance probability

$$q = \frac{L(y, M(\xi^*, \eta, X\mathbb{I}_\varepsilon))}{L(y, M(\xi, \eta, X\mathbb{I}_\varepsilon))} \frac{\pi_\xi(\xi^*)}{\pi_\xi(\xi)} \frac{\xi^*}{\xi}.$$

   With probability q set $\xi = \xi^*$.

   (b) Sample

$$\sigma^2 \mid \eta, \xi \sim \text{InvGamma}\Big(\frac{a_0 + n}{2}, \frac{y^\top M(\xi, \eta, X\mathbb{I}_\varepsilon)^{-1} y + b_0}{2}\Big).$$

   (c) Sample

$$\beta \mid \eta, \xi, \sigma^2 \sim \mathcal{N}\big(\Sigma_\varepsilon^{-1}(X\mathbb{I}_\varepsilon)^\top y, \sigma^2\Sigma_\varepsilon^{-1}\big)$$

   with $\Sigma_\varepsilon = (X\mathbb{I}_\varepsilon)^\top(X\mathbb{I}_\varepsilon) + \xi \operatorname{diag}(\eta)$, using the algorithm of Bhattacharya et al. (2016).

3. Return $(\eta, \xi, \sigma^2, \beta)$.

(ii) The invariant distribution of the ULA chain of Appendix C.3.1 is $\pi^{(\infty)}$.

(iii) $\pi^{(\infty)} \overset{\text{COT}}{\rightsquigarrow} \pi$.

Proof. We first consider the case $A = I_d$ and $\mu = 0$.

For claim (i), the steps BAB form a velocity Verlet integrator of Hamiltonian dynamics. By e.g. Apers et al. (2024, Section 2.3.1), these dynamics are an exact time-discretization of Hamiltonian dynamics that leave the Hamiltonian $H(x, z) = \frac{1}{2}x^\top\Sigma^{-1}\big(I - \frac{h^2}{4}\Sigma^{-1}\big)x + \frac{1}{2}\|z\|^2$ invariant. The O step leaves the marginal distribution $\mathcal{N}_d(0_d, I_d)$ invariant. It follows that the invariant distribution of the OBABO chain is

$\mathcal{N}_d\big(\mu, (I_d - \frac{h^2}{4}\Sigma^{-1})^{-1}\Sigma\big) \otimes \mathcal{N}_d(0_d, I_d)$.

For claim (ii), we use that ULA is a particular case of OBABO with $\eta = 0$.

For claim (iii), we use that $(I_d - \frac{h^2}{4}\Sigma^{-1})^{-1}\Sigma \succeq \Sigma$, then apply Proposition 5.3.5(i).

Finally, to deal with the case of general (A, μ), we use that the process $(\bar{X}^{(t)}, \bar{Z}^{(t)})_{t\geq 0} = (A^{-1/2}X^{(t)} - \mu, A^{-1/2}Z^{(t)})_{t\geq 0}$ is an OBABO chain with preconditioner $\bar{A} = I_d$. Transforming back to the original process provides the claimed results. □

C.4.2 Overdispersion of approximate Gibbs sampler for half-t regression

Algorithm 8 explicitly zeroes the columns of the design matrix X with *a posteriori* weakest signal (via $X\mathbb{I}_\varepsilon$ in step 2). Compared to the exact algorithm of Biswas et al. (2022) (Algorithm 8 with $(\varepsilon, X\mathbb{I}_\varepsilon) = (0, X)$), this allows for faster computation in high-dimensional settings, and causes Algorithm 8 to sample from an overdispersed version of the exact posterior distribution of the regression coefficients β, as we now explain.

Inspecting how step 2 in Algorithm 8 changes as the level of approximation $\varepsilon \geq 0$ increases, we see that $\mathbb{I}_\varepsilon$ becomes sparser, so the sequences $(M(\xi, \eta, X\mathbb{I}_\varepsilon)^{-1})_{\varepsilon\geq 0}$ and $(\Sigma_\varepsilon^{-1})_{\varepsilon\geq 0}$ increase in the Loewner order. Therefore, the update $\sigma^2 \mid \eta, \xi$ increases in the usual stochastic order and the update $\beta \mid \eta, \xi, \sigma^2$ becomes more dispersed and the active components $\mathbb{I}_\varepsilon\beta$ become more outwardly shifted.[3] This indicates that stationary distribution of β spreads out as ε increases.

C.4.3 Proof of Proposition 5.5.1

Proposition 5.5.1 is an immediate consequence of the following result and Proposition 5.3.5(i).

[3]For the inactive components $(I_d - \mathbb{I}_\varepsilon)\beta$, since they correspond to a weak signal, the dispersion term in the update $\beta \mid \eta, \xi, \sigma^2$ dominates.

Theorem C.4.2. *Let $(X^{(t)})_{t\geq 0}$ be the AR(1) process with recursion*

$$X^{(t+1)} - \mu = B(X^{(t)} - \mu) + AZ^{(t)}, \;\; Z^{(t)} \sim \mathcal{N}_d(0_d, I_d).$$

Let the spectral radius $\rho(B) < 1$ and let $\mu^{(t)} = \mathbb{E}[X^{(t)}]$ and $\Sigma^{(t)} = \mathrm{Var}(X^{(t)})$. The following claims hold:

(i) The process converges to the stationary distribution $\pi^{(\infty)} = \mathcal{N}(\mu^{(\infty)}, \Sigma^{(\infty)})$, where $\mu^{(\infty)} = \mu$ and $\Sigma^{(\infty)} = \sum_{n\geq 0} B^n AA^\top (B^n)^\top$.

(ii) $\mu^{(t)} - \mu^{(\infty)} = B^t(\mu^{(0)} - \mu^{(\infty)})$ and $\Sigma^{(t)} - \Sigma^{(\infty)} = B^t(\Sigma^{(0)} - \Sigma^{(\infty)})(B^t)^\top$ for all $t \geq 0$.

(iii) If $\Sigma^{(0)} \succeq \Sigma^{(\infty)}$, then $\Sigma^{(t)} \succeq \Sigma^{(\infty)}$ for all $t \geq 0$.

(iv) If $X^{(0)}$ is Gaussian, then $X^{(t)}$ is Gaussian for all $t \geq 0$.

Proof. Taking means and variances in the autoregression, we obtain

$$\mu^{(t+1)} - \mu = B(\mu^{(t)} - \mu), \quad \Sigma^{(t+1)} = B\Sigma^{(t)}B^\top + AA^\top.$$

For claim (i), the convergence part is well-known (Tjøstheim, 1990). The stationary distribution $\pi^{(\infty)} = \mathcal{N}(\mu^{(\infty)}, \Sigma^{(\infty)})$ is a fixed point of the autoregression; the solutions $\mu^{(\infty)} = \mu$ and $\Sigma^{(\infty)} = \sum_{n\geq 0} B^n AA^\top (B^n)^\top$ can be seen by inspection.

