MIT Open Access Articles

Transport Map Accelerated Markov Chain Monte Carlo

Massachusetts Institute of Technology

DSpace@MIT

Transport Map Accelerated Markov Chain Monte Carlo[*]

Matthew D. Parno[†] and Youssef M. Marzouk[‡]

Abstract. We introduce a new framework for efficient sampling from complex probability distributions, using a combination of transport maps and the Metropolis–Hastings rule. The core idea is to use deterministic couplings to transform typical Metropolis proposal mechanisms (e.g., random walks, Langevin methods) into non-Gaussian proposal distributions that can more effectively explore the target density. Our approach adaptively constructs a lower triangular transport map—an approximation of the Knothe–Rosenblatt rearrangement—using information from previous Markov chain Monte Carlo (MCMC) states, via the solution of an optimization problem. This optimization problem is convex regardless of the form of the target distribution and can be solved efficiently without gradient information from the target probability distribution; the target distribution is instead represented via samples. Sequential updates enable efficient and parallelizable adaptation of the map even for large numbers of samples. We show that this approach uses inexact or truncated maps to produce an adaptive MCMC algorithm that is ergodic for the exact target distribution. Numerical demonstrations on a range of parameter inference problems show order-of-magnitude speedups over standard MCMC techniques, measured by the number of effectively independent samples produced per target density evaluation and per unit of wallclock time.

Key words. adaptive MCMC, Bayesian inference, measure transformation, Knothe–Rosenblatt rearrangement, optimal transport

AMS subject classifications. 62F15, 65C05, 65C40

DOI. 10.1137/17M1134640

1. Introduction. Markov chain Monte Carlo (MCMC) algorithms provide an enormously flexible approach for sampling from complex target probability distributions, using only evaluations of an unnormalized probability density [19, 54, 36, 9]. Within this general framework, the Metropolis–Hastings algorithm [41, 25] is one of the most broadly applicable and well-studied sampling strategies. It combines a simple proposal distribution with an accept/reject step to create the transition kernel for a Markov chain that has the desired target as its stationary distribution. Under some additional technical conditions on the proposal and on the

target density π, the Markov chain is also ergodic [56].

This paper introduces a new approach to the design of Metropolis–Hastings algorithms based on the adaptive construction of *transport maps* between the target probability distribution and a simple reference distribution. These maps are monotone and typically nonlinear transformations of the target distribution that render it easier to sample, much as a preconditioner expedites the solution of a linear system. To put our approach into context, we first recall some challenges underlying MCMC sampling and current methods for addressing them.

Effective MCMC proposal mechanisms seek to make successive iterates of the Markov chain as independent as possible. When estimating an expectation over the target distribution, efficient "mixing" in this sense reduces the variance of estimates computed from the MCMC samples. A useful intuition is that effective MCMC proposals aim to approximate the target distribution at least locally (e.g., in the case of random-walk Metropolis or Langevin proposals) or perhaps globally (e.g., in the case of Metropolis independence samplers). Consider, for example, a Gaussian proposal density centered at the current state of the chain, as in a random-walk Metropolis algorithm. The adaptive Metropolis scheme of [24] sequentially updates the covariance of this proposal in order to reflect the covariance of π. In a similar fashion, [3] uses the empirical covariance of the target to scale proposals in a Metropolis-adjusted Langevin algorithm (MALA), which also uses the gradient of π to push the proposal mean towards regions of higher target density.

Many other MCMC algorithms use local derivative information to improve sampling of the target distribution. Hamiltonian Monte Carlo methods, as in [48] and [27], propose samples via trajectories of a Hamiltonian dynamical system defined on an augmented state space. Computing these trajectories requires many evaluations of the gradient of the target density, but can produce large steps that have high acceptance probability. The stochastic Newton method of [38] uses higher-order derivative information, in the form of approximate Hessians of the local log-posterior, to scale a Gaussian proposal in high dimensions. The geometrically motivated approach of [22] also uses higher-order derivative information to define a local metric for both Langevin proposals and Hamiltonian dynamics on a Riemannian manifold. Contrasting with these schemes but also related to our work are adaptive Metropolis independence samplers [2], which construct a global approximation of the target using, for example, Gaussian mixtures. This approximation is updated recursively from past MCMC samples using a stochastic approximation scheme.

The theory of optimal transport has a rich history dating back to Monge [44], who—motivated by logistical problems involving earthworks—sought a deterministic transformation from one probability measure to another that would minimize an expected transport cost. This cost is defined by a function $c(\theta, r)$ that reflects the cost of transporting a unit of mass from θ to r. A transformation that solves the Monge problem is called an optimal transport map and induces a deterministic coupling of the two probability measures. A relaxation of the Monge problem to more general couplings was introduced by Kantorovich [30, 66], yet under certain conditions, a minimizer of the Kantorovich formulation also solves the Monge problem, i.e., is an optimal transport map. For a contemporary development of this subject, see [68, 67] and [52].

Optimal transport between discrete measures has been used for Bayesian inference in [53], where the solution of a discrete assignment problem yields a consistent ensemble transforma-

tion scheme to replace resampling in the context of a Bayesian filter. This problem differs from those considered here, however, as we focus on transport between continuous probability measures. In [47] continuous transport maps were introduced that characterize the Bayesian posterior distribution as a pushforward of the prior distribution. In this formulation, the transport map is used to generate independent samples from a distribution that in principle can be made arbitrarily close to π. However, constructing sufficiently accurate maps can be computationally taxing. The implicit sampling approach of [15, 14, 46] and the randomize-then-optimize approach of [5] compute the action of certain transport maps sample by sample, without representing the maps explicitly. But these samples do not come from π and thus require reweighting in order to represent the target. Implementing either of these approaches requires access to gradients of π.

In this paper, we will use *approximate* transport maps to achieve *exact* sampling from the target distribution by integrating transport maps with MCMC. We *reverse* the direction of the maps computed in [47] and adaptively construct our maps (now from the target to a simple reference distribution) by solving an optimization problem based on MCMC samples. We will show that the optimization problem has a remarkably simple structure: it is convex regardless of the form of the target distribution and separable across dimensions of the parameter space; it also affords substantial opportunities for parallel computation and efficient sequential updating. Moreover, computing derivatives of the optimization objective requires no derivative information from the target probability density. We will analyze the scheme from the theoretical perspective of adaptive MCMC, allowing us to establish ergodicity of the resulting chain. The transport map constructed in this way aims to represent the entire target distribution as the pullback of a Gaussian reference measure, and in that sense our approach is a global one. Unlike adaptive Metropolis independence samplers, however, we approximate the target density not by choosing from a particular family of densities, but by building an invertible transformation between the target distribution and a reference distribution. Critically, this structure enables us to use both local proposals and global/independence proposals, and to transition naturally between the two as the transport map becomes more accurate. The transport map is not tied to any particular type of MCMC proposal; it instead provides a framework for improving many standard proposal schemes.

The remainder of this paper is organized as follows. Section 2 will provide relevant background on transport maps and explain how suitable maps can be constructed from samples. Section 3 will formulate the map-based MCMC approach, while section 4 will introduce adaptive strategies. A theoretical convergence analysis is provided in section 5. Section 6 compares the performance of map-based MCMC with that of existing state-of-the-art samplers on a range of test problems.

2. Construction of transport maps. Transport maps will be used in sections 3 and 4 to define a new class of MCMC methods. This section first introduces transport maps in the context of optimal transportation (section 2.1) and then describes a practical method for constructing maps from samples (section 2.2).

2.1. Optimal transportation. Consider two Borel probability measures on $\mathbb{R}^n$, μ_θ and μ_r. We will refer to these as the *target* and *reference* measures, respectively, and associate them with random variables $\theta \sim \mu_\theta$ and $r \sim \mu_r$. A transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ is a deterministic

transformation that pushes forward μ_θ to μ_r, yielding

$$\mu_r = T_\sharp \mu_\theta. \tag{1}$$

In other words, $\mu_r(A) = \mu_\theta\left(T^{-1}(A)\right)$ for any Borel set $A \subseteq \mathbb{R}^n$. In terms of the random variables, we write $r \stackrel{d}{=} T(\theta)$, where $\stackrel{d}{=}$ denotes equality in distribution. The transport map induces a *deterministic coupling* of probability measures [68].

Of course, there can be infinitely many transport maps between two probability measures. On the other hand, it is possible that no transport map exists: consider the case where μ_θ has a point mass (an "atom") but μ_r does not. If a transport map exists, one way of regularizing the problem and finding a unique map is to introduce a cost function $c(\theta, r)$ on $\mathbb{R}^n \times \mathbb{R}^n$ that represents the work needed to move one unit of mass from θ to r. Using this cost function, the total cost of pushing μ_θ to μ_r is

$$C(T) = \int_{\mathbb{R}^n} c\left(\theta, T(\theta)\right) \, d\mu_\theta(\theta). \tag{2}$$

Minimization of this cost subject to the constraint $\mu_r = T_\sharp \mu_\theta$ is called the Monge problem, after [44]. A transport map satisfying the measure constraint (1) and minimizing the cost in (2) is an *optimal* transport map. The celebrated result of [8], later generalized by [40], shows that this map exists, is unique, and is monotone μ_θ-almost everywhere (a.e.) when μ_θ is atomless and the cost function $c(\theta, r)$ is quadratic. Generalizations of this result to other cost functions and spaces have been established in [13, 1, 18, 6].

The choice of cost function in (2) naturally influences the structure of the map. For illustration, consider the Gaussian case of $\theta \sim N(0, I)$ and $r \sim N(0, \Sigma)$ for some positive definite covariance matrix Σ. The associated transport map is linear: $T = S\theta$, where the matrix S is any square root of Σ. When the transport cost is quadratic, $c(\theta, r) = |\theta - r|^2$, S is the symmetric square root obtained from the eigendecomposition of Σ, $\Sigma = V\Lambda V^\top$, and $S = V\Lambda^{1/2}V^\top$ [49]. If the cost is instead taken to be the following weighted quadratic

$$c(\theta, r) = \sum_{i=1}^{n} t^{i-1} |\theta_i - r_i|^2, \ t > 0, \tag{3}$$

then, as $t \to 0$, the optimal map becomes lower triangular and equal to the Cholesky factor of Σ. Generalizing to non-Gaussian μ_θ and μ_r, as $t \to 0$, the optimal maps T_t obtained with the cost function (3) are shown by [11] and [7] to converge to the *Knothe–Rosenblatt* (KR) rearrangement [59, 32] between probability measures. The KR map exists and is uniquely defined if μ_θ is absolutely continuous with respect to Lebesgue measure. The KR map also has several useful properties: the Jacobian matrix of T is lower triangular and has positive diagonal entries μ_θ-a.e. Because of this triangular structure, the Jacobian determinant and the inverse of the map are easy to evaluate. This is an important computational advantage that we exploit in section 2.2.

We will employ lower triangular maps in our MCMC construction, but without directly appealing to the transport cost in (3). While this cost is meaningful for theoretical analysis and even numerical continuation schemes [11], we find that for small t, the sequence of weights

$\{t^i\}$ quickly produces numerical underflow as the parameter dimension n increases. Instead, we will directly impose the lower triangular structure and search for a map $\widetilde{T}$ that *approximately* satisfies the measure constraint, i.e., for which $\mu_r \approx \widetilde{T}_\sharp \mu_\theta$. This approach is a key difference between our construction and standard optimal transportation.

Numerical challenges with (3) are not the only reason to seek approximate maps. Suppose that the target measure μ_θ is a posterior or some other intractable distribution, but let the reference μ_r be something simpler, e.g., a Gaussian distribution with identity covariance. In this case, the complex structure of μ_θ is captured by the map T. Sampling and other tasks can then be performed with the simple reference distribution instead of the more complicated distribution. In particular, if a map exactly satisfying (1) were available, sampling the target distribution μ_θ would simply require drawing a sample $r' \sim \mu_r$ and pushing it to the target space with $\theta' = T^{-1}(r')$. This concept was employed by [47] for posterior sampling. Depending on the structure of the reference and the target, however, finding an exact map may be computationally challenging. In particular, if the target contains many nonlinear dependencies that are not present in the reference distribution, the *representation* of the map T (e.g., in some canonical basis) can become quite complex. Hence, it is desirable to work with approximations to T. Below we will demonstrate that even approximate maps can capture the key structure of the target distribution and thus be used to construct more efficient MCMC proposals.

Another reason for seeking approximate transport maps is regularity. There is an extensive theory on the regularity of optimal transport—with much that is understood, along with some open questions [10]. Since we are only concerned with approximate measure transformations, we can impose regularity conditions that may not hold for the optimal map or the KR map. In particular, we will require that $\widetilde{T}$ and its inverse have continuous derivatives on $\mathbb{R}^n$, i.e., that $\widetilde{T}$ be a C^1-diffeomorphism. Later we will impose additional constraints on the derivatives of $\widetilde{T}$, which will prove useful for our theoretical analysis of map-based MCMC.

2.2. Constructing maps from samples. As noted above, we will seek transport maps that have a lower triangular structure, i.e.,

$$(4) \qquad T(\theta_1, \theta_2, \ldots, \theta_n) = \begin{bmatrix} T_1(\theta_1) \\ T_2(\theta_1, \theta_2) \\ \vdots \\ T_n(\theta_1, \theta_2, \ldots, \theta_n) \end{bmatrix},$$

where θ_i denotes the ith component of θ and $T_i : \mathbb{R}^i \to \mathbb{R}$ is the ith component of the map T. For simplicity, we assume that both the target and reference measures are absolutely continuous on $\mathbb{R}^n$, with densities π and p, respectively. This assumption precludes the existence of atoms in μ_θ and thus makes the KR coupling well-defined. To find a useful approximation of the KR coupling, we will define a map-induced density $\tilde{\pi}(\theta)$ and minimize the distance between this map-induced density and the target density $\pi(\theta)$. The next three subsections describe the setup of this optimization problem.

Note that when the reference measure is a standard Gaussian (as we shall prescribe below), the construction of a map from target samples to the reference is a goal shared by the iterative Gaussianization scheme of [34] and the density estimation schemes of [65, 64]. Both of these approaches *compose* a series of simple maps (e.g., sigmoid-type functions of one variable in

[64]) in order to achieve the desired transformation, but can require a large number of such layers in order to converge. Also, the resulting maps are not triangular. Here, we seek to develop a more expressive all-at-once approximation of the triangular KR map.

2.2.1. Optimization objective. Let p be the probability density associated with the reference measure μ_r, and consider a transformation $\widetilde{T}(\theta)$ that is monotone and differentiable μ_θ-a.e. (In section 2.2.2 we will discuss constraints to ensure monotonicity; moreover, we will employ maps that are everywhere differentiable by construction.) Now consider the pullback of μ_r through $\widetilde{T}$. The density of this pullback measure is

$$
(5) \qquad \tilde{\pi}(\theta) = p(\widetilde{T}(\theta))|\det \nabla \widetilde{T}(\theta)|,
$$

where $\nabla T(\theta)$ is the Jacobian of the map, evaluated at θ, and $|\det \nabla \widetilde{T}(\theta)|$ is the absolute value of the Jacobian determinant. We call $\tilde{\pi}$ the *map-induced density*.

