Transport map unadjusted Langevin algorithms: learning and discretizing perturbed samplers

Benjamin J. Zhang[*†1], Youssef M. Marzouk[†2], and Konstantinos Spiliopoulos[‡3]

[1]Department of Mathematics and Statistics, University of Massachusetts Amherst
[2]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology
[3]Department of Mathematics and Statistics, Boston University

October 24, 2024

Abstract

Langevin dynamics are widely used in sampling high-dimensional, non-Gaussian distributions whose densities are known up to a normalizing constant. In particular, there is strong interest in unadjusted Langevin algorithms (ULA), which directly discretize Langevin dynamics to estimate expectations over the target distribution. We study the use of transport maps that approximately normalize a target distribution as a way to precondition and accelerate the convergence of Langevin dynamics. We show that in continuous time, when a transport map is applied to Langevin dynamics, the result is a Riemannian manifold Langevin dynamics (RMLD) with metric defined by the transport map. We also show that applying a transport map to an irreversibly-perturbed ULA results in a geometry-informed irreversible perturbation (GiIrr) of the original dynamics. These connections suggest more systematic ways of learning metrics and perturbations, and also yield alternative discretizations of the RMLD described by the map, which we study. Under appropriate conditions, these discretized processes can be endowed with non-asymptotic bounds describing convergence to the target distribution in 2-Wasserstein distance. Illustrative numerical results complement our theoretical claims.

Key words. Langevin dynamics, transport maps, Bayesian inference, Markov chain Monte Carlo, geometry-informed irreversiblility

AMS subject classifications. 62D99, 60H35

1 Introduction

Langevin dynamics frequently appear in sampling methods for high dimensional, non-Gaussian probability distributions. Given access to the gradient of the logarithm of the target density, many Markov chain Monte Carlo (MCMC) methods use Langevin dynamics as a core ingredient. Particular recent interest is in unadjusted Langevin algorithms (ULA), in which the empirical average over a single chain is used to approximate expectations with respect to the invariant distribution, without a Metropolis-Hastings correction.

While standard ULA has proven to be a valuable tool in simulation, it often exhibits slow convergence. Many methods have been proposed to improve the convergence properties of standard ULA. One such class of methods involves simulating from appropriate reversible and irreversible perturbations of the original Markov process, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9]. Such methods have been shown to be effective in accelerating convergence while guaranteeing improved performance of the corresponding ergodic estimator; see Section 2.2.

A different class of sampling methods that has been explored in the literature employs transport maps; see, for instance, [10, 11, 12, 13, 14, 15, 16, 17]. Transport maps provide functional representations of complex random variables by transforming them, deterministically and invertibly, to a simple reference random variable of the same dimension—for example, a standard Gaussian. We discuss these methods in Section 2.3. An "exact" transport map S_{exact} achieves this measure transformation exactly, i.e., it pushes forward a complex target distribution π to a desired reference distribution η_{ref}, a relationship we denote as $(S_{\text{exact}})_\sharp \pi = \eta_{\text{ref}}$. If such a map is constructed, one can efficiently sample from the target π by evaluating S_{exact}^{-1} on samples drawn from the reference distribution.

A crucial starting point for this work is the observation that no matter what form of transport map S and construction procedure are chosen, the pushforward $S_\sharp \pi$ obtained in practice will generally **not** be equal to *the desired reference distribution*—due to limitations of the (finite) approximation space in which the map is sought, the finite number of samples used for estimation, behavior of the associated numerical optimization procedure, etc. We demonstrate this idea in Figure 1, which represents an example analyzed in detail in Section 4.2. In this case, using the transport map S directly for sampling will lead to bias, as $S_\sharp^{-1}\eta_{\text{ref}}$ will differ from the target measure π. As we will see below, combining any such *approximate map* with Langevin dynamics can mitigate this bias. We will also see that approximate maps can *accelerate* the convergence of Langevin sampling for π.
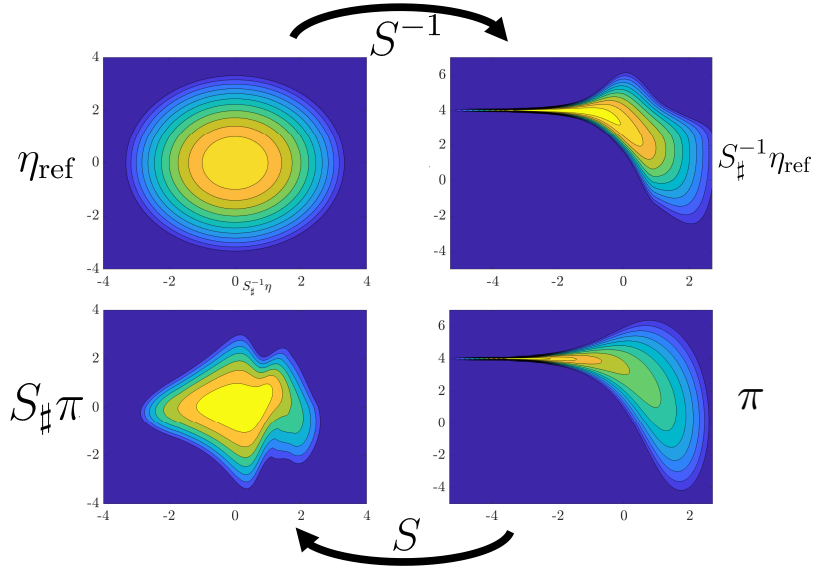

Figure 1: The pushforward distribution $S_\sharp \pi$ obtained with any approximate map S (bottom left) is not equal to the *desired* reference distribution η_{ref} (top left). This is a cartoon representation of the example studied in Section 4.2.

Our contribution. The contribution of this paper is threefold. First, we rigorously demonstrate that applying a transport map to standard overdamped Langevin dynamics (OLD) corresponds to a Riemannian manifold Langevin dynamics (RMLD) [1], which is equivalent to a reversible perturbation of the OLD [7]. Moreover, we show that applying a transport map to an irreversibly-perturbed OLD system results in a geometry-informed irreversible perturbation (GiIrr) [8] of the original dynamics. The resulting reversible or irreversible dynamics are naturally linked to the chosen transport map in a precise way; see Section 3.2. A key consequence of this link is a *systematic* way of learning reversible perturbations—by learning maps.

Second, while transport map transformations of the OLD are equivalent to RMLD or GiIrr in continuous time, we show that when *discretized*, the resulting discrete-time Markov chains have different properties. Specifically, we present the transport map unadjusted Langevin algorithm (TMULA), which is the stochastic process produced by applying a transport map to an Euler-Maruyama discretization of the OLD. The differences between a TMULA and the process produced by a direct discretization of RMLD depend on

2

how strongly the transport map departs from linearity. We discuss these discretization issues in Section 3.3. A schematic representation of these observations is in Figure 2; see Section 2 for notational details. The simulation studies of Section 4 demonstrate that TMULA typically produces better samplers than the discretization of the equivalent RMLD or GiIrr.

$$\eta = S_\sharp \pi \iff \eta(x) = \pi(S^{-1}(x)) \det \mathbf{J}_{S^{-1}}(x)$$
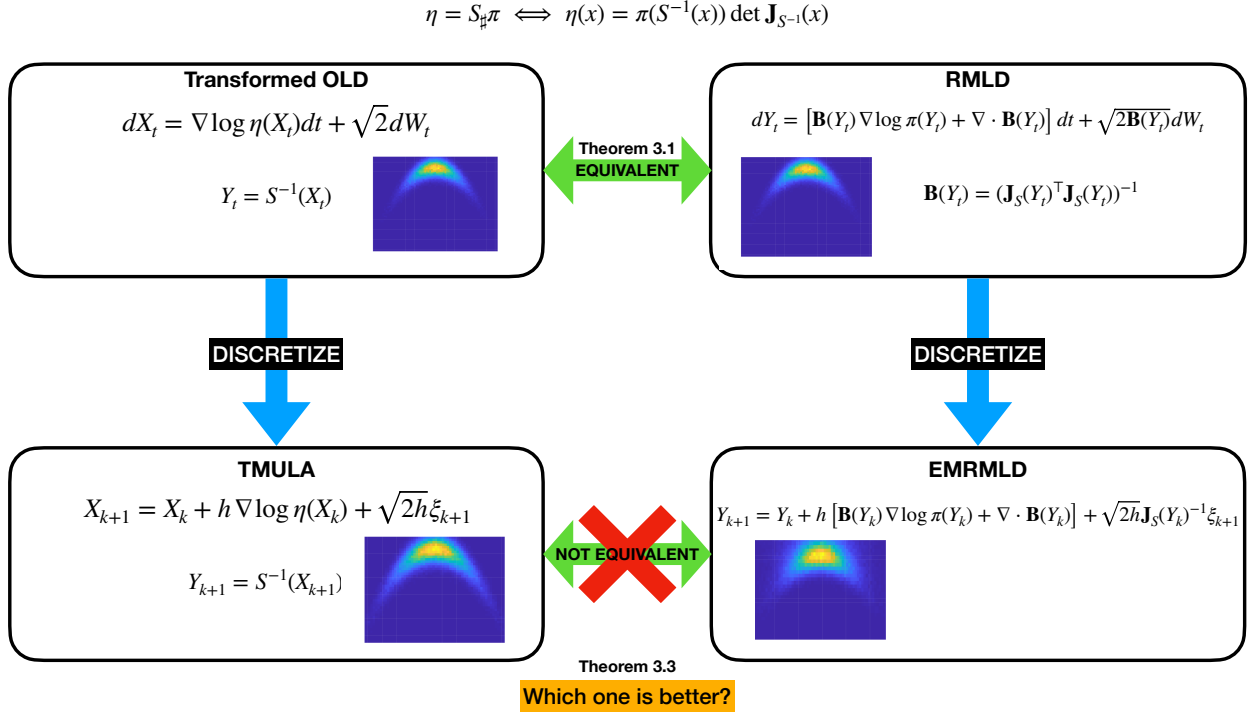

Figure 2: TMULA and RMLD are equivalent in continuous time, but not necessarily after discretization. See Section 3.3 for details. The banana-shaped density in each panel corresponds to each algorithm discussed in Section 4.1. The bottom left panel shows that discretizing TMULA produces a more accurate approximation of the target than discretizing RMLD.

Third, we prove theoretical guarantees of convergence of TMULA in the 2–Wasserstein metric under the assumptions that (a) the transport map is sufficiently monotone (Assumption 3.1) and (b) the logarithm of the pushforward of the target measure under the chosen transport map is strongly convex and has Lipschitz gradients (Assumption 3.2); see Section 3.4. In Remark 3.3, we demonstrate that even when one samples from a log-concave distribution with ULA, it is often practically advantageous to use a transport map when available. Moreover, our theory provides guidance on the ideal choice of reference measure when constructing maps. However, we emphasize that in this paper we do not address the general question of how to judiciously construct transport maps to optimize sampling performance. This question and some others that are left for future work are summarized in Section 5.

Related work. To set these contributions in context, let us discuss the connections between TMULA and other mapped Langevin processes. The mirrored-Langevin [18] and mirror-Langevin dynamics (MLD) [19, 20] are particular versions of TMULA in which the transport map is defined as the gradient of a convex scalar function. MLD [18] is most similar to the dynamics we study here, as the reference process is a Langevin process forced by an isotropic Brownian motion, while the MLD in [20] has a state-dependent diffusion matrix. The transformed unadjusted Langevin algorithm (TULA) is another example of a mapped Langevin process [21], concerned with sampling heavy-tailed distributions. This algorithm uses specific isotropic maps, inspired by the approach in in [22] for random-walk Metropolis samplers, to transform the heavy-tailed target distribution to have sub-exponential tails. Standard ULA is then used to sample from the transformed distribution. In [9], the *optimal* reversible perturbation is characterized as the minimizer of a Riemannian Poincaré constant. The existence of such a minimizer is provided in terms of a so-called moment map.

MLD, TULA, and the design of an optimal Riemannian metric are cases of TMULA when the map or reversible perturbation is defined *a priori*, often for the specific purpose of transforming the support or tails of the target distribution. Meanwhile, a great deal of recent work has focused on *learning* a flexible and expressive transport map from an unnormalized density and/or from samples. For example, normalizing flows—popular in both variational inference and generative modeling—define an invertible map via a composition of simple transformations, parameterized by neural networks; these maps can be learned using evaluations of the unnormalized target density [15], using samples [17, 12, 16], or both [23]. Other classes of maps used in practice include triangular maps [24, 13] or approximations of optimal transport maps [25].

We emphasize that the results in this paper are not tied to any particular class of monotone maps or map construction algorithm. Our goal is to study the impact of a transport map (e.g., one that approximately normalizes a target distribution) on perturbations and discretization of the overdamped Langevin process for that distribution. In fact, one can envision many algorithms that include TMULA as a core element, in which a map is learned either *a priori* via samples or density evaluations, or instead learned and refined on-the-fly [26, 27], and then applied to an overdamped Langevin process. Our numerical studies employ a specific computationally tractable class of monotone triangular maps that approximate the Knothe–Rosenblatt rearrangement (7) (see Section 2.3 and [28, 13, 24] for a quick review), but many other numerical choices are possible.

The rest of the paper is organized as follows. In Section 2, we review reversible and irreversible perturbations to ULA as well as transport maps. Section 3 contains our main results on TMULA, i.e., continuous time equivalence to RMLD or GiIrr, their differences when discretized, and the theoretical results on convergence in Wasserstein metric; proofs are deferred to Appendices A and B. In Section 4, we present four simulation studies to showcase the performance of TMULA. In Section 5, we discuss some future avenues for research.

Glossary of acronyms

For readability, we provide a glossary of the most frequently used acronyms in this paper in Table 1.

Acronym	Meaning	Reference
OLD	Overdamped Langevin dynamics	(1)
ULA	Unadjusted Langevin algorithm	(3), (4)
RMLD	Riemannian manifold Langevin dynamics	(5)
GiIrr	Geometry-informed irreversible perturbation	[8]
TMRMLD	Transport map defined RMLD	Thm 3.1
TMULA	Transport map unadjusted Langevin algorithm	Sec 3.1, (11)
EMRMLD	Euler-Maruyama discretization of TMRMLD	(12)
ATM	Adaptive transport maps	[24], [29]
KSD	Kernelized Stein discrepancy	[30]
UILA	Unadjusted implicit Langevin algorithm	Sec 4.3, [31]
TMUILA	Transport map unadjusted implicit Langevin algorithm	Sec 4.3, (16)

Table 1: Glossary of acronyms used in this paper in order of appearance.

