Learning Paths for Dynamic Measure Transport: A Control Perspective

Aimee Maurais
Massachusetts Institute of Technology
Cambridge, MA 02139
maurais@mit.edu

Bamdad Hosseini
University of Washington
Seattle, WA 98195
bamdadh@uw.edu

Youssef Marzouk
Massachusetts Institute of Technology
Cambridge, MA 02139
ymarz@mit.edu

Abstract

We bring a control perspective to the problem of identifying paths of measures for sampling via dynamic measure transport (DMT). We highlight the fact that commonly used paths may be poor choices for DMT and connect existing methods for learning alternate paths to mean-field games. Based on these connections we pose a flexible family of optimization problems for identifying tilted paths of measures for DMT and advocate for the use of objective terms which encourage smoothness of the corresponding velocities. We present a numerical algorithm for solving these problems based on recent Gaussian process methods for solution of partial differential equations and demonstrate the ability of our method to recover more efficient and smooth transport models compared to those which use an untilted reference path.

1 Introduction

Sampling from a target probability distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ is a fundamental task in modern machine learning, enabling, e.g., uncertainty quantification in Bayesian inference [16] and generation of convincing synthetic data [8, 29, 13]. Many recent sampling algorithms are grounded in a *dynamic measure transport (DMT)* framework, which typically makes use of a stochastic differential equation (SDE)

$$\mathrm{d}X_t = v(X_t, t)\,\mathrm{d}t + \sigma \mathrm{d}W_t, \quad t \in [0,T], \quad X_0 \sim \eta, \tag{1}$$

where $v : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is the *drift* or *velocity*, $\sigma \geq 0$ is a fixed noise level, W_t is white noise, η is a reference measure, and $T > 0$ is a stopping time. Broadly speaking, the goal is to design the dynamics (1) such that $X_T \sim \pi$. In practice, due to limitations of data and computation, we ask for an approximate process $\widehat{X}_t$ such that $\mathrm{Law}(\widehat{X}_T) \approx \pi$. This can often be cast as a learning problem for an approximate drift $\widehat{v} \approx v$. With $\widehat{v}$ in hand, we can generate approximate samples from π by simulating (1) with $\widehat{v}$ to transform samples from η into approximate samples from π.

The SDE (1) (an ODE for $\sigma = 0$) induces a *path of distributions* $(\rho(t))_{t \in [0,T]}$, where $\rho(t) = \mathrm{Law}(X_t)$, satisfying $\rho(0) = \eta$ and $\rho(T) = \pi$. In some DMT approaches, such as neural ODEs and continuous normalizing flows [10, 17], this path is implicit or of little concern, but in more recent methods, such as diffusion models or stochastic interpolants [30, 1, 26, 9, 34], the path is explicit and at the heart of the methodology. In these latter methods the drift $\widehat{v}$ is identified not only such that $\mathrm{Law}(\widehat{X}_T) \approx \pi$, *but*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Frontiers in Probabilistic Inference: Sampling Meets Learning.

also such that $\mathrm{Law}(\widehat{X}_t) \approx \rho(t)$*, for all* $t \in [0, T]$. As ρ and v must jointly satisfy a Fokker–Planck equation (FPE) corresponding to (1), the entire problem of DMT can be cast as one of approximately solving the FPE; some recent techniques are based precisely on this idea [31, 26, 25].

2 Good and bad paths of measures

In this article we consider the following question:

> *Can we identify a problem-dependent path of densities $\rho(t)$ for which an associated drift v and sample trajectories X_t can be well approximated?*

Our motivation stems from the fact that some DMT approaches can be used with virtually *any* tractable path of measures so long as the required "ingredients" for approximating v are available. For instance, stochastic interpolants [1, 2] use paths given by the law of a random variable interpolation which can be constructed rather arbitrarily. Likewise, density-driven DMT approaches often use the geometric annealing path between η and π, but there are some, e.g., [25, 26, 31, 34], that could, in principle, be used with any path of measures with an accessible log-derivative. Within these flexible frameworks it is not often clear which paths are best, especially given that canonical paths like the McCann interpolant [27] are typically intractable. The current practice in DMT approaches that allow a choice of path is seemingly to choose one which is easy to write down: in stochastic interpolants [2, 23, 24] the default path corresponds to a linear interpolation between reference and target random variables, and in density-driven settings practitioners tend to employ the geometric annealing path.

2.1 Issues with the geometric annealing path

The geometric annealing path, given by $\mu(t) \propto \eta^{1-t}\pi^t$, $t \in [0, 1]$ is convenient for density-driven DMT because it has a log-derivative which is independent of normalizing constants. Moreover, it possesses Fisher–Rao gradient flow structure [14, 12] and variational characterizations (e.g., [4, Theorem 4.9]). It may, however, be problematic for DMT with certain combinations of η and π. This issue was, to our knowledge, first highlighted in Máté and Fleuret [25]. We demonstrate this phenomenon via the example $\eta = \mathcal{N}(0, 1)$ and $\pi = \frac{2}{3}\mathcal{N}(-8, 1) + \frac{1}{3}\mathcal{N}(4, 1)$ in the top row of Figure 1. The evolution of $\mu(t)$ is dominated by transport from η to the closest mode of π until $t \approx 0.8$, at which point "teleportation of mass" from the lesser to the greater mode begins. Capturing this teleportation with DMT is difficult; the velocity we identify by numerically solving the FPE (see Section 4) almost completely fails to place samples in the left mode. The physics-informed neural network (PINN) approach used in Máté and Fleuret [25] faced similar difficulties with analogous examples. Even if an algorithm could learn a velocity achieving transport along $\mu(t)$ for this (η, π), such a velocity would be large and irregular; see Chemseddine et al. [9] for results in this vein.

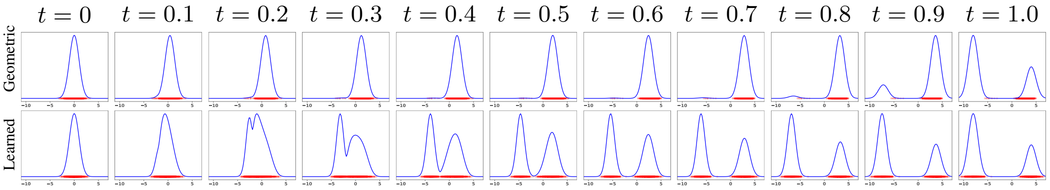


Figure 1: Geometric annealing path (top) and path resulting from solving our proposed control problem (9) (bottom) for the example $\eta = \mathcal{N}(0, 1)$ and $\pi = \frac{2}{3}\mathcal{N}(-8, 1) + \frac{1}{3}\mathcal{N}(4, 1)$. Samples generated by the respective velocity fields are plotted overtop in red.

2.2 A fix and an explanation

The approach taken in Máté and Fleuret [25] to correct teleportation behavior of $\mu(t)$ is to add a *perturbation* $f : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$ to the log of the geometric mixture,

$$\log \mu^f(\cdot, t) = (1 - t) \log \eta(\cdot) + t \log \pi(\cdot) + t(1 - t) f(\cdot, t) - \log Z(t), \tag{2}$$

where $Z(t) = \int_{\mathbb{R}^d} \eta^{1-t}\pi^t e^{t(1-t)f(\cdot,t)}\,\mathrm{d}x$ is the normalizing constant. The interpolation (2) ensures $\mu^f(0) = \eta$ and $\mu^f(1) = \pi$ and corresponds to a tilting $\mu^f(\cdot,t) \propto \mu(\cdot,t)e^{t(1-t)f(\cdot,t)}$ of μ. In [25], f is learned alongside a velocity field v by minimizing a PINN loss corresponding to the continuity equation for ODE transport along the path (2). This optimization problem is strongly ill-posed—there are infinitely many fs one could use in (2), and even for fixed f there are infinitely many valid velocities v. Yet, remarkably, the f and v [25] recovers are quite well-behaved [25, Figure 8].

