Continuum-marginal optimal transport: A mesh-free kernel method

Anonymous authors
Paper under double-blind review

Abstract

In this paper we study *continuum-marginal optimal transport*. Given a time-continuous family of probability marginals, the problem is to recover the minimum-energy velocity field whose flow reproduces every marginal. This problem is the continuum limit of the classical two-marginal Benamou–Brenier formulation, and also the deterministic limit of the Nelson problem of stochastic optimal transport. We propose a practical mesh-free solver for this problem. The weak continuity equation is embedded in a reproducing kernel Hilbert space, yielding a sample-only objective that requires no spatial discretization. The velocity is parametrized by any linear-in-parameters dictionary or neural network, and is optimized by mini-batch stochastic methods. Synthetic experiments confirm that the method achieves accurate drift recovery and marginal consistency. The same computational framework also applies to the stochastic Nelson problem.

1 Introduction

The Monge optimal transport problem, in its classical form, consists of finding a deterministic map that sends a source probability distribution to a target probability distribution with minimum cost. Benamou and Brenier Benamou & Brenier (2000) gave a dynamic reformulation of this problem, in which the static map is replaced by a time-dependent velocity field: one seeks a pair of a velocity and a density which minimizes the kinetic energy, subject to the continuity equation and to the given distributions as endpoint constraints. The optimal velocity field generates deterministic ODE paths that realize the Monge map, and no stochasticity is involved. This two-marginal dynamic formulation has been studied extensively in computational optimal transport.

From two marginals to a continuum of marginals. In many applications, the evolving density is observed not only at the endpoints but also at *all* intermediate times. Examples include particle image velocimetry, crowd monitoring, and density-functional molecular dynamics, in which the density is observed continuously while individual trajectories are not tracked. Motivated by such situations, we consider the following generalization of the Benamou–Brenier problem. Let $\mu = \{\mu_t\}_{0\le t\le T}$ be a given continuous family of probability measures on $\mathbb{R}^d$. We seek a velocity field u that minimizes

$$\int_0^T \int_{\mathbb{R}^d} |u(t,x)|^2 \, p(t,x)\, dx\, dt \tag{1}$$

over all u for which the continuity equation

$$\partial_t p + \operatorname{div}(up) = 0$$

holds, together with $p(t,\cdot) = d\mu_t/dx$ for every $t \in [0,T]$. We call (1) the *continuum-marginal optimal transport* problem, following the terminology of Mikami Mikami (2021).[1] As observed by Mikami Mikami (2021),

[1]For brevity, in the rest of the paper we also use the shorthand *all-time optimal transport* (all-time OT), which reflects Mikami's own phrasing "marginals at all times". The shorthand reads more naturally when the term recurs many times in the body. We retain the formal name *continuum-marginal* in the title, abstract and keywords, where its descriptive precision is needed, and emphasize that the two refer to the same problem.

this problem arises as the continuum limit of the classical two-marginal Benamou–Brenier formulation, obtained when the marginal constraint is imposed at all times $t \in [0, T]$, not only at the endpoints. When only μ_0 and μ_T are prescribed, (1) reduces to the classical W_2^2 cost. When the full continuum of marginals is prescribed, the feasible set is smaller, and the optimal velocity contains dynamical information that cannot be recovered from the endpoints alone. Mikami Mikami (2021) established existence and uniqueness of the minimizer as well as the gradient structure $u^*(t, x) = \nabla_x \psi(t, x)$. The optimal paths are ODE solutions and can be regarded as *continuous-time Monge maps*. Problem (1) is also the small-noise limit ($\sigma \to 0$) of the *Nelson problem* Nelson (1985); Mikami (1990; 2021). The Nelson problem is the stochastic analogue of (1), in which the ODE $\dot{X}_t = u(t, X_t)$ is replaced by the Itô SDE $dX_t = u\, dt + \sigma\, dW_t$ and the continuity equation is replaced by the Fokker–Planck equation, under the same marginal constraints $X_t \sim \mu_t$ for every $t \in [0, T]$. The Nelson problem is also called a Schrödinger bridge with marginals fixed at all times. Hence (1) serves as a deterministic counterpart that connects Monge transport and the Schrödinger bridge.

Related work. The two-marginal Benamou–Brenier formulation Benamou & Brenier (2000) is solved on a fixed Eulerian grid by augmented Lagrangian or proximal methods Papadakis et al. (2014). Flow matching Lipman et al. (2023) and stochastic interpolants Albergo & Vanden-Eijnden (2023) learn an ODE velocity from source and target distributions by neural networks, and minibatch OT variants Tong et al. (2024) improve scalability. These methods treat only the *two-marginal* setting and do not impose intermediate marginal constraints.

Discrete multi-marginal approaches. Several approaches are available when marginals are observed at a finite set of time points. Waddington-OT (WOT) Schiebinger et al. (2019) chains entropic OT couplings between consecutive snapshots and reconstructs the drift by barycentric projection. The multi-marginal Monge formulation of Nakano & Saito (2026) parametrizes $K-1$ deterministic transport maps and penalizes an MMD distance between the pushed-forward and observed marginals. Like the present paper, the method of Nakano & Saito (2026) seeks a minimum-energy velocity field, but under a finite number of marginal constraints. Albergo et al. Albergo et al. (2024) extended stochastic interpolants to a multimarginal generative model: given $K+1$ densities $\rho_0, \ldots, \rho_K$, they defined a barycentric interpolant on the K-simplex and learned a velocity field via conditional-expectation regression. Their method gives *a* velocity that reproduces the marginals, but it does not minimize the kinetic energy. It therefore addresses a different problem from the one considered here. In contrast, the present paper imposes a continuum of marginal constraints and explicitly minimizes the Benamou–Brenier energy, recovering the unique Monge-optimal velocity field. In our experiments, we compare our method with Nakano & Saito (2026) and with WOT. These are the two existing methods that take all-time marginal data as input and produce a velocity field as output. The method of Nakano & Saito (2026) minimizes a piecewise-affine multi-marginal OT cost, and WOT applies the barycentric projection of pairwise entropic couplings. Neither method directly minimizes the Benamou–Brenier energy. A comparison with Albergo et al. (2024) would measure whether optimal-transport structure matters beyond marginal reproduction; this is a complementary question and is left to future work.

Stochastic counterpart. For the Nelson problem introduced above, Lavenant et al. Lavenant & Schiebinger (2024) developed a convex variational framework (global Waddington-OT) for stochastic trajectory inference, and proved consistency of the estimator. Their approach uses entropy regularization with respect to a Brownian reference measure, and so requires $\sigma > 0$. It does not apply to the deterministic case $\sigma = 0$. For the two-endpoint stochastic problem, Schrödinger bridge algorithms De Bortoli et al. (2021) also provide practical solutions, but these again target the stochastic setting.

Problem (1) is the natural deterministic foundation of these stochastic formulations (see (Mikami, 2021, Theorem 3.7)), but no practical numerical method has been available for it. Grid-based OT solvers scale poorly with the spatial dimension, and flow matching methods do not incorporate intermediate marginal information.

Our contribution. In this paper, we propose a practical mesh-free solver for (1). We replace the continuity-equation constraint by its weak form and embed the residual in a reproducing kernel Hilbert space (RKHS) Schölkopf & Smola (2002). The squared RKHS norm then admits a closed-form expression that involves only kernel evaluations and values of the velocity at sample points, and is estimated by a

sample average from mini-batches. The velocity $u(t,x)$ is parametrized by a linear-in-parameters dictionary or by a neural network, and is optimized by stochastic methods without any spatial discretization. The same computational framework also applies to the stochastic setting ($\sigma > 0$): the continuity equation is replaced by the Fokker–Planck equation, a Laplacian term is added to the RKHS operator, and the resulting problem is the Nelson problem. We treat this stochastic case only as a proof-of-concept extension (see Sections 3.5, 4.7); a full theoretical analysis is left for future work.

Outline. Section 2 formulates the continuum-marginal OT problem and summarizes the theoretical foundations from Mikami (2021). Section 3 derives the RKHS-based objective and provides explicit kernel formulas, with Section 3.5 treating the stochastic extension to the Nelson problem. Section 4 presents numerical experiments, beginning with the original deterministic case and then demonstrating the stochastic extension.

Notation. For a set A in a topological space, a σ-finite measure ν on A, and $p \geq 1$, we write $L^p(A;\mathbb{R}^m,\nu)$ for the space of Borel measurable functions $f\colon A \to \mathbb{R}^m$ with $\int_A |f|^p \, d\nu < \infty$; when ν is the Lebesgue measure on $\mathbb{R}^k$ we abbreviate $L^p(\mathbb{R}^k) := L^p(\mathbb{R}^k;\mathbb{R},\nu)$. For an open set $\Omega \subset \mathbb{R}^n$, $C^k(\Omega)$ denotes the space of k-times continuously differentiable functions; $C_b^k(\Omega)$ is the subspace whose derivatives up to order k are bounded; $C_0^\infty(\Omega)$ is the space of infinitely differentiable functions with compact support. We write ∇_x and Δ_x for the gradient and Laplacian in the spatial variable $x \in \mathbb{R}^d$, and $\mathrm{div}_x(\cdot)$ for the divergence. For a mean vector $m \in \mathbb{R}^d$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, $\mathcal{N}(m,\Sigma)$ denotes the d-dimensional Gaussian distribution with mean m and covariance Σ; in the one-dimensional case we write $\mathcal{N}(m,\sigma^2)$ with variance σ^2.

2 Problem formulation

Let $\mu = \{\mu_t\}_{0\leq t\leq T}$ be a given continuous family of Borel probability measures on $\mathbb{R}^d$, each admitting a density $p(t,x) = d\mu_t/dx$. We consider the ODE $\dot{X}_t = u(t,X_t)$, where the drift u belongs to

$$\mathcal{U} := \big\{ u \in L^1([0,T]\times\mathbb{R}^d;\mathbb{R}^d, p(t,x)\,dt\,dx) \;:\; \mathrm{div}_x(pu) \in L^1((0,T)\times\mathbb{R}^d) \text{ in the weak sense} \big\}.$$

The divergence condition ensures that the distributional residual of the continuity equation is a function, not merely a distribution; without some such condition the problem is not well-posed for rough drifts. Denote by $\mathcal{U}_0$ the set of admissible velocity fields: $u \in \mathcal{U}_0$ if $u \in \mathcal{U}$ and for any $\varphi \in C_b^1([0,T]\times\mathbb{R}^d)$ and $t \in [0,T]$,

$$\int_{\mathbb{R}^d} \varphi(t,x)\,p(t,x)\,dx - \int_{\mathbb{R}^d} \varphi(0,x)\,p(0,x)\,dx = \int_0^t \int_{\mathbb{R}^d} \big(\partial_s\varphi + u(s,x)^\mathsf{T}\nabla_x\varphi\big)p(s,x)\,dx\,ds. \tag{2}$$

This is the weak formulation of the continuity equation

$$\partial_t p + \mathrm{div}(up) = 0. \tag{3}$$

Our problem is:

$$V_0(\mu) := \inf_{u\in\mathcal{U}_0} \int_{[0,T]\times\mathbb{R}^d} |u(t,x)|^2\,p(t,x)\,dx\,dt. \tag{P}$$

When only two marginals μ_0 and μ_T are prescribed, (P) reduces to the *Benamou–Brenier formulation* of the W_2 optimal transport Benamou & Brenier (2000). Problem (P) can equivalently be stated over absolutely continuous paths. Indeed, we have

$$V_0(\mu) = \inf \mathbb{E}\int_0^T |\dot{X}(t)|^2\,dt$$

where the infimum is taken over all absolutely continuous $(X(t))_{0\leq t\leq T}$ such that $\mathbb{P}\circ X(t)^{-1} = \mu_t$ for all $t \in [0,T]$, since any admissible velocity–density pair (u,p) can be lifted to a pathwise ODE solution by the *superposition principle* of Ambrosio (Mikami, 2021, Theorem 3.9).

The theoretical properties of (P) are established by Mikami (Mikami, 2021, Section 3.2.2).

Proposition 1 ((Mikami, 2021, Theorem 3.8, Proposition 3.7)). *Suppose $V_0(\mu) < \infty$. Then (P) admits a unique minimizer $u^* \in \mathcal{U}_0$, determined $\mu_t(dx)\,dt$-a.e. Every optimal path satisfies*

$$X(t) = X(0) + \int_0^t u^*(s, X(s))\,ds, \quad t \in [0,T], \quad a.s. \tag{4}$$

Moreover, the minimizer has gradient structure: $u^*(t,x) = \nabla_x \psi(t,x)$ for some $\psi : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ that is locally weakly differentiable in x.

If only μ_0 and μ_T are prescribed, the minimizer of (P) gives the Benamou–Brenier velocity field, and the optimal paths are the McCann displacement interpolants McCann (1997). With the full continuum of marginals prescribed, the feasible set is strictly smaller and the optimal velocity encodes richer dynamical information that cannot be recovered from the endpoints alone.

Problem (P) arises as the zero-noise limit of the *Nelson problem* (stochastic optimal transportation with marginals fixed at all times, in Mikami's Mikami (2021) phrasing) Nelson (1985); Mikami (1990; 2021). Write $V_\varepsilon(\mu)$ for the value of the Nelson problem with diffusion coefficient $\sigma = \sqrt{\varepsilon}$. Mikami (Mikami, 2021, Theorem 3.7) shows that

$$\lim_{\varepsilon \to 0} V_\varepsilon(\mu) = V_0(\mu) \tag{5}$$

and any weak limit point of a minimizer of $V_\varepsilon(\mu)$ is a minimizer of $V_0(\mu)$. In the two-marginal case, Theorem 2.9 of Mikami (2021) further shows that the optimal drift converges to the Monge map $D\varphi(\cdot) - \mathrm{Id}$, with φ the Brenier convex potential. Thus (P) sits at the deterministic endpoint of a continuous spectrum that connects optimal transport ($\sigma = 0$) to the Schrödinger bridge ($\sigma > 0$).

3 Methods

3.1 Kernel embedding of the weak continuity equation

The weak continuity equation (2) can be rewritten as: for any $t \in [0,T]$ and $\varphi \in C_0^\infty([0,T] \times \mathbb{R}^d)$,

$$\int_0^t \int_{\mathbb{R}^d} p(s,x) \mathcal{A}^u \varphi(s,x)\,ds\,dx = \int_{\mathbb{R}^d} p(t,x)\varphi(t,x)\,dx - \int_{\mathbb{R}^d} p(0,x)\varphi(0,x)\,dx, \tag{6}$$

where

$$\mathcal{A}^u \varphi = \partial_t \varphi + u^\mathsf{T} \nabla_x \varphi.$$

Remark. In the stochastic extension ($\sigma > 0$, Section 3.5), the operator becomes $\mathcal{A}^u \varphi = \partial_t \varphi + u^\mathsf{T} \nabla_x \varphi + (\sigma^2/2)\Delta_x \varphi$. All derivations below carry through with the Laplacian term included; we present the $\sigma = 0$ case for clarity.

Let $K : \mathbb{R}^{d+1} \to \mathbb{R}$ be translation-invariant, $K(\xi, \eta) = K(\xi - \eta)$. We require the following:

(A1) $K \in C(\mathbb{R}^{d+1}) \cap L^1(\mathbb{R}^{d+1})$ is positive definite, and the induced reproducing kernel Hilbert space $\mathcal{H}$ is universal on every compact subset of $[0,T] \times \mathbb{R}^d$.

(A2) The Fourier transform $\widehat{K}$ of K satisfies

$$\int_{\mathbb{R}^{d+1}} (1 + |z|^2)\widehat{K}(z)\,dz < \infty.$$

By Bochner's theorem, (A1) implies $\widehat{K} \geq 0$, and the inversion formula

$$K(\xi) = (2\pi)^{-(d+1)/2} \int_{\mathbb{R}^{d+1}} e^{iz^\mathsf{T}\xi} \widehat{K}(z)\,dz, \qquad \xi \in \mathbb{R}^{d+1},$$

holds with $i = \sqrt{-1}$ the imaginary unit. Universality on every compact set is equivalent (for translation-invariant kernels) to $\operatorname{supp} \widehat{K} = \mathbb{R}^{d+1}$, see Sriperumbudur et al. Sriperumbudur et al. (2010). We denote the RKHS inner product and norm by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$.

We also impose the following regularity condition on the marginal flow:

(A3) $p \in C([0,T]; L^1(\mathbb{R}^d)) \cap L^\infty([0,T]\times\mathbb{R}^d)$, each $p(t,\cdot)$ is a probability density on $\mathbb{R}^d$, and $\partial_t p \in L^1((0,T)\times\mathbb{R}^d)$ in the weak sense.

Under (A3), for every $u \in \mathcal{U}$ the *residual*

$$D^u(s,x) := \partial_s p(s,x) + \operatorname{div}_x\big(p(s,x)u(s,x)\big) \tag{7}$$

belongs to $L^1((0,T)\times\mathbb{R}^d)$, since both summands do ($\partial_s p \in L^1$ by (A3), and $\operatorname{div}_x(pu) \in L^1$ by the definition of $\mathcal{U}$).

The weak continuity equation (6) is stated for the natural test class C_0^∞. The proposition below extends the test class to the entire RKHS $\mathcal{H}$, so that the residual functional can later be characterized as an element of $\mathcal{H}$ itself.

Proposition 2. *Assume* (A1), (A2) *and* (A3). *Then for any $u \in \mathcal{U}$ and any $t \in [0,T]$, the weak continuity equation* (6) *holds for all $\varphi \in C_0^\infty([0,T]\times\mathbb{R}^d)$ if and only if*

$$\int_0^t\!\!\int_{\mathbb{R}^d} p(s,x)\mathcal{A}^u\varphi\, ds\, dx = \int_{\mathbb{R}^d} p(t,x)\varphi(t,x)\, dx - \int_{\mathbb{R}^d} p(0,x)\varphi(0,x)\, dx, \quad \varphi \in \mathcal{H}. \tag{8}$$

Proof. By (A2) and (Wendland, 2005, Theorem 10.45), every $\varphi \in \mathcal{H}$ is bounded with bounded space-time partial derivatives of order one (resp. two in the $\sigma > 0$ case). Combined with (A3), the bilinear functional

$$L_t(\varphi) := \int_0^t\!\!\int_{\mathbb{R}^d} p(s,x)\mathcal{A}^u\varphi(s,x)\, ds\, dx - \int_{\mathbb{R}^d} p(t,x)\varphi(t,x)\, dx + \int_{\mathbb{R}^d} p(0,x)\varphi(0,x)\, dx \tag{9}$$

is well-defined for every $\varphi \in C_b^1([0,T]\times\mathbb{R}^d)$. Indeed, (A3) gives $p \in L^1([0,T]\times\mathbb{R}^d)$, which combined with the boundedness of $\partial_s\varphi$ yields $p\,\partial_s\varphi \in L^1$; and the defining property $u \in L^1([0,T]\times\mathbb{R}^d; \mathbb{R}^d, p\,dt\,dx)$ of $\mathcal{U}$ combined with the boundedness of $\nabla\varphi$ yields $p\,u^\mathsf{T}\nabla\varphi \in L^1([0,T]\times\mathbb{R}^d)$, since $\int |p\,u^\mathsf{T}\nabla\varphi|\, dt\, dx \le \|\nabla\varphi\|_\infty \int |u|\, p\, dt\, dx < \infty$. The spatial boundary at infinity vanishes because $p(s,\cdot) \in L^1(\mathbb{R}^d)$ for every s and φ is bounded. Hence $D^u \in L^1((0,t)\times\mathbb{R}^d)$ and the integration by parts in s and x yields

$$L_t(\varphi) \;=\; -\int_0^t\!\!\int_{\mathbb{R}^d} D^u(s,x)\,\varphi(s,x)\, ds\, dx \tag{10}$$

for every $\varphi \in C_b^1([0,T]\times\mathbb{R}^d)$.

Assume (6). Then, equivalently, by (10), $\int D^u\varphi = 0$ for every $\varphi \in C_0^\infty([0,T]\times\mathbb{R}^d)$, hence $D^u = 0$ a.e. on $(0,t)\times\mathbb{R}^d$. Substituting back into (10) gives $L_t(\varphi) = 0$ for every $\varphi \in \mathcal{H} \subset C_b^1([0,T]\times\mathbb{R}^d)$, which is exactly (8).