2   Langevin samplers and transport maps

We review the overdamped Langevin dynamics (OLD) and techniques for accelerating their convergence via reversible and irreversible perturbations. We also provide an overview of measure transport.

2.1 Overdamped Langevin dynamics and the unadjusted Langevin algorithm (ULA)

Let Y be a $\mathbb{R}^d$–valued random variable with unnormalized target density π. A stochastic differential equation (SDE) that has π as its invariant density is the so-called OLD:

$$\mathrm{d}Y(t) = \nabla \log \pi(Y(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W(t). \tag{1}$$

Here, $W(t)$ is a standard Brownian motion on $\mathbb{R}^d$. By ergodicity, expectations of test functions $\phi : \mathbb{R}^d \to \mathbb{R}^d$ with respect to π can be approximated by the time average of $\phi(y)$ over a single trajectory, e.g., [32, Section 2.4]:

$$\mathbb{E}_\pi[\phi(Y)] = \int_{\mathbb{R}^d} \phi(y)\pi(y)\mathrm{d}y = \lim_{T \to \infty} \frac{1}{T} \int_0^T \phi(Y(t))\mathrm{d}t. \tag{2}$$

To practically use this relation to estimate expectations with respect to π, the Langevin dynamics must be discretized. An Euler-Maruyama discretization of (1) yields

$$Y_{k+1} = Y_k + h\nabla \log \pi(Y_{k+1}) + \sqrt{2h}\xi_{k+1}, \tag{3}$$

where $\xi^{k+1} \sim \mathcal{N}(0, \mathbf{I})$ are independent, h is the step size, and $Y_k = Y(kh)$; see [33] for a general overview of SDE discretization schemes. The expectation is then approximated by

$$\mathbb{E}_\pi[\phi(Y)] \approx \frac{1}{K} \sum_{k=0}^{K-1} \phi(Y_k). \tag{4}$$

The algorithm computing this estimator is known as the *unadjusted Langevin algorithm* (ULA), and has been used in problems ranging from molecular simulation to Bayesian inference [34]. While the relation in (2) is exact, discretization introduces bias into the estimator. Some previous research focuses on removing this bias by considering the *Metropolis-adjusted Langevin algorithm* (MALA), in which the Langevin dynamics is used as a proposal distribution within the Metropolis-Hastings algorithm [35, 1]. Recently, there has been renewed interest in ULA, especially for high-dimensional problems in machine learning. In particular, [36, 34] provide a collection of convergence results for ULA on strongly log-concave distributions. They derive inequalities showing fast convergence of ULA to the invariant distribution in the 2-Wasserstein and total variation distances, and use those results to obtain bounds on the mean-squared error of ULA for Lipschitz and bounded test functions [34]. Related work extends these results to distributions that satisfy log-Sobolev inequalities [37, 38, 39].

2.2 Accelerating Langevin-based samplers via perturbations

The OLD in (1) is not the only stochastic dynamical system that has π as its invariant density. In fact, there are infinitely many such systems, some of which can converge to π more quickly than OLD [32]. The work of [3, 5, 6, 7] show that appropriately chosen reversible or irreversible perturbations to the OLD can accelerate its convergence to the invariant distribution.

2.2.1 Reversible and irreversible perturbations

Let $\mathbf{B}$ be a matrix-valued valued function that maps $y \in \mathbb{R}^d$ to symmetric positive definite matrices $\mathbf{B}(y) \in \mathbb{R}^{d \times d}$. A reversibly perturbed overdamped Langevin dynamics is given by

$$\mathrm{d}Y(t) = [\mathbf{B}(Y(t))\nabla \log \pi(Y(t)) + \nabla \cdot \mathbf{B}(Y(t))]\,\mathrm{d}t + \sqrt{2\mathbf{B}(Y(t))}\mathrm{d}W(t), \tag{5}$$

where $(\nabla \cdot \mathbf{B}(y))_i = \sum_{j=1}^d \frac{\partial \mathbf{B}_{ij}(y)}{\partial y_j}$. In [7], the authors show that if $\mathbf{B}(y) - \mathbf{I} \succ \mathbf{0}$, then accelerated convergence to the invariant density will be achieved. The degree of accelerated convergence depends on the choice of $\mathbf{B}$, but, in general, there is not a known optimal choice of $\mathbf{B}$. One well-studied reversible perturbation is the *Riemannian manifold Langevin dynamics* (RMLD), in which the matrix $\mathbf{B}(y) = \mathbf{G}(y)^{-1}$ is defined in

terms of the metric of a Riemannian manifold. In particular, when the target distribution is the posterior of a Bayesian inference problem, [1] chooses the metric $\mathbf{G}(y)$ to be the expected Fisher information matrix plus the Hessian of the log-prior. We refer the reader to [40, 41] for further discussion of RMLD and related MCMC methods and to [42] for an exploration of Bayesian inference from an optimization perspective.

With regard to irreversible perturbations, consider the SDE

$$\mathrm{d}Y(t) = [\nabla \log \pi(Y(t)) + \psi(Y(t))]\,\mathrm{d}t + \sqrt{2}\mathrm{d}W(t). \tag{6}$$

An irreversible perturbation corresponds with choosing the term γ such that the resulting SDE is not time-reversible, while preserving the invariant distribution of the unperturbed system. By using the Fokker–Planck equation of (6), one can show that to preserve the invariant measure, ψ should be chosen such that $\nabla \cdot (\psi(y)\pi(y)) = 0$ for all y. A simple and frequently used choice is $\psi(y) = \mathbf{C}\nabla \log \pi(y)$, where $\mathbf{C}$ is a constant skew-symmetric matrix, i.e., $\mathbf{C} = -\mathbf{C}^\top$. This irreversible perturbation is computationally attractive, as applying it requires only one additional matrix-vector multiplication for each step of the Langevin dynamics. The work [8] proposes an irreversible perturbation tailored to an already reversibly-perturbed OLD, and empirically shows the advantages of using a state-dependent skew-symmetric matrix in the irreversible perturbation. The resulting so-called *geometry-informed irreversible perturbation* is $\psi(y) = \mathbf{C}(y)\nabla \log \pi(y) + \nabla \cdot \mathbf{C}(y)$, where $\mathbf{C}(y) = -\mathbf{C}(y)^\top$. Theoretical results show that, in continuous time, the performance of the Langevin sampler cannot be hurt by the choice of the irreversible perturbation; see [3, 5, 6, 7]. We also refer the interested reader to the works of [4, 43, 8] for further results related to irreversible perturbations.

2.3   Transport maps

Transport maps arise from deterministic couplings of probability measures; see [13, 44] for an overview. Let smooth probability densities η and π correspond to probability measures μ_η and μ_π, respectively. A coupling of the two measures is a joint measure ζ whose marginals are μ_η and μ_π. A coupling is deterministic if there exists a measurable function T, called a transport map, such that for all μ_π-measurable sets A, $\mu_\pi(A) = \mu_\eta(T^{-1}(A))$ [44]. This is denoted $\mu_\pi = T_\sharp \mu_\eta$,[1] and μ_π is the pushforward of μ_η through the map T. If T is invertible, then $\pi(y) = \eta(T^{-1}(y)) \det \mathbf{J}_{T^{-1}}(y)$, where $\mathbf{J}_T$ is the Jacobian of T.

Transport maps provide a method for expressing complex random variables in terms of simpler ones. For example, if we choose $X \sim \eta \equiv \mathcal{N}(0, \mathbf{I})$, and construct a map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\pi = T_\sharp \eta$, then we can produce independent realizations of random variable $Y \sim \pi$ by drawing independent realizations of the standard normal and evaluating the transport map at the sample points, i.e., $T(X) \sim \pi$. In general, there are infinitely many maps that can induce a deterministic coupling of X and Y [44], and there are many computational methods for constructing them, e.g., via optimal transport [10, 25, 45], triangular transport [24, 11, 14], normalizing flows [12, 16, 46], and so on.

2.3.1   Monotone triangular transport maps

While the sampling algorithms and theoretical results of this paper will apply to *any* sufficiently smooth and invertible transport map, our numerical illustrations will employ a specific, computationally tractable class: triangular monotone maps. Let $Y = (y_1, \ldots, y_d)$ and consider a map S of the form

$$S(Y) = \begin{bmatrix} S_1(y_1) \\ \vdots \\ S_d(y_1, y_2, \ldots, y_d) \end{bmatrix}, \tag{7}$$

where the ith component of the map is monotone increasing in y_i, i.e., $\frac{\partial S_i}{\partial y_i} > 0$, $\forall i$. This monotonicity condition ensures that S is invertible, i.e., that there exists a map $T = S^{-1}$. A triangular monotone map satisfying $S_\sharp \mu_\pi = \mu_\eta$ is guaranteed to exist when μ_π and μ_η are atomless (which is always the case in this paper, as we assume both measures to have Lebesgue densities on $\mathbb{R}^d$). A map S of the form (7) is often called a Knothe–Rosenblatt (KR) rearrangement; see [47, 48, 28, 13].

---

[1] We will also abuse notation by writing $\pi = T_\sharp \eta$.

Enforcing triangular structure offers several advantages. First, the pushforward density of π through S can be easily computed since the determinant of the Jacobian of S is the product of diagonal entries. Second, the triangular structure of the map allows the inverse to be easily computed. Evaluating $S^{-1}(x)$ involves solving a sequence of d one-dimensional root finding problems, and since the ith entry of each map component is monotone with respect to y_i, the roots are unique. Third, triangular maps enjoy certain sparsity and decomposability properties that follow from the conditional independence structure of π, particularly when η is chosen to be a product measure. These properties are useful when building transport maps in high dimensions, but do depend on the ordering of the variables in the triangular construction; see [49] for details. We refer the interested reader to [24] for a construction of triangular maps that is based only on samples of π and guarantees monotonicity everywhere.

Most methods for computing monotone triangular maps minimize the Kullback–Leibler (KL) divergence between the target π and approximating distribution $T_\sharp \eta^0$, for some fixed choice of η^0; the direction of the KL divergence depends on what information is available. [10] constructs maps given unnormalized evaluations of the density π, by minimizing $D_{\mathrm{KL}}(T_\sharp \eta^0 \| \pi)$ plus a penalty term chosen so that the minimizer is monotone with high probability; the map is represented in a linear space (for instance in the span of a chosen set of polynomials). [24], in contrast, introduces a triangular map parameterization that guarantees monotonicity everywhere, and shows how to construct the transport map given only samples of π. The i-th component function of the map S (7) is chosen to be of the form $S_i(y_1, \ldots, y_i) = f(y_1, \ldots, y_{i-1}, 0) + \int_0^{y_i} g(\partial_i f^i(y_1, \ldots, y_{i-1}, t)) \mathrm{d}t$, where $f^i : \mathbb{R}^i \to \mathbb{R}$ is represented in a polynomial or wavelet basis and $g : \mathbb{R} \to \mathbb{R}_{>0}$ is a bijective and strictly positive function, such as a softplus or shifted exponential linear unit. The parameters of the map are then learned by minimizing an empirical approximation of $D_{\mathrm{KL}}(\pi \| S_\sharp^{-1} \eta)$. This is equivalent to maximum likelihood estimation of the map: given samples $\{Z^i\}_{i=1}^N \sim \pi$, we find $\hat{S} = \arg\max \sum_{i=1}^N \log S_\sharp^{-1} \eta(Z^i)$, where η is a standard Gaussian density and the optimization is over the space of functions $\{f^i\}_{i=1}^d$ in the monotone parameterization described above. Crucially, under appropriate conditions on g, this optimization problem is smooth and has *no spurious local optima*. Moreover, [24] describes a greedy algorithm for enriching the function spaces in which one seeks each f^i, yielding a semi-parametric estimation procedure. We employ this approach to build the maps used in our numerical studies, with code available in the `ATM` GitHub repository [29].

3   Transport map unadjusted Langevin algorithm (TMULA)

Now we present our main methodology, TMULA. The main idea behind TMULA is to apply a transport map to the target distribution so that the pushforward of the target through the map becomes easier to sample using ULA. We show that, in continuous time, this construction can be understood as creating a reversible perturbation to the original OLD, and that it also naturally gives rise to geometrically-informed irreversible perturbations. Comparing TMULA to a direct discretization of the equivalent RMLD, we show that TMULA provides a different and in some cases more accurate approximation of the desired target measure. Finally, building on recent analysis of ULA, we show that in discrete time, under appropriate conditions on the map and target, TMULA converges geometrically to the target distribution in $\mathcal{W}_2$ and that the rate of convergence can be accelerated relative to the original OLD.

3.1   Transforming Langevin dynamics

Suppose we are given a continuously differentiable unnormalized target density π with support on $\mathbb{R}^d$ and an invertible twice-continuously differentiable map $S : \mathbb{R}^d \to \mathbb{R}^d$. TMULA consists in applying ULA to the resulting pushforward density $\eta := S_\sharp \pi$.[2] Samples from π are then obtained by applying the inverse map $T = S^{-1}$ to the trajectory produced by such an ULA.

Different choices of the transport map S yield different versions of TMULA. Recall that $\eta(x) = (S_\sharp \pi)(x) = \pi(T(x)) \det \mathbf{J}_T(x)$, where $\mathbf{J}_T(x)$ is the Jacobian of T. Computing the gradient of $\log \eta(x)$ requires computing

---

[2]Following the discussion in Section 1, we note that in general η will differ from the ideal desired reference η^0 used in any given map construction procedure. Once a map is constructed, the desired reference is no longer directly relevant; we work only with π, S, and the resulting η from the numerical algorithm.

the gradient of the log determinant of S. We have

$$\nabla_x \log \eta(x) = \nabla_x(\log \pi(T(x)) \det \mathbf{J}_T(x)) = \mathbf{J}_T^\top \nabla_y \log \pi(T(x)) - \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_S(y),$$

where $y = T(x)$. While in principle the map S only needs to be invertible and twice continuously differentiable, for computational purposes S should be chosen so that $\nabla_y \log \det \mathbf{J}_S(y)$ can be computed efficiently. For example, [18] defines S as the gradient of a convex scalar function, but only considers cases in which the determinant of the Hessian of this function can be computed exactly. If S is constructed with a normalizing flow [12, 16], there are many parameterizations that allow efficient computation of the log determinant. Triangular maps, as described in Section 2.3.1, in general allow for fast computation not only of the log determinant but also its gradient. In the triangular case, we have $\log \det \mathbf{J}_T(x) = -\log \det \mathbf{J}_S(y) = -\sum_{i=1}^{d} \log \frac{\partial S_i}{\partial y_i}$, and so

$$\nabla_x \log \eta(x) = \mathbf{J}_S^{-\top}(T(x)) \left(\nabla_y \log \pi(T(x)) - \sum_{i=1}^{d} \left(\frac{\partial S_i}{\partial y_i}(T(x)) \right)^{-1} \mathbf{H}_i(T(x)) \right), \tag{8}$$

where $\mathbf{H}_i(y) = \left[\frac{\partial^2 S_i}{\partial y_1 \partial y_i}(y), \ldots, \frac{\partial^2 S_i}{\partial y_d \partial y_i}(y) \right]^\top$; see also [26].