Obtaining a nice path by minimizing a PINN loss over neural networks is not a given; in replicating the results of [25] we found that considerable tuning was necessary. This behavior and the ill-posedness of the underlying optimization problem suggest that implicit regularization is occurring. In fact, the interpolation (2) can alternately be grounded in an *explicit* regularization approach. Many generative models which make use of DMT can be identified with solutions of *mean-field games* (MFGs), which are infimizations of structured cost functionals over paths of measures ρ and drifts v jointly satisfying a FPE [35]. A particular MFG which fits into the framework of Zhang and Katsoulakis [35] is

$$\inf_{v,\rho}\left\{D_{\mathrm{KL}}(\rho(1)\|\pi) + \int_0^1 (1-t)D_{\mathrm{KL}}(\rho(t)\|\eta) + tD_{\mathrm{KL}}(\rho(t)\|\pi)\,\mathrm{d}t + \int_0^1 \mathbb{E}_{\rho(t)}[L(X_t, v(X_t,t))]\,\mathrm{d}t\right\}$$
$$\text{s.t. } \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho v) = 0, \quad \rho(0) = \eta. \quad (3)$$

In (3), the terminal cost $D_{\mathrm{KL}}(\rho(1)\|\pi)$ encourages $\rho(1) \approx \pi$, while $\int_0^1 \mathbb{E}_{\rho(t)}[L(X_t, v(X_t,t))]\,\mathrm{d}t$ is an action cost used to penalize v; a typical choice is $L(x,v) = \frac{1}{2}|v|^2$. We choose *interaction costs* $\mathcal{I}_t(\rho) = (1-t)D_{\mathrm{KL}}(\rho\|\eta) + tD_{\mathrm{KL}}(\rho\|\pi)$, $t \in [0,1]$, because they are minimized by $\mu(t) \propto \eta^{1-t}\pi^t$ [4]. Thus, the solution $\rho(t)$ to (3) will be close to $\mu(t)$ to the extent that it does not incur large action costs. The optimality conditions for (3) imply that

$$\log\rho(\cdot,t) = (1-t)\log\eta(\cdot) + t\log\pi(\cdot) - \frac{\partial U(\cdot,t)}{\partial t} + H(\cdot,\nabla U(\cdot,t)) - c(t), \quad (4)$$

i.e., $\rho(t)$ is a tilting of the geometric mixture $\mu(t)$, similar to the model posed in (2). In (4), U is the value function and H is the Hamiltonian; see Appendix B for details.

3 Path identification via regularization

Given the surprising performance of the learned interpolation approach [25] and its connection to mean-field games [35] or related control problems, we propose to identify tilted paths of measures

$$\log\rho^g(x,t) = \log\rho^{\mathrm{ref}}(x,t) + g(x,t) - \log Z(t),$$

and corresponding velocity fields v for ODE transport by solving control problems of the form

$$\inf_{v\in\mathcal{V},g\in\mathcal{G}} \|v\|_{\mathcal{V}}^2 + \lambda_g\|g\|_{\mathcal{G}}^2 \quad \text{s.t. } -\nabla\cdot(v\rho^g) = \rho^g(\partial_t\log\rho^g), \quad \rho^g \propto \rho^{\mathrm{ref}}e^g, \quad g(\cdot,0) = g(\cdot,1) \equiv 0. \quad (5)$$

In the above, $\rho^{\mathrm{ref}} : [0,1] \to \mathcal{P}(\mathbb{R}^d)$ is a reference path of measures (such as $\mu(t)$), $g : [0,1]\times\mathbb{R}^d \to \mathbb{R}$ is a perturbation taken in a Banach space $\mathcal{G}$, $Z(t) = \int_{\mathbb{R}^d} \rho^{\mathrm{ref}}(x,t)e^{g(x,t)}\,\mathrm{d}x$ is the normalizing constant, and $\lambda_g > 0$ is a regularization parameter. We additionally take $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ in a Banach space $\mathcal{V}$. We justify the formulation (5) as follows:

- Parametrizing ρ^g as a tilting is tractable and expressive. Tilted measures are already used to obtain diffusion-based samplers from unnormalized densities via stochastic optimal control [36, 18, 6, 33], and fine-tuning of diffusion models is frequently cast as one of sampling from a tilting of the distribution of the base model (e.g., [15]).
- Equation (5) captures a wider range of penalties on v than those that arise in MFGs [35]. Note that if we take $\mathcal{V} = L^2([0,1],V)$ to be a Bochner space, where V is an appropriate Banach space, we obtain a penalty $\|v\|_{\mathcal{V}}^2 = \int_0^1 \|v(\cdot,t)\|_V^2\,\mathrm{d}t$ akin to the action cost in a MFG (3). Action costs in MFGs, however, must be of the form $\int_0^1 \mathbb{E}_{\rho_t}[L(X_t, v_t(X_t))]\,\mathrm{d}t$, precluding the use of, e.g., Sobolev or reproducing kernel Hilbert space (RKHS) norms to regularize v. Recent works suggest that *smoothness* plays an important role in convergence of learned DMT models [7, 32], and we argue that it is important to capture this explicitly.

- The constraints in (5) enforce $\rho^g(1) = \pi$ rather than encouraging $\rho^g(1) \approx \pi$ via a terminal cost. Our formulation may thus be better able to capture the regularization phenomena occurring in the approach of Máté and Fleuret [25] and is also more relevant to other DMT approaches which enforce $\rho(1) = \pi$, such as stochastic interpolants [1, 24, 23].

3.1 Comparison to other control approaches for sampling and path identification

Several works use a stochastic optimal control (SOC) approach to construct samplers (given an unnormalized density) as solutions to a Schrödinger bridge problem [36, 18, 6, 33]. The SOC formulation, which can also be cast as a mean-field game [35], is

$$\min_{u\in\mathcal{U}} \mathbb{E}\left[\int_0^1 \tfrac{1}{2}\|u(X_t^u,t)\|^2\,\mathrm{d}t + \log\frac{\rho^{\mathrm{ref}}(\cdot,1)}{\pi}(X_1^u)\right] \quad \text{s.t.} \quad \mathrm{d}X_t^u = \sigma(t)u(X_t^u,t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t, \qquad X_0^u = 0, \tag{6}$$

where $\mathcal{U}$ is a set of allowable controls, $\sigma : [0,1] \to \mathbb{R}^{d\times d}$ is a diffusion coefficient, and $\rho_{\mathrm{ref}}(\cdot,1)$ is the $t=1$ density of the uncontrolled process, e.g.,

$$\mathrm{d}X_t = \sigma(t)\,\mathrm{d}W_t, \quad t\in[0,1], \quad X_0 = 0. \tag{7}$$

The motivation for adopting (6) is that one can show, via Girsanov's theorem, that the optimally controlled process $(X_t^{u^*})_{t\in[0,1]}$ has terminal distribution $\rho^{u^*}(\cdot,1) = \pi$. The path measure of the process $(X_t^{u^*})_{t\in[0,1]}$ is in fact the Schrödinger bridge (SB) between $\eta = \delta_0$ and π with base process (7). While we also use control in our framework (5), we seek a path of measures resulting in *smooth* dynamics, whereas the SB seeks a path of measures which is as close as possible, in KL divergence, to a reference path while satisfying desired terminal and initial conditions. When the SB problem is cast as an SOC problem, the L_2 norm of the drift is penalized, which promotes small magnitude but not necessarily smoothness, and the terminal condition is replaced with a terminal cost. Our approach also differs from (6) in that we focus on ODEs rather than SDEs; we assume that η has a density (i.e., is not a Dirac); and we use explicit boundary conditions to ensure $\rho^g(1) = \pi$.