Conversely, assume (8), i.e. $L_t(\varphi) = 0$ for every $\varphi \in \mathcal{H}$. Extend $D^u \in L^1((0,t)\times\mathbb{R}^d)$ by zero to $\mathbb{R}^{d+1}$. For any $y \in \mathbb{R}^{d+1}$ the section $K(\cdot,y)$ belongs to $\mathcal{H}$ by the reproducing property, so applying (10) with $\varphi = K(\cdot,y)$ gives

$$\int_{\mathbb{R}^{d+1}} K(z-y)D^u(z)\, dz = 0, \quad y \in \mathbb{R}^{d+1}. \tag{11}$$

Taking the Fourier transform of (11), we get $\widehat{K}\,\widehat{D^u} = 0$ a.e. Since $D^u \in L^1(\mathbb{R}^{d+1})$, the function $\widehat{D^u}$ is continuous. This together with $\operatorname{supp}\widehat{K} = \mathbb{R}^{d+1}$ yields $\widehat{D^u} = 0$ on $\mathbb{R}^{d+1}$, hence $D^u = 0$ a.e. Inserting this into (10) for any $\varphi \in C_0^\infty([0,T]\times\mathbb{R}^d)$ yields (6). □

Remark. The argument uses $\mathcal{H}$ only through the family $\{K(\cdot,y) : y \in \mathbb{R}^{d+1}\}$. Universality enters as the characteristic-kernel property $\operatorname{supp}\widehat{K} = \mathbb{R}^{d+1}$, which makes convolution by K injective on $L^1(\mathbb{R}^{d+1})$. The argument extends to the stochastic case ($\sigma > 0$) verbatim with the operator $\mathcal{A}^u$ of (34) and the Laplacian-augmented residual; (A2) is then used in full to ensure $\mathcal{H} \hookrightarrow C_b^2$.

3.2 Reproducing-kernel residual and the relaxed problem

By the reproducing property, the linear functional L_t of (9) is itself an element of $\mathcal{H}$. Concretely, for any $u \in L^1([0,T]\times\mathbb{R}^d;\mathbb{R}^d, p(t,x)\,dt\,dx)$ and $s \in [0,T]$, define

$$H_s^u(t,x) := \int_0^s \int_{\mathbb{R}^d} p(r,y)\mathcal{A}_{r,y}^u K(t,x,r,y)\,dr\,dy - \int_{\mathbb{R}^d} p(s,y)K(t,x,s,y)\,dy + \int_{\mathbb{R}^d} p(0,y)K(t,x,0,y)\,dy,$$

where $K(t,x,s,y) = K((t,x)-(s,y))$ and $\mathcal{A}_{t,x}^u = \mathcal{A}^u$ that acts on the variable (t,x).

Theorem 1. *Suppose that* (A1), (A2) *and* (A3) *hold. Then $H_s^u \in \mathcal{H}$ for any $s \in [0,T]$ and $u \in L^1([0,T]\times\mathbb{R}^d;\mathbb{R}^d, p(t,x)\,dt\,dx)$. Moreover, $u \in \mathcal{U}_0$ if and only if $H_s^u = 0$ for every $s \in [0,T]$.*

Proof. Step (i). Since $u \in \mathcal{U}$ is in $L^1([0,T]\times\mathbb{R}^d; p\,dt\,dx)$, by Fubini there exists a Lebesgue-null set $\mathcal{N} \subset [0,T]$ such that for every $t \in [0,T]\setminus\mathcal{N}$ the section $u(t,\cdot) \in L^1(\mu_t)$, hence $u(t,\cdot) < \infty$ μ_t-almost everywhere. Fix any such $t \in [0,T]\setminus\mathcal{N}$; the values of Φ_t on the null set $\mathcal{N}$ do not affect the Bochner integral in Step (iii). We claim

$$\Phi_t(\cdot) := \int_{\mathbb{R}^d} p(t,x)\mathcal{A}_{t,x}^u K(\cdot,t,x)\,dx \in \mathcal{H}. \tag{12}$$

By Lemma 10.44 and Theorem 10.45 in Wendland (2005), for any $(t,x) \in [0,T]\times\mathbb{R}^d$ we have $\mathcal{A}_{t,x}^u K(\cdot,t,x) \in \mathcal{H}$. Let $\{X_t^{(n)}\}_{n=1}^\infty$ be i.i.d. copies of $X_t \sim \mu_t$ and define the Monte-Carlo approximation

$$\psi_n(\xi) = \frac{1}{n}\sum_{j=1}^n L(\xi; X_t^{(j)}), \quad \xi = (s,y),$$

with $L(\xi; X_t^{(j)}) = \mathcal{A}_{t,x}^u K(s,y,t,x)\big|_{x=X_t^{(j)}}$, so that $\psi_n \in \mathcal{H}$. By the strong law of large numbers,

$$\int_{\mathbb{R}^d} p(t,x)\mathcal{A}_{t,x}^u K(s,y,t,x)\,dx = \mathbb{E}\left[\mathcal{A}_{t,x}^u K(s,y,t,X_t)\right] = \lim_{n\to\infty}\psi_n(\xi), \quad \text{a.s.} \tag{13}$$

By the reproducing property, for $m,n \in \mathbb{N}$,

$$\langle \psi_m, \psi_n\rangle = \frac{1}{mn}\sum_{j=1}^m\sum_{\ell=1}^n \mathcal{A}_{t,x}^u\mathcal{A}_{t,y}^u K(t,y,t,x)\big|_{y=X_t^{(j)}, x=X_t^{(\ell)}}.$$

Denote $i = \sqrt{-1}$. For $\xi = (s,y), \eta = (t,x) \in \mathbb{R}^{d+1}$, the Fourier inversion formula

$$K(\xi-\eta) = (2\pi)^{-(d+1)/2}\int_{\mathbb{R}^{d+1}} e^{iz^\top(\xi-\eta)}\widehat{K}(z)\,dz$$

combined with (A2) allows the differentiations $\mathcal{A}_\eta^u$ and $\mathcal{A}_\xi^u$ to commute with the Fourier integral. Setting $\xi^{(j)} = (t, X_t^{(j)})$, we have

$$\langle \psi_m, \psi_n\rangle = (2\pi)^{-(d+1)/2}\int_{\mathbb{R}^{d+1}} \left(\frac{1}{m}\sum_{j=1}^m \mathcal{A}_\xi^u e^{iz^\top\xi^{(j)}}\right)\overline{\left(\frac{1}{n}\sum_{\ell=1}^n \mathcal{A}_\eta^u e^{iz^\top\xi^{(\ell)}}\right)}\,\widehat{K}(z)\,dz.$$

The second-moment hypothesis (A2) gives an integrable dominating function (the integrand is of order $|z|^2\,\widehat{K}(z)$), so by the dominated convergence theorem

$$\lim_{m,n\to\infty}\langle \psi_m, \psi_n\rangle = (2\pi)^{-(d+1)/2}\int_{\mathbb{R}^{d+1}} \left|\mathbb{E}[\mathcal{A}_\xi^u e^{iz^\top\Xi}]\right|^2\widehat{K}(z)\,dz = \mathbb{E}\left[\mathcal{A}_\xi^u\mathcal{A}_\eta^u K(\Xi,\tilde{\Xi})\right],$$

with $\Xi = (t, X_t)$ and $\tilde{\Xi} = (t, \tilde{X}_t)$ independent copies. Hence $\{\psi_n\}$ is Cauchy in $\mathcal{H}$ and converges to some $\Phi_t^* \in \mathcal{H}$. Using the reproducing property and (13),

$$\Phi_t^*(\xi) = \langle \Phi_t^*, K(\cdot, \xi)\rangle = \lim_n \langle \psi_n, K(\cdot, \xi)\rangle = \lim_n \psi_n(\xi) = \Phi_t(\xi), \quad \xi \in \mathbb{R}^{d+1},$$

so $\Phi_t = \Phi_t^* \in \mathcal{H}$, establishing (12).

Step (ii). For any $s \in [0, T]$ we claim

$$\Theta_s(\cdot) := \int_{\mathbb{R}^d} p(s, y) K(\cdot, s, y)\, dy \in \mathcal{H}. \tag{14}$$

The argument is the same as Step (i) with $\mathcal{A}_{t,x}^u$ suppressed. Take an i.i.d. sequence $\{Y_s^{(j)}\}_{j=1}^{\infty}$ with $Y_s^{(1)} \sim \mu_s$ and set $\theta_n(\xi) := (1/n)\sum_{j=1}^n K(\xi, s, Y_s^{(j)})$. Then,

$$\langle \theta_m, \theta_n\rangle = \frac{1}{mn}\sum_{j=1}^m \sum_{\ell=1}^n K\big((s, Y_s^{(j)}), (s, Y_s^{(\ell)})\big) \longrightarrow \mathbb{E}\big[K((s, Y_s), (s, \tilde{Y}_s))\big],$$

with $Y_s, \tilde{Y}_s$ independent copies of μ_s. Hence $\{\theta_n\}$ is Cauchy with limit $\Theta_s \in \mathcal{H}$.

Step (iii). We claim

$$\Psi_s(\cdot) := \int_0^s \int_{\mathbb{R}^d} p(r, y) \mathcal{A}_{r,y}^u K(\cdot, r, y)\, dy\, dr = \int_0^s \Phi_r\, dr \in \mathcal{H}. \tag{15}$$

By Step (i), $\Phi_r \in \mathcal{H}$ for a.e. $r \in [0, T]$. Since $\mathcal{H}$ is separable under (A1) (the kernel K is translation-invariant with $\widehat{K} \in L^1$, so $\mathcal{H}$ embeds into a separable space of continuous functions), the weak measurability of $r \mapsto \Phi_r$—which follows from the measurability of each evaluation $r \mapsto \langle \varphi, \Phi_r\rangle = \int p(r, y)\mathcal{A}_{r,y}^u \varphi(r, y)\, dy$ for $\varphi \in \mathcal{H}$—implies strong measurability by Pettis' theorem. For the norm estimate, write

$$\|\Phi_r\|_{\mathcal{H}} \;\le\; \int_{\mathbb{R}^d} p(r, y)\, \big\|\mathcal{A}_{r,y}^u K(\cdot, r, y)\big\|_{\mathcal{H}}\, dy \;\le\; C\int_{\mathbb{R}^d} p(r, y)\big(1 + |u(r, y)|\big) dy,$$

where C depends only on $\int(1 + |z|)\widehat{K}(z)\, dz < \infty$ via (A2) and the differential reproducing identity; the finiteness of this first-moment integral follows from (A2) by Cauchy–Schwarz, $\int(1 + |z|)\widehat{K}(z)\, dz \le \big(\int(1 + |z|^2)\widehat{K}(z)\, dz\big)^{1/2}\big(\int \widehat{K}(z)\, dz\big)^{1/2} < \infty$. By (A3) and $u \in L^1(p\, dt\, dx)$ this bound is in $L^1([0, T])$, so $\int_0^s \|\Phi_r\|\, dr < \infty$. Therefore the Bochner integral $\int_0^s \Phi_r\, dr$ defines an element of $\mathcal{H}$, and its evaluation at any ξ agrees with $\Psi_s(\xi)$ by the reproducing property applied under the integral sign. Combining Steps (i)–(iii), we get

$$H_s^u = \Psi_s - \Theta_s + \Theta_0 \in \mathcal{H}.$$

Step (iv). For every $\varphi \in \mathcal{H}$,

$$\langle \varphi, H_s^u\rangle = \int_0^s \int_{\mathbb{R}^d} p(r, y)\mathcal{A}_{r,y}^u \varphi(r, y)\, dy\, dr - \int_{\mathbb{R}^d} p(s, y)\varphi(s, y)\, dy + \int_{\mathbb{R}^d} p(0, y)\varphi(0, y)\, dy. \tag{16}$$

The three summands on the right-hand side are treated similarly. Since $\langle \varphi, \cdot\rangle$ is a bounded linear functional on $\mathcal{H}$, it commutes with Bochner integration of any integrable $\mathcal{H}$-valued map (Hytönen et al., 2016, Theorem 3.7). Applied to the map $y \mapsto p(s, y)K(\cdot, s, y)$, whose Bochner integral is $\Theta_s \in \mathcal{H}$ by Step (ii), and using the reproducing property $\langle \varphi, K(\cdot, s, y)\rangle = \varphi(s, y)$, we obtain

$$\langle \varphi, \Theta_s\rangle = \Big\langle \varphi, \int p(s, y)K(\cdot, s, y)\, dy\Big\rangle = \int p(s, y)\langle \varphi, K(\cdot, s, y)\rangle\, dy = \int p(s, y)\varphi(s, y)\, dy.$$

For the Stein-operator term, the differential form of the reproducing property (Wendland, 2005, Theorem 10.45) $\langle \varphi, \mathcal{A}_{r,y}^u K(\cdot, r, y)\rangle = \mathcal{A}_{r,y}^u \varphi(r, y)$ together with commutation of $\langle \varphi, \cdot\rangle$ with the Bochner integral gives

$$\langle \varphi, \Psi_s\rangle = \int_0^s \int p(r, y)\langle \varphi, \mathcal{A}_{r,y}^u K(\cdot, r, y)\rangle\, dy\, dr = \int_0^s \int p(r, y)\mathcal{A}_{r,y}^u \varphi(r, y)\, dy\, dr.$$

The Bochner integrability required in each step follows from the norm estimates in Steps (i)–(iii), which are uniform on $[0, s]$ under (A2) and (A3). Subtracting the Θ_s term and adding the Θ_0 term gives (16).

Step (v). The right-hand side of (16) is precisely the functional $L_s(\varphi)$ of (9), which characterizes $u \in \mathcal{U}_0$ via (8) of Proposition 2: $u \in \mathcal{U}_0$ if and only if $L_s(\varphi) = 0$ for every $\varphi \in \mathcal{H}$ and every $s \in [0, T]$. By (16) this is equivalent to $\langle \varphi, H_s^u \rangle = 0$ for every $\varphi \in \mathcal{H}$ and every s, which, since $\mathcal{H}$ is a Hilbert space, is equivalent to $H_s^u = 0$ for every $s \in [0, T]$. □

Since $u \in \mathcal{U}_0$ is equivalent to $\|H_t^u\| = 0$ for every $t \in [0, T]$, we replace the hard constraint of (P) by a quadratic penalty and consider the following minimization problem:

$$V_\lambda(\mu) := \inf_{u \in \mathcal{U}} J_\lambda(u) \tag{P_λ}$$

with

$$J_\lambda(u) = \int_0^T \int_{\mathbb{R}^d} |u(t,x)|^2 p(t,x)\, dx\, dt + \lambda \int_0^T \|H_t^u\|^2\, dt.$$

The weight $\lambda > 0$ trades off kinetic energy against admissibility. The subsequent subsections derive a sample-based estimator of the penalty $\int_0^T \|H_t^u\|^2\, dt$.

Let $\{\varepsilon_n\}_{n=1}^\infty$ and $\{\lambda_n\}_{n=1}^\infty$ be positive sequences satisfying $\lim_{n\to\infty} \varepsilon_n = 0$ and $\lim_{n\to\infty} \lambda_n = +\infty$, respectively, and take $u_n \in \mathcal{U}$ to be an ε_n-optimal solution for (P_λ) with $\lambda = \lambda_n$:

$$J_{\lambda_n}(u_n) \le V_{\lambda_n}(\mu) + \varepsilon_n. \tag{17}$$

L^2-convergence of u_n to the optimal u^* will require the following additional structural assumption on the marginal flow.

(A4) The problem (P) admits a unique minimizer $u^* \in \mathcal{U}_0$ which has gradient structure: there exists a potential $\psi \in C_b^1([0, T] \times \mathbb{R}^d)$ such that

$$u^*(t,x) = \nabla_x \psi(t,x) \qquad \text{for } p(t,x)\, dt\, dx\text{-a.e. } (t,x).$$

Remark. Assumption (A4), and the discussion below, concern the deterministic case $\sigma = 0$, which is the setting of Theorem 2. The corresponding structural assumption and convergence analysis in the stochastic Nelson setting ($\sigma > 0$) are beyond the scope of this paper; Section 4.7 presents the stochastic case only as a computational demonstration that the same estimator framework extends when the operator $\mathcal{A}^u$ is augmented with the diffusion term.

Existence, uniqueness and the gradient structure in (A4) are guaranteed by the theory of Mikami (Mikami, 2000, Theorem 2.4) (see also (Mikami, 2021, Proposition 3.7, Theorem 3.8)). Specifically, in the setting of Mikami (2000), if the marginal flow arises as the solution of the Fokker–Planck equation $\partial_t p = \Delta_x p + \mathrm{div}_x(p\, \nabla \Psi)$ for a confining potential Ψ with $\Psi \in C^4(\mathbb{R}^d; [0, \infty))$ whose second–fourth derivatives are bounded (Mikami's (A.1)–(A.3)), and if the initial density $p_0 = p(0, \cdot)$ is twice continuously differentiable with bounded derivatives (Mikami's (A.4)) and satisfies $\inf_x (|x|^2+1)^{-1} \log p_0(x) > -\infty$ and $\sup_x (|x|+1)^{-1} |\nabla \log p_0(x)| < \infty$ (Mikami's (A.5)–(A.6)), then the unique minimizer of (P) is explicitly given by $u^*(t,x) = -\nabla_x \big(\log p(t,x) + \Psi(x)\big)$, so $\psi(t,x) = -\log p(t,x) - \Psi(x)$. The regularity $\psi \in C_b^1$ holds when $\log p$ and Ψ have bounded first-order space-time derivatives; this is a mild strengthening of Mikami's linear-growth bound (Mikami, 2000, Theorem 2.2) that holds in the confining regimes used in our experiments. Crucially, (A4) does *not* require $\psi \in \mathcal{H}$: only the weaker condition $\psi \in C_b^1$ is needed, since the test function class in the weak formulation can be enlarged from $\mathcal{H}$ to C_b^1 by the same kind of density/mollification argument used to establish Proposition 2.

Finally, a small but worth-noting point on the quantifier in (A4): we phrase the gradient identity $u^* = \nabla_x \psi$ as holding $p(t,x)\, dt\, dx$-a.e. rather than *everywhere*, while Mikami's theorems above give the identity pointwise in $\{p > 0\}$. The two formulations agree on the set $\{p > 0\}$, which is the only set relevant to the variational problem (P); on $\{p = 0\}$ the value of u^* is immaterial because the kinetic energy integrand vanishes there. We adopt the a.e. formulation because it is the natural one for the $L^2(p\, dt\, dx)$ weak convergence argument in Step (iv) of the proof of Theorem 2.

Theorem 2. *Suppose that* (A1)–(A3) *hold and* $V_0(\mu) < \infty$. *Then*

$$\lim_{n\to\infty} \lambda_n \int_0^T \|H_t^{u_n}\|^2 \, dt = 0, \tag{18}$$

$$\lim_{n\to\infty} J_{\lambda_n}(u_n) = V_0(\mu), \tag{19}$$

$$\sup_{\lambda>0} \inf_{u\in\mathcal{U}} J_\lambda(u) = V_0(\mu). \tag{20}$$

If moreover (A4) *holds, then*

$$u_n \longrightarrow u^* \quad \text{in } L^2([0,T]\times\mathbb{R}^d;\mathbb{R}^d, p(t,x)\,dt\,dx), \quad \text{as } n\to\infty. \tag{21}$$

Remark (Verification of (A4) in the numerical experiments). Theorem 2 is established for $\sigma = 0$ and requires the gradient structure (A4): $u^* = \nabla_x\psi$ with $\psi \in C_b^1$. By Remark 3.2, sufficient regularity of $\nabla_x\psi$ (uniform boundedness) is guaranteed in the confining regime of Mikami (Mikami, 2021, Theorem 3.8) for $\sigma = 0$ flows. In the deterministic numerical experiments of Section 4 (Exps 1–5), the prescribed drift u^* admits an explicit C_b^1 potential—$\psi(x) = 2x$ for Exp 1, $\psi(t,x) = \pi\cos(\pi t)\,x$ for Exp 2, $\psi(x) = -\log\cosh(2x)$ for Exp 3 (with $|\nabla\psi| \le 2$), and the analogous separable forms in dimension two for Exps 4–5—so (A4) holds (in Exp 5 it holds for the x_1-component; the x_2-component is constrained only up to a divergence-free deformation). Exp 6 (EB scRNA-seq) is real data, where (A4) is not verifiable a priori. Exp 7 (stochastic Nelson, $\sigma = 1$) is presented in Section 4.7 as a computational demonstration that the estimator framework extends to $\sigma > 0$; the convergence theory of Theorem 2 is restricted to $\sigma = 0$ and does *not* apply there.

