Adaptive Algorithms for Continuous-Time Transport: Homotopy-Driven Sampling and a New Interacting Particle System

Aimee Maurais & Youssef Marzouk
Center for Computational Science and Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139
{maurais, ymarz}@mit.edu

Abstract

We propose a new dynamic algorithm which transports samples from a reference distribution to a target distribution in unit time, given access to the target-to-reference density ratio. Our approach is to seek a sequence of transport maps that push forward the reference along a path given by a geometric mixture of the two densities. We take the maps to be simply parameterized, local, sample-driven optimal transport maps which we identify by approximately solving a root-finding problem formulated using importance weights. When feature functions for the maps are taken to be kernels, we obtain a novel interacting particle system from which we derive finite-particle and mean-field ODEs. In discrete time, we introduce an adaptive algorithm for simulating this interacting particle system which adjusts the ODE time steps based on the quality of the transport, automatically uncovering a good "schedule" for traversing the geometric mixture of densities.

1 Introduction

In this work we consider the problem of *sampling via transport*: given a target distribution π_1 on $\mathbb{R}^d$ and a reference π_0 on $\mathbb{R}^d$ from which we can sample, our goal is to find $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\# \pi_0 = \pi_1$, i.e., $\{X_0^{(j)}\}_{j=1}^J \sim \pi_0 \Rightarrow \{T(X_0^{(j)})\}_{j=1}^J \sim \pi_1$. We assume that π_0 and π_1 both admit densities and that we can evaluate the (unnormalized) *density ratio*[1] $\frac{\pi_1}{\pi_0}$ but do not have samples of π_1 with which to train the map or access to gradients (including the score) of π_1. The target-to-reference density ratio is available when the density of π_1 is known and π_0 is chosen to be some "standard" reference (e.g., Gaussian), but is also accessible in the Bayesian setting so long as the likelihood function is known: therein $\pi_1 \propto \ell\, \pi_0$ for some likelihood $\ell(x) = \pi(y^*|x)$, and hence the ratio can be computed $\frac{\pi_1}{\pi_0} \propto \ell$.

The canonical sampling approach employing a density ratio is importance sampling [30], which transforms an unweighted ensemble of samples of π_0 into a *weighted* ensemble, enabling the estimation of expectations under π_1. Importance sampling is the foundation for sequential Monte Carlo methods [11], but is frequently plagued by issues of weight degeneracy and ensemble collapse, necessitating large ensemble sizes [36] or interventions such as resampling [24] and MCMC rejuvenation.

Importance weights can alternately be used as *ingredients* to build transport maps which, when applied to samples from π_0, yield *uniformly* weighted approximate samples from π_1. This strategy is employed in the analysis step of the ensemble transform particle filter of [34], wherein importance

[1] For the remainder of this paper the terms "density" and "ratio" refer to unnormalized quantities unless otherwise stated.

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

weights are used to define the marginals of a discrete optimal transport (OT) problem and the resulting OT coupling used as a Bayesian prior-to-posterior transformation. While this approach is shown to be consistent and has the benefit of being nonparametric, the transformation obtained is linear and transformed samples cannot leave the convex hull of the originals. However, the basic premise of [34]—using importance weights to define optimal transport problems—is nonetheless the inspiration for the approach we use here to find local OT maps within a new homotopy method for transport.

In our method we transport samples from π_0 to π_1 in unit time by applying a sequence of transport maps that push forward π_0 along a discretization of the geometric mixture $\pi_t \propto \pi_0^{1-t}\pi_1^t$, $t \in [0,1]$. We obtain the map at each step based on the fact that for Δt sufficiently small, the sample-driven OT map [23] which pushes π_t to $\pi_{t+\Delta t}$ can be well-approximated by a perturbation of the identity map with a linear combination of gradients of "feature functions." We identify the coefficients of this combination by solving a linearized discretization of the Monge–Ampère equations formulated using importance weights. When the feature functions are chosen to be kernels, this approach gives rise to a novel interacting particle system with intriguing continuous-time and mean-field limits. In discrete time, we introduce an adaptive algorithm for realizing this sequence of maps which adjusts the time increments based on the *quality* of the approximate transport. In a sense we thus use local, sample-driven OT as an adaptive time-stepper to propagate samples along the prescribed (non-optimal) geometric mixture path.

The paper is organized as follows: in Section 2 we review existing approaches to sampling via transport. In Section 3 we introduce our method, encompassing our adaptation of sample-driven OT (Section 3.1), algorithmic formulations (Section 3.2), and choice of feature functions (Section 3.3). We examine the continuous-time and mean-field limits of our algorithm with kernel features in Section 4 and provide a numerical demonstration in Section 5. We close in Section 6.

2 Background

Sampling via measure transport is an active area of research, with many computational approaches [28, 22, 31, 42] appearing in recent years. Most practical transport maps are parameterized, and thus a crucial part of realizing them is selecting an appropriately rich function class within which to search for the map. Common map approximation classes include polynomials [32, 3], radial basis functions [39], composed simple transformations [35, 31, 22], neural networks [6, 41, 2], and reproducing kernel Hilbert spaces [26, 23]. Determining an appropriate basis to represent a transport map can be challenging, especially when the target and reference distributions are high-dimensional or differ from each other considerably. For this reason it may be necessary to employ, e.g., adaptive feature selection algorithms [3] or dimension reduction techniques [38, 9, 7].

As an alternative to searching for a single, potentially highly complex transport map which pushes the reference π_0 directly to the target π_1, one can instead prescribe a *path* of distributions $(\pi_t)_{t\in[0,1]}$ having the target and reference as endpoints and seek a sequence of maps $T_1, \ldots, T_N$ which push samples along a discretization of the path, e.g.,

$$\pi_0 \xrightarrow{T_1} \pi_{\frac{1}{N}} \xrightarrow{T_2} \cdots \xrightarrow{T_{N-1}} \pi_{\frac{N-1}{N}} \xrightarrow{T_N} \pi_1. \tag{1}$$

The composed map $T = T_N \circ T_{N-1} \circ \cdots \circ T_1$ thus pushes forward π_0 to π_1. This approach underlies flow, diffusion, and bridge techniques for generative modeling, e.g., [10, 23, 25, 27, 45, 1, 37], wherein access to samples from both π_0 and π_1 is almost invariably required for training (with [43, 20] being recent exceptions). In the setting where π_1 is known only through its unnormalized density, there are a number of *infinite-time* compositional sampling algorithms which have their grounding as approximate Wasserstein gradient flows or Langevin diffusions, e.g., [26, 17, 18, 8], but in practice we cannot actually run these iterations for infinite time.

Finite-time samplers of unnormalized densities which employ the homotopy approach (1) frequently take π_t to be the geometric mixture $\pi_t \propto \pi_0^{1-t}\pi_1^t = \pi_0(\frac{\pi_1}{\pi_0})^t$, $t \in [0,1]$ or a reparameterization thereof. This mixture may be referred to as the "power posterior" path and appears, for example, in annealed importance sampling [29, 4] and parallel tempering [19, 15, 40]. In Bayesian computation this path is sometimes referred to as "tempered likelihood" and has been used as the basis for algorithms which generate (approximate) posterior samples [33, 21, 13] or posterior densities [12].

3 Method: time-stepping with local optimal transport

We employ the power posterior in a dynamical approach to sampling, seeking a sequence of maps $T_1, \ldots, T_N$ that push samples along a discretization of the homotopy

$$\pi_t \propto \pi_0^{1-t} \pi_1^t = \pi_0 \left(\frac{\pi_1}{\pi_0}\right)^t, \quad t \in [0,1] \tag{2}$$

given only samples $\{X_0^{(j)}\}_{j=1}^J \overset{\text{i.i.d.}}{\sim} \pi_0$ and access to the density ratio $\frac{\pi_1}{\pi_0}$. As in [34], *incremental importance weights* will inform inform our search for the maps $T_1, \ldots, T_N$. Our high-level approach is summarized in Algorithm 1, where for simplicity we assume a uniform time-step of Δt such that $1 = N\Delta t$ for some $N \in \mathbb{N}$. We will revisit this assumption in Section 3.2.