Another related recent work is Hernandez et al. [19], which considers action-minimization problems for identifying paths between probability measures. Like our framework, [19] includes more general costs via an interaction energy term and enforces $\rho(0) = \eta$ and $\rho(1) = \pi$ via explicit boundary conditions. The motivation in [19] is to enable obstacle avoidance and to incorporate other application-specific costs in settings such as robotics, whereas our aim is principled design of DMT-based samplers. Numerically, [19] recasts the action-minimization problem as a static transport problem and lifts to a space of parametric pushforward measures, which is quite different from the dynamic PDE-constrained optimization approach we adopt here.

4 Numerical approach & experiments

Here we consider (5) with $\mathcal{V} = \mathcal{H}_v$ and $\mathcal{G} = \mathcal{H}_g$ as follows: $\mathcal{H}_g$ is a scalar-valued RKHS [5] with kernel $K_g : Y \times Y \to \mathbb{R}$, where $Y = \mathbb{R}^d \times [0,1]$, and $\mathcal{H}_v$ is a *vector-valued* RKHS [3, 21]. We take $\mathcal{H}_v$ to be curl-free and identify $v = \nabla u$, where u is an element of a scalar-valued RKHS $\mathcal{H}_u$ with kernel $K_u : Y \times Y \to \mathbb{R}$ (in our one-dimensional example this is WLOG). The problem (5) is then

$$\inf_{u\in\mathcal{H}_u, g\in\mathcal{H}_g} \|u\|_{\mathcal{H}_u}^2 + \lambda_g\|g\|_{\mathcal{H}_g}^2 \quad \text{s.t.} \; -\nabla\cdot(\rho^g\nabla u) = \rho^g(\partial_t\log\rho^g), \quad \rho^g \propto \mu e^g, \qquad g(\cdot,0) = g(\cdot,1) \equiv 0. \tag{8}$$

To solve (8) we employ the Gaussian-process PDE (GP-PDE) solution method of [11]. In brief, we enforce the PDE constraint and the boundary condition $g(\cdot,0) = g(\cdot,1) = 0$ at finite sets of collocation points on the interior and boundary of X. Representer theorems for u and g [28] simplify $\|g\|_{\mathcal{H}_g}^2$ and $\|u\|_{\mathcal{H}_u}^2$ and we relax the constraints, ultimately obtaining the equivalent discrete problem

$$\inf_{\substack{\mathbf{z}_u\in\mathbb{R}^{(d+1)J},\, \mathbf{c}\in\mathbb{R}^N \\ \mathbf{z}_g\in\mathbb{R}^{(d+1)J+J_b}}} \mathbf{z}_u^\top K_u(\varphi,\varphi)^{-1}\mathbf{z}_u + \lambda_g\mathbf{z}_g^\top K_g(\psi,\psi)^{-1}\mathbf{z}_g + \lambda_{\mathrm{pde}}\sum_{j=1}^{J}\left|F_j(z_j^1,\mathbf{z}_j^2,\mathbf{z}_j^3,z_j^4,\mathbf{c})\right|^2 + \lambda_{\mathrm{bc}}\sum_{j=1}^{J_b}|z_j^5|^2, \tag{9}$$

where $\lambda_{\text{pde}}, \lambda_{\text{bc}} > 0$ are regularization parameters, $\{F_j : j \in [J]\}$ encode the PDE constraint, and $\mathbf{z}_u$ and $\mathbf{z}_g$ completely parametrize the optimal u and g. We use a Levenberg-Marquardt algorithm to solve (9) with a Cholesky change-of-variables as advocated in [20]. As proof of concept, we use (8) and (9) to find a path ρ^g and velocity $v_g = \nabla u_g$ for ODE transport between $\eta = \mathcal{N}(0, 1)$ and $\pi = \frac{2}{3}\mathcal{N}(-8, 1) + \frac{1}{3}\mathcal{N}(4, 1)$, with $\rho^{\text{ref}} = \mu$. For comparison, we use a GP-PDE approach to directly compute a velocity field $v_{\text{ref}} = \nabla u_{\text{ref}}$ for transport along ρ^{ref}. Both approaches use the same collocation points and kernels; in particular, we take $K_u((x, t), (x', t')) = K_g((x, t), (x', t')) = K_x(x, x')K_t(t, t')$, where K_x and K_t are kernels on $\mathbb{R}$. See Appendices C and D for further details.

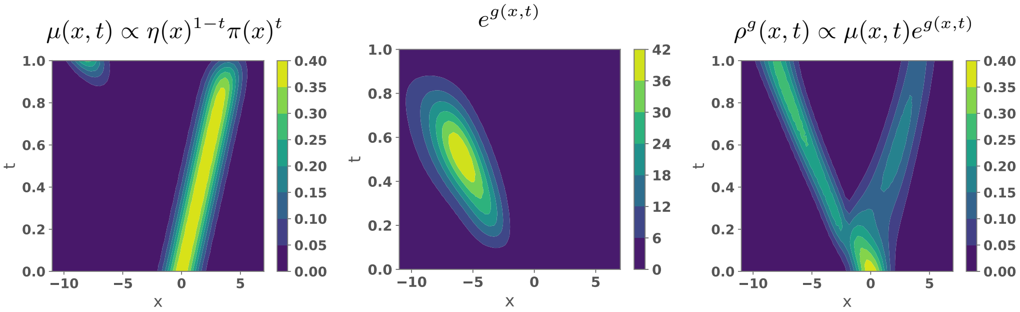


Figure 2: Space-time plots of the reference path $\mu(x, t) \propto \eta(x)^{1-t}\rho(x)^t$ (left), the tilting $e^{g(x,t)}$ (center), and the path $\rho^g(x, t) \propto \mu(x, t)e^{g(x,t)}$ resulting from (9) (right).

In Figures 1 and 2 we show the two paths, $\rho^{\text{ref}} = \mu$ and ρ^g, and in Figure 1 we show samples generated using the corresponding velocities, v_{ref} and v_g. The tilting e^g recovered from (9) eliminates the teleportation present in μ, leading to better-quality samples generated by v_g. In Figure 3 we display the trajectories of particles sampled from η and transported by v_{ref}, v_g, and the velocity corresponding to the McCann interpolant [27] (computed analytically in this 1D example). We see that, in addition to placing more samples in the left mode of π than v_{ref}, the learned velocity ∇u_g is spatially smoother than the McCann velocity. This result is similar in flavor to that of Tsimpos et al. [32, Figure 3], wherein a time-rescaling is applied to the McCann interpolant to obtain a smoother velocity field. Our approach differs from [32] in that we do not use the McCann interpolant as a starting point and that the path of densities itself, rather than just the schedule, is allowed to deviate from the reference. In Figure 5 we plot the spatial RKHS norms $\|u_g(\cdot, t)\|_{\mathcal{H}_x}$ and $\|u_{\text{ref}}(\cdot, t)\|_{\mathcal{H}_x}$, where $\mathcal{H}_x$ is the RKHS with kernel K_x, as a function of t. We see that $\|u_{\text{ref}}(\cdot, t)\|_{\mathcal{H}_x}$ increases by more than tenfold over the course of $[0, 1]$ in order to capture the teleportation in ρ^{ref}, while $\|u_g\|_{\mathcal{H}_x}$ stays relatively constant. We assess the quality of the samples generated by v_g and v_{ref} in Table 1. While v_g does not sample perfectly, it still represents a dramatic improvement over v_{ref}.