Proof. Throughout the proof, for brevity, write $\mathcal{J}_n := \int_0^T \|H_t^{u_n}\|^2\,dt \ge 0$ and $\mathcal{E}_n := \int_0^T \int |u_n|^2 p\,dx\,dt$, so $J_{\lambda_n}(u_n) = \mathcal{E}_n + \lambda_n\mathcal{J}_n$. Since $u^* \in \mathcal{U}_0$, Theorem 1 gives $H_t^{u^*} = 0$ for every t, hence $J_\lambda(u^*) = V_0(\mu)$ for every λ. Together with (17) this yields

$$J_{\lambda_n}(u_n) \le V_{\lambda_n}(\mu) + \varepsilon_n \le J_{\lambda_n}(u^*) + \varepsilon_n = V_0(\mu) + \varepsilon_n. \tag{22}$$

Step (i). Proof of (18). We follow the penalty-method argument of Nakano (Nakano, 2026, Theorem 3.1). Assume for contradiction that

$$\limsup_{n\to\infty} \lambda_n\mathcal{J}_n = 5\delta$$

for some $\delta > 0$. Then extract a subsequence along which the lim sup is attained. By (22), the sequence $\{J_{\lambda_n}(u_n)\}$ is bounded, so a further subsequence (relabelled $\bar\lambda_m = \lambda_{n_m}$, $\bar u_m = u_{n_m}$, $\bar{\mathcal{J}}_m = \mathcal{J}_{n_m}$, $\bar\varepsilon_m = \varepsilon_{n_m}$) satisfies $\lim_m J_{\bar\lambda_m}(\bar u_m) = \kappa$ for some finite $\kappa \le V_0(\mu)$ and $\lim_m \bar\lambda_m\bar{\mathcal{J}}_m = 5\delta$. Choose m_0 with $\kappa < J_{\bar\lambda_{m_0}}(\bar u_{m_0}) + \delta$, then $m_1 > m_0$ with $J_{\bar\lambda_{m_1}}(\bar u_{m_1}) < \kappa + \delta$, $\bar\lambda_{m_1} > 7\bar\lambda_{m_0}$, and $3\delta + \bar\varepsilon_{m_0} < \bar\lambda_{m_1}\bar{\mathcal{J}}_{m_1} < 7\delta$. Using (17) and the fact that u_{m_1} is feasible for $(P_{\bar\lambda_{m_0}})$,

$$\begin{aligned}
\kappa &< J_{\bar\lambda_{m_0}}(\bar u_{m_0}) + \delta \le V_{\bar\lambda_{m_0}}(\mu) + \bar\varepsilon_{m_0} + \delta \le J_{\bar\lambda_{m_0}}(\bar u_{m_1}) + \bar\varepsilon_{m_0} + \delta \\
&= \mathcal{E}_{m_1} + \frac{\bar\lambda_{m_0}}{\bar\lambda_{m_1}}\bar\lambda_{m_1}\bar{\mathcal{J}}_{m_1} + \bar\varepsilon_{m_0} + \delta < \mathcal{E}_{m_1} + \frac{\bar\lambda_{m_1}\bar{\mathcal{J}}_{m_1}}{7} + \bar\varepsilon_{m_0} + \delta \\
&< \mathcal{E}_{m_1} + 2\delta + \bar\varepsilon_{m_0} < \mathcal{E}_{m_1} + \bar\lambda_{m_1}\bar{\mathcal{J}}_{m_1} - \delta = J_{\bar\lambda_{m_1}}(\bar u_{m_1}) - \delta < \kappa,
\end{aligned}$$

a contradiction. Hence $\lim_n \lambda_n\mathcal{J}_n = 0$, proving (18).

Step (ii). Proof of (19). From (22) and (18), $\limsup_n J_{\lambda_n}(u_n) \le V_0(\mu)$.

We first argue that $\mathcal{E}_n \to V_0(\mu)$. The upper bound $\mathcal{E}_n \le J_{\lambda_n}(u_n) \le V_0(\mu) + \varepsilon_n$ gives immediately

$$\limsup_{n\to\infty} \mathcal{E}_n \;\le\; V_0(\mu). \tag{23}$$

This shows $\{u_n\}$ is bounded in the Hilbert space $L^2(p\,dt\,dx) := L^2([0,T]\times\mathbb{R}^d;\mathbb{R}^d, p(t,x)\,dt\,dx)$. For the matching lower bound, let $\{u_{n_k}\}$ be any subsequence along which $\mathcal{E}_{n_k} \to \liminf_n \mathcal{E}_n$. Since $\{u_{n_k}\}$ is still

bounded, by reflexivity we may pass to a further subsequence (relabelled u_{n_k}) with $u_{n_k} \rightharpoonup u_\infty$ weakly in $L^2(p\,dt\,dx)$. For every $\varphi \in \mathcal{H}$, Theorem 1 gives $\langle \varphi, H_t^{u_{n_k}} \rangle = L_t(\varphi; u_{n_k})$. By Cauchy–Schwarz, $|\langle \varphi, H_t^{u_{n_k}} \rangle| \le \|\varphi\|\, \|H_t^{u_{n_k}}\|$. Since $\lambda_{n_k} \to \infty$ and $\lambda_{n_k} \mathcal{J}_{n_k} \to 0$ by (18), we have $\mathcal{J}_{n_k} \to 0$, that is $\int_0^T \|H_t^{u_{n_k}}\|^2\, dt \to 0$; passing to a further subsequence (still denoted u_{n_k}) we may assume $\|H_t^{u_{n_k}}\| \to 0$ for a.e. $t \in [0,T]$, and therefore $\langle \varphi, H_t^{u_{n_k}} \rangle \to 0$ for those t. The functional $L_t(\varphi; \cdot) = \int_0^t \int p(s,x) \big(\partial_s \varphi + u(s,x)^\mathsf{T} \nabla_x \varphi\big) dx\, ds - \int p(t,x)\varphi(t,x) dx + \int p(0,x)\varphi(0,x) dx$ is, in its dependence on u, the linear form $u \mapsto \int_0^t \int p\, u^\mathsf{T} \nabla_x \varphi\, dx\, ds$. By (A2) and Wendland (Wendland, 2005, Theorem 10.45), every $\varphi \in \mathcal{H}$ has $\nabla_x \varphi \in C_b \subset L^\infty$; combined with $p \in L^1 \cap L^\infty$ from (A3) this makes $u \mapsto L_t(\varphi; u)$ a bounded linear functional on $L^2(p\,dt\,dx)$. Passing to the weak limit therefore gives $L_t(\varphi; u_\infty) = 0$ for every $\varphi \in \mathcal{H}$ and a.e. $t \in [0,T]$. Since $t \mapsto L_t(\varphi; u_\infty)$ is continuous on $[0,T]$ (by $p \in C([0,T]; L^1)$ from (A3) and the boundedness of φ, $\partial_t \varphi$, $\nabla_x \varphi$), the equality $L_t(\varphi; u_\infty) = 0$ extends from a.e. t to every $t \in [0,T]$. By Proposition 2, u_∞ satisfies the weak continuity equation (6) against every $\varphi \in C_0^\infty([0,T] \times \mathbb{R}^d)$. This extends to every $\varphi \in C_b^1$ by a standard mollification argument: for $\varphi \in C_b^1$ choose $\varphi_n \in C_0^\infty$ with $\varphi_n \to \varphi$, $\partial_t \varphi_n \to \partial_t \varphi$ and $\nabla_x \varphi_n \to \nabla_x \varphi$ pointwise, with $\sup_n (\|\varphi_n\|_\infty + \|\partial_t \varphi_n\|_\infty + \|\nabla_x \varphi_n\|_\infty) < \infty$; then $p \in L^1$ and $u_\infty \in L^1(p\,dt\,dx)$ allow dominated convergence in every term of (2). Hence $u_\infty \in \mathcal{U}_0$. By weak lower semicontinuity of the norm and (23),

$$\mathcal{E}_\infty := \int_0^T \int |u_\infty(t,x)|^2 p(t,x)\, dx\, dt \le \liminf_k \mathcal{E}_{n_k} = \liminf_n \mathcal{E}_n.$$

Since $u_\infty \in \mathcal{U}_0$, also $\mathcal{E}_\infty \ge V_0(\mu)$, so

$$V_0(\mu) \le \mathcal{E}_\infty \;\le\; \liminf_{n \to \infty} \mathcal{E}_n.$$

Together with (23) this gives $\mathcal{E}_n \to V_0(\mu)$ and u_∞ is a minimizer of (P). Combined with (18), $J_{\lambda_n}(u_n) = \mathcal{E}_n + \lambda_n \mathcal{J}_n \to V_0(\mu)$.

Step (iii). Proof of (20). Since $u^* \in \mathcal{U}_0$, Theorem 1 gives $\|H_t^{u^*}\| = 0$, so for every $\lambda > 0$,

$$\inf_{u \in \mathcal{U}} J_\lambda(u) \le J_\lambda(u^*) = V_0(\mu),$$

whence $\sup_{\lambda > 0} \inf_u J_\lambda(u) \le V_0(\mu)$. Conversely, by (17), $J_{\lambda_n}(u_n) \le \inf_u J_{\lambda_n}(u) + \varepsilon_n \le \sup_\lambda \inf_u J_\lambda(u) + \varepsilon_n$, and letting $n \to \infty$ using (19) yields $V_0(\mu) \le \sup_\lambda \inf_u J_\lambda(u)$. This proves (20).

Step (iv). L^2-convergence under (A4). Assume (A4) and let $\psi \in C_b^1([0,T] \times \mathbb{R}^d)$ with $u^* = \nabla_x \psi$. The weak limit identified in Step (ii) equals u^* by uniqueness in (A4). Since every weakly convergent subsequence of the bounded sequence $\{u_n\} \subset L^2(p\,dt\,dx)$ has this same limit, the full sequence satisfies

$$u_n \rightharpoonup u^* \qquad \text{weakly in } L^2(p\,dt\,dx; \mathbb{R}^d). \tag{24}$$

Moreover $\mathcal{E}_n \to V_0(\mu) = \int |u^*|^2 p\, dx\, dt$.

Since $\psi \in C_b^1([0,T] \times \mathbb{R}^d)$ and $p \in L^1([0,T] \times \mathbb{R}^d)$, we have $\nabla_x \psi \in L^\infty([0,T] \times \mathbb{R}^d) \cap L^2(p\,dt\,dx; \mathbb{R}^d)$, and the weak convergence (24) yields

$$\begin{aligned} \lim_{n \to \infty} \int_0^T \int_{\mathbb{R}^d} p(t,x) u_n^\mathsf{T}(t,x) \nabla_x \psi(t,x)\, dx\, dt &= \int_0^T \int_{\mathbb{R}^d} p(t,x) (u^*(t,x))^\mathsf{T} \nabla_x \psi(t,x)\, dx\, dt \\ &= \int_0^T \int_{\mathbb{R}^d} p(t,x) |u^*(t,x)|^2\, dx\, dt, \end{aligned}$$

where we used $u^* = \nabla_x \psi$ in the last equality. Combining this cross-term convergence with norm convergence $\mathcal{E}_n \to \int |u^*|^2 p\, dx\, dt$,

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^d} |u_n(t,x) - u^*(t,x)|^2 p(t,x)\, dx\, dt \\ &= \mathcal{E}_n - 2 \int_0^T \int_{\mathbb{R}^d} p(t,x) u_n^\mathsf{T}(t,x) u^*(t,x)\, dx\, dt + \int_0^T \int_{\mathbb{R}^d} p(t,x)\, |u^*(t,x)|^2\, dx\, dt \;\longrightarrow\; 0, \end{aligned}$$

which is (21). □

Remark. Step (iv) proceeds by weak convergence of $\{u_n\}$ in $L^2(p\,dt\,dx)$ (inherited from Step (ii)) rather than via the RKHS pairing $\langle \psi, H_T^{u_n} \rangle$, since the argument only requires $\nabla_x \psi \in L^2(p\,dt\,dx)$, a consequence of $\psi \in C_b^1$ and $p \in L^1$, and does not need the stronger property $\psi \in \mathcal{H}$. The role of the vanishing penalty $\|H_t^{u_n}\| \to 0$ was played earlier, in Step (ii), where it ensured that any weak limit of $\{u_n\}$ lies in $\mathcal{U}_0$.

3.3 Representations of the penalty term

We derive a computable expression for the penalty $\int_0^T \|H_t^u\|^2\,dt$ appearing in (P_λ). The reproducing property of $\mathcal{H}$ states

$$\mathcal{A}_{t,x}^u \varphi(t,x) = \langle \varphi,\, \mathcal{A}_{t,x}^u K(\cdot,(t,x)) \rangle, \quad \varphi \in \mathcal{H}.$$

Using this, the Bochner-integral identity for Hilbert-valued integrands (Hytönen et al., 2016, Proposition 1.2.4) together with the Bochner integrability established in Step (iii) of the proof of Theorem 1 yield, for any Borel measurable $Q \subset [0,T] \times \mathbb{R}^d$,

$$\Big\| \int_Q p(y) \mathcal{A}_y^u K(\cdot, y)\, dy \Big\|^2 = \int_{Q \times Q} p(y) p(y')\, \mathcal{A}_y^u \mathcal{A}_{y'}^u K(y,y')\, dy'\, dy. \tag{25}$$

Proposition 3. *Let $(X_t)_{0 \le t \le T}$ and $(\tilde{X}_t)_{0 \le t \le T}$ be independent continuous processes such that $X_t, \tilde{X}_t \sim \mu_t$ for any $t \in [0,T]$. Then, under* (A1), (A2) *and* (A3) *we have*

$$\|H_t^u\|^2 \;=\; I_1(t) + I_2(t) + I_3(t) - 2\,I_4(t) - 2\,I_5(t) + 2\,I_6(t), \quad 0 \le t \le T, \tag{26}$$

where

$$\begin{aligned}
I_1(t) &= \int_0^t \int_0^t \mathbb{E}\big[\mathcal{A}_s^u \mathcal{A}_r^u K(s, X_s, r, \tilde{X}_r)\big]\, ds\, dr, \quad & I_2(t) &= \mathbb{E}\big[K(t, X_t, t, \tilde{X}_t)\big], \\
I_3(t) &= \mathbb{E}\big[K(0, X_0, 0, \tilde{X}_0)\big], & I_4(t) &= \int_0^t \mathbb{E}\big[\mathcal{A}_s^u K(t, X_t, s, \tilde{X}_s)\big]\, ds, \\
I_5(t) &= \mathbb{E}\big[K(0, X_0, t, \tilde{X}_t)\big], & I_6(t) &= \int_0^t \mathbb{E}\big[\mathcal{A}_s^u K(0, X_0, s, \tilde{X}_s)\big]\, ds.
\end{aligned} \tag{27}$$

Proof. Decompose $H_t^u = A_1 + A_2 + A_3 \in \mathcal{H}$ where

$$\begin{aligned}
A_1 &= \int_0^t \int_{\mathbb{R}^d} p(r,x) \mathcal{A}_{r,y}^u K(\cdot, r, x)\, dx\, dr, \\
A_2 &= -\int_{\mathbb{R}^d} p(t,x) K(\cdot, t, x)\, dx, \\
A_3 &= \int_{\mathbb{R}^d} p(0,x) K(\cdot, 0, x)\, dx.
\end{aligned} \tag{28}$$

Expanding the squared norm gives

$$\|H_t^u\|^2 = \|A_1\|^2 + \|A_2\|^2 + \|A_3\|^2 + 2\langle A_1, A_2 \rangle + 2\langle A_1, A_3 \rangle + 2\langle A_2, A_3 \rangle. \tag{29}$$

We evaluate each contribution using the reproducing property. By (25) applied on $Q = [0,t] \times \mathbb{R}^d$,

$$\|A_1\|^2 = \int_0^t \int_0^t \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(s,x) p(r,y) \mathcal{A}_{s,x}^u \mathcal{A}_{r,y}^u K(s,x,r,y)\, dx\, dy\, ds\, dr \;=\; I_1(t),$$

on identifying $p(r,y)dy$ with the law of $\tilde{X}_r \sim \mu_r$. Next, a direct use of $\langle K(\cdot,y), K(\cdot,y') \rangle = K(y,y')$ yields

$$\|A_2\|^2 = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(t,x) p(t,x') K(t,x,t,x')\, dx\, dx' = I_2(t)$$

and $\|A_3\|^2 = I_3(t)$. Applying the reproducing property with $\varphi = A_1 \in \mathcal{H}$, we find

$$\begin{aligned}\langle A_1, A_2\rangle &= -\int_{\mathbb{R}^d} p(t,x')\langle A_1, K(\cdot,t,x')\rangle\, dx' = -\int_{\mathbb{R}^d} p(t,x')\, A_1(t,x')\, dx' \\ &= -\int_0^t \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p(r,x)p(t,x')\, \mathcal{A}^u_{r,y} K(t,x',r,x)\, dx\, dx'\, dr = -I_4(t).\end{aligned}$$

By the same reasoning with A_2 replaced by $-A_3$ and the time argument $t \mapsto 0$,

$$\langle A_1, A_3\rangle = \int_{\mathbb{R}^d} p(0,x')A_1(0,x')\, dx' = I_6(t).$$

Further, directly,

$$\langle A_2, A_3\rangle = -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p(t,x)p(0,x')K(t,x,0,x')\, dx\, dx' = -I_5(t).$$

Collecting all six contributions in (29) yields (26). □

Hereafter we further assume that $K(y,y') = \phi(|y-y'|)$ for some even $\phi \in C^2(\mathbb{R})$. Since ϕ' is odd, $\phi'(0) = 0$, and we define

$$\phi_1(r) := \frac{\phi'(r)}{r}, \qquad \phi_2(r) := \frac{\phi_1'(r)}{r}.$$

Both ϕ_1 and ϕ_2 extend continuously to $r = 0$. With $\sigma = 0$, the operator is $\mathcal{A}^u = \partial_t + u^\mathsf{T}\nabla_x$, and we write $\tilde{u}(y) = (1, u_1(y), \ldots, u_d(y))^\mathsf{T}$ for $y = (t,x)$. Then

$$\mathcal{A}^u_y K(y',y) = \phi_1(|y-y'|)\, \tilde{u}(y)^\mathsf{T}(y-y'), \tag{30}$$

and

$$\mathcal{A}^u_y \mathcal{A}^u_{y'} K(y,y') = -\phi_2(|y-y'|)\,(y-y')^\mathsf{T}\tilde{u}(y)\,\tilde{u}(y')^\mathsf{T}(y-y') - \phi_1(|y-y'|)\,\tilde{u}(y)^\mathsf{T}\tilde{u}(y'). \tag{31}$$

Remark. In the stochastic extension ($\sigma > 0$, Section 3.5), the operator $\mathcal{A}^u$ acquires a Laplacian term and (31) gains additional terms involving $\phi_3(r) := \phi_2'(r)/r$, $\phi_4(r) := \phi_3'(r)/r$, and powers of σ^2.

3.4 Sample estimator of the penalty

We construct a consistent (Monte Carlo) estimator of $\int_0^T \|H_t^u\|^2\, dt$ for the approximate problem. The quantities $I_2(t), I_3(t), I_5(t)$ in (27) do not depend on u and are therefore dropped when optimizing over u. The remaining u-dependent part is

$$\mathcal{Q}(u) \;:=\; \int_0^T \big[\, I_1(t) - 2\, I_4(t) + 2\, I_6(t)\,\big]\, dt. \tag{32}$$

By the change of the order of integration, the bulk contribution $\int_0^T I_1(t)\, dt$ is given by

$$\int_0^T I_1(t)\, dt = \int_0^T\int_0^t\int_0^t \mathbb{E}[\mathcal{A}^u_s \mathcal{A}^u_r K]\, ds\, dr\, dt = \int_0^T\int_0^T w(s,r)\, \mathbb{E}[\mathcal{A}^u_s \mathcal{A}^u_r K(s, X_s, r, \tilde{X}_r)]\, ds\, dr,$$

where $w(s,r) := T - s \vee r$.

Similarly, the boundary contribution $\int_0^T \big[-2I_4(t) + 2I_6(t)\big] dt$ is given by

$$\begin{aligned}\int_0^T \big[-2I_4(t) + 2I_6(t)\big] dt &= \int_0^T \Big[-2\int_0^t \mathbb{E}[\mathcal{A}^u_s K(t, X_t, s, \tilde{X}_s)]\, ds + 2\int_0^t \mathbb{E}[\mathcal{A}^u_s K(0, X_0, s, \tilde{X}_s)]\, ds\Big] dt \\ &= 2\int_0^T\int_s^T \mathbb{E}\big[\mathcal{A}^u_s \big\{K(0, X_0, s, \tilde{X}_s) - K(t, X_t, s, \tilde{X}_s)\big\}\big]\, dt\, ds.\end{aligned}$$

To compute the expectations in (32), we draw i.i.d. samples $t_1, \dots, t_M \sim \mathrm{Unif}[0,T]$ and, for each t_m, draw iid samples $X_{t_m}^{(1)}, \dots, X_{t_m}^{(N)} \sim \mu_{t_m}$. Additionally draw i.i.d. samples $X_0^{(1)}, \dots, X_0^{(N_0)} \sim \mu_0$. Write $y_{m,i} := (t_m, X_{t_m}^{(i)})$.