Algorithm 1 Tempered transport with importance weights

```
Require: Reference ensemble $\{X_0^{(j)}\}_{j=1}^J \overset{\text{i.i.d.}}{\sim} \pi_0$, density ratio $\pi_1/\pi_0$, timestep $\Delta t \in (0,1]$
1: $N \leftarrow 1/\Delta t, \quad t \leftarrow 0$
2: for $n = 1, \ldots, N$ do
3:     Compute importance weights: $w_t^{(j)} = \frac{(\frac{\pi_1}{\pi_0}(X_t^{(j)}))^{\Delta t}}{\sum_{i=1}^J (\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}}, \quad j \in \{1, \ldots, J\}$
4:     Estimate map $T_n : \mathbb{R}^d \to \mathbb{R}^d$ using the empirical measure $\sum_{j=1}^J w_t^{(j)} \delta_{X_t^{(j)}}$ such that
           $T_{n\#}\pi_{X_t} = \pi_{X_{t+\Delta t}} \approx \pi_0 \left(\frac{\pi_1}{\pi_0}\right)^{t+\Delta t}$
5:     Transport samples: $X_{t+\Delta t}^{(j)} = T_n(X_t^{(j)}), \quad j \in \{1, \ldots, J\}$
6:     $t \leftarrow t + \Delta t$
7: end for
Ensure: $\{X_1^{(j)}\}_{j=1}^J \sim \pi_{X_1} \approx \pi_1$
```

3.1 Sample-driven optimal transport with importance weights

In this work we take the maps T_n in line 5 of Algorithm 1 to be *local, sample-driven optimal transport maps* as formulated by Kuang and Tabak [23]. Their approach, modified for our setting in which target samples are unavailable, is as follows: at the outset of an iteration of Algorithm 1 we have samples $\{X_t^{(j)}\}_{j=1}^J \sim \pi_t$ which we would like to push forward to $\pi_{t+\Delta t} \propto \pi_t(\frac{\pi_1}{\pi_0})^{\Delta t}$. Given that π_t and $\pi_{t+\Delta t}$ both admit densities, there are many maps $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $T_\#\pi_t = \pi_{t+\Delta t}$. The optimal transport approach [44], which we will approximate in our algorithm, is to seek the map which *minimizes expected transport cost*,

$$\min_{T_\#\pi_t = \pi_{t+\Delta t}} \mathbb{E}_{\pi_t}[\|T(X_t) - X_t\|^2]. \tag{3}$$

Owing to the choice of quadratic cost, one can show that the optimal map in (3) is the unique convex gradient which pushes forward π_t to $\pi_{t+\Delta t}$ [5]. That is, if we find $T = \nabla\phi$ satisfying $T_\#\pi_t = \pi_{t+\Delta t}$ with $\phi : \mathbb{R}^d \to \mathbb{R}$ convex, we have found the optimal transport map. Thus, we can obtain the optimal transport map by seeking $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ convex such that $\nabla\phi_\#\pi_t = \pi_{t+\Delta t}$. The push-forward condition $\nabla\phi_\#\pi_t = \pi_{t+\Delta t}$ can be written as a Monge–Ampère PDE [16]

$$\pi_{t+\Delta t}(\nabla\phi(x)) \det(\nabla^2\phi(x)) = \pi_t(x),$$

and interpreted in weak form as

$$\int_{\mathbb{R}^d} f(\nabla\phi(x))\, \mathrm{d}\pi_t(x) = \int_{\mathbb{R}^d} f(y)\, \mathrm{d}\pi_{t+\Delta t}(y) \quad \forall f : \mathbb{R}^d \to \mathbb{R} \text{ continuous.} \tag{4}$$

In our setting we arguably do not have enough information to find a map $T = \nabla\phi$ which exactly satisfies $T_\#\pi_t = \pi_{t+\Delta t}$, so we discretize the weak-form (4) over finitely many continuous feature functions $f_1, \ldots, f_M : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathbb{R}^d} f_m(\nabla\phi(x))\, \mathrm{d}\pi_t(x) = \int_{\mathbb{R}^d} f_m(y)\, \mathrm{d}\pi_{t+\Delta t}(y), \quad m = 1, \ldots, M, \tag{5}$$

and approximate the expectations on either side using samples from π_t and self-normalized importance weights,

$$\frac{1}{J}\sum_{j=1}^{J} f_m(\nabla\phi(X_t^{(j)})) = \sum_{j=1}^{J} w_t^{(j)} f_m(X_t^{(j)}), \quad w_t^{(j)} = \frac{(\frac{\pi_1}{\pi_0}(X_t^{(j)}))^{\Delta t}}{\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}}, \quad m = 1,\dots,M. \tag{6}$$

Kuang and Tabak [23] refer to the relationship (6) as *sample equivalence* and denote it by $\{\nabla\phi(X_t^{(j)})\}_{j=1}^J \sim \{w_t^{(j)} X_t^{(j)}\}_{j=1}^J$. Because we have discretized the Monge–Ampère equations (4) over finite samples and feature functions, a solution $\nabla\phi$ to (6) is not guaranteed to be unique or optimal. Thus, the *sample-driven* OT problem as formulated in [23] is to find a minimum cost map $\nabla\phi$ which satisfies sample-equivalence,

$$\min_{\{\nabla\phi(X_t^{(j)})\}_{j=1}^J \sim \{w_t^{(j)} X_t^{(j)}\}_{j=1}^J} \sum_{j=1}^{J} \left\| X_t^{(j)} - \nabla\phi(X_t^{(j)}) \right\|^2. \tag{7}$$

Without imposition of further restrictions, the problem (7) will not yield a smooth transport map. Thus Kuang and Tabak [23] suggest parameterizing the potential ϕ by the feature functions themselves,

$$\phi_{\mathbf{s}}(x) = \frac{\|x\|^2}{2} + \sum_{m=1}^{M} s_m f_m(x) \implies \nabla\phi_{\mathbf{s}}(x) = x + \sum_{m=1}^{M} s_m \nabla f_m(x), \tag{8}$$

where $s_1,\dots,s_M \in \mathbb{R}$ are coefficients to be optimized. With this parameterization the sample-based optimal transport problem (7) is reduced to a finite-dimensional constrained optimization over the coefficients $\mathbf{s} \equiv (s_1,\dots,s_M)$. Owing to the relationship between the feature functions $f_1,\dots,f_M$ and the parameterization (8), the optimal $\mathbf{s}$ can be identified via root-finding: define $F : \mathbb{R}^d \to \mathbb{R}^M$ by $F(x) = (f_1(x),\dots,f_M(x))^\top$ with Jacobian $\nabla F(x) \in \mathbb{R}^{M\times d}$ and denote by $\mathbf{a}$ and $\mathbf{b}$ the feature means over the unweighted and weighted reference ensembles,

$$\mathbf{a} = \frac{1}{J}\sum_{j=1}^{J} F(X_t^{(j)}), \quad \mathbf{b} = \sum_{j=1}^{J} w_t^{(j)} F(X_t^{(j)}) \in \mathbb{R}^M.$$

For $\mathbf{s} \in \mathbb{R}^M$, define $G : \mathbb{R}^M \to \mathbb{R}^M$ to be the feature means over $\{\nabla\phi_{\mathbf{s}}(X_t^{(j)})\}_{j=1}^J$,

$$G(\mathbf{s}) = \frac{1}{J}\sum_{j=1}^{J} F(\nabla\phi_{\mathbf{s}}(X_t^{(j)})) = \frac{1}{J}\sum_{j=1}^{J} F(X_t^{(j)} + \mathbf{s}^\top \nabla F(X_t^{(j)})).$$

In order for sample-equivalence to be satisfied, we need to find $\mathbf{s}^*$ such that $G(\mathbf{s}^*) = \mathbf{b}$.