If the measure constraint $\mu_r = \widetilde{T}_\sharp \mu_\theta$ were exactly satisfied, the map-induced density $\tilde{\pi}$ would equal the target density π. This suggests finding $\widetilde{T}$ by minimizing a distance or divergence. Out of many possibilities [21], here we use the Kullback–Leibler (KL) divergence, which takes the form

$$
(6) \qquad \begin{aligned}
D_{\mathrm{KL}}(\pi \| \tilde{\pi}) &= \mathbb{E}_\pi \left[\log \left(\frac{\pi(\theta)}{\tilde{\pi}(\theta)} \right) \right] \\
&= \mathbb{E}_\pi \left[\log \pi(\theta) - \log p \left(\widetilde{T}(\theta) \right) - \log \left| \det \nabla \widetilde{T}(\theta) \right| \right].
\end{aligned}
$$

The KL divergence, in this particular direction, is a widely adopted objective (cf. expectation propagation [42] and adaptive importance sampling [17, 60]) that offers computational advantages discussed below and in section 2.4. Minimizing this KL divergence also favors approximations $\tilde{\pi}$ that "cover" the target π [37]. (Of course, other divergences might emphasize different aspects of matching between $\tilde{\pi}$ and π, and thus be better suited for certain applications.) We can now find transport maps by solving the following optimization problem:

$$
(7) \qquad \min_{T \in \mathcal{T}} \mathbb{E}_\pi \left[-\log p \left(T(\theta) \right) - \log |\det \nabla T(\theta)| \right],
$$

where $\mathcal{T}$ is some space of lower-triangular functions from $\mathbb{R}^n$ to $\mathbb{R}^n$. If $\mathcal{T}$ is large enough to include the KR map, then the solution of this optimization problem will exactly satisfy (1). Note that we have removed the $\log \pi(\theta)$ term in (6) from the optimization objective (7), as it is independent of T. If the exact coupling condition is satisfied, however, then the quantity inside the expectation in (6) becomes constant in θ. If π is unnormalized, this constant is in fact the log of the normalizing constant of π.

One benefit of using KL divergence in the direction specified above is that we can use Monte Carlo samples (in particular, MCMC samples) to approximate the expectation with respect to π. Furthermore, as we will show below, this direction allows us to dramatically simplify the solution of (7) when p is Gaussian. Suppose that we have K samples from π, denoted by $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}\}$. Taking a sample-average approximation (SAA) [31], we

replace the objective in (7) with its Monte Carlo estimate and, for this fixed set of samples, solve the corresponding deterministic optimization problem:

$$(8) \qquad \widetilde{T} = \operatorname*{argmin}_{T \in \mathcal{T}} \frac{1}{K} \sum_{k=1}^{K} \left[-\log p\left(T(\theta^{(k)}) \right) - \log \left| \det \nabla T(\theta^{(k)}) \right| \right].$$

The solution $\widetilde{T}$ is an approximation to the exact transport map for two reasons: first, we have used an approximation of the expectation operator, and second, we have restricted the feasible domain of the optimization problem to $\mathcal{T}$. The specification of $\mathcal{T}$ is the result of constraints, discussed in section 2.2.2, and of the finite-dimensional parameterization of the map, discussed in section 2.3.

2.2.2. Constraints. To write the map-induced density $\tilde{\pi}$ as in (5), it is sufficient that $\widetilde{T}$ be differentiable and monotone, i.e., $(\theta' - \theta)^{\top}(\widetilde{T}(\theta') - \widetilde{T}(\theta)) \geq 0$ for distinct points $\theta, \theta' \in \mathbb{R}^n$. Since we assume that μ_θ has no atoms, to ensure that the pushforward $\widetilde{T}_\sharp \mu_\theta$ also has no atoms we only need to require that $\widetilde{T}$ be strictly monotone. To show ergodicity of the MCMC samplers constructed in sections 3 and 4, however, we will need to impose the stricter condition that $\widetilde{T}$ be bi-Lipschitz,

$$(9) \qquad \lambda_{\min} \|\theta' - \theta\| \leq \|\widetilde{T}(\theta') - \widetilde{T}(\theta)\| \leq \lambda_{\max} \|\theta' - \theta\|$$

for some $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$. This condition implies that $\widetilde{T}$ is differentiable almost everywhere. But the maps we will employ are, by construction, everywhere differentiable and lower triangular, and hence the lower Lipschitz condition in (9) is equivalent to a lower bound on the map derivative,

$$(10) \qquad \frac{\partial \widetilde{T}_i}{\partial \theta_i} \geq \lambda_{\min}, \ i = 1, \ldots, n.$$

Since $\widetilde{T}$ is lower triangular, the Jacobian $\nabla \widetilde{T}$ is also lower triangular, and (10) ensures that the Jacobian is positive definite. Because the Jacobian determinant is then positive, we can remove the absolute value from the determinant terms in (7), (8), and related expressions. This is an important step towards arriving at a convex optimization problem (see section 2.2.3). We stress that while a nonzero $\lambda_{\min}$ is required for our theoretical analysis, it does not need to be tuned in order to apply the algorithm in practice; typically we just choose a very small value, e.g., $\lambda_{\min} = 10^{-8}$. An explicit value for $\lambda_{\max}$ can also be prescribed, but can instead be defined implicitly through the construction described next.

Many representations of $\widetilde{T}$ (e.g., polynomial expansions) will yield maps with unbounded derivatives as $\|\theta\| \to \infty$. Clearly, such maps would not satisfy the upper bound in (9). Fortunately, a simple correction ensures (9) is satisfied. Let $\widetilde{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function whose derivatives grow without bound as $\|\theta\| \to \infty$, but are finite within a ball $B(0, R)$ of radius $R < \infty$. We can satisfy (9) by setting $\widetilde{T}^R(\theta) = \widetilde{T}(\theta)$ over $B(0, R)$ and forcing $\widetilde{T}^R(\theta)$ to be linear outside of this ball. More precisely, let $w(\theta) := R \frac{\theta}{\|\theta\|}$ be the projection of θ to the closest point in $B(0, R)$, and let $d(\theta) := \frac{\theta}{\|\theta\|} \cdot \nabla \widetilde{T}(w(\theta))$ be the

directional derivative of $\widetilde{T}$ at the ball boundary. We then define $\widetilde{T}^R(\theta)$ in terms of $\widetilde{T}(\theta)$ as

$$
(11) \qquad \widetilde{T}^R(\theta) = \begin{cases} \widetilde{T}(\theta), & \|\theta\| \le R, \\ \widetilde{T}\left(w(\theta)\right) + d(\theta)(\theta - w(\theta)), & \|\theta\| > R. \end{cases}
$$

Note that a continuously differentiable $\widetilde{T}(\theta)$ will yield a continuously differentiable $\widetilde{T}^R(\theta)$. Moreover, if $\widetilde{T}(\theta)$ satisfies the lower bound in (9), $\widetilde{T}^R(\theta)$ will satisfy both the lower and upper bounds in (9).

When a finite number of samples are used in the Monte Carlo sum of (8), R can usually be chosen so that all the samples lie in $B(0, R)$, and hence $\widetilde{T}$ can be evaluated directly. In this setting, a value of R need not be explicitly prescribed. However, our asymptotic convergence theory requires finite derivatives of the map as $\|\theta\| \to \infty$ in order to achieve the correct tail behavior, which is guaranteed by using $\widetilde{T}^R$ as in (11).

Unfortunately, we cannot generally enforce the lower bound in (10) over the entire support of the target measure. A weaker, but practically enforceable, alternative is to require the map to be increasing at each sample used to approximate the KL divergence. In other words, we use the constraints

$$
(12) \qquad \left. \frac{\partial \widetilde{T}_i}{\partial \theta_i} \right|_{\theta^{(k)}} \ge \lambda_{\min} \quad \forall i \in \{1, 2, \ldots, n\}, \ \forall k \in \{1, 2, \ldots, K\}.
$$

In practice, we find that (12) is usually sufficient to ensure the monotonicity of a map represented by a finite basis expansion. When K is small, however, the pointwise constraint in (12) may not be an adequate representation of the global constraint (10), and the map may not be monotone over the entire support of the target density. When this occurs, the value of inverse map $\widetilde{T}^{-1}(r)$ may not be unique. To overcome this issue in our implementation, we choose the value that is closest to the current sample mean. In our tests, this "trick" is infrequently used and does not seem to impact convergence of the algorithm. We also mention ongoing work to develop monotone parameterizations of triangular maps [39, 61]; these parameterizations can *guarantee* global monotonicity at the expense of a slightly more challenging optimization problem.

2.2.3. Convexity and separability of the optimization problem. Now we consider the task of minimizing the objective in (8). The $1/K$ term can immediately be discarded, and the derivative constraints above let us remove the absolute value from the determinant term. While one could tackle the resulting minimization problem directly, we can simplify it further by exploiting the structure of the reference density and the triangular map.

First, we let $r \sim N(0, I)$. This choice of reference distribution yields

$$
(13) \qquad \log p(r) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} r_i^2.
$$

Next, the lower triangular Jacobian $\nabla \widetilde{T}$ simplifies the determinant term in (8) to give

$$
(14) \qquad \log \left| \det \nabla \widetilde{T}(\theta) \right| = \log \left(\det \nabla \widetilde{T}(\theta) \right) = \log \left(\prod_{i=1}^{n} \frac{\partial \widetilde{T}_i}{\partial \theta_i} \right) = \sum_{i=1}^{n} \log \frac{\partial \widetilde{T}_i}{\partial \theta_i}.
$$

The objective function in (8) now becomes

$$(15) \qquad C(\widetilde{T}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}) - \log \left. \frac{\partial \widetilde{T}_i}{\partial \theta_i} \right|_{\theta^{(k)}} \right].$$

This objective is *separable*: it is a sum of n terms, each involving a single component $\widetilde{T}_i$ of the map. The constraints in (12) are also separable; there are K constraints for each $\widetilde{T}_i$, and no constraint involves multiple components of the map. Hence the entire optimization problem separates into n individual optimization problems, one for each dimension of the parameter space. Moreover, each optimization problem is *convex*: the objective is convex and the feasible domain is closed (note the $\geq$ operator in the linear constraints (12)) and convex.

In practice, we must solve the optimization problem over some finite-dimensional space of candidate maps. Let each component of the map be written as $\widetilde{T}_i(\theta; \gamma_i)$, $i = 1, \ldots, n$, where $\gamma_i \in \mathbb{R}^{M_i}$ is a vector of parameters, e.g., coordinates in some basis. The complete map is then defined by the parameters $\bar{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_n]$. Note that there are distinct parameter vectors for each component of the map. The optimization problem over the parameters remains separable, with each of the n different subproblems given by

$$(16) \qquad \begin{aligned} \min_{\gamma_i} \quad & \sum_{k=1}^{K} \left[\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right] \\ \text{s.t.} \quad & \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \geq \lambda_{\min}, \quad k \in \{1, 2, \ldots, K\}, \end{aligned}$$

for $i = 1, \ldots, n$. All of these optimization subproblems can be solved in parallel without evaluating the target density $\pi(\theta)$. Since the map components $\widetilde{T}_i$ are linear in the coefficients γ_i, each finite-dimensional problem is still convex.

2.3. Map parameterization. In this work, we parameterize each component of the map $\widetilde{T}_i$ with a multivariate polynomial expansion. Each multivariate polynomial $\psi_{\mathbf{j}}$ is defined as

$$(17) \qquad \psi_{\mathbf{j}}(\theta) = \prod_{i=1}^{n} \varphi_{j_i}(\theta_i),$$

where $\mathbf{j} = (j_1, j_2, \ldots, j_n) \in \mathbb{N}_0^n$ is a multi-index and φ_{j_i} is a univariate polynomial of degree j_i. The univariate polynomials can be chosen from any family of orthogonal polynomials (e.g., Hermite, Legendre, Jacobi); even monomials are sufficient for the present purposes.[1] Using these multivariate polynomials, we express the map as a finite expansion of the form

$$(18) \qquad \widetilde{T}_i(\theta; \gamma_i) = \sum_{\mathbf{j} \in \mathcal{J}_i} \gamma_{i,\mathbf{j}} \, \psi_{\mathbf{j}}(\theta),$$

---

[1]In principle, there is some advantage to choosing polynomials that are orthogonal with respect to the input distribution μ_θ, as in polynomial chaos approaches [20, 35]. In the present context, however, we only have samples from μ_θ, and this distribution is almost certainly not one of the canonical distributions found in the Wiener–Askey scheme [71]. Thus μ_θ-orthogonal polynomials are not readily available, and there is little reason to be picky about the choice of polynomial basis.

where $\mathcal{J}_i$ is a set of multi-indices defining the polynomial terms in the expansion. Notice that the cardinality of the multi-index set defines the dimension of each parameter vector γ_i, i.e., $M_i = |\mathcal{J}_i|$. An appropriate choice of each multi-index set $\mathcal{J}_i$ will force the entire map $\widetilde{T}$ to be lower triangular.

One simple choice of the multi-index set corresponds to a total-order polynomial basis, where the maximum degree of each multivariate polynomial is bounded by some integer $p \geq 0$:

$$\mathcal{J}_i^{TO} = \{\mathbf{j} : \|\mathbf{j}\|_1 \leq p, \ j_k = 0 \ \forall k > i\}.$$

The first constraint in this set limits the polynomial order, while the second constraint, $j_k = 0$ for all $k > i$, applied over all $i = 1, \ldots, n$ components of the map, forces $\widetilde{T}$ to be lower triangular. A smaller multi-index set for large n can be obtained by removing all the mixed terms in the basis:

$$\mathcal{J}_i^{NM} = \{\mathbf{j} : \|\mathbf{j}\|_1 \leq p, \ j_k j_m = 0 \ \forall k \neq m, \ j_k = 0 \ \forall k > i\}.$$

An even more parsimonious option is to use diagonal maps, via the multi-index sets

$$\mathcal{J}_i^D = \{\mathbf{j} : \|\mathbf{j}\|_1 \leq p, \ j_k = 0 \ \forall k \neq i\}.$$

We will occasionally use a union of low degree $\mathcal{J}_i^{TO}$ and high degree $\mathcal{J}_i^D$ to define expressive map expansions with a tractable number of terms.

Finally, we emphasize that *any* parameterization of the map that is linear in the coefficients $\bar{\gamma}$ can be used in the optimization problems defined earlier. While the examples in this paper will focus on polynomial maps, we have also had good success representing the map as a summation of linear terms and radial basis functions [50].

2.4. Solving the map optimization problem. Since the map $\widetilde{T}_i(\theta; \gamma_i)$ is linear in the expansion coefficients γ_i, the objective in (15) can be evaluated using efficient matrix-matrix and matrix-vector operations. We first construct matrices $F_i, G_i \in \mathbb{R}^{K \times M_i}$ with components defined by $[F_i]_{k,\mathbf{j}} = \psi_{\mathbf{j}}(\theta^{(k)})$ and $[G_i]_{k,\mathbf{j}} = \frac{\partial \psi_{\mathbf{j}}}{\partial \theta_i}\big|_{\theta^{(k)}}$ for all $\mathbf{j} \in \mathcal{J}_i$. Recall that K is the number of samples in our Monte Carlo approximation of the optimization objective. Using these matrices and the expansion (18), we can rewrite (15) as

$$
\begin{aligned}
(19) \qquad & \min_{\gamma_i} \quad \frac{1}{2}\gamma_i^\top (F_i^\top F_i)\gamma_i - c^\top \log(G_i\gamma_i) \\
& \text{s.t.} \quad G_i\gamma_i \geq \lambda_{\min},
\end{aligned}
$$

where c is a K-dimensional vector of ones and the log is taken componentwise. Clearly, the objective can be evaluated with efficient numerical linear algebra routines.

Beyond efficient evaluations, the only difference between (19) and a simple quadratic program is the log term in the objective. However, the quadratic term often dominates the log term, making a standard Newton optimizer with backtracking line search quite efficient. In practice, starting with an identity map, we usually observe convergence in fewer than ten Newton iterations. Notice also that the log term in (19) acts as a barrier function for the constraints.

3. Map-based MCMC proposals. Now we will show how a transport map can be used to modify the Metropolis–Hastings algorithm by equivalently transforming either the target distribution or the proposal mechanism. In this section, we assume that a fixed transport map $\widetilde{T}$ is in hand. Of course, this map must somehow be constructed, and hence the fixed-map approach described here is just an intermediate step in our exposition. The next section (section 4) will use the optimization approaches of section 2 to iteratively build such a map in an adaptive MCMC framework.