There are MN sample points in total, which we index as y_p, $p = 1, \dots, MN$, with $y_{(m-1)N+i} = y_{m,i}$. For two indices $p = (m-1)N + i$ and $q = (l-1)N + j$, define the weight $\hat{w}_{pq} := T - t_p \vee t_q$. Here $\mathbf{1}[A]$ denotes the indicator, equal to 1 if condition A holds and 0 otherwise. The estimator of $\mathcal{Q}(u)$ is

$$\begin{aligned}\hat{\mathcal{Q}}(u) = \frac{T^2}{M^2N^2} \sum_{\substack{p,q=1 \\ p\neq q}}^{MN} \hat{w}_{pq}\, (\mathcal{A}_y^u \mathcal{A}_{y'}^u K)(y_p,\, y_q) + \frac{2T}{MNN_0} \sum_{p=1}^{MN} \sum_{k=1}^{N_0} (T - t_p)\, \mathcal{A}_{y_p}^u\, K(z_{0,k},\, y_p) \\ - \frac{2T^2}{M^2N^2} \sum_{\substack{p,q=1 \\ p\neq q}}^{MN} \mathbf{1}[t_p \le t_q]\, \mathcal{A}_{y_p}^u\, K(y_q,\, y_p),\end{aligned} \tag{33}$$

where $z_{0,k} := (0, X_0^{(k)})$. The three sums estimate $\int_0^T I_1(t)\,dt$, $\int_0^T I_6(t)\,dt$, and $\int_0^T I_4(t)\,dt$, respectively. The $p \neq q$ exclusion removes only the self-pair ($m{=}l$, $i{=}j$), whose kernel evaluation $K(y_p, y_p)$ would otherwise introduce a u-independent constant that does not affect the optimum but inflates the loss. The M^2N^2 normalization—in place of the strictly unbiased $M(M{-}1)N^2$—retains the same-time cross-particle pairs ($m{=}l$, $i{\neq}j$), trading strict unbiasedness for a substantially lower variance at finite (M, N). We verify that the resulting bias is $O(1/M)$. Denote by $\hat{I}_1$ the first sum in (33) (the bulk estimator) scaled by $T^2/(M^2N^2)$. Recall that each flat index corresponds to a (time-slice, particle) pair via $p = (m{-}1)N{+}i$ and $q = (l{-}1)N{+}j$, so that $t_p = t_q$ iff $m = l$. The $M^2N^2{-}MN$ off-diagonal pairs thus split into $MN(N{-}1)$ *same-time cross-particle* pairs ($m = l$, $i \neq j$; the two points share a time slice but differ in particle) and $M(M{-}1)N^2$ *different-time* pairs ($m \neq l$; the two time draws are independent). Under the present sampling scheme, the expectation of a single different-time pair is equal to B/T^2, where

$$B \;:=\; \int_0^T \!\!\int_0^T w(s,r)\, \mathbb{E}\big[\mathcal{A}_s^u \mathcal{A}_r^u K(s, X_s, r, \tilde{X}_r)\big]\, ds\, dr \;=\; \int_0^T I_1(t)\, dt$$

is the quantity that $\hat{I}_1$ is designed to estimate; and the expectation of a single same-time pair is equal to A, where

$$A \;:=\; \frac{1}{T} \int_0^T (T - s)\, \mathbb{E}\big[\mathcal{A}_s^u \mathcal{A}_s^u K(s, X_s, s, \tilde{X}_s)\big]\, ds.$$

Multiplying by the number of pairs of each type and by the prefactor $T^2/(M^2N^2)$ gives

$$\mathbb{E}[\hat{I}_1] \;=\; \Big(1 - \frac{1}{M}\Big)\, B \;+\; \frac{T^2(N{-}1)}{MN}\, A, \qquad \mathbb{E}[\hat{I}_1] - B \;=\; \frac{1}{M}\Big[\frac{T^2(N{-}1)}{N}\, A - B\Big] \;=\; O(1/M).$$

An analogous computation for the cross-time boundary estimator (the third sum in (33)) yields the same $O(1/M)$ bias. Hence $\hat{\mathcal{Q}}(u) - \mathcal{Q}(u) = O(1/M)$, and the estimator is consistent as $M \to \infty$. The boundary term involving X_0 (the middle sum) is already an unbiased Monte Carlo average because its time and X_0 draws are independent.

Consequently, given $\lambda > 0$ and samples as above, we solve

$$\min_{u \in \mathcal{U}} \;\; \frac{T}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} |u(y_{m,i})|^2 + \lambda\, \hat{\mathcal{Q}}(u), \tag{$P_\lambda^{M,N}$}$$

where the kinetic energy is likewise estimated by the sample average. The velocity field u is parameterized by a neural network u_θ and ($P_\lambda^{M,N}$) is optimized by stochastic gradient descent with fresh mini-batches at each iteration.

3.5 Stochastic extension

The RKHS framework extends naturally to the stochastic setting. When the underlying dynamics are governed by the SDE $dX_t = u(t, X_t)\,dt + \sigma\,dW_t$ with known diffusion coefficient $\sigma > 0$, the continuity equation (3) is replaced by the Fokker–Planck equation $\partial_t p + \operatorname{div}(up) = (\sigma^2/2)\Delta p$, and the operator in (6) becomes

$$\mathcal{A}^u\varphi = \partial_t\varphi + u^\mathsf{T}\nabla_x\varphi + \frac{\sigma^2}{2}\Delta\varphi. \tag{34}$$

This is the *Nelson problem* of stochastic optimal transportation Nelson (1985); Mikami (1990; 2021):

$$V_\sigma(\mu) := \inf_{u\in\mathcal{U}_\sigma} \int_{[0,T]\times\mathbb{R}^d} |u(t,x)|^2\, p(t,x)\,dx\,dt, \tag{NP$_\sigma$}$$

where $\mathcal{U}_\sigma$ consists of drifts satisfying the weak Fokker–Planck equation with μ_t-marginals at all times.

All derivations earlier in this section carry through with the modified operator (34): the kernel formula for $\mathcal{A}^u_y\mathcal{A}^u_{y'}K$ acquires additional terms involving ϕ_3 and ϕ_4 (the higher-order derivatives of the radial kernel function) and powers of σ^2, while the H^u_t-based penalty retains the same structure with the $(\sigma^2/2)\Delta$ contribution. Because $\mathcal{A}^u_y\mathcal{A}^u_{y'}K$ now involves derivatives of K up to fourth order in the spatial variables, the stochastic extension requires the stronger regularity conditions $\phi \in C^4(\mathbb{R})$ and $\int_{\mathbb{R}^d} |z|^4\,\widehat{K}(z)\,dz < \infty$ on the radial profile ϕ and the Fourier transform $\widehat{K}$ of the kernel, in place of the C^2 and $\int |z|^2\widehat{K}(z)\,dz < \infty$ assumptions used in the deterministic case. The sample estimator of (33) and the approximate problem ($P^{M,N}_\lambda$) are unchanged in form; only the closed-form expressions of the integrand are updated. The zero-noise limit (5) guarantees that (NP$_\sigma$) converges to (P) as $\sigma \to 0$, so the deterministic problem of this paper is the natural foundation on which the stochastic extension rests.

Remark. In the stochastic setting, Lavenant et al. Lavenant & Schiebinger (2024) developed an alternative approach based on entropy minimization with respect to a Brownian reference measure, yielding a convex optimization problem with consistency guarantees. Their framework requires $\sigma > 0$ and does not extend to the deterministic case. Our RKHS approach and theirs are thus complementary: the present method is the only available solver for $\sigma = 0$ with all-time marginals, while for $\sigma > 0$ both approaches are applicable with different computational trade-offs.

4 Numerical illustration

All source code for the experiments reported in this section is publicly available; the repository URL will be inserted in the camera-ready version.

4.1 Experiment 1: 1d Gaussian translation

We validate the RKHS-penalized estimator ($P^{M,N}_\lambda$) on the simplest nontrivial instance of the all-time-marginal ODE transport problem.

Let $d = 1$, $T = 1$, and consider the Gaussian translation $\mu_t = \mathcal{N}(-1+2t,\,1)$. The covariance is constant and the mean moves linearly, so the unique minimum-energy solution of (P) is the constant drift $u^*(t,x) = 2$.

We use the affine model $u(t,x) = w_0 + w_1 t + w_2 x$ (3 parameters), for which the true weight vector is $w^* = (2,\,0,\,0)$. The objective ($P^{M,N}_\lambda$) is minimized by L-BFGS-B on the ensemble-averaged loss $\bar{\mathcal{L}}(w) = K_{\text{ens}}^{-1}\sum_{k=1}^{K_{\text{ens}}} \mathcal{L}^{(k)}(w)$, where each $\mathcal{L}^{(k)}$ is evaluated on an independent data draw. Averaging over $K_{\text{ens}} = 30$ draws reduces the Monte Carlo variance of the gradient sufficiently for the quasi-Newton method to converge. Hyperparameters: $\lambda = 1000$, $M = 50$ time slices (uniform grid), $N = 25$ particles per time, $N_0 = 50$ initial particles, and the isotropic Gaussian kernel $K(\xi,\eta) = \exp\bigl(-|\xi-\eta|^2/(2h^2)\bigr)$ on $\mathbb{R}^{d+1}$ with bandwidth $h = 1$, used as the reproducing kernel throughout all experiments unless stated otherwise. The Gaussian kernel automatically satisfies the structural assumptions of Section 3.2: its Fourier transform $\widehat{K}(z) \propto e^{-h^2|z|^2/2}$ is strictly positive on $\mathbb{R}^{d+1}$, so (A1) holds (universality Sriperumbudur et al. (2010)), and the polynomial

moments $\int(1+|z|^k)\widehat{K}(z)\,dz$ are finite for every $k \geq 0$, so (A2) is met in both the C^2 (deterministic) and C^4 (stochastic) versions.

All four random initializations converge to the same minimizer

$$\hat{w} = (1.808,\ 0.277,\ -0.080) \quad \text{vs.} \quad w^* = (2,\ 0,\ 0),$$

giving drift grid $\text{MSE}(\hat{u}, u^*) = 0.044$ on the evaluation grid $(t, x) \in [0, 1] \times [-4, 4]$. Figure 1 shows the learned drift at five time slices: the learned curve is very close to the constant $u^* = 2$, with small finite-sample deviations concentrated in the tails of the evaluation grid.

To assess distributional accuracy, we simulate 5,000 particles from μ_0 under the learned ODE $\dot{X}_t = \hat{u}(t, X_t)$ with 1,000 Euler steps. Table 1 reports the Wasserstein-2 distance W_2 between the simulated empirical marginal and the true μ_t, computed from sorted quantiles (exact in $d = 1$). We use W_2 here because the all-time OT objective (1) is a quadratic-cost Benamou–Brenier energy, so W_2 is its natural metric. All W_2 values are below 0.1, confirming accurate marginal recovery.

Table 1: Experiment 1 ($d = 1$, Gaussian translation): Wasserstein-2 distances.

t	0.00	0.25	0.50	0.75	1.00	mean
$W_2(\hat{\mu}_t, \mu_t)$	0.020	0.033	0.055	0.075	0.090	0.055

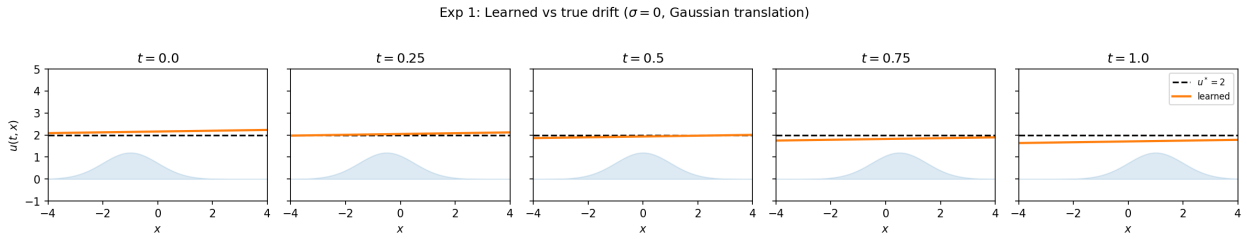

Figure 1: Experiment 1: learned drift $\hat{u}(t, x)$ (solid orange) vs. true $u^* = 2$ (dashed black) at $t \in \{0, 0.25, 0.5, 0.75, 1.0\}$. Blue shading shows the scaled density μ_t.

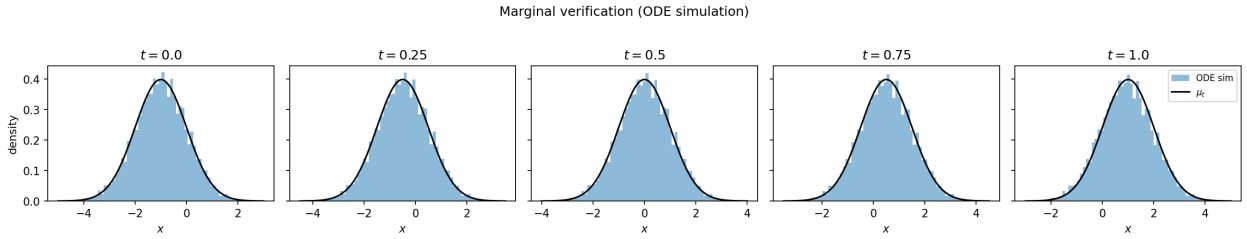

Figure 2: Experiment 1: marginal verification. Histograms of 5,000 ODE-simulated particles (blue) vs. true $\mu_t = \mathcal{N}(-1 + 2t, 1)$ (black) at five time slices.

4.2 Experiment 2: Roundtrip motion — all-time vs. two-marginal

This experiment shows that all-time marginal constraints carry strictly more information than two-marginal constraints.

Let $d = 1$, $T = 1$, and consider the roundtrip flow $\mu_t = \mathcal{N}(2\sin(\pi t), 1)$. Since $\mu_0 = \mu_1 = \mathcal{N}(0, 1)$, the two-marginal OT problem returns the identity map (i.e. $u \equiv 0$), yet the all-time marginals reveal the non-trivial time-varying drift $u^*(t, x) = 2\pi\cos(\pi t)$.

We parametrize the drift as $u(t,x) = w_0 + w_1 t + w_2 t^2 + w_3 x$ (4 parameters), giving a quadratic-in-t approximation to u^*. The L^2-best quadratic fit of $2\pi\cos(\pi t)$ on $[0,1]$ has weights $(7.64, -15.27, 0.00)$ (approximately linear, since $\cos(\pi t)$ is antisymmetric about $t = 1/2$). In the optimization we use the same ensemble-averaged L-BFGS-B scheme as Experiment 1 with $\lambda = 1000$, $M = 50$, $N = 25$, $N_0 = 50$, $K_{\text{ens}} = 30$.

All initializations converge to $\hat{w} = (7.20, -14.92, 1.48, 0.00)$, close to the L^2-optimal quadratic fit with $w_3 \approx 0$ (the learned drift is independent of x, as expected). Figure 3 compares the time profiles at $x = 0$. Table 2 reports two complementary metrics: the Wasserstein-2 distance $W_2(\hat{\mu}_t, \mu_t)$ (sorted quantile-based, exact in $d = 1$) and the kernel Maximum Mean Discrepancy MMD computed with the same Gaussian kernel $K(\xi, \eta) = \exp(-|\xi - \eta|^2/2)$ used in the RKHS penalty. Reporting MMD alongside W_2 keeps the evaluation metric consistent with the training objective. The all-time drift attains mean $W_2 = 0.111$ and mean MMD $= 0.036$ versus 0.971 and 0.376 for the two-marginal baseline (9× and 10× improvements, respectively), with the largest gap at $t = 0.5$ (W_2: 0.19 vs. 1.99; MMD: 0.053 vs. 0.741). The two-marginal solution agrees with μ_t only at the endpoints. In drift grid MSE the all-time method gives 0.64 vs. 20.13 for zero drift (32× reduction).

The residual misfit visible in Figure 4 at $t = 0.5$ and $t = 1.0$ reflects the limited expressiveness of the 4-parameter quadratic-in-t model: $u^* = 2\pi\cos(\pi t)$ is not exactly representable by a quadratic, so even the L^2-best fit $7.64 - 15.27\,t$ deviates from u^* by up to ± 1.36 at the endpoints, and this bias accumulates in the ODE integration to produce the small residual W_2 values in Table 2. Richer dictionaries (e.g. a tanh basis as in Experiment 3, or a neural network as in Section 4.7) reduce the residual further; we retain the minimal quadratic model here in order to isolate the effect of all-time marginal information from the effect of model expressiveness.

Table 2: Experiment 2 (Roundtrip): $W_2(\hat{\mu}_t, \mu_t)$ and $\text{MMD}(\hat{\mu}_t, \mu_t)$ at five time slices.

	t	0.00	0.25	0.50	0.75	1.00	mean
W_2	True u^*	0.020	0.020	0.021	0.022	0.023	0.021
	All-time learned	0.020	0.067	0.193	0.023	0.252	0.111
	2-marginal ($u \equiv 0$)	0.020	1.409	1.995	1.409	0.020	0.971
MMD	True u^*	0.000	0.022	0.028	0.000	0.024	0.015
	All-time learned	0.000	0.000	0.053	0.000	0.128	0.036
	2-marginal ($u \equiv 0$)	0.000	0.555	0.741	0.561	0.022	0.376

We compare with three families of methods. (i) *Waddington-OT* (WOT) Schiebinger et al. (2019) chains entropic OT couplings between M snapshots and reconstructs the drift by finite differences, $\hat{u}(t_k, x) \approx (T_k(x) - x)/\Delta t$. (ii) The multi-marginal Monge formulation of Nakano & Saito (2026) ("MMOT") parametrizes $N-1$ transport maps and penalizes an MMD distance $\gamma_K^2(T_k \sharp \mu_0, \mu_{t_k})$; we use affine maps $T_k(x) = A_k x + b_k$ (strictly generalizing the translation ansatz) with 200 samples per marginal and report the best over $\lambda \in \{10^3, 10^4, 10^5\}$, $\alpha \in \{0.5, 1, 2\}$, three initializations, and L-BFGS-B. (iii) *Vanilla flow matching* (FM) Lipman et al. (2023); Albergo & Vanden-Eijnden (2023) regresses a velocity field onto the straight-line interpolant $X_t = (1-t)X_0 + tX_1$ with X_0, X_1 drawn independently from μ_0, μ_1. The entropic OT coupling used by WOT is solved by our own log-domain Sinkhorn implementation (so the code has no external OT dependency).