Kuang and Tabak [23] demonstrate that if the Jacobian of G at $\mathbf{s} = 0$ is nonsingular (for which a necessary condition is $M \leq dJ$), G is a bijection from a neighborhood U about $\mathbf{s} = \mathbf{0}$ to a neighborhood V about $G(\mathbf{0}) = \mathbf{a}$. If $\mathbf{b} \in V$, then the potential $\phi_{\mathbf{s}}$ parameterized with $\mathbf{s}^* = G^{-1}(\mathbf{b})$ gives the *global minimum* of the sample-based OT problem (7) restricted to maps of the form (8). Furthermore, [23] shows that if the feature functions are in C^2, then $\phi_{\mathbf{s}^*}$ is locally convex.

Existence of the solution $\mathbf{s}^* = G^{-1}(\mathbf{b})$ and regularity of the importance weights $w_t^{(j)}$ ultimately depend on how close $\mathbf{b}$ is to $\mathbf{a}$, which can be managed by choice of time-step Δt. As $\Delta t \to 0$ the importance weights approach uniformity and $\mathbf{b} \to \mathbf{a}$. Thus, for sufficiently small Δt, a solution $\mathbf{s}^* = G^{-1}(\mathbf{b})$ will exist and the weights $w_t^{(j)}$ will not suffer from degeneracy.

3.2 Adaptive algorithm

Given sufficiently small Δt, in order to find the optimal sample-driven OT map (8) from π_t to $\pi_{t+\Delta t}$ we must solve the root-finding problem $G(\mathbf{s}^*) = \mathbf{b}$. To enable fast updating within Algorithm 1, we restrict ourselves to just *one* step of Newton's method starting from $\mathbf{s}_0 = \mathbf{0}$ and approximate

$$\mathbf{s}^* \approx -\left(\frac{1}{J}\sum_{i=1}^{J} \nabla F(X_t^{(i)}) \nabla F(X_t^{(i)})^\top\right)^{-1} \sum_{k=1}^{J} (\tfrac{1}{J} - w_t^{(k)}) F(X_t^{(k)}), \tag{9}$$

keeping in mind that if the feature functions $f_1, \dots, f_m$ are chosen to be C^1, $G(\mathbf{s})$ will be locally linear and thus the approximation (9) will become exact as $\Delta t \to 0$. This choice also enables us to easily write our algorithm as an interacting particle system, which we discuss in Sections 3.3.2 and 4.

This sample-driven OT approach ((7) and (8)) and local linear approximation (9) we have outlined for finding the local maps $T_1, \dots, T_N$ give rise to a fixed-timestep instantiation of Algorithm 1, which we have included in Appendix A.1 (Algorithm 3). However, the transport maps employed in Algorithm 3 will not achieve exact sample equivalence because we are approximating the solution to

$$\frac{1}{J}\sum_{j=1}^{J} F(X_t^{(j)} + (\nabla F(X_t^{(j)})^\top \mathbf{s}) = \sum_{j=1}^{J} w_t^{(j)} F(X_t^{(j)})$$

with the first Newton iterate starting from $\mathbf{s}_0 = 0$. When the incremental importance weights $w_t^{(j)}$ are sufficiently close to uniformity this local linear approximation should be fairly accurate, but it is hard to know *a priori* how small Δt must be taken to ensure that this is the case. For this reason we suggest performing the time-stepping in the tempered transport update *adaptively*, adjusting the increment Δt on the fly such that a sample-equivalence accuracy criterion is met at each step. An implementation of this idea is given in Algorithm 2.

Algorithm 2 Adaptive tempered transport with sample-driven OT maps

```
Require: Reference ensemble $\{X_0^{(j)}\}_{j=1}^{J} \overset{\text{i.i.d.}}{\sim} \pi_0$, density ratio $\pi_1/\pi_0$, tolerance $\epsilon > 0$, maximum
    time step $\Delta t_{\max} > 0$, features $f_1, \dots, f_M : \mathbb{R}^d \to \mathbb{R}$
$\Delta t \leftarrow \Delta t_{\max}/2$,    $t \leftarrow 0$
while $t < 1$ do,
    success ← false,    $\Delta t \leftarrow \min(\Delta t_{\max},\, 1-t,\, 2\Delta t)$
    while success = false do
        Compute importance weights: $w_t^{(j)} = \frac{(\frac{\pi_1}{\pi_0}(X_t^{(j)}))^{\Delta t}}{\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}}$,    $j \in \{1, \dots, J\}$
        Approximate solution to sample-driven OT with one Newton step:
            $\mathbf{s}_t = -\left(\frac{1}{J}\sum_{i=1}^{J} \nabla F(X_t^{(i)}) \nabla F(X_t^{(i)})^\top\right)^{-1} \sum_{k=1}^{J} (\frac{1}{J} - w_t^{(k)}) F(X_t^{(k)})$
        Compute sample-equivalence error
            $\ell(\mathbf{s}_t) = \frac{1}{M}\left\| \frac{1}{J}\sum_{j=1}^{J} F(X_t^{(j)} + (\nabla F(X_t^{(j)})^\top \mathbf{s}_t) - \sum_{j=1}^{J} w_t^{(j)} F(X_t^{(j)}) \right\|^2$
        if $\ell(\mathbf{s}_t) < \epsilon$ then
            Transport: $X_{t+\Delta t}^{(j)} = X_t^{(j)} + \begin{pmatrix} \nabla F_1(X_t^{(j)}) & \cdots & \nabla F_M(X_t^{(j)}) \end{pmatrix} \mathbf{s}_t$
            success ← true,    $t \leftarrow t + \Delta t$
        else
            $\Delta t \leftarrow \Delta t/2$
        end if
    end while
end while
Ensure: $\{X_1^{(j)}\}_{j=1}^{J} \sim \pi_{X_1} \approx \pi_1$
```

An interesting benefit of Algorithm 2 is that it does not prescribe a particular *schedule* for traversing the path $\pi_t \propto \pi_0^{1-t}\pi_1^t$. For instance, Algorithms 1 and 3 are written assuming a uniform schedule as in (1) but we could easily modify the importance increments and adopt a non-uniform schedule

$$\pi_0 \xrightarrow{T_1} \pi_{\tau(\frac{1}{N})} \xrightarrow{T_2} \cdots \xrightarrow{T_{N-1}} \pi_{\tau(\frac{N-1}{N})} \xrightarrow{T_N} \pi_1,$$

for some continuous, strictly increasing $\tau : [0,1] \to [0,1]$ satisfying $\tau(0) = 0$ and $\tau(1) = 1$ (e.g., $\tau(t) = t^p$ for $p \in \mathbb{N}$, $\tau(t) = \sin(\frac{\pi t}{2})$, etc.). Algorithm 2 *adapts the schedule and the number of steps* *to the problem*, saving us the need to guess what a suitable $\tau(\cdot)$ and N might be for a given π_0 and π_1. We will see examples of schedules traced out by our adaptive Algorithm 2 in Section 5.

3.3 Implementation choice: feature functions

Employing Algorithm 2 requires identifying feature functions $f_1, \dots, f_M$ for enforcement of sample-equivalence at each step. While this choice may be application-specific, here we suggest two generally applicable families from which to select features: multivariate polynomials and kernels.