For claim (ii), since $\Sigma^{(\infty)}$ is a fixed point of the autoregression, it holds that $\Sigma^{(\infty)} = B\Sigma^{(\infty)}B^\top + AA^\top$. Subtracting this off from the autoregression, we obtain that $\Sigma^{(t+1)} - \Sigma^{(\infty)} = B(\Sigma^{(t)} - \Sigma^{(\infty)})B^\top$. Similarly, $\mu^{(t+1)} - \mu^{(\infty)} = B(\mu^{(t)} - \mu^{(\infty)})$. The claim follows by induction.

Claim (iii) follows from claim (ii).

Claim (iv) follows from the closure of Gaussians under affine transformations. □

C.4.4 Verifying the claims of Remark 5.5.2

Underdamped Langevin. We consider the underdamped Langevin diffusion (ULD)

$$\mathrm{d}\begin{bmatrix} X^{(t)} \\ Z^{(t)} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} A^{-1}Z^{(t)} \\ \nabla \log \pi(X^{(t)})\mathrm{d}t - \gamma Z^{(t)} \end{bmatrix} + \begin{bmatrix} 0 \\ (\gamma A)^{1/2}\mathrm{d}W_t \end{bmatrix}$$

with stationary distribution $\pi \otimes \mathcal{N}_d(0_d, A)$, where $(W_t)_{t\geq 0}$ is Brownian motion, $\gamma \in (0, \infty)$ is a friction parameter, and $A \succ 0$ is a preconditioner.

We now verify that overdispersion persists in the X-coordinate when the target is $\pi = \mathcal{N}_d(\mu, \Sigma)$. Suppose that $X^{(0)}$ is drawn independently from $Z^{(0)} \sim \mathcal{N}_d(0_d, A)$. Since the stationary and initial distributions factorize over the X- and Z-components, and furthermore since any time-discretization of the ULD is an AR(1) process, Theorem C.4.2(ii) provides

$$\begin{bmatrix} \Sigma^{(t)} - \Sigma & * \\ * & * \end{bmatrix} = B_t \begin{bmatrix} \Sigma^{(0)} - \Sigma & 0 \\ 0 & 0 \end{bmatrix} B_t^\top, \text{ for some } B_t \text{ and for all } t \geq 0,$$

where the blocks represent the X- and Z-components, where $\Sigma^{(t)} := \mathrm{Var}(X^{(t)})$, and where $*$ denotes an arbitrary entry. Therefore, $\Sigma^{(0)} \succeq \Sigma^{(\infty)}$ implies that $\Sigma^{(t)} \succeq \Sigma^{(\infty)}$ for all $t \geq 0$, as desired.

Random scan Gibbs. For random scan Gibbs samplers targeting Gaussians, we can prove the following result related to overdispersion over time.

Proposition C.4.3. *Let $(X^{(t)})_{t\geq 0}$ be a random scan Gibbs sampler targeting $\pi^{(\infty)} = \mathcal{N}_d(\mu^{(\infty)}, \Sigma^{(\infty)})$. The following claims hold:*

(i) If $\pi^{(0)}$ is Gaussian, then $\pi^{(t)}$ is a mixture of Gaussian distributions for all $t \geq 0$, say $\pi^{(t)} := \sum_{k=1}^{K^{(t)}} p_k \mathcal{N}(\mu_k^{(t)}, \Sigma_k^{(t)})$.

(ii) Let $\pi^{(0)} = \mathcal{N}(\mu^{(0)}, \Sigma^{(0)})$. *If* $\Sigma^{(0)} \succeq \Sigma^{(\infty)}$, *then* $\Sigma_k^{(t)} \succeq \Sigma^{(\infty)}$ *for all* (t, k). *Therefore,* $\pi^{(t)} \overset{\text{PCA}}{\rightsquigarrow} \pi^{(\infty)}$ *for all* $t \geq 0$.

Proof. Representing the random scan Gibbs kernel as a mixture of Gibbs steps, we can write the evolution of the chain as

$$X^{(t+1)} = \sum_{m=1}^{M} \mathbb{1}_{\{M^{(t)}=m\}} \left(B_m X^{(t)} + A_m Z^{(t)}\right), \;\; Z^{(t)} \sim \mathcal{N}_d(0_d, I_d) \tag{C.4.1}$$

where $M^{(t)} \sim \text{Categorical}(p_{1:k})$ selects the mixture component, and where each of the components is a $\pi^{(\infty)}$-invariant Gibbs step.

For claim (i), $\pi^{(t)}$ is a Gaussian mixture for all $t \geq 0$ because linear-Gaussian mixture kernels are closed under Gaussian mixtures.

For claim (ii), we argue by induction. The base case $t = 0$ is trivial. Fixing $t \geq 0$, the recursion (C.4.1) implies that for all k, there exist (ℓ, m) such that $\Sigma_k^{(t+1)} = B_m \Sigma_\ell^{(t)} B_m^\top + A_m A_m^\top$. Since all kernels are $\pi^{(\infty)}$-invariant, $\Sigma^{(\infty)}$ is a fixed point of the recursion (C.4.1), hence $\Sigma_k^{(t+1)} - \Sigma^{(\infty)} = B_m(\Sigma_\ell^{(t)} - \Sigma^{(\infty)})B_m^\top$. Therefore, $\Sigma_\ell^{(t)} \succeq \Sigma^{(\infty)}$ implies $\Sigma_k^{(t+1)} \succeq \Sigma^{(\infty)}$. By induction, $\Sigma_k^{(t)} \succeq \Sigma^{(\infty)}$ for all (t, k). Finally, because $\overset{\text{PCA}}{\rightsquigarrow}$ is partially closed under mixtures, it follows that $\pi^{(t)} \overset{\text{PCA}}{\rightsquigarrow} \pi^{(\infty)}$ for all $t \geq 0$. □

C.5 Estimating the convergence of Markov chains

C.5.1 Plug-in method with time-averaging

We present a refinement of the MCMC convergence rate estimation method of Section 5.5 that applies to overdispersed initializations only. The method proceeds as follows.