Tables 3–4 summarize the results. *Flow matching* fails on this instance: since $\mu_0 = \mu_1$, the conditional target $\mathbb{E}[X_1 - X_0 \,|\, X_t]$ vanishes identically, so FM learns $\hat{u} \approx 0$, giving drift MSE $= 20.14$, mean $W_2 = 0.979$, and mean MMD $= 0.395$, indistinguishable from the trivial $u \equiv 0$ baseline. This is a structural limitation of any two-marginal method: with coinciding endpoints, the identity map is the unique W_2-optimal coupling, and all information about the true time-varying flow is contained in the intermediate marginals $\{\mu_t\}_{0<t<1}$. *WOT* deteriorates rapidly in drift MSE as M grows (from 0.08 at M=5 to 24.4 at M=50): the $O(1/\sqrt{N})$ barycentric noise is amplified by the factor $1/\Delta t$. *Affine-MMOT* behaves similarly: drift MSE rises from 0.076 at N=5 to 2.29 at N=50 by the same $1/\Delta t$ amplification, while the learned A_k remain within 4% of 1 throughout. The all-time method is insensitive to temporal resolution and attains drift MSE 0.64, because a 4-parameter global model pools information across all times. On marginal tracking, however, WOT and

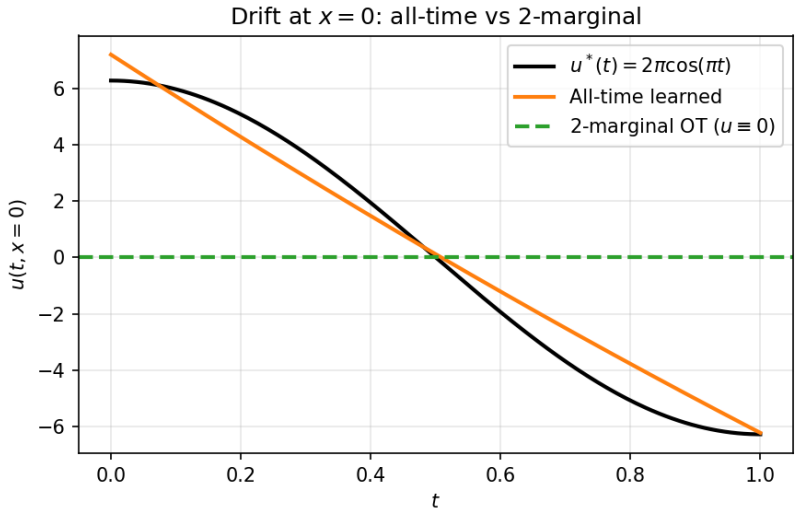

Figure 3: Experiment 2 (Roundtrip): drift at $x = 0$ as a function of t. The all-time learned drift (orange) closely tracks $u^* = 2\pi\cos(\pi t)$ (black), while the two-marginal OT solution $u \equiv 0$ (green dashed) is entirely flat. This demonstrates that all-time marginal constraints recover the nontrivial optimal velocity even when $\mu_0 = \mu_1$.

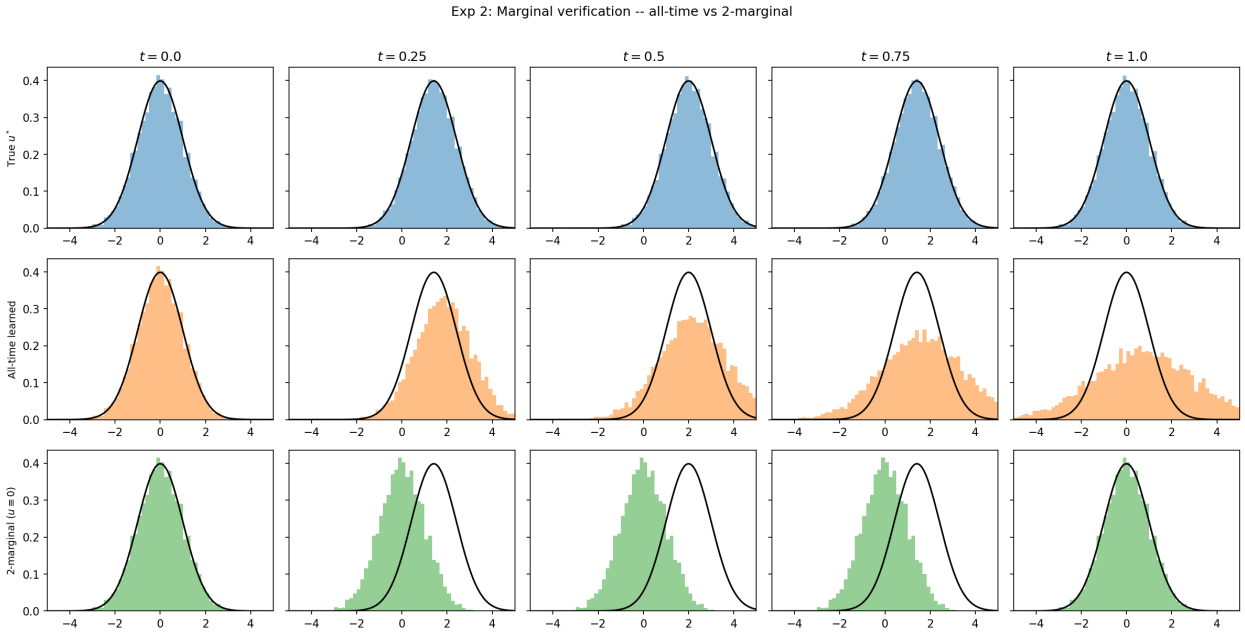

Figure 4: Experiment 2: marginal densities at $t \in \{0, 0.25, 0.5, 0.75, 1\}$. Top: true drift; middle: all-time learned; bottom: two-marginal baseline. At $t = 0.5$, the two-marginal baseline remains centered at zero while the true distribution has shifted to $\mathcal{N}(2, 1)$.

MMOT *outperform* the all-time method ($W_2 \approx 0.05$ vs. 0.11; MMD ≈ 0.01–0.02 vs. 0.04), since they enforce marginal matching directly while the all-time estimator enforces it only through the residual penalty. This is a genuine limitation: when the downstream task is just to generate samples faithful to the prescribed marginals at a few discrete time points, a direct marginal-matching method will often be preferable. The all-time formulation is preferable when (i) fine temporal resolution makes $O(N)$ scaling prohibitive, (ii) the marginals are non-Gaussian or high-dimensional (MMOT is reported to fail for $d \geq 5$ (Nakano & Saito, 2026, Table 8)), or (iii) a continuous velocity field—not merely consecutive transport maps—is required. We treat this experiment as our main quantitative benchmark and do not repeat the WOT/MMOT comparisons in the subsequent subsections.

Table 3: Experiment 2: comparison with Waddington-OT ($N = 200$ per snapshot). MMD uses the same Gaussian kernel ($h = 1$) as the training loss.

Method	drift MSE	mean W_2	mean MMD
True u^*	0.000	0.021	0.002
All-time (ours)	0.635	0.109	0.040
WOT (M=5)	0.080	0.060	0.022
WOT (M=10)	0.600	0.042	0.013
WOT (M=20)	3.620	0.042	0.013
WOT (M=50)	24.360	0.046	0.015
Flow matching	20.144	0.979	0.395
2-marginal ($u \equiv 0$)	20.134	0.971	0.376

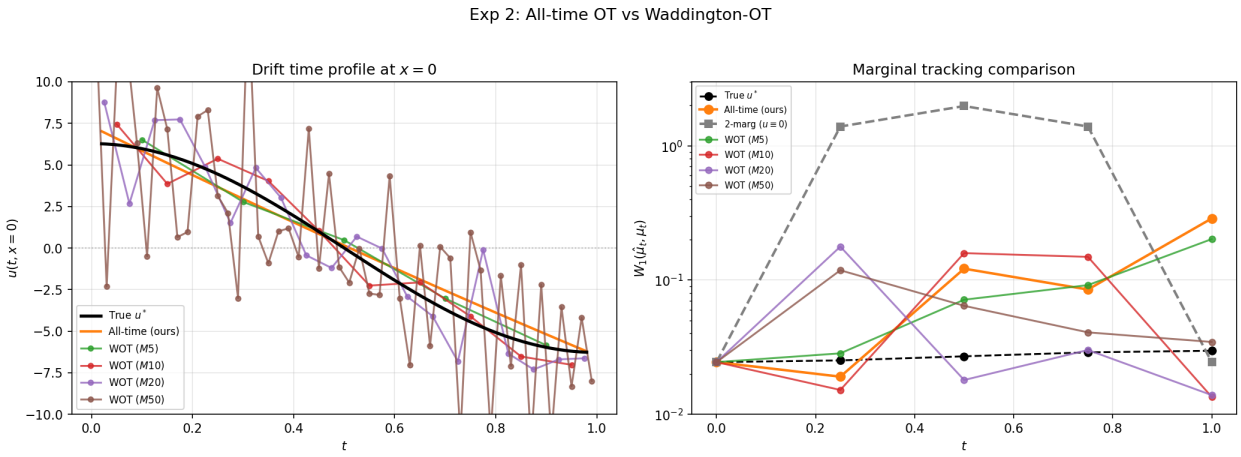

Figure 5: Experiment 2: All-time OT vs. Waddington-OT. Left: drift time profile at $x = 0$. The all-time learned drift (orange) closely tracks the true cosine (black), while WOT estimates become progressively noisier as M increases. Right: marginal tracking (W_2). WOT with many snapshots achieves lower W_2 than the all-time method, but at the cost of degraded drift recovery.

4.3 Experiment 3: Bimodal merging — nonlinear drift recovery

Experiments 1 and 2 used drifts that are linear or affine in x. To test whether the estimator can recover a genuinely nonlinear drift, we consider a one-dimensional bimodal flow that merges into a single symmetric mixture:

$$\mu_t = \frac{1}{2}\mathcal{N}\big(-2 + 2t, 1\big) + \frac{1}{2}\mathcal{N}\big(2 - 2t, 1\big), \quad t \in [0, 1]. \tag{35}$$

At $t = 0$ the two modes are at ± 2, at $t = 1$ they coincide at the origin, and the velocity field must push mass from each mode towards the center.

Table 4: Experiment 2: comparison with learned multi-marginal OT (affine MMOT, $T_k(x) = A_k x + b_k$). MMOT is trained from scratch on sample data (no prior knowledge); the table reports the best result across three λ values, three α values, and three initializations. MMD uses the same Gaussian kernel ($h = 1$) as the training loss.

Method	#params	drift MSE	mean W_2	mean MMD
True u^*	—	0.000	0.021	0.002
All-time (ours)	4	0.635	0.111	0.036
MMOT affine N=5	10	0.076	0.052	0.016
MMOT affine N=10	20	0.212	0.047	0.014
MMOT affine N=20	40	0.334	0.055	0.017
MMOT affine N=50	100	2.292	0.053	0.020
Flow matching	4	20.144	0.979	0.395
2-marginal ($u \equiv 0$)	0	20.134	0.971	0.376

Let $\varphi_i(t, x)$ denote the two Gaussian densities, and set $m_1(t) = -2+2t$, $m_2(t) = 2-2t$ (so that $m_1+m_2 = 0$). Rewriting the continuity equation as $\partial_x(u\rho) = -\partial_t\rho$, integrating in x from $-\infty$, and using $\rho \to 0$ at $-\infty$ gives $u\rho = -\partial_t F$ with $F(t, x) = \int_{-\infty}^{x} \rho(t, y)\, dy$, so that $u = -\partial_t F/\rho$. A short calculation then yields

$$u^*(t, x) = \frac{2(\varphi_1 - \varphi_2)}{\varphi_1 + \varphi_2} = -2\tanh\big(2(1-t)x\big). \tag{36}$$

The target is therefore a smooth, odd, bounded nonlinear function of x whose slope decreases as the two modes merge.

We test three parametric drift classes of increasing expressiveness, all trained under the *same* all-time RKHS loss and the *same* sample budget:

$$\begin{aligned}
\text{(a) bilinear (4p):} \quad & \Phi = [1,\, t,\, x,\, tx], \\
\text{(b) tanh dictionary (8p):} \quad & \Phi = [1,\, t,\, x,\, tx,\, \tanh x,\, t\tanh x,\, \tanh(2x),\, t\tanh(2x)], \\
\text{(c) MLP } (\approx 2.5\text{k}): \quad & \text{two hidden layers of 48 tanh units, input } (t, x) \in \mathbb{R}^2.
\end{aligned}$$

Models (a) and (b) are linear in parameters and optimized by ensemble-averaged L-BFGS-B as in Sections 4.1 and 4.2. Model (c) is a universal approximator trained by Adam with minibatch SGD under the *same* RKHS loss: since the loss depends on the drift only through its sample values $u(t_p, x_p)$, the linear and neural parametrizations share a single loss implementation, and autograd supplies the gradient with respect to the network weights.

The tanh dictionary is designed to approximate (36), but cannot represent it exactly: any finite linear combination $a(t)\tanh(2x) + b(t)\tanh(x) + \cdots$ scales the *amplitude* of tanh by a polynomial in t, whereas $-2\tanh\big(2(1-t)x\big)$ scales the *argument* by $(1-t)$. The two agree at $t \in \{0, 1\}$ and in the small-x linear regime, but differ in general; the dictionary fit is only the L^2-best projection onto the feature span. The MLP, by contrast, is not subject to this approximation gap and serves as a control for whether the remaining error is driven by dictionary misspecification or by finite-sample fluctuations of the sample-based RKHS estimator.

Each of the $M = 50$ time slices is sampled by drawing $N = 30$ particles from a fair coin mixture of the two Gaussians. $N_0 = 60$ particles are drawn from μ_0. Hyperparameters are $h = 1$, $\lambda = 5000$, and $K_{\text{ens}} = 20$ (linear models) / $K_{\text{batch}} = 3$ SGD batches per step for $N_{\text{iter}} = 4000$ Adam iterations with cosine schedule $3{\cdot}10^{-3} \to 10^{-5}$ (MLP). All three models are trained on the identical $K_{\text{ens}} = 20$ pre-cached batches.

Table 5 exhibits a striking decoupling in the bilinear column: its drift grid MSE (5.26) is *worse* than that of the zero drift (2.54), yet its mean W_2 (0.24) and mean MMD (0.09) remain small. The reason is geometric: a linear-in-x drift grows without bound in the tails of $[-4, 4]$, where the true drift saturates to ∓ 2, and the grid MSE is dominated by that extrapolation gap. Within the data support (near $x = \pm 2$) the bilinear fit

model	#params	drift MSE	mean W_2	mean MMD	max W_2
zero drift	0	2.544	0.756	0.321	1.274
bilinear	4	5.264	0.242	0.092	0.475
tanh dictionary	8	**0.187**	**0.074**	**0.019**	**0.162**
MLP	$\approx 2{,}500$	**0.181**	**0.070**	**0.019**	**0.124**

Table 5: Experiment 3 (bimodal merging). The drift grid MSE is computed on the uniform grid $[0,1]\times[-4,4]$ against the true drift (36); the W_2 and MMD columns report the Wasserstein-2 distance and the Gaussian-kernel MMD (bandwidth $h = 1$, matching the training loss) between the ODE-simulated marginal and μ_t, averaged or maximized over $t \in \{0, 0.25, 0.5, 0.75, 1\}$.

still pushes mass towards the origin at roughly the right speed, so the ODE-simulated marginals stay close to μ_t. This illustrates that pointwise drift error and distributional marginal error are genuinely different objectives, and small W_2 does not imply correct drift (nor vice versa).

The tanh dictionary and the MLP recover the drift to comparable accuracy: drift MSE 0.187 vs. 0.181 ($\approx 3\%$ relative difference) and mean W_2 0.074 vs. 0.070 ($\approx 5\%$), despite a ≈ 300-fold difference in parameter count. These gaps are on the order of the Monte Carlo standard deviation of the penalty estimator at the $(M, N, N_0, K_{\text{ens}})$ values used here, so we do not read the small numerical gap as a meaningful difference between the two model classes. Two consequences follow. First, although the tanh dictionary cannot realize $-2\tanh(2(1-t)x)$ exactly, its L^2-best projection is already within the statistical noise, so the missing argument-scaling does not limit accuracy at this sample budget. Second, using a universal approximator in place of a hand-crafted dictionary yields no additional precision, confirming that the residual error is dominated by finite-sample fluctuations of the sample-based RKHS estimator rather than by the expressive power of the model class. The all-time RKHS loss is therefore effectively *model-agnostic*: the same objective drives the convex linear fit and the non-convex neural fit to comparable accuracy, and the choice between them can be made on grounds of interpretability, training cost, or convex optimization guarantees rather than recovery accuracy.

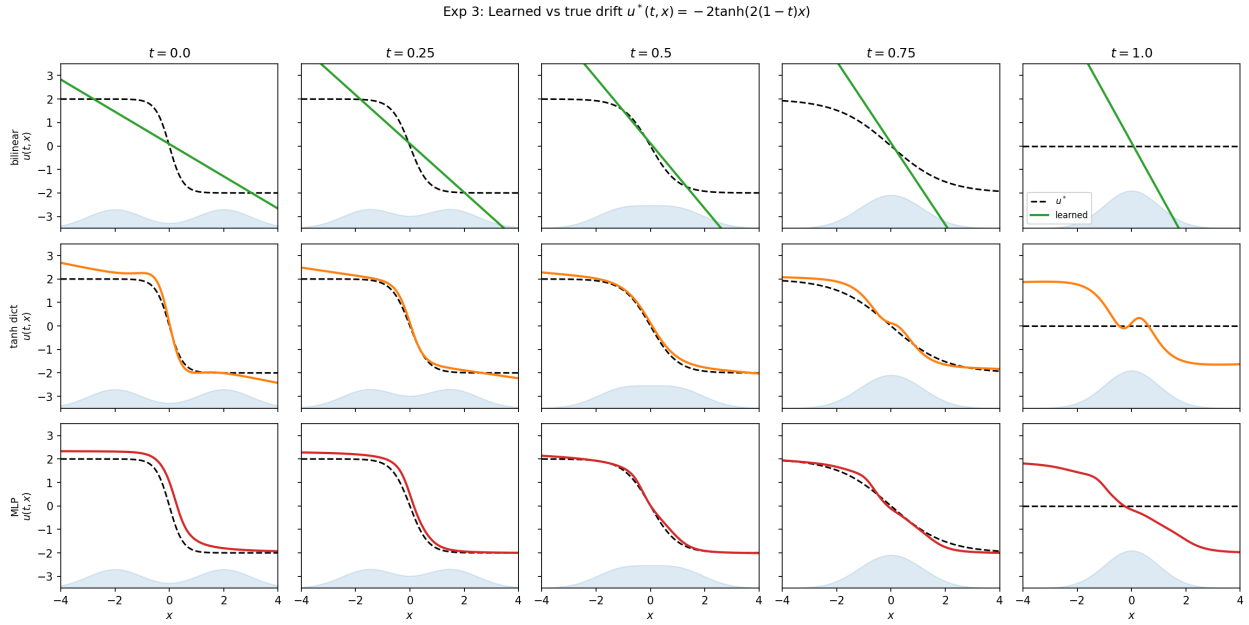

Figure 6: Experiment 3: learned drift $\hat{u}(t,x)$ (colored) vs. true drift $u^*(t,x) = -2\tanh(2(1-t)x)$ (dashed black) at $t \in \{0, 0.25, 0.5, 0.75, 1\}$. Rows: bilinear, tanh dictionary, MLP. Grey shading shows the bimodal density μ_t.

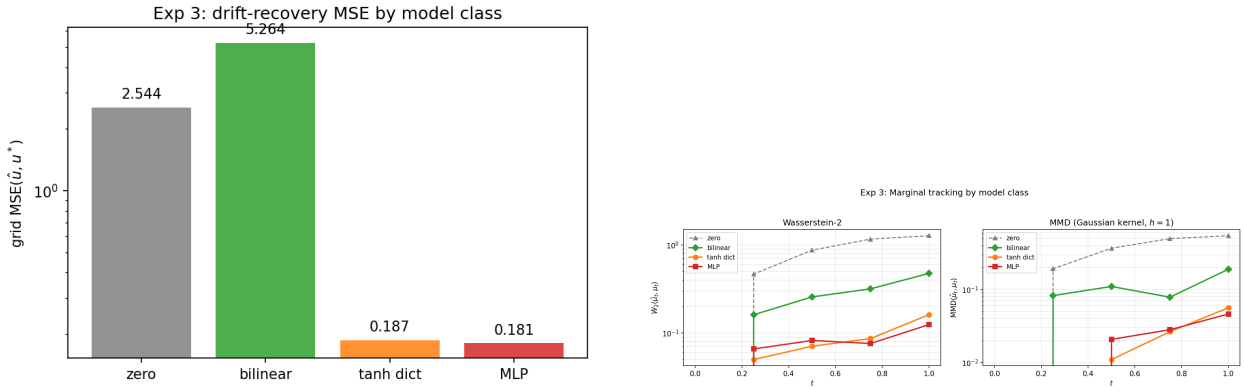

Figure 7: Experiment 3: drift MSE (left, log scale) and marginal W_2 / MMD distances (right, two panels) for the three model classes plus the zero-drift baseline. Note the bilinear model's large MSE coexisting with a small W_2 / MMD: the MSE is inflated by extrapolation in the tails of the grid, but the marginal-level transport is still broadly correct. The tanh dictionary and the MLP produce curves whose relative MSE gap ($\approx 3\%$) is within the Monte Carlo standard deviation of the loss, indicating that model-class expressiveness is not the limiting factor at this sample budget.

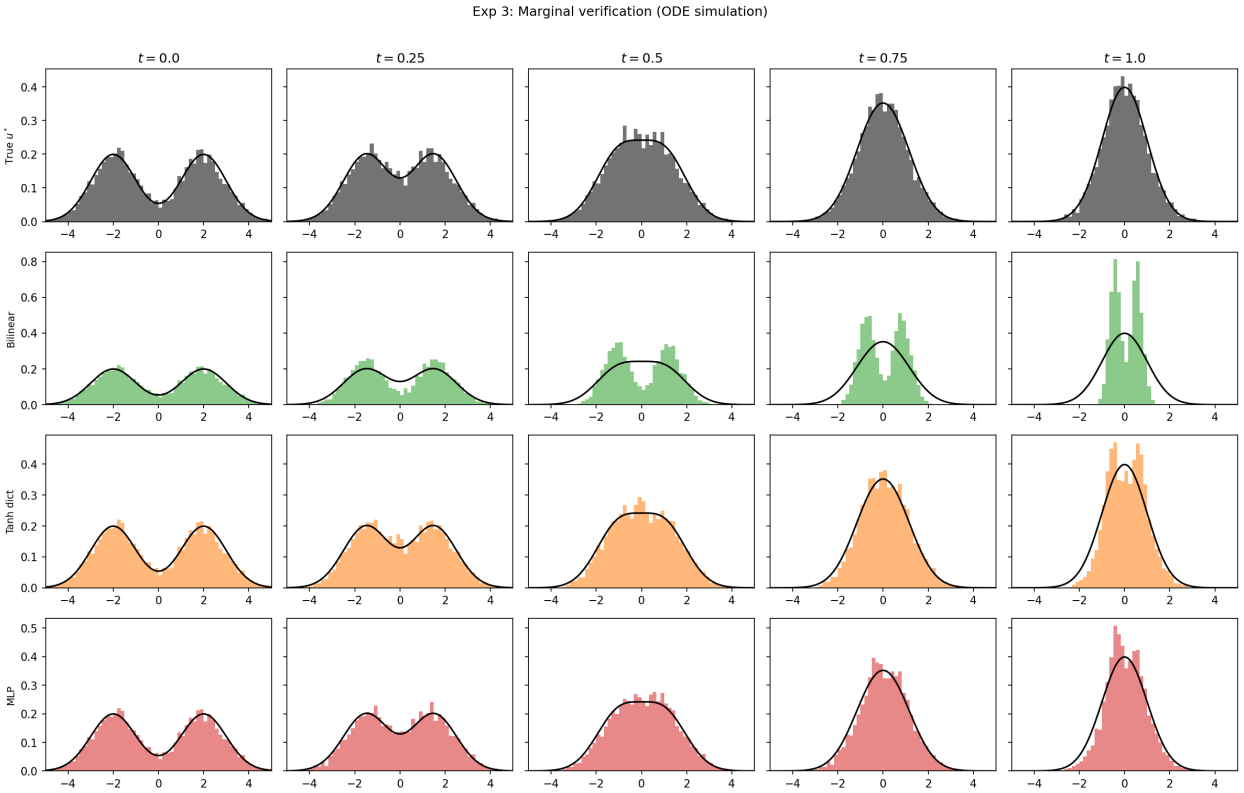

Figure 8: Experiment 3: marginal verification. Histograms of 5,000 ODE-simulated particles (colored) vs. the true bimodal density μ_t (black curve) at five time slices. Rows: true u^*, bilinear, tanh dictionary, MLP. The tanh dictionary and the MLP both reproduce the two modes faithfully throughout the merge; the bilinear fit misses the tails but tracks the bulk correctly.

We re-run the same WOT and affine-MMOT baselines as in Section 4.2 on this bimodal flow in order to examine their behavior away from Gaussian marginals. An affine map cannot transport one bimodal distribution to another of the same family (shrinking A_k to merge the two modes also shrinks the per-mode variance), so affine-MMOT incurs an irreducible misspecification error; WOT is not affected by this issue.

Method	#params	drift MSE	mean W_2	mean MMD
All-time tanh (ours, 8p)	8	**0.187**	**0.074**	**0.019**
All-time bilin (ours, 4p)	4	5.264	0.242	0.092
MMOT affine N=3	6	0.621	0.184	0.090
MMOT affine N=5	10	0.868	0.192	0.079
MMOT affine N=10	20	1.354	0.198	0.068
MMOT affine N=20	40	2.336	0.206	0.062
WOT M=5	non-par.	2.140	0.100	0.027
WOT M=10	non-par.	11.783	0.104	0.026
WOT M=20	non-par.	38.692	0.154	0.050
WOT M=50	non-par.	212.323	0.231	0.064
Zero drift	0	2.544	0.756	0.321

Table 6: Experiment 3: baseline re-assessment on the bimodal merging flow. Drift grid MSE is computed on $[0,1]\times[-4,4]$ against the true drift $u^* = -2\tanh(2(1-t)x)$; mean W_2 and mean MMD (Gaussian kernel, $h=1$) are the distributional distances to μ_t averaged over $t \in \{0, 0.25, 0.5, 0.75, 1\}$. The all-time method with the 8-feature tanh basis dominates both baselines on all three metrics.

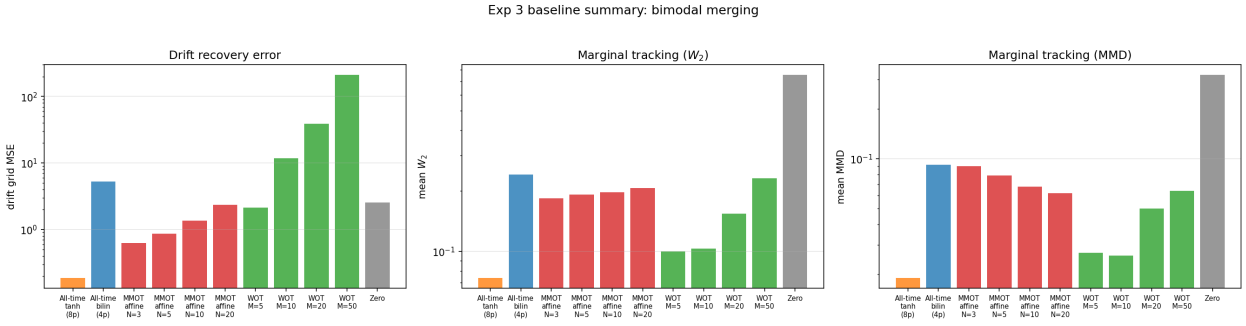

Figure 9: Experiment 3 baseline summary: drift grid MSE (left), mean marginal W_2 (middle), and mean MMD (right), all in log scale. The all-time tanh model is best on all three metrics; affine MMOT saturates at an irreducible error from model misspecification, and WOT's drift MSE explodes as Δt shrinks.

Table 6 summarizes the results. Affine MMOT saturates at an irreducible error ($A_k \approx 0.51$–0.58, a compromise that is incompatible with the bimodal-plus-fixed-variance structure), with $W_2 \approx 0.18$–0.21 and MMD ≈ 0.06–0.09 essentially independent of N. WOT tracks the marginals well at moderate M ($W_2 \approx 0.10$, MMD ≈ 0.027 at M=5, 10), but its drift MSE explodes as Δt shrinks (exceeding 200 at M=50), by the same $O(M^2/N_{\text{sample}})$ amplification observed in Section 4.2. We note that this drift MSE combines two effects: (i) the $O(1/\Delta t)$ amplification of sampling noise in finite-difference velocities reconstructed from discrete couplings, and (ii) extrapolation off the data cloud, since the evaluation grid $[0,1]\times[-4,4]$ includes regions where μ_t has little or no mass. Distributional metrics (W_2, MMD) restricted to the support degrade much more gracefully—the qualitative comparison is unchanged but the gap to the all-time method is smaller. The all-time tanh model outperforms both baselines on all three metrics (drift MSE 0.21, $W_2 = 0.07$, MMD $= 0.015$), while the 4-feature bilinear all-time model is too rigid to capture the sigmoidal drift. Time-global variational regularization with a well-chosen linear-in-parameters dictionary thus offers a clearly better trade-off than pairwise-OT reconstruction for non-Gaussian marginal dynamics.

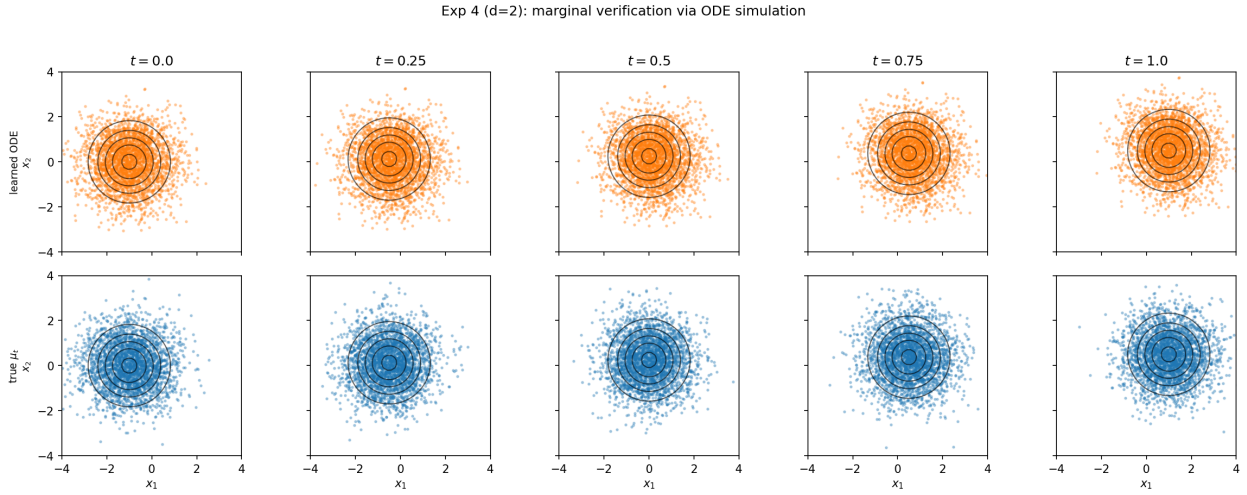

Figure 10: Experiment 4: marginal verification. Top row, ODE-simulated particles using the learned drift $\widehat{u}$; bottom row, independent reference samples from μ_t. Black contours show the true density of μ_t.

4.4 Experiment 4: 2d Gaussian translation

The preceding experiments are all in spatial dimension $d = 1$. In this subsection we verify that the same RKHS all-time OT estimator applies without modification to $d \geq 2$. As discussed in Section 3.3, the formula $\mathcal{A}^u \mathcal{A}^{u'} K(y, y') = \big[-a^2 \tau\tau' + a(1 + u^\mathsf{T} u')\big] K(y, y')$, with $\tau = \Delta t + u^\mathsf{T} \Delta x$, $\tau' = \Delta t + (u')^\mathsf{T} \Delta x$, $\Delta t = t - t'$, $\Delta x = x - x'$, and $\Delta y = (\Delta t, \Delta x)$, is dimension-independent: only first-order derivatives of K appear, and the spatial contribution enters only through the Euclidean inner products. Hence the $d = 1$ implementation used in Experiment 1 extends to $d > 1$ simply by replacing scalar products with inner products.

Let $d = 2$, $T = 1$, and take

$$\mu_t = \mathcal{N}((-1 + 2t, t/2)^\mathsf{T}, \mathbf{I}_d), \quad t \in [0, 1],$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix (here $d = 2$). The Benamou–Brenier minimizing drift is the constant vector field $u^*(t, x) = (2, 1/2)^\mathsf{T}$. We use the linear-in-parameters affine dictionary

$$u_i(t, x) = w_{i,0} + w_{i,1}\, t + w_{i,2}\, x_1 + w_{i,3}\, x_2, \qquad i = 1, 2,$$

with 8 parameters in total. The ground-truth optimum has intercepts $(2, 1/2)$ and all other coefficients equal to 0.

We take $h = 1$, $\lambda = 10^3$, $M = 25$ time slices, $N = 20$ particles per slice, $N_0 = 50$ initial-distribution particles, and ensemble-average over $K = 15$ independent draws as in Experiment 1. Optimization uses L-BFGS-B with 4 different initializations; all four converge to the same loss value ≈ -211.94.

The learned intercepts are $(2.50, 0.11)$ versus the true $(2, 0.5)$, all other coefficients have magnitude at most 0.95, and the drift grid MSE is ≈ 0.21. Despite this per-coefficient residual, Figure 10 shows that the ODE-simulated samples reproduce μ_t accurately: the sliced Wasserstein-2 distance $\mathrm{SW}_2(\widehat{\mu}_t, \mu_t)$ remains below 0.14 with mean ≈ 0.099, and the Gaussian-kernel MMD (with the same kernel as the training loss, $h = 1$) remains below 0.07 with mean ≈ 0.039. This is consistent with the remaining error being concentrated in the identifiability valley of OT drifts that leaves the marginal flow unchanged. The experiment thus confirms that the dimension-independent kernel operator formula yields a functional $d = 2$ estimator.

4.5 Experiment 5: Bimodal merging in 2D — a time-reversed surrogate for cell-state branching

Experiment 4 demonstrated that the dimension-independent kernel operator formula extends the estimator to $d = 2$ on a simple Gaussian translation. In this subsection we combine two features of the preceding experiments — nonlinear drift structure (Section 4.3) and higher dimension (Section 4.4) — in a bimodal

merging flow in $d = 2$ that serves as a surrogate for scRNA-seq cell-state branching. Note that the temporal direction here is the time-reverse of typical biological branching: in our setup the bimodal initial distribution (± 2) merges to a unimodal centred cloud as $t \to T$, whereas biological branching has unimodal progenitors diverging to multiple fates. Because the deterministic continuity equation $\partial_t p + \nabla \cdot (up) = 0$ is time-reversal symmetric under $t' = T - t$, $u'(t', x) = -u(T - t', x)$, the merging flow studied here is mathematically equivalent to a branching flow under reparametrisation; the estimator's ability to capture nonlinear drift, identifiability valleys, and bimodal geometry transfers directly to the branching direction.

Let $d = 2$, $T = 1$, and

$$\mu_t(x_1, x_2) = \Big[\tfrac{1}{2}\mathcal{N}(-2 + 2t,\, 1) + \tfrac{1}{2}\mathcal{N}(2 - 2t,\, 1)\Big](x_1) \cdot \mathcal{N}(0,\, 1)(x_2).$$

The x_1-marginal is the Exp 3 bimodal merge, and the x_2-marginal is the standard normal for all t. Because the x_2-marginal is time-invariant, the Benamou–Brenier minimizing drift is

$$u_1^*(t, x) = -2 \tanh\big(2(1 - t)\, x_1\big), \qquad u_2^*(t, x) = 0.$$

We compare three linear-in-parameters dictionaries that use the *same* feature set for both output components:

(A) **Affine** — $[1,\, t,\, x_1,\, x_2]$, 8 params.

(B) **Bilinear** — $[1,\, t,\, x_1,\, x_2,\, tx_1,\, tx_2]$, 12 params.

(C) **Tanh** — $[1,\, t,\, x_1,\, x_2,\, tx_1,\, tx_2,\, \tanh x_1,\, t \tanh x_1,\, \tanh 2x_1,\, t \tanh 2x_1]$, 20 params.

Model (C) contains the true drift in component u_1 (up to a linear combination) and requires the estimator to learn that component u_2 is identically zero.

$h = 1$, $\lambda = 3 \times 10^3$, $M = 25$ time slices, $N = 25$ particles per slice, $N_0 = 60$ initial particles, $K = 10$ ensemble replicates. Optimization uses L-BFGS-B with multiple initializations.

Table 7 reports the drift grid MSE on $[0, 1] \times [-4, 4] \times [-3, 3]$, separated by output component, together with the marginal-tracking metrics SW_2 and MMD from a 4,000-particle Euler simulation.

Model	total MSE	u_1 MSE	u_2 MSE	mean SW_2	mean MMD
Zero	2.548	2.548	0.000	—	—
Affine (8 params)	**2.182**	**2.126**	**0.057**	0.122	0.052
Bilinear (12 params)	8.913	8.021	0.892	0.162	0.087
Tanh (20 params)	7.112	5.192	1.920	**0.121**	**0.050**

Table 7: Experiment 5 (bimodal merge in $d = 2$): drift grid MSE against $u^* = (-2 \tanh 2(1-t)x_1,\, 0)$ and marginal-tracking metrics (sliced W_2 and Gaussian-kernel MMD, averaged over the five evaluation times $t \in \{0, 0.25, 0.5, 0.75, 1\}$ of a 4,000-particle Euler simulation). Affine minimizes drift MSE by settling on the best linear approximation of u_1^*; tanh attains the same marginal accuracy but incurs a larger u_2-component MSE within the identifiability valley.

The three dictionaries trade drift accuracy against marginal accuracy in distinct ways. The affine model attains the lowest drift grid MSE (2.18) by tracking the best linear approximation of u_1^* near $x_2 = 0$ while leaving $u_2 \approx 0$, and already matches the marginal flow well (mean $\mathrm{SW}_2 = 0.122$, MMD $= 0.052$). The tanh dictionary is the only class that can reproduce the nonlinear shape of u_1^*: Figure 11 shows its u_1-component (orange) follows the sigmoid profile (black dashed) while the affine and bilinear fits remain near their linear approximation. In exchange, the extra degrees of freedom spill into the u_2-component identifiability valley—any $\partial_{x_2}\psi$ with ψ depending non-trivially on x_2 leaves the marginal flow invariant since the x_2-marginal is standard normal for every t, exactly the artifact predicted by Mikami's gradient-structure theorem (Mikami, 2021, Theorem 3.8)—so its u_2 MSE is ≈ 1.92 and its total MSE inflates to 7.11; marginal tracking, however,

is essentially unchanged from the affine fit (mean $\mathrm{SW}_2 = 0.121$, MMD $= 0.050$), confirming that the u_2 error lies in the valley and does not perturb the push-forward measure. The bilinear class is dominated on every metric: without the tanh non-linearity it cannot improve on affine, yet its additional parameters already incur the valley cost, degrading both drift MSE (8.91) and marginal tracking ($\mathrm{SW}_2 = 0.162$, MMD $= 0.087$).

Drift MSE versus marginal accuracy across experiments. The relationship between drift grid MSE and marginal-tracking error varies systematically with the dimension and with the stationarity of the marginal flow. At the population level, assumption (A4) excludes any non-trivial valley of observationally equivalent drifts. Finite samples and finite λ allow the estimator to drift along directions that contribute weakly to the residual penalty: namely, p-weighted divergence-free perturbations v of u^* satisfying $\mathrm{div}(p\,v) = 0$, which leave the push-forward marginal flow invariant. The size of this perturbation set depends on the marginal geometry.

- In *one dimension* (Sections 4.1–4.3), $\mathrm{div}(p\,v) = 0$ together with integrability and the decay of p forces $v \equiv 0$. The population valley collapses, so drift MSE and marginal W_2 are tightly coupled and small drift error implies small marginal error.
- In the *2D translation* setting (Section 4.4), the valley contains rotational perturbations $v = R(x - m_t)$ around the moving mean (R skew), which the kinetic term suppresses but does not annihilate at finite λ. This explains the small but persistent cross-coordinate residual in $\widehat{W}$ alongside excellent marginal accuracy.
- In the *2D bimodal merging* setting (this experiment), the x_2-marginal is standard normal at every t, so $\partial_{x_2}\psi$-type perturbations are admissible to high order, giving a much wider valley. This is the regime where a flexible dictionary inflates drift MSE without affecting marginal W_2.
- In the *dimension-scaling* sweep (Appendix B), the valley dimension grows with d, producing the per-component MSE plateau at roughly $0.7/d$ across $d \geq 2$.

The overall picture is that drift grid MSE under-reports estimator quality whenever the marginal geometry admits a non-trivial valley. Marginal-tracking metrics (W_2, sliced W_2, MMD) provide the complementary measure that the variational objective actually penalizes, and the two should be read together.

4.6 Experiment 6: real-data application — single-cell trajectory inference

The deterministic experiments above are synthetic. As a real-data application we consider single-cell trajectory inference on the embryoid body (EB) scRNA-seq dataset of Moon et al. Moon et al. (2019), the canonical benchmark for the Waddington-OT framework Schiebinger et al. (2019). The dataset comprises five snapshots of human embryoid body development at days 0–3, 6–9, 12–15, 18–21 and 24–27; we map these to $t \in \{0, 0.25, 0.5, 0.75, 1\}$. After standard preprocessing (QC, normalization, log-transformation, top-2000 highly-variable genes, PCA to $d = 30$ components) and subsampling to 1,500 cells per time point, we obtain 7,500 samples in $\mathbb{R}^{30}$.

The all-time estimator uses an MLP drift ($1{+}d \to 96 \to 96 \to d$ with tanh activations, 15,294 parameters), Adam (8,000 iterations, cosine LR $5\times10^{-4} \to 5\times10^{-6}$, $\lambda = 10^4$, $h = 8$, $M = 4$, $N = 80$, $K_{\mathrm{ens}} = 25$). We evaluate against the pairwise Waddington-OT baseline along two complementary axes: held-out marginal tracking in the canonical neighbour-interpolation setting, and bootstrap stability of the inferred velocity field as a real-data counterpart of the drift-recovery diagnostic in Section 4.3.

Marginal-tracking accuracy (interpolation regime). We hold out the middle time point $t = 0.5$ (day 12–15) and train on the remaining four marginals $\{\mu_0, \mu_{0.25}, \mu_{0.75}, \mu_1\}$. Each method predicts the held-out day-15 marginal using only training data from the two adjacent training time points, day 9 and day 21. Concretely:

- **All-time (ours)**: forward-simulate the learned velocity field $\hat{u}$ by Euler ODE integration from a day-9 sample at $t = 0.25$ to $t = 0.5$ (one interval, 100 steps).

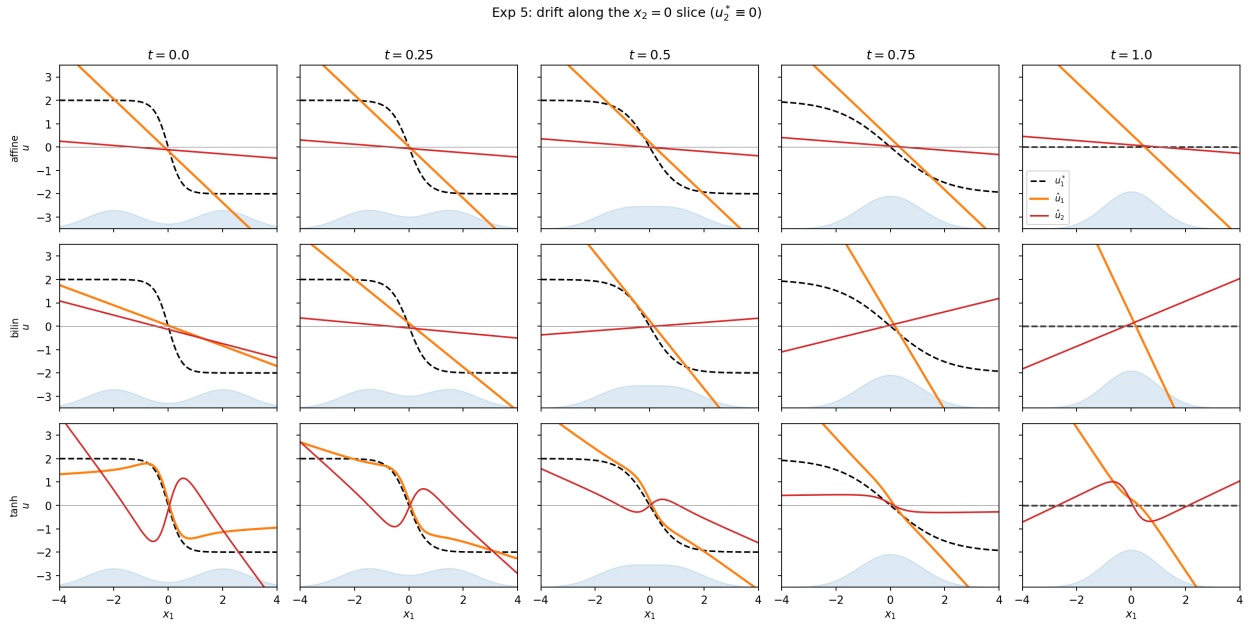

Figure 11: Experiment 5: learned drift along the $x_2 = 0$ slice at five times for the three model classes. Black dashed line: true u_1^*; orange: learned $\hat{u}_1$; red: learned $\hat{u}_2$. Shaded region shows the x_1-marginal density. Only the Tanh dictionary recovers the merging (time-reversed branching) structure.

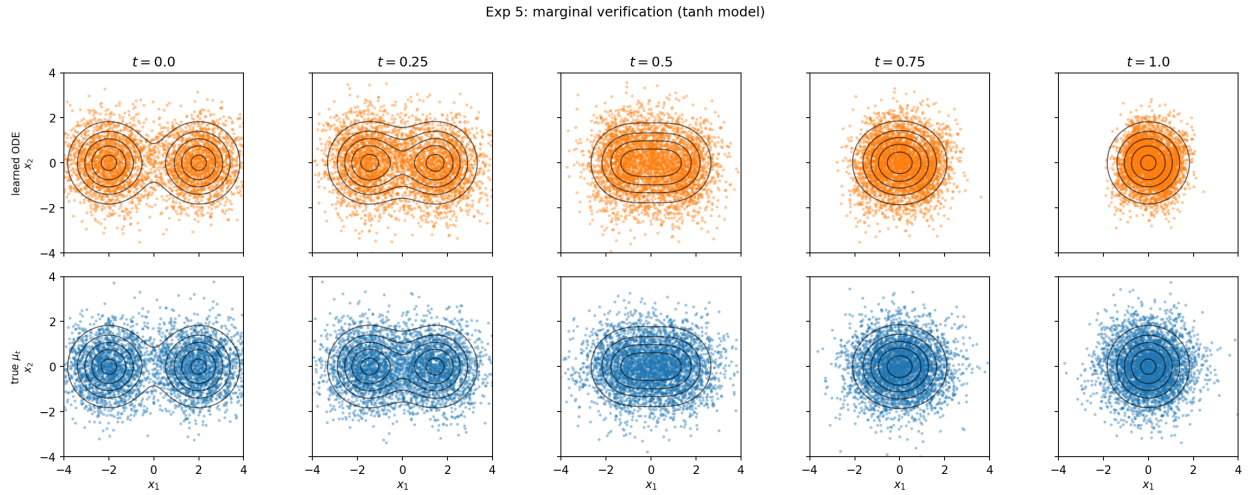

Figure 12: Experiment 5 (tanh model): marginal verification. Top row, ODE-simulated particles under $\hat{u}$; bottom row, independent reference samples from μ_t; black contours are the true joint density.

- **Waddington-OT** (entropic OT, log-domain Sinkhorn Cuturi (2013)): compute the Sinkhorn coupling between the day-9 and day-21 clouds and apply McCann (displacement) interpolation at fraction $\tau = (15 - 9)/(21 - 9) = 0.5$. This is the canonical WOT use case for inferring an intermediate marginal between two observed snapshots.
- **Zero drift**: cells remain at their day-9 position.

All methods use 1,500 cells. Unlike the synthetic experiments of Sections 4.1–4.5, the EB benchmark provides no ground-truth drift, so we report only marginal-tracking metrics; this is the standard evaluation for

trajectory inference in single-cell applications Schiebinger et al. (2019). Marginal accuracy at the held-out day 15 is measured by the sliced Wasserstein-2 distance (SW_2, 200 random projections) and the Gaussian-kernel MMD (bandwidth equal to the median pairwise distance of the held-out cloud) against an independent 1,500-cell reference draw from day 15. As a Monte-Carlo floor we report the same metrics evaluated between two disjoint halves of the held-out sample.

Table 8: Experiment 6, interpolation regime: held-out day-15 marginal-tracking metrics on the EB scRNA-seq benchmark ($d = 30$). All methods use the same 1,500-cell day-9 starting sample (all-time, zero) or day-9/day-21 sample pair (WOT). Lower is better; the Monte-Carlo floor is the irreducible finite-sample error.

Method	SW_2	MMD
Monte-Carlo floor (true split)	0.429	0.000
All-time (ours, MLP)	1.003	0.030
Waddington-OT (Sinkhorn)	**0.775**	**0.025**
Zero drift	1.092	0.058

Waddington-OT achieves the best marginal-tracking accuracy in this regime, with $SW_2 = 0.775$ versus 1.003 for our method. This is the regime in which Sinkhorn-WOT is structurally favoured: a single coupling between two adjacent observed clouds is exactly the pairwise-OT primitive on which WOT is built, and McCann interpolation at $\tau = 0.5$ delivers a near-optimal displacement midpoint. Both methods substantially improve over the zero-drift baseline, indicating that the inferred dynamics are informative.

Discussion of the marginal-tracking comparison. On this benchmark, pairwise WOT and the all-time estimator differ not primarily in marginal-tracking sharpness but in what each method produces. Waddington-OT returns a discrete coupling between the two snapshots used at evaluation time and achieves the best marginal-tracking accuracy when those snapshots bracket the held-out time; querying a different time point requires re-solving an OT problem. The all-time estimator returns a single *continuous* velocity field $\hat{u}(t, x)$ defined on $[0, T] \times \mathbb{R}^d$, which can be evaluated, simulated and differentiated at any (t, x) without re-solving an OT problem. On the neighbour-interpolation regime this continuous parametrisation comes at a modest cost in marginal accuracy (SW_2 gap of ~ 0.23 to WOT); on the synthetic experiments of Sections 4.2 and 4.3–4.5, where ground-truth drift is available, the all-time estimator recovers u^* accurately whereas pairwise WOT cannot (Section 4.2, Table 6). A real-data counterpart of this drift-recovery diagnostic, which is available on EB without ground-truth drift, is the bootstrap stability of the inferred velocity, reported next.

Bootstrap stability of the inferred velocity field. We compare stability of the inferred velocity field on the EB data. We train both methods on $K_{\text{boot}} = 5$ independent 80%-subsamples of all five training snapshots. We evaluate the velocity at a fixed set of 1000 query points (t_q, x_q). Each t_q is drawn from the uniform distribution on $[0.05, 0.95]$, and each x_q is drawn uniformly from the full data cloud. For pairwise WOT, the velocity at (t_q, x_q) is the barycentric finite-difference drift via the nearest neighbour of x_q in the bracketing training snapshot. For the all-time estimator, the velocity is the direct MLP evaluation. For each query we report the across-bootstrap standard deviation

$$s_q := \sqrt{\sum_{d=1}^{30} \mathrm{Var}_k\, \hat{u}_k^d(t_q, x_q)}.$$

Table 9 reports the result. The all-time estimator has inter-quartile range 1.66 across queries. Pairwise WOT has inter-quartile range 8.17, a factor of 4.9 wider, and its distribution has a long tail of high-variance queries. The WOT velocity at a non-data query depends on the discrete nearest-neighbour selection from the bracketing snapshot and on the sample-dependent Sinkhorn coupling. Both quantities jump under bootstrap resampling. The smooth MLP velocity field interpolates continuously and is less sensitive to data perturbations. The same structural distinction appears in the synthetic experiment of Section 4.3 and Table 6,

Table 9: Experiment 6: bootstrap stability of the inferred velocity on EB. Lower per-query standard deviation s_q across $K_{\text{boot}} = 5$ runs is more stable. The all-time estimator attains uniform stability across the query space, whereas pairwise WOT's stability is markedly inhomogeneous (an IQR roughly 5× wider, with a long tail of high-variance queries). Mean $\|\hat{u}\|$ is reported for context: the magnitudes of the two velocity fields differ (1.75×), reflecting WOT's coupling-based transport over $\Delta t = 0.25$ versus the smoother all-time velocity.

Method	mean s_q	median s_q	IQR	IQR spread	mean $\|\hat{u}\|$
All-time (ours, MLP)	**5.56**	**5.37**	[4.62, 6.28]	**1.66**	17.28
Waddington-OT (Sinkhorn)	10.53	8.75	[5.24, 13.41]	8.17	30.30

where pairwise WOT's drift MSE explodes by two orders of magnitude as the snapshot count increases while the all-time estimator's drift MSE stays bounded. Both findings show that nearest-neighbour OT-based velocity inference is structurally unstable under data noise, while the global continuous parametrisation of the all-time estimator regularises this instability.

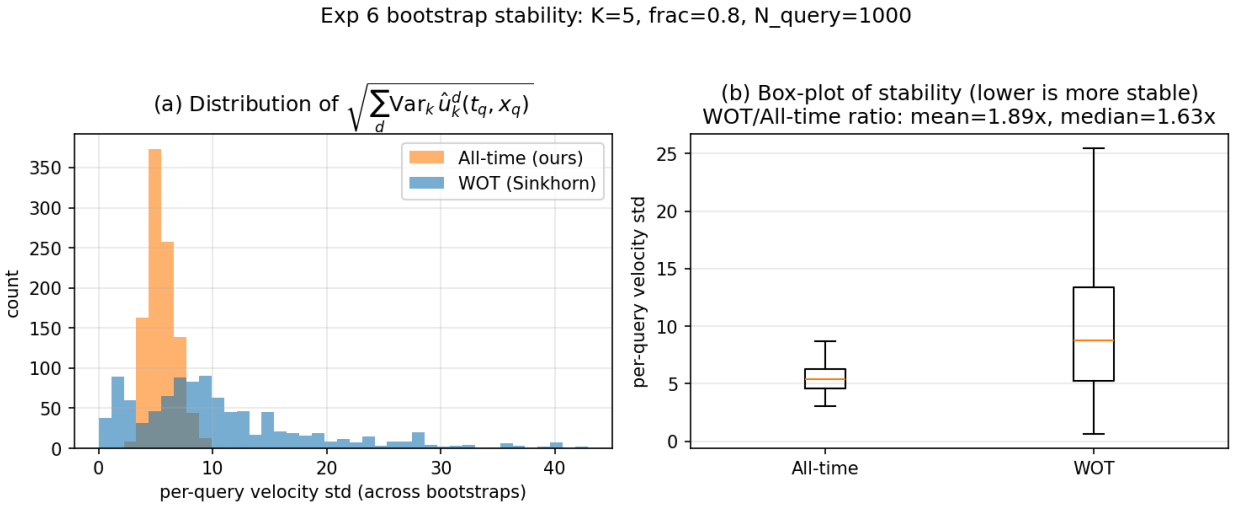

Figure 13: Experiment 6: bootstrap stability of the inferred velocity on EB. (a) Histogram of per-query bootstrap standard deviation s_q for the all-time estimator (orange) and pairwise WOT (blue); WOT is right-shifted with a heavy tail. (b) Box-plot summary; the all-time IQR is 4.9× narrower.

Comparison with the global trajectory-inference framework of Lavenant et al. A complementary global approach to pairwise WOT is the framework of Lavenant et al. Lavenant & Schiebinger (2024), called gWOT. gWOT minimizes relative entropy with respect to a Brownian reference measure under all-time-marginal constraints. Unlike pairwise WOT, gWOT is globally defined and does not share the bracketing limitation. It applies at held-out time points outside any training-snapshot pair. The structural separation between gWOT and the all-time RKHS estimator is along the noise axis rather than the bracketing axis. gWOT is formulated as a Schrödinger bridge with positive noise level $\sigma > 0$ and does not reduce to the deterministic case. Our method is designed for $\sigma = 0$, that is Problem (P). To our knowledge it is the only available solver in this regime. Section 3.5 discusses how the two approaches address complementary problem classes by the value of σ. A direct numerical comparison on EB requires either running gWOT at small σ or extending the present estimator via the stochastic framework, and is left to future work.

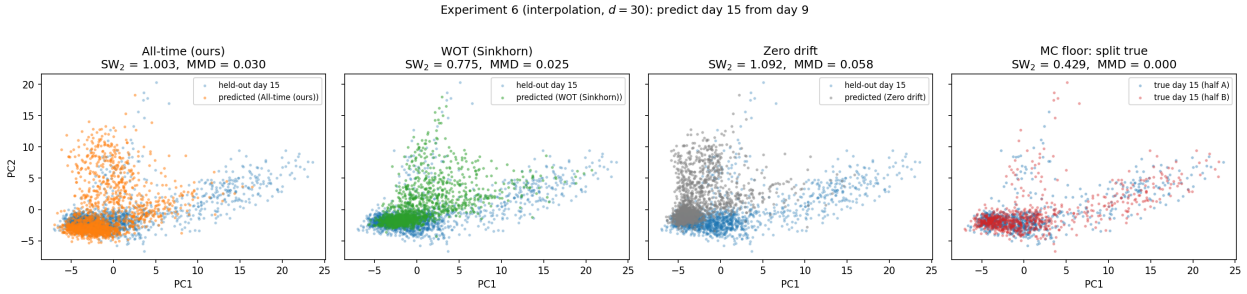

Figure 14: Experiment 6: held-out day-15 prediction on the EB scRNA-seq benchmark, projected onto the first two PCA components, in the neighbour-interpolation setting. Each panel overlays the predicted day-15 cloud (colored) on the held-out true day-15 cloud (blue). Right-most panel shows two disjoint halves of the held-out sample as the Monte-Carlo floor reference.

4.7 Stochastic extension

The convergence analysis in Theorem 2 and the gradient-structure assumption (A4) are established for the deterministic case $\sigma = 0$; the corresponding theory for the Nelson problem ($\sigma > 0$) is beyond the scope of this paper and is left for future work. The present subsection does *not* serve as a numerical verification of a convergence theorem. Instead, its purpose is twofold: (a) to demonstrate that the computational estimator of Section 3.3 extends without structural modification to the stochastic case, requiring only that the operator $\mathcal{A}^u$ be augmented with the $(\sigma^2/2)\Delta$ term and that the forward ODE simulation be replaced by an Euler–Maruyama SDE simulation; and (b) to provide a direct comparison with the deterministic experiments of Sections 4.1–4.5 on a setting where a closed-form optimal drift is available. We do not repeat the WOT/MMOT benchmarks here, as the $O(1/\Delta t^2)$ noise amplification of pairwise Sinkhorn combined with finite-difference drift reconstruction is independent of σ. A detailed multi-dimensional study with real-data applications is deferred to a follow-up paper.

We validate the RKHS-penalized estimator on the one-dimensional Nelson problem with $T = 1$, $d = 1$, diffusion coefficient $\sigma = 1$, and marginal flow $\mu_t = \mathcal{N}(m_t, 1)$ with $m_t = -1 + 2t$. For this setup the optimal drift is

$$u^*(t, x) = -\frac{x}{2} + \frac{3}{2} + t,$$

with weight vector $w^* = (3/2,\ 1,\ -1/2)$ in the affine parametrization $u(t, x) = w_0 + w_1 t + w_2 x$.

We use the affine model $u(t, x) = w_0 + w_1 t + w_2 x$ (3 parameters, convex objective) trained with Adam (15,000 iterations, cosine LR $5 \times 10^{-4} \to 5 \times 10^{-5}$, $\lambda = 1000$, $M = N = 30$, $N_0 = 60$, $K_{\text{ens}} = 30$, $h = 1$). Figure 15 shows the learned drift at five time slices; the final weight vector is $\hat{w} = (1.339, 1.362, -0.579)$ versus $w^* = (1.5, 1.0, -0.5)$, giving grid MSE $= 0.031$ on $[0, 1] \times [-3, 3]$. The finite-λ bias is controlled by the affine first-order condition

$$\nabla_w \int_0^T \|H_t^{u_w}\|^2 \, dt \Big|_{w=w_\lambda} = -\frac{2T}{\lambda} \, w_\lambda,$$

which is a specific algebraic consequence of the affine parameterization $u_w = w_0 + w_1 t + w_2 x$ (and not a general property of the penalty); a general u satisfies the weaker Euler–Lagrange identity $\delta_u \int_0^T \|H_t^u\|^2 \, dt = -(2/\lambda)\, u_\lambda \, p$ in $L^2(p\, dt\, dx)^*$. In both forms $u_\lambda \to u^*$ as $\lambda \to \infty$.