5 Conclusion

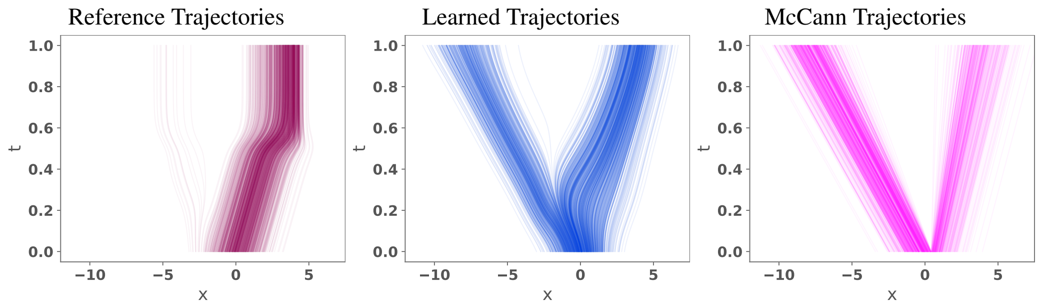
We have presented a flexible, general control framework (5) for identifying paths of measures for DMT as tiltings of accessible reference paths. Our framework enables the promotion of smoothness of the associated dynamics via penalization with, e.g., Sobolev or RKHS norms, and can serve as the basis for a range of numerical implementations. We have used one such implementation to generate proof-of-concept results demonstrating clear benefits of using a learned path with smooth dynamics, but looking ahead we are considering other formulations based on alternate functional penalties, for example, Bochner space norms on v and g. We anticipate that our framework will enable us to discrern the relative roles of spatial and temporal regularity in influencing the tractability of a given path, and ultimately inform better choices of path in sampling applications, like Bayesian inference and data assimilation, where annealing is often employed and the reference η cannot be modified.

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 and by Google through an MIT Schwarzman College of Computing Future Research Cohort Fellowship. BH was supported by the National Science Foundation grants DMS-233767 (CAREER: Gaussian Processes for Scientific Machine Learning: Theoretical Analysis and Computational Algorithms), and DMS-2208535 (Machine Learning for Bayesian Inverse Problems). AM thanks Jannis Chemseddine, Markos Katsoulakis, Gabriele Steidl, and Benjamin Zhang for helpful discussions and thanks Alexander Hsu and Juan Felipe Osorio Ramierez for help with the GP-PDE implementation.

A Additional numerical results

Here we provide additional figures and tables corresponding to the experiment in Section 4.

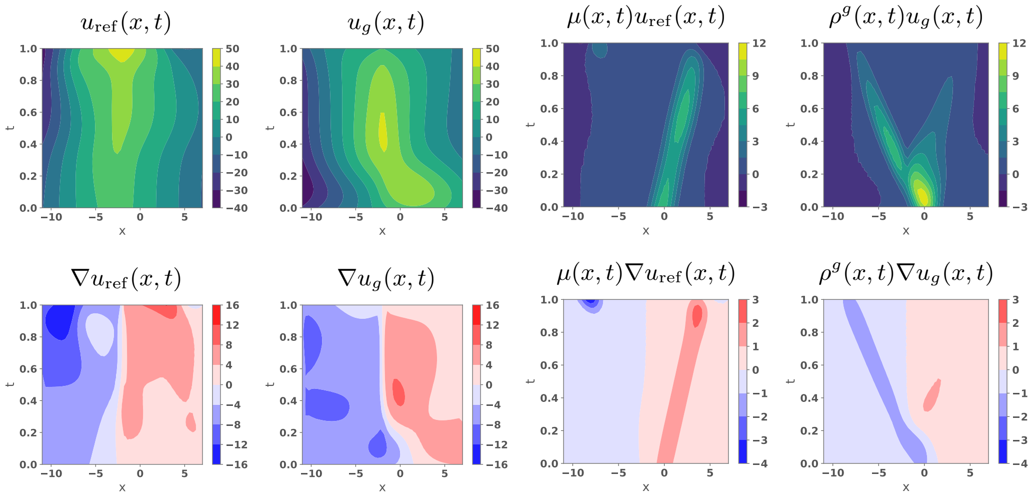


Reference Trajectories

Learned Trajectories

McCann Trajectories

Figure 3: Trajectories corresponding to three different velocity fields for DMT between η and π: the reference velocity $v_{\text{ref}} = \nabla u_{\text{ref}}$ (left), the learned velocity $v_g = \nabla u_g$ (center), and the McCann interpolant velocity (right). The learned velocity v_g places more mass in the left mode than the reference velocity v_{ref} and is spatially smoother than the McCann interpolant velocity.



$u_{\text{ref}}(x,t)$

$u_g(x,t)$

$\mu(x,t)u_{\text{ref}}(x,t)$

$\rho^g(x,t)u_g(x,t)$

$\nabla u_{\text{ref}}(x,t)$

$\nabla u_g(x,t)$

$\mu(x,t)\nabla u_{\text{ref}}(x,t)$

$\rho^g(x,t)\nabla u_g(x,t)$

Figure 4: Potentials u_{ref} and u_g and velocity fields $v_{\text{ref}} = \nabla u_{\text{ref}}$ and $v_g = \nabla u_g$ corresponding to the geometric path $\rho^{\text{ref}} = \mu$ and the path ρ^g obtained from (9). In the first two columns of panels we show the absolute potentials/velocities, and in the second two columns we show the potentials/velocities weighted by their respective probability densities, which better capture how the mass is moving.

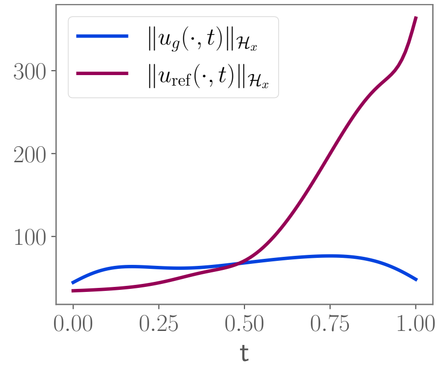


Figure 5: Spatial RKHS norms of $u_g(\cdot,t)$ (blue) and $u_{\text{ref}}(\cdot,t)$ (red) as a function of time.

	Fraction in left mode	Relative Error in Mean ↓	Relative Error in Variance ↓	MMD ↓	$\|u\|_{\mathcal{H}}$
Reference Interpolation	0.005	1.80	0.96	0.743	770.
Learned Interpolation	0.375	0.88	0.016	0.137	136
Ground Truth Samples	*0.654*	*0.040*	*0.024*	7.21×10^{-4}	*n/a*

Table 1: Quality metrics evaluated on 1000 samples generated by v_g and by v_{ref}. We evaluate the same metrics on 1000 ground-truth samples from π for comparison. In truth $2/3$ of the mass of π belongs in the left mode, the mean of π is -4 and the variance of π is 33.