We simulate n replicate Markov chains with marginals $(\pi^{(t)})_{t \geq 0}$ up to a large time $T \gg 1$. We collect the samples from $\pi^{(t)}$ with equal weight in $\pi_n^{(t)}$, for $t \geq 0$. We then

estimate

$$L_{T,t} \lessapprox \mathcal{W}_2^2(\pi^{(T)}, \pi^{(t)}) \lessapprox U_{T,t}$$

when $\pi^{(t)}$ is more dispersed than $\pi^{(T)}$, where

$$U_{T,t} := \mathcal{W}_2^2(\pi_n^{(T)}, \pi_n^{(t)}) - \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} \mathcal{W}_2^2(\pi_n^{(T)}, \pi_n^{(S)}),$$
$$L_{T,t} := \left[\mathcal{W}_2(\pi_n^{(T)}, \pi_n^{(t)}) - \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} \mathcal{W}_2(\pi_n^{(T)}, \pi_n^{(S)}) \right]_{\pm}^2 =: \left[\bar{L}_{T,t}\right]_{\pm}^2 .$$

We quantify the uncertainty of $\{U_{T,t}, L_{T,t}\}$ as described in Appendix C.2.4.

The estimators are valid when the MCMC algorithm has reached stationarity by time S and has thereafter mixed *at least once* by time T. In practice, we trace $\mathcal{W}_2^2(\pi_n^{(T)}, \pi_n^{(t)})$ from $t = 0$ until one integrated autocorrelation time before T, then choose $\mathcal{S}$ as the interval of stationarity of this trace.

If the trace does not become stationary, we increase T. Pilot runs with small n can help speed up the search for a large enough T. Another failure mode is when the trace increases towards stationarity, indicating that time-marginals are underdispersed, and that the sample splitting method of Section 5.5 should be used instead.

C.5.2 p-Wasserstein lagged coupling bound

We extend the coupling-based bound of Biswas et al. (2019) to general Wasserstein distances of arbitrary orders $p \geq 1$, as in equation (5.2.1).

Suppose that we wish to estimate the convergence of a Markov chain with kernel P and initialization $\pi^{(0)}$ towards the stationary distribution $\pi^{(0)}P^\infty$. We consider a construction based on a joint Markov kernel $\tilde{P}((x,y),\cdot)$ with marginals $(P(x,\cdot), P(y,\cdot))$ and a lag parameter $\ell \in \mathbb{N}$: we sample a coupled pair of Markov chains $(\bar{X}^{(t)}, X^{(t)})_{t \geq 0}$ evolving under $\tilde{P}$ that is initialized at $(\bar{X}^{(0)}, X^{(0)}) \in \Gamma(\pi^{(0)}P^\ell, \pi^{(0)})$. Then, by the

triangle and coupling inequalities, we obtain the bound

$$\mathcal{W}_p(\pi^{(0)}P^{\infty}, \pi^{(t)}) \leq \sum_{j\geq 0} \mathcal{W}_p(\pi^{(0)}P^{t+(j+1)\ell}, \pi^{(0)}P^{t+j\ell}) \leq \sum_{j\geq 0} \mathbb{E}\left[c(\bar{X}^{(t+j\ell)}, X^{(t+j\ell)})^p\right]^{1/p}.$$

We estimate this bound by sampling i.i.d. replicates of the ℓ-lag coupling construction, replacing expectations by empirical averages. To ensure that the estimator can be computed in finite time, an elegant solution is to design the joint Markov kernel $\tilde{P}$ such that the chains coalesce in finite time, see Biswas et al. (2019); Jacob et al. (2020b) for coalescive coupling strategies.

The method is appealing, as it only requires keeping track of one-dimensional summary statistics. The bound is informative when sufficiently contractive couplings $\tilde{P}$ can be devised. Choosing the lag ℓ large sharpens the bound by eliminating the inefficiency introduced by the triangle inequality, as demonstrated empirically in Biswas et al. (2019).

C.6 Numerical experiments

C.6.1 Benchmark of assignment problem solvers

Figure C.6.1 compares the assignment problem solvers of Bonneel et al. (2011) and Guthe and Thuerck (2021), and contrasts them against the time spent computing the cost matrix using the linear algebra library Eigen (Guennebaud et al., 2010). We see that both methods scale closer to $O(n^2)$ in practice, and that the method of Guthe and Thuerck (2021) outperforms that of Bonneel et al. (2011) and allows for problems to be solved at sample size $n = 1000$ in around 0.1 seconds, and at $n = 10000$ in 10 seconds.

Figure C.6.1 also shows the wall-time of the Flapjack algorithm (Appendix C.2.1). We see that this scales as $O(n^3)$, but that at sample size $n = 1000$ it only takes around 2 seconds.

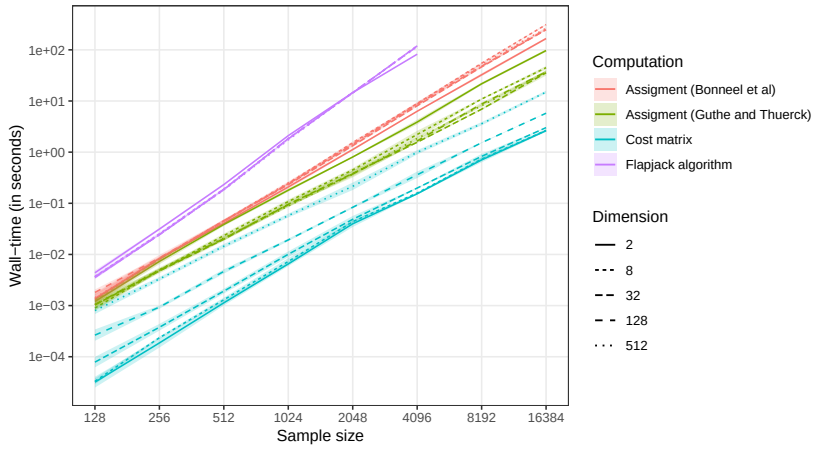

Figure C.6.1: Benchmark of single-core assignment problem solvers. We solved for $\mathcal{W}_2^2(\mu_n, \nu_n)$ with $\mu = \mathcal{N}_d(0_d, I_d)$ and $\nu = \mathcal{N}_d(0_d, 4I_d)$ in various dimensions d and at various sample sizes n. For each dimension, empirical means and standard deviations based on 8 replicates are shown.

C.6.2 Quality of approximate inference methods

Asymptotic bias of unadjusted MCMC algorithms

For the plug-in estimators $\{U, L\}$ we used a sample size of $n = 1024$, based on independent samples for simplicity, and we obtained empirical means and standard deviations from 256 replicates. For the coupling bound, we used $(B, I) = (1000, 2000)$ and $K = 10$.