Regardless of the particular choice of the invertible map, applying ULA to $\eta = S_\sharp \pi$ with step size h yields a discrete-time Markov chain,

$$X_{k+1} = X_k + h \nabla_x \log \eta(X_k) + \sqrt{2h} \xi_{k+1}, \ \xi_{k+1} \sim \mathcal{N}(0, \mathbf{I}),$$

which produces trajectories that approximately sample from η, while $Y_k = T(X_k)$ will produce trajectories that approximately sample π.

The idea of transforming a Langevin sampler is not new. Indeed, there have been many instances of using some type of invertible mapping to "reshape" the target distribution so that sampling schemes, such as ULA or MCMC, will perform better. For example, [22] proposes a class of isotropic invertible transformations for sub-exponentially light densities that makes them super-exponentially light. The random-walk Metropolis algorithm can then be shown to be geometrically ergodic for the transformed densities. In the same spirit, [21] uses the same transformations for heavy-tailed densities but focuses on the convergence properties of ULA. [18, 19, 20] consider Langevin sampling for densities with bounded support, transforming to densities supported on $\mathbb{R}^d$. These papers consider specific maps defined as gradients of convex functions, so that the inverse map is given by a Legendre transformation.

In the efforts just described, however, the map S is not tailored to the target density at hand, other than to its support or the decay rate of its tails. In contrast, we are interested in more general situations where an invertible map (in general an *approximate* map) *can be* learned and tailored to reshape the entire density. Recent literature has seen many algorithmic realizations of this idea. [11] learns both normalizing flows and triangular maps from unnormalized target densities, and uses these maps to transform the target density before applying MCMC. [14] takes a similar approach, but proposes using a computationally cheaper approximation of the target density to learn the map, before performing MCMC on a transformation of the true target density. Other work seeks to learn and update a map on-the-fly using samples produced during MCMC: [26] does so with triangular maps, and proves ergodicity of the resulting adaptive MCMC approach; [27] pursues a similar construction with normalizing flows. These efforts have all demonstrated strong empirical performance, but so far there is little theoretical understanding of these performance gains. The use of general transport maps with ULA also has not, to our knowledge, been studied.

More fundamentally, we emphasize that details of how the map is learned or constructed are not our focus here. Rather, we are interested in the underlying question of how the performance of ULA is affected by the introduction of a mapping: how this relates to other perturbations of the Langevin dynamics, what finite-sample convergence guarantees can be derived, and what is the impact of a map on asymptotic bias.

3.2   Connections to perturbations of overdamped Langevin dynamics

One interpretation of TMULA is that the map S "preconditions" the target density to improve the performance of ULA. Earlier, we discussed how reversible perturbations can be interpreted as preconditioning the gradient of the log-density so that Langevin dynamics converges faster. We now make a precise connection

between these two types of preconditioning in the continuous-time setting. Specifically, we show that the stochastic process induced by OLD on η, then transformed by T, is an RMLD [1]. Moreover, we also show that a transport map applied to an irreversibly perturbed OLD produces a GiIrr [8]. The following results assume only that the map is invertible and twice-continuously differentiable. Proofs are given in Appendix A.

3.2.1 Transport maps induce reversible perturbations

Let π be a continuously differentiable probability density with support on $\mathbb{R}^d$ and let S be an invertible, twice-continuously differentiable map, with $T \coloneqq S^{-1}$. Define the density η as the pushforward of π under S, i.e., $\eta \coloneqq S_\sharp \pi$, and let $X(t)$ be the overdamped Langevin dynamics on η:

$$\mathrm{d}X(t) = \nabla \log \eta(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W(t). \tag{9}$$

Theorem 3.1. *(TM + LD = TMRMLD)* *The diffusion process $Y(t) = T(X(t))$, where $X(t)$ is the process (9), evolves according to*

$$dY(t) = [\mathbf{B}(Y(t))\nabla \log \pi(Y(t)) + \nabla \cdot \mathbf{B}(Y(t))]\, dt + \sqrt{2\mathbf{B}(Y(t))}\, dW(t), \tag{10}$$

with $\mathbf{B}(y) = (\mathbf{J}_S^\top \mathbf{J}_S(y))^{-1} = \mathbf{J}_T \mathbf{J}_T^\top(y)$ and $\sqrt{\mathbf{B}(y)} = \mathbf{J}_S^{-1}(y) = \mathbf{J}_T(y)$.

This result implies that prescribing a map S from π to some $\eta = S_\sharp \pi$ is the same as choosing a reversible perturbation of the OLD for π. This link provides a systematic way of building reversible perturbations—beyond, for example, standard choices based on local curvature [1, 50]—by constructing a map S that transforms π to a desired reference distribution. We will comment in Section 3.4 on what constitutes a "good" choice of reference in this setting. In the other direction, the perspective that certain reversible perturbations correspond to transformed Langevin processes allows us to analyze the quality and convergence properties of such RMLD algorithms by leveraging the analysis of ULA; see again Section 3.4.

3.2.2 Transport maps induce geometry-informed irreversibility

Irreversible perturbations to Langevin dynamics are known to accelerate convergence to the equilibrium distribution by taking advantage of the anisotropy of the target distribution. A natural question to ask is: what is the stochastic process that results from applying a transport map to a reference Langevin dynamics that has an irreversible perturbation? The following Theorem states that the output is a geometry-informed irreversible perturbation of the RMLD derived in Theorem 3.1.

Theorem 3.2. *(TM + Irr = TMGiIrr)* *Let $X(t) \in \mathbb{R}^d$ evolve according to*

$$dX(t) = (\mathbf{I} + \mathbf{D})\nabla \log \eta(X(t))dt + \sqrt{2}dW(t)$$

where $\mathbf{D}$ is a constant skew-symmetric matrix, i.e., $\mathbf{D}^\top = -\mathbf{D}$. Then the stochastic process $Y(t) = T(X(t))$ evolves according to

$$dY(t) = [\mathbf{P}(Y(t))\nabla \log \pi(Y(t)) + \nabla \cdot \mathbf{P}(Y(t))]\, dt + \sqrt{2\mathbf{B}(Y(t))}\, dW(t),$$

with $\mathbf{P}(y) = \mathbf{B}(y) + \mathbf{C}(y)$, $\mathbf{B}(y)$ is defined in (10), $\mathbf{C}(y) = \mathbf{J}_T(y)\mathbf{D}\mathbf{J}_T^\top(y)$, $\mathbf{J}_T(y) = \mathbf{J}_S(y)^{-1}$.

Remark 3.1. *(Not all metrics correspond to maps.) Theorem 3.1 showed that a transport map applied to an overdamped Langevin dynamics is equivalent to a Riemannian manifold Langevin dynamics where the metric is related to the Jacobian of the transport map. The converse of this statement, however, is not true. That is, not all RMLD metrics correspond with a transport map. Let us demonstrate this via a simple counterexample. For Bayesian inverse problems, a common heuristic is to choose the metric as the sum of the expected Fisher information matrix and the negative Hessian of the log-prior [1, 50]. In Section 4.2, we use this heuristic, as applied in [1], and find that the metric $\mathbf{G}(y_1, y_2)$ is of the form*

$$\mathbf{G}(y_1, y_2) = \mathbf{B}(y_1, y_2)^{-1} = \begin{bmatrix} 2N\beta + e^{y_2} & 0 \\ 0 & Ne^{-2y_2} + 1/3 \end{bmatrix},$$

9

LD	$\mathrm{d}Y_t = \nabla \log \pi(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$
RMLD	$\mathrm{d}Y_t = [\mathbf{B}(Y_t)\nabla \log \pi(Y_t) + \nabla \cdot \mathbf{B}(Y_t)]\,\mathrm{d}t + \sqrt{2\mathbf{B}(Y_t)}\mathrm{d}W_t$
Irr	$\mathrm{d}Y_t = (\mathbf{I} + \mathbf{D})\nabla \log \pi(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$
GiIrr	$\mathrm{d}Y_t = [(\mathbf{B}(Y_t) + \mathbf{C}(Y_t))\nabla \log \pi(Y_t) + \nabla \cdot (\mathbf{B}(Y_t) + \mathbf{C}(Y_t))]\,\mathrm{d}t + \sqrt{2\mathbf{B}(Y_t)}\mathrm{d}W_t$

Table 2: Summary of the SDEs that share the same invariant density $\pi(y)$. Here, $\mathbf{B}(y) = \mathbf{B}(y)^\top$, $\mathbf{D} = -\mathbf{D}^\top$, and $\mathbf{C}(y) = -\mathbf{C}(y)^\top$

TM + LD	TMRMLD
$\begin{cases} \mathrm{d}X_t = \nabla \log \eta(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t \\ Y_t = S^{-1}(X_t) \end{cases}$	$\begin{cases} \mathrm{d}Y_t = [\mathbf{B}(Y_t)\nabla \log \pi(Y_t) + \nabla \cdot \mathbf{B}(Y_t)]\,\mathrm{d}t + \sqrt{2\mathbf{B}(Y_t)}\mathrm{d}W_t \\ \mathbf{B}(Y_t) = (\mathbf{J}_S(Y_t)^\top \mathbf{J}_S(Y_t))^{-1} \end{cases}$

Table 3: Summary of Theorem 3.1 stating that these two systems are equivalent. Here, $\eta = S_\sharp \pi$.

where N, β are some positive parameters. This implies that the Jacobian is

$$\mathbf{J}_S(y_1, y_2) = \begin{bmatrix} \sqrt{2N\beta + e^{y_2}} & 0 \\ 0 & \sqrt{Ne^{-2y_2} + 1/3.} \end{bmatrix}$$

If there exists a map that has this matrix as its Jacobian, we find that

$$\frac{\partial S_1}{\partial y_1} = \sqrt{2N\beta + e^{y_2}} \implies S_1(y_1, y_2) = y_1\sqrt{2N\beta + e^{y_2}} + C_1(y_2)$$

$$\frac{\partial S_1}{\partial y_2} = 0 \implies S_1(y_1, y_2) = C_2(y_1),$$

where $C_1(y_2)$ and $C_2(y_1)$ are functions of y_2 and y_1 alone, respectively. These two conditions on $S_1(y_1, y_2)$ cannot be satisfied simultaneously, and therefore, there exists no map that corresponds with this metric.

3.3 Comparing discretizations

Now we study discretizations of the perturbed Langevin systems presented in Section 3.2. Theorem 3.1 showed that there are two equivalent characterizations of the Riemannian manifold Langevin dynamics for π—either as a transformation $T = S^{-1}$ of an overdamped Langevin system for $S_\sharp \pi$, or as a reversible perturbation on π with matrix $(\mathbf{J}_S^\top \mathbf{J}_S)^{-1}$. To simulate (10), we can take advantage of these two characterizations and derive two different discretizations, i.e., two different discrete-time Markov chains.

3.3.1 One-step analysis

The first discretization we consider is simply TMULA, summarized as follows:

$$\begin{aligned} Y_{k+1} := F_{\mathrm{TMULA}}(Y_k, \xi_{k+1}) &= T\left(X_k + h\nabla_x \log \eta(X_k) + \sqrt{2h}\xi_{k+1}\right) \\ &= T\left(S(Y_k) + h\mathbf{J}_S(Y_k)^{-\top}\left[\nabla_y \log \pi(Y_k) - \nabla_y \log \det \mathbf{J}_S(Y_k)\right] + \sqrt{2h}\xi_{k+1}\right). \end{aligned} \tag{11}$$

TM + Irr	TMGiIrr
$\begin{cases} \mathrm{d}X_t = (\mathbf{I} + \mathbf{D})\nabla \log \eta(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t \\ Y_t = S^{-1}(X_t) \end{cases}$	$\begin{cases} \mathrm{d}Y_t = [\mathbf{P}(Y_t)\nabla \log \pi(Y_t) + \nabla \cdot \mathbf{P}(Y_t)]\,\mathrm{d}t + \sqrt{2\mathbf{B}(Y_t)}\mathrm{d}W_t \\ \mathbf{B}(Y_t) = (\mathbf{J}_S(Y_t)^\top \mathbf{J}_S(Y_t))^{-1} \\ \mathbf{P}(Y_t) = \mathbf{B}(Y_t) + \mathbf{J}_S(Y_t)^{-1}\mathbf{D}\mathbf{J}_S(Y_t)^{-\top} \end{cases}$

Table 4: Summary of Theorem 3.2 stating that these two systems are equivalent. Here, $\eta = S_\sharp \pi$.

10

As an alternative, we consider the discretization given by a direct application of the Euler-Maruyama scheme to (10):

$$Y_{k+1} := F_{\text{EMRMLD}}(Y_k, \xi_{k+1})$$
$$= Y_k + h(\mathbf{J}_S^\top \mathbf{J}_S(Y_k))^{-1} \nabla_y \log \pi(Y_k) + h\nabla \cdot (\mathbf{J}_S^\top \mathbf{J}_S(Y_k))^{-1} + \sqrt{2h}\mathbf{J}_S(Y_k)^{-1}\xi_{k+1}. \tag{12}$$

Here, $\xi_{k+1} \sim \mathcal{N}(0, \mathbf{I})$. The next result relates F_{TMULA} with F_{EMRMLD}.