B Optimality conditions for mean-field game

In Section 2.2 we introduce the mean-field game

$$\inf_{v,\rho}\left\{D_{\text{KL}}(\rho(1)\|\pi)+\int_0^1(1-t)D_{\text{KL}}(\rho(t)\|\eta)+tD_{\text{KL}}(\rho(t)\|\pi)\,\mathrm{d}t+\int_0^1\mathbb{E}_{\rho(t)}[L(X_t,v(X_t,t))]\,\mathrm{d}t\right\}$$
$$\text{s.t. }\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho v)=0,\quad \rho(0)=\eta. \tag{10}$$

The optimality conditions for this game consist of a coupled system of a Hamilton-Jacobi-Bellman equation (11) and a continuity equation ,

$$-\frac{\partial U(x,t)}{\partial t}+H(x,\nabla U(x,t))=\log\rho(x,t)-(1-t)\log\eta(x)-t\log\pi(x)+c(t) \tag{11}$$

$$\frac{\partial\rho(x,t)}{\partial_t}-\nabla\cdot(\rho(x,t)\nabla_2 H(x,\nabla U(x,t)))=0 \tag{12}$$

$$U(x,1)=b+\log\frac{\rho(\cdot,1)}{\pi}(x),\quad \rho(\cdot,0)=\eta, \tag{13}$$

where $H(x,p)=\sup_v[-p^\top v-L(x,v)]$ is the Hamiltonian[1], $U:\mathbb{R}^d\times[0,1]\to\mathbb{R}$ is the value function, $b\in\mathbb{R}$ is constant, and $c:[0,1]\to\mathbb{R}$ is a time-varying constant. Equations (11) to (13) follow from standard results in control theory [22, 35].

C GP-PDE computational approach

C.1 For the reference solution

Before describing the GP-PDE solution approach to (8), we first describe the kernel collocation approach used to approximately solve the elliptic equation

$$-\nabla\cdot(\mu(x,t)\nabla u_{\text{ref}}(x,t))=\mu(x,t)\left(\log\tfrac{\eta}{\pi}(x)-\mathbb{E}_{\mu(\cdot,t)}[\log\tfrac{\eta}{\pi}]\right), \tag{14}$$

[1] e.g., if $L(x,v)=\frac{1}{2}|v|^2$, then $H(x,p)=\frac{1}{2}|p|^2$ and $\nabla_2 H(x,p)=p$.

which recovers a velocity field ∇u_{ref} for transport along the geometric mixture $\mu(t) = \eta^{1-t}\pi^t$. This approach is used as a basis for comparison to (8) and is a building block of the approach to (8).

We denote the linear operator on the LHS of (14) by $\mathcal{L}u \coloneqq -\nabla \cdot (\mu \nabla u)$, and denote the right-hand-side of (14) by $f(x,t) \coloneqq \mu(x,t)(\log \frac{\eta}{\pi}(x) - \mathbb{E}_{\mu(\cdot,t)}[\log \frac{\eta}{\pi}])$. Thus, the PDE (14) reads $\mathcal{L}u_{\mathrm{ref}} = f$.

We only enforce the PDE (14) at a set of collocation points $\{(x_j, t_j)\}_{j=1}^J \subseteq Y$, obtaining

$$(\mathcal{L}u_{\mathrm{ref}})(x_j, t_j) = f(x_j, t_j), \quad j = 1, \ldots, J. \tag{15}$$

We now seek a solution $u_{\mathrm{ref}} : Y \to \mathbb{R}$ to (15), where $Y = \mathbb{R}^d \times [0,1]$, in the RKHS $\mathcal{H}_u$ with kernel K_u having *minimum norm*,

$$u_{\mathrm{ref}} = \operatorname*{arg\,min}_{u \in \mathcal{H}_u} \|u\|_{\mathcal{H}_u}^2 \quad \text{s.t. } (\mathcal{L}u)(x_j, t_j) = f(x_j, t_j), \quad j = 1, \ldots, J. \tag{16}$$

For $j = 1, \ldots, J$, let $\phi_j : \mathcal{H}_u \to \mathbb{R}$ denote the linear functional

$$\phi_j(u) = (\mathcal{L}u)(x_j, t_j),$$

and let $\phi = (\phi_1, \ldots, \phi_{\mathrm{J}}) : \mathcal{H}_u \to \mathbb{R}^J$ be the linear feature map comprised of $\phi_1, \ldots, \phi_J$. Denoting $\mathbf{f} = (f(x_1, t_1), \ldots, f(x_J, t_J)) \in \mathbb{R}^J$, the problem (16) reads

$$u_{\mathrm{ref}} = \operatorname*{arg\,min}_{u \in \mathcal{H}_u} \|u\|_{\mathcal{H}_u}^2 \quad \text{s.t. } \phi(u) = \mathbf{f}. \tag{17}$$

Equation (17) is an *optimal recovery problem* and has a well-known solution arising from representer theorems on RKHS (e.g., Owhadi and Scovel [28], see also Chen et al. [11], Jalalian et al. [20]), namely

$$u_{\mathrm{ref}}(\cdot) = K_u(\cdot, \phi) K_u(\phi, \phi)^{-1} \mathbf{f}. \tag{18}$$

In (18) $K_u : Y \to \mathbb{R}^{1 \times J}$ is a vector field with elements

$$K_u(\cdot, \phi) = \begin{pmatrix} K_u(\cdot,\phi)_1 & \cdots & K_u(\cdot,\phi)_J \end{pmatrix}, \quad K_u(y, \phi)_i = \phi_i(K_u(y, \cdot)), \quad i = 1, \ldots, J, \tag{19}$$

and $K_u(\phi, \phi) \in \mathbb{R}^{J \times J}$ is a symmetric matrix with entries

$$K_u(\phi, \phi)_{ij} = \phi_i(K_u(\cdot, \phi)_j), \quad i, j \in \{1, \ldots, J\}. \tag{20}$$

Moreover, the RKHS norm of the optimal recovery solution u^* is

$$\|u_{\mathrm{ref}}\|_{\mathcal{H}_u}^2 = \mathbf{f}^\top K_u(\phi, \phi)^{-1} \mathbf{f}.$$

In our experiments, we approximate the unknown expectations $\mathbb{E}_{\mu(t)}[\log \frac{\pi}{\eta}]$ appearing in (14) using quadrature, since our examples are one-dimensional.

C.2 For the control problem

Now we return to the problem in Equation (8),

$$\inf_{u \in \mathcal{H}_u, g \in \mathcal{H}_g} \|u\|_{\mathcal{H}_u}^2 + \lambda \|g\|_{\mathcal{H}_g}^2 \quad \text{s.t. } -\nabla \cdot (\rho^g \nabla u) = \rho^g (\partial_t \log \rho^g), \quad \rho^g \propto \mu e^g, \quad g(\cdot, 0) = g(\cdot, 1) \equiv 0. \tag{21}$$