Tall data

Model. We considered the logistic regression model with likelihood

$$y_i \mid x_i, \beta \sim \text{Bern}(F(x_i^\top \beta)) \text{ independently for observations } i \in [n],$$

where $x_i \in \mathbb{R}^d$ and $F(z) = 1/(1 + e^{-z})$. We approximately followed the guidelines of Gelman et al. (2008), centering the covariates and scaling them to scale 0.5, adding an intercept, and imposing the prior $\beta \sim \mathcal{N}_d(0_d, 25I_d)$. Chopin and Ridgway (2017) lists the posterior log-density, score and Hessian.

MCMC. The MCMC algorithms were preconditioned using the inverse-Hessian at the target mode β^*, and used the proposal covariance $d^{-1/3}[\nabla^2 \log \pi(\beta^*)]^{-1}$, resulting in an $\approx 90\%$ acceptance rate for the MALA kernels. We initialized the MCMC algorithms at the mode β^* and discarded $B = 100$ iterations as burn-in.

Estimators. The parameters used in the main text can be found below. We also experimented with setting the thinning to $T = 1$, and found that nearly identical point estimates $\{V, L\}$ were obtained.

Pima dataset. For the plug-in estimators $\{V, L\}$, we used $(K, I) = (16, 100)$ with thinning $T = 5$ for an overall sample size of $n = 1600$. For the coupling bound, we used $(K, I) = (32, 500)$. We estimated that the coupling reduced the variance of V by factors of roughly $(1.1, 1.5, 1.6, 1.5)$ for (SGLD, SGLD-cv, Laplace, VI).

DS1 dataset. For the plug-in estimators $\{V, L\}$, we used $(K, I) = (16, 200)$ with thinning $T = 10$ for an overall sample size of $n = 3200$. For the coupling bound, we used $(K, I) = (32, 2000)$. We estimated that the coupling reduced the variance of V by factors of roughly $(1.0, 2.2, 1.6, 1.2)$ for (SGLD, SGLD-cv, Laplace, VI).

High-dimensional Bayesian linear regression

The model and sampler are detailed in Appendix C.3.3.

Model. We set $a_0 = b_0 = 1$. Since the model does not have an intercept, we centered the covariates and responses.

MCMC. We set $\sigma_{\text{MH}} = 0.8$. We initialized the MCMC algorithms from the prior and we discarded $B = 1000$ iterations as burn-in.

Estimators. For the plug-in estimators $\{U, L\}$, we used $(K, I) = (100, 100)$ for an overall sample size of $n = 10000$, with thinning $T = 50$. For the coupling bound, we used

$(K, I) = (100, 5000)$. We estimated that the coupling reduced the variance of U by factors of roughly $\{22, 18, 7.0, 3.4, 1.7\}$ in order of increasing $\varepsilon \in \{0.0003, 0.001, 0.003, 0.01, 0.03\}$.

C.6.3 Convergence of MCMC algorithms

Additional investigations

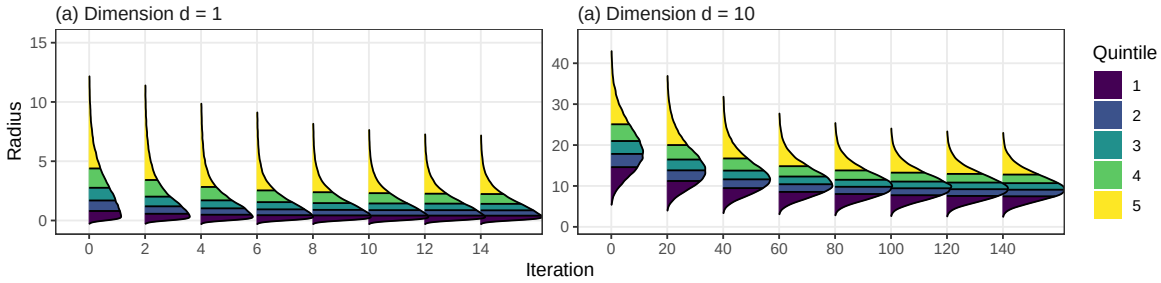

Figure C.6.2: Density plots for the radial component of $\pi^{(t)}$ of a RWM algorithm targeting a multivariate logistic target in various dimensions. See Appendix C.6.3 for details.

Multivariate logistic target. We consider a RWM algorithm with spherical Gaussian proposals with standard deviation h targeting a multivariate logistic target with density $\pi^{(\infty)}(x) \propto e^{-\|x\|}/(1 + e^{-\|x\|})^2$. We initialize the sampler from $\pi^{(0)} \overset{\text{COT}}{\rightsquigarrow} \pi^{(\infty)}$ with density $\pi^{(0)}(x) \propto \pi^{(\infty)}(x/2)$.

Our goal is to verify that overdispersion persists in the sense of $\overset{\text{COT}}{\rightsquigarrow}$. Since the target $\pi^{(\infty)}$ and time-marginal $\pi^{(t)}$ are spherically symmetric, by Proposition 5.3.5, we can verify $\pi^{(t)} \overset{\text{COT}}{\rightsquigarrow} \pi^{(\infty)}$ by checking the dispersion of their radial components.

Figure C.6.2 displays the radial density of $\pi^{(t)}$ against time t, where we considered dimensions, step sizes and acceptance rates of $(d, h, \alpha) \in \{(1, 3, 0.53), (10, 2.5, 0.24)\}$. Since the separation of any two pairs of quantiles gradually concentrates as $t \to \infty$, we conclude that $\pi^{(t)} \overset{\text{COT}}{\rightsquigarrow} \pi^{(\infty)}$ is approximately satisfied for all $t \geq 0$.

Bimodal target. We explore the target $\pi^{(\infty)} = \frac{1}{2}\mathcal{N}(-5, 1) + \frac{1}{2}\mathcal{N}(5, 1)$ by RWM algorithms with Gaussian proposals with standard deviation h and various initializations

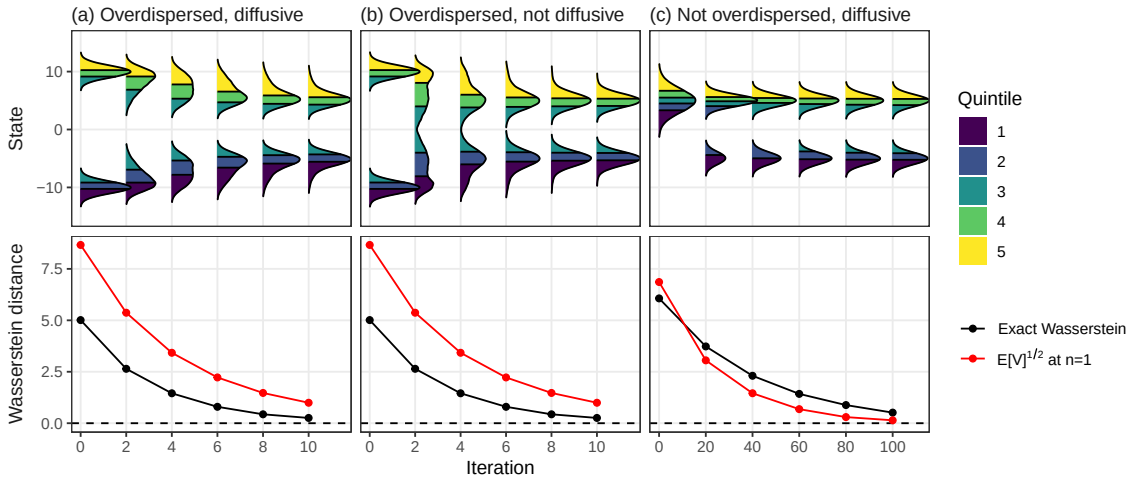

Figure C.6.3: The effect of multimodality on the convergence of an MCMC algorithm. See Appendix C.6.3 for details.