A simple Metropolis–Hastings algorithm [25, 41] generates a new state $\theta^{(k+1)}$ from the current state $\theta^{(k)}$ in two steps. First, a sample θ' is drawn from a proposal density $q_{\theta,\bar{\gamma}}(\cdot|\theta^{(k)})$. Then an accept-reject step is performed: $\theta^{(k+1)}$ is set to θ' with probability $\alpha(\theta', \theta^{(k)})$ and to $\theta^{(k)}$ with probability $1 - \alpha(\theta', \theta^{(k)})$, where

$$(20) \qquad \alpha\left(\theta', \theta^{(k)}\right) = \min\left\{1, \frac{\pi(\theta')q_{\theta,\bar{\gamma}}(\theta^{(k)}|\theta')}{\pi(\theta^{(k)})q_{\theta,\bar{\gamma}}(\theta'|\theta^{(k)})}\right\}.$$

The choice of proposal $q_{\theta,\bar{\gamma}}$ controls the dependence between successive states in the MCMC chain through both the acceptance rate and the step size. Knowledge of the target density π is helpful in designing proposals to make large moves that simultaneously have a high acceptance probability. The scheme presented here encodes information about the target distribution via a transport map $\widetilde{T}$.

3.1. MCMC with a fixed transport map. Assume that we have an approximate transport map $\widetilde{T}$ between a standard Gaussian reference and the target measure μ_θ, i.e., $\mu_r \approx \widetilde{T}_\sharp \mu_\theta$. The pushforward of the target measure through this map will not be Gaussian. But a map that reduces the optimization objective of section 2 will make the pushforward closer (in this particular sense) to a standard Gaussian than the original target. We will then use MCMC to sample this pushforward distribution, with a proposal $q_r(r'|r)$. The proposal q_r may be chosen quite freely, and examples below will encompass both local and independence proposals. Equivalently, one can view this process from the perspective of the target space by considering the pullback through the map $\widetilde{T}$ of the proposal q_r; this map-induced proposal is applied to the original target density π. Below we will describe our algorithm from this second perspective, but the first perspective of transforming or "preconditioning" the target density may also provide useful intuition.

Let $q_r(r'|r)$ be a standard Metropolis–Hastings proposal on the reference space. The pullback of this proposal through $\widetilde{T}$ induces a target space proposal density written as

$$(21) \qquad q_{\theta,\bar{\gamma}}(\theta'|\theta) = q_r\left(\widetilde{T}(\theta')|\widetilde{T}(\theta)\right)\left|\det \nabla\widetilde{T}(\theta')\right|,$$

where $\bar{\gamma}$ denotes the dependency of this proposal on the map parameters. To perform MCMC, we need the ability to evaluate this proposal density and to draw samples from it. The expression (21) provides an easy way of evaluating the proposal density. Sampling from the proposal $q_{\theta,\bar{\gamma}}(\cdot|\theta)$ involves three steps: (1) use the current target state θ to compute the current reference state, $r = \widetilde{T}(\theta)$; (2) draw a sample $r' \sim q_r(r'|r)$ from the reference proposal; and (3) evaluate the inverse map at r' to obtain a sample from the target proposal: $\theta' = \widetilde{T}^{-1}(r')$. These steps are given as lines 4–6 of Algorithm 1 and illustrated in Figure 1. Ignoring the adaptation
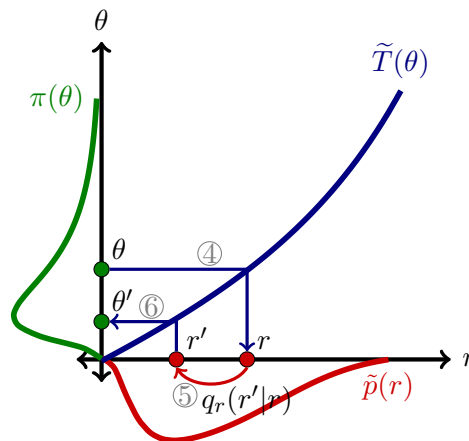
Figure 1. *Illustration of the Metropolis–Hastings proposal process in transport map–accelerated MCMC. The gray circled numbers on each arrow correspond to the line number in Algorithm* 1.

in lines 9–13, Algorithm 1 is equivalent to a standard Metropolis–Hastings algorithm on the target distribution, using $q_{\theta,\bar{\gamma}}(\theta'|\theta)$ as a proposal.

Because of the map's lower triangular structure, evaluating the inverse map $\widetilde{T}^{-1}(r)$ only requires n one-dimensional nonlinear solves. These one-dimensional problems can be tackled efficiently with a simple Newton method or, if the map is represented with polynomials, with a bisection solver based on Sturm sequences [69]. We utilize the latter approach because of its robustness.

3.2. Derivative-based proposals. An important feature of our approach is that the map-induced proposal $q_{\theta,\bar{\gamma}}(\theta'|\theta)$ requires derivative information from the target density $\pi(\theta)$ if and only if the reference proposal $q_r(r'|r)$ explicitly requires derivative information. We also note that Algorithm 1 does not require $\pi(\theta)$ to take any particular form (e.g., to be a Bayesian posterior or to result from a Gaussian prior). The ability to work with arbitrary target distributions for which derivative information may not be available is a distinction from many recent sampling approaches, such as Riemannian manifold MCMC [22], the No-U-Turn Sampler of [27], or optimization-based samplers such as implicit sampling or RTO [46, 5]. That said, though our approach can perform quite well without derivative information, we can still accommodate proposals that employ it.

The reference proposal q_r is applied to the pushforward distribution of the target π through the map $\widetilde{T}$. Let $\tilde{p}$ denote the corresponding pushforward density. Taking advantage of the map's lower triangular structure, we can write the logarithm of this density as

$$(22) \qquad \log \tilde{p}(r) = \log \pi \left(\widetilde{T}^{-1}(r) \right) + \sum_{i=1}^{n} \log \frac{\partial \widetilde{T}_i^{-1}}{\partial r_i}.$$

We will use the chain rule to obtain the gradient of this expression. First, make the substitution

$r = \widetilde{T}(\theta)$ and take the gradient with respect to θ:

$$(23) \qquad \nabla_\theta \log \tilde{p}\left(\widetilde{T}(\theta)\right) = \nabla_\theta \log \pi(\theta) - \sum_{i=1}^{n} \left(\frac{\partial \widetilde{T}_i}{\partial \theta_i}\right)^{-1} H_i(\theta),$$

where H_i is a row vector of second derivatives coming from the determinant term: $H_i(\theta) = \left[\begin{array}{cccc} \frac{\partial^2 \widetilde{T}_i}{\partial \theta_1 \partial \theta_i} & \frac{\partial^2 \widetilde{T}_i}{\partial \theta_2 \partial \theta_i} & \cdots & \frac{\partial^2 \widetilde{T}_i}{\partial \theta_n \partial \theta_i} \end{array} \right]$. Accounting for our change of variables, we now have an expression for the reference gradient given by

$$(24) \qquad \nabla_r \log \tilde{p}(r) = \left(\nabla_\theta \log \pi(\theta) - \sum_{i=1}^{n} \left(\frac{\partial \widetilde{T}_i}{\partial \theta_i}\right)^{-1} H_i(\theta)\right) \left[\nabla \widetilde{T}(\theta)\right]^{-1}.$$

Note that this expression is only valid at $\theta = \widetilde{T}^{-1}(r)$.

The lower triangular structure allows us not only to expand the determinant and obtain (24), but also to apply the inverse Jacobian $(\nabla T(\theta))^{-1}$ easily through forward substitution. Furthermore, computing the Jacobian $\nabla \widetilde{T}(\theta)$ or the second derivatives in $H_i(\theta)$ is trivial when polynomials or other standard basis functions are used to parameterize the map.

4. Adaptive transport map MCMC. Given more samples of the target distribution, we can construct a more accurate transport map, which in turn yields a more efficient map-accelerated proposal. Hence, we adaptively construct the map $\widetilde{T}$ as the MCMC chain progresses.

4.1. Adaptive algorithm overview. In our adaptive MCMC approach, we initialize the sampler with a simple map $\widetilde{T}_0$ and update the map every K_U steps using the previous states of the MCMC chain. The map update uses these samples to define the optimization problem (16), the solution of which yields a new map. This approach is conceptually similar to the adaptive Metropolis algorithm of [24]. In [24], however, previous states are used to update the covariance matrix of a Gaussian proposal; in the present case, previous states are used to construct a nonlinear transport map that yields more general non-Gaussian proposals.

The most straightforward version of our adaptive algorithm would find the coefficients γ_i for each component of the map by solving (16) directly. However, when the number of existing samples K is small or if the initial steps of the chain mix poorly, the Monte Carlo sum in (16) will be a poor approximation of the true integral, producing maps that do not capture the structure of π. This is a standard issue in adaptive MCMC, and though it does not matter asymptotically, it can impact practical performance with finite samples. One way to overcome this problem is to start adapting the map only after some initial exploration of the parameter space, i.e., after drawing a sufficient number of MCMC samples using the initial map $\widetilde{T}_0$. A more efficient alternative, however, is to introduce a regularization term $g(\gamma_i)$ into the objective, allowing the map to start adapting much earlier. The purpose of this term is to ensure that the map does not prematurely focus on one region of the target space, making it more difficult for the chain to explore the entire support of π. Regularization yields the

following modified objective:

$$
\begin{aligned}
(25) \quad & \min_{\gamma_i} \quad g(\gamma_i) + \sum_{k=1}^{K} \left[\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right] \\
& \text{s.t.} \quad \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \geq \lambda_{\min} \quad \forall k \in \{1, 2, \dots, K\}.
\end{aligned}
$$

In practice, we choose $g(\gamma_i)$ to prevent $\widetilde{T}$ from deviating too strongly from the identity map, particularly when K is small. Thus, we use a simple quadratic penalty function centered on the coefficients of the identity map: letting γ_i^{Id} denote the coefficients of the identity map, we put $g(\gamma_i) = k_R \|\gamma_i - \gamma_i^{\mathrm{Id}}\|^2$, where k_R is a user-defined regularization parameter. We have found that in most problems, small values of k_R yield similar performance. (In the numerical examples below, we mostly set $k_R = 10^{-4}$.) Because we have discarded the $1/K$ coefficient scaling the Monte Carlo sum in (25), the second term of the objective overwhelms the regularization term as the number of samples grows, and the value of k_R eventually becomes unimportant.

Other forms of regularization might also be effective. For instance, if additional problem structure such as the covariance of π were known, it could also be incorporated into the regularization term. Alternatively, one could add *upper bounds* on the map derivative to the constraints in (25); these bounds would prevent the map-induced proposal distribution $\widetilde{T}_\sharp^{-1} \mu_r$ from becoming too narrow. Finally, note that because of the Metropolis–Hastings correction, the map does *not* actually need to converge to an exact transformation. Hence one could retain a regularization term that does not decay as the number of samples increases, e.g., by reintroducing a $1/K$ prefactor to the sum in (25). The only drawback of these stronger regularizations is that they constrain the potential expressiveness of the map, even at large K, and thus might sacrifice efficiency for robustness.

Lines 9–13 of Algorithm 1 show how we incorporate the map update into our adaptive MCMC framework.

4.2. Sequential map updates. At first glance, updating the map every K_U MCMC iterations might seem computationally taxing. Fortunately, the form of the optimization problem in (25) allows for efficient updates. When K_U is small relative to the current number of steps K, the objective function in (25) changes little between updates, and the previous map coefficients provide a good initial guess for the new optimization problem. Thus new optimal coefficients can be found in only a few Newton iterations—sometimes only one or two. As the timing results in section 6 show, even for long chains (large K), the advantage of using the map to define $q_{\theta, \bar{\gamma}}$ greatly outweighs the computational costs of sequential map updates.

We also note that the optimization could be performed with stochastic approximation techniques [33, 2], in which case each map update would use only a portion of the chain, and would have a cost independent of K. Our tests with K up to 5×10^5 have shown SAA to be more efficient, but even longer chains might favor a stochastic approximation approach.

4.3. Monitoring map convergence. As the map in Algorithm 1 is adapted, the pushforward of π through the map becomes closer to the reference Gaussian, and the best choice of reference proposal $q_r(r|r')$ will evolve as well. A small-scale random-walk proposal may

---

Algorithm 1: MCMC algorithm with adaptive map.

Input: Initial state θ_0, initial vector of transport map parameters $\bar{\gamma}_0$, reference proposal $q_r(\cdot|r^{(k)})$, number of steps K_U between map adaptations, total number of steps L.

Output: MCMC samples of the target distribution, $\left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(L)}\right\}$.

1 Set state $\theta^{(1)} = \theta_0$

2 Set parameters $\bar{\gamma}^{(1)} = \bar{\gamma}_0$

3 for $k \leftarrow 1 \ldots L-1$ **do**

4   Compute the reference state, $r^{(k)} = \widetilde{T}(\theta^{(k)}; \bar{\gamma}^{(k)})$

5   Sample the reference proposal, $r' \sim q_r(\cdot|r^{(k)})$

6   Compute the target proposal sample, $\theta' = \widetilde{T}^{-1}(r'; \bar{\gamma}^{(k)})$

7   Calculate the acceptance probability:

$$\alpha = \min\left\{1, \frac{\pi(\widetilde{T}^{-1}(r'; \bar{\gamma}^{(k)}))}{\pi(\widetilde{T}^{-1}(r^{(k)}; \bar{\gamma}^{(k)}))} \frac{q_r\left(r^{(k)}|r'\right)}{q_r\left(r'|r^{(k)}\right)} \frac{\det[\nabla\widetilde{T}^{-1}(r'; \bar{\gamma}^{(k)})]}{\det[\nabla\widetilde{T}^{-1}(r^{(k)}; \bar{\gamma}^{(k)})]}\right\}$$

8   Set $\theta^{(k+1)}$ to θ' with probability α; else set $\theta^{(k+1)} = \theta^{(k)}$

9   **if** $(k \bmod K_U) = 0$ **then**

10   **for** $i \leftarrow 1$ **to** n **do**

11   Update $\gamma_i^{(k+1)}$ by solving (25) with $\left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k+1)}\right\}$

12   **else**

13 $\bar{\gamma}^{(k+1)} = \bar{\gamma}^{(k)}$

14 return *Target samples* $\left\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(L)}\right\}$

---

be appropriate at early iterations, but a larger and perhaps position-independent proposal may be advantageous as the map captures more of the target distribution's structure. By monitoring the difference between $\tilde{p}$ (22) and the uncorrelated standard Gaussian density, we can adapt the reference proposal q_r to better explore the changing $\tilde{p}$.

To this end, it is important to have an indicator of the map's current accuracy. In the discussion following (6), we noted that $\log \pi - \log p \circ \widetilde{T} - \log|\det \nabla\widetilde{T}|$ becomes a constant function of θ when an exact transformation (1) between the target and reference is achieved. A useful way to monitor the map's convergence is then to calculate the variance of this quantity,

$$(26) \qquad \sigma_M^2 = \mathbb{Var}_\theta\left[\log \pi(\theta) - \log p\left(\widetilde{T}(\theta)\right) - \log\left|\det \nabla\widetilde{T}(\theta)\right|\right].$$

A variance of zero indicates that the map is exact: $\tilde{p}$ is a standard Gaussian. Asymptotically, as $\sigma_M^2 \to 0$, the KL divergence (6) becomes $2\sigma_M^2$ [47].

4.4. Choice of reference proposal. Until now, we have left the choice of reference proposal $q_r(r'|r)$ rather open. Indeed, any nonadaptive proposal, including both independence
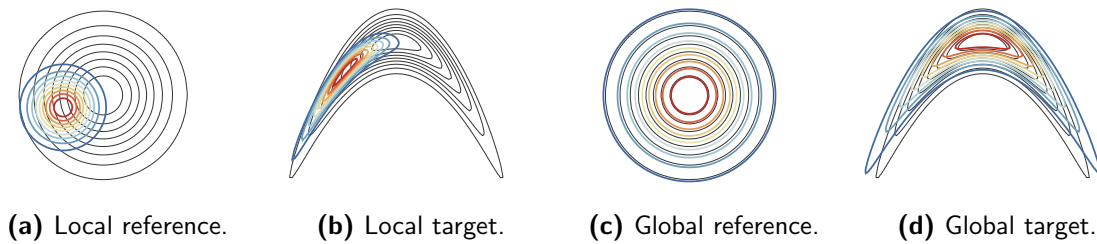
(a) Local reference.　　**(b)** Local target.　　**(c)** Global reference.　　**(d)** Global target.