Theorem 3.3. *It holds that*

$$F_{TMULA}(Y_k, \xi_{k+1}) - F_{EMRMLD}(Y_k, \xi_{k+1}) = hN_k(Y_k, \xi_{k+1}) + \mathcal{O}(h^{3/2}),$$

where the i-th component of $N_k(Y_k, \xi_{k+1})$ is a mean-zero non-Gaussian random variable

$$N_k^{(i)}(Y_k, \xi_{k+1}) = \xi_{k+1}^\top \nabla^2 T_i(Y_k)\xi_{k+1} - \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2}(Y_k). \tag{13}$$

Moreover,

$$\mathbb{E}\left[\|F_{TMULA}(Y_k, \xi_{k+1}) - F_{EMRMLD}(Y_k, \xi_{k+1})\|^2\right]$$
$$= h^2\left[\sum_{i,j=1}^d \left(\frac{\partial^2 T_i}{\partial x_j^2}\right)^2 + \sum_{i,j,l=1}^d \left(\frac{\partial^2 T_i}{\partial x_j \partial x_l}\right)^2\right] + \mathcal{O}(h^{5/2}).$$

We prove this result in Appendix C. This result shows that these two discretization schemes share the same first moment, while the discrepancies for higher moments depend on the regularity of the transport map T. The two discretizations are identical when the transport map has second and higher derivatives equal to zero, corresponding to a linear map. Otherwise, in contrast to EMRMLD, TMULA typically takes *non-Gaussian* steps on the support of π.

3.3.2 Approximations of the invariant measure

The ultimate goal of our study of Langevin algorithms is to approximate expectations with respect to π. We comment on how accurately the numerical invariant distributions of EMRMLD and TMULA approximate the true target distribution. Our discussion here is based on the results of [51].

Given a discrete-time stochastic process $\{Y_k\}$ with step size h which approximates a diffusion process $\{Y_t\}$ that has invariant measure $\pi(y)$ on $\mathbb{R}^d$, the asymptotic bias of the ergodic estimator for test function $\phi : \mathbb{R}^d \to \mathbb{R}$ is

$$e(\phi, h) = \lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} \phi(Y_k) - \int_{\mathbb{R}^d} \phi(y)\pi(y)\mathrm{d}y.$$

In [51], it is shown that when the discrete-time process is constructed using a discretization of local weak order p, then the asymptotic bias is of the form $e(\phi, h) = -\lambda_p h^p + \mathcal{O}(h^{p+1})$, where

$$\lambda_p = \int_0^\infty \int_{\mathbb{R}^d} \left(\frac{1}{(p+1)!}\mathcal{L}^{p+1} - A_p\right) u(y, t)\pi(y)\,\mathrm{d}y\,\mathrm{d}t, \tag{14}$$

$\mathcal{L}$ is the generator of the Langevin process, A_p is a linear operator that depends on the choice of discretization scheme, and $u(y, t)$ solves the Kolmogorov backward equation (KBE) $\partial_t u = \mathcal{L}u$ with initial condition $u(y, 0) = \phi(y)$. Both TMULA and EMRMLD are weak order $p = 1$ discretization schemes since the latter is a standard Euler-Maruyama discretization, while the former is an EM discretization of the reference Langevin process that is then mapped forward through transport map T. Using this result, one could show that TMULA is a better discretization scheme for constructing ergodic estimators than EMRMLD by showing that λ_1 for TMULA is smaller in magnitude than that of EMRMLD. While this is generally difficult to

do, we will examine a setting where these constants λ_1 can be computed exactly. Later we will apply this analysis to a particular target π and observable ϕ, and show explicitly that TMULA yields a smaller value.

The following computations requires that map S pushes forward π exactly to a standard normal distribution η. Recall that we are considering a reversibly perturbed Langevin dynamics which has generator

$$\mathcal{L}\phi(y) = \langle b(y), \nabla\phi(y)\rangle + \text{Tr}\left[(\mathbf{J}_S^\top\mathbf{J}_S)^{-1}\nabla^2\phi(y)\right],$$

where $b(y) = (\mathbf{J}_S^\top\mathbf{J}_S)^{-1}\nabla\log\pi(y) + \nabla \cdot [(\mathbf{J}_S^\top\mathbf{J}_S)^{-1}](y)$. Studying the asymptotic bias of TMULA can be reduced to analyzing the ergodic estimator for the test function $\phi \circ T$ for the overdamped Langevin process on $\eta(x)$. Here, $\mathcal{L}^{\text{TMULA}}\phi = \langle\tilde{b},\nabla\phi\rangle + \Delta\phi$, where $\tilde{b}(x) = \nabla\log\eta(x)$; using results from [52] we have,

$$A_1^{\text{TMULA}}\phi = \frac{1}{2}\tilde{b}^\top\left[\nabla^2\phi\right]\tilde{b} + \sum_{i=1}^d \phi^{(3)}(e_i, e_i, \tilde{b}) + \frac{1}{2}\sum_{i,j=1}^d \phi^{(4)}(e_i, e_i, e_j, e_j),$$

where $\phi^{(3)}$ and $\phi^{(4)}$ are multilinear forms of the third and fourth derivative terms and e_i is the canonical basis in $\mathbb{R}^d$.

Likewise, for EMRMLD, one can derive that

$$A_1^{\text{EMRMLD}}\phi = \frac{1}{2}b^\top\left[\nabla^2\phi\right]b + \frac{1}{6}\sum_{i,j,k=1}^d \frac{\partial^3\phi}{\partial x_i\partial x_j\partial x_k}(b_i\mathbf{Q}_{jk} + b_j\mathbf{Q}_{ik} + b_k\mathbf{Q}_{ij})$$

$$+ \frac{1}{8}\sum_{i,j,k,l=1}^d \frac{\partial^4\phi}{\partial x_i\partial x_j\partial x_k\partial x_l}\sum_{a,b=1}^d g_{ia}g_{jb}g_{ka}g_{lb}$$

where $\mathbf{Q} = (\mathbf{J}_S^\top\mathbf{J}_S)^{-1}$ and $g_{ij} = (\mathbf{J}_S^{-1})_{ij}$. Since S pushes π to a standard normal distribution exactly, the solution to the KBE can be computed exactly via Hermite polynomial expansions.

In Section 4.1, we will show an example where the constants λ_1 from (14) controlling the asymptotic bias of both schemes are analytically computable in this way and verifiable via simulation—demonstrating, for this example, that TMULA has smaller bias than EMRMLD.

3.4   Convergence in Wasserstein distance

In this section, we visit the convergence guarantees provided for ULA for log-concave distributions. We show that TMULA converges at a geometric rate in the 2-Wasserstein distance to the target distribution π under proper assumptions. The main idea is that the transport map allows us to consider log-concavity of the *pushforward* density $\eta := S_\sharp\pi$ and to transfer convergence rates back to the target. Define $U(x) := -\log\eta(x) = -\log S_\sharp\pi(x)$. We make the following assumptions on S and U.

Assumption 3.1. *The map S is sufficiently monotone in the sense that there is a $\rho > 0$ such that $\|S(z) - S(z')\| \geq \rho\|z - z'\|$.*

If S satisfies Assumption 3.1, then $T = S^{-1}$ is globally Lipschitz with constant $1/\rho$.

Assumption 3.2. *The function $U(x) = -\log\eta(x)$ is m-strongly convex, with L-Lipschitz gradients. That is, for all $x, y \in \mathbb{R}^d$, there exist m and L such that*

$$U(y) \geq U(x) + \langle\nabla U(x), y - x\rangle + \frac{m}{2}\|x - y\|^2, \qquad \|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|.$$

Recall the 2-Wasserstein distance [18]

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{W:W_\sharp\mu=\nu}\int_{\mathbb{R}^d}\|x - W(x)\|^2 d\mu. \tag{15}$$

Denote η^k and π^k to be the distributions of the discrete-time process X^k and Y^k, respectively, at time step k. Let the time step $h \in (0, \frac{1}{m+L})$ and $\kappa = \frac{2mL}{m+L}$. Then we have the following result, which follows naturally from Theorem 5 of [34].

12

Theorem 3.4. *Let S and $\eta = S_\sharp \pi$ satisfy Assumptions 3.1 and 3.2. Then*

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{1}{\rho^2}\left(1 - \frac{\kappa h}{2}\right)^k\left(2\|y - y^\star\|^2 + \frac{2d}{m} - C\right) + \frac{C}{\rho^2},$$

where $C = \frac{2L^2 d}{\kappa}[h(\kappa^{-1} + h)]\left(2 + \frac{L^2 h}{m} + \frac{L^2 h^2}{6}\right).$

The proof is in Appendix B. An important consequence of Theorem 3.4 is that, via the link between transport and reversible perturbations established earlier, we now have *non-asymptotic convergence results* for certain reversibly perturbed *discretized* Langevin dynamics. These non-asymptotic bounds apply to TMULA, rather than to a direct (e.g., Euler–Maruyama) discretization of RMLD; recall by Theorem 3.1 that these two formulations are equivalent in continuous time, but as we have argued their behavior after discretization is not necessarily the same.

The map S in TMULA is a degree of freedom of the method, and can be refined; hence a natural trade-off emerges. A simple transport map might be easy to compute with, but could yield limited acceleration compared to ULA for π. A complex transport map might add additional computational overhead, but could yield faster convergence by controlling the constants m and L of the pushforward density η. Along these lines, another useful consequence of Theorem 3.4 is that it provides guidance for the choice of S, which we discuss in the next remark.

Remark 3.2. *(What is the optimal η?) We may gain some insight into what η should look like so that maximal acceleration can be achieved according to Theorem 3.4. The rate of convergence can be optimized by considering the term $r = 1 - \frac{\kappa h}{2}$ alone. Choosing h to be as large as possible, the rate is $r = 1 - \frac{mL}{(m+L)^2}$. It turns out this rate is optimized (minimizing r) only if η is an isotropic Gaussian. Fix $m > 0$, we study the function $r(L) = 1 - \frac{mL}{(m+L)^2}$ for $L \in [m, \infty)$. Taking the derivative with respect to L yields $\frac{\partial r}{\partial L} = \frac{m(L-m)}{(m+L)^3}$, which is always positive for $L > m$. Thus, the greatest rate of convergence is attained when $L = m$, which implies that $\nabla^2 U(x) = \mathbf{I}$, and therefore implies that η is identically a standard normal. This result supports the intuition that constructing a map S such that η is as close as possible to an isotropic Gaussian is best.*

Next we briefly make a useful observation on the relation of our results to target distributions that satisfy a log-Sobolev inequality (LSI). Convergence results (e.g., in Kullback–Leibler (KL) divergence) for ULA have also been developed for target distributions that satisfy LSI, e.g., see [53, 37, 38, 39]. Our goal here is to explore the connection between such target distributions and the transport map S. We present a lemma showing that if S satisfies Assumption 3.1 and if its inverse has a globally bounded Jacobian, then if η satisfies a LSI, π must also satisfy a LSI. As is known in the literature, see for example [53], LSI implies exponential convergence in KL divergence of the law of the Langevin diffusion that has η as its invariant distribution to the target distribution η.

We recall that a measure satisfies the log-Sobolev inequality with parameter $\alpha > 0$ when, for every smooth function $f : \mathbb{R}^d \to \mathbb{R}^+$, it holds that

$$\text{Ent}_\eta[f] \leq \frac{1}{2\alpha}\int_{\mathbb{R}^d}\frac{\|\nabla f(y)\|^2}{f(y)}\eta(y)dy,$$

where $\text{Ent}_\eta[f] = \int_{\mathbb{R}^d} f(y)\log f(y)\eta(y)dy - \int_{\mathbb{R}^d} f(y)\eta(y)dy\log\left(\int_{\mathbb{R}^d} f(y)\eta(y)dy\right)$. In this case, we write that η satisfies LSI(α).

Lemma 3.1. *Assume that the map S satisfies Assumption 3.1 with a differentiable inverse $T = S^{-1}$, $T^{-1}(\mathbb{R}^d) = \mathbb{R}^d$, and that the Jacobian of T is uniformly bounded in Frobenius norm,*

$$\sup_{x \in \mathbb{R}^d}\sum_{i=1}^d\sum_{k=1}^d\left|\frac{\partial T_k}{\partial x_i}(x)\right|^2 = \sup_{x \in \mathbb{R}^d}\|J_T(x)\|_2^2 \leq \frac{1}{\rho^2}.$$

Then, if η satisfies LSI(α), we have that the target measure $\pi = T_\sharp\eta$ satisfies LSI($\alpha\rho^2$).

Lemma 3.1 shows that there is no free lunch regarding to exponential convergence to equilibrium. By this we mean that even if we produce an η that satisfies a log-Sobolev inequality, the original target measure must

also satisfy a log-Sobolev inequality if the map $T = S^{-1}$ has all of its partial derivatives uniformly bounded. If this uniform bound on the Jacobian of T does not hold, however, then the original target measure π might indeed not satisfy the log-Sobolev inequality. As we shall see in Remark 3.3 and in Section 4, however, there are advantages of employing a transport map even if π satisfies an LSI.

Remark 3.3. *Here we argue that even if the target π has nice convexity properties, transforming it with a map S via Theorem 3.4 can improve the convergence rate. Consider a two-dimensional example where the target density is Gaussian and hence strongly log-concave. In particular, let $\pi = \mathcal{N}(0, diag(1/m, 1/L))$ for $m < L$. Then $U(x) = -\log \pi(x)$ is m-strongly convex and ∇U is L-Lipschitz. The map S that transforms π to a standard Gaussian is $S(x_1, x_2) = (\sqrt{m}x_1, \sqrt{L}x_2)$. This map S satisfies Assumption 3.1 with $\rho = \sqrt{m}$.*

For the original distribution, since $L > m$, the rate $r = 1 - (\kappa h)/2$ is not optimal. For the transformed distribution η we have $L = m = 1$, and thus we can have $\kappa = 1$ and for h at its maximum value ($h = 1/2$) we get $r = 3/4$, which is the minimal attainable value according to Remark 3.2. By Theorem 3.4, we pay a constant penalty of $1/\rho^2 = 1/m$ multiplying r^k. But it is better to make r small in exchange for a constant factor penalty, since doing so maximizes the rate of geometric convergence.

4 Numerical examples

We empirically study TMULA samplers and compare them to other Langevin-type samplers, such as ULA and standard discretizations of RMLD.