$\pi^{(0)}$. We consider scenarios:

(a) Step size $h = 2$, overdispersed initialization $\pi^{(0)} = \frac{1}{2}\mathcal{N}(-10, 1) + \frac{1}{2}\mathcal{N}(10, 1)$.

(b) Step size $h = 6$, overdispersed initialization .

(c) Step size $h = 4$, initialization $\pi^{(0)} = \mathcal{N}(5, 2)$ located in one of the modes.

Figure C.6.3 displays marginal density plots and compares $\mathbb{E}[V]^{1/2}$ at sample size one to the true Wasserstein $\mathcal{W}_2^2(\pi^{(\infty)}, \pi^{(t)})$. In settings (a) and (b), the marginals are overdispersed with respect to the target and the estimator V is conservative. In setting (c), the marginals are not overdispersed with respect to the target and the estimator V is not conservative, however V is still able to distinguish the marginals from the target.

Gaussian Gibbs sampler

Model. The Gaussian target has precision matrix $\Omega \in \mathbb{R}^{d\times d}$ whose only non-zero entries are $\Omega_{ii} = 1 + \rho^2$ and $\Omega_{i,i\pm1} = -\rho$ for $i \in [d]$, where we identify the indices $(0, d+1)$ as $(d, 1)$.

Estimators. Plug-in estimators $\{U, L\}$ were computed using the method of Appendix C.5.1, based on $n = 1024$ chains, $S \in [2000, 5000]$ and with a thinning factor of

5. As samples from the target $\pi^{(\infty)}$ could be drawn, we set $T = \infty$ for simplicity.

Mixing time of Langevin algorithms

Model. The model and MCMC parameters are as in Appendix C.6.2.

Estimators. Plug-in estimators $\{U, L\}$ were computed using the method of Appendix C.5.1, based on $n = 1024$ chains and $S \in [300, 1000]$. As samples from the target $\pi^{(\infty)}$ could be drawn, we set $T = \infty$ for simplicity.

We estimated the exact mixing time under the assumption that $\pi^{(t)}$ is Gaussian for all $t \geq 0$. This is true for ULA and OBABO, whereas for MALA and the Horowitz method this results in a very slight underestimate of the exact mixing time.

Stochastic volatility model

Model. We considered the stochastic volatility model

$$\begin{aligned} x_1 &\sim \mathcal{N}\left(0, \sigma^2/(1-\varphi^2)\right), \\ x_{t+1} \mid x_t &\sim \mathcal{N}(\varphi x_t, \sigma^2), \qquad \forall t \in [d-1], \\ y_t \mid x_t &\sim \mathcal{N}\left(0, \beta^2 \exp(x_t)\right), \quad \forall t \in [d]. \end{aligned}$$

We fixed $(\beta, \sigma, \varphi) = (0.65, 0.15, 0.98)$ and simulated the data $y_{1:d}$ from the model. Liu (2001, Section 9.6.2) lists the posterior log-density and score.

RWM. Plug-in estimators $\{U, L\}$ were computed using the method of Appendix C.5.1, based on $n = 1024$ chains, $T = 1.5 \times 10^6$, $S \in [5 \times 10^5, 1.25 \times 10^6]$ and thinning every 500 iterations.

MALA. Plug-in estimators $\{U, L\}$ were computed were computed using the method of Appendix C.5.1, based on $n = 1024$ chains, $T = 3 \times 10^4$, $S \in [10^4, 2.5 \times 10^4]$ and thinning every 15 iterations.

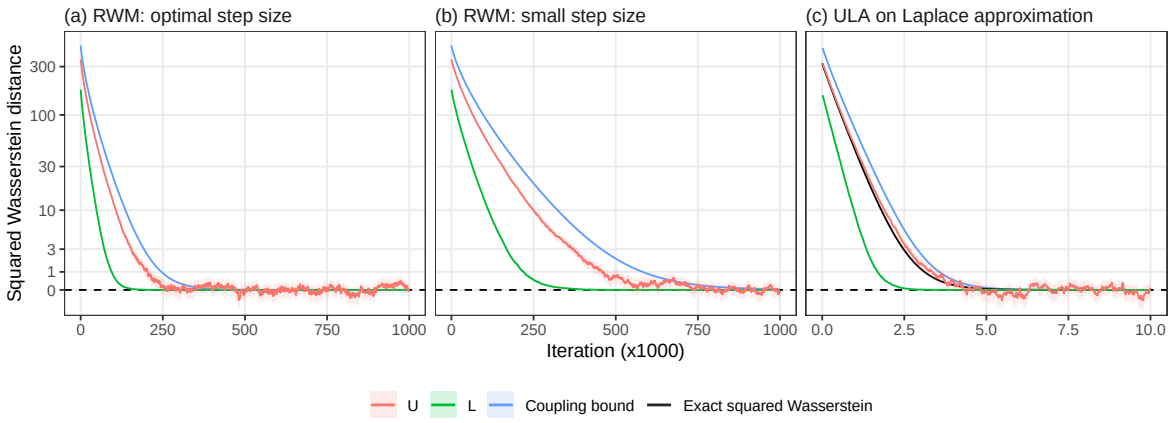

Figure C.6.4: Additional experiments with samplers targeting the stochastic volatility model or its Laplace approximation. See Appendix C.6.3 for details.

Fisher-MALA. Plug-in estimators $\{U, L\}$ were computed were computed using the method of Appendix C.5.1, based on $n = 1024$ chains, $T = 1.25 \times 10^4$, $S \in [7.5 \times 10^3, 9 \times 10^3]$ and thinning every 5 iterations.