3.3.1 Multivariate polynomials

A straightforward choice is to take F to be a collection of multivariate polynomials, $F(x) = (P_{\boldsymbol{\alpha}_1}(x), \dots, P_{\boldsymbol{\alpha}_M}(x))^\top$, with each multivariate polyomial $P_{\boldsymbol{\alpha}} : \mathbb{R}^d \to \mathbb{R}$ obtained as the product of univariate polynomials according to a multi-index $\boldsymbol{\alpha} \in (\mathbb{N} \cup \{0\})^d$. That is, $P_{\boldsymbol{\alpha}}(x) = \prod_{i=1}^d P_{\alpha_i}(x_i)$, with $P_{\alpha_1}, \dots P_{\alpha_d}$ univariate polynomials of degrees $\alpha_1, \dots, \alpha_d$. Any number of bases can be used to define the univariate P_{α_i}, but, for ease of exposition, consider a total-degree multiindex set $\{\boldsymbol{\alpha} : 0 < \|\boldsymbol{\alpha}\|_1 \leq p\}^2$ for some $p \geq 1$, with the univariate polynomials taken to be monomials. This choice causes sample equivalence (6) to enforce *moment-matching* up to order p. For instance, setting $p = 2$ corresponds to matching the sample *mean* and *covariance* across the weighted and push-forward ensembles,

$$\frac{1}{J} \sum_{i=1}^{J} \nabla \phi_{\mathbf{s}}(X_t^{(i)}) = \sum_{i=1}^{J} w_t^{(i)} X_t^{(i)} \quad \text{and} \quad \frac{1}{J} \sum_{i=1}^{J} \nabla \phi_{\mathbf{s}}(X_t^{(i)}) (\nabla \phi_{\mathbf{s}}(X_t^{(i)}))^\top = \sum_{i=1}^{J} w_t^{(i)} X_t^{(i)} (X_t^{(i)})^\top,$$

and results in affine transport maps. Indeed, Kuang and Tabak [23] note that when the feature functions consist of first and second moments, the optimal map $\nabla \phi_{\mathbf{s}^*}$ takes the form of an optimal transport map between two Gaussians, which can equivalently be viewed as a symmetrized ensemble Kalman update. Thus in the Bayesian setting this choice of features recovers what may be viewed as a deterministic form of ensemble Kalman inversion (EKI) [13, 21].

Within our tempered sample-driven OT framework with total-order polynomial feature bases we have the opportunity to enforce matching of even *higher* moments than the mean and covariance matrix, yielding a family of algorithms which can be seen as generalizations of this deterministic EKI.

3.3.2 Kernels

As an alternative to polynomials, we can take the feature functions to be *kernels* placed at each ensemble member, $f_j(x) = K(x, X_t^{(j)})$, $j = 1, \dots J$, for some positive-definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The sample-equivalence we enforce under this choice is

$$\frac{1}{J} \sum_{i=1}^{J} K(\nabla \phi_{\mathbf{s}}(X_t^{(i)}), X_t^{(j)}) = \sum_{i=1}^{J} w_t^{(i)} K(X_t^{(i)}, X_t^{(j)}), \quad j = 1, \dots, j. \tag{10}$$

Equation (10) can be interpreted as requiring that a kernel density estimate (KDE) of the push-forward of $\{X_t^{(i)}\}_{i=1}^J$ under $\nabla \phi$ match the importance-weighted KDE of π_{X_t} at each of $X_t^{(1)}, \dots X_t^{(J)}$, or equivalently requiring that the Monge–Ampère equations be satisfied for all functions in the finite-dimensional RKHS spanned by $\{K(\cdot, X_t^{(1)}), \dots, K(\cdot, X_t^{(J)})\}$.

Using the single Newton step approximation (9), the iteration represented in Algorithm 2 is

$$X_{t+\Delta t}^{(j)} = X_t^{(j)} - \Big(\nabla_1 K(X_t^{(j)}, X_t^{(1)}) \quad \cdots \quad \nabla_1 K(X_t^{(j)}, X_t^{(J)}) \Big) M_t^{-1} \sum_{k=1}^{J} \left(\tfrac{1}{J} - w_t^{(k)} \right) \begin{pmatrix} K(X_t^{(k)}, X_t^{(1)}) \\ \vdots \\ K(X_t^{(k)}, X_t^{(J)}) \end{pmatrix},$$

$$j = 1, \dots, J, \quad t \in [0, 1], \quad \{X_0^{(j)}\}_{j=1}^J \overset{\text{i.i.d.}}{\sim} \pi_0, \tag{11}$$

[2] We exclude $P_{\mathbf{0}} \equiv 1$ because expectations of constant functions are the same across all probability measures.

with $w_t^{(k)} = \frac{(\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{\Delta t}}{\sum_{i=1}^J (\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}}$ and $M_t \in \mathbb{R}^{J \times J}$ defined

$$(M_t)_{\ell,m} = \frac{1}{J} \sum_{i=1}^{J} \langle \nabla_1 K(X_t^{(i)}, X_t^{(\ell)}),\, \nabla_1 K(X_t^{(i)}, X_t^{(m)}) \rangle, \quad \ell, m = 1, \ldots, J.$$

Combining a kernel feature basis with the updating scheme (9) yields a curious interacting particle system (11) somewhat resembling Stein variational gradient descent (SVGD) [26]—with the major distinction that (11) is to be run for unit time, while SVGD is an infinite-time iteration. We explore this connection further by examining the *continuous-time* and *mean-field* limits of (11) in Section 4.

4 Continuous-time and mean-field limits

In this section we consider the continuous-time ($\Delta t \to 0$) and mean-field ($J \to \infty$) limits of the particle system (11), with proofs deferred to Appendix A.2. We begin with the continuous-time limit.

Theorem 1 *In the limit as $\Delta t \to 0$, Equation* (11) *approaches the ODE*

$$\begin{aligned} \dot{X}_t^{(j)} = \Big(\nabla_1 K(X_t^{(j)}, X_t^{(1)}) \quad \cdots \quad \nabla_1 K(X_t^{(j)}, X_t^{(J)}) \Big)\, M_t^{-1} \cdot \\ \frac{1}{J} \sum_{k=1}^{J} \left(\log \frac{\pi_1}{\pi_0}(X_t^{(k)}) - \frac{1}{J} \sum_{i=1}^{J} \log \frac{\pi_1}{\pi_0}(X_t^{(i)}) \right) \begin{pmatrix} K(X_t^{(k)}, X_t^{(1)}) \\ \vdots \\ K(X_t^{(k)}, X_t^{(J)}) \end{pmatrix}, \end{aligned} \tag{12}$$

with $t \in [0,1]$ and $\{X_0^{(j)}\}_{j=1}^J \overset{\text{i.i.d.}}{\sim} \pi_0$.

Understanding the performance of direct discretization of the ODE (12) as an alternative to to the sample-driven OT update (11) is a subject of ongoing work.

Next we take the number of particles $J \to \infty$ and examine the mean-field limit of Equation (12)

Theorem 2 *In the limit as $J \to \infty$, the, interacting particle system* (12) *approaches the mean-field ODE*

$$\dot{X}_t^{(j)} = \mathbb{E}_{X \sim \pi_t} \left[M_{\pi_t}^{-1} \nabla_1 K(X_t^{(j)}, X) K_{\pi_t} \left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t} \left[\log \frac{\pi_1}{\pi_0} \right] \right)(X) \right], \tag{13}$$

where K_{π_t} is the kernel integral operator

$$K_{\pi_t} f(x) = \int_{\mathbb{R}^d} f(z) K(x,z)\, \mathrm{d}\pi_t(z)$$

and M_{π_t} is the integral operator

$$\begin{aligned} M_{\pi_t} f(x) &= \int_{\mathbb{R}^d} f(z) \mathbb{E}_{X_t \sim \pi_t} [\langle \nabla_1 K(X_t, x),\, \nabla_1 K(X_t, z) \rangle]\, \mathrm{d}\pi_t(z) \\ &= \iint\limits_{\mathbb{R}^d \times \mathbb{R}^d} f(z) \langle \nabla_1 K(y,x),\, \nabla_1 K(y,z) \rangle\, \mathrm{d}\pi_t(y) \mathrm{d}\pi_t(z). \end{aligned}$$

To our knowledge, these finite-particle (12) and mean-field (13) ODEs are new. Interpretation and analysis of these finite-particle and mean-field ODEs is the subject of a forthcoming paper.