Recall that $\mathcal{H}_g$ is a scalar-valued RKHS with kernel $K_g : Y \times Y \to \mathbb{R}$ and $\mathcal{H}_u$ is a scalar-valued RKHS with kernel $K_u : Y \times Y \to \mathbb{R}$. This problem is similar to (16) except that the constraint is nonlinear in u and g jointly. We proceed similarly as before, only enforcing the PDE constraint, which can be equivalently written

$$\begin{aligned} F(x,t;g,u) \equiv \log \tfrac{\pi}{\eta}(x) + \partial_t g(x,t) - \mathbb{E}_{\xi \sim \rho^g(t)} \left[\log \tfrac{\pi}{\eta}(\xi) + \partial_t g(\xi, t) \right] \\ - \langle (1-t) \nabla \log \eta(x) + t \nabla \log \pi(x) + \nabla g(x,t), \nabla u(x,t) \rangle - \Delta u(x,t) = 0, \end{aligned} \tag{22}$$

at the same finite set of points $\{(x_j, t_j)\}_{j=1}^J \subseteq \mathbb{R}^d \times [0,1]$ used for the reference method. Likewise, we enforce the boundary conditions $g(\cdot, 0) = g(\cdot, 1) \equiv 0$ at finite sets of points on the boundary, $\{(x_j^0, 0)\}_{j=1}^{J_0}$ and $\{(x_j^1, 1)\}_{j=1}^{J_1}$, obtaining

$$\inf_{u \in \mathcal{H}_u, g \in \mathcal{H}_g} \|u\|_{\mathcal{H}_u}^2 + \lambda \|g\|_{\mathcal{H}_g}^2 \quad \text{s.t.} \quad \begin{cases} F(x_j, t_j; g, u) = 0 & j \in \{1, \ldots, J\} \\ g(x_j^0, 0) = 0 & j \in \{1, \ldots, J_0\} \\ g(x_j^1, 0) = 0 & j \in \{1, \ldots, J_1\}, \end{cases} \tag{23}$$

The first set of constraints in (23) can be expanded

$$\begin{aligned} F(x_j, t_j; g, u) &= \log \tfrac{\pi}{\eta}(x_j) + \partial_t g(x_j, t_j) - C(t_j) \\ &\quad - \langle (1-t_j)\nabla \log \eta(x_j) + t_j \nabla \log \pi(x_j) + \nabla g(x_j, t_j), \nabla u(x_j, t_j)\rangle - \Delta u(x_j, t_j) = 0, \\ &\qquad j = 1, \ldots, J., \end{aligned} \tag{24}$$

where

$$C(t_j) \coloneqq \mathbb{E}_{\xi \sim \rho^g(t_j)} \left[\log \tfrac{\pi}{\eta}(\xi) + \partial_t g(\xi, t_j) \right].$$

$C(t)$ is the time-derivative of the log normalizing constant of ρ^g and is typically unknown; in our implementation we learn needed evaluations of C (at all distinct t_j in our collocation point set), which we denote by $\mathbf{c} \in \mathbb{R}^N$, simultaneously with u and g; see also Máté and Fleuret [25, Lemma 1]. Notice that the constraints (24) only depend on the values of $\partial_t g$, ∇g, ∇u, and Δu at $\{(x_j, t_j)\}_{j=1}^J$. Likewise, the boundary constraints in (23) only depend on the values of g at $\{(x_j^0, 0)\}_{j=1}^{J_0} \cup \{(x_j^1, 1)\}_{j=1}^{J_1} \equiv \{(x_j^b, t_j^b)\}_{j=1}^{J_b}$, where $J_b = J_0 + J_1$. As such we denote these values

$$\left.\begin{aligned} \partial_t g(x_j, t_j) &\coloneqq z_j^1 \in \mathbb{R} \\ \nabla g(x_j, t_j) &\coloneqq \mathbf{z}_j^2 \in \mathbb{R}^d \\ \nabla u(x_j, t_j) &\coloneqq \mathbf{z}_j^3 \in \mathbb{R}^d \\ \Delta u(x_j, t_j) &\coloneqq z_j^4 \in \mathbb{R}, \end{aligned}\right\}, \quad j = 1, \ldots, J \tag{25}$$

and

$$g(x_j^b, t_j^b) \coloneqq z_j^5 \in \mathbb{R}, \quad j = 1, \ldots J_b.$$

For brevity, we also introduce notation for the known quantities in (23),

$$\begin{gathered} \log \tfrac{\pi}{\eta}(x_j) \coloneqq \ell_j \in \mathbb{R} \\ (1-t_j)\nabla \log \eta(x_j) + t_j \nabla \log \pi(x_j) \coloneqq \mathbf{s}_j \in \mathbb{R}^d, \quad j = 1, \ldots J. \end{gathered}$$

With this notation in hand, the collocation Equations (23) and (24) can be written

$$F_j(z_j^1, \mathbf{z}_j^2, \mathbf{z}_j^3, z_j^4, \mathbf{c}) \equiv \ell_j + z_j^1 - C(t_j) - \left\langle \mathbf{s}_j + \mathbf{z}_j^2, \mathbf{z}_j^3 \right\rangle - z_j^4 = 0, \quad j \in \{1, \ldots, J\}, \tag{26}$$

and

$$z_j^5 = 0, \quad j \in \{1, \ldots, J_b\}. \tag{27}$$

Thus, fulfilling the constraints of (23) consists in identifying suitable values of z_j^1, $\mathbf{z}_j^2$, $\mathbf{z}_j^3$, and z_j^4, $j \in \{1, \ldots, J\}$, z_j^5, $j \in \{1, \ldots, J_b\}$, and $\mathbf{c} \in \mathbb{R}^N$. Therefore we return to (23), replacing the constraints with the collocation Equations (26) and (27) and obtaining a *bilevel* optimization problem

$$\begin{aligned} \inf_{\substack{z_j^1, \mathbf{z}_j^2, \mathbf{z}_j^3, z_j^4,\, j \in [J] \\ z_j^5,\, j \in [J_b] \\ \mathbf{c} \in \mathbb{R}^N}} &\left\{ \inf_{u \in \mathcal{H}_u, g \in \mathcal{H}_g} \|u\|_{\mathcal{H}_u}^2 + \lambda \|g\|_{\mathcal{H}_g}^2 \quad \text{s.t.} \quad \begin{Bmatrix} \partial_t g(x_j, t_j) & = z_j^1 \in \mathbb{R},\ j \in [J] \\ \nabla g(x_j, t_j) & = \mathbf{z}_j^2 \in \mathbb{R}^d,\ j \in [J] \\ \nabla u(x_j, t_j) & = \mathbf{z}_j^3 \in \mathbb{R}^d,\ j \in [J] \\ \Delta u(x_j, t_j) & = z_j^4 \in \mathbb{R},\ j \in [J] \\ g(x_j^b, t_j^b) & = z_j^5 \in \mathbb{R},\ j \in [J_b] \end{Bmatrix} \right\} \\ &\text{s.t. } F_j(z_j^1, \mathbf{z}_j^2, \mathbf{z}_j^3, z_j^4, \mathbf{c}) = 0,\ j \in [J], \quad z_j^5 = 0, \quad j \in [J_b]. \end{aligned} \tag{28}$$

The inner problem in (28), being separable in u and g, has a solution analogous to (18),

$$u^*(\cdot) = K_u(\cdot, \varphi) K_u(\varphi, \varphi)^{-1} \mathbf{z}_u, \quad g^*(\cdot) = K_g(\cdot, \psi) K_u(\psi, \psi)^{-1} \mathbf{z}_g. \tag{29}$$

We use $\mathbf{z}_u \in \mathbb{R}^{J(d+1)}$ to denote

$$\mathbf{z}_u = \begin{pmatrix} z_1^4 & \cdots & z_J^4 & (\mathbf{z}_1^3)^\top & \cdots & (\mathbf{z}_J^3)^\top \end{pmatrix}^\top, \tag{30}$$

and $\mathbf{z}_g \in \mathbb{R}^{J(d+1)+J_b}$ to denote

$$\mathbf{z}_g = \begin{pmatrix} z_1^1 & \cdots & z_J^1 & z_1^5 & \cdots & z_{J_b}^5 & (\mathbf{z}_1^2)^\top & \cdots & (\mathbf{z}_J^2)^\top \end{pmatrix}^\top .$$