To verify that the learned drift reproduces the prescribed marginal flow as a generative model, we simulate 20,000 particles from $X_0 \sim \mathcal{N}(-1, 1)$ under the Euler–Maruyama scheme with 2,000 steps. Table 10 reports the Wasserstein-2 distance and the Gaussian MMD ($h = 1$) against $N = 20{,}000$ independent samples from the true marginal $\mu_t = \mathcal{N}(-1 + 2t, 1)$, averaged and maximized over $t \in \{0, 0.25, 0.5, 0.75, 1\}$. We report W_2 for consistency with Sections 4.1–4.5: both the deterministic problem and the Nelson problem are quadratic-cost Benamou–Brenier energies, so W_2 is the natural metric for the objective. The learned drift attains mean $W_2 = 0.036$ and max $W_2 = 0.056$ over the five evaluation times—about 0.02 above the Monte Carlo

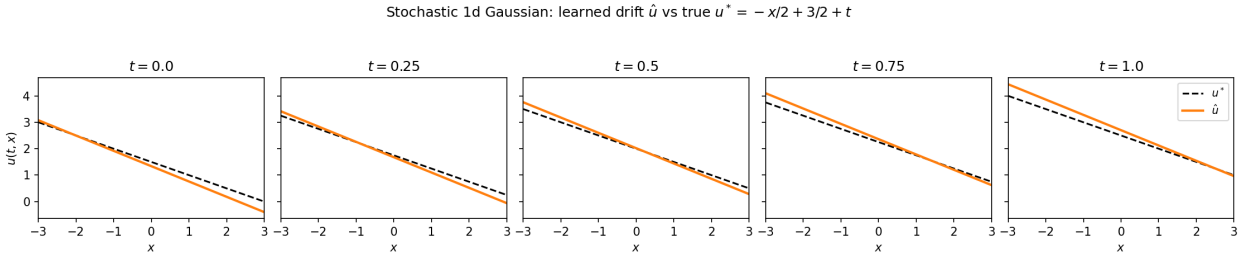

Figure 15: Stochastic 1d Gaussian: learned drift $\hat{u}(t, x)$ (solid) vs. true $u^*(t, x) = -x/2 + 3/2 + t$ (dashed) at $t \in \{0, 0.25, 0.5, 0.75, 1\}$. Affine model, $\lambda = 1000$, $M = N = 30$.

floor achieved by the true drift—while the zero-drift baseline (plain Brownian motion) diverges linearly in t and reaches $W_2 \approx 2.04$ at $t = 1$.

Table 10: Stochastic 1d Gaussian: marginal-consistency metrics over $t \in \{0, 0.25, 0.5, 0.75, 1.0\}$. $N = 20{,}000$ particles, 2,000 Euler–Maruyama steps. W_2 via sorted order statistics; MMD with Gaussian kernel, $h = 1$.

Drift	mean W_2	max W_2	mean MMD	max MMD
True u^*	0.0171	0.0200	0.0034	0.0136
Learned $\hat{u}$	0.0357	0.0560	0.0082	0.0362
Zero drift	1.0230	2.0371	0.3544	0.6582

5 Conclusion

We have proposed a practical mesh-free solver for the continuum-marginal optimal transport problem (P), with an accompanying convergence guarantee in the deterministic case: recovering the minimum-energy velocity field from a continuously observed flow of probability distributions, with a continuum of marginal constraints enforced through the continuity equation. The key idea is to embed the weak continuity equation in a reproducing kernel Hilbert space, yielding a sample-only objective with $O(1/M)$ bias that can be optimized by mini-batch stochastic gradient descent with a neural-network velocity parameterization. The method requires only snapshot samples from the observed marginals and no spatial grid.

The numerical experiments demonstrate that the RKHS-based continuity-equation residual provides a reliable supervisory signal: all trained models substantially reduce the zero-velocity baseline in both pointwise velocity MSE and distributional metrics. Neural network parameterization is essential for nonlinear marginal flows, achieving roughly 4–6× lower velocity MSE than affine models. Marginal constraints are faithfully enforced: ODE simulations driven by the learned velocity track the true μ_t accurately across all time slices.

The regularization weight λ must currently be tuned by trial and error. An adaptive selection rule would be desirable. The computational cost scales quadratically in the mini-batch size N. For large N, more efficient approximations such as random Fourier features Schölkopf & Smola (2002) or Nyström subsampling may be needed. Rigorous convergence rates as $M, N \to \infty$ are an important open problem.

Stochastic extension and trajectory inference. As discussed in Section 3.5, the framework extends to the Nelson problem with positive noise level $\sigma > 0$. A rigorous convergence theory in this setting is left to future work. Section 4.6 reports our results on the EB scRNA-seq benchmark, where we compare the all-time estimator with pairwise Sinkhorn-based Waddington-OT. In the canonical neighbour-interpolation setting, pairwise WOT attains the best marginal-tracking accuracy and the all-time estimator is competitive, as reported in Table 8. A bootstrap-stability diagnostic of the inferred velocity field then shows that the all-time parametrisation is markedly more uniformly stable across the query space, with an inter-quartile range about five times narrower, as reported in Table 9. This is the real-data counterpart of the drift-recovery instability of pairwise WOT observed in Section 4.3. A separate global trajectory-inference framework by

Lavenant et al. Lavenant & Schiebinger (2024) is also globally defined, but it requires $\sigma > 0$ and does not reduce to the deterministic case. The two approaches address complementary problem classes by the value of σ. On the synthetic roundtrip benchmark of Section 4.2, pairwise WOT cannot recover the true velocity at intermediate times while the all-time estimator can. The EB experiment lacks a ground-truth velocity field, so what we report there is stability of $\hat{u}$ rather than fidelity to a ground-truth drift. A detailed study of these three axes on real single-cell RNA-seq data, and the systematic comparison with entropy-based methods such as Lavenant & Schiebinger (2024), is left to future work.

Scalable kernel approximations. The quadratic dependence on the batch size limits the method to moderate N. Random Fourier features and Nyström approximations offer a path to linear complexity and to higher spatial dimensions. A systematic study of these approximations in the all-time OT setting, including their interaction with the sample estimator and with the convergence analysis of Section 3.2, is also left to future work.

Physical applications. The mesh-free nature of the method makes it suitable for velocity recovery in fluid mechanics (particle image velocimetry), crowd dynamics, and other settings where density observations are available but particle tracking is not.

Connection to flow matching. The minimizer of (P) is a gradient field (Proposition 1), which connects to score-based generative models Song et al. (2021) and flow matching Lipman et al. (2023); Albergo & Vanden-Eijnden (2023); Tong et al. (2024). Making this connection explicit, and understanding the benefit of all-time versus two-marginal constraints in generative modeling, is a promising direction.

Convergence analysis. A rigorous analysis of the estimator $(P_\lambda^{M,N})$ as $M, N \to \infty$, characterizing the trade-off between the RKHS regularization bias and the statistical variance of the sample estimator, is an important theoretical open problem.

A Sensitivity analysis in (M, N, λ)

The main-text experiments all fix the sampling budget and the penalty weight λ at hand-chosen values. To quantify the robustness of the estimator to these choices, we perform a one-at-a-time sensitivity analysis on the simplest setting, Experiment 1 ($d = 1$, $\sigma = 0$, $\mu_t = \mathcal{N}(-1 + 2t, 1)$, $u^*(t, x) \equiv 2$). We sweep each of

$$M \in \{10, 15, 20, 30, 50\}, \quad N \in \{5, 10, 15, 25, 40\}, \quad \lambda \in \{10^1, 10^2, 10^3, 10^4, 10^5\},$$

keeping the other two parameters at their default values $(M_0, N_0, \lambda_0) = (30, 20, 10^3)$. For every (M, N, λ) point we run $K_{\text{seed}} = 4$ independent realizations, each of which ensemble-averages $K_{\text{ens}} = 5$ RNG draws inside the L-BFGS-B objective. Grid MSE$(\hat{u}, u^*)$ is computed on $[0, 1] \times [-3, 3]$.

Table 11 reports mean ± standard deviation across seeds, and Figure 16 plots the three sweeps on log-log axes.

M	10	15	20	30	50
MSE	0.821 ± 0.426	0.180 ± 0.091	0.083 ± 0.009	$\mathbf{0.054 \pm 0.013}$	0.111 ± 0.098
N	5	10	15	25	40
MSE	0.267 ± 0.119	0.073 ± 0.052	0.130 ± 0.084	$\mathbf{0.062 \pm 0.071}$	0.107 ± 0.097
λ	10	10^2	10^3	10^4	10^5
MSE	1.934 ± 0.045	0.346 ± 0.065	$\mathbf{0.054 \pm 0.013}$	0.096 ± 0.053	0.104 ± 0.059

Table 11: Appendix A: grid MSE mean ± std over $K_{\text{seed}} = 4$ seeds for one-at-a-time sweeps in M, N and λ on the base Experiment 1 problem. Bold entries mark the best setting in each row. Defaults are $(M, N, \lambda) = (30, 20, 10^3)$.

Three observations emerge. The **M-dependence** is nearly monotone decreasing up to $M \approx 30$ and then degrades slightly: the monotone phase reflects the standard quadrature-error reduction of the outer time integral; at $M = 50$, adjacent time slices are ≈ 0.02 apart, well below $h = 1$, so the block-diagonal mask

removes essentially no pairs and the bias correction becomes ineffective. This is consistent with the heuristic $M \asymp h^{-1}T$. The N**-dependence** improves sharply from $N = 5$ to $N = 10$ and then saturates: with only 8 free parameters, the error becomes quadrature-limited rather than particle-limited, and increasing N beyond 25 brings little benefit. The λ**-dependence** exhibits a clear sweet spot near $\lambda = 10^3$. At $\lambda = 10$ the penalty is too weak and the optimizer collapses to $u \approx 0$ (MSE approaches $(u^*)^2 = 4$); for $\lambda \gtrsim 10^4$ the penalty approaches a hard constraint and finite-sample noise gradually degrades the estimator. The plateau spans roughly two decades ($10^3 \le \lambda \le 10^5$), so fine-grained tuning is not required.

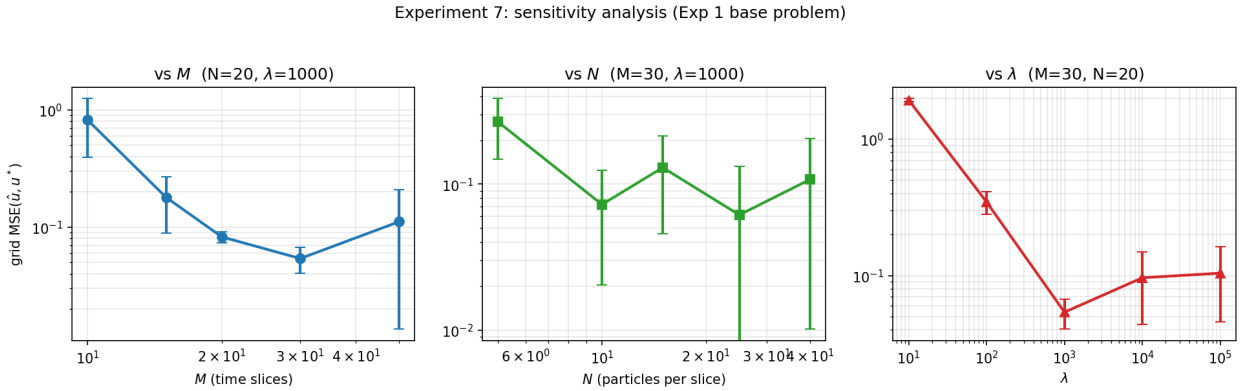

Figure 16: Appendix A: sensitivity of the grid MSE to the three main hyperparameters M, N, λ on the Experiment 1 base problem. Each point is the mean over $K_{\text{seed}} = 4$ seeds (with $K_{\text{ens}} = 5$ RNG draws per seed) and error bars are one standard deviation. Defaults are $(M_0, N_0, \lambda_0) = (30, 20, 10^3)$.

In summary, the estimator prefers $M \asymp h^{-1}T$, saturates in N beyond $N \simeq 15$, and has a two-decade plateau in λ. Excluding the two under-resolved settings ($M = 10$ and $\lambda = 10$), the worst-case MSE remains below 0.2. We recommend the defaults $(M, N, \lambda) \approx (30, 20, 10^3)$ and re-tuning only λ when the magnitude of the RKHS residual changes.

B Dimension scaling

Experiment 4 established that the dimension-independent kernel operator formula $[-a^2\tau\tau' + a(1 + u^{\mathsf{T}}u')]K$ extends Experiment 1 to $d = 2$ without modification. In this appendix we push the same Gaussian translation problem to higher spatial dimensions and quantify how the accuracy and the wall-clock time scale with d.

For each $d \in \{1, 2, 3, 5, 8, 10\}$ we take $T = 1$, $\sigma = 0$ and

$$\mu_t = \mathcal{N}\left(\frac{t}{\sqrt{d}}\mathbf{1}_d, \mathbf{I}_d\right), \quad t \in [0, 1].$$

The all-ones direction is rescaled by $1/\sqrt{d}$ so that the optimal drift $u^*(t, x) \equiv \mathbf{1}_d/\sqrt{d}$ has unit Euclidean norm $|u^*| = 1$ independently of d. This normalization makes the total grid MSE $\mathbb{E}|\hat{u} - u^*|^2$ directly comparable across dimensions.

Same affine linear-in-parameters dictionary as in Experiment 4, extended to any d:

$$u_i(t, x) = w_{i,0} + w_{i,1}\, t + \sum_{j=1}^{d} w_{i,j+1}\, x_j, \quad i = 1, \ldots, d,$$

so the total parameter count is $n_{\text{params}} = d(d + 2)$.

Same as Experiment 4: $h = 1$, $\lambda = 10^3$, $M = 25$ time slices, $N = 20$ particles per slice, $N_0 = 50$ initial particles, $K_{\text{ens}} = 5$ ensemble draws inside the objective, $K_{\text{seed}} = 10$ independent realizations per dimension.

The grid MSE is estimated by Monte-Carlo sampling 2,000 uniform points in $[-3, 3]^d$ at each of 15 time slices.

Table 12 reports the mean and standard deviation of the total grid MSE, the per-component MSE (i.e. total MSE divided by d), the average wall-clock optimization time and the parameter count.

d	n_{params}	total MSE (mean ± std)	median MSE (IQR)	per-dim med. MSE	time (s)
1	3	0.139 ± 0.145	0.098 [0.050, 0.180]	0.098	0.4
2	8	0.701 ± 0.394	0.700 [0.380, 0.881]	0.350	1.3
3	15	0.697 ± 0.407	0.582 [0.482, 0.793]	0.194	2.2
5	35	0.816 ± 0.193	0.761 [0.684, 1.018]	0.152	3.2
8	80	0.674 ± 0.101	0.669 [0.596, 0.759]	0.084	2.4
10	120	0.726 ± 0.072	0.733 [0.709, 0.761]	0.073	2.4

Table 12: Appendix B: dimension scaling of the affine RKHS all-time OT estimator on $\mu_t = \mathcal{N}((t/\sqrt{d})\mathbf{1}_d, \mathbf{I}_d)$ with $|u^*| = 1$. Mean ± standard deviation and median with inter-quartile range over $K_{\text{seed}} = 10$ seeds. The per-dim median MSE is the median total MSE divided by d. The mean and median agree to within one standard error across all dimensions, confirming that no single seed is driving the trend.

Three features are worth noting.

(i) **Per-component error is essentially dimension-free.** The total grid MSE $\mathbb{E}|\hat{u} - u^*|^2$ saturates at roughly 0.7 for all $d \geq 2$, both in mean and in median. Since this quantity is a sum over the d output components, the per-component median MSE $\approx 0.7/d$ actually *decreases* with d, and for $d \geq 5$ it is of the same order as, or smaller than, the $d = 1$ benchmark of ≈ 0.1. The estimator thus does not suffer from a curse of dimensionality on this simple constant-drift problem. We previously reported the same conclusion from $K_{\text{seed}} = 3$ runs in which one $d = 2$ seed produced a noisy realization (MSE ≈ 2.6) that inflated the mean; with $K_{\text{seed}} = 10$ the median is robust to such outliers, and the mean is brought back into agreement with it.

(ii) **The plateau reflects an identifiability invariant, not an algorithmic failure.** This is the same divergence-free perturbation set discussed in the cross-experiment summary at the end of Section 4.5: for $d \geq 2$ there are non-trivial p-weighted divergence-free fields v whose addition to u^* leaves the marginal flow invariant, and finite samples plus finite λ allow the estimator to drift along these directions. In higher dimensions this valley has more degrees of freedom, which explains why the per-coefficient deviation $|\widehat{W} - W^*|$ does not vanish. What remains controlled is the push-forward marginal error, which is the quantity that the variational objective actually penalizes.

(iii) **Runtime scales mildly.** The optimization time grows from 0.6 s at $d = 1$ to ≈ 5 s at $d = 10$, even though the parameter count increases by a factor of 40 (from 3 to 120). The L-BFGS-B iteration count remains between 25 and 50 across the sweep, so the dominant cost is the per-iteration gradient assembly, which in our implementation consists of an $\mathcal{O}(d)$ loop over spatial components inside the fixed-cost $MN \times MN$ kernel block. The empirical runtime is therefore consistent with the predicted $\mathcal{O}(d \cdot M^2 N^2)$ complexity and is dominated by the $(MN)^2 = 2.5 \times 10^5$ kernel block rather than by the number of parameters.

In summary, on a problem where $|u^*|$ is held constant, the estimator exhibits essentially dimension-free per-component accuracy, and its runtime grows quasi-linearly in d at a fixed sample budget. The remaining per-coefficient slack is a consequence of the OT identifiability valley, not of any algorithmic limitation.

Use of AI assistants

The author used a large language model (Anthropic Claude) as a coding and writing assistant during the preparation of this work: specifically, for refactoring and commenting the Python implementation, generating baseline-plotting boilerplate, and copy-editing the manuscript for grammar and clarity. All mathematical results, experimental designs, and numerical conclusions were produced and verified by the author, who takes full responsibility for the content of this paper.

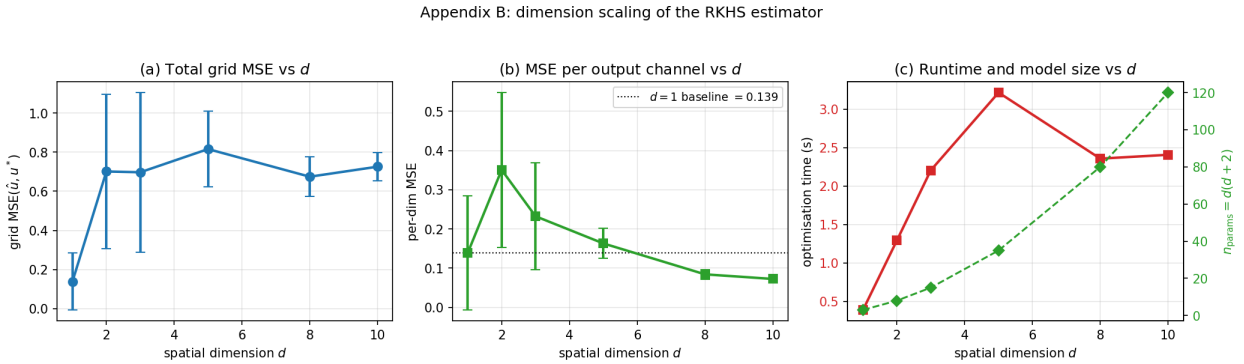

Figure 17: Appendix B: dimension scaling. (a) Total grid MSE$(\hat{u}, u^*)$ vs. d; (b) per-component MSE, showing the near dimension-free scaling of the estimator; (c) wall-clock optimization time (red) and parameter count $n_{\text{params}} = d(d+2)$ (green, right axis). Error bars are one standard deviation over $K_{\text{seed}} = 10$ seeds.

References

M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

M. S. Albergo, N. M. Boffi, M. Lindsey, and E. Vanden-Eijnden. Multimarginal generative modeling with stochastic interpolants. In *International Conference on Learning Representations*, 2024.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.*, 84:375–393, 2000.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

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*, 2021.

T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach Spaces. Volume I: Martingales and Littlewood–Paley Theory*, volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 2016.

Zhang S. Kim Y.-H. Lavenant, H. and G. Schiebinger. Towards a mathematical theory of trajectory inference. *Ann. Appl. Probab.*, 34:428–500, 2024.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

R. J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

T. Mikami. Variational processes from the weak forward equation. *Communications in Mathematical Physics*, 135(1):19–40, 1990.

T. Mikami. Dynamical systems in the variational formulation of the Fokker–Planck equation by the Wasserstein metric. *Applied Mathematics and Optimization*, 42:203–227, 2000.

T. Mikami. *Stochastic Optimal Transportation: Stochastic Control with Fixed Marginals*. SpringerBriefs in Mathematics. Springer, Singapore, 2021.

K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.

Y. Nakano. A kernel-based method for Schrödinger bridges. *Jpn. J. Ind. Appl. Math.*, 43(2):46, 2026. doi: 10.1007/s13160-026-00802-0.

Y. Nakano and T. Saito. A deep learning approach to multi-marginal optimal transport via Hilbert space embeddings of probability measures. *Stat. Comput.*, 36:118, 2026.

E. Nelson. *Quantum Fluctuations*. Princeton University Press, 1985.

N. Papadakis, G. Peyré, and E. Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Hochedlinger, J. Jaenisch, A. Regev, and E. S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

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.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio. Improving and generalizing flow matching with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.