5 Numerical example

To demonstrate the efficacy of our adaptive tempered transport Algorithm 2, we apply it to a two-dimensional problem in which π_1 is a Bayesian posterior $\pi_1 \propto \pi_0 \pi_\ell(y^* \mid \cdot)$, and the density ratio is $\frac{\pi_1}{\pi_0} = \pi_\ell(y^* \mid \cdot)$. We take $\pi_0 = \mathcal{N}(0, I)$ to be the prior distribution of $X_0 \in \mathbb{R}^2$ and $Y \in \mathbb{R}$ related to X_0 by $Y = \|X_0\| + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$. In our experiments we set $y^* = 2$ and $\sigma_\varepsilon = \frac{1}{2}$, and the posterior density of $X_0 \mid Y = y^*$ is the "donut"

$$\pi_1(x) \propto \exp\left(-\tfrac{1}{2} x^\top x - \tfrac{1}{2\sigma_\varepsilon^2} (y^* - \|x\|)^2 \right).$$

We apply our adaptive Algorithm 2 to approximately transport $J = 500$ samples of π_0 to π_1 in unit time. We choose feature bases $\{F_m\}_{m=1}^M$ as either total-degree Hermite polynomial bases, varying the degree $p \in \{1, \ldots, 8\}$, or Gaussian kernels centered at a random subset of the ensemble members $\{X_t^{(j)}\}_{j=1}^J$. We vary the total number of kernel basis functions in $\left\{\frac{J}{8}, \frac{2J}{8}, \ldots, \frac{7J}{8}, J\right\}$ and select the bandwidth according to the median heuristic [26]. For each feature basis setting we vary the sample equivalence error tolerance ϵ (line 9 of Algorithm 2) in $\{10^{-1}, 10^{-2}, \ldots, 10^{-8}\}$.

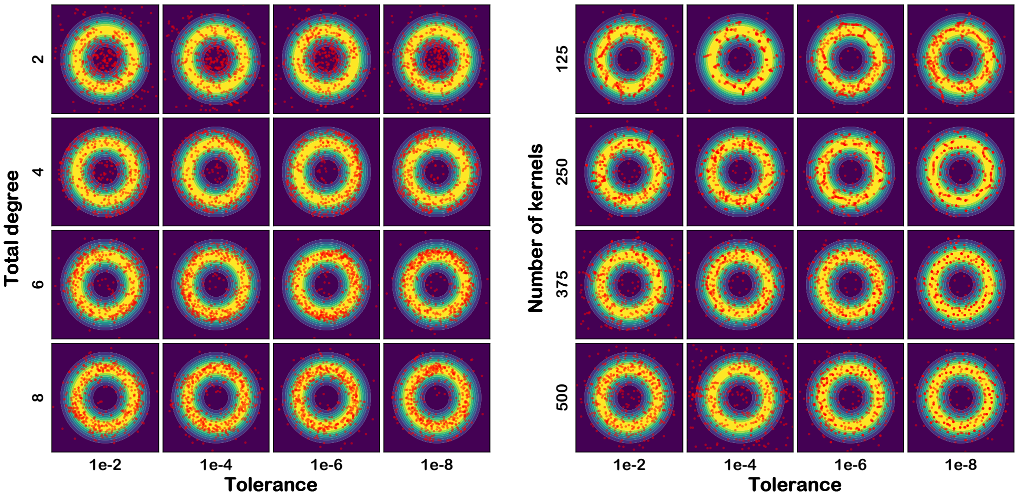

Figure 1: Ensembles at time $t = 1$ for total order polynomial feature bases of varying degree (left) and kernel feature bases of varying size (right) and varying sample-equivalence error tolerance.

In Figure 1 we show the particle ensembles at $t = 1$ overlaid atop the true density π_1 for a subset of the (feature basis, tolerance) settings considered (for complete results see Appendix A.3). For sufficiently rich feature bases and strict tolerances the samples generated by the adaptive algorithm are visually of good quality, with lower tolerances generally leading to better samples. Interestingly, for total-degree polynomial bases sample quality appears to increase with total degree up to a certain threshold (perhaps degree 6) and then decrease, perhaps as overfitting effects become present. For kernel basis functions, sample quality tends to improve with increasing numbers of kernels.

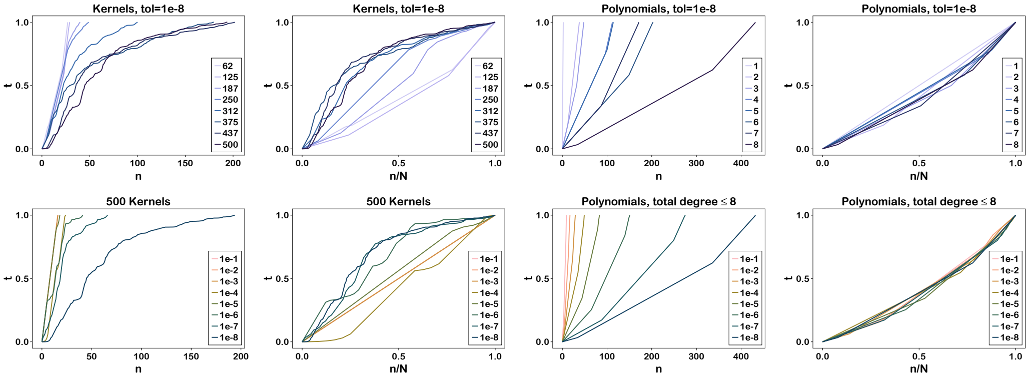

Figure 2: Examples of tempering schedules discovered by the adaptive Algorithm 2 grouped by tolerance (top row) and basis complexity (bottom row). The first and third columns show the evolution of time t with number of steps $n \in \{0, \ldots, N\}$, revealing that more complex bases and stricter tolerances tend to require smaller increments Δt. The second and fourth columns show the evolution with time t of the *normalized* step number $n/N \in [0, \ldots, 1]$ to allow for easier comparison of the "shapes" of the tempering schedules.

In Figure 2 we show the schedules—that is, the choice of number of steps N and waypoints $\{t_1, \ldots, t_{N-1}\}$—traced out by the adaptive algorithm for a selection of feature bases and tolerances. It is interesting that for low tolerances and rich bases the schedules selected by the adaptive algorithm for kernel bases tend to be concave, while those selected for polynomial bases tend to be linear or even slightly convex. Understanding the nature of these schedules is an area of ongoing investigation.

6 Future work

There are many questions surrounding the methods presented in this paper which remain to be answered. The kernel implementation of our algorithm in particular yields a curious new interacting particle system, suggesting that it may be possible to write an SVGD-like [26, 14] algorithm which samples in finite time rather than infinite time. We are working to understand the continuous-time (12) and mean-field (13) limits further and to answer questions of, e.g., consistency, optimal schedules, and performance with both finite and infinite numbers of particles.

Acknowledgments and Disclosure of Funding

AM and YM were supported by the Office of Naval Research, SIMDA (Sea Ice Modeling and Data Assimilation) MURI, award number N00014-20-1-2595 (Dr. Reza Malek-Madani and Dr. Scott Harper). AM was additionally supported by the NSF Graduate Research Fellowship under Grant No. 1745302.