Figure 2. *Example proposals in the reference space and the target space. Plots of both $q_r(r'|r)$ and $q_\theta(\theta'|r) = q_r(r'|r)|\nabla \widetilde{T}(\theta')|$ are shown for local and independence (global) proposals. The black contours depict the target distributions, while the colored contours illustrate the proposal densities.*

proposals and random-walk proposals, could be used within our framework. Figure 2 shows some typical proposals on both the reference space and the target space. In this section, we describe a few reference proposals that we will use in our numerical demonstrations, with particular attention to how they are implemented within the transport map framework. This selection is far from exhaustive and is only intended to indicate how the transport map can dictate the choice of reference proposal.

Metropolis-adjusted Langevin (MALA) proposal. Discretizing an appropriate Langevin equation yields a proposal of the form

$$(27) \qquad q_{\text{MALA}}(r'|r) = \mathcal{N}\left(r + \frac{(\Delta\tau)^2}{2}\nabla_r \log \tilde{p}(r),\, (\Delta\tau)^2\, I\right),$$

with a step size $(\Delta\tau)^2$ and a symmetric positive definite matrix Σ [57].

Delayed rejection proposals. The delayed-rejection (DR) MCMC scheme of [43] allows several proposals to be attempted during each MCMC step. With such a multistage proposal, we can try a larger or more aggressive proposal at the first stage, followed by more conservative proposals likely to produce accepted moves. We use this scheme to define $q_r(r'|r)$ in two ways.

Our first instantiation of DR employs a standard Gaussian as an independence proposal in the first stage, followed by a Gaussian random-walk proposal in the second stage. Our motivation for this global-then-local strategy is the evolving nature of $\tilde{p}(r)$. Initially, $\tilde{p}(r)$ will resemble the target density, which is more efficiently sampled by the random-walk proposal; we need samples to be accepted in order to build a good map. As the map adapts, however, $\tilde{p}(r)$ will approach a standard normal density, which can be efficiently explored by the position-independent first stage. DR naturally trades off between these alternatives. Figure 2 illustrates the difference between local and independence proposals for a simple banana-shaped distribution. Our second instantiation of DR employs two symmetric random-walk proposals, the first with a larger variance and the second with a smaller variance.

5. Convergence analysis. This section investigates conditions under which our adaptive algorithm yields an ergodic chain. Proofs of the lemmas are deferred to Appendix B.

5.1. The need for bounded derivatives. Consider a random-walk proposal on the reference space $q_r(r'|r) = N(r, \sigma^2 I)$ with some fixed variance σ^2. For illustration, assume that

the target density is a standard normal distribution: $\pi(\theta) = N(0, I)$. The random walk Metropolis (RWM) algorithm is geometrically ergodic for any density satisfying the following two conditions (see Theorem 4.3 of [28]):

$$\text{(28)} \qquad \limsup_{\|\theta\| \to \infty} \frac{\theta}{\|\theta\|} \cdot \nabla \log \pi(\theta) = -\infty$$

and

$$\text{(29)} \qquad \lim_{\|\theta\| \to \infty} \frac{\theta}{\|\theta\|} \cdot \frac{\nabla \log \pi(\theta)}{\|\nabla \log \pi(\theta)\|} < 0.$$

Densities that satisfy (28) are called super-exponentially light. It is easy to show that our example Gaussian density satisfies these conditions. In Algorithm 1, however, instead of applying the RWM proposal to π directly, we apply the RWM proposal to the map-induced density in (22). If the conditions in (11) are not satisfied, we can show that even when π is Gaussian, any monotone polynomial map with degree greater than one results in a density $\tilde{p}(r)$ that is no longer super-exponentially light. For example, let $\widetilde{T}$ have a maximum polynomial degree of $M > 1$, with M odd. Then

$$\limsup_{\|r\| \to \infty} \frac{r}{\|r\|} \cdot \nabla \log \tilde{p}(r) = \limsup_{\|r\| \to \infty} \frac{1}{\|r\|} \sum_{i=1}^{n} r_i \left(\frac{\partial \widetilde{T}_i^{-1}}{\partial r_i} \right)^{-1} \frac{\partial^2 \widetilde{T}_i^{-1}}{\partial r_i^2}$$

$$\text{(30)} \qquad\qquad\qquad = \limsup_{\|r\| \to \infty} \frac{n}{\|r\|} \left(\frac{1}{M} - 1 \right) = 0.$$

Clearly, the map-induced density is not super-exponentially light. We have therefore jeopardized the geometric ergodicity of our sampler on a simple Gaussian target. Additional restrictions on the map are needed to ensure convergence.

The loss of geometric ergodicity in (30) is due to the unbounded derivatives of nonlinear polynomial maps, which do not satisfy (9). Unbounded derivatives of $\widetilde{T}$ imply that $\widetilde{T}^{-1}$ has derivatives that approach zero as $\|r\| \to \infty$, which leads to (30). More intuitively, without an upper bound on their derivatives, polynomial maps move too much weight to the tails of $\tilde{p}$. In the next section, we show that the conditions in (9) ensure the ergodicity of Algorithm 1, even with map adaptation.

5.2. Convergence of the adaptive algorithm. Our goal in this section is to show that the adaptive Algorithm 1 produces samples that can be used in Monte Carlo approximations. We thus need to show that Algorithm 1 is ergodic for the target density $\pi(\theta)$.

Assume that the target density is finite, continuous, and super-exponentially light. (Note that certain densities which are not super-exponentially light can be transformed to super-exponentially light densities using the techniques from [29].) Also assume that the reference proposal $q_r(r'|r)$ is Gaussian with bounded mean. Furthermore, let Γ be the space of the map parameters $\bar{\gamma}$ such that $\widetilde{T}(\theta; \bar{\gamma})$ satisfies the bi-Lipschitz condition given by (9).

The map at iteration k of the MCMC chain is defined by the coefficients $\bar{\gamma}^{(k)}$. Let $P_{\bar{\gamma}^{(k)}}$ be the transition kernel of the chain at iteration k, constructed from the map $\widetilde{T}(\theta; \bar{\gamma}^{(k)})$, the

target space proposal in (21), and the Metropolis–Hastings kernel:

$$(31) \qquad P_{\bar{\gamma}^{(k)}}(\theta, \mathcal{A}) = \int_{\mathcal{A}} \left(\alpha(\theta', \theta) q_{\theta, \bar{\gamma}^{(k)}}(\theta'|\theta) + (1 - r(\theta)) \delta_{\theta}(\theta') \right) d\theta'.$$

Here $q_{\theta, \bar{\gamma}^{(k)}}$ is the map-induced proposal density from (21), $\alpha(\theta', \theta)$ is the acceptance probability defined in (20), and $r(\theta) = \int \alpha(\theta', \theta) q_{\theta, \bar{\gamma}^{(k)}}(\theta'|\theta) d\theta'$. Now, following [55] and [4], we can establish the ergodicity of our adaptive algorithm by showing that it satisfies two conditions: diminishing adaptation and containment. Diminishing adaptation is defined as follows.

Definition 5.1 (diminishing adaptation). *For any starting point $x^{(0)}$ and initial set of map parameters $\bar{\gamma}^{(0)}$, a transition kernel $P_{\bar{\gamma}^{(k)}}$ satisfies the diminishing adaptation condition when*

$$(32) \qquad \lim_{k \to \infty} \sup_{x \in \mathbb{R}^n} \left\| P_{\bar{\gamma}^{(k)}}(x, \cdot) - P_{\bar{\gamma}^{(k+1)}}(x, \cdot) \right\|_{TV} = 0 \qquad \text{in probability,}$$

where $\| \cdot \|_{TV}$ denotes the total variation norm.

Instead of working with the containment condition directly (see [4] or [55]), we will show that our adaptive MCMC algorithm instead satisfies the simultaneous strongly aperiodic geometric ergodicity condition.

Definition 5.2 (SSAGE). *Simultaneous strongly aperiodic geometric ergodicity (SSAGE) is the condition that there exist a measurable set $C \in \mathcal{B}(\mathbb{R}^D)$, a drift function $V : \mathbb{R}^n \to [1, \infty)$, and scalars $\delta > 0$, $\lambda < 1$, and $b < \infty$ such that $\sup_{x \in C} V(x) < \infty$ and the following two conditions hold:*
1. *(Minorization.) For each vector of map parameters $\bar{\gamma} \in \Gamma$, there is a probability measure $\nu_{\bar{\gamma}}(\cdot)$ defined on $C \subset \mathbb{R}^n$ with $P_{\gamma}(x, \cdot) \geq \delta \nu_{\bar{\gamma}}(\cdot)$ for all $x \in C$.*
2. *(Simultaneous drift.) $\int_{\mathbb{R}^n} V(x) P_{\bar{\gamma}}(x, dx) \leq \lambda V(x) + b I_C(x)$ for all $\bar{\gamma} \in \Gamma$ and $x \in \mathbb{R}^n$.*

By Theorem 3 of [55], SSAGE ensures the containment condition. The following three lemmas establish diminishing adaptation and SSAGE. In the following, let $C = B(0, R_C)$ be a ball of radius $R_C > 0$ and let $V(x) = k_v \pi^{-\alpha}(x)$ for some $\alpha \in (0, 1)$ and $k_v = \sup_x \pi^\alpha(x)$. Also, assume that $\pi(x) > 0$ for all $x \in C$. For this choice of $V(x)$ and our assumption that $\pi(x) > 0$ for $x \in C$, we have that $\sup_{x \in C} V(x) < \infty$.

Because the reference proposal is Gaussian with bounded mean, we can find two scalars k_1 and k_2, and two zero-mean Gaussian densities g_1 and g_2, such that the reference proposal is bounded as

$$(33) \qquad k_1 g_1(r' - r) \leq q_r(r'|r) \leq k_2 g_2(r' - r).$$

The bounds in (9) then imply that the target space proposal can also be bounded. This result is captured in Lemma 5.3.

Lemma 5.3 (bounded target space proposal). *For any map coefficients $\bar{\gamma} \in \Gamma$, the map-induced proposal $q_{\theta, \bar{\gamma}}(\theta'|\theta)$ is bounded as*

$$(34) \qquad k_L g_L(\theta' - \theta) \leq q_{\theta, \bar{\gamma}}(\theta'|\theta) \leq k_U g_U(\theta' - \theta),$$

where $k_L = k_1 \lambda_{\min}^n$, $k_U = k_2 \lambda_{\max}^n$, $g_L(x) = g_1(\lambda_{\max} x)$, and $g_U(x) = g_2(\lambda_{\min} x)$.

The upper and lower bounds in (34) are key to our proof of convergence. In fact, with these bounds, the proofs of Lemmas 5.5 and 5.6 below closely follow the proof of Proposition 2.1 in [3]. Again, proofs of these results are left to the appendix.

Lemma 5.4 (diminishing adaptation of Algorithm 1). *Let the map parameters* $\bar{\gamma}$ *be restricted to a compact subset of* Γ. *Then the sequence of transition kernels defined by the update step in lines* 9–13 *of Algorithm* 1 *satisfies the diminishing adaptation condition.*

Lemma 5.5 (minorization condition for Algorithm 1). *There is a scalar* δ *and a set of probability measures* $\nu_{\bar{\gamma}}$ *defined on* C *such that* $P_{\bar{\gamma}}(x, \cdot) \geq \delta \nu_{\bar{\gamma}}(\cdot)$ *for all* $x \in C$ *and* $\bar{\gamma} \in \Gamma$.

Lemma 5.6 (drift condition for Algorithm 1). *For all points* $x \in \mathbb{R}^n$ *and all feasible map parameters* $\bar{\gamma} \in \Gamma$, *there are scalars* λ *and* b *such that* $\int_{\mathbb{R}^n} V(x) P_{\bar{\gamma}}(x, dx) \leq \lambda V(x) + b I_C(x)$.

With Lemmas 5.4–5.6 in hand, Theorem 5.7 finally yields the ergodicity of our adaptive algorithm.

Theorem 5.7 (ergodicity of Algorithm 1). *Algorithm* 1 *is ergodic for the target distribution* $\pi(\theta)$ *when* $\bar{\gamma}$ *is constrained to a compact set within which* $\widetilde{T}(\theta; \bar{\gamma})$ *is guaranteed to satisfy* (9) *for all* $\theta \in \mathbb{R}^n$.

Proof. Lemmas 5.5 and 5.6 ensure that SSAGE is satisfied, which subsequently ensures containment. The diminishing adaptation property from Lemma 5.4 combined with SSAGE implies ergodicity by Theorem 3 of [55]. ∎

6. Numerical examples. Here we compare the performance of Algorithm 1 with that of several existing MCMC methods, including delayed rejection adaptive Metropolis (DRAM) [23], simplified manifold MALA (sMMALA) [22], adaptive MALA (AMALA) [3], and the No-U-Turn Sampler (NUTS) [27]. For a full comparison, we will pair transport maps with several different reference proposal mechanisms: a random walk (TM+RW), both varieties of delayed rejection discussed in section 4.4 (denoted by TM+DRG for the global/independence proposal and TM+DRL for local proposals), and a MALA proposal (TM+LA). To explore the strengths and weaknesses of each algorithm, we consider three test problems that provide a range of target distributions.

Throughout our results, the minimum effective sample size (ESS) over all parameter dimensions is used to evaluate MCMC performance. We run *multiple* independent chains for each sampler, extract the median integrated autocorrelation time for each dimension, then take the worst case over dimensions; details on this ESS evaluation are provided in Appendix A. Larger effective sample sizes correspond to smaller variances of estimates computed from MCMC samples. To illustrate the computational cost of each method, we also report the ESS normalized by run time and by the number of function evaluations. Posterior density evaluations and gradient evaluations are summed when normalizing by "function evaluation."

6.1. Biochemical oxygen demand model. In water quality monitoring, the simple biochemical oxygen demand (BOD) model given by $B(t) = \theta_0(1 - \exp(-\theta_1 t))$ is often fit to observations of $B(t)$ at early times (e.g., $t < 5$) [63]. In this example, we wish to infer θ_0 and θ_1 given N observations at times $\{t_1, t_2, \ldots, t_N\}$. We use 20 observations evenly spread over $[1, 5]$, with additive Gaussian errors, $y(t_i) = \theta_0(1 - \exp(-\theta_1 t_i)) + e$, where $e \sim N(0, \sigma_B^2)$ and $\sigma_B^2 = 2 \times 10^{-4}$.

Table 1

Performance of MCMC samplers on the BOD problem. τ_{max} is the maximum integrated autocorrelation time, where the maximum is taken over all dimensions; ESS is the corresponding minimum effective sample size. Results are averaged over multiple independent runs of each sampler, and σ_τ is the empirical standard deviation of τ_{max} over these runs.

Method	$\tau_{\max}$	σ_τ	ESS	ESS/sec	ESS/eval	Rel. ESS/sec	Rel. ESS/eval
DRAM	59.2	24.6	551	1.04e-01	4.23e-03	1.00	1.00
NUTS	14.7	1.0	2214	4.97e-02	1.20e-03	0.48	0.28
sMMALA	84.4	14.4	385	1.05e-03	2.57e-03	0.01	0.61
AMALA	42.1	11.7	771	1.46e-01	5.14e-03	1.40	1.22
TM+DRG	2.1	0.7	15660	1.44e+00	1.61e-01	13.85	38.06
TM+DRL	4.5	0.5	7174	6.13e-01	5.90e-02	5.89	13.95
TM+RWM	5.0	0.2	6558	7.98e-01	8.73e-02	7.67	20.64
TM+LA	854.9	340.3	38	2.94e-03	2.53e-04	0.03	0.06

Our synthetic data come from evaluating $B(t_i)$ with $\theta_0 = 1$ and $\theta_1 = 0.1$ and sampling e. Using a uniform improper prior over $\mathbb{R}^2$, we have the target posterior given by

$$(35) \qquad \log \pi(\theta_0, \theta_1) = -2\pi\sigma_B^2 - \frac{1}{2}\sum_{i=1}^{2}\left[\theta_0(1 - \exp(-\theta_1 t_i)) - y(t_i)\right]^2.$$

It is easy to obtain gradients of the posterior density, allowing us to again compare many different MCMC algorithms. For each algorithm, we run 30 independent chains starting at the posterior mode, which is computed with an LBFGS optimization algorithm. Each chain is run for 7.5×10^4 iterations, with the first 1×10^4 iterations discarded as burn-in. Results are shown in Table 1.