The covariance structure of Fisher-MALA was adapted using the default recursion of Titsias (2023), diminishing the adaptation at the rate t^{-1} with the iteration t. The global scale parameter h_t^2 was updated with adaptation diminishing at a rate $t^{-2/3}$ after 1000 iterations, using the recursion $h_{t+1}^2 = h_t^2 + \ell(\alpha_t - \alpha^*) \cdot \min(1, 100t^{-2/3})$ based on the current acceptance probability α_t, the target acceptance probability $\alpha^* = 0.574$ (Roberts and Rosenthal, 1998), and the default learning rate $\ell = 0.015$ of Titsias (2023).

Additional experiments. Figure C.6.4 displays the results of additional experiments. We repeated the RWM experiments in the main text, replacing the coupling with the contractive GCRN coupling of Papp and Sherlock (2025b), finding that the coupling bound became effective but that the proposed estimator U was even sharper. We also considered an ULA targeting a Laplace approximation to the SVM (same parameters as MALA; we used a CRN coupling), finding that U was remarkably close to the exact squared Wasserstein distance.

Bibliography

Agapiou, S., Roberts, G. O., and Vollmer, S. J. (2018). Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models. *Bernoulli*, 24(3):1726–1786.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342.

Andrieu, C., Lee, A., Power, S., and Wang, A. Q. (2024). Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. *The Annals of Applied Probability*, 34(4):4022–4071.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Apers, S., Gribling, S., and Szilágyi, D. (2024). Hamiltonian Monte Carlo for efficient Gaussian sampling: long and random steps. *Journal of Machine Learning Research*, 25(348):1–30.

Atchadé, Y. and Wang, L. (2023). A fast asynchronous Markov chain Monte Carlo sampler for sparse bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1492–1516.

Atchadé, Y. F. and Jacob, P. E. (2024). Unbiased Markov chain Monte Carlo: what, why, and how.

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. (2019). Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference: the minimum distance approach*. CRC press, New York, 1 edition.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Besag, J. (1994). Comments on "Representations of knowledge in complex systems" by U. Grenander and M.I. Miller. *Journal of the Royal Statistical Society: Series B*, 56(4):549–581.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534.

Beskos, A., Roberts, G., Thiery, A., and Pillai, N. (2018). Asymptotic analysis of the random walk Metropolis algorithm on ridged densities. *The Annals of Applied Probability*, 28(5):2966–3001.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Bhattacharya, A., Linero, A., and Oates, C. J. (2024). Grand challenges in Bayesian computation. *Bulletin of the International Society for Bayesian Analysis (ISBA)*, 31(3).

Biswas, N., Bhattacharya, A., Jacob, P. E., and Johndrow, J. E. (2022). Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):973–996.

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, volume 32, pages 7391–7401.

Biswas, N. and Mackey, L. (2024). Bounding Wasserstein Distance with Couplings. *Journal of the American Statistical Association*, 119(548):2947–2958.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bobkov, S. and Ledoux, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261(1259).

Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539–563.

Bolley, F. and Villani, C. (2005). Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Sciences de Toulouse*, 14(3):331–352.

Bonneel, N., van de Panne, M., Paris, S., and Heidrich, W. (2011). Displacement Interpolation Using Lagrangian Mass Transport. *ACM Transactions on Graphics*, 30(6):1–12.

Bortolato, E. (2024). *Advancements in Distribution Sampling for Statistical Inference*. PhD thesis, University of Padua.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3):1209–1250.

Bou-Rabee, N. and Vanden-Eijnden, E. (2010). Pathwise Accuracy and Ergodicity of Metropolized Integrators for SDEs. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(5):655–696.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors (2011). *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC, New York, 1 edition.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Charlier, B., Feydy, J., Glaunès, J. A., Collin, F.-D., and Durif, G. (2021). Kernel Operations on the GPU, with Autodiff, without Memory Overflows. *Journal of Machine Learning Research*, 22(74):1–6.

Chen, M.-F. and Li, S.-F. (1989). Coupling Methods for Multidimensional Diffusion Processes. *The Annals of Probability*, 17(1):151–177.

Chewi, S. and Pooladian, A.-A. (2023). An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *Comptes Rendus. Mathématique*, 361:1471–1482.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems*, volume 33.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer Cham.

Chopin, N. and Ridgway, J. (2017). Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation. *Statistical Science*, 32(1):64–87.

Christensen, O. F., Roberts, G. O., and Rosenthal, J. S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268.

Connor, S. (2007). *Coupling: Cutoffs, CFTP and Tameness*. PhD thesis, University of Warwick.

Corenflos, A., Sutton, M., and Chopin, N. (2023). Debiasing piecewise deterministic markov process samplers using couplings. *arXiv preprint arXiv:2306.15422*.

Corenflos, A. and Särkkä, S. (2022). The coupled rejection sampler. *arXiv preprint arXiv:2201.09585*.

Correa, J. R. and Romero, M. (2021). On the asymptotic behavior of the expectation of the maximum of i.i.d. random variables. *Operations Research Letters*, 49(5):785–786.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Statistical Science*, 28(3):424–446.

Craiu, R. V. and Meng, X.-L. (2022). Double happiness: Enhancing the coupled gains of L-lag coupling via control variates. *Statistica Sinica*, 34(4).

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26.

Dantzig, G. B. (1951). Application of the simplex method to a transportation problem. In *Activity Analysis of Production and Allocation*, pages 359–373.

Darling, R. and Norris, J. (2008). Differential equation approximations for Markov chains. *Probability Surveys*, 5:37 – 79.

Dau, H.-D. and Chopin, N. (2023). On backward smoothing algorithms. *The Annals of Statistics*, 51(5):2145–2169.

Deb, N., Ghosal, P., and Sen, B. (2021). Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections. In *Advances in Neural Information Processing Systems*, pages 29736–29753.

Dehaene, G. and Barthelmé, S. (2018). Expectation propagation in the large data limit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):199–217.

del Barrio, E., González-Sanz, A., and Loubes, J.-M. (2024). Central limit theorems for general transportation costs. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):847–873.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Annals of Probability*, 47(2):926–951.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Dobson, M., Li, Y., and Zhai, J. (2021). Using coupling methods to estimate sample quality of stochastic differential equations. *SIAM/ASA Journal on Uncertainty Quantification*, 9(1):135–162.

Doeblin, W. (1938). Exposé de la théorie des chaines simples constantes de Markoff à un nombre fini d'états. *Mathématique de l'Union Interbalkanique*, 2(77-105):78–80.

Douc, R., Jacob, P. E., Lee, A., and Vats, D. (2023). Solving the Poisson equation using coupled Markov chains. *arXiv preprint arXiv:2206.05691*.

Douc, R., Moulines, E., Soulier, P., and Priouret, P. (2018). *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer Nature, Cham, 1 edition.