In (29) $\varphi : \mathcal{H}_u \to \mathbb{R}^{J(d+1)}$ is the linear feature map

$$\begin{aligned} \varphi(\cdot) &= \begin{pmatrix} \varphi^1(\cdot) & \cdots & \varphi^J(\cdot) & \varphi^{11}(\cdot) & \cdots & \varphi^{1d}(\cdot) & \cdots\cdots & \varphi^{J1}(\cdot) & \cdots & \varphi^{Jd}(\cdot) \end{pmatrix}^\top \\ &\equiv \begin{pmatrix} \varphi^1(\cdot) & \cdots & \varphi^J(\cdot) & \varphi^{J+1}(\cdot) & \cdots & \cdots & \cdots & \cdots & \cdots & \varphi^{J(d+1)}(\cdot) \end{pmatrix}^\top , \end{aligned} \tag{31}$$

where the component linear functionals

$$\varphi^i(u) = \Delta u(x_i, t_i), \quad \varphi^{ij}(u) = (\nabla u(x_i, t_i))_j, \quad i \in \{1, \dots, J\},\ j \in \{1, \dots, d\}, \tag{32}$$

give rise to the elements of $\mathbf{z}_u$. Similarly, $\psi : H_g \to \mathbb{R}^{J(d+1)+J_b}$ is the linear feature map

$$\begin{aligned} \psi &= \left(\psi^1, \dots, \psi^J, \psi^{\mathrm{J+1}}, \dots, \psi^{J+J_b}, \psi^{11}, \dots, \psi^{1d}, \dots, \psi^{J1}, \dots, \psi^{Jd}\right)^\top \\ &\equiv \begin{pmatrix} \psi^1 & \dots & \psi^{J+J_b} & \psi^{J+J_b+1} & \dots & \dots & \dots & \dots & \dots & \psi^{(d+1)J+J_b} \end{pmatrix}^\top , \end{aligned} \tag{33}$$

where the component linear functionals

$$\psi^i(g) = \partial_t g(x_i, t_i), \quad \psi^{ij}(u) = (\nabla g(x_i, t_i))_j, \quad i \in \{1, \dots, J\},\ j \in \{1, \dots, d\}$$
$$\psi^{J+i}(g) = g(x_i^b, t_i^b), \quad i \in \{1, \dots, J_b\}.$$

give rise to the elements of $\mathbf{z}_g$.

The vector fields $K_u : Y \to \mathbb{R}^{1 \times J(d+1)}$ and $K_g(\cdot, \psi) : Y \to \mathbb{R}^{1 \times J(d+1)+J_b}$ are defined analogously to (19), and the symmetric matrices $K(\varphi, \varphi) \in \mathbb{R}^{J(d+1) \times J(d+1)}$ and $K_g(\psi, \psi) \in \mathbb{R}^{(J(d+1)+J_b) \times (J(d+1)+J_b)}$ are defined analogously to (20).

The norm of u^* in (29) is $\|u^*\|_{\mathcal{H}_u}^2 = \mathbf{z}_u^\top K_u(\varphi, \varphi)^{-1} \mathbf{z}_u$ and the norm of g^* is $\|g^*\|_{\mathcal{H}_g}^2 = \mathbf{z}_g^\top K_g(\psi, \psi)^{-1} \mathbf{z}_g$. These norms define the optimal value of the inner problem in (28) such that the problem reduces to

$$\inf_{\substack{\mathbf{z}_u \in \mathbb{R}^{(d+1)J} \\ \mathbf{z}_g \in \mathbb{R}^{(d+1)J+J_b} \\ \mathbf{c} \in \mathbb{R}^N}} \mathbf{z}_u^\top K_u(\varphi, \varphi)^{-1} \mathbf{z}_u + \lambda \mathbf{z}_g^\top K_g(\psi, \psi)^{-1} \mathbf{z}_g \qquad \text{s.t. } F_j(z_j^1, \mathbf{z}_j^2, \mathbf{z}_j^3, z_j^4, \mathbf{c}) = 0,\ j \in [J], \quad z_j^5 = 0, \quad j \in [J_b]. \tag{34}$$

Following the relaxation approach of Chen et al. [11], in practice we exchange the constrained problem (34) for the penalized unconstrained problem

$$\inf_{\substack{\mathbf{z}_u \in \mathbb{R}^{(d+1)J} \\ \mathbf{z}_g \in \mathbb{R}^{(d+1)J+J_b} \\ \mathbf{c} \in \mathbb{R}^N}} \mathbf{z}_u^\top K_u(\varphi, \varphi)^{-1} \mathbf{z}_u + \lambda \mathbf{z}_g^\top K_g(\psi, \psi)^{-1} \mathbf{z}_g + \lambda_{\mathrm{pde}} \sum_{j=1}^{J} \left| F_j(z_j^1, \mathbf{z}_j^2, \mathbf{z}_j^3, z_j^4, \mathbf{c}) \right|^2 + \lambda_{\mathrm{bc}} \sum_{j=1}^{J_b} |z_j^5|^2. \tag{35}$$

Problems of the form (35) can be solved via Gauss-Newton or Levenberg-Marquardt algorithms; we take the approach of Jalalian et al. [20, Appendix C.2] and employ Levenberg-Marquardt with Cholesky changes of variables $\mathbf{w}_u = L_u^{-1} \mathbf{z}_u$ and $\mathbf{w}_g = L_g^{-1} \mathbf{z}_g$, where $K_u(\varphi, \varphi) = L_u L_u^\top$ and $K_g(\psi, \psi) = L_g L_g^\top$ are the Cholesky factorizations of $K_u(\varphi, \varphi)$ and $K_g(\psi, \psi)$.

D Experimental details

In the experiment of Section 4, our collocation points $\{(x_j, t_j)\}_{j=1}^J$ are the tensor-product of a uniform spatial grid over the interval $[-2s - 3,\ s + 3]$ and a uniform time grid over the interval $[0, 1]$. We take $N_x = 50$ spatial points and $N_t = 51$ time points, for a total of $J = N_x N_t = 2550$ space-time collocation points. Additionally, the boundary points $\{(x_j^b, t_j^b)\}_{j=1}^{J_b}$ are the tensor product between the same uniform spatial grid and $\{0, 1\}$ for a total of $J_b = 2N_x$ boundary points.

We take $K_u((x,t),(x',t')) = K_g((x,t),(x',t')) = K_x(x,x')K_t(t,t')$, where K_x and K_t are SPD kernels on $\mathbb{R}$. We choose K_x and K_t to both be Matern kernels,

$$K(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x-x'\|}{\sigma}\right) K_\nu \left(\sqrt{2\nu}\frac{\|x-x'\|}{\sigma}\right),$$

where Γ is the Gamma function and K_ν is the modified Bessel function of the second kind. We take the smoothness $\nu = 5/2$. We set the lengthscale of K_t to be $\sigma_t = 1/\sqrt{N_t}$ and the lengthscale of K_x to be $\sigma_x = 180/Nx$. We initialize the unknowns in Equation (9) at $\mathbf{z}_u = \mathbf{0} \in \mathbb{R}^{J(d+1)}$, $\mathbf{z}_g = \mathbf{0} \in \mathbb{R}^{J(d+1)+J_b}$, and $\mathbf{c} = \mathbf{0} \in \mathbb{R}^N$. For the first few iterations of optimization we dynamically adjust the regularization parameters to balance the terms of the loss, ultimately settling on $\lambda_g = 51.8$, $\lambda_{\text{pde}} = 2.63 \times 10^5$, and $\lambda_{\text{bc}} = 6.01 \times 10^4$.