In this example, we represent the map with total-order Hermite polynomials of degree three. The additional nonlinear terms help capture the changing posterior correlation structure shown in Figure 3(a), which is challenging for standard samplers to explore. Methods like DRAM and AMALA may capture the global covariance, but this covariance is often not representative of the local structure and does not provide enough information for efficient posterior sampling. Other methods, like sMMALA and NUTS, use derivative information to capture local geometry, but the local geometry varies considerably and is not sufficiently representative of the global structure, making it difficult for these samples to take large jumps through the parameter space. Our transport map proposals, on the other hand, are capable of capturing the global non-Gaussian structure of Figure 3(a); in fact, the pushforward of this target density through the map becomes much more Gaussian, as shown in Figure 3(b). Map-based methods with global independence proposals (e.g., TM+DRG) can then efficiently "jump" across the entire parameter space, yielding the much shorter integrated autocorrelation times shown in Table 1.

Another interesting result in Table 1 is the poor performance of TM+LA. In this example, the basic MALA algorithm was not able to sufficiently explore the space on its own (or, equivalently, with an initial identity map); hence, poor exploration in the early stages of Algorithm 1 hindered good adaptation and resulted in the inefficient sampling shown here. Because of this poor performance, the TM+LA algorithm will not be employed in our other
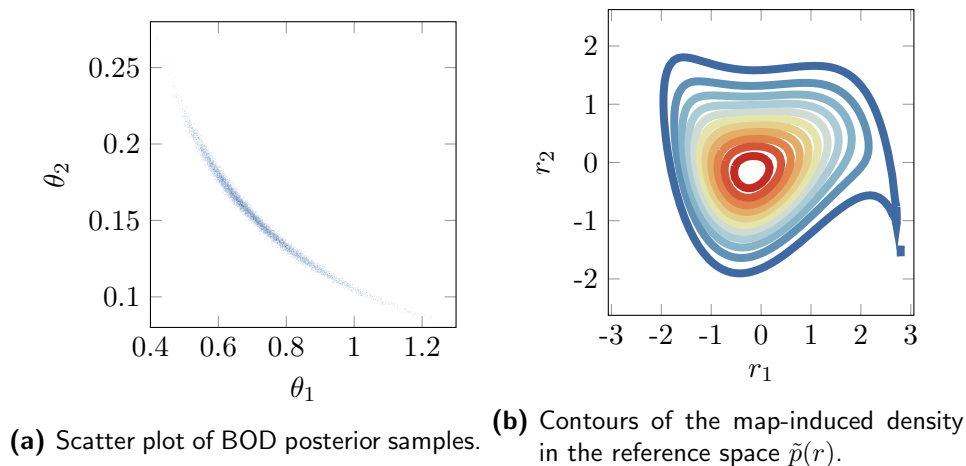
(a) Scatter plot of BOD posterior samples.

(b) Contours of the map-induced density in the reference space $\tilde{p}(r)$.

Figure 3. *The narrow high-density region and changing correlation structure of the target distribution on the left is difficult for many samplers. The transport map approach, after adaptation, pushes forward the original target to the distribution shown on the right, which can be sampled much more effectively.*

test problems.

6.2. Predator-prey system. The previous example has a posterior density whose derivatives are easy to evaluate in closed form. However, many realistic inference problems involve complex likelihoods for which derivative information is expensive to compute. This example illustrates such a situation; we consider parameter inference in an ODE model of a predator-prey system,

$$
\begin{aligned}
\frac{dP}{dt} &= rP\left(1 - \frac{P}{K}\right) - s\frac{PQ}{a+P}, \\
\frac{dQ}{dt} &= u\frac{PQ}{a+P} - vQ,
\end{aligned}
$$

(36)

where (P, Q) are the prey and predator populations and r, K, s, a, u, and v are model parameters. See [58] for model details and the ecological meaning of these parameters. In addition to these six parameters, we infer the initial conditions $P(0)$ and $Q(0)$ from five noisy observations of both P and Q at times regularly spaced on $[0, 50]$. The observations are perturbed with independent Gaussian observational errors with mean zero and variance 10. We generate the data using the following "true" parameter values:

$$
(37) \qquad [P^*(0), Q^*(0), r^*, K^*, s^*, a^*, u^*, v^*]^T = [50, 5, 0.6, 100, 1.2, 25, 0.5, 0.3]^T.
$$

The MCMC chain is run on a set of parameters θ that are scaled by these true parameters.

The prior for this problem is uniform over the intersection of a hypercube in parameter space, $[0.001, 50]^8$, and the set of parameters that produce cyclic solutions. The cyclic solution requirement can be enforced by examining the Jacobian of (36) at its fixed points. A fixed point, denoted by $[P_f, Q_f]$, must satisfy $P_f > 0$ and $Q_f > 0$, and the Jacobian on the right-hand side of (36) must have eigenvalues with positive real components when evaluated at $[P_f, Q_f]$ [62].
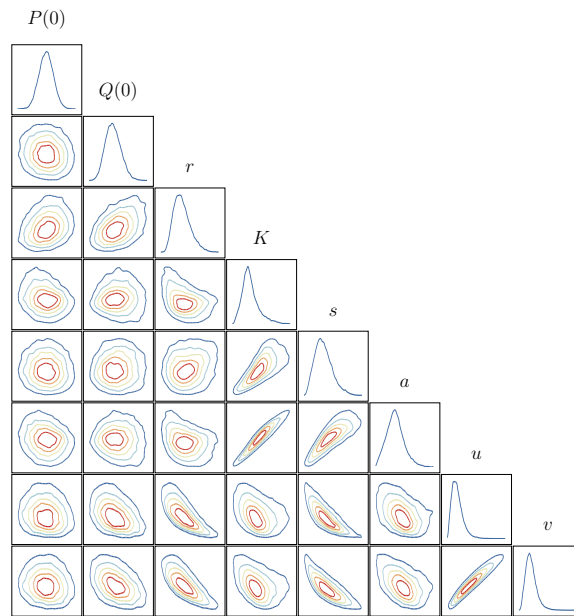
© 2018 U.S. Government

Figure 4. *Posterior distribution for the predator-prey inference example. The lack of sharp abrupt edges in the posterior indicates that the posterior density is significantly different from the uniform prior.*
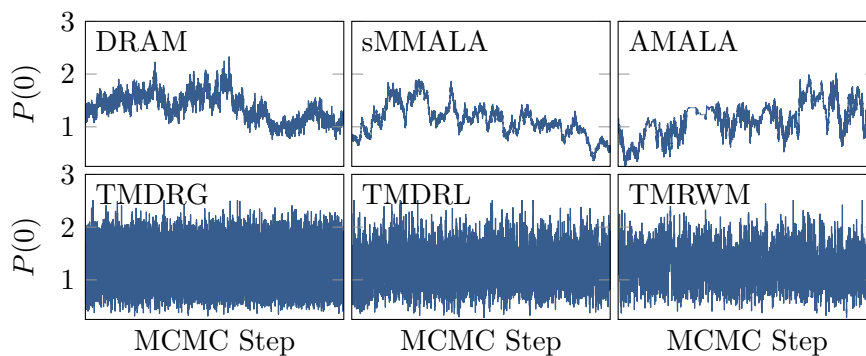

Figure 5. *Trace of MCMC chains for the parameter $P(0)$ on the predator-prey problem. These plots show the 5×10^4 steps occurring just after 2×10^5 burn-in steps, for a realization of the long-chain cases. The map-accelerated approaches show significantly better mixing.*

The posterior distribution of the parameters is shown in Figure 4. While not as narrow as the BOD posterior, this target distribution is non-Gaussian and its various marginals have changing local correlation structures. Figure 5 shows trace plots for each algorithm, while Table 2 shows a performance comparison of the samplers. Results are computed for two different chain lengths: chains of 1.2×10^5 steps, with the first 5×10^4 steps discarded as burn-in; and longer chains with 5×10^5 total steps, discarding the first 2×10^5 as burn-in. The longer chains are intended as a check to validate the performance conclusions drawn from shorter chains in the other examples. The transport map algorithms used multivariate Hermite polynomials of total degree three.

© 2018 U.S. Government

Table 2

Performance of MCMC samplers on the predator-prey parameter inference problem. Column headings are as described in Table 1. The "long" results use a chain of 5×10^5 total steps, while the "short" results use chains of length 5×10^4. The long and short chains were generated on different platforms, so the timing results should not be compared directly. Also, because the chains are different lengths, the raw ESS values should also not be compared directly. The relative results are normalized by DRAM-Short values for short chains and by DRAM-Long values for long chains.

Method	Chain length	$\tau_{\max}$	σ_τ	ESS	ESS/sec	ESS/eval	Rel. ESS/sec	Rel. ESS/eval
DRAM	5e4	4131.5	2613.7	8	7.0e-06	1.7e-04	1.0	1.0
	5e5	6673.8	5950.5	22	1.6e-05	2.7e-05	1.0	1.0
sMMALA	5e4	1913.4	521.8	18	2.9e-06	3.7e-04	0.43	2.2
	5e5	6365.7	3508.8	23	4.5e-06	2.2e-05	0.28	0.81
AMALA	5e4	1244.6	858.3	26	3.9e-06	5.4e-04	0.56	3.2
	5e5	4323.8	3611.8	34	5.9e-06	2.6e-05	0.37	0.95
TM+DRG	5e4	27.3	26.3	1280	1.4e-04	2.6e-02	20	150
	5e5	18.0	19.5	8344	9.3e-04	1.2e-02	59	420
TM+DRL	5e4	32.8	16.7	1067	1.2e-04	2.1e-02	17	130
	5e5	24.7	7.5	6081	6.7e-04	8.3e-03	42	300
TM+RWM	5e4	42.9	21.3	790	9.2e-05	1.6e-02	13	93
	5e5	32.7	15.6	4585	5.4e-04	1.1e-02	34	390

For the shorter chains, each algorithm was started at the posterior mode, and 30 independent runs of each sampler were used to generate the results. The longer chains were started with random initial points taken from the prior, and 100 independent runs of each sampler were performed. All derivative information was computed by solving the forward sensitivity equations corresponding to (36). Even though we would expect NUTS to have a large effective sample size on this problem, NUTS was not included here because of the intractable number of gradient evaluations it required. Our initial tests indicated that roughly 40 days would be required to run our full numerical comparison with NUTS.

As in the BOD example, map-accelerated algorithms using independence proposals have dramatically shorter integrated autocorrelation times. For the longer chains, TM+DRG yields an ESS about 380 times larger than that of DRAM. Moreover, in terms of ESS per posterior evaluation, TM+DRG is 420 times more efficient than DRAM. We also observe good agreement between the longer-chain and shorter-chain results; trends are the same in both cases. Overall, the gradient-based methods showed relatively poor performance. sMMALA in particular suffers from nearly singular metrics. We found that tuning the step size in sMMALA was difficult. On the other hand, the derivative-free methods were easier to tune and had much better performance. Even when normalized by run time, the ESS/sec of TM+DRG is still more than one order of magnitude larger than that of DRAM. While posterior evaluations in this example are not trivially cheap, the ESS/evaluation represents the limiting behavior of the algorithm as evaluations become the dominant cost of an MCMC step; here we see improvements of at least two orders of magnitude over the baseline schemes.

Results for the longer chains are generally more favorable for the transport map approaches. We believe this is caused by two factors: first, the burn-in is smaller in relative terms for the longer chains, which reduces wasted computational effort; second, the adaptive

proposals have more time to accurately characterize the posterior. In longer trace plots, we observe that the adaptation is negligible after approximately 1×10^5 steps, which suggests that the different burn-in lengths dominate the difference between long and short chains.

6.3. Maple sap exudation. This section presents an inference problem based on the system of differential-algebraic equations introduced in [12] to describe microscale sap dynamics in a maple tree during spring freeze-thaw cycles. The posterior in this 10-dimensional problem is particularly challenging to explore, and helps illustrate aspects of the map adaptation process. The nonlinear forward model has three state variables describing the positions of gas, liquid, and ice interfaces ($s_{gi}(t)$, $s_{iw}(t)$, and $r(t)$) as well as a state variable $U(t)$ representing the volume of melted ice. These variables are related via the following differential-algebraic equations:

$$2\rho_i s_{gi}(t) \dot{s}_{gi}(t) = \frac{\rho_w}{\pi L^f} \dot{U}(t) - 2(\rho_w - \rho_i) s_{iw}(t) \dot{s}_{iw}(t), \tag{38}$$

$$\lambda \rho_w \dot{s}_{iw}(t) = -\kappa(x) \partial_x T(x, t) \quad \text{at} \quad x = s_{iw}(t), \tag{39}$$

$$N \dot{U}(t) = -\frac{KA}{\rho_w g W} \left[p_w^v(t) - p_g^f(t) - RT(R^f, t) c_s^v \right], \tag{40}$$

$$r(t) \dot{r}(t) = -\frac{N \dot{U}(t)}{2\pi L^v}. \tag{41}$$

In addition to the state equations, the model is closed with five algebraic relations:

$$p_g^f(t) = p_g^f(0) \left(\frac{s_{gi}(0)}{s_{gi}(t)} \right)^2, \tag{42}$$

$$p_w^v(t) = p_g^v(x, t) + \frac{\sigma}{r(t)} \qquad\qquad\qquad \text{at } x = R^f + R^v - r, \tag{43}$$

$$p_g^v(t) = \frac{\rho_g^v(x, t) R T_g^v(x, t)}{M_g} \qquad\qquad \text{at } x = R^f + R^v - r, \tag{44}$$

$$c_g^v(t) = \frac{H}{M_g} \rho_g^v(x, t) \qquad\qquad\qquad \text{at } x = R^f + R^v - r, \tag{45}$$

$$\rho_g^v(x, t) = \frac{\rho_g^v(x, 0) V_g^v(0) - M_g c_g^v(\tilde{t}) \left(V^v - V_g^v(\tilde{t}) \right)\big|_{\tilde{t}=0}^{t}}{V_g^v(t)} \qquad \text{at } x = R^f + R^v - r. \tag{46}$$

In this system, $T(x, t)$ is a transient temperature field, $[\rho_i, \rho_w, \lambda, R, g, \sigma, H, M_g]$ are physical constants, and the parameters $[V^v, V_g^v, N, K, A, W, L^f, L^v, c_s^v]$ are inference targets. The initial conditions $s_{gi}(0)$, $s_{iw}(0)$, and $r(0)$ are also inference targets. For additional details on the model and its parameters, see Appendix C.