We first describe how sample quality is measured in the following numerical examples. We compute the bias, variance, asymptotic variance, and mean-squared errors of ergodic estimators for some chosen test functions. Given a test function $\phi(Y) : \mathbb{R}^d \to \mathbb{R}$, let $\bar{\phi}_K = \frac{1}{K} \sum_{k=0}^{K-1} \phi(Y_k)$ be the estimator of $\mathbb{E}_\pi[\phi(Y)]$. Here, $\{Y_k\}$ are produced by a single chain of a discretized Langevin process. The asymptotic variance is computed through the method of batch means [54, 8]. The MSE is estimated by computing 100 independent trajectories and then evaluating the relevant observables for each of those trajectories. The true value is either computed analytically, if available, or through a Monte Carlo estimate with 10^8 samples. In addition, we also measure sample quality by computing the kernelized Stein discrepancy (KSD). The KSD is a computable expression that can approximate certain integral probability metrics for a certain class of functions defined through the action of the Stein operator on a reproducing kernel Hilbert space; see [30, 8] for details on the implementation of KSD. For the KSD computations, we simulate 100 independent trajectories of a particular system and then compute the KSD using 10000 sample points along those trajectories.

Each example presented below has a different purpose. In Section 4.1 we present a simple non-Gaussian example that allows us to explicitly compare the asymptotic bias of an Euler-Maruyama discretization of RMLD with that of TMULA, in a setting where they are equivalent in continuous time, using the theory of Section 3.3.2 and via simulation. In Section 4.2 we present a funnel distribution that arises from a Bayesian inference problem. The goal of this example is to compare multiple algorithms: ULA, RMLD with standard metrics in previous literature, RMLD with a transport-derived metric, TMULA, and TMULA with irreversible perturbations. In Section 4.3 we present a more complicated, higher-dimensional example, a "hybrid Rosenbrock" distribution. We show in this example that even though ULA is not able to produce samples from the narrow region of the distribution, TMULA can do so when implemented with a split-step implicit Langevin algorithm. Finally, in Section 4.4, we consider a multimodal distribution. Our goal here is to demonstrate some of numerical challenges that arise when using transport to reshape and sample from multimodal targets.

4.1 Evaluating discretizations on a simple Banana distribution

Consider a banana distribution with, up to an additive constant, $\log \pi(y) = -y_1^2/s^2 - (y_2 + by_1^2 - 100b)^2$. For this example, we can write explicitly a transport map S that pushes the distribution to a standard Gaussian:

$$S(y_1, y_2) = \begin{bmatrix} y_1/s \\ y_2 + by_1^2 - 100b \end{bmatrix}.$$

We choose $s = 4$ and $b = 0.01$, and consider the observable $\phi(y_1, y_2) = y_1^2 + y_1 + y_2^2 + y_2$. Using the symbolic algebra toolbox in MATLAB, we find that $\lambda_1^{\text{TMULA}} = -0.62$ while $\lambda_1^{\text{EMRMLD}} = 34.69$, which
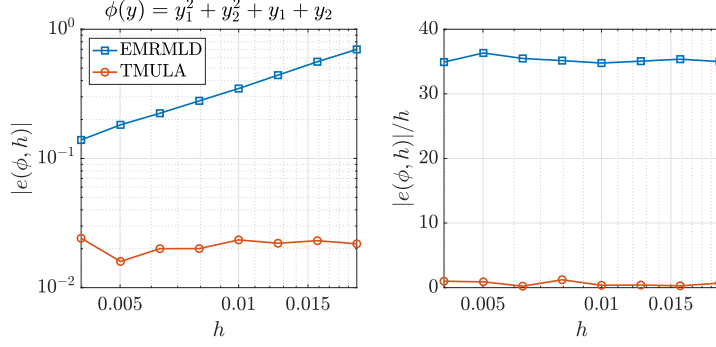
Figure 3: Empirical estimates of $|e(\phi, h)|$ and $e(\phi, h)/h$ for TMULA and EMRMLD as a function of h. The constants λ_1 from (14) are estimated by averaging over $e(\phi, h)/h$. The empirical estimates match well with the theoretical values of λ_1 reported in Section 4.1.

clearly shows that TMULA results in a better ergodic estimator for this test function. We confirm these results numerically by simulating several chains with total simulation time $T = 10^6$, for a range of step sizes h. We show the empirical estimates of the asymptotic bias in Figure 3. The slopes of the error plots, which correspond to λ_1 values in (14), are estimated by computing the term $e(\phi, h)/h$, yielding $\lambda_1^{\text{EMRMLD}} \approx 35.26$ and $\lambda_1^{\text{TMULA}} \approx -0.6345$. These values very closely match our theoretical predictions.

4.2 A funnel density: parameters of a normal distribution

We study a Bayesian inference problem similar to the one studied in ([1], Section 5). Suppose $X \sim N(\mu, \sigma^2)$. Given a dataset $\mathbf{X} = \{X_i\}_{i=1}^N$, we infer the mean and variance. We assume that the prior on μ is a normal distribution with $\mu = 0$ and $\sigma^2 = 3$, and that the prior on σ is a Gamma distribution with shape parameter $\alpha = 0.75$ and rate parameter $\beta = 0.5$. Furthermore, to make the resulting posterior distribution have infinite support, we consider the variable transformation $\gamma = \log \sigma$. Up to an additive constant, the log-posterior distribution is

$$\log \pi(\mu, \gamma | \mathbf{X}) = -N\gamma - \frac{1}{2}e^{-2\gamma} \sum_{i=1}^{N}(X_i - \mu)^2 - \mu^2/6 + \alpha\gamma - \beta e^\gamma.$$

One interesting feature about this posterior distribution is that there is a narrow funnel-shaped region that is difficult for ULA to sample from unless a very small step size is used.

We consider five different Langevin dynamics that all have the posterior distribution as their invariant distribution, and compare their sampling performance. They are denoted ULA, RMLD, TMULA, TMRMLD, and TMRMLD + Irr. ULA is the discretization of the standard Langevin dynamics. RMLD is a Riemannian manifold Langevin dynamics where the metric is chosen according to the one in [1]. In [1] the authors choose the metric to be the sum of the expected Fisher information matrix and the negative Hessian of the log-prior.

$$\mathbf{B}(\mu, \gamma) = \begin{bmatrix} \frac{1}{2N\beta + e^\gamma} & 0 \\ 0 & \frac{1}{Ne^{-2\gamma} + 1/3} \end{bmatrix}.$$

Using samples generated by one trajectory of ULA, a monotone triangular S that pushes the target to an approximate normal distribution is learned using ATM with 20000 sample points and where f is trained over a total order Hermite polynomial basis of order 3 that is rectified so that the Jacobian is positive definite (see Section 2.3) [24]. The sample points are generated by simulating a single ULA chain with a step size of 10^{-6}. Note that this step size is much smaller than the simulations we run to evaluate each system. This is so that ULA produces training samples in the narrow funnel of the distribution. In TMULA we simulate trajectories of ULA on the pushforward density $S_\sharp \pi$ and map each trajectory into trajectories on π through S^{-1}. TMRMLD is a RMLD where the metric is chosen according to $\mathbf{B}(\mu, \gamma)^{-1} = \mathbf{J}_S^\top \mathbf{J}_S(\mu, \gamma)$. While the analytical map of this posterior distribution to a standard normal is not known, we show that a metric that

15

arises from the construction of an *approximate* triangular transport map can still result in better sample quality. In continuous time, Theorem 3.1 shows that `TMULA` and `TMRMLD` are equivalent, but our analysis in Section 3.3 suggests that these two approaches may yield different results when discretized. We empirically explore their differences. In `TMULA + Irr` we also consider the discretization of the system in Theorem 3.2 with the same discretization as `TMULA`.

Figure 4 shows how the map S can be used to sample from the posterior distribution π. Observe that when applied directly to a standard normal distribution, the map produces an approximate target distribution that is visually quite different from the target distribution. In contrast, by performing Langevin on the pushforward of the map S on the target distribution, we can produce samples from the posterior. In Figure 5 we show sample points produced by `ULA` and `TMULA` with the same step size $h = 8 \times 10^{-3}$. Notice that `ULA` is unable to produce samples from the narrow region of this distribution.

We plot the KSD for four of the systems in Figure 6. We do not report the KSD for the `ULA` system as it does not produce samples from the narrow region of the posterior and leads to poor estimates of the KSD. Notice that all the systems with the map-defined metric outperform the metric defined in [1], and that the `TMULA` discretization outperforms the direct Euler-Maruyama discretization of `TMRMLD` (EMRMLD). These findings are further supported by Table 6 and Figure 7, which show the asymptotic variance and MSE convergence plots for test functions $\phi_1(\mu, \gamma) = \exp(\gamma)$, $\phi_2(\mu, \gamma) = \gamma + \mu$, and $\phi_3(\mu, \gamma) = \gamma^2 + \mu^2$.
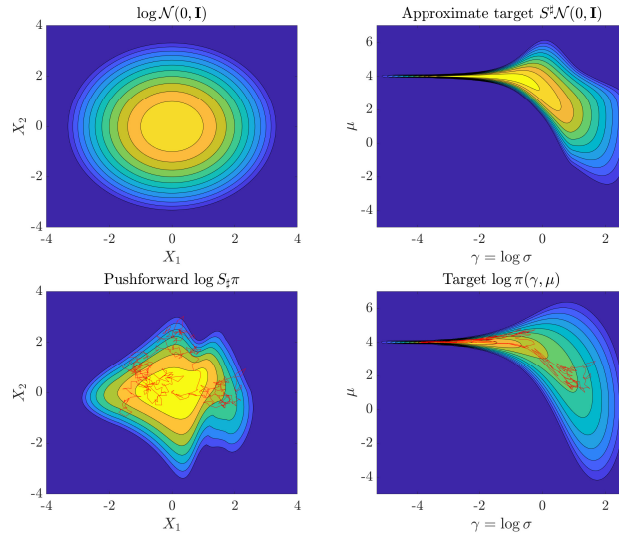

Figure 4: Posterior for parameters of a normal distribution example. Top right figure shows the pullback of normal distribution through map S. Figures in the bottom row show the pushforward of the target π through map S and the true distribution π. Red curves show the Langevin trajectory on the pushforward distribution and on the target distribution.

	$\mathbb{E}[\mathrm{AVar}_{\phi_1}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_1}]$	$\mathbb{E}[\mathrm{AVar}_{\phi_2}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_2}]$	$\mathbb{E}[\mathrm{AVar}_{\phi_3}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_3}]$
`ULA`	8.759	1.797	1.957	0.4774	195.4	35.30
`RMLD`	25.46	7.550	28.82	2.860	1558	184.8
`TMRMLD`	1.344	**0.2057**	2.655	0.3705	108.7	15.48
`TMULA`	1.444	0.2061	2.480	0.3475	114.8	14.00
`TMULA + Irr`	**1.243**	0.2131	**1.961**	**0.2851**	**92.72**	**12.89**

Table 5: Asymptotic variance estimates for the funnel distribution.

4.3 Hybrid Rosenbrock distribution

We demonstrate our methodology on a seven-dimensional hybrid Rosenbrock distribution. This distribution was originally constructed as a generalization of the 2D Rosenbrock distribution for the purposes of testing
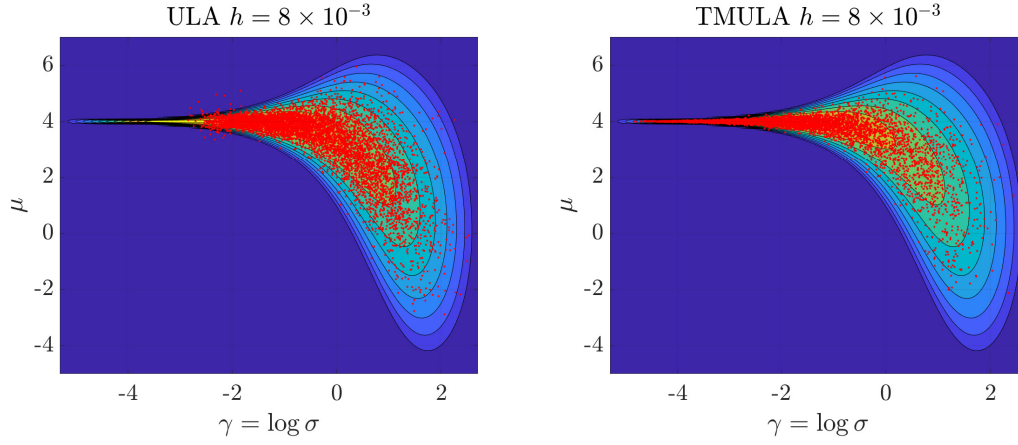
Figure 5: Posterior for parameters of a normal distribution. Left figure shows the samples produced by ULA, right figure shows produced by TMULA. For this step size, note that ULA is unable to produce samples from the narrow region of the distribution.
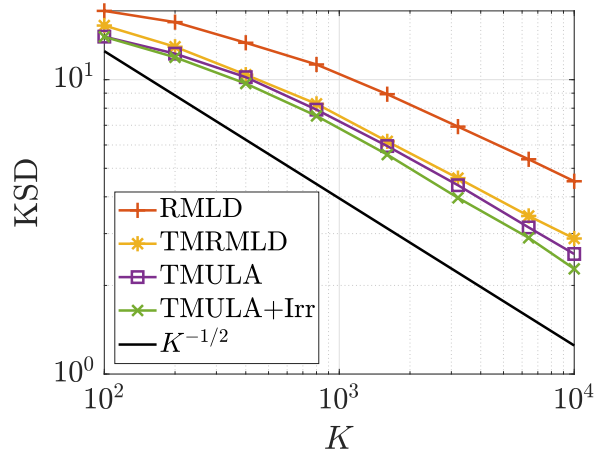