Downey, P. J. (1990). Distribution-free bounds on the expectation of the maximum with scheduling applications. *Operations Research Letters*, 9(3):189–201.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Durmus, A. and Moulines, E. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1367–1376.

Eberle, A. (2014). Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377.

Eberle, A. (2016). Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166:851–886.

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010.

Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics*, 9(3):586–596.

Fearnhead, P., Bierkens, J., Pollock, M., and Roberts, G. O. (2018). Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo. *Statistical Science*, 33(3):386–412.

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., and Maji, U. (2021). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, and Kanpur, India. R package version 1.5-0.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070.

Forsythe, G. E. and Leibler, R. A. (1950). Matrix Inversion by a Monte Carlo Method. *Mathematical Tables and Other Aids to Computation*, 4(31):127–129.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738.

Frigessi, A., Gåsemyr, J., and Rue, H. (2000). Antithetic coupling of two Gibbs sampler chains. *The Annals of Statistics*, 28(4):1128–1149.

Gelbrich, M. (1990). On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, 3 edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning Generative Models with Sinkhorn Divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2025). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.3-3.

Gerber, M. and Lee, A. (2020). Discussion on the paper by Jacob, O'Leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483.

Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, chapter 5. CRC Press.

Giovagnoli, A. and Wynn, H. P. (1995). Multivariate dispersion orderings. *Statistics & Probability Letters*, 22(4):325–332.

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

Glynn, P. W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations research*, 40(3):505–520.

Goodman, J. B. and Lin, K. K. (2009). Coupling control variates for Markov chain Monte Carlo. *Journal of Computational Physics*, 228(19):7127–7136.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1292–1301. PMLR.

Gozlan, N. and Léonard, C. (2007). A large deviation approach to some transportation cost inequalities. *Probability Theory and Related Fields*, 139:235–283.

Grushka, E. (1972). Characterization of exponentially modified gaussian peaks in chromatography. *Analytical Chemistry*, 44(11):1733–1738.

Guennebaud, G., Jacob, B., et al. (2010). Eigen v3. http://eigen.tuxfamily.org.

Guthe, S. and Thuerck, D. (2021). Algorithm 1015: A Fast Scalable Solver for the Dense Linear (Sum) Assignment Problem. *ACM Trans. Math. Softw.*, 47(2).

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302.

Horowitz, A. M. (1991). A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2):247–252.

Huggins, J., Kasprzak, M., Campbell, T., and Broderick, T. (2020). Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR.

Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.

Jacob, P. E. (2020). Lecture notes on Couplings and Monte Carlo.

Jacob, P. E., Lindsten, F., and Schön, T. B. (2020a). Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*.

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020b). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Johndrow, J., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61.

Johnson, V. E. (1996). Studying Convergence of Markov Chain Monte Carlo Algorithms Using Coupled Sample Paths. *Journal of the American Statistical Association*, 91(433):154–166.

Johnson, V. E. (1998). A Coupling-Regeneration Scheme for Diagnosing Convergence in Markov Chain Monte Carlo Algorithms. *Journal of the American Statistical Association*, 93(441):238–248.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320.

Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Kallenberg, O. (2021). *Foundations of Modern Probability*. Springer.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Kassraie, P., Pooladian, A.-A., Klein, M., Thornton, J., Niles-Weed, J., and Cuturi, M. (2024). Progressive Entropic Optimal Transport Solvers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kelly, L. J., Ryder, R. J., and Clarté, G. (2023). Lagged couplings diagnose Markov chain Monte Carlo phylogenetic inference. *The Annals of Applied Statistics*, 17(2):1419–1443.

Komarek, P. and Moore, A. W. (2003). Fast robust logistic regression for large sparse datasets with binary outputs. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4, pages 163–170.

Kong, A. (1992). A note on importance sampling using standardized weights. Technical report.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18(14):1–45.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Kuntz, J., Ottobre, M., and Stuart, A. M. (2019). Diffusion limit for the random walk Metropolis algorithm out of stationarity. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 55(3):1599–1648.

Lawson, J. D. and Lim, Y. (2001). The Geometric Mean, Matrices, Metrics, and More. *The American Mathematical Monthly*, 108(9):797–812.

Lindvall, T. (1992). *Lectures on the coupling method*. Wiley, New York.

Lindvall, T. and Rogers, L. C. G. (1986). Coupling of Multidimensional Diffusions by Reflection. *The Annals of Probability*, 14(3):860–872.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, pages 27–40.

Livingstone, S. and Zanella, G. (2022). The Barker Proposal: Combining Robustness and Efficiency in Gradient-Based MCMC. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):496–523.

Ludkin, M. and Sherlock, C. (2022). Hug and hop: a discrete-time, nonreversible Markov chain Monte Carlo algorithm. *Biometrika*, 110(2):301–318.

Ma, Y.-A., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient MCMC. *Advances in Neural Information Processing Systems*, 28.

Mangoubi, O. and Smith, A. (2019). Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 586–595. PMLR.

Mangoubi, O. and Smith, A. (2021). Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions: Continuous dynamics. *The Annals of Applied Probability*, 31(5):2019–2045.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2024). Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998.

Margossian, C. C., Hoffman, M. D., Sountsov, P., Riou-Durand, L., Vehtari, A., and Gelman, A. (2024). Nested $\widehat{R}$: Assessing the Convergence of Markov Chain Monte Carlo When Running Many Short Chains. *Bayesian Analysis*.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–323.

McDiarmid, C. (1989). On the method of bounded differences. In Siemons, J., editor, *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, pages 148–188. Cambridge University Press.

McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17(4):301–315.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2019). Unbiased Smoothing using Particle Independent Metropolis-Hastings. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2378–2387. PMLR.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891.

Miller, R. G. (1974). The Jackknife–A Review. *Biometrika*, 61(1):1–15.

Mills-Tettey, G. A., Stentz, A., and Dias, M. B. (2007). The Dynamic Hungarian Algorithm for the Assignment Problem with Changing Costs. Technical Report CMU-RI-TR-07-27.

Minka, T. P. (2001). Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704.

Monmarché, P. (2021). High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electronic Journal of Statistics*, 15(2):4117–4166.

Munkres, J. (1957). Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Neal, R. M. (1999). Circularly-coupled Markov chain sampling. Technical report, University of Toronto, Department of Statistics.

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, chapter 5. CRC Press.

Negrea, J. (2022). *Approximations and scaling limits of Markov chains with applications to MCMC and approximate inference*. PhD thesis, University of Toronto.

Nemeth, C. and Fearnhead, P. (2021). Stochastic gradient Markov chain Monte Carlo. *Journal of the American Statistical Association*, 116(533):433–450.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Springer.