A Supplementary Material

A.1 Fixed-timestep tempered, sample-driven transport algorithm

Algorithm 3 Tempered transport with sample-driven OT

```
Require: Reference ensemble $\{X_0^{(j)}\}_{j=1}^J \overset{\text{i.i.d.}}{\sim} \pi_0$, density ratio $\pi_1/\pi_0$, timestep $\Delta t \in (0,1]$,
    features $f_1, \dots, f_M : \mathbb{R}^d \to \mathbb{R}$
$N \leftarrow 1/\Delta t$
$t \leftarrow 0$
for $n = 1, \dots, N$ do
    Compute importance weights: for $j \in \{1, \dots, J\}$
        $w_t^{(j)} = \frac{(\frac{\pi_1}{\pi_0}(X_t^{(j)}))^{\Delta t}}{\sum_{i=1}^J (\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}}$
    Approximate solution to sample-driven OT with one Newton step:
        $\mathbf{s}_t = -\left(\frac{1}{J}\sum_{i=1}^J \nabla F(X_t^{(i)})\nabla F(X_t^{(i)})^\top\right)^{-1} \sum_{k=1}^J (\frac{1}{J} - w_t^{(k)}) F(X_t^{(k)})$
    Transport samples:
        $X_{t+\Delta t}^{(j)} = X_t^{(j)} + \begin{pmatrix} \nabla F_1(X_t^{(j)}) & \cdots & \nabla F_M(X_t^{(j)}) \end{pmatrix} \mathbf{s}_t$
    $t \leftarrow t + \Delta t$
end for
Ensure: $\{X_1^{(j)}\}_{j=1}^J \sim \pi_{X_1} \approx \pi_1$
```

A.2 Proofs

A.2.1 Proof of Theorem 1

Notice that time only enters the update equation (11) through the importance weights $w_t^{(k)}$. To obtain the continuous time limiting ODE we rearrange, divide by Δt on both sides, and take $\Delta t \to 0$,

$$\lim_{\Delta t \to 0} \frac{X_{t+\Delta t}^{(j)} - X_t^{(j)}}{\Delta t} =$$

$$\lim_{\Delta t \to 0} - \left(\nabla_1 K(X_t^{(j)}, X_t^{(1)}) \quad \cdots \quad \nabla_1 K(X_t^{(j)}, X_t^{(J)}) \right) M_t^{-1} \sum_{k=1}^{J} \frac{\frac{1}{J} - w_t^{(k)}}{\Delta t} \begin{pmatrix} K(X_t^{(k)}, X_t^{(1)}) \\ \vdots \\ K(X_t^{(k)}, X_t^{(J)}) \end{pmatrix}.$$

Examining the terms above involving Δt, we see that for $k \in \{1, \ldots, J\}$ we have

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{\frac{1}{J} - w_t^{(k)}}{\Delta t} &= \lim_{\Delta t \to 0} \frac{\frac{1}{J} - \frac{(\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{\Delta t}}{\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}}}{\Delta t} = - \lim_{\Delta t \to 0} \frac{\frac{(\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{\Delta t}}{\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}} - \frac{(\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{0}}{\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{0}}}{\Delta t} \\
&= - \frac{\mathrm{d}}{\mathrm{d}\Delta t} \left. \frac{(\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{\Delta t}}{\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t}} \right|_{\Delta t = 0} \\
&= - \left. \frac{(\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{\Delta t} \log \frac{\pi_1}{\pi_0}(X_t^{(k)}) \sum_{i=1}^{J} (\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t} - (\frac{\pi_1}{\pi_0}(X_t^{(k)}))^{\Delta t} \sum_{i=1}^{J} (\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t} \log \frac{\pi_1}{\pi_0}(X_t^{(i)})}{\left(\sum_{i=1}^{J}(\frac{\pi_1}{\pi_0}(X_t^{(i)}))^{\Delta t} \right)^2} \right|_{\Delta t = 0} \\
&= - \frac{J \log \frac{\pi_1}{\pi_0}(X_t^{(k)}) - \sum_{i=1}^{j} \log \frac{\pi_1}{\pi_0}(X_t^{(i)})}{J^2} = - \frac{1}{J} \left(\log \frac{\pi_1}{\pi_0}(X_t^{(k)}) - \frac{1}{J} \sum_{i=1}^{J} \log \frac{\pi_1}{\pi_0}(X_t^{(i)}) \right).
\end{aligned}$$

Hence, the ODE arising from the limit of (11) as $\Delta t \to 0$ is

$$\begin{aligned}
\dot{X}_t^{(j)} = &\left(\nabla_1 K(X_t^{(j)}, X_t^{(1)}) \quad \cdots \quad \nabla_1 K(X_t^{(j)}, X_t^{(J)}) \right) M_t^{-1} \\
&\frac{1}{J} \sum_{k=1}^{J} \left(\log \frac{\pi_1}{\pi_0}(X_t^{(k)}) - \frac{1}{J} \sum_{i=1}^{J} \log \frac{\pi_1}{\pi_0}(X_t^{(i)}) \right) \begin{pmatrix} K(X_t^{(k)}, X_t^{(1)}) \\ \vdots \\ K(X_t^{(k)}, X_t^{(J)}) \end{pmatrix}, \qquad (14)
\end{aligned}$$

with initial condition $\{X_0^{(j)}\}_{j=1}^{J} \overset{\text{i.i.d.}}{\sim} \pi_0$.

A.2.2 Proof of Theorem 2

Notice that as $J \to \infty$ and for any $x \in \mathbb{R}^d$ the sum

$$\frac{1}{J} \sum_{k=1}^{J} \left(\log \frac{\pi_1}{\pi_0}(X_t^{(k)}) - \sum_{i=1}^{J} \log \frac{\pi_1}{\pi_0}(X_t^{(i)}) \right) K(X_t^{(k)}, x), \quad X_t^{(k)} \overset{\text{i.i.d.}}{\sim} \pi_t$$

approaches the projection of $\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}[\log \frac{\pi_1}{\pi_0}]$ onto the reproducing kernel Hilbert space (RKHS) defined by K with respect to π_t, evaluated at x,

$$K_{\pi_t} \left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}[\log \frac{\pi_1}{\pi_0}] \right)(x) \equiv \int_{\mathbb{R}^d} (\log \frac{\pi_1}{\pi_0}(z) - \mathbb{E}_{\pi_t}[\log \frac{\pi_1}{\pi_0}]) K(z, x) \mathrm{d}\pi_t(z).$$

Hence, as $J \to \infty$ the vector

$$\frac{1}{J} \sum_{k=1}^{J} \left(\log \frac{\pi_1}{\pi_0}(X_t^{(k)}) - \sum_{i=1}^{J} \log \frac{\pi_1}{\pi_0}(X_t^{(i)}) \right) \begin{pmatrix} K(X_t^{(k)}, X_t^{(1)}) \\ \cdots \\ K(X_t^{(k)}, X_t^{(J)}) \end{pmatrix}$$

can be replaced by the function $x \mapsto K_{\pi_t}\left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}[\log \frac{\pi_1}{\pi_0}]\right)(x)$.

Similarly, as $J \to \infty$, M_t can be viewed as a kernel $M_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$M_t(z, z') = \mathbb{E}_{X_t \sim \pi_t} \langle \nabla_1 K(X_t, z),\, \nabla_1 K(X_t, z') \rangle,$$

which can be applied to functions on $\mathbb{R}^d$ as a convolution-type operator with respect to π_t,

$$M_{\pi_t} f(x) = \int_{\mathbb{R}^d} f(z) M_t(x, z) \, \mathrm{d}\pi_t(z) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(z) \langle \nabla_1 K(y, x),\, \nabla_1 K(y, z) \rangle \, \mathrm{d}\pi_t(y) \mathrm{d}\pi_t(z).$$

$M_{\pi_t}^{-1}$ is the inverse operator to M_{π_t}. We see in (12) that we have a choice in whether we view $M_{\pi_t}^{-1}$ as acting on $\nabla K(X_t^{(j)}, \cdot)$ or on $K_{\pi_t}\left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}[\log \frac{\pi_1}{\pi_0}]\right)$. Thus the mean-field limit of (12) can be written

$$\begin{aligned} \dot{X}_t^{(j)} \overset{J \to \infty}{\to} & \int_{\mathbb{R}^d} M_{\pi_t}^{-1} \nabla_1 K(X_t^{(j)}, x) K_{\pi_t}\left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}\left[\log \frac{\pi_1}{\pi_0}\right]\right)(x) \, \mathrm{d}\pi_t(x) \\ &= \mathbb{E}_{X \sim \pi_t}\left[M_{\pi_t}^{-1} \nabla_1 K(X_t^{(j)}, X) K_{\pi_t}\left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}\left[\log \frac{\pi_1}{\pi_0}\right]\right)(X)\right], \end{aligned}$$

or equivalently

$$\dot{X}_t^{(j)} \overset{J \to \infty}{\to} \mathbb{E}_{X \sim \pi_t}\left[\nabla_1 K(X_t^{(j)}, X) M_{\pi_t}^{-1} K_{\pi_t}\left(\log \frac{\pi_1}{\pi_0}(\cdot) - \mathbb{E}_{\pi_t}\left[\log \frac{\pi_1}{\pi_0}\right]\right)(X)\right].$$

A.3 Additional numerical results

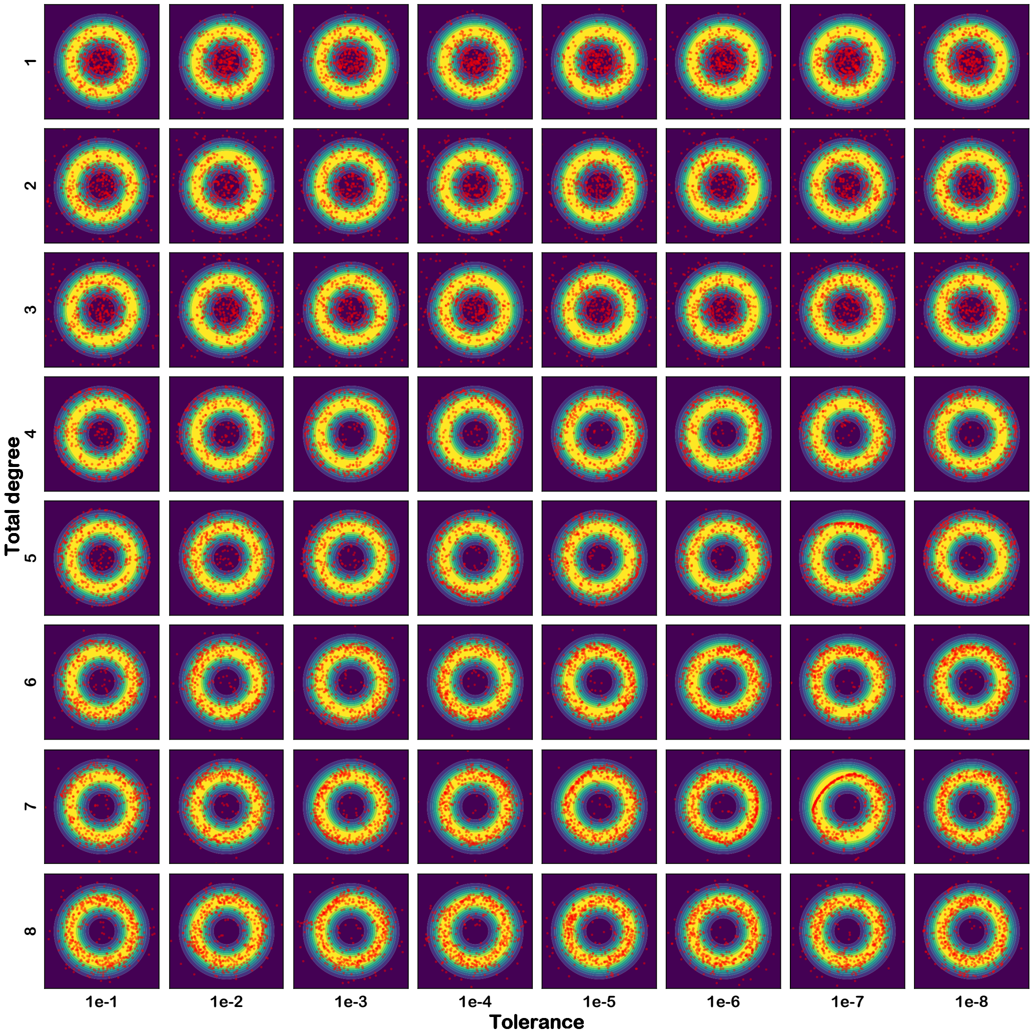

Figure 3: Ensembles at time $t = 1$ for total order polynomial feature bases of varying degree and sample-equivalence error tolerances.

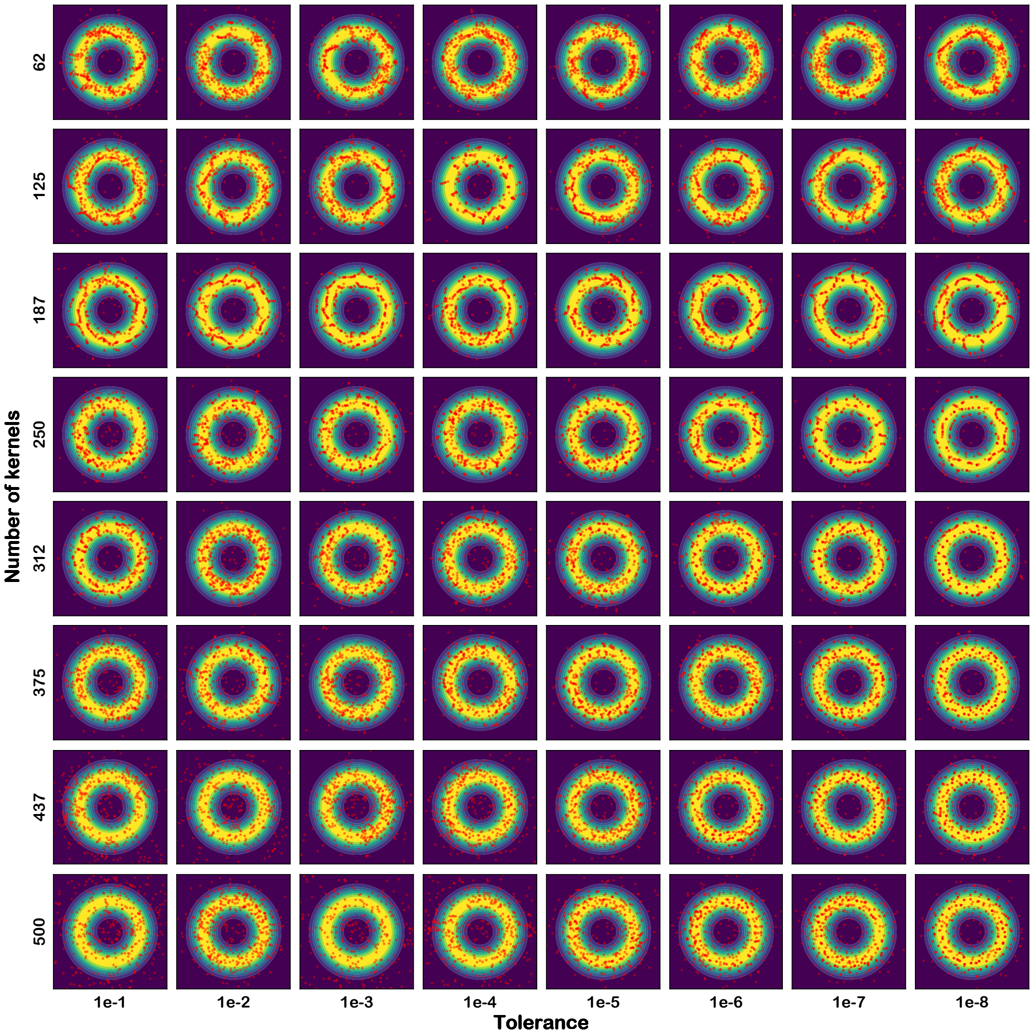

Figure 4: Ensembles at time $t = 1$ for kernel feature bases with varying number of kernels and sample-equivalence error tolerance.