Figure 6: KSD for the funnel distribution. Note that we do not plot the KSD of the `ULA` process as it is unable to produce samples from the narrow region of the posterior and will produce poor estimates of the KSD.
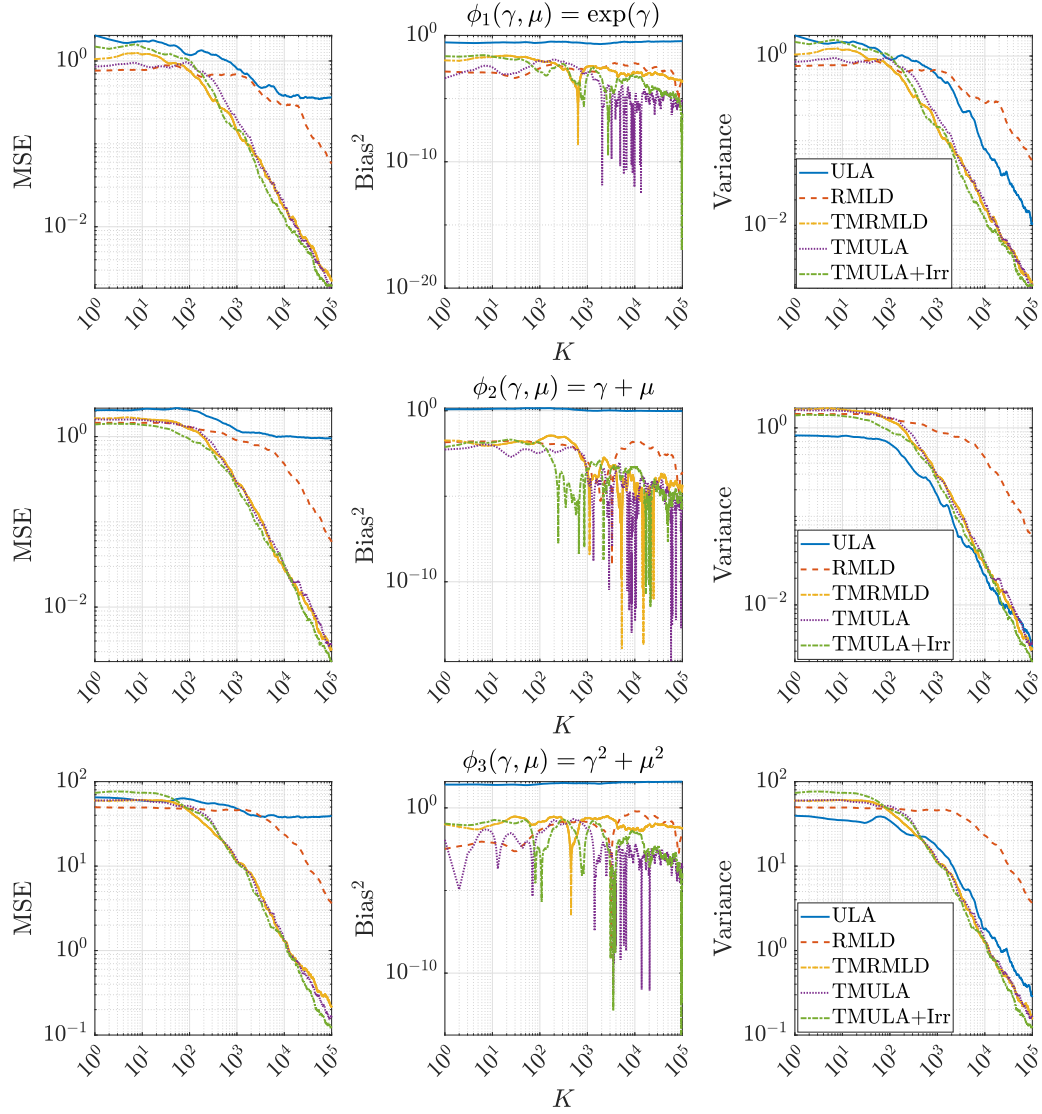
Figure 7: MSE, bias, and variance for the funnel distribution.

sampling schemes [55]. The density is defined as

$$\pi(\mathbf{y}) \propto \exp\left\{-a(y_1 - \mu)^2 - \sum_{j=1}^{n_2}\sum_{i=2}^{n_1} b_{ji}(y_{j,i} - y_{j,i-1}^2)^2\right\}$$

where $\mu, y_{j,i} \in \mathbb{R}$, $a, b_{j,i}$ are strictly positive, and the dimension of the problem is $d = (n_1 - 1)n_2 + 1$. We choose $n_1 = 4$, $n_2 = 2$, $\mu = 1$, $a = 30$, and $b_{j,i} = 20$ for all (j, i). It is important to note that the discrete-time stochastic process that ULA produces for this density is transient, as the drift of the resulting SDE has a cubic term. This phenomenon is well-documented and so ULA cannot be used to reliably produce samples [31]. While a perfect transport map is able to alleviate this problem, an approximate map may not be able to eliminate all the higher order terms, which may cause the stochastic process generated by TMULA to be transient. To solve this problem, for this example we use the split-step implicit Langevin algorithm, which is an implicit algorithm. A deterministic implicit Euler step is first performed before noise from the discretized Brownian motion is added. See [56] for details about this method. In Equation 16 we show how we incorporate the implicit scheme with `TMULA`. We refer to this resulting implicit scheme as TMUILA.

$$
\begin{cases}
S(Y^*) = S(Y^k) + h\mathbf{J}_S^\top(Y^*)^{-1}\left[\nabla_Y \log \pi(Y^*) - \sum_{i=1}^{d}\left(\frac{\partial S_i}{\partial y_i}(Y^*)\right)^{-1} H_i(Y^*)\right] \\
X_{k+1} = S(Y^*) + \sqrt{2h}\xi^{k+1} \\
Y_{k+1} = T(X_{k+1}),
\end{cases}
\tag{16}
$$

where $H_i(Y^k) = \left[\frac{\partial^2 S_i}{\partial y_1 \partial y_i}, \cdots, \frac{\partial^2 S_i}{\partial y_d \partial y_i}\right]^\top$, where $\xi^{k+1} \sim \mathcal{N}(0, \mathbf{I})$.

A transport map is learned with 2500 training samples. In this example, the training samples can be obtained directly since the true map that normalizes the hybrid Rosenbrock distribution is known. We plot the training samples in left of Figure 8. The function f in the transport map is optimized over the span of Hermite polynomials with total order 2 that is then rectified so that its Jacobian is always positive definite.

We compare implicit unadjusted Langevin (UILA) with the implicit version of TMULA (TMUILA) both with step size $h = 0.01$. In right of Figure 8, we compare the KSD of the two systems. Notice that the KSD of UILA plateaus quickly, which implies that the samples are quite biased while TMUILA does not exhibit this phenomenon in our experiments. This implies that the use of the transport map allows us to use larger step sizes when simulating TMUILA. These results are further supported by the asymptotic variance estimates in Table 5 and the MSE plots in Figure 9 test functions $\phi_1(Y) = \sum_{i=1}^{7} Y^i$ and $\phi_s(Y) = \sum_{i=1}^{7} (Y^i)^2$.
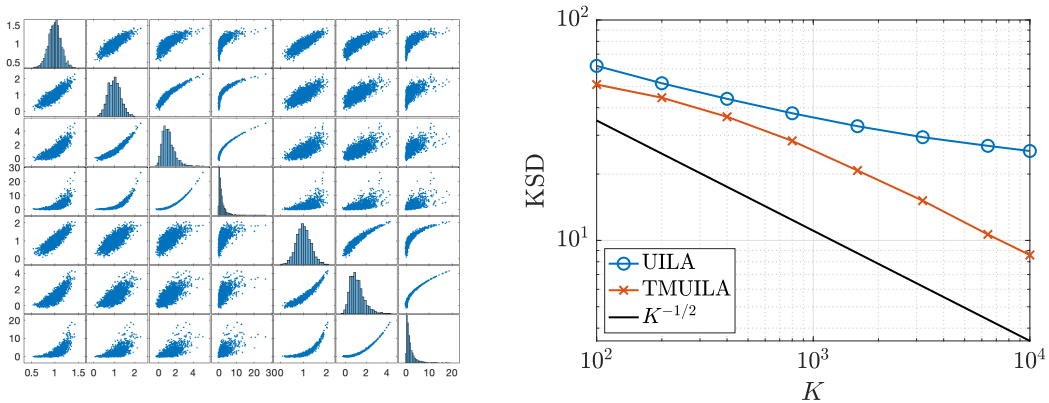

Figure 8: Left: Training samples for the hybrid Rosenbrock example. Right: Kernelized stein discrepancy for the hybrid Rosenbrock distribution.
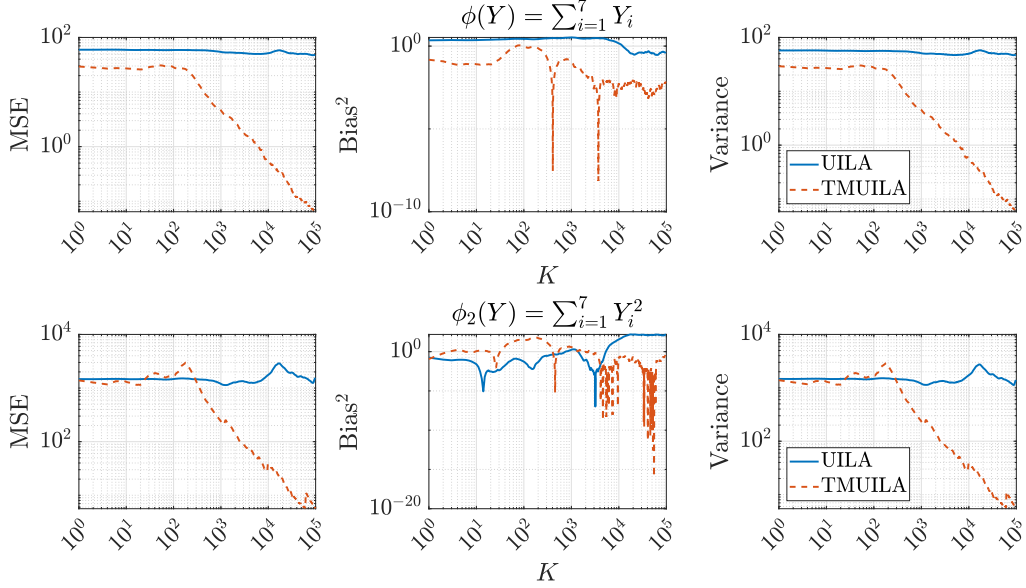
19

Figure 9: MSE for hybrid Rosenbrock distribution.

	$\mathbb{E}[\mathrm{AVar}_{\phi_1}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_1}]$	$\mathbb{E}[\mathrm{AVar}_{\phi_2}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_2}]$
UILA	6762	2663	6.957×10^6	5.185×10^6
TMUILA	**65.03**	**28.54**	**6506**	**1284**

Table 6: Asymptotic variance estimates for the hybrid Rosenbrock distribution.

4.4 Multimodal distributions

We consider a multimodal example. Efficient Langevin sampling of not strongly log-concave distributions is an active area of research. In contrast to the previous examples, we only take note of some qualitative features and challenges of the method for sampling from distributions with multiple modes. The target distribution is a mixture of four Gaussians with means located at $(-4, -4), (4, -4), (-4, 4), (4, 4)$ all with identity covariance matrices. The weights for each Gaussian are $w_1 = 0.337, w_2 = 0.050, w_3 = 0.284, w_4 = 0.328$. In Figure 10 we plot the training samples and the density of the target.

Two maps are learned adaptively with `ATM` using $N = 200$ and $N = 2000$ samples [29, 24]. When applying TMULA with the learned maps, we often observed that the method would still get stuck in certain modes of the target distribution despite the fact that the learned map would push the samples to what appears to be a unimodal distribution. In Figure 11, we show the pushforward samples through the learned maps with $N = 200$ and $N = 2000$ training samples. Notice that while the scatterplot of samples pushed forward through the learned maps looks quite normal and unimodal, the pushforward *densities* have separatrices that make it difficult for TMULA to transition between the multiple modes. More samples can reduce these separations; the $N = 2000$ case shows reduced intensity of the separatrices.

We conjecture that this phenomenon further demonstrates the difficulty of sampling from multimodal distributions with Langevin dynamics. While the transport map can normalize the shape of the distribution, it has challenges in resolving the transitions between modes. We note that this phenomenon is very sensitive to the errors of the map. In Figure 12, we plot the exact and learned maps (with $N = 200$). Notice that the learned map approximates the true normalizing map quite well, even though it still results in a pushforward density that is difficult to sample using Langevin dynamics. We conjecture that separatrices follow from the fact that, when learning maps from samples, it is intrinsically challenging to capture the correct slope (or the Jacobian determinant) of S in regions of low density π, because there is necessarily a lack of sample information in such regions. Further research into this phenomenon and techniques for learning maps that avoid these separatrices are ongoing. Our focus here is on how learned maps interact with ULA sampling,

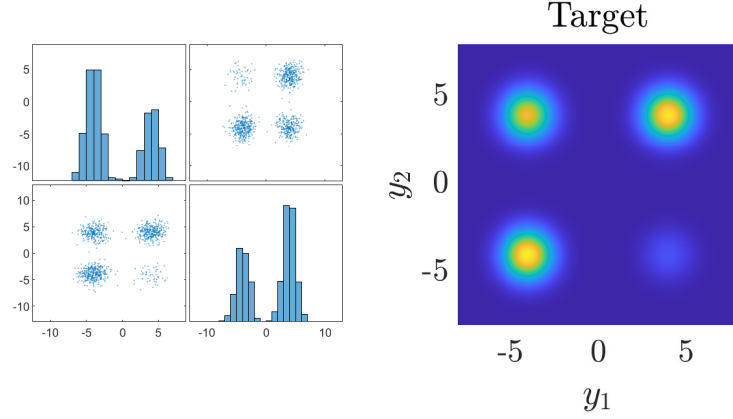and we leave further investigation of how best to obtain maps for multi-modal targets to other work.

Figure 10: Gaussian mixture target distribution. Left figure shows samples, right figure shows target density.
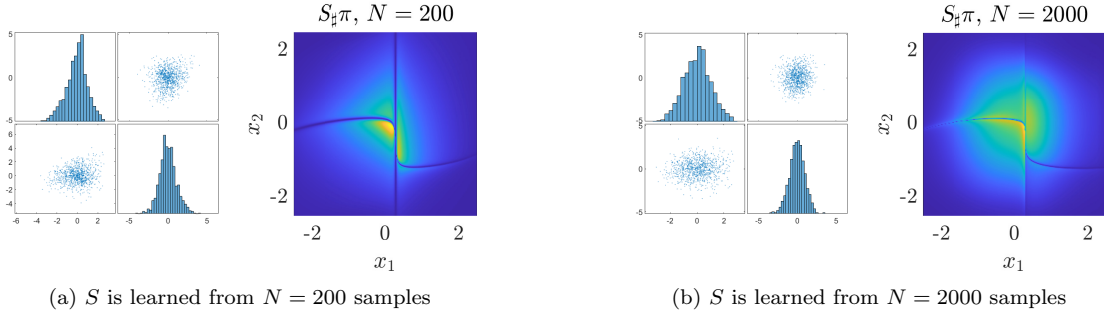

(a) S is learned from $N = 200$ samples

(b) S is learned from $N = 2000$ samples

Figure 11: Mixture of Gaussians. Notice that while the pushforward samples in each figure look quite normal, the pushforward density $S_\sharp \pi$ (computed exactly using the change-of-variables formula) shows boundaries separating the Gaussian into four regions.
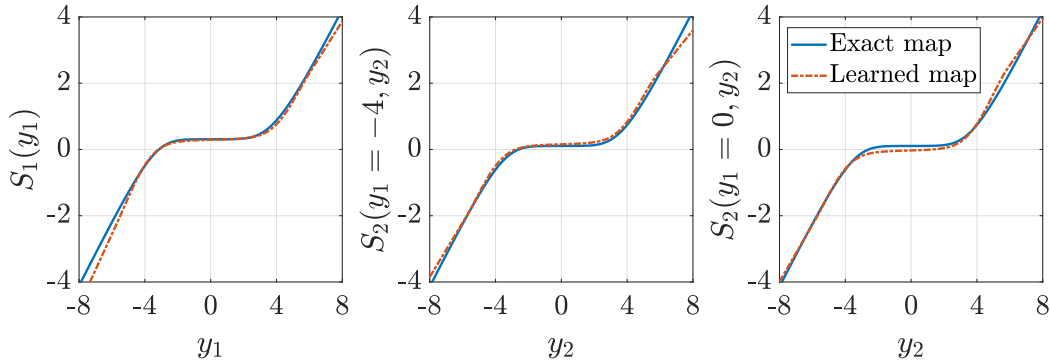

Figure 12: Comparison showing that a transport map can be well approximated but still lead to a problematic pushforward density. Here, $N = 200$.

5 Outlook

In this paper we studied the effects of transport maps on the unadjusted Langevin algorithm. We discussed how the properties of transport maps and target distributions affect the resulting TMULA algorithm. The general question of how to judiciously construct transport maps to optimize sampling performance is left for future work. We discuss many open and interesting questions that work towards maps for optimally improving sampling.

First, building on the results of Section 3.4, it would be useful to improve our theoretical understanding of how to characterize the transport map within a given approximate class (i.e., with finite expressivity) that maximizes the efficiency of TMULA sampling, and to develop practical training objectives for learning such a map. Second, in the numerical example of Section 4.4 we demonstrated the issues that approximate transport maps face in the presence of multimodality. It would be useful to pinpoint the essence of this challenge and to devise workarounds. Third, a transport can sometimes produce a pushforward density with tails lighter than Gaussian, such that the resulting discretized Langevin process is transient. It would be desirable to understand this phenomenon better.

A   Proofs for Section 3.2

Proof of Theorem 3.1. We first derive the SDE of the diffusion process $Z(t) = T(X(t))$. By Itô's lemma [57], the k–th component of $Z(t)$ can be written as

$$\mathrm{d}Z(t) = \mathbf{J}_T \nabla_X \log \eta(S(Z(t)))\mathrm{d}t + c(Z(t))\mathrm{d}t + \sqrt{2}\mathbf{J}_T \mathrm{d}W(t), \tag{17}$$

where $c_k(Z) = \sum_{i=1}^d \frac{\partial^2 T_k}{\partial x_i^2} = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2 T_k}{\partial y_j \partial x_i} \cdot \frac{\partial T_j}{\partial x_i}$. The second equality is true by the multivariate chain rule and will be useful later.

To construct the reversibly perturbed OLD on π, we apply the reversible perturbation to $\{Y(t)\}$ via the matrix $\mathbf{B} = (\mathbf{J}_S^\top \mathbf{J}_S)^{-1}$. After substituting $\pi(y) = \eta(S(y)) \det \mathbf{J}_S(y)$, we have

$$\mathrm{d}Y(t) = (\mathbf{J}_S^\top \mathbf{J}_S)^{-1} \nabla_y \log \left[\eta(S(Y(t))) \det(\mathbf{J}_S)\right] \mathrm{d}t + \nabla_y \cdot (\mathbf{J}_S^\top \mathbf{J}_S)^{-1}\mathrm{d}t + \sqrt{2}\mathbf{J}_S^{-1}\mathrm{d}W(t). \tag{18}$$

First note that the diffusion term is equivalent to that of Equation (17), so we only need to compare the drift terms. Next, notice that

$$\frac{\partial}{\partial y_k} \log(\eta(S(y))) = \sum_{i=1}^d \frac{\partial S_i}{\partial y_k} \frac{\partial}{\partial x_i} \log(\eta(S(y))).$$

Therefore, $\nabla_y \log \eta(S(y)) = \mathbf{J}_S^\top \nabla_x \log \eta(S(y))$ and

$$(\mathbf{J}_S^\top \mathbf{J}_S)^{-1} \nabla_y \log \eta(S(Y)) = \mathbf{J}_S^{-1} \nabla_x \log \eta(S(y))$$

. This exactly matches the first part of the drift term in Equation (17).

For the divergence term, first note the following identity. Let $\mathbf{A}$ and $\mathbf{D}$ be $d \times d$ matrix-valued functions. Then observe that

$$(\nabla \cdot \mathbf{A}\mathbf{D})_k = \sum_{j=1}^d \frac{\partial}{\partial y_j}(\mathbf{A}\mathbf{D})_{kj} = \sum_{j=1}^d \sum_{i=1}^d \frac{\partial}{\partial y_j}\mathbf{A}_{ki}\mathbf{D}_{ij}$$

$$= \sum_{i=1}^d \sum_{j=1}^d \frac{\partial \mathbf{A}_{ki}}{\partial y_j}\mathbf{D}_{ij} + \mathbf{A}_{ki}\frac{\partial \mathbf{D}_{ij}}{\partial y_j} = \sum_{i=1}^d \sum_{j=1}^d \left[\frac{\partial \mathbf{A}_{ki}}{\partial y_j}\mathbf{D}_{ij}\right] + (\mathbf{A}\nabla \cdot \mathbf{D})_k.$$

Applying this identity to $\mathbf{B} = (\mathbf{J}_S^\top \mathbf{J}_S)^{-1} = \mathbf{J}_T \mathbf{J}_T^\top$, we obtain

$$(\nabla_y \cdot \mathbf{J}_T \mathbf{J}_T^\top)_k = \sum_{i,j=1}^d \left[\frac{\partial (\mathbf{J}_T)_{ki}}{\partial y_j} (\mathbf{J}_T)_{ij}^\top \right] + (\mathbf{J}_T \nabla_y \cdot \mathbf{J}_T^\top)_k$$

$$= \sum_{i,j=1}^d \left[\frac{\partial^2 T_k}{\partial y_j \partial x_i} \cdot \frac{\partial T_j}{\partial x_i} \right] + (\mathbf{J}_T \nabla_y \cdot \mathbf{J}_T^\top)_k.$$

Notice that the first term is exactly the Itô correction term. Hence, we only need to show that $(\mathbf{J}_S^\top \mathbf{J}_S)^{-1} \nabla_y \log \det \mathbf{J}_S + \mathbf{J}_T \nabla_y \cdot \mathbf{J}_T^\top = 0$. To avoid issues with the chain rule relating Y and X, it suffices to show that $\nabla_y \log \det \mathbf{J}_S + \mathbf{J}_S^\top \nabla_y \cdot (\mathbf{J}_S^\top)^{-1} = 0$.

For the first term, observe that

$$\frac{\partial}{\partial y_k} \log \det \mathbf{J}_S = \sum_{i,j=1}^d ((\mathbf{J}_S^\top)^{-1})_{ij} \left(\frac{\partial \mathbf{J}_S}{\partial y_k} \right)_{ij} = \sum_{i,j=1}^d \left(\frac{\partial^2 S_i}{\partial y_k \partial y_j} \right) ((\mathbf{J}_S^\top)^{-1})_{ij}$$

$$= \sum_{i,j=1}^d \left(\frac{\partial \mathbf{J}_S}{\partial y_j} \right)_{ik} ((\mathbf{J}_S^\top)^{-1})_{ij}$$

Letting $:j$ denote the j–th column vector, we see

$$(\mathbf{J}_S^\top \nabla \cdot (\mathbf{J}_S^\top)^{-1})_k = - \left(\mathbf{J}_S^\top \sum_{j=1}^d \left[(\mathbf{J}_S^\top)^{-1} \frac{\partial \mathbf{J}_S^\top}{\partial y_j} (\mathbf{J}_S^\top)^{-1} \right]_{:j} \right)_k$$

$$= - \left(\sum_{j=1}^d \left[\frac{\partial \mathbf{J}_S^\top}{\partial y_j} (\mathbf{J}_S^\top)^{-1} \right]_{:j} \right)_k$$

$$= - \sum_{j=1}^d \left[\frac{\partial \mathbf{J}_S^\top}{\partial y_j} (\mathbf{J}_S^\top)^{-1} \right]_{kj}$$

$$= - \sum_{j,i=1}^d \left(\frac{\partial \mathbf{J}_S^\top}{\partial y_j} \right)_{ki} ((\mathbf{J}_S^\top)^{-1})_{ij}.$$

which exactly the negative of the derivative of the log determinant of $\mathbf{J}_S$. We therefore conclude that Equation (17) exactly matches (18). $\square$

Proof of Theorem 3.2. Apply Itô's formula to get

$$\mathrm{d}Y_k(t) = \sum_{i=1}^d \frac{\partial T_k}{\partial x_i} \left[\frac{\partial}{\partial x_i} \log \eta(X(t)) + (\mathbf{D}\nabla \log \eta(X(t)))_i \right] \mathrm{d}t + \sum_{i=1}^d \frac{\partial^2 T_k}{\partial x_i^2} \mathrm{d}t$$

$$+ \sqrt{2} \sum_{i=1}^d \frac{\partial T_k}{\partial x_i} \mathrm{d}W_i(t).$$

The first, third, and fourth terms are identical to the ones in Theorem 3.1 when irreversibility is not considered. Therefore, we only need to address the second term. Note that

$$\mathbf{J}_T \mathbf{D} \nabla_x \log \eta(x) = \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top \nabla_y \log \eta(S(y))$$

$$= \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top \nabla_y \log \left(\eta(S(y)) \det \mathbf{J}_S(y) \right) - \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_S(y)$$

$$= \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top \nabla_y \log \pi(y) - \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_S(y).$$

We only need to show that the last term is equal to $\nabla_y \cdot \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top$. From the proof of the previous Theorem, first observe that

$$(\nabla_y \cdot \mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top)_k = \sum_{i,j=1}^{d} \left[\frac{\partial}{\partial y_j} (\mathbf{J}_T \mathbf{D})_{ki} \frac{\partial T_j}{\partial x_i} \right] + (\mathbf{J}_T \mathbf{D} \nabla_y \cdot \mathbf{J}_T^\top)_k \tag{19}$$

and that $\mathbf{J}_T \mathbf{D} \nabla_y \cdot \mathbf{J}_T^\top = -\mathbf{J}_T \mathbf{D} \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_S$. Therefore, we need to show that the first term on the right hand side in (19) above is identically zero. We compute

$$\sum_{i,j=1}^{d} \left[\frac{\partial}{\partial y_j} (\mathbf{J}_T \mathbf{D})_{ki} \frac{\partial T_j}{\partial x_i} \right] = \sum_{i,j,l=1}^{d} \left[\mathbf{D}_{li} \frac{\partial^2 T_k}{\partial y_j \partial x_l} \frac{\partial T_j}{\partial x_i} \right] = \sum_{i,l=1}^{d} \left[\mathbf{D}_{li} \frac{\partial^2 T_k}{\partial x_i \partial x_l} \right]$$

$$= \sum_{i,l=1,i>l}^{d} \mathbf{D}_{li} \left[\frac{\partial^2 T_k}{\partial x_i \partial x_l} - \frac{\partial^2 T_k}{\partial x_l \partial x_i} \right] = 0.$$

$\square$

B   Proofs for Section 3.4

The proof is straightforward: we use the same approach as taken in [18] and relate the convergence of ULA on $S_\sharp \pi$ to convergence on π by relating the Wasserstein distance between η^k and η with that of π^k and π [34].