We describe the model parameters with a random variable θ taking values in $\mathbb{R}^{10}$. As detailed in Appendix C, the components of θ are scaled and combined to obtain the model parameters and initial conditions. We choose θ so that each component of the prior $\pi(\theta)$ is independent. In particular, the prior is given by $\theta_{1:3} \sim U[-1, 1]^3$ and $\theta_{4:10} \sim N(0, I)$. Noisy observations of $p_w^v(t)$ at 100 times equally spaced over $t \in [0, 1209600]$ are combined with an independent additive Gaussian error model to define the likelihood function $\pi(\theta|d)$. The
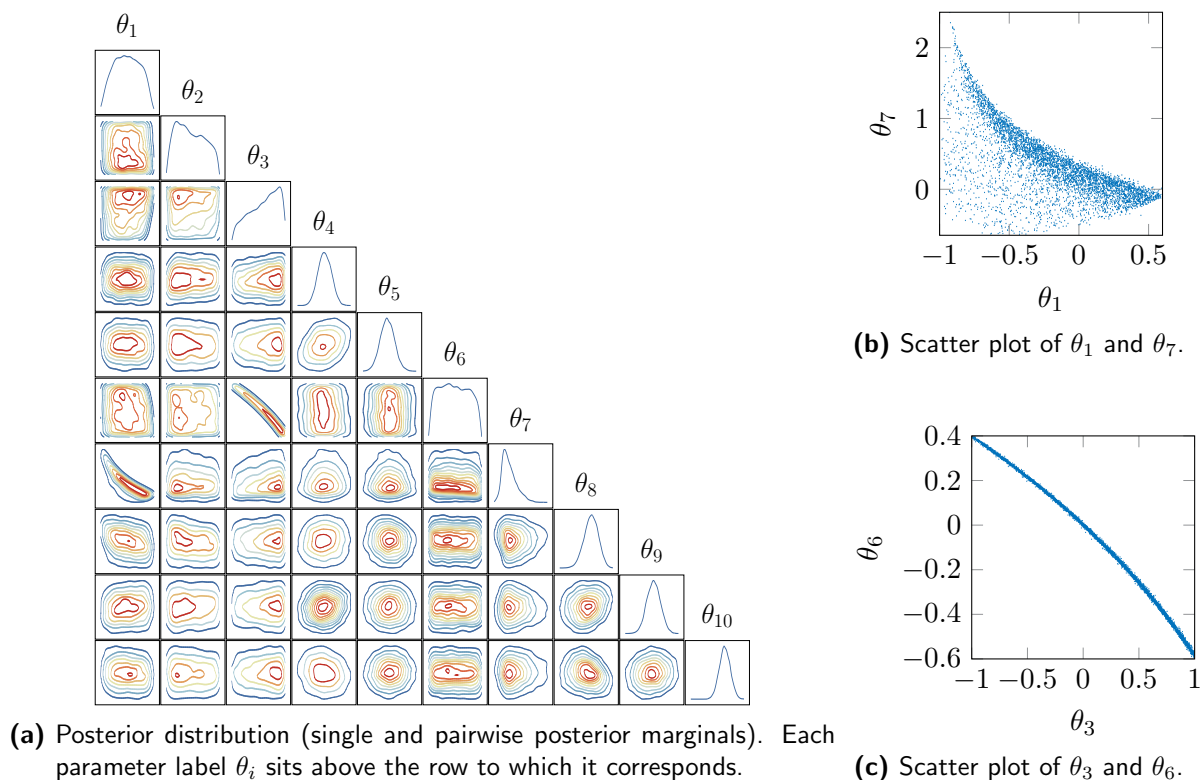
(a) Posterior distribution (single and pairwise posterior marginals). Each parameter label θ_i sits above the row to which it corresponds.

(b) Scatter plot of θ_1 and θ_7.

(c) Scatter plot of θ_3 and θ_6.

Figure 6. *Posterior distribution of the maple sap exudation example. The kernel density estimates in Figure* 6(a) *misrepresent some sharp edges and narrow regions of the posterior, as illustrated in the scatter plots of Figures* 6(b) *and* 6(c)*.*

additive errors are identically distributed with zero mean and a standard deviation of 1000 pascals.

The posterior distribution is illustrated in Figure 6, and the performance of several algorithms is summarized in Table 3. We restricted this study to derivative-free MCMC samplers, due to the complexity of computing derivative information with the maple forward model. To obtain our performance results, we ran MCMC chains of 2×10^5 steps each, discarding the first 1×10^5 samples as burn-in. As before, the ESS values reported in Table 3 represent a minimum over all ten components of each chain, calculated after burn-in. Yet the number of evaluations and the run times reported in the table reflect the cost of all 2×10^5 steps, *including* burn-in. Hence these are conservative numbers that include the computational effort required for adaptation. 50 repetitions of each sampler were used to obtain these performance evaluations. We again used cubic total-degree polynomial maps.

Many of the two-dimensional marginal plots in Figure 6 are close to Gaussian; however, the complicated relationships between (θ_1, θ_7) and between (θ_3, θ_6) yield a difficult posterior for MCMC methods. The very tight and curved joint distribution shown in Figure 6(c) is particularly challenging to capture and sample. At the early stages of adaptation, both DRAM and the transport map proposals are nearly isotropic and require very small steps

Table 3

Performance of MCMC samplers on the maple parameter inference problem. Column headings are as described in Table 1.

Method	$\tau_{\max}$	σ_τ	ESS	ESS/sec	ESS/eval	Rel. ESS/sec	Rel. ESS/eval
DRAM	2571.4	1410.0	19	2.2e-06	5.6e-05	1.0	1.0
TM+DRG	1144.2	494.8	43	4.4e-06	1.2e-04	2.0	2.1
TM+DRL	460.1	170.0	108	1.2e-05	3.3e-04	5.4	5.9
TM+RWM	1129.7	775.9	44	8.0e-06	8.9e-04	3.7	15.8

to have a nonzero acceptance rate. As the methods adapt, however, the proposals begin to capture the strong correlation between θ_3 and θ_6 and larger steps can be employed. The nonlinear dependencies are much better captured by the transport map proposals, resulting in the order-of-magnitude performance gains shown in Table 3.

In contrast with the previous two examples, the TM+DRG method is not the top performer in this comparison. The previous examples had simpler target distributions where the transport map could capture nearly all of the problem structure, allowing the independence proposal in TM+DRG to efficiently explore the parameter space. The maple model's posterior, however, is much more challenging and cannot be entirely characterized with a cubic map; thus, the global proposals are less effective. In this example, TM+DRL is the best-performing variant of the algorithm because it uses only local proposals and is not as sensitive to map deficiencies.

With challenging target distributions like this one, small initial proposal steps are needed to begin sampling. However, small initial steps do not adequately explore the parameter space, yielding an inaccurate finite-sample approximation to the KL divergence in (25). Without the regularization term in (25), one may then obtain transport maps that place too much probability mass on the relatively small region explored by the initial chain. A sufficiently large regularization term prevents this, but can also result in a slower adaptation process. We started adapting the map after 5×10^3 steps of the chain and found that $k_R = 2 \times 10^{-5}$ was sufficiently large to ensure the proposal did not become too small when the starting isotropic random-walk proposal was tuned to have a 1% acceptance rate. However, when the initial proposal was shrunk to obtain a 30% acceptance rate, we needed a much larger value of $k_R \approx 1 \times 10^{-2}$.

7. Conclusions. We have introduced a new MCMC approach that uses transport maps to accelerate sampling from challenging target distributions. Our approach adaptively constructs nonlinear transport maps from MCMC samples, via the solution of a convex and separable optimization problem. From one perspective, the resulting maps transform the target to a reference distribution that is increasingly Gaussian and isotropic, and hence easier to sample. From a complementary perspective, the maps transform simple proposal mechanisms into non-Gaussian proposals on the target. Our maps are by construction invertible and continuously differentiable functions between the reference and target spaces, and hence they allow broad flexibility in choosing reference-space MCMC proposals. Yet building the maps themselves requires no derivative information from the target distribution.

The efficiency of our approach is primarily a result of capturing nonlinear dependencies

and non-Gaussian structure in the posterior and, when possible, exploiting this knowledge with global independence proposals (e.g., TM+DRG). Of course, sequentially updating the transport map introduces an additional computational cost, which may become important in simple problems. As shown in the BOD example, however, our methods can be more efficient on strongly non-Gaussian problems, even when the target density is trivial to evaluate. On more complex posteriors, as in the ODE and DAE examples of sections 6.2 and 6.3, the efficiency gains can be even more significant, both in terms of effective sample size per posterior evaluation and effective sample size per unit of wallclock time. It is also important to point out that our current implementation does not exploit the many levels of parallelism afforded by the map construction algorithm: solution of the optimization problem (25) can be made embarrassingly parallel over parameter dimensions, and additional parallelism can be introduced over samples.[2]

While the present work used polynomials to represent the transport map, this is not an essential aspect of the framework. In fact, the optimization problem for the map coefficients in (25) will be unchanged for any map representation that is linear in the coefficients; we have experimented with other bases, e.g., radial basis functions, to good effect. Moreover, both polynomials and radial basis functions could be embedded in the *monotone parameterizations* recently proposed in [39, 61]; adopting these parameterizations may improve numerical robustness, particularly in the small-sample regime. Extending the transport map approach to higher-dimensional problems may also require a more parsimonious choice of basis (versus the total-order bases used here). Recent results on the sparsity of triangular transports [61, 45] may be useful in this regard. The map regularization term can also affect performance, especially in early iterations of MCMC with challenging targets. Alternative forms of regularization, e.g., additional constraints on the gradients or Jacobian determinant of the map, or entropic regularization of optimal transport as in [16], could also be investigated. We also note that the transport map defines a Riemannian metric on the parameter space, locally given by $(\nabla \widetilde{T}(\theta))^{\top}(\nabla \widetilde{T}(\theta))$. This suggests links between map-accelerated sampling and differential geometric MCMC methods, which we plan to explore.

Appendix A. ESS calculation details. Here we describe the calculation of the maximum integrated autocorrelation time $\tau_{\max}$ used throughout our results. Assume we are given M independent MCMC chains on an n-dimensional parameter space. Then let $\tau_{i,j}$ be the integrated autocorrelation time of dimension j on chain i. This value is computed by applying the Fourier transform method from [70] to each dimension of each chain independently. We then define $\tau_{\max}$ as

$$\tag{47} \tau_{\max} = \max_{j\in\{1,\dots,n\}} \left[\operatorname*{median}_{i\in\{1,\dots,M\}} (\tau_{i,j}) \right],$$

where the median is taken over the chains and the maximum (worst case) is taken over dimensions.

Effective sample size (ESS) is calculated similarly. Let $\mathrm{ESS}_{i,j} = \frac{K}{2\tau_{i,j}}$, where K is the

---

[2]Our implementation is freely available in MUQ [51]. This work used commit 7417f35 from MUQ's Git repository.

number of post-burn-in samples in each chain. The reported ESS is then given by

$$\text{(48)} \qquad \text{ESS} = \min_{j \in \{1,\dots,n\}} \left[\underset{i \in \{1,\dots,M\}}{\text{median}} \left(\frac{K}{2\tau_{i,j}} \right) \right].$$

Note that while ESS uses only the samples produced after the burn-in period, normalized values of ESS reported in section 6 (e.g., ESS per function evaluation and ESS per second of wallclock time) use *all* function evaluations or computational time in evaluating the denominator. Thus the cost of burn-in is reflected in these normalized performance metrics.

Appendix B. Proof of ergodicity. Section 5 of the paper provides an overview of the convergence properties of our map-accelerated MCMC algorithm. In this appendix, we include some of the associated technical analysis. In particular, we provide detailed proofs of Lemmas 5.3 and 5.4. The remaining results needed for Theorem 5.7 are direct extensions of the proof of Lemma 6.1 in [3].

B.1. Bounded target proposal. The goal of this section is to prove Lemma 5.3 by finding two zero-mean Gaussian densities that bound the map-induced target space proposal density $q_{\theta,\bar{\gamma}}$. We assume throughout this appendix that the target density $\pi(\theta)$ is finite, continuous, and super-exponentially light. (See (28) for the definition of super-exponentially light.) We also assume that the reference proposal density $q_r(r'|r)$ is a Gaussian random walk with a location-dependent bounded drift term $m(r)$ and fixed covariance Σ. Such a proposal takes the form

$$\text{(49)} \qquad q_r(r'|r) = N(r + m(r), \Sigma).$$

Given this proposal density, we can follow [3] and show that there exist two zero-mean Gaussian densities g_1 and g_2, as well as two scalars k_1 and k_2, such that $0 < k_1 < k_2 < \infty$ and

$$\text{(50)} \qquad k_1 g_1(r' - r) \leq q_r(r'|r) \leq k_2 g_2(r' - r).$$

Now, we will use the bi-Lipschitz condition in (9) to bound the target space proposal $q_{\theta,\bar{\gamma}}$ as required by Lemma 5.3.

Proof of Lemma 5.3. The following steps yield an upper bound:

$$
\begin{aligned}
q_{\theta,\bar{\gamma}}(\theta'|\theta) &= q_r(\widetilde{T}(\theta')|\widetilde{T}(\theta))|\det \nabla \widetilde{T}(\theta')| \\
&\leq q_r(\widetilde{T}(\theta')|\widetilde{T}(\theta))\lambda_{\max}^n \\
&\leq k_2 g_2(\widetilde{T}(\theta') - \widetilde{T}(\theta))\lambda_{\max}^n \\
&\leq (k_2 \lambda_{\max}^n) g_2\left(\lambda_{\min}(\theta' - \theta)\right) \\
\text{(51)} \qquad &= k_U g_U\left(\theta' - \theta\right),
\end{aligned}
$$

where g_U is another zero-mean Gaussian. Moving from the second line to the third line above is a consequence of (9). Moving from the third line to the fourth line uses the lower bound in (9) and the fact that g_2 is a Gaussian with zero mean, which implies that $g_2(x_1) > g_2(x_2)$ when $\|x_1\| < \|x_2\|$. Notice that k_U does not depend on the particular coefficients of the map

© 2018 U.S. Government

$\widetilde{T}$; it only depends on the Lipschitz constant in (9). A similar process can be used to obtain the following lower bound:

$$
\begin{aligned}
q_{\theta,\bar{\gamma}}(\theta'|\theta) &= q_r(\widetilde{T}(\theta')|\widetilde{T}(\theta))|\det \nabla\widetilde{T}(\theta')| \\
&\geq q_r(\widetilde{T}(\theta')|\widetilde{T}(\theta))\lambda_{\min}^n \\
&\geq k_1 g_1(\widetilde{T}(\theta') - \widetilde{T}(\theta))\lambda_{\min}^n \\
&\geq (k_1\lambda_{\min}^n)\, g_1\left(\lambda_{\max}\left(\theta' - \theta\right)\right) \\
&= k_L g_L\left(\theta' - \theta\right).
\end{aligned}
$$

(52)

Lemma 5.3 follows directly from (51) and (52).   ∎

B.2. SSAGE. With (51) and (52) in hand, the proof of Lemma 6.1 in [3] yields Lemma 5.5: the minorization component of the SSAGE condition. Thus, to show SSAGE, we only need to establish Lemma 5.6. Our proof of Lemma 5.6 is built on the intermediate Lemmas B.1 and B.2 provided below and on the proof of Lemma 6.2 in [3].

For the arguments below, we will use the Metropolis–Hastings transition kernel given by

$$
P_{\bar{\gamma}}(x, dy) = \alpha_{\bar{\gamma}}(x, y)q_{\theta,\bar{\gamma}}(y|x)dy + r_{\bar{\gamma}}(x)\delta_x(dy),
$$

where

$$
r_{\bar{\gamma}}(x) = 1 - \int \alpha_{\bar{\gamma}}(x, y)q_{\theta,\bar{\gamma}}(y|x)dy,
$$

and α is the Metropolis–Hastings acceptance probability given by

$$
\alpha_{\bar{\gamma}}(x, y) = \min\left\{1, \frac{\pi(y)q_{\theta,\bar{\gamma}}(x|y)}{\pi(x)q_{\theta,\bar{\gamma}}(y|x)}\right\}.
$$

We will also use the set of guaranteed acceptance, given by

$$
A_{\bar{\gamma}}(x) = \left\{y \in \mathbb{R}^n : \pi(y)q_{\theta,\bar{\gamma}}(x|y) \geq \pi(x)q_{\theta,\bar{\gamma}}(y|x)\right\},
$$

and the set of possible rejection, simply defined as the complement of the set above:

$$
R_{\bar{\gamma}}(x) = A_{\bar{\gamma}}(x)^C.
$$

Lemma B.1. *Let $V(x) = c_V\pi^{-\alpha}(x)$ be a drift function defined by some $\alpha \in (0, 1)$. The constant $c_V = \sup_x \pi^\alpha(x)$ is chosen so that $\inf_x V(x) = 1$. Then the following holds:*

(53)
$$
\limsup_{\|x\|\to\infty} \sup_{\bar{\gamma}} \frac{\int_{\mathbb{R}^n} V(y)P_{\bar{\gamma}}(x, dy)}{V(x)} < \limsup_{\|x\|\to\infty} \sup_{\bar{\gamma}} \int_{R_{\bar{\gamma}}(x)} q_{\theta,\bar{\gamma}}(y|x)dy.
$$

Proof. First, we decompose the left-hand side of (53) into

$$
\frac{\int_{\mathbb{R}^n} V(y)P_{\bar{\gamma}}(x, dy)}{V(x)} = \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}q_{\theta,\bar{\gamma}}(y|x)dy + \int_{R_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}\frac{\pi(y)q_{\theta,\bar{\gamma}}(x|y)}{\pi(x)q_{\theta,\bar{\gamma}}(y|x)}q_{\theta,\bar{\gamma}}(y|x)dy
$$

(54)
$$
+ \int_{R_{\bar{\gamma}}(x)} \left(1 - \frac{\pi(y)q_{\theta,\bar{\gamma}}(x|y)}{\pi(x)q_{\theta,\bar{\gamma}}(y|x)}\right) q_{\theta,\bar{\gamma}}(y|x)dy.
$$

Following the proof of Lemma 6.2 in [3], we can show that the first two integrals in (54) go to zero as $\|x\| \to \infty$. With that, we have

$$\limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, dy)}{V(x)} = \limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \int_{R_{\bar{\gamma}}(x)} \left(1 - \frac{\pi(y) q_{\theta, \bar{\gamma}}(x|y)}{\pi(x) q_{\theta, \bar{\gamma}}(y|x)}\right) q_{\theta, \bar{\gamma}}(y|x) dy$$

$$(55) \qquad\qquad\qquad < \limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \int_{R_{\bar{\gamma}}(x)} q_{\theta, \bar{\gamma}}(y|x) dy.$$

The inequality results from the fact that $[\pi(y) q_{\theta, \bar{\gamma}}(x|y)]/[\pi(x) q_{\theta, \bar{\gamma}}(y|x)] < 1$ when $y \in R_{\bar{\gamma}}(x)$. ∎

Lemma B.2. *The proposal has a nonzero probability of acceptance, i.e.,*

$$(56) \qquad\qquad \int_{R_{\bar{\gamma}}(x)} q_{\theta, \bar{\gamma}}(y|x) dy < 1.$$

Proof. A nonzero probability of acceptance occurs if and only if there is a measurable set $W(x) \subset A_{\bar{\gamma}}(x)$. To show that $W(x)$ exists, consider a small ball of radius R around x. Since g_L and g_U are zero mean and have positive variance, this implies

$$(57) \qquad\qquad \inf_{y \in B(x,R)} \inf_{\bar{\gamma}} \frac{q_{\theta, \bar{\gamma}}(x|y)}{q_{\theta, \bar{\gamma}}(y|x)} \geq \inf_{y \in B(x,R)} \frac{k_L g_L(x-y)}{k_U g_U(y-x)} \geq c_0$$

for some $c_0 > 0$. Because $\pi(x)$ is super-exponentially light, for any $u \in (0, R)$, there exists a radius r_4 such that $\|x\| > r_4$ implies

$$\pi\left(x - u\frac{x}{\|x\|}\right) \geq \frac{\pi(x)}{c_0}.$$

Subsequently, for any map coefficients, the acceptance probability for $x_1 = x - u\frac{x}{\|x\|}$ is one, which implies that $x_1 \in A_{\bar{\gamma}}(x)$. Now, define $W(x)$ as

$$W(x) = \left\{x_1 - a\zeta, \, 0 < a < R - u, \, \zeta \in S^{n-1}, \, \left\|\zeta - \frac{x_1}{\|x_1\|}\right\| < \frac{\epsilon}{2}\right\},$$

where ϵ is an arbitrarily small scalar and S^{n-1} is the unit sphere in $\mathbb{R}^n$ dimensions. Note that $\|\zeta - x_1/\|x_1\|\| < \frac{\epsilon}{2}$ ensures that $W(x)$ is a cone of points closer to the origin than x_1. Now, using the final paragraph of the proof of Lemma 6.2 in [3], the curvature condition from (29) ensures that the target density is larger in $W(x)$ than at x_1. Since x_1 was accepted, this means that everything in $W(x)$ will also be accepted and that $W(x) \subseteq A_{\bar{\gamma}}(x)$. Subsequently, we obtain

$$\lim_{\|x\| \to \infty} \int_{R_{\bar{\gamma}}(x)} q_{\theta, \bar{\gamma}}(y|x) dy = \lim_{\|x\| \to \infty} \left(1 - \int_{A_{\bar{\gamma}}(x)} q_{\theta, \bar{\gamma}}(y|x) dy\right)$$

$$\leq \lim_{\|x\| \to \infty} \left(1 - \int_{W(x)} q_{\theta, \bar{\gamma}}(y|x) dy\right)$$

$$(58) \qquad\qquad\qquad < 1,$$

where we have used the fact that $W(x)$ is a measurable subset of $A_{\bar{\gamma}}(x)$ for large x. ∎

With Lemmas B.1 and B.2 in hand, we can now proceed to the proof of Lemma 5.6 (the drift condition) from the main text.

Proof of Lemma 5.6. Recall our choice of drift function: $V(x) = c_V \pi^{-\alpha}(x)$ for $\alpha \in (0,1)$. Using this function and the definitions of $P_{\bar{\gamma}}$, $R_{\bar{\gamma}}$, and $A_{\bar{\gamma}}$ we can show that

$$
\begin{aligned}
\frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, dy)}{V(x)} &= \frac{\int_{\mathbb{R}^n} \pi^{-\alpha}(y) P_{\bar{\gamma}}(x, dy)}{\pi^{-\alpha}(x)} \\
&\quad + \int_{R_{\bar{\gamma}}(x)} \left(1 - \frac{\pi(y) q_{\theta,\bar{\gamma}}(x|y)}{\pi(x) q_{\theta,\bar{\gamma}}(y|x)} \right) q_{\theta,\bar{\gamma}}(y|x) dy \\
&= \int_{R_{\bar{\gamma}}(x)} q_{\theta,\bar{\gamma}}(y|x) dy + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&\quad + \int_{R_{\bar{\gamma}}(x)} \left(\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} - 1 \right) \frac{\pi(y) q_{\theta,\bar{\gamma}}(x|y)}{\pi(x) q_{\theta,\bar{\gamma}}(y|x)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&\quad + \int_{R_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} \frac{\pi(y) q_{\theta,\bar{\gamma}}(x|y)}{\pi(x) q_{\theta,\bar{\gamma}}(y|x)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&\leq 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&\quad + \int_{R_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} \frac{\pi(y) q_{\theta,\bar{\gamma}}(x|y)}{\pi(x) q_{\theta,\bar{\gamma}}(y|x)} q_{\theta,\bar{\gamma}}(y|x) dy.
\end{aligned}
\tag{59}
$$

Within the region of possible rejection $R_{\bar{\gamma}}(x)$, the acceptance rates are all in $[0,1]$, which allows us to further simplify (59) to obtain

$$
\begin{aligned}
\frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, dy)}{V(x)} &\leq 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy + \int_{R_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&< 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy + \int_{R_{\bar{\gamma}}(x)} \frac{q_{\theta,\bar{\gamma}}^{-\alpha}(y|x)}{q_{\theta,\bar{\gamma}}^{-\alpha}(x|y)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&= 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy + \int_{R_{\bar{\gamma}}(x)} q_{\theta,\bar{\gamma}}^{1-\alpha}(y|x) q_{\theta,\bar{\gamma}}^{\alpha}(x|y) dy \\
&\leq 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy \\
&\leq 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy + k_U^2 \int_{R_{\bar{\gamma}}(x)} g_U(y - x) dy \\
&= 1 + C_R + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta,\bar{\gamma}}(y|x) dy,
\end{aligned}
\tag{60}
$$

where we have used the density upper bound in (51) and C_R is a finite constant. A similar

application of (51) over $A_{\bar{\gamma}}(x)$ yields

$$\frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, dy)}{V(x)} \leq 1 + C_R + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^\alpha(x)}{\pi^\alpha(y)} q_{\theta, \bar{\gamma}}(y|x) dy$$

$$\leq 1 + C_R + \int_{A_{\bar{\gamma}}(x)} q_{\theta, \bar{\gamma}}^\alpha(x|y) q_{\theta, \bar{\gamma}}^{1-\alpha}(y|x) dy$$

$$\leq 1 + C_R + k_U^2 \int_{A_{\bar{\gamma}}(x)} g_U(x - y) dy$$

(61) $< \infty.$

Using Lemmas B.1 and B.2, we also have that

(62) $$\limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, dy)}{V(x)} < \limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \int_{R_{\bar{\gamma}}(x)} q_{\theta, \bar{\gamma}}(y|x) dy < 1.$$

From the proof of Lemma 6.2 in [3], which resembles the proofs in [28], Lemma 5.6 follows from simultaneously satisfying the bounds (61) and (62). ∎

B.3. Diminishing adaptation. In addition to SSAGE and containment, Theorem 5.7 requires diminishing adaptation (Definition 5.1). The following proof establishes the diminishing adaptation proposed in Lemma 5.4.

Proof of Lemma 5.4. The proof of this lemma relies on continuity of the map with respect to $\bar{\gamma}$ and the convergence of (25) as the number of samples $K \to \infty$. Note that we do not require (25) (or (8)) to converge to the minimizer of the true KL divergence.

When the MCMC chain is not at an adaptation step, $\bar{\gamma}^{(k+1)} = \bar{\gamma}^{(k)}$. Thus, to show diminishing adaptation, we need to show that the difference between transition kernels at step K and $K + K_U$ decreases as $K \to \infty$. Mathematically, we require

(63) $$\lim_{K \to \infty} \mathbb{P} \left(\sup_{x \in \mathbb{R}^n} \left\| P_{\bar{\gamma}^{(K)}}(x, \cdot) - P_{\bar{\gamma}^{(K+K_U)}}(x, \cdot) \right\|_{TV} \geq \delta_1 \right) = 0$$

for any $\delta_1 > 0$. Because the maps are linear in $\bar{\gamma}$, for a fixed x, the mapping from $\bar{\gamma}$ to $P_{\bar{\gamma}}(x, A)$ is continuous for any A. Combined with the fact that $q_{\theta, \bar{\gamma}}$ is bounded, we have that (63) will be satisfied when

(64) $$\lim_{K \to \infty} \mathbb{P} \left(\left\| \gamma_i^{(K+K_U)} - \gamma_i^{(K)} \right\| \geq \delta \right) = 0$$

for any $\delta > 0$ and all $i \in \{1, 2, \ldots, n\}$. We now turn to proving (64).

Recall that $\bar{\gamma}^{(K)}$ is the minimizer of (25), which is based on a K-sample Monte Carlo approximation of the KL divergence. To notationally simplify (25), we will now use the convention that $\log(0) = -\infty$ and define the objective functions $f_i^{(K)}(\gamma_i)$ and $f_i^{(K+K_U)}(\gamma_i)$ as

(65) $$f_i^{(K)}(\gamma_i) = \frac{1}{K} g(\gamma_i) + \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right],$$

(66) $$f_i^{(K+K_U)}(\gamma_i) = \frac{1}{K} g(\gamma_i) + \frac{1}{K} \sum_{k=1}^{K+K_U} \left[\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right].$$

From (25), it should be clear that

$$\gamma_i^{(K)} = \operatorname{argmin} f_i^{(K)}(\gamma_i), \tag{67}$$

$$\gamma_i^{(K+K_U)} = \operatorname{argmin} f_i^{(K+K_U)}(\gamma_i) \tag{68}$$

for all $i = \{1, 2, \ldots, n\}$.[3] Combining these expressions, we have

$$f_i^{(K+K_U)}(\gamma_i) = f_i^{(K)}(\gamma_i) + \frac{1}{K} \sum_{k=K+1}^{K+K_U} \left[\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right]. \tag{69}$$

From Markov's inequality, we then have

$$\mathbb{P} \left[\left| f_i^{(K+K_U)}(\gamma_i) - f_i^{(K)}(\gamma_i) \right| \geq \delta_2 \right] \leq \frac{1}{K\delta_2} \mathbb{E} \left[\left| \sum_{k=K+1}^{K+K_U} \left(\frac{1}{2} \widetilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \widetilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right) \right| \right] \tag{70}$$

for any $\delta_2 > 0$ and all γ_i. Notice that the expectation on the right-hand side of this expression is finite because the map is bi-Lipschitz (9), the proposal density is bounded by Gaussian densities (see Lemma 5.3), and the map is linear for large $\|\theta\|$ (see (11)). Thus,

$$\lim_{K\to\infty} \mathbb{P} \left[\left| f_i^{(K+K_U)}(\gamma_i) - f_i^{(K)}(\gamma_i) \right| \geq \delta_2 \right] = 0 \quad \forall \gamma_i. \tag{71}$$

We now show that this implies the convergence of $\|\gamma_i^{(K+K_U)} - \gamma_i^{(K)}\|$. First, consider a set $\mathcal{C}^{(K)}$ that depends on δ_2 and takes the form

$$\mathcal{C}^{(K)} = \left\{ \gamma_i \, : \, f_i^{(K)}(\gamma_i) - \delta_2 \leq f_i^{(K)}(\gamma_i^{(K)}) + \delta_2 \right\}. \tag{72}$$

By definition, $\mathcal{C}^{(K)}$ will always contain $\gamma_i^{(K)}$. Recall that $f_i^{(K)}$ is convex and admits a unique global minimizer. Thus, as $\delta_2 \to 0$, the set $\mathcal{C}^{(K)}$ will collapse on $\gamma_i^{(K)}$ and the maximum distance between any two points in $\mathcal{C}^{(K)}$ will go to zero. This implies that for any $\delta > 0$, there exists a δ_2 such that

$$\sup_{\gamma_i, \gamma_i' \in \mathcal{C}^{(K)}} \|\gamma_i - \gamma_i'\| < \delta. \tag{73}$$

We will now combine this expression with (71). Notice that for any $\delta_2 > 0$, (71) implies that

$$\lim_{K\to\infty} \mathbb{P} \left(\gamma_i^{(K+K_U)} \in \mathcal{C}^{(K)} \right) = 1. \tag{74}$$

Combining this result with (73) yields

$$\lim_{K\to\infty} \mathbb{P} \left(\left\| \gamma_i^{(K+K_U)} - \gamma_i^{(K)} \right\| \geq \delta \right) = 0, \tag{75}$$

which is the desired condition in (64). ∎

---

[3]Using the factor $\frac{1}{K}$ in both (65) and (66) is intentional. Multiplying the objective in (25) by any positive scalar will not affect the solution, and the common value of $\frac{1}{K}$ used here simplifies the results later on.

Appendix C. Maple exudation model details. The forward model in section 6.3 is a complicated system of differential-algebraic equations describing maple sap dynamics. Here we give a minimal description of the model. Interested readers should consult the original derivation in [12].

In addition to the differential-algebraic system defined by (38)–(46), the volumes V^v and $V_g^v(t)$ are given by

$$V^v = \pi (R^v)^2 L^v,$$

$$V_g^v(t) = \pi r(t)^2 L^v,$$

$$N = \frac{2\pi (R^f + R^v + W)}{2R^f + W}.$$

The system is solved using MUQ [51], which in turn links to SUNDIALS [26]. The initial conditions for the state variables s_{gi}, s_{iw}, and $r(t)$ are derived from a steady state solution. We put $U(0) = 0$.