Nguyen, T. D., Trippe, B. L., and Broderick, T. (2022). Many processors, little time: MCMC for partitions via optimal transport couplings . In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3483–3514. PMLR.

Øksendal, B. (1998). *Stochastic Differential Equations*. Springer-Verlag, 5 edition.

O'Leary, J. (2021). Couplings of the Random-Walk Metropolis algorithm. *arXiv preprint arXiv:2102.01790*.

O'Leary, J. and Wang, G. (2024). Metropolis-Hastings transition kernel couplings. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):1101–1124.

Orlin, J. B. (1997). A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129.

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431.

Papp, T. P. and Sherlock, C. (2025a). Centered plug-in estimation of Wasserstein distances.

Papp, T. P. and Sherlock, C. (2025b). Scalable couplings for the random walk Metropolis algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):772–795.

Paty, F.-P., d'Aspremont, A., and Cuturi, M. (2020). Regularity as regularization: smooth and strongly convex Brenier potentials in optimal transport. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1222–1232.

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5–6):355–607.

Pinto, R. L. and Neal, R. M. (2001). Improving Markov chain Monte Carlo estimators by coupling to an approximating chain. Technical report, University of Toronto, Department of Statistics.

Piponi, D., Hoffman, M., and Sountsov, P. (2020). Hamiltonian Monte Carlo swindles. In *International Conference on Artificial Intelligence and Statistics*, pages 3774–3783. PMLR.

Pitman, J. W. (1976). On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35(4):315–322.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Prado, E., Nemeth, C., and Sherlock, C. (2024). Metropolis–Hastings with Scalable Subsampling. *arXiv preprint arXiv:2407.19602*.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Reutter, A. and Johnson, V. E. (1995). General strategies for assessing convergence of MCMC algorithms using coupled sample paths. Technical report, Duke University, Institute of Statistics & Decision Sciences.

Rhee, C.-H. (2013). *Unbiased estimation with biased samplers*. PhD thesis, Stanford University.

Rhee, C.-H. and Glynn, P. W. (2015). Unbiased estimation with square root convergence for sde models. *Operations Research*, 63(5):1026–1043.

Rippl, T., Munk, A., and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer, New York, 2 edition.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal Scaling of Discrete Approximations to Langevin Diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1):255–268.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317.

Roberts, G. O. and Tweedie, R. L. (1996a). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.

Rosenthal, J. S. (1997). Faithful Couplings of Markov Chains: Now Equals Forever. *Advances in Applied Mathematics*, 18(3):372–381.

Rosenthal, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4(2):207–236.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC.

Ruiz, F. J. R., Titsias, M. K., Cemgil, T., and Doucet, A. (2021). Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. In *Proceedings*

of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, volume 161, pages 707–717. PMLR.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Birkhäuser, Cham, 1 edition.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Sejourne, T., Vialard, F.-X., and Peyré, G. (2022). Faster unbalanced optimal transport: Translation invariant Sinkhorn and 1-d Frank-Wolfe. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 4995–5021. PMLR.

Shaked, M. (1982). Dispersive ordering of distributions. *Journal of Applied Probability*, 19(2):310–320.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer.

Sherlock, C. (2013). Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule. *Journal of Applied Probability*, 50(1):1–15.

Sherlock, C. and Roberts, G. (2009). Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC press.

Sixta, S., Rosenthal, J. S., and Brown, A. (2025). Estimating MCMC convergence rates using common random number simulation. *arXiv preprint arXiv:2309.15735v3*.

Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability & Its Applications*, 1(3):261–290.

Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265. American Medical Informatics Association.

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25.

Solomon, J., Greenewald, K., and Nagaraja, H. (2022). k-variance: A clustered notion of variance. *SIAM Journal on Mathematics of Data Science*, 4(3):957–978.

Staudt, T. and Hundrieser, S. (2024). Convergence of Empirical Optimal Transport in Unbounded Settings. *Bernoulli*, advance publication.

Strassen, V. (1965). The Existence of Probability Measures with Given Marginals. *The Annals of Mathematical Statistics*, 36(2):423–439.

Tarjan, R. E. (1997). Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728.

Titsias, M. (2023). Optimal preconditioning and Fisher adaptive Langevin sampling. In *Advances in Neural Information Processing Systems*, volume 36, pages 29449–29460.

Tjøstheim, D. (1990). Non-linear time series and markov chains. *Advances in Applied Probability*, 22(3):587–611.

Trench, W. F. (1999). Asymptotic distribution of the spectra of a class of generalized Kac–Murdock–Szegö matrices. *Linear Algebra and its Applications*, 294(1):181–192.

van den Boom, W., Jasra, A., De Iorio, M., Beskos, A., and Eriksson, J. G. (2022). Unbiased approximation of posteriors via coupled particle Markov chain Monte Carlo. *Statistics and Computing*, 32(3):36.

Van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796.

Vanetti, P. and Doucet, A. (2020). Discussion on the paper by Jacob, O'leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585.

Vats, D. and Flegal, J. M. (2021). Lugsail lag windows for estimating time-average covariance matrices. *Biometrika*, 109(3):735–750.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337.

Vats, D. and Knudson, C. (2021). Revisiting the Gelman–Rubin Diagnostic. *Statistical Science*, 36(4):518–529.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718.

Villani, C. (2003). *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, Providence, RI.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer.

Von Neumann, J. (1949). On rings of operators. reduction theory. *Annals of Mathematics*, 50(2):401–485.

Walker, A. J. (1977). An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):253–256.

Wang, B. and Titterington, D. M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *International workshop on artificial intelligence and statistics*, pages 373–380.

Wang, G., Blanchet, J., and Glynn, P. W. (2024). When are Unbiased Monte Carlo Estimators More Preferable than Biased Ones? *arXiv preprint 2404.01431*.

Wang, G., O'Leary, J., and Jacob, P. (2021). Maximal Couplings of the Metropolis-Hastings Algorithm. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Wang, T. and Wang, G. (2023). Unbiased Multilevel Monte Carlo Methods for Intractable Distributions: MLMC Meets MCMC. *Journal of Machine Learning Research*, 24(249):1–40.

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

Welling, M. and Teh, Y. W. (2011). Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.

Wu, K., Schmidler, S., and Chen, Y. (2022). Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270):1–63.

Xu, K., Fjelde, T. E., Sutton, C., and Ge, H. (2021). Couplings for multinomial Hamiltonian Monte Carlo. In *International conference on artificial intelligence and statistics*, pages 3646–3654. PMLR.

Yang, J., Roberts, G. O., and Rosenthal, J. S. (2020). Optimal scaling of random-walk Metropolis algorithms on general target distributions. *Stochastic Processes and their Applications*, 130(10):6094–6132.

Zhou, X. (2018). On the Fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*.