Lemma B.1. *Let η^k and π^k be the distribution of the discrete-time process X^k and Y^k at time step k, respectively.*

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{1}{\rho^2} \mathcal{W}_2^2(\eta^k, \eta). \tag{20}$$

Proof. We use the fact that S is ρ-strongly monotone:

$$\mathcal{W}_2^2(\eta^k, \eta) = \mathcal{W}_2^2(S_\sharp \pi^k, S_\sharp \pi) = \inf_W \int \|x - W(x)\|^2 \mathrm{d}S_\sharp \pi^k(x)$$

$$= \inf_W \int \|S(x) - W \circ S(x)\|^2 \mathrm{d}\pi^k(x)$$

$$\geq \rho^2 \inf_W \int \|x - S^{-1} \circ W \circ S(x)\|^2 \mathrm{d}\pi^k(x).$$

Since W is such that $W_\sharp S_\sharp \pi^k = S_\sharp \pi$, this implies that $S_\sharp^{-1} W_\sharp S_\sharp \pi^k = \pi$, and we therefore have

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{1}{\rho^2} \mathcal{W}_2^2(\eta^k, \eta).$$

$\square$

Proof of Theorem 3.4. We apply Theorem 5 of [34] to ULA on η, and then use Lemma B.1 to derive a bound for π. Observe that

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{1}{\rho^2} \mathcal{W}_2^2(\eta^k, \eta) \leq \frac{1}{\rho^2} \left[\left(1 - \frac{\kappa h}{2} \right)^k \left(2\|y - y^\star\|^2 + \frac{2d}{m} - C \right) + C \right],$$

where $C = \frac{2L^2 d}{\kappa} [h(\kappa^{-1} + h)] \left(2 + \frac{L^2 h}{m} + \frac{L^2 h^2}{6} \right).$

$\square$

24

C Proofs for Section 3.3

Proof of Theorem 3.3. Consider the i–th component of Z_{k+1} and its Taylor expansion around the value $Z_k^{(i)} = T_i(X_k)$. Ignoring terms of order $h^{3/2}$ and higher, we have

$$
\begin{aligned}
Z_{k+1}^{(i)} =& T_i\left(X_k + h\nabla_x \log \eta(X_k) + \sqrt{2h}\xi_{k+1}\right) \\
=& T_i(X_k) + (\nabla_x T_i(X_k))^\top \left(h\nabla_x \log \eta(X_k) + \sqrt{2h}\xi_{k+1}\right) \\
&+ \frac{1}{2}\left(h\nabla_x \log \eta(X_k) + \sqrt{2h}\xi_{k+1}\right)^\top \nabla_x^2 T_i \left(h\nabla_x \log \eta(X_k) + \sqrt{2h}\xi_{k+1}\right) \\
&+ \mathcal{O}(h^{3/2}) \\
=& Z_k^{(i)} + h(\nabla_x T_i)^\top \left(\mathbf{J}_T^\top \nabla_y \log \pi(T(X_k)) + \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_T\right) + \sqrt{2h}(\mathbf{J}_T \xi_{k+1})^{(i)} \\
&+ h\xi_{k+1}^{\ \top}\nabla_x^2 T_i \xi_{k+1} + \mathcal{O}(h^{3/2}) \\
=& Z_k^{(i)} + h(\mathbf{J}_T \mathbf{J}_T^\top \nabla_y \log \pi(Z_k))^{(i)} + h(\mathbf{J}_T \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_T)^{(i)} + h\sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2} \\
&+ \sqrt{2h}(\mathbf{J}_T \xi^m)_k + hN_k^{(i)} + \mathcal{O}(h^{3/2}),
\end{aligned}
$$

where $N_k^{(i)} = \xi_{k+1}^\top \nabla_x^2 T_i \xi_{k+1} - \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2}$. Note that $N_k^{(i)}$ is a mean zero non-Gaussian random variable. (In fact, it is a sum of scaled and centered χ^2 random variables). Recall that $\xi_{k+1} \sim \mathcal{N}(0, \mathbf{I}_d)$, so $\mathbb{E}[\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}] = \delta_{jl}$ and

$$
\mathbb{E}\left[\xi_{k+1}^\top \nabla_x^2 T_i \xi_{k+1}\right] - \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2} = \sum_{j=1}^d \sum_{l=1}^d \mathbb{E}\left[\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}\right]\frac{\partial^2 T_i}{\partial x_j \partial x_l} - \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2} = 0.
$$

In the proof of Theorem 3.1, we showed $(\nabla_y \cdot \mathbf{J}_T \mathbf{J}_T^\top)^{(i)} = \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2} + \mathbf{J}_T \mathbf{J}_T^\top \nabla_y \log \det \mathbf{J}_T$, and therefore,

$$
\begin{aligned}
Z_{k+1}^{(i)} = Z_k^{(i)} + h(\mathbf{J}_T \mathbf{J}_T^\top \nabla_y \log \pi(Z_k))^{(i)} + h(\nabla_y \cdot \mathbf{J}_T \mathbf{J}_T^\top)^{(i)} + \sqrt{2h}(\mathbf{J}_T \xi_{k+1})^{(i)} \\
+ N_k^{(i)} + \mathcal{O}(h^{3/2}).
\end{aligned}
$$

After dropping the non-Gaussian and higher order terms, and substituting $\mathbf{J}_S^{-1}$ for $\mathbf{J}_T$, we arrive at the desired result.

As for the mean-squared error, note that $\mathbb{E}\left[\|N_k\|^2\right] = \sum_{i=1}^d \mathbb{E}\left[\left|N_k^{(i)}\right|^2\right]$, so we compute

$$
\begin{aligned}
\mathbb{E}\left[\left|N_k^{(i)}\right|^2\right] &= \mathbb{E}\left[\left(\xi_{k+1}^\top \nabla^2 T_i \xi_{k+1} - \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{jl} \frac{\partial^2 T_i}{\partial x_j \partial x_l}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)} - \sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2}\right)^2\right].
\end{aligned}
$$

Expanding the expression in the expectation, we have

$$
\begin{aligned}
&\left(\sum_{jl} \frac{\partial^2 T_i}{\partial x_j \partial x_l}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}\right)^2 + \left(\sum_{j=1}^d \frac{\partial^2 T_i}{\partial x_j^2}\right)^2 - 2\left(\sum_{jl} \frac{\partial^2 T_i}{\partial x_j \partial x_l}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}\right)\left(\sum_{i=1}^d \frac{\partial^2 T_i}{\partial x_i^2}\right) \\
&= \sum_{jlmn} \frac{\partial^2 T_i}{\partial x_j \partial x_l}\frac{\partial^2 T_i}{\partial x_m \partial x_n}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}\xi_{k+1}^{(m)}\xi_{k+1}^{(n)} + \sum_{jl} \frac{\partial^2 T_i}{\partial x_j^2}\frac{\partial^2 T_i}{\partial x_l^2} \\
&\quad - 2\sum_{jlm} \frac{\partial^2 T_i}{\partial x_j \partial x_l}\frac{\partial^2 T_i}{\partial x_m^2}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}.
\end{aligned}
$$

The expectation of the summands of the first term are only nonzero if pairs of indices are equal or if all the indices are equal. Using the fact that $\mathbb{E}\left[\left(\xi_{k+1}^{(j)}\right)^4\right] = 3$ and $\mathbb{E}\left[\left(\xi_{k+1}^{(j)}\right)^2\left(\xi_{k+1}^{(l)}\right)^2\right] = 1$,

$$\mathbb{E}\left[\sum_{jlmn}\frac{\partial^2 T_i}{\partial x_j \partial x_l}\frac{\partial^2 T_i}{\partial x_m \partial x_n}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}\xi_{k+1}^{(m)}\xi_{k+1}^{(n)}\right]$$

$$= 3\sum_j\left(\frac{\partial^2 T_i}{\partial x_j^2}\right)^2 + \sum_{\substack{jl\\j\neq l}}\frac{\partial^2 T_i}{\partial x_j^2}\frac{\partial^2 T_i}{\partial x_l^2} + \sum_{\substack{jl\\j\neq l}}\left(\frac{\partial^2 T_i}{\partial x_j \partial x_l}\right)^2$$

$$= \sum_j\left(\frac{\partial^2 T_i}{\partial x_j^2}\right)^2 + \sum_{jl}\left[\frac{\partial^2 T_i}{\partial x_j^2}\frac{\partial^2 T_i}{\partial x_l^2} + \left(\frac{\partial^2 T_i}{\partial x_j \partial x_l}\right)^2\right]$$

For the third term, we have $\mathbb{E}\left[-2\sum_{jlm}\frac{\partial^2 T_i}{\partial x_j \partial x_l}\frac{\partial^2 T_i}{\partial x_m^2}\xi_{k+1}^{(j)}\xi_{k+1}^{(l)}\right] = -2\sum_{j,l}\frac{\partial^2 T_i}{\partial x_j^2}\frac{\partial^2 T_i}{\partial x_l^2}$. Summing terms together, we obtain

$$\mathbb{E}\left[\left|N_k^{(i)}\right|^2\right] = \sum_{j=1}^d\left(\frac{\partial^2 T_i}{\partial x_j^2}\right)^2 + \sum_{j=1}^d\sum_{l=1}^d\left(\frac{\partial^2 T_i}{\partial x_j \partial x_l}\right)^2.$$

To obtain the desired result, we sum over the above expression for all $i = 1, \ldots d$. $\square$

Proof of Lemma 3.1. The one-dimensional statement of this lemma is given without details in Proposition 5.4.3 of [53]. Here we provide for completeness the details for the general multidimensional case. Since η satisfies $\mathrm{LSI}(\alpha)$, a change of variables leads to

$$\mathrm{Ent}_\pi[f] = \mathrm{Ent}_\eta[f \circ T] \le \frac{1}{2\alpha}\int_{\mathbb{R}^d}\frac{\|\nabla(f \circ T)(x)\|^2}{(f \circ T)(x)}\eta(x)dx$$

$$= \frac{1}{2\alpha}\int_{\mathbb{R}^d}\frac{\|\mathbf{J}_T^\top(x)\nabla f(T(x))\|^2}{f(T(x))}\pi(T(x))\det\mathbf{J}_T(x)dx$$

$$= \frac{1}{2\alpha}\int_{T^{-1}(\mathbb{R}^d)}\frac{\|\mathbf{J}_T^\top(T^{-1}(y))\nabla f(y)\|^2}{f(y)}\pi(y)\left[\det\mathbf{J}_T(T^{-1}(y))\det\mathbf{J}_{T^{-1}}(y)\right]dy.$$

Since, we shall have that $\det\mathbf{J}_T(T^{-1}(y))\det\mathbf{J}_{T^{-1}}(y) = 1$ and due to $T^{-1}(\mathbb{R}^d) = \mathbb{R}^d$:

$$\mathrm{Ent}_\pi[f] \le \frac{1}{2\alpha}\int_{\mathbb{R}^d}\frac{\|\mathbf{J}_T^\top(T^{-1}(y))\nabla f(y)\|^2}{f(y)}\pi(y)dy.$$

Notice that the Cauchy-Schwarz inequality gives

$$\|\mathbf{J}_T^\top(T^{-1}(y))\nabla f(y)\|^2 = \sum_{i=1}^d\left|\sum_{k=1}^d\frac{\partial T_k}{\partial x_i}(x)\bigg|_{x=T^{-1}(y)}\partial_{y_k}f(y)\right|^2 \tag{21}$$

$$\le \|\mathbf{J}_T^\top(T^{-1}(y))\|_2^2\|\nabla f(y)\|^2.$$

Since $\sup_{y \in \mathbb{R}^d}\|\mathbf{J}_T^\top(T^{-1}(y))\|_2^2 \le \frac{1}{\rho^2}$, we shall have

$$\mathrm{Ent}_\pi[f] \le \frac{1}{2\alpha\rho^2}\int_{\mathbb{R}^d}\frac{\|\nabla f(y)\|^2}{f(y)}\pi(y)dy$$

which means that π satisfies $\mathrm{LSI}(\alpha\rho^2)$, completing the proof of the lemma. $\square$

References

[1] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[2] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating Gaussian diffusions. *The Annals of Applied Probability*, pages 897–913, 1993.

[3] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating diffusions. *Annals of Applied Probability*, 15(2):1433–1444, 2005.

[4] Tony Lelievre, Francis Nier, and Grigorios A Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013.

[5] Luc Rey-Bellet and Konstantinos Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081, 2015.

[6] Luc Rey-Bellet and Konstantinos Spiliopoulos. Variance reduction for irreversible Langevin samplers and diffusion on graphs. *Electronic Communications in Probability*, 20, 2015.

[7] Luc Rey-Bellet and Konstantinos Spiliopoulos. Improving the convergence of reversible samplers. *Journal of Statistical Physics*, 164(3):472–494, 2016.

[8] Benjamin J Zhang, Youssef M Marzouk, and Konstantinos Spiliopoulos. Geometry-informed irreversible perturbations for accelerated convergence of Langevin dynamics. *Statistics and Computing*, 32(5):1–22, 2022.

[9] Tiangang Cui, Xin Tong, and Olivier Zahm. Optimal riemannian metric for poincar {\'e} inequalities and how to ideally precondition langevin dymanics. *arXiv preprint arXiv:2404.02554*, 2024.

[10] Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[11] Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, and Youssef Marzouk. Greedy inference with structure-exploiting lazy maps. *Advances in Neural Information Processing Systems*, 33:8330–8342, 2020.

[12] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

[13] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. *Sampling via Measure Transport: An Introduction*, pages 1–41. Springer International Publishing, Cham, 2016.

[14] Benjamin Peherstorfer and Youssef Marzouk. A transport-based multifidelity preconditioner for Markov chain Monte Carlo. *Advances in Computational Mathematics*, 45(5):2321–2348, 2019.

[15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[16] George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

[17] Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[18] Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

[19] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-Langevin diffusions. *Advances in Neural Information Processing Systems*, 33:19573–19585, 2020.

[20] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Conference on Learning Theory*, pages 3814–3841. PMLR, 2020.

[21] Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. Heavy-tailed sampling via transformed unadjusted Langevin algorithm. *arXiv preprint arXiv:2201.08349*, 2022.

[22] Leif T Johnson and Charles J Geyer. Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *The Annals of Statistics*, pages 3050–3076, 2012.

[23] Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pages 5361–5370. PMLR, 2020.

[24] Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. On the representation and learning of monotone triangular transport maps. *arXiv preprint arXiv:2009.10303v2*, 2022.

[25] Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. OT-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232, 2021.

[26] Matthew D Parno and Youssef M Marzouk. Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.

[27] Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022.

[28] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[29] Ricardo Baptista. Adaptive transport maps. https://github.com/baptistar/ATM, 2022.

[30] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PLMR, 2017.

[31] Bruno Casella, Gareth Roberts, and Osnat Stramer. Stability of partially implicit Langevin schemes and their MCMC variants. *Methodology and Computing in Applied Probability*, 13:835–854, 2011.

[32] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

[33] Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

[34] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[35] Gareth O Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[36] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[37] Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

[38] Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

[39] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[40] Samuel Livingstone and Mark Girolami. Information-geometric Markov chain Monte Carlo methods using diffusions. *Entropy*, 16(6):3074–3102, 2014.

[41] Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.

[42] Nicolas Garcia Trillos and Daniel Sanz-Alonso. The Bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56, 2020.

[43] Michela Ottobre, Natesh S Pillai, and Konstantinos Spiliopoulos. Optimal scaling of the MALA algorithm with irreversible proposals for gaussian targets. *Stochastics and Partial Differential Equations: Analysis and Computations*, pages 1–51, 2019.

[44] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[45] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

[46] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[47] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[48] Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.

[49] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[50] James Martin, Lucas C Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[51] Assyr Abdulle, Gilles Vilmart, and Konstantinos C Zygalakis. High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM Journal on Numerical Analysis*, 52(4):1600–1622, 2014.

[52] Konstantinos C Zygalakis. On the existence and the applications of modified equations for stochastic differential equations. *SIAM Journal on Scientific Computing*, 33(1):102–130, 2011.

[53] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Springer International Publishing Switzerland, 2014.

[54] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

[55] Filippo Pagani, Martin Wiegand, and Saralees Nadarajah. An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. *Scandinavian Journal of Statistics*, 49(2):657–680, 2022.

[56] Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

[57] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.