The temperature field is assumed to be quasi-steady and is defined by the heat equation

(76) $\qquad \partial_x \left(\kappa(x) \partial_x T(x,t) \right) = 0 \qquad\qquad$ for $x \in (\, s_{iw}(t),\, R^f + 2R^v\,)$,

$\qquad\qquad T(x,t) = 0 \qquad\qquad$ at $x = s_{iw}(t)$,

$\qquad [.7em] \kappa_w \partial_x T(x,t) = h(T_a(t) - T(x,t)) \qquad$ at $x = R^f + 2R^v$,

where $T_a(t)$ is a transient temperature forcing at the edge of the computational domain ($x = R^f + 2R^v$), $h = 10$ is a heat transfer coefficient, and the thermal conductivity is defined piecewise as

$$\kappa(x) = \begin{cases} \kappa_w, & x \in [\, s_{iw}(t),\, R_f + R_v - r(t)\,), \\[0.5em] \kappa_g, & x \in [\, R_f + R_v - r(t),\, R_f + R_v + r(t)\,), \\[0.5em] \kappa_w, & x \in [\, R_f + R_v + r(t),\, R_f + 2R_v\,], \end{cases}$$

where κ_w is the thermal conductivity of water and κ_g is the thermal conductivity of air. At any particular time, it is straightforward to solve (76) analytically, yielding a piecewise linear temperature field.

The inference parameters θ are related to the model parameters in (38)–(46) using the transformations in Table 4; variables with an overbar are default parameters taken from [12] and are shown in Table 5. Values for the remaining physical constants are listed in Table 6.

Table 4

Relationship between inference targets θ and model parameters for the maple problem.

Model variable	Transformation from θ
$s_{gi}(0)$	$(0.5\theta_2 + 0.5)\exp\left[0.2\log(\bar{R}^f)\theta_7 + \log(\bar{R}^f)\right]$
$s_{iw}(0)$	$\exp\left[0.2\log(\bar{R}^f)\theta_7 + \log(\bar{R}^f)\right]$
$r(0)$	$(0.5\theta_2 + 0.5)\exp\left[0.2\log(\bar{R}^v)\theta_8 + \log(\bar{R}^v)\right]$
$p_g^f(0)$	$50{\times}10^3\theta_3 + 150{\times}10^3$
K	$0.2\log(\bar{K})\theta_4 + \log(\bar{K})$
W	$0.2\log(\bar{W})\theta_5 + \log(\bar{W})$
c_s^v	$0.2\log(\bar{c}_s^v)\theta_6 + \log(\bar{c}_s^v)$
R^f	$0.2\log(\bar{R}^f)\theta_7 + \log(\bar{R}^f)$
R^v	$0.2\log(\bar{R}^v)\theta_8 + \log(\bar{R}^v)$
L^f	$0.2\log(\bar{L}^f)\theta_9 + \log(\bar{L}^f)$
L^v	$0.2\log(\bar{L}^v)\theta_{10} + \log(\bar{L}^v)$

Table 5

Default values used to generate synthetic data and to scale the inference parameters in the maple problem.

Symbol	Value	Units	Description
$\bar{s}_{gi}(0)$	$0.7\bar{R}^f$	m	Initial location of gas-ice interface in fiber.
$\bar{s}_{iw}(0)$	$\bar{R}^f$	m	Initial location of ice-water interface in fiber.
$\bar{r}(0)$	$0.3\bar{R}^v$	m	Initial radius of vessel gas bubble.
$\bar{p}_g^f(0)$	$200{\times}10^3$	Pa	Initial gas pressure in fiber.
$\bar{K}$	$1.98{\times}10^{-14}$	$\mathrm{m\,s^{-1}}$	Hydraulic conductivity of fiber-vessel wall.
$\bar{W}$	$3.64{\times}10^{-6}$	m	Thickness of fiber-vessel wall.
$\bar{c}_s^v$	58.4	$\mathrm{mol\,m^{-3}}$	Sucrose concentration in vessel sap.
$\bar{R}^f$	$3.5{\times}10^{-6}$	m	Fiber radius.
$\bar{R}^v$	$2{\times}10^{-5}$	m	Vessel radius.
$\bar{L}^f$	$1{\times}10^{-3}$	m	Fiber length.
$\bar{L}^v$	$5{\times}10^{-4}$	m	Vessel length.

Table 6

Physical constants used in the maple exudation model.

Symbol	Value	Units	Description
ρ_i	917	$\mathrm{kg\,m^{-3}}$	Density of water ice.
ρ_w	1000	$\mathrm{kg\,m^{-3}}$	Density of water.
λ	$3.34{\times}10^5$	$\mathrm{J\,kg^{-1}}$	Latent heat of fusion for water.
R	8.314	$\mathrm{J\,mol^{-1}\,K^{-1}}$	Universal gas constant.
g	9.81	$\mathrm{m\,s^{-2}}$	Acceleration due to gravity.
σ	0.0756	$\mathrm{N\,m^{-1}}$	Surface tension of water.
H	0.0274	–	Henry's constant for air and water.
M_g	0.0290	$\mathrm{kg\,mol^{-1}}$	Molar mass of air.
κ_w	0.580	$\mathrm{W\,m^{-1}\,K^{-1}}$	Thermal conductivity of water.
κ_g	0.0243	$\mathrm{W\,m^{-1}\,K^{-1}}$	Thermal conductivity of air.

REFERENCES

[1] L. Ambrosio and N. Gigli, *A user's guide to optimal transport*, in Modelling and Optimisation of Flows on Networks, Springer, 2013, pp. 1–155.

[2] C. Andrieu and E. Moulines, *On the ergodicity properties of some adaptive MCMC algorithms*, Ann. Appl. Probab., 16 (2006), pp. 1462–1505, https://doi.org/10.1214/105051606000000286.

[3] Y. F. Atchadé, *An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift*, Methodol. Comput. Appl. Probab., 8 (2006), pp. 235–254, https://doi.org/10.1007/s11009-006-8550-0.

[4] Y. Bai, G. Roberts, and J. Rosenthal, *On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms*, tech. report, University of Warwick, 2009, http://wrap.warwick.ac.uk/35203/.

[5] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, *Randomize-then-Optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1895–A1910, https://doi.org/10.1137/140964023.

[6] P. Bernard and B. Buffoni, *Optimal Mass Transportation and Mather Theory*, preprint, https://arxiv.org/abs/math/0412299, 2004.

[7] N. Bonnotte, *From Knothe's rearrangement to Brenier's optimal transport map*, SIAM J. Math. Anal., 45 (2013), pp. 64–87, https://doi.org/10.1137/120874850.

[8] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417, http://onlinelibrary.wiley.com/doi/10.1002/cpa.3160440402/abstract.

[9] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds., *Handbook of Markov Chain Monte Carlo*, Chapman and Hall, 2011.

[10] L. A. Caffarelli, *The regularity of mappings with a convex potential*, J. Amer. Math. Soc., 5 (1992), pp. 99–104.

[11] G. Carlier, A. Galichon, and F. Santambrogio, *From Knothe's transport to Brenier's map and a continuation method for optimal transport*, SIAM J. Math. Anal., 41 (2010), pp. 2554–2576, https://doi.org/10.1137/080740647.

[12] M. Ceseri and J. M. Stockie, *A mathematical model of sap exudation in maple trees governed by ice melting, gas dissolution, and osmosis*, SIAM J. Appl. Math., 73 (2013), pp. 649–676, https://doi.org/10.1137/120880239.

[13] T. Champion and L. De Pascale, *The Monge problem in $\mathbb{R}^d$*, Duke Math. J., 157 (2011), pp. 551–572.

[14] A. J. Chorin, M. Morzfeld, and X. Tu, *Implicit particle filters for data assimilation*, Commun. Appl. Math. Comput. Sci., 5 (2010), pp. 221–240, https://msp.org/camcos/2010/5-2/camcos-v5-n2-s.pdf#page=74, http://msp.org/camcos/2010/5-2/p03.xhtml.

[15] A. J. Chorin and X. Tu, *Implicit sampling for particle filters*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 17249–17254.

[16] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems, 2013, pp. 2292–2300.

[17] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, *A tutorial on the cross-entropy method*, Ann. Oper. Res., 134 (2005), pp. 19–67.

[18] D. Feyel and A. S. Üstünel, *Monge-Kantorovitch measure transportation and Monge-Ampere equation on Wiener space*, Probab. Theory Related Fields, 128 (2004), pp. 347–385.

[19] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed., Chapman and Hall, 2003.

[20] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer, 1991.

[21] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Internat. Statist. Rev., 70 (2002), pp. 419–435.

[22] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 1–37.

[23] H. Haario, M. Laine, A. Mira, and E. Saksman, *DRAM: Efficient adaptive MCMC*, Statist. Comput., 16 (2006), pp. 339–354.

[24] H. Haario, E. Saksman, and J. Tamminen, *An adaptive Metropolis algorithm*, Bernoulli, 7 (2001),

pp. 223–242, http://projecteuclid.org/euclid.bj/1080222083.

[25] W. HASTINGS, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109, http://biomet.oxfordjournals.org/content/57/1/97.short.

[26] A. C. HINDMARSH, P. N. BROWN, K. E. GRANT, S. L. LEE, R. SERBAN, D. E. SHUMAKER, AND C. S. WOODWARD, *SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers*, ACM Trans. Math. Software, 31 (2005), pp. 363–396.

[27] M. HOFFMAN AND A. GELMAN, *The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo*, J. Mach. Learn. Res., 15 (2014), pp. 1351–1381, http://www.stat.columbia.edu/~gelman/research/published/nuts.pdf.

[28] S. R. F. JARNER AND E. HANSEN, *Geometric ergodicity of Metropolis algorithms*, Stochastic Process. Appl., 85 (2000), pp. 341–361, https://doi.org/10.1016/S0304-4149(99)00082-4.

[29] L. JOHNSON AND C. GEYER, *Variable transformation to obtain geometric ergodicity in the random walk Metropolis algorithm*, Ann. Statist., (2012), pp. 1–30, http://projecteuclid.org/euclid.aos/1361542074.

[30] L. V. KANTOROVICH, *On the transfer of masses*, Dokl. Akad. Nauk. SSSR, 37 (1942), pp. 227–229.

[31] A. J. KLEYWEGT, A. SHAPIRO, AND T. HOMEM-DE-MELLO, *The sample average approximation method for stochastic discrete optimization*, SIAM J. Optim., 12 (2001), pp. 479–502, https://doi.org/10.1137/S1052623499363220.

[32] H. KNOTHE, *Contributions to the theory of convex bodies*, Michigan Math. J., 4 (1957), pp. 39–52, https://doi.org/10.1307/mmj/1028990175.

[33] H. KUSHNER AND G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, 2003.

[34] V. LAPARRA, G. CAMPS-VALLS, AND J. MALO, *Iterative Gaussianization: From ICA to random rotations*, IEEE Trans. Neural Networks, 22 (2011), pp. 1–13, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5720319.

[35] O. LE MAITRE AND O. M. KNIO, *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*, Springer, 2010.

[36] J. S. LIU, *Monte Carlo Strategies in Scientific Computing*, Springer, 2004.

[37] D. J. MACKAY, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[38] J. MARTIN, L. C. WILCOX, C. BURSTEDDE, AND O. GHATTAS, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. 1460–1487, https://doi.org/10.1137/110845598.

[39] Y. MARZOUK, T. MOSELHY, M. PARNO, AND A. SPANTINI, *Sampling via measure transport: An introduction*, in Handbook of Uncertainty Quantification, R. Ghanem, D. Higdon, and H. Owhadi, eds., Springer, 2016, https://doi.org/10.1007/978-3-319-11259-6_23-1.

[40] R. MCCANN, *Existence and uniqueness of monotone measure-preserving maps*, Duke Math. J., 80 (1995), pp. 309–323, www.math.toronto.edu/mccann/papers/monotone.ps.

[41] N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1092, https://doi.org/10.1063/1.1699114.

[42] T. P. MINKA, *Expectation propagation for approximate Bayesian inference*, in Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 2001, pp. 362–369.

[43] A. MIRA, *On Metropolis-Hastings algorithms with delayed rejection*, Metron, 59 (2001), pp. 231–241, ftp://luna.sta.uniroma1.it/RePEc/articoli/2001-LIX-3_4-16.pdf.

[44] G. MONGE, *Mémoire sur la théorie des déblais et de remblais*, in Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, 1781, pp. 666–704.

[45] R. MORRISON, R. BAPTISTA, AND Y. MARZOUK, *Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting*, in Advances in Neural Information Processing Systems 30, 2017, pp. 2356–2366.

[46] M. MORZFELD, X. TU, E. ATKINS, AND A. J. CHORIN, *A random map implementation of implicit filters*, J. Comput. Phys., 231 (2012), pp. 2049–2066, https://doi.org/10.1016/j.jcp.2011.11.022.

[47] T. MOSELHY AND Y. MARZOUK, *Bayesian inference with optimal maps*, J. Comput. Phys., 231 (2012), pp. 7815–7850, http://adsabs.harvard.edu/abs/2011arXiv1109.1516E.

[48] R. M. Neal, *MCMC using Hamiltonian dynamics*, in Handbook of Markov Chain Monte Carlo, S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds., Taylor and Francis, Boca Raton, FL, 2011, pp. 113–162.

[49] I. Olkin and F. Pukelsheim, *The distance between two random vectors with given dispersion matrices*, Linear Algebra Appl., 48 (1982), pp. 257–263.

[50] M. Parno, *Transport Maps for Accelerated Bayesian Computation*, Ph.D. thesis, Massachusetts Institute of Technology, 2014.

[51] M. Parno, A. Davis, and P. Conrad, *MIT Uncertainty Quantification (MUQ) library*, 2014, https://bitbucket.org/mituq/muq.

[52] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Volume* I: *Theory*, Vol. 1, Springer, 1998.

[53] S. Reich, *A nonparametric ensemble transform method for Bayesian inference*, SIAM J. Sci. Comput., 35 (2013), pp. A2013–A2024, https://doi.org/10.1137/130907367.

[54] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, 2004.

[55] G. O. Roberts and J. S. Rosenthal, *Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms*, J. Appl. Probab., 44 (2007), pp. 458–475, http://www.jstor.org/stable/27595854.

[56] G. O. Roberts and J. S. Rosenthal, *General state space Markov chains and MCMC algorithms*, Probab. Surv., 1 (2004), pp. 20–71, https://doi.org/10.1214/154957804100000024.

[57] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363, https://doi.org/10.2307/3318418.

[58] L. Rockwood, *Introduction to Population Ecology*, 1st ed., Wiley-Blackwell, 2006.

[59] M. Rosenblatt, *Remarks on a multivariate transformation*, Ann. Math. Statist., 23 (1952), pp. 470–472, https://www.jstor.org/stable/10.2307/2236692.

[60] D. Sanz-Alonso, *Importance Sampling and Necessary Sample Size: An Information Theory Approach*, preprint, https://arxiv.org/abs/1608.08814, 2016.

[61] A. Spantini, D. Bigoni, and Y. M. Marzouk, *Inference via Low-Dimensional Couplings*, preprint, https://arxiv.org/abs/1703.06131, 2017.

[62] S. Strogatz, *Nonlinear Dynamics and Chaos*, Westview Press, 2001.

[63] A. B. Sullivan, D. M. Snyder, and S. A. Rounds, *Controls on biochemical oxygen demand in the upper Klamath River, Oregon*, Chem. Geol., 269 (2010), pp. 12–21, https://doi.org/10.1016/j.chemgeo.2009.08.007.

[64] E. Tabak and C. Turner, *A family of nonparametric density estimation algorithms*, Comm. Pure Appl. Math., 66 (2013), pp. 145–164, http://onlinelibrary.wiley.com/doi/10.1002/cpa.21423/full.

[65] E. G. Tabak and E. Vanden-Eijnden, *Density estimation by dual ascent of the log-likelihood*, Commun. Math. Sci., 8 (2010), pp. 217–233.

[66] A. M. Vershik, *Long history of the Monge-Kantorovich transportation problem*, Math. Intelligencer, 35 (2013), pp. 1–9, https://doi.org/10.1007/s00283-013-9380-x.

[67] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.

[68] C. Villani, *Optimal Transport: Old and New*, Springer-Verlag, 2009.

[69] H. Wilf, *A global bisection algorithm for computing the zeros of polynomials in the complex plane*, J. Assoc. Comput. Mach., 25 (1978), pp. 415–420, http://dl.acm.org/citation.cfm?id=322084.

[70] U. Wolff, *Monte Carlo errors with less errors*, Comput. Phys. Comm., 156 (2004), pp. 143–153.

[71] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644, https://doi.org/10.1137/S1064827501387826.