The ensembles appearing in Figure 1 and for which the metrics in Table 1 were computed consist of 1000 particles each and were generated using the forward Euler method with a uniform step-size $\Delta t = 0.01$.

We make use of the implementation of the GP-PDE approach provided by Jalalian et al. [20]. All experiments were run on one Nvidia A100 GPU, although they could be feasibly run on a standard CPU (e.g., on a laptop) as well.

References

[1] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden, “Stochastic Interpolants: A Unifying Framework for Flows and Diffusions,” Mar. 2023.

[2] M. S. Albergo and E. Vanden-Eijnden, “Building Normalizing Flows with Stochastic Interpolants,” in The Eleventh International Conference on Learning Representations, Sep. 2022.

[3] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, “Kernels for Vector-Valued Functions: A Review,” Foundations and Trends® in Machine Learning, vol. 4, no. 3, pp. 195–266, Jun. 2012.

[4] S.-i. Amari, Information geometry and its applications. Springer, 2016, vol. 194.

[5] A. Berlinet and C. Thomas-Agnan, Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media, 2011.

[6] J. Berner, L. Richter, and K. Ullrich, “An optimal control perspective on diffusion-based generative modeling,” Transactions on Machine Learning Research, Oct. 2023.

[7] E. Beyler and F. Bach, “Convergence of Deterministic and Stochastic Diffusion-Model Samplers: A Simple Analysis in Wasserstein Distance,” Aug. 2025.

[8] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, “A survey on generative diffusion models,” IEEE transactions on knowledge and data engineering, vol. 36, no. 7, pp. 2814–2830, 2024.

[9] J. Chemseddine, C. Wald, R. Duong, and G. Steidl, “Neural Sampling from Boltzmann Densities: Fisher-Rao Curves in the Wasserstein Geometry,” Oct. 2024.

[10] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural Ordinary Differential Equations,” in Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc., 2018.

[11] Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart, “Solving and learning nonlinear PDEs with Gaussian processes,” Journal of Computational Physics, vol. 447, p. 110668, Dec. 2021.

[12] 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.

[13] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” IEEE signal processing magazine, vol. 35, no. 1, pp. 53–65, 2018.

[14] C. Domingo-Enrich and A.-A. Pooladian, “An Explicit Expansion of the Kullback-Leibler Divergence along its Fisher-Rao Gradient Flow,” Transactions on Machine Learning Research, Mar. 2023.

[15] C. Domingo-Enrich, M. Drozdzal, B. Karrer, and R. T. Q. Chen, “Adjoint Matching: Fine-tuning Flow and Diffusion Generative Models with Memoryless Stochastic Optimal Control,” Jan. 2025.

[16] R. Ghanem, D. Higdon, H. Owhadi et al., Handbook of uncertainty quantification. Springer New York, 2017, vol. 6.

[17] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, “FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models,” in International Conference on Learning Representations, Sep. 2018.

[18] A. Havens, B. K. Miller, B. Yan, C. Domingo-Enrich, A. Sriram, B. Wood, D. Levine, B. Hu, B. Amos, B. Karrer, X. Fu, G.-H. Liu, and R. T. Q. Chen, “Adjoint Sampling: Highly Scalable Diffusion Samplers via Adjoint Matching,” May 2025.

[19] S. G. Hernandez, P. Chen, and H. Zhou, “PDPO: Parametric Density Path Optimization,” May 2025.

[20] Y. Jalalian, J. F. O. Ramirez, A. Hsu, B. Hosseini, and H. Owhadi, “Data-Efficient Kernel Methods for Learning Differential Equations and Their Solution Operators: Algorithms and Error Analysis,” Apr. 2025.

[21] H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren, “Operator-valued Kernels for Learning from Functional Response Data,” Journal of Machine Learning Research, vol. 17, no. 20, pp. 1–54, 2016.

[22] J.-M. Lasry and P.-L. Lions, “Mean field games,” Japanese Journal of Mathematics, vol. 2, no. 1, pp. 229–260, Mar. 2007.

[23] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow Matching for Generative Modeling,” in The Eleventh International Conference on Learning Representations, Sep. 2022.

[24] X. Liu, C. Gong, and Q. Liu, “Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow,” in The Eleventh International Conference on Learning Representations, Sep. 2022.

[25] B. Máté and F. Fleuret, “Learning Interpolations between Boltzmann Densities,” May 2023.

[26] A. Maurais and Y. Marzouk, “Sampling in Unit Time with Kernel Fisher-Rao Flow,” in Proceedings of the 41st International Conference on Machine Learning. PMLR, Jul. 2024, pp. 35 138–35 162.

[27] R. J. McCann, “A Convexity Principle for Interacting Gases,” Advances in Mathematics, vol. 128, no. 1, pp. 153–179, Jun. 1997.

[28] H. Owhadi and C. Scovel, Optimal Recovery Splines, ser. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2019, p. 154–159.

[29] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, “Normalizing flows for probabilistic modeling and inference,” Journal of Machine Learning Research, vol. 22, no. 57, pp. 1–64, 2021.

[30] 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, Oct. 2020.

[31] J. Sun, J. Berner, L. Richter, M. Zeinhofer, J. Müller, K. Azizzadenesheli, and A. Anandkumar, “Dynamical Measure Transport and Neural PDE Solvers for Sampling,” Jul. 2024.

[32] P. Tsimpos, R. Zhi, J. Zech, and Y. Marzouk, “Optimal scheduling of dynamic transport,” in Proceedings of Thirty Eighth Conference on Learning Theory, ser. Proceedings of Machine Learning Research, N. Haghtalab and A. Moitra, Eds., vol. 291. PMLR, 30 Jun–04 Jul 2025, pp. 5441–5505. [Online]. Available: https://proceedings.mlr.press/v291/tsimpos25a.html

[33] F. Vargas, W. S. Grathwohl, and A. Doucet, “Denoising Diffusion Samplers,” in The Eleventh International Conference on Learning Representations, Sep. 2022.

[34] L. Wang and N. Nüsken, “Measure transport with kernel mean embeddings,” Sep. 2024.

[35] B. J. Zhang and M. A. Katsoulakis, “A mean-field games laboratory for generative modeling,” Oct. 2023.

[36] Q. Zhang and Y. Chen, “Path Integral Sampler: A stochastic control approach for sampling,” Mar. 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract succinctly summarizes the contributions of our paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This is a workshop paper and the results are preliminary. The conclusion discusses avenues for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: While we do not have any formal theorems/proofs, the mathematical connections we draw are well-explained and justified.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithms and hyperparameters are detailed in Section 4 and Appendices C and D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: While our code is not ready to be released at this time, we will provide a link to a repository containing code to reproduce the experiments in the camera-ready version.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix D

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our algorithms and their initializations are deterministic and thus error bars are not needed. For testing the performance of our learned velocities in sampling, we used a large enough ensemble (1000 samples in one dimension) that any statistical errors are very small.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No human subjects were involved in this research. This reseach does not rely on any datasets.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The methods proposed are primarily concerned with sampling from probability measures known through their densities, and as such the potential for direct societial impacts is low.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the paper corresponding to the code we used for the GP-PDE implementation and will also acknowledge the creators in the camera-ready version.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: New code will be released with the camera-ready version.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used in any aspect of the methods of this paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.