References

[1] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden, “Stochastic Interpolants: A Unifying Framework for Flows and Diffusions,” no. arXiv:2303.08797, Mar. 2023.

[2] R. Baptista, B. Hosseini, N. B. Kovachki, and Y. Marzouk, “Conditional Sampling with Monotone GANs: From Generative Models to Likelihood-Free Inference,” no. arXiv:2006.06755, Jun. 2023.

[3] R. Baptista, Y. Marzouk, and O. Zahm, “On the representation and learning of monotone triangular transport maps,” *Foundations of Computational Mathematics*, vol. in press, 2023, arXiv:2009.10303.

[4] R. Brekelmans, V. Masrani, T. Bui, F. Wood, A. Galstyan, G. V. Steeg, and F. Nielsen, “Annealed Importance Sampling with q-Paths,” no. arXiv:2012.07823, Dec. 2020.

[5] Y. Brenier, “Polar factorization and monotone rearrangement of vector-valued functions,” *Communications on pure and applied mathematics*, vol. 44, no. 4, pp. 375–417, 1991.

[6] C. Bunne, A. Krause, and M. Cuturi, “Supervised training of conditional monge maps,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 6859–6872, 2022.

[7] P. Chen, K. Wu, J. Chen, T. O’Leary-Roseberry, and O. Ghattas, “Projected stein variational newton: A fast and scalable bayesian inference method in high dimensions,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart, “Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance,” Jul. 2023.

[9] B. Dai and U. Seljak, “Sliced Iterative Normalizing Flows,” in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 2352–2364.

[10] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, “Diffusion schrödinger bridge with applications to score-based generative modeling,” in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 17 695–17 709. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/940392f5f32a7ade1cc201767cf83e31-Abstract.html

[11] P. Del Moral, A. Doucet, and A. Jasra, “Sequential monte carlo samplers,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 68, no. 3, pp. 411–436, 2006.

[12] B. M. Dia, “A Continuation Method in Bayesian Inference,” *SIAM/ASA Journal on Uncertainty Quantification*, pp. 646–681, Jun. 2023.

[13] Z. Ding and Q. Li, “Ensemble Kalman inversion: Mean-field limit and convergence analysis,” *Statistics and Computing*, vol. 31, no. 1, p. 9, Jan. 2021.

[14] A. Duncan, N. Nüsken, and L. Szpruch, “On the geometry of Stein variational gradient descent,” *Journal of Machine Learning Research*, vol. 24, no. 56, pp. 1–39, 2023.

[15] D. J. Earl and M. W. Deem, “Parallel tempering: Theory, applications, and new perspectives,” *Physical Chemistry Chemical Physics*, vol. 7, no. 23, pp. 3910–3916, 2005.

[16] L. C. Evans, “Partial differential equations and monge-kantorovich mass transfer,” *Current developments in mathematics*, vol. 1997, no. 1, pp. 65–126, 1997.

[17] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, “Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler,” *SIAM Journal on Applied Dynamical Systems*, vol. 19, no. 1, pp. 412–441, Jan. 2020.

[18] A. Garbuno-Inigo, N. Nüsken, and S. Reich, “Affine Invariant Interacting Langevin Dynamics for Bayesian Inference,” *SIAM Journal on Applied Dynamical Systems*, Jul. 2020.

[19] C. J. Geyer, “Markov chain monte carlo maximum likelihood,” 1991.

[20] J. Heng, V. De Bortoli, and A. Doucet, “Diffusion Schr\"{o}dinger Bridges for Bayesian Computation,” no. arXiv:2308.14106, Aug. 2023.

[21] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, “Ensemble Kalman methods for inverse problems,” *Inverse Problems*, vol. 29, no. 4, p. 045001, Apr. 2013.

[22] I. Kobyzev, S. J. Prince, and M. A. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.

[23] M. Kuang and E. G. Tabak, “Sample-Based Optimal Transport and Barycenter Problems,” *Communications on Pure and Applied Mathematics*, vol. 72, no. 8, pp. 1581–1630, 2019.

[24] H. R. Künsch, “Recursive Monte Carlo filters: Algorithms and theoretical analysis,” *The Annals of Statistics*, vol. 33, no. 5, pp. 1983 – 2021, 2005. [Online]. Available: https://doi.org/10.1214/009053605000000426

[25] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” no. arXiv:2210.02747, 2023. [Online]. Available: http://arxiv.org/abs/2210.02747

[26] Q. Liu and D. Wang, “Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm,” in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[27] X. Liu, C. Gong, and Q. Liu, “Flow straight and fast: Learning to generate and transfer data with rectified flow,” no. arXiv:2209.03003, 2022. [Online]. Available: http://arxiv.org/abs/2209.03003

[28] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, “Sampling via measure transport: An introduction,” *Handbook of uncertainty quantification*, vol. 1, p. 2, 2016.

[29] R. M. Neal, “Annealed importance sampling,” *Statistics and computing*, vol. 11, pp. 125–139, 2001.

[30] A. B. Owen, *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

[31] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, “Normalizing flows for probabilistic modeling and inference,” *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 2617–2680, 2021.

[32] M. Ramgraber, R. Baptista, D. McLaughlin, and Y. Marzouk, “Ensemble transport smoothing–part 2: nonlinear updates,” *arXiv preprint arXiv:2210.17435*, 2022.

[33] S. Reich, “A dynamical systems framework for intermittent data assimilation,” *BIT Numerical Mathematics*, vol. 51, no. 1, pp. 235–249, Mar. 2011.

[34] ——, “A Nonparametric Ensemble Transform Method for Bayesian Inference,” *SIAM Journal on Scientific Computing*, vol. 35, no. 4, pp. A2013–A2024, Jan. 2013. [Online]. Available: http://epubs.siam.org/doi/10.1137/130907367

[35] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[36] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, “Obstacles to high-dimensional particle filtering,” *Monthly Weather Review*, vol. 136, no. 12, pp. 4629–4640, 2008.

[37] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *International conference on learning representations*, 2021.

[38] A. Spantini, D. Bigoni, and Y. Marzouk, “Inference via low-dimensional couplings,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2639–2709, 2018.

[39] A. Spantini, R. Baptista, and Y. Marzouk, “Coupling Techniques for Nonlinear Ensemble Filtering,” *SIAM Review*, vol. 64, no. 4, pp. 921–953, Nov. 2022.

[40] S. Syed, V. Romaniello, T. Campbell, and A. Bouchard-Côté, "Parallel tempering on optimized paths," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 033–10 042.

[41] A. Taghvaei and B. Hosseini, "An Optimal Transport Formulation of Bayes' Law for Nonlinear Filtering Algorithms," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, Dec. 2022, pp. 6608–6613.

[42] N. G. Trillos, B. Hosseini, and D. Sanz-Alonso, "From Optimization to Sampling Through Gradient Flows," Feb. 2023.

[43] F. Vargas, W. Grathwohl, and A. Doucet, "Denoising diffusion samplers," *arXiv preprint arXiv:2302.13834*, 2023.

[44] C. Villani, *Topics in Optimal Transportation*. American Mathematical Soc., Aug. 2021.

[45] C. Xu, X. Cheng, and Y. Xie, "Optimal transport flow and infinitesimal density ratio estimation," no. arXiv:2305.11857. [Online]. Available: http://arxiv.org/abs/2305.11857