A Computational Framework for Solving Wasserstein Lagrangian Flows

Kirill Neklyudov^{*,1}**, Rob Brekelmans**^{*,1}**, Alexander Tong**^{4}**, Lazar Atanackovic**^{1,2}
Qiang Liu^{3}**, Alireza Makhzani**^{1,2}
^{1} Vector Institute ^{2} University of Toronto ^{3} UT Austin
^{4} Mila – Quebec AI Institute, Université de Montréal

Abstract

The dynamical formulation of the optimal transport can be extended through various choices of the underlying geometry (*kinetic energy*), and the regularization of density paths (*potential energy*). These combinations yield different variational problems (*Lagrangians*), encompassing many variations of the optimal transport problem such as the Schrödinger bridge, unbalanced optimal transport, and optimal transport with physical constraints, among others. In general, the optimal density path is unknown, and solving these variational problems can be computationally challenging. Leveraging the dual formulation of the Lagrangians, we propose a novel deep learning based framework approaching all of these problems from a unified perspective. Our method does not require simulating or backpropagating through the trajectories of the learned dynamics, and does not need access to optimal couplings. We showcase the versatility of the proposed framework by outperforming previous approaches for the single-cell trajectory inference, where incorporating prior knowledge into the dynamics is crucial for correct predictions.

1 Introduction

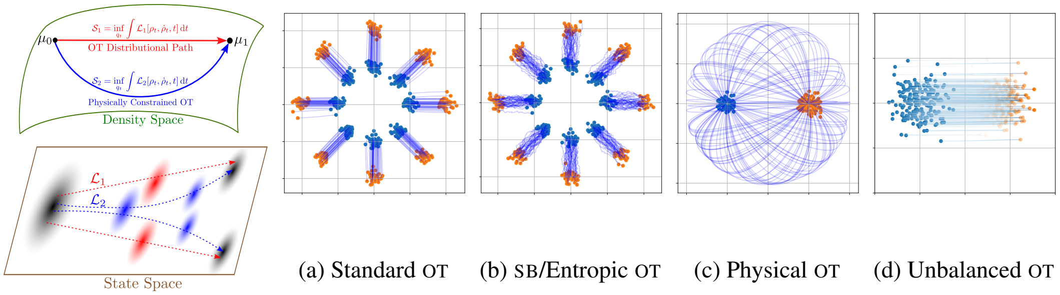
The problem of *trajectory inference*, or recovering the population dynamics of a system from samples of its temporal marginal distributions, is a problem arising throughout the natural sciences [25, 29]. A particularly important application is analysis of single-cell RNA-sequencing data [58, 57, 55], which provides a heterogeneous snapshot of a cell population at a high resolution, allowing high-throughput observation over tens of thousands of genes [43]. However, since the measurement process ultimately leads to cell death, we can only observe temporal changes of the *marginal* or *population* distributions of cells as they undergo treatment, differentiation, or developmental processes of interest. To understand these processes and make future predictions, we are interested in both (i) interpolating the evolution of marginal cell distributions between observed timepoints and (ii) modeling the full trajectories at the individual cell level.

However, when inferring trajectories over cell distributions, there exist multiple cell dynamics that yield the same population marginals. This presents an ill-posed problem, which highlights the need for trajectory inference methods to be able to flexibly incorporate different types of prior information on the cell dynamics. Commonly, such prior information is specified via posing a variational problem on the space of marginal distributions, where previous work on measure-valued splines [14, 8, 16, 20, 13] are examples which seek minimize the acceleration of particles.

We propose a general framework for using deep neural networks to infer dynamics and solve marginal interpolation problems, using Lagrangian action functionals on manifolds of probability

[*] Joint main-authorship. Contact: k.necludov@gmail.com; {rob.brekelmans, makhzani}@vectorinstitute.ai.

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



(a) Standard OT (b) SB/Entropic OT (c) Physical OT (d) Unbalanced OT

Figure 1: Our *Wasserstein Lagrangian Flows* are action-minimizing curves for various choices of Lagrangian $\mathcal{L}_i[\rho_t, \dot\rho_t, t]$ on the space of densities, which each translate to optimal state-space dynamics. Toy examples of dynamics resulting from various potential or kinetic energy terms are given in (a)-(d). We may also constrain Wasserstein Lagrangian flows to match intermediate data marginals $\rho_{t_i} = \mu_{t_i}$ and combine energy terms to define a suitable notion of interpolation between given μ_{t_i}.

densities that can flexibly incorporate various types of prior information. We consider Lagrangians of the form $\mathcal{L}[\rho_t, \dot\rho_t, t] = \mathcal{K}[\rho_t, \dot\rho_t, t] - \mathcal{U}[\rho_t, t]$, referring to the first term as a *kinetic energy* and the second as a *potential energy*. Our methods can be used to solve a diverse family of problems defined by the choice of these energies and constraints on the evolution of ρ_t. More explicitly, we specify

- A *kinetic energy* which, in the primary examples considered in this paper, corresponds to a *geometry* on the space of probability measures. We primarily consider the Riemannian structures corresponding to the Wasserstein-2 and Wasserstein Fisher-Rao metrics.
- A *potential energy*, which is a functional of the density, for example the expectation of a physical potential encoding prior knowledge or even a nonlinear functional.
- A collection of *marginal* constraints which are inspired by the availability of data in the problem of interest. For optimal transport (OT), Schrödinger Bridge (SB), or generative modeling tasks, we are often interested in interpolating between two endpoint marginals given by a data distribution and/or a tractable prior distribution. For applications in trajectory inference, we may incorporate multiple constraints to match the observed temporal marginals, given via data samples. Notably, in the limit of data sampled infinitely densely in time, we recover the Action Matching (AM) framework of Neklyudov et al. [47].

Within our *Wasserstein Lagrangian Flows* framework, we propose tractable dual objectives to solve (i) *standard Wasserstein-2* OT (Ex. 4.1, Benamou & Brenier [7], Villani [66]), (ii) *entropy regularized* OT or Schrödinger Bridge (Ex. 4.4, Léonard [31], Chen et al. [15], (iii) *physically constrained* OT (Ex. 4.3, Tong et al. [61], Koshizuka & Sato [28]), and (iv) *unbalanced* OT (Ex. 4.2, Chizat et al. [17]) (Sec. 4). Our framework also allows for combining energy terms to incorporate features of the above problems as inductive biases for trajectory inference. In Sec. 5, we showcase the ability of our methods to accurately solve Wasserstein Lagrangian flow optimizations, and highlight how testing different Lagrangians can improve results in single-cell RNA-sequencing applications. We discuss benefits of our approach compared to related work in Sec. 6.

2 Background

2.1 Wasserstein-2 Geometry

For two given densities with finite second moments $\mu_0, \mu_1 \in \mathcal{P}_2(\mathcal{X})$, the Wasserstein-2 OT problem is defined, in the Kantorovich formulation, as a cost-minimization problem over joint distributions or 'couplings' $\pi \in \Pi(\mu_0, \mu_1) = \{\pi(x_0, x_1) \mid \int \pi(x_0, x_1) dx_1 = \mu_0,\ \int \pi(x_0, x_1) dx_0 = \mu_1\}$, i.e.

$$W_2(\mu_0, \mu_1)^2 := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int \|x_0 - x_1\|^2 \pi(x_0, x_1) dx_0 dx_1 \,. \tag{1}$$

The dynamical formulation of Benamou & Brenier [7] gives an alternative perspective on the W_2 OT problem as an optimization over a vector field v_t that transports samples according to an ODE $\dot x_t = v_t$. The evolution of the samples' density ρ_t, under transport by v_t, is governed by the *continuity equation* $\dot\rho_t = -\nabla \cdot (\rho_t v_t)$ (Figalli & Glaudo [23] Lemma 4.1.1), and we have

$$W_2(\mu_0, \mu_1)^2 = \inf_{\rho_t} \inf_{v_t} \int_0^1 \int \frac{1}{2} \|v_t\|^2 \rho_t \, dx_t dt \quad \text{s.t. } \dot\rho_t = -\nabla \cdot (\rho_t v_t), \quad \rho_0 = \mu_0, \ \rho_1 = \mu_1, \tag{2}$$

where $\nabla \cdot ()$ is the divergence operator. The W_2 transport cost can be viewed as providing a Riemannian manifold structure on $\mathcal{P}_2(\mathcal{X})$ (Otto [48], Ambrosio et al. [3], see also Figalli & Glaudo [23] Ch. 4). Introducing Lagrange multipliers s_t to enforce the constraints in Eq. (2), we obtain the condition

$v_t = \nabla s_t$ (see App. B.1), which is suggestive of the result from Ambrosio et al. [3] characterizing the tangent space $T_\rho^{W_2}\mathcal{P}_2 = \{\dot\rho \mid \int \dot\rho \, dx_t = 0\}$ via the continuity equation,

$$T_\rho^{W_2}\mathcal{P}_2(\mathcal{X}) = \{\dot\rho \mid \dot\rho = -\nabla \cdot (\rho \nabla s)\}. \tag{3}$$

We also write the cotangent space as $T_\rho^{*W_2}\mathcal{P}_2(\mathcal{X}) = \{[s] \mid s \in \mathcal{C}^\infty(\mathcal{X})\}$, where $\mathcal{C}^\infty(\mathcal{X})$ denotes smooth functions and $[s]$ is an equivalence class up to addition by a constant. For two curves $\mu_t, \rho_t : [-\epsilon, \epsilon] \mapsto \mathcal{P}_2(\mathcal{X})$ passing through $\rho := \rho_0 = \mu_0$, the Otto metric is defined

$$\langle \dot\mu_t, \dot\rho_t \rangle_{T_\rho}^{W_2} = \langle s_{\dot\mu_t}, s_{\dot\rho_t} \rangle_{T_\rho^*}^{W_2} = \int \langle \nabla s_{\dot\mu_t}, \nabla s_{\dot\rho_t} \rangle \rho \, dx. \tag{4}$$

2.2 Wasserstein Fisher-Rao Geometry

Building from the dynamical formulation in Eq. (2), Chizat et al. [17, 18], Kondratyev et al. [26], Liero et al. [33, 34] consider additional terms allowing for birth and death of particles, or teleportation of probability mass. In particular, consider extending the continuity equation to include a 'growth term' $g_t : \mathcal{X} \to \mathbb{R}$ whose norm is regularized in the cost,

$$WFR_\lambda(\mu_0, \mu_1)^2 = \inf_{\rho_t} \inf_{v_t, g_t} \int_0^1 \int \left(\frac{1}{2}\|v_t\|^2 + \frac{\lambda}{2} g_t^2 \right) \rho_t \, dx_t dt, \tag{5}$$

subject to $\dot\rho_t = -\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t, \rho_0 = \mu_0, \rho_1 = \mu_1$. We call this the Wasserstein Fisher-Rao (WFR) distance, since considering *only* the growth terms recovers the non-parametric Fisher-Rao metric [17, 6]. We also refer to Eq. (5) as the *unbalanced* OT problem on the space of unnormalized densities $\mathcal{M}(\mathcal{X})$, since the growth terms need not preserve normalization $\int \dot\rho_t dx_t = \int \lambda g_t \rho_t dx_t \neq 0$ without further modifications (see e.g. Lu et al. [42]).

Kondratyev et al. [26] define a Riemannian structure on $\mathcal{M}(\mathcal{X})$ via the WFR distance. Introducing Lagrange multipliers s_t and eliminating v_t, g_t in Eq. (5) yields the optimality conditions $v_t = \nabla s_t$ and $g_t = s_t$. In analogy with Sec. 2.1, this suggests characterizing the tangent space via the tuple $(s_t, \nabla s_t)$ and defining the metric as a characterization of the tangent space

$$T_\rho^{WFR_\lambda}\mathcal{M}(\mathcal{X}) = \{\dot\rho \mid \dot\rho = -\nabla \cdot (\rho \nabla s) + \lambda \rho s\} \tag{6}$$

$$\langle \dot\mu_t, \dot\rho_t \rangle_{T_\rho}^{WFR_\lambda} = \langle s_{\dot\mu_t}, s_{\dot\rho_t} \rangle_{T_\rho^*}^{WFR_\lambda} = \int \big(\langle \nabla s_{\dot\mu_t}, \nabla s_{\dot\rho_t} \rangle + \lambda \, s_{\dot\mu_t} s_{\dot\rho_t} \big) \rho \, dx. \tag{7}$$

2.3 Action Matching

Finally, Action Matching (AM) [47] considers only the inner optimizations in Eq. (2) or Eq. (5) as a function of v_t or (v_t, g_t), assuming a distributional path μ_t is given via samples. In the W_2 case, to solve for the velocity $v_t = \nabla s_{\dot\mu_t}$ which corresponds to μ_t via the continuity equation or Eq. (3), Neklyudov et al. [47] optimize the objective

$$\mathcal{A}[\mu_t] = \sup_{s_t} \int s_1 \mu_1 \, dx_1 - \int s_0 \mu_0 \, dx_0 - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 \right) \mu_t \, dx_t dt, \tag{8}$$

over $s_t : \mathcal{X} \times [0,1] \to \mathbb{R}$ parameterized by a neural network, with similar objectives for WFR_λ. To foreshadow our exposition in Sec. 3, we view Action Matching as maximizing a lower bound on the *action* $\mathcal{A}[\mu_t]$ or *kinetic energy* of the curve $\mu_t : [0,1] \to \mathcal{P}_2(\mathcal{X})$ of densities. In particular, at the optimal $s_{\dot\mu_t}$ satisfying $\dot\mu_t = -\nabla \cdot (\mu_t \nabla s_{\dot\mu_t})$, the value of Eq. (8) becomes

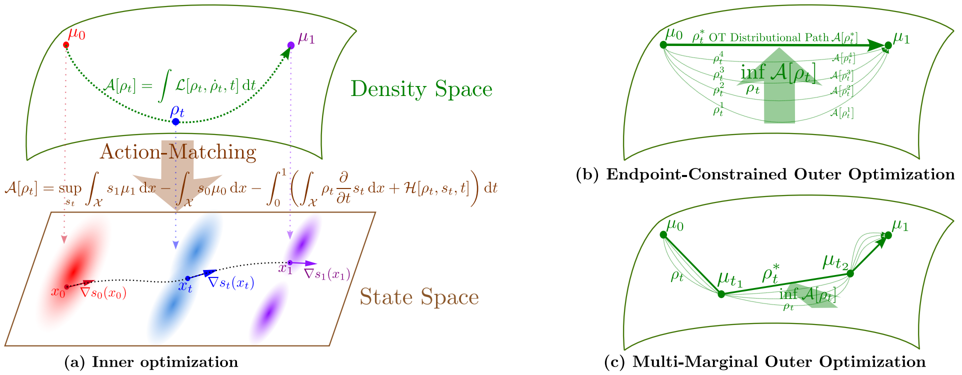
$$\mathcal{A}[\mu_t] = \int_0^1 \frac{1}{2} \langle \dot\mu_t, \dot\mu_t \rangle_{T_{\mu_t}}^{W_2} dt = \int_0^1 \frac{1}{2} \langle s_{\dot\mu_t}, s_{\dot\mu_t} \rangle_{T_{\mu_t}^*}^{W_2} dt = \int_0^1 \int \frac{1}{2} \|\nabla s_t\|^2 \mu_t \, dx_t dt. \tag{9}$$

Our proposed framework for Wasserstein Lagrangian Flows considers minimizing the action functional over distributional paths, and our computational approach will include AM as a component.

3 Wasserstein Lagrangian Flows

In this section, we develop computational methods for optimizing Lagrangian action functionals on the space of (unnormalized) densities $\mathcal{P}(\mathcal{X})$.[2] Lagrangian actions are commonly used to define a cost function on the ground space $\mathcal{X}$, which is then 'lifted' to the space of densities via an optimal transport distance (Villani [66] Ch. 7). We propose to formulate Lagrangians $\mathcal{L}[\rho_t, \dot\rho_t, t]$ *directly* in the density space, which includes OT with ground-space Lagrangian costs as a special case (App. B.1.2), but *also* allows us to consider kinetic and potential energies which depend on the density and thus cannot be expressed using a ground-space Lagrangian. In particular, we consider kinetic energies capturing space-dependent birth-death terms (as in WFR_λ, Ex. 4.2) and potential energies capturing information about the distribution of particles (as in the SB problem, Ex. 4.4).

[2] For convenience, we describe our methods using a generic $\mathcal{P}(\mathcal{X})$ (which may represent $\mathcal{P}_2(\mathcal{X})$ or $\mathcal{M}(\mathcal{X})$).


μ_0
μ_1
$\mathcal{A}[\rho_t] = \int \mathcal{L}[\rho_t, \dot\rho_t, t]\,\mathrm{d}t$
Density Space
ρ_t
Action-Matching
$\mathcal{A}[\rho_t] = \sup_{s_t} \int_{\mathcal{X}} s_1\mu_1\,\mathrm{d}x - \int_{\mathcal{X}} s_0\mu_0\,\mathrm{d}x - \int_0^1 \Big(\int_{\mathcal{X}} \rho_t \frac{\partial}{\partial t} s_t\,\mathrm{d}x + \mathcal{H}[\rho_t, s_t, t]\Big)\,\mathrm{d}t$
x_0
$\nabla s_0(x_0)$
x_t
$\nabla s_t(x_t)$
x_1
$\nabla s_1(x_1)$
State Space
μ_0
ρ_t^* OT Distributional Path $\mathcal{A}[\rho_t^*]$
μ_1
ρ_t^4
ρ_t^3
ρ_t^2
ρ_t^1
$\inf_{\rho_t} \mathcal{A}[\rho_t]$
$\mathcal{A}[\rho_t^4]$
$\mathcal{A}[\rho_t^3]$
$\mathcal{A}[\rho_t^2]$
$\mathcal{A}[\rho_t^1]$
μ_0
μ_1
ρ_t
μ_{t_1}
ρ_t^*
μ_{t_2}
$\inf_{\rho_t} \mathcal{A}[\rho_t]$

(a) Inner optimization

(b) Endpoint-Constrained Outer Optimization

(c) Multi-Marginal Outer Optimization

Figure 2: For different definitions of Lagrangian $\mathcal{L}[\rho_t, \dot\rho_t, t]$ or Hamiltonian $\mathcal{H}[\rho_t, s_t, t]$ on the space of densities, we obtain different action functionals $\mathcal{A}[\rho_t]$. Here, we show state-space velocity and optimal density paths for the W_2 geometry and OT problem. (a) The action functional for each curve can be evaluated using Action Matching (inner optimization in Thm. 1), which is performed in the state-space. (b,c) Minimization of the action functional (outer optimization in Thm. 1) is performed on the space of densities satisfying two endpoint constraints and possible intermediate constraints.

3.1 Wasserstein Lagrangian and Hamiltonian Flows

We consider Lagrangian action functionals on the space of densities, defined in terms of a *kinetic energy* $\mathcal{K}[\rho_t, \dot\rho_t, t]$, which captures any dependence on the velocity of a curve $\dot\rho_t$, and a *potential energy* $\mathcal{U}[\rho_t, t]$ which depends only on the position ρ_t,

$$\mathcal{L}[\rho_t, \dot\rho_t, t] = \mathcal{K}[\rho_t, \dot\rho_t, t] - \mathcal{U}[\rho_t, t]. \tag{10}$$

Throughout, we will assume $\mathcal{L}[\rho_t, \dot\rho_t, t]$ is lower semi-continuous (lsc) and strictly convex in $\dot\rho_t$.

Our goal is to solve for *Wasserstein Lagrangian Flows*, by optimizing the given Lagrangian over curves of densities $\rho_t : [0,1] \to \mathcal{P}(\mathcal{X})$ which are constrained to pass through M given points $\mu_{t_i} \in \mathcal{P}(\mathcal{X})$ at times t_i. We define the *action* of a curve $\mathcal{A}_{\mathcal{L}}[\rho_t] = \int_0^1 \mathcal{L}[\rho_t, \dot\rho_t, t]dt$ as the time-integral of the Lagrangian and seek the action-minimizing curve subject to the constraints

$$\mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) \coloneqq \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \mathcal{A}_{\mathcal{L}}[\rho_t] \coloneqq \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot\rho_t, t]dt \quad \text{s.t.} \;\; \rho_{t_i} = \mu_{t_i} \quad \forall\; 0 \le i \le M-1 \tag{11}$$

where $\Gamma(\{\mu_{t_i}\}) = \{\rho_t : [0,1] \to \mathcal{P}(\mathcal{X}) \mid \rho_0 = \mu_0,\; \rho_1 = \mu_1,\; \rho_{t_i} = \mu_{t_i} \quad (\forall\; 1 \le i \le M-2)\}$ indicates the set of curves matching given constraints. We note $M = 2$ as an important special case.

Our objectives for solving Eq. (11) are based on the Hamiltonian $\mathcal{H}$ associated with the chosen Lagrangian. In particular, consider a cotangent vector $s_t \in T^*\mathcal{P}_2(\mathcal{X})$ or $s_t \in T^*\mathcal{M}(\mathcal{X})$, which is identified with a linear functional on the tangent space $s_t[\cdot] : \dot\rho_t \mapsto \int s_t \dot\rho_t dx_t$ via the canonical duality bracket. We define the *Hamiltonian* $\mathcal{H}[\rho_t, s_t, t]$ via the Legendre transform

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{\dot\rho_t \in T_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t\, dx_t - \mathcal{L}[\rho_t, \dot\rho_t, t] = \mathcal{K}^*[\rho_t, s_t, t] + \mathcal{U}[\rho_t, t], \tag{12}$$

where the sign of $\mathcal{U}[\rho_t, t]$ changes and $\mathcal{K}^*[\rho_t, s_t, t]$ translates the kinetic energy to the dual space. A primary example is when $\mathcal{K}[\rho_t, \dot\rho_t, t] = \frac{1}{2}\langle \dot\rho_t, \dot\rho_t \rangle_{T_{\rho_t}}$ is given by a Riemannian metric in the tangent space (such as for W_2 or WFR_λ), then $\mathcal{K}^*[\rho_t, s_t, t] = \frac{1}{2}\langle s_t, s_t \rangle_{T^*_{\rho_t}}$ is the same metric written in the cotangent space (see App. B.1 for detailed derivations for all examples considered in this work).

Finally, under our assumptions, $\mathcal{L}[\rho_t, \dot\rho_t, t]$ can also be written using the Legendre transform, $\mathcal{L}[\rho_t, \dot\rho_t, t] = \sup_{s_t \in T^*_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t\; dx_t - \mathcal{H}[\rho_t, s_t, t]$. The following theorem forms the basis for our computational approach, and can be derived using the Legendre transform and integration by parts in time (see App. A for proof and Fig. 2 for visualization).

Theorem 1. *For a Lagrangian $\mathcal{L}[\rho_t, \dot\rho_t, t]$ which is lsc and strictly convex in $\dot\rho_t$, the optimization*

$$\mathcal{S} = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \mathcal{A}_{\mathcal{L}}[\rho_t] = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \int_0^1 \mathcal{L}[\rho_t, \dot\rho_t, t]dt$$

is equivalent to the following dual

$$\mathcal{S} = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1\mu_1\, dx_1 - \int s_0\mu_0\, dx_0 - \int_0^1 \Big(\int \frac{\partial s_t}{\partial t}\rho_t dx_t + \mathcal{H}[\rho_t, s_t, t]\Big) dt \tag{13}$$

where, for $s_t \in T^*_{\rho_t}\mathcal{P}$, the Hamiltonian $\mathcal{H}[\rho_t, s_t, t]$ is the Legendre transform of $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ (Eq 12). In particular, the action $\mathcal{A}_{\mathcal{L}}[\rho_t]$ of a given curve is the solution to the inner optimization,

$$\mathcal{A}_{\mathcal{L}}[\rho_t] = \sup_{s_t} \int s_1\mu_1 \, dx_1 - \int s_0\mu_0 \, dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt. \tag{14}$$

In line with our goal of defining Lagrangian actions *directly* on $\mathcal{P}(\mathcal{X})$ instead of via $\mathcal{X}$, Thm. 1 operates *only* in the abstract space of densities. See App. C for a detailed discussion.

Finally, the solution to Eq. (11) can also be expressed as a *Wasserstein Hamiltonian flow* [19], with the optimality conditions $\frac{\partial \rho_t}{\partial t} = \frac{\delta}{\delta s_t}\mathcal{H}[\rho_t, s_t, t]$ and $\frac{\partial s_t}{\partial t} = -\frac{\delta}{\delta \rho_t}\mathcal{H}[\rho_t, s_t, t]$ (see Sec. 6).

To further analyze Thm. 1 and set the stage for our computational approach in Sec. 3.2, we consider the two optimizations in Eq. (13) as (i) ***evaluating*** the action functional $\mathcal{A}_{\mathcal{L}}[\rho_t]$ for a given curve ρ_t, and (ii) ***optimizing*** the action over curves $\rho_t \in \Gamma(\{\mu_{t_i}\})$ satisfying the desired constraints.

3.1.1 Inner Optimization: Evaluating $\mathcal{A}_{\mathcal{L}}[\rho_t]$ using Action Matching

We immediately recognize the similarity of Eq. (14) to the AM objective in Eq. (8) for $\mathcal{H}[\rho_t, s_t, t] = \int \frac{1}{2}\|\nabla s_t\|^2 \rho_t dx_t$, which suggests a generalized notion of Action Matching as an inner loop to evaluate $\mathcal{A}_{\mathcal{L}}[\rho_t]$ for a given $\rho_t \in \Gamma(\{\mu_{t_i}\})$ in Thm. 1. For all t, the optimal cotangent vector $s_{\dot{\rho}_t}$ corresponds to the tangent vector $\dot{\rho}_t$ of the given curve via the Legendre transform or Eq. (14).

Neklyudov et al. [47] assume access to samples from a *continuous* curve of densities μ_t which, from our perspective, corresponds to the limit as the number of constraints $M \to \infty$. Since $\rho_t \in \Gamma(\{\mu_{t_i}\})$ has no remaining degrees of freedom in this case, the outer optimization over ρ_t can be ignored and expectations in Eq. (8) are written directly under μ_t. However, this assumption is often unreasonable in applications such as trajectory inference, where data is sampled discretely in time.

3.1.2 Outer Optimization over Constrained Distributional Paths

In our settings of interest, the outer optimization over curves $\mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}) = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \mathcal{A}_{\mathcal{L}}[\rho_t]$ is thus necessary to *interpolate* between M given marginals using the inductive bias encoded in the Lagrangian $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$. Crucially, our parameterization of ρ_t in Sec. 3.2.2 will enforce $\rho_t \in \Gamma(\{\mu_{t_i}\})$ by design, given access to samples from μ_{t_i}. Nevertheless, upon reaching an optimal ρ_t, our primary object of interest is the dynamics model corresponding to $\dot{\rho}_t$ and parameterized by the optimal $s_{\dot{\rho}_t}$ in Eq. (14), which may be used to transport particles or predict individual trajectories.

3.2 Computational Approach for Solving Wasserstein Lagrangian Flows

In this section, we describe our computational approach to solving for a class of Wasserstein Lagrangian Flows, which is summarized in Alg. 1.

3.2.1 Linearizable Kinetic and Potential Energies

Despite the generality of Thm. 1, we restrict attention to Lagrangians with the following property.

Definition 3.1 ((Dual) Linearizability)**.** *A Lagrangian $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ is* dual linearizable *if the corresponding Hamiltonian $\mathcal{H}[\rho_t, s_t, t]$ can be written as a linear functional of the density ρ_t. In other words, $\mathcal{H}[\rho_t, s_t, t]$ is* linearizable *if there exist functions $K^*(x_t, s_t, t)$, and $U(x_t, s_t, t)$ such that*

$$\mathcal{H}[\rho_t, s_t, t] = \int \Big(K^*(x_t, s_t, t) + U(x_t, s_t, t) \Big) \rho_t dx_t. \tag{15}$$

This property suggests that we only need to draw samples from ρ_t and need not evaluate its density, which allows us to derive an efficient parameterization of curves satisfying $\rho_t \in \Gamma(\{\mu_{t_i}\})$ below. [3]

As examples, note that the WFR_λ or W_2 metrics as the Lagrangian yield a linear Hamiltonian $\mathcal{H}[\rho_t, s_t, t] = \mathcal{K}^*[\rho_t, s_t, t] = \frac{1}{2}\langle s_t, s_t \rangle^{WFR_\lambda}_{T^*_{\rho_t}} = \int (\frac{1}{2}\|\nabla s_t\|^2 + \frac{\lambda}{2}s_t^2)\rho_t dx_t$, with $\lambda = 0$ for W_2. Potential energies $\mathcal{U}[\rho_t, t] = \int V_t(x_t)\rho_t dx_t$ which are linear in ρ_t (Ex. 4.3) clearly satisfy Def. 3.1. However, nonlinear potential energies as in Ex. 4.4 require reparameterization to be linearizable.

[3] In App. B.3, we highlight the Schrödinger Equation as a special case of our framework which does not appear to admit a linear dual problem. In this case, optimization of Eq. (13) may require explicit modeling of the density ρ_t corresponding to a given set of particles x_t (e.g. see Pfau et al. [51]).

Algorithm 1 Learning Wasserstein Lagrangian Flows

```
Require: samples from the marginals $\mu_0, \mu_1$, parametric model $s_t(x;\theta)$, generator from $\rho_t(x;\eta)$
  for learning iterations do
    sample from marginals $\{x_0^i\}_{i=1}^n \sim \mu_0,\ \{x_1^i\}_{i=1}^n \sim \mu_1$, sample time $\{t^i\}_{i=1}^n \sim \textsc{uniform}[0,1]$
    $x_t^i = (1-t^i)x_0^i + t^i x_1^i + t^i(1-t^i)\textsc{nnet}(t^i, x_0^i, x_1^i; \eta)$
    $-\textsc{grad}_\eta = \nabla_\eta \frac{1}{n}\sum_{i=1}^n \Big[\frac{\partial s_{t^i}}{\partial t}(x_t^i(\eta);\theta) + K^*\big(x_t^i(\eta), s_{t^i}(x_t^i(\eta);\theta), t^i\big) + U\big(x_t^i(\eta), s_{t^i}(x_t^i(\eta);\theta), t^i\big)\Big]$

    for Wasserstein gradient steps do
      $x_t^i \leftarrow x_t^i + \alpha \cdot t^i(1-t^i)\nabla_x\Big[\frac{\partial s_{t^i}}{\partial t}(x_t^i;\theta) + K^*\big(x_t^i, s_{t^i}(x_t^i;\theta), t^i\big) + U\big(x_t^i, s_{t^i}(x_t^i;\theta), t^i\big)\Big]$
    end for
    $\textsc{grad}_\theta = \nabla_\theta \frac{1}{n}\sum_{i=1}^n \Big[s_1(x_1^i;\theta) - s_0(x_0^i;\theta) - \frac{\partial s_{t^i}}{\partial t}(x_t^i;\theta) - K^*\big(x_t^i, s_{t^i}(x_t^i;\theta), t^i\big) - U\big(x_t^i, s_{t^i}(x_t^i;\theta), t^i\big)\Big]$

    update parameters using gradients $\textsc{grad}_\eta, \textsc{grad}_\theta$
  end for
  return cotangent vectors $s_t(x;\theta)$
```

3.2.2 Parameterization and Optimization

For any Lagrangian optimization with a linearizable dual objective as in Def. 3.1, we consider parameterizing the cotangent vectors s_t and the distributional path ρ_t. We parameterize s_t as a neural network $s_t(x;\theta)$ which takes t and x as inputs with parameters θ, and outputs a scalar. Inspired by the fact that we only need to draw samples from ρ_t for these problems, we parameterize the distribution path $\rho_t(x;\eta)$ as a generative model, where the samples are generated as follows

$$x_t = (1-t)x_0 + tx_1 + t(1-t)\textsc{nnet}(t, x_0, x_1; \eta), \quad x_0 \sim \mu_0, \quad x_1 \sim \mu_1. \tag{16}$$

Notably, this preserves the endpoint marginals μ_0, μ_1. For multiple constraints, we can modify our sampling procedure to interpolate between two intermediate dataset marginals, with neural network parameters η shared across timesteps

$$x_t = \frac{t_{i+1}-t}{t_{i+1}-t_i}x_{t_i} + \frac{t-t_i}{t_{i+1}-t_i}x_{t_{i+1}} + \left(1 - \left(\frac{t_{i+1}-t}{t_{i+1}-t_i}\right)^2 - \left(\frac{t-t_i}{t_{i+1}-t_i}\right)^2\right)\textsc{nnet}(t, x_{t_i}, x_{t_{i+1}}; \eta)\,.$$

For linearizable dual objectives as in Eq. (13) and Eq. (15), we optimize

$$\begin{aligned}\textsc{loss}(\theta,\eta) = \min_\eta \max_\theta \int s_1(x_1;\theta)\mu_1\, dx_1 - \int s_0(x_0;\theta)\mu_0\, dx_0 \qquad\qquad (17)\\ - \int_0^1 \int \left(\frac{\partial s_t}{\partial t}(x_t;\theta) + K^*\big(x_t, s_t(x_t;\theta), t\big) + U\big(x_t, s_t(x_t;\theta), t\big)\right)\rho_t(x_t;\eta)dx_t dt,\end{aligned}$$

where the optimization w.r.t. η is performed via the re-parameterization trick. An alternative to parametrizing the distributional path ρ_t is to perform minimization of Eq. (17) via the Wasserstein gradient flow, i.e. the samples x_t from the initial path ρ_t are updated as follows

$$x_t' = x_t + \alpha \cdot t(1-t)\nabla_x\left[\frac{\partial s_t}{\partial t}(x_t;\theta) + K^*\big(x_t, s_t(x_t;\theta), t\big) + U\big(x_t, s_t(x_t;\theta), t\big)\right], \tag{18}$$

where α is a hyperparameter regulating the step-size, and the coefficient $t(1-t)$ guarantees the preservation of the endpoints. In practice, we find that combining both the parametric and nonparametric approaches works best. The pseudo-code for the resulting algorithm is given in Alg. 1.

Discontinuous Interpolation The support of the optimal distribution path ρ_t might be disconnected, as in Fig. 1a. Thus, it may be impossible to interpolate continuously between independent samples from the marginals while staying in the support of the optimal path. To allow for a discontinuous interpolation, we pass a discontinuous indicator variable $\mathbb{1}[t < 0.5]$ to the model $\rho_t(x;\eta)$. This indicator is crucial to ensure our parameterization is expressive enough to approximate any suitable distributional path including, for example, the optimal OT path (see proof in App. D).

Proposition 2. *For any absolutely-continuous distributional path* $\rho_t : [0,1] \mapsto \mathcal{P}_2(\mathcal{X})$ *on the* W_2 *manifold, there exists a function* $\textsc{nnet}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta)$ *such that Eq.* (16) *samples from* ρ_t.

4 Examples of Wasserstein Lagrangian Flows

We now analyze the Lagrangians, dual objectives, and Hamiltonian optimality conditions corresponding to several important examples of Wasserstein Lagrangian flows. We present various kinetic and potential energy terms using their motivating examples and $M = 2$ endpoint constraints.

However, note that we may combine various energy terms to construct Lagrangians $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$, and optimize subject to multiple constraints, as we consider in our experiments in Sec. 5.

Example 4.1 (W_2 **Optimal Transport**)**.** The Benamou-Brenier formulation of W_2 optimal transport in Eq. (2) is the simplest example of our framework, with no potential energy and the kinetic energy defined by the Otto metric $\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \frac{1}{2}\langle \dot{\rho}_t, \dot{\rho}_t \rangle^{W_2}_{T_{\rho_t}} = \mathcal{H}[\rho_t, s_{\dot{\rho}_t}, t] = \frac{1}{2}\int \|\nabla s_{\dot{\rho}_t}\|^2 \rho_t dx_t$. In contrast to Eq. (2), note that our Lagrangian optimization in Eq. (11) is over ρ_t only, while solving the dual objective introduces the second optimization to identify $s_{\dot{\rho}_t}$ such that $\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_{\dot{\rho}_t})$. Our dual objective for solving the standard optimal transport problem with quadratic cost becomes

$$\mathcal{S}_{OT} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 d\mu_1 - \int s_0 d\mu_0 - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 \right) \rho_t dx_t dt, \tag{19}$$

where the Hamiltonian optimality conditions $\frac{\partial \rho_t}{\partial t} = \frac{\delta}{\delta s_t}\mathcal{H}[\rho_t, s_t, t]$, $\frac{\partial s_t}{\partial t} = -\frac{\delta}{\delta \rho_t}\mathcal{H}[\rho_t, s_t, t]$ [19] recover the characterization of W_2 geodesics via the continuity and Hamilton-Jacobi equations [7],

$$\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_t) \qquad \frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 = 0. \tag{20}$$

It is well known that optimal transport plans (or Wasserstein-2 geodesics) are 'straight-paths' in the Euclidean space [66]. For the flow induced by a vector field ∇s_t, we calculate the acceleration, or second derivative with respect to time, as

$$\ddot{X}_t = \nabla \left[\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 \right] = 0, \tag{21}$$

where zero acceleration is achieved if $\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 = c, \forall t$, as occurs at optimality in Eq. (20).

Example 4.2 (**Unbalanced Optimal Transport**)**.** The *unbalanced* OT problem arises from the WFR_λ geometry, and is useful for modeling mass teleportation and changes in total probability mass when cell birth and death occur as part of the underlying dynamics [58, 42]. Viewing the dynamical formulation of WFR in Eq. (5) as a Lagrangian optimization,

$$WFR_\lambda(\mu_0, \mu_1)^2 = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt = \int_0^1 \frac{1}{2}\langle \dot{\rho}_t, \dot{\rho}_t \rangle^{WFR_\lambda}_{T_{\rho_t}} dt \text{ s.t. } \rho_0 = \mu_0, \rho_1 = \mu_1.$$

Compared to Eq. (5), our Lagrangian formulation again optimizes over ρ_t only, and solving the dual requires finding $s_{\dot{\rho}_t}$ such that $\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_{\dot{\rho}_t}) + \lambda \rho_t s_{\dot{\rho}_t}$ as in Eq. (6). We optimize the objective

$$\mathcal{S}_{uOT} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 d\mu_1 - \int s_0 d\mu_0 - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 + \frac{\lambda}{2} s_t^2 \right) \rho_t dx_t dt,$$

where we recognize the WFR_λ cotangent metric from Eq. (7) in the final term, $\mathcal{H}[\rho_t, s_t, t] = \mathcal{K}^*[\rho_t, s_t, t] = \frac{1}{2}\langle s_t, s_t \rangle^{WFR_\lambda}_{T^*_{\rho_t}} = \frac{1}{2}\int \left(\|\nabla s_t\|^2 + \lambda\, s_t^2 \right) \rho_t \, dx_t$.

Example 4.3 (**Physically Constrained Optimal Transport**)**.** A popular technique for incorporating inductive bias from biological or geometric prior information into trajectory inference methods is to consider spatial potentials $\mathcal{U}[\rho_t, t] = \int V_t(x_t) \rho_t dx_t$ [61, 28, 53], which are already linear in the density. In this case, we may consider *any* linearizable kinetic energy (see App. B.1). For the W_2 transport case, our objective is

$$\mathcal{S}_{pOT} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 d\mu_1 - \int s_0 d\mu_0 - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 + V_t \right) \rho_t dx_t dt,$$

with the optimality conditions

$$\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_t), \qquad \frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 + V_t = 0, \qquad \ddot{X}_t = \nabla \left[\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 \right] = -\nabla V_t. \tag{22}$$

As in Eq. (21), the latter condition implies that the acceleration is given by the gradient of the spatial potential $V_t(x_t)$. We describe the potentials used in our experiments on scRNA datasets in Sec. 5.

Example 4.4 (**Schrödinger Bridge**)**.** For many problems of interest, such as scRNA sequencing [58], it may be useful to incorporate stochasticity into the dynamics as prior knowledge. For Brownian-motion diffusion processes with known coefficient σ, the dynamical Schrödinger Bridge (SB) problem [45, 31, 15] is given by

$$\mathcal{S}_{SB} = \inf_{\rho_t, v_t} \int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t dx_t dt \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) + \frac{\sigma^2}{2}\Delta \rho_t, \;\; \rho_0 = \mu_0, \;\; \rho_1 = \mu_1. \tag{23}$$

Table 1: Results for high-dim PCA representation of single-cell data for corresponding datasets. We report Wasserstein-1 distance averaged over left-out marginals. All results are averaged over 5 independent runs. Results with citations are taken from corresponding papers.

	dim=5	**dim=50**		**dim=100**	
Method	**EB**	**Cite**	**Multi**	**Cite**	**Multi**
exact OT	0.822	37.569	47.084	42.974	53.271
WLF-OT (ours)	0.814 ± 0.002	38.253 ± 0.071	47.736 ± 0.110	44.769 ± 0.054	55.313 ± 0.754
OT-CFM (more parameters)	$0.822 \pm$ 3.0e-4	37.821 ± 0.010	47.268 ± 0.017	44.013 ± 0.010	54.253 ± 0.012
OT-CFM [63]	$\mathbf{0.790 \pm 0.068}$	38.756 ± 0.398	47.576 ± 6.622	45.393 ± 0.416	54.814 ± 5.858
I-CFM [63]	0.872 ± 0.087	41.834 ± 3.284	49.779 ± 4.430	48.276 ± 3.281	57.262 ± 3.855
WLF-UOT ($\lambda = 1$, ours)	$\mathbf{0.800 \pm 0.002}$	$\mathbf{37.035 \pm 0.079}$	45.903 ± 0.161	$\mathbf{43.530 \pm 0.067}$	53.403 ± 0.168
WLF-SB (ours)	$0.816 \pm$ 7.7e-4	39.240 ± 0.068	47.788 ± 0.111	46.177 ± 0.083	55.716 ± 0.058
[SF]^2 M-Geo [62]	1.221 ± 0.38	38.524 ± 0.293	$\mathbf{44.795 \pm 1.911}$	44.498 ± 0.416	$\mathbf{52.203 \pm 1.957}$
[SF]^2 M-Exact [62]	$\mathbf{0.793 \pm 0.066}$	40.009 ± 0.783	45.337 ± 2.833	46.530 ± 0.426	52.888 ± 1.986
WLF-(OT + potential, ours)	0.651 ± 0.002	36.167 ± 0.031	38.743 ± 0.060	42.857 ± 0.045	47.365 ± 0.051
WLF-(UOT + potential, $\lambda = 1$, ours)	$\mathbf{0.634 \pm 0.001}$	$\mathbf{34.160 \pm 0.041}$	$\mathbf{36.131 \pm 0.023}$	$\mathbf{41.084 \pm 0.043}$	$\mathbf{45.231 \pm 0.010}$

To model the SB problem, we consider the following potential energy with the W_2 kinetic energy,

$$\mathcal{U}[\rho_t, t] = -\frac{\sigma^4}{8} \int \left\|\nabla \log \rho_t\right\|^2 \rho_t dx_t, \tag{24}$$

which arises from the entropy $\mathcal{F}[\rho_t] = -H[\rho_t] = \int (\log \rho_t - 1)\, \rho_t dx_t$ via $\nabla \frac{\delta}{\delta \rho_t} \mathcal{F}[\rho_t] = \nabla \log \rho_t$. We assume time-independent σ to simplify $\mathcal{U}[\rho_t, t]$, but consider time-varying σ_t in Ex. B.2.

To transform the potential energy term into a dual-linearizable form for the SB problem, we consider the reparameterization $\Phi_t = s_t + \frac{\sigma^2}{2} \log \rho_t$, which translates between the drift ∇s_t of the probability flow ODE and the drift $\nabla \Phi_t$ of the Fokker-Planck equation [60]. With detailed derivations in App. B.2, the dual objective becomes

$$\mathcal{S}_{SB} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} \int \Phi_1 d\mu_1 - \int \Phi_0 d\mu_0 - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \left\|\nabla \Phi_t\right\|^2 + \frac{\sigma^2}{2} \Delta \Phi_t \right) \rho_t dx_t dt. \tag{25}$$

5 Experiments

We apply our methods for trajectory inference of single-cell RNA sequencing data, including the Embryoid body (**EB**) dataset [46], CITE-seq (**Cite**) and Multiome (**Multi**) datasets [11], and melanoma treatment dataset of [9, 49].

Potential for Physically-Constrained OT For all tasks, we consider the simplest possible model of the physical potential accelerating the cells. For each marginal except the first and the last ones, we estimate the acceleration of its mean using finite differences. The potential for the corresponding time interval is then $V_t(x) = -\langle x, a_t \rangle$, where a_t is the estimated acceleration of the mean value. For leave-one-out tasks, we include the mean of the left out marginal since the considered data contains too few marginals (4 for Cite and Multi) for learning a meaningful model of the acceleration.

Leave-One-Out Marginal Task To test the ability of our approaches to approximate interpolating marginal distributions, we follow [61] and evaluate models using a leave-one-timepoint-out strategy. In particular, we train on all marginals except at time t_i, and evaluate by computing the Wasserstein-1 distance between the predicted marginal ρ_{t_i} and the left-out marginal μ_{t_i}. For preprocessing and baselines, we follow Tong et al. [62, 63] (see App. E.1 for details).

In Table 1, we report results on EB, Cite, and Multi datasets. First, we see that our proposed WLF-OT method achieves comparable results to related approaches: OT-CFM and I-CFM [63], which use minibatch OT couplings or independent samples of the marginals, respectively. For OT-CFM, we reproduce the results using a larger model to match the performance of the exact OT solver [24]. These models represent dynamics with minimal prior knowledge, and thus serve as a baseline when compared against dynamics incorporating additional priors.

Next, we consider Lagrangians encoding various prior information. WLF-SB (ours), [SF]^2 M-Exact [62], and SB-CFM [63] incorporate stochasticity into the dynamics by solving the SB problem; [SF]^2 M-Geo takes advantage of the data manifold geometry by learning from OT couplings generated with the approximate geodesic cost; our WLF-UOT incorporates probability mass teleportation using the WFR kinetic energy. In Table 1, we see that WLF-UOT yields consistent performance improvements across datasets. Finally, we observe that a good model of the potential function can drastically improve performance, using either W_2 or WFR kinetic energy.

Table 2: Results for train/test splits of 5-dim PCA on EB dataset, with the setting and baseline results taken from Koshizuka & Sato [28, Table 1]. We report W1 distance between test μ_{t_i} and ρ_{t_i} obtained by running dynamics from $\mu_{t_{i-1}}$.

Model	t_1	t_2	t_3	t_4	Mean
Neural SDE [32]	0.69	0.91	0.85	0.81	0.82
TrajectoryNet [61]	0.73	1.06	0.90	1.01	0.93
IPF (GP) [65]	0.70	1.04	0.94	0.98	0.92
IPF (NN) [21]	0.73	0.89	0.84	0.83	0.82
SB-FBSDE [12]	0.56	0.80	1.00	1.00	0.84
NLSB [28]	0.68	0.84	0.81	0.79	0.78
WLF-OT	0.65	0.78	0.76	0.75	0.74
WLF-SB	0.63	0.79	0.77	0.74	**0.73**
WLF-(OT + potential)	0.64	0.77	0.76	0.76	**0.73**
WLF-UOT ($\lambda = 0.1$)	0.64	0.84	0.80	0.81	0.77
WLF-(UOT + potential, $\lambda = 0.1$)	0.67	0.80	0.78	0.78	0.76

Table 3: Results in the setting of Pariset et al. [49, Table 1] (uDSB) for melanoma treatment data with 3 marginals and train/test splits. We report test MMD and W2 distance between μ_1 and ρ_1 obtained by running dynamics from μ_0.

Model	MMD	W_2
SB-FBSDE [12]	1.86e-2	6.23
uDSB (no growth) [49]	1.86e-2	6.27
uDSB (w/growth) [49]	1.75e-2	6.11
WLF-OT (no growth)	**5.04e-3**	5.20
WLF-UOT ($\lambda = 0.1$)	9.16e-3	**5.01**

Comparison with SB Baselines on EB Dataset To compare against a broader class of baselines for the SB problem, we consider the setting of Koshizuka & Sato [28, Table 1] on the EB dataset. Instead of leaving out one marginal, we divide the data using a train/test split and evaluate the W1 distance between the test μ_{t_i} and ρ_{t_i} obtained by running dynamics from the previous test $\mu_{t_{i-1}}$. In Table 2, we find that WLF-SB outperforms several SB baselines from recent literature (see Sec. 6).

Comparison with UOT Baseline on Melanoma Dataset To test the ability of our WLF-OT approach to account for cell birth and death, we consider the 50-dim. setting of Pariset et al. [49, Table 1] for melanoma cells undergoing treatment with a cancer drug. In Table 3, we show that WLF-OT and WLF-UOT can outperform the unbalanced baseline (uDSB) from Pariset et al. [49].

6 Related Work

Wasserstein Hamiltonian Flows Chow et al. [19] develop the notion of a Hamiltonian flow on the Wasserstein manifold and consider several of the same examples discussed here. While the Hamiltonian and Lagrangian formalisms describe the same integral flow through optimality conditions for $(\rho_t, \dot{\rho}_t)$ and (ρ_t, s_t), Chow et al. [19], Wu et al. [67] emphasize solving the Cauchy problem suggested by the Hamiltonian perspective. Our approach recovers the Hamiltonian flow (ρ_t, s_t) in the cotangent bundle at optimality, but does so by solving a variational problem.

Flow Matching and Diffusion Schrödinger Bridge Methods Flow Matching methods [38, 35, 1, 2, 63, 62] learn a marginal vector field corresponding to a mixture-of-bridges process parameterized by a coupling and interpolating bridge [59]. When samples from the endpoint marginals are coupled via an OT plan, Flow Matching solves a dynamical optimal transport problem [52]. Rectified Flow obtains couplings using ODE simulation with the goal of straight-path trajectories for generative modeling [38, 40], which is extended to SDEs in bridge matching methods [59, 50]. Diffusion Schrödinger Bridge (DSB) methods [21, 12] also update the couplings iteratively based on learned forward and backward SDEs, and have recently been adapted to solve the unbalanced OT problem in Pariset et al. [49]. Finally, Liu et al. [36, 37] consider extending DSB or bridge matching methods to solve physically-constrained SB problems. Unlike the above methods, our approach does not require optimal couplings to sample from the intermediate marginals, and thus avoids both simulating ODEs or SDEs and running minibatch (regularized) OT solvers.

Optimal Transport with Lagrangian Cost Input-convex neural networks [4] provide an efficient approach to static OT [44, 27, 9, 10] but are limited to Euclidean cost. Several works extend to other costs using static [22, 53, 64] or dynamical formulations [39, 28]. The most general way to define a transport cost is using a Lagrangian action in the state-space (Villani [66] Ch. 7). While we focus on lifted Lagrangians in density space, our framework also encompasses OT with state-space Lagrangian costs (App. B.1.2).

7 Conclusion

In this work, we demonstrated that many variations of optimal transport, such as Schrödinger bridge, unbalanced OT, or OT with physical constraints can be formulated as Lagrangian action minimization on the density manifold. We proposed a computational framework for this minimization by deriving a dual objective in terms of cotangent vectors, which correspond to a vector field on the state-space and can be parameterized via a neural network. We studied the problem of trajectory inference in biological systems, and showed that we can incorporate prior knowledge of the dynamics while respecting marginal constraints on observed data, resulting in significant improvement in several benchmarks. We expect our approach can extend to other natural science domains such as quantum mechanics and social sciences by incorporating new priors for learning the underlying dynamics.

References

[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *International Conference on Learning Representations*, 2022.

[2] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint 2303.08797*, 2023.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[4] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

[5] Vladimir Igorevich Arnol'd. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013.

[6] Martin Bauer, Martins Bruveris, and Peter W Michor. Uniqueness of the fisher–rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

[7] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[8] Jean-David Benamou, Thomas O Gallouët, and François-Xavier Vialard. Second-order models for optimal transport and cubic splines on the wasserstein space. *Foundations of Computational Mathematics*, 19:1113–1143, 2019.

[9] Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, pp. 2021–12, 2021.

[10] Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

[11] Daniel Burkhardt, Jonathan Bloom, Robrecht Cannoodt, Malte D Luecken, Smita Krishnaswamy, Christopher Lance, Angela O Pisco, and Fabian J Theis. Multimodal single-cell integration across time, individuals, and batches. In *NeurIPS Competitions*, 2022.

[12] Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward sdes theory. In *International Conference on Learning Representations*, 2021.

[13] Tianrong Chen, Guan-Horng Liu, Molei Tao, and Evangelos A Theodorou. Deep momentum multi-marginal Schrödinger bridge. *arXiv preprint arXiv:2303.01751*, 2023.

[14] Yongxin Chen, Giovanni Conforti, and Tryphon T Georgiou. Measure-valued spline curves: An optimal transport viewpoint. *SIAM Journal on Mathematical Analysis*, 50(6):5947–5968, 2018.

[15] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge. *Siam Review*, 63(2):249–313, 2021.

[16] Sinho Chewi, Julien Clancy, Thibaut Le Gouic, Philippe Rigollet, George Stepaniants, and Austin Stromme. Fast and smooth interpolation on wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pp. 3061–3069. PMLR, 2021.

[17] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[18] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[19] Shui-Nee Chow, Wuchen Li, and Haomin Zhou. Wasserstein hamiltonian flows. *Journal of Differential Equations*, 268(3):1205–1219, 2020.

[20] Julien Clancy and Felipe Suarez. Wasserstein-fisher-rao splines. *arXiv preprint arXiv:2203.15728*, 2022.

[21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[22] Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable computation of monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[23] Alessio Figalli and Federico Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. 2021.

[24] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

[25] Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pp. 2417–2426. PMLR, 2016.

[26] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite radon measures. *Advances in Differential Equations*, 21(11/12): 1117–1164, 2016.

[27] Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021.

[28] Takeshi Koshizuka and Issei Sato. Neural lagrangian Schrödinger bridge: Diffusion modeling for population dynamics. In *The Eleventh International Conference on Learning Representations*, 2022.

[29] Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

[30] Flavien Léger and Wuchen Li. Hopf–cole transformation via generalized schrödinger bridge problem. *Journal of Differential Equations*, 274:788–827, 2021.

[31] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[32] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

[33] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: The Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

[34] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations*, 2022.

[36] Guan-Horng Liu, Tianrong Chen, Oswin So, and Evangelos Theodorou. Deep generalized schrödinger bridge. *Advances in Neural Information Processing Systems*, 35:9374–9388, 2022.

[37] Guanhorng Liu, Yaron Lipman, Maximilian Nickel, Brian Karrer, Evangelos A. Theodorou, and Ricky T.Q. Chen. Generalized schrödinger bridge matching, 9 2023.

[38] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[39] Shu Liu, Shaojun Ma, Yongxin Chen, Hongyuan Zha, and Haomin Zhou. Learning high dimensional wasserstein geodesics. *arXiv preprint arXiv:2102.02992*, 2021.

[40] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *International Conference on Learning Representations*, 2022.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[42] Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

[43] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161 (5):1202–1214, 2015.

[44] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

[45] Toshio Mikami. Optimal transportation problem as stochastic mechanics. *Selected Papers on Probability and Statistics, Amer. Math. Soc. Transl. Ser*, 2(227):75–94, 2008.

[46] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37 (12):1482–1492, 2019.

[47] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, 2023.

[48] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26:101–174, 2001.

[49] Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced diffusion Schrödinger bridge. *arXiv preprint arXiv:2306.09099*, 2023.

[50] Stefano Peluchetti. Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling. *arXiv preprint arXiv:2304.00917*, 2023.

[51] David Pfau, James S Spencer, Alexander GDG Matthews, and W Matthew C Foulkes. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Physical Review Research*, 2(3):033429, 2020.

[52] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky Chen. Multisample flow matching: Straightening flows with minibatch couplings. *International Conference on Machine Learning*, 2023.

[53] Aram-Alexandre Pooladian, Carles Domingo-Enrich, Ricky TQ Chen, and Brandon Amos. Neural optimal transport with lagrangian costs. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

[55] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[56] Benjamin Schachter. *An Eulerian Approach to Optimal Transport with Applications to the Otto Calculus*. University of Toronto (Canada), 2017.

[57] Geoffrey Schiebinger. Reconstructing developmental landscapes and trajectories from single-cell data. *Current Opinion in Systems Biology*, 27:100351, 2021.

[58] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[59] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023.

[60] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[61] Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

[62] Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023.

[63] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[64] Théo Uscidda and Marco Cuturi. The monge gap: A regularizer to learn all transport maps. *arXiv preprint arXiv:2302.04953*, 2023.

[65] Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

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

[67] Hao Wu, Shu Liu, Xiaojing Ye, and Haomin Zhou. Parameterized wasserstein hamiltonian flow. *arXiv preprint arXiv:2306.00191*, 2023.

A General Dual Objectives for Wasserstein Lagrangian Flows

In this section, we derive the general forms for the Hamiltonian dual objectives arising from Wasserstein Lagrangian Flows. We prove Thm. 1 and derive the general dual objective in Eq. (13) of the main text, before considering the effect of multiple marginal constraints in App. A.1. We defer explicit calculation of Hamiltonians for important special cases to App. B.

Theorem 1. *For a Lagrangian $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ which is lsc and strictly convex in $\dot{\rho}_t$, the optimization*

$$\mathcal{S} = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \mathcal{A}_{\mathcal{L}}[\rho_t] = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt$$

is equivalent to the following dual

$$\mathcal{S} = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1 \mu_1 \, dx_1 - \int s_0 \mu_0 \, dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt \tag{13}$$

where, for $s_t \in T^*_{\rho_t}\mathcal{P}$, the Hamiltonian $\mathcal{H}[\rho_t, s_t, t]$ is the Legendre transform of $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ (Eq 12). In particular, the action $\mathcal{A}_{\mathcal{L}}[\rho_t]$ of a given curve is the solution to the inner optimization,

$$\mathcal{A}_{\mathcal{L}}[\rho_t] = \sup_{s_t} \int s_1 \mu_1 \, dx_1 - \int s_0 \mu_0 \, dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt. \tag{14}$$

Recall the definition of the Legendre transform for $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ strictly convex in $\dot{\rho}_t$,

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{\dot{\rho}_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, dx_t - \mathcal{L}[\rho_t, \dot{\rho}_t, t] \tag{26}$$

$$\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \sup_{s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, dx_t - \mathcal{H}[\rho_t, s_t, t] \tag{27}$$

Proof. We prove the case of $M = 2$ here and the case of $M > 2$ below in App. A.1.

Denote the set of curves of marginal densities ρ_t with the prescribed endpoint marginals as $\Gamma(\mu_0, \mu_1) = \{\rho_t | \ \rho_t \in \mathcal{P}(\mathcal{X}) \ \forall t, \rho_0 = \mu_0, \rho_1 = \mu_1\}$. The result follows directly from the definition of the Legendre transform in Eq. (26) and integration by parts in time in step (i),

$$\begin{aligned}
\mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) &= \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt \quad \text{s.t.} \quad \rho_0 = \mu_0, \qquad \rho_1 = \mu_1 \\
&= \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt \\
&= \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}} \int_0^1 \left(\int s_t \dot{\rho}_t \, dx_t - \mathcal{H}[\rho_t, s_t, t] \right) dt \\
&\overset{(i)}{=} \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 \rho_1 dx_1 - \int s_0 \rho_0 dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt \\
&\overset{(ii)}{=} \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 \mu_1 dx_1 - \int s_0 \mu_0 dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt
\end{aligned} \tag{28}$$

which is the desired result. In (ii), we use the fact that $\rho_0 = \mu_0$, $\rho_1 = \mu_1$ for $\rho_t \in \Gamma(\mu_0, \mu_1)$. Finally, note that $s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}$ simply identifies s_t as a cotangent vector and does not impose meaningful constraints on the form of $s_t \in \mathcal{C}^\infty(\mathcal{X})$, so we drop this from the optimization in step (i). □

A.1 Multiple Marginal Constraints

Consider multiple marginal constraints in the Lagrangian action minimization problem for $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ strictly convex in $\dot{\rho}_t$,

$$\begin{aligned}
\mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) &= \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt \quad \text{s.t.} \quad \rho_{t_i} = \mu_{t_i} \ (\forall \ 0 \le i \le M-1) \\
&= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt
\end{aligned} \tag{29}$$

As in the proof of Thm. 1, the dual becomes

$$
\begin{aligned}
\mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) &= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}} \int_0^1 \left(\int s_t \dot{\rho}_t \, dx_t - \mathcal{H}[\rho_t, s_t, t] \right) dt \\
&= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1 \rho_1 dx_1 - \int s_0 \rho_0 dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt \\
&= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1 \mu_1 dx_1 - \int s_0 \mu_0 dx_0 - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt
\end{aligned}
$$

where the intermediate marginal constraints do not affect the result. Crucially, as discussed in Sec. 3.2.2, our sampling approach satisfies the marginal constraints $\rho_{t_i}(x_{t_i}) = \mu_{t_i}(x_{t_i})$ by design.

Piecewise Lagrangian Optimization Note that the concatenation of dual objectives for $M = 3$, or action-minimization problems between $\{\mu_{0,t_1}\}$ and $\{\mu_{t_1,1}\}$ yields the following dual objective

$$
\begin{aligned}
&\mathcal{S}_{\mathcal{L}}(\{\mu_{0,t_1}\}) + \mathcal{S}_{\mathcal{L}}(\{\mu_{t_1,1}\}) \qquad (30) \\
&\quad = \inf_{\rho_t \in \Gamma(\{\mu_0, \mu_{t_1}\})} \sup_{s_t} \int s_{t_1} \mu_{t_1} dx_{t_1} - \int s_0 \mu_0 dx_0 + \int_0^{t_1} \left(\int \frac{\partial s_t}{\partial t} \rho_t \, dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt \\
&\quad + \inf_{\rho_t \in \Gamma(\{\mu_{t_1}, \mu_1\})} \sup_{s_t} \int s_1 \mu_1 dx_1 - \int s_{t_1} \mu_{t_1} dx_{t_1} + \int_{t_1}^{1} \left(\int \frac{\partial s_t}{\partial t} \rho_t \, dx_t + \mathcal{H}[\rho_t, s_t, t] \right) dt
\end{aligned}
$$

After telescoping cancellation and taking the union of the constraints, we see that our computational approach yields a piece-wise solution to the multi-marginal problem, with $\mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) = \sum_{i=0}^{M-2} \mathcal{S}_{\mathcal{L}}(\{\mu_{t_i, t_{i+1}}\})$.

B Tractable Objectives for Special Cases

In this section, we calculate Hamiltonians and explicit dual objectives for important special cases of Wasserstein Lagrangian Flows, including those in Sec. 4.

We consider several important kinetic energies in App. B.1, including the W_2 and WFR_λ metrics (App. B.1.1) and the case of OT costs defined by general ground-space Lagrangians (App. B.1.2). In App. B.2, we provide further derivations to obtain a linear dual objective for the Schrödinger Bridge problem. Finally, we highlight the lack of dual linearizability for the case of the Schrödinger Equation App. B.3 Ex. B.3.

B.1 Dual Kinetic Energy from W_2, WFR, or Ground-Space Lagrangian Costs

Thm. 1 makes progress toward a dual objective *without* considering the continuity equation or dynamics in the ground space, by instead invoking the Legendre transform $\mathcal{H}[\rho_t, s_t, t]$ of a given Lagrangian $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ which is strictly convex in $\dot{\rho}_t$. However, to derive $\mathcal{H}[\rho_t, s_t, t]$ and optimize objectives of the form Eq. (13), we will need to represent the tangent vector on the space of densities $\dot{\rho}_t$, for example using a vector field v_t and growth term g_t as in Eq. (5).

Given a Lagrangian $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$, we seek to solve the optimization

$$
\mathcal{H}[\rho_t, s_t, t] = \sup_{\dot{\rho}_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, dx_t - \mathcal{L}[\rho_t, \dot{\rho}_t, t] = \sup_{\dot{\rho}_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, dx_t - \mathcal{K}[\rho_t, \dot{\rho}_t, t] + \mathcal{U}[\rho_t, t] \tag{31}
$$

Since the potential energy does not depend on $\dot{\rho}_t$, we focus on *kinetic energies* $\mathcal{K}[\rho_t, \dot{\rho}_t, t]$ which are linear in the density (see Def. 3.1). We consider two primary examples, the WFR_λ metric $\mathcal{K}[\rho_t, \dot{\rho}_t, t]$ using the continuity equation with growth term dynamics, and kinetic energies defined by expectations of ground-space Lagrangian costs under ρ_t (see App. B.1.2, Villani [66] Ch. 7, Ex. B.1 below),

$$
WFR_\lambda : \quad \mathcal{K}[\rho_t, \dot{\rho}_t, t] = \int \left(\frac{1}{2} \|v_t\|^2 + \frac{\lambda}{2} g_t^2 \right) \rho_t dx_t, \qquad \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t \tag{32}
$$

$$
L(\gamma_t, \dot{\gamma}_t, t) : \quad \mathcal{K}[\rho_t, \dot{\rho}_t, t] = \int L(x_t, v_t, t) \rho_t dx_t, \qquad \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) \tag{33}
$$

where $(x_t, v_t) = (\gamma_t, \dot{\gamma}_t)$ and we recover the W_2 kinetic energy for $L[x_t, v_t, t] = \frac{1}{2}\|v_t\|^2$ or $\lambda = 0$.

We proceed with common derivations, writing $\mathcal{K}[\rho_t, \dot{\rho}_t, t] = \int K(x_t, v_t, g_t, t)\rho_t dx_t$ and simplifying Eq. (31) using the more general dynamics in Eq. (32)

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{\dot{\rho}_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, dx_t - \mathcal{K}[\rho_t, \dot{\rho}_t, t] + \mathcal{U}[\rho_t, t] \tag{34}$$

$$= \sup_{(v_t, g_t)} \int s_t\big(-\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t\big) \, dx_t - \mathcal{K}[\rho_t, \dot{\rho}_t, t] + \mathcal{U}[\rho_t, t] \tag{35}$$

Integrating by parts, we have

$$= \sup_{(v_t, g_t)} \int \big(\langle \nabla s_t, v_t\rangle \rho_t + \lambda \rho_t s_t g_t\big) \, dx_t - \mathcal{K}[\rho_t, \dot{\rho}_t, t] + \mathcal{U}[\rho_t, t]. \tag{36}$$

We now focus on the special cases in Eq. (32) and Eq. (33).

B.1.1 Wasserstein Fisher-Rao and W_2

For $\mathcal{K}[\rho_t, \dot{\rho}_t, t] = \int \big(\frac{1}{2}\|v_t\|^2 + \frac{\lambda}{2} g_t^2\big)\rho_t dx_t$, we proceed from Eq. (36),

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{(v_t, g_t)} \int \Big(\langle \nabla s_t, v_t\rangle \rho_t + \lambda \rho_t s_t g_t\Big) \, dx_t - \int \Big(\frac{1}{2}\|v_t\|^2 + \frac{\lambda}{2} g_t^2\Big)\rho_t dx_t + \mathcal{U}[\rho_t, t] \tag{37}$$

Eliminating v_t and g_t implies

$$v_t = \nabla s_t \qquad\qquad g_t = s_t \tag{38}$$

where $v_t = \nabla s_t$ also holds for the W_2 case with $\lambda = 0$. Substituting into Eq. (37), we obtain a Hamiltonian with a dual kinetic energy $\mathcal{K}^*[\rho_t, \dot{\rho}_t, t]$ below that is linear in ρ_t and matches the metric expressed in the cotangent space $\frac{1}{2}\langle s_t, s_t\rangle_{T_{\rho_t}}^{WFR_\lambda}$,

$$\mathcal{H}[\rho_t, s_t, t] = \int \Big(\frac{1}{2}\|\nabla s_t\|^2 + \frac{\lambda}{2} s_t^2\Big)\rho_t \, dx_t + \mathcal{U}[\rho_t, t] = \frac{1}{2}\langle s_t, s_t\rangle_{T_{\rho_t}}^{WFR_\lambda} + \mathcal{U}[\rho_t, t]. \tag{39}$$

We make a similar conclusion for the W_2 metric with $\lambda = 0$, where the dual kinetic energy is $\mathcal{K}^*[\rho_t, \dot{\rho}_t, t] = \frac{1}{2}\langle s_t, s_t\rangle_{T_{\rho_t}}^{W_2} = \frac{1}{2}\int \|\nabla s_t\|^2 \rho_t \, dx_t$.

B.1.2 Lifting Ground-Space Lagrangian Costs to Kinetic Energies

We first consider using Lagrangians in the ground space to define costs associated with action-minimizing curves $\gamma^*(x_0, x_1)$. As in Villani [66] Thm. 7.21, we can consider using this cost to define an optimal transport costs between densities. We show that this corresponds to a special case of our Wasserstein Lagrangian Flows framework with kinetic energy $\mathcal{K}[\rho_t, \dot{\rho}_t, t] = \int L(x_t, v_t, t)\rho_t dx_t$ as in Eq. (33). However, as discussed in Sec. 3, defining our Lagrangians $\mathcal{L}[\rho_t, \dot{\rho}_t, t]$ *directly* on the space of densities allows for more generality using kinetic energies which include growth terms or potential energies which depend on the density.

Lagrangian and Hamiltonian Mechanics in the Ground-Space We begin by reviewing action-minimizing curves in the ground space, which forms the basis the Lagrangian formulation of classical mechanics [5]. For curves $\gamma(t) : [0, 1] \to \mathcal{X}$ with velocity $\dot{\gamma}_t \in \mathcal{T}_{\gamma(t)}\mathcal{X}$, we consider evaluating a Lagrangian function $L(\gamma_t, \dot{\gamma}_t, t)$ along the curve to define the *action* as the time integral $\mathcal{A}(\gamma) = \int_0^1 L(\gamma_t, \dot{\gamma}_t, t)dt$. Given two endpoints $x_0, x_1 \in \mathcal{X}$, we consider minimizing the action along all curves with the appropriate endpoints $\gamma \in \Pi(x_0, x_1)$,

$$c(x_0, x_1) = \inf_{\gamma \in \Pi(x_0, x_1)} \mathcal{A}(\gamma) = \inf_{\gamma_t} \int_0^1 L(\gamma_t, \dot{\gamma}_t, t)dt \quad \text{s.t. } \gamma_0 = x_0, \ \gamma_1 = x_1 \tag{40}$$

We refer to the optimizing curves $\gamma^*(x_0, x_1)$ as *Lagrangian flows* in the ground-space, which satisfy the Euler-Lagrange equation $\frac{d}{dt}\frac{\partial}{\partial \dot{\gamma}_t} L(\gamma_t, \dot{\gamma}_t, t) = \frac{d}{d\gamma_t} L(\gamma_t, \dot{\gamma}_t, t)$ as a stationarity condition.

We will assume that $L(\gamma_t, \dot{\gamma}_t, t)$ is strictly convex in the velocity $\dot{\gamma}_t$, in which case we can obtain an equivalent, *Hamiltonian* perspective via convex duality. Considering momentum variables p_t, we define the Hamiltonian $H(\gamma_t, p_t, t)$ as the Legendre transform of L with respect to $\dot{\gamma}_t$,

$$H(\gamma_t, p_t, t) = \sup_{\dot{\gamma}_t} \; \langle \dot{\gamma}_t, p_t \rangle - L(\gamma_t, \dot{\gamma}_t, t) \tag{41}$$

The Euler-Lagrange equations can be written as Hamilton's equations in the phase space

$$\dot{\gamma}_t = \frac{\partial}{\partial p_t} H(\gamma_t, p_t, t) \qquad\qquad \dot{p}_t = -\frac{\partial}{\partial \gamma_t} H(\gamma_t, p_t, t). \tag{42}$$

We proceed to consider Lagrangian actions in the ground-space as a way to construct optimal transport costs over distributions.

Example B.1 (Ground-Space Lagrangians as OT Costs). The cost function $c(x_0, x_1)$ is a degree of freedom in specifying an optimal transport distance between probability densities $\mu_0, \mu_1 \in \mathcal{P}(\mathcal{X})$ in Eq. (1). Beyond $c(x_0, x_1) = \|x_0 - x_1\|^2$, one might consider defining the OT problem using a cost $c(x_0, x_1)$ induced by a Lagrangian $L(\gamma_t, \dot{\gamma}_t, t)$ in the ground space $\gamma_t \in \mathcal{X}$, as in Eq. (40) (Villani [66] Ch. 7). In particular, a coupling $\pi(x_0, x_1)$ should assign mass to endpoints (x_0, x_1) based on the Lagrangian cost of their action-minimizing curves $\gamma^*(x_0, x_1)$ Translating to a dynamical formulation (Villani [66] Thm. 7.21) and using notation $(\gamma_t, \dot{\gamma}_t) = (x_t, v_t)$, the OT problem is

$$W_L(\mu_0, \mu_1) = \inf_{(x_t, v_t)} \int_0^1 \int L(x_t, v_t, t) \rho_t dx_t dt \quad \text{s.t. } \operatorname{law}(x_t) = \rho_t, \; \operatorname{law}(x_0) = \mu_0, \; \operatorname{law}(x_1) = \mu_1. \tag{43}$$

which we may also view as an optimization over the distribution of marginals ρ_t under which x_t is evaluated (see, e.g. Schachter [56] Def. 3.4.1)

$$W_L(\mu_0, \mu_1) = \inf_{\rho_t} \inf_{v_t} \int_0^1 \int L(x_t, v_t, t) \rho_t dx_t dt \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t), \; \rho_0 = \mu_0, \; \rho_1 = \mu_1. \tag{44}$$

We can thus view the OT problem as 'lifting' the Lagrangian cost on the ground space $\mathcal{X}$ to a distance in the space of probability densities $\mathcal{P}_2(\mathcal{X})$ via the kinetic energy $\mathcal{K}[\rho_t, \dot{\rho}_t, t] = \int L(x_t, v_t, t) \rho_t dx_t$ (see below). Of course, the Benamou-Brenier dynamical formulation of W_2-OT in Eq. (2) may be viewed as a special case with $L(\gamma_t, \dot{\gamma}_t, t) = L(x_t, v_t, t) = \frac{1}{2}\|v_t\|^2$.

Wasserstein Lagrangian and Hamiltonian Perspective Recognizing the similarity with the Benamou-Brenier formulation in Ex. 4.1, we consider the Wasserstein Lagrangian optimization with two endpoint marginal constraints,

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) &= \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \int_0^1 \mathcal{K}[\rho_t, \dot{\rho}_t, t] - \mathcal{U}[\rho_t, t] dt \qquad (45) \\ &= \inf_{\rho_t} \int_0^1 \left(\int L(x_t, v_t, t) \rho_t dx_t - \mathcal{U}[\rho_t, t] \right) dt \quad \text{s.t.} \quad \rho_0 = \mu_0, \qquad \rho_1 = \mu_1 \end{aligned}$$

Parameterizing the tangent space using the continuity equation as in Eq. (33) or Eq. (44), we can derive the Wasserstein Hamiltonian from Eq. (36) with $\lambda = 0$ (no growth dynamics). Including a potential energy $\mathcal{U}[\rho_t, t]$, we have

$$\begin{aligned} \mathcal{H}[\rho_t, s_t, t] &= \sup_{v_t} \int \langle \nabla s_t, v_t \rangle \rho_t \, dx_t - \mathcal{K}[\rho_t, \dot{\rho}_t, t] + \mathcal{U}[\rho_t, t]. \qquad (46) \\ &= \sup_{v_t} \int \langle \nabla s_t, v_t \rangle \rho_t \, dx_t - \int L(x_t, v_t, t) \rho_t dx_t + \mathcal{U}[\rho_t, t] \qquad (47) \\ &= \int \left(\sup_{v_t} \; \langle \nabla s_t, v_t \rangle - L(x_t, v_t, t) \right) \rho_t \, dx_t + \mathcal{U}[\rho_t, t] \qquad (48) \end{aligned}$$

which is simply a Legendre transform between velocity and momentum variables in the ground space (Eq. (41)). We can finally write,

$$\mathcal{H}[\rho_t, s_t, t] = \int H(x_t, \nabla s_t, t) \rho_t dx_t + \mathcal{U}[\rho_t, t] \tag{49}$$

which implies the dual kinetic energy is simply the expectation of the Hamiltonian $\mathcal{K}^*[\rho_t, s_t, t] = \int H(x_t, \nabla s_t, t) \rho_t dx_t$ and is clearly linear in the density ρ_t.

We leave empirical exploration of various Lagrangian costs for future work, but note that $H(x_t, \nabla s_t, t)$ in Eq. (49) must be known or optimized using Eq. (48) to obtain a tractable objective.

B.2 Schrödinger Bridge

In this section, we derive potential energies and tractable objectives corresponding to the Schrödinger Bridge problem

$$S_{SB} = \inf_{\rho_t, v_t} \int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t dx_t dt \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) - \frac{\sigma^2}{2}\Delta\rho_t \quad \rho_0 = \mu_0, \;\; \rho_1 = \mu_1. \tag{50}$$

which we will solve using the following (linear in ρ_t) dual objective from Eq. (25)

$$\mathcal{S}_{SB} = \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\|\nabla\Phi_t\|^2 + \frac{\sigma^2}{2}\Delta\Phi_t \right) \rho_t dx_t dt.$$

Lagrangian and Hamiltonian for SB We consider a potential energy of the form,

$$\mathcal{U}[\rho_t, t] = -\frac{\sigma^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t dx_t \tag{51}$$

which, alongside the W_2 kinetic energy, yields the full Lagrangian

$$\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \frac{1}{2}\langle \dot{\rho}_t, \dot{\rho}_t \rangle^{W_2}_{T_{\rho_t}} + \frac{\sigma^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t dx_t. \tag{52}$$

As in Eq. (34)-(37), we parameterize the tangent space using the continuity equation $\dot{\rho}_t = -\nabla \cdot (\rho_t v_t)$ and vector field v_t in solving for the Hamiltonian,

$$\begin{aligned} \mathcal{H}[\rho_t, s_t, t] &= \sup_{\dot{\rho}_t} \int s_t \dot{\rho}_t dx_t - \mathcal{L}[\rho_t, \dot{\rho}_t, t] \qquad (53) \\ &= \sup_{v_t} \int \langle \nabla s_t, v_t \rangle \rho_t dx_t - \frac{1}{2}\int \|v_t\|^2 \rho_t dx_t - \frac{\sigma_t^4}{8}\int \|\nabla \log \rho_t\|^2 \rho_t dx_t + \int \left(\frac{\partial}{\partial t} \frac{\sigma_t^2}{2} \right) \log \rho_t \; \rho_t dx_t \end{aligned}$$

which implies $v_t = \nabla s_t$ as before. Substituting into the above, the Hamiltonian becomes

$$\mathcal{H}[\rho_t, s_t, t] = \frac{1}{2}\int \|\nabla s_t\|^2 \rho_t dx_t - \frac{\sigma^4}{8}\int \|\nabla \log \rho_t\|^2 \rho_t dx_t. \tag{54}$$

which is of the form $\mathcal{H}[\rho_t, s_t, t] = \mathcal{K}^*[\rho_t, s_t, t] + \mathcal{U}[\rho_t, t]$ and matches Léger & Li [30] Eq. 8. As in Thm. 1, the dual for the Wasserstein Lagrangian Flow with the Lagrangian in Eq. (52) involves the Hamiltonian in Eq. (54),

$$\mathcal{S}_{\mathcal{L}} = \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \sup_{s_t} \int s_1 \mu_1 dx_1 - \int s_0 \mu_0 dx_0 - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 - \frac{\sigma^4}{8}\int \|\nabla \log \rho_t\|^2 \right) \rho_t \; dx_t \tag{55}$$

However, this objective is nonlinear in ρ_t and requires access to $\nabla \log \rho_t$. To linearize the dual objective, we proceed using a reparameterization in terms of the Fokker-Planck equation, or using the Hopf-Cole transform, in the following proposition.

Proposition 3. *The solution to the Wasserstein Lagrangian flow*

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) &= \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] dt \qquad \textit{s.t.} \quad \rho_0 = \mu_0, \qquad \rho_1 = \mu_1 \qquad (56) \\ \textit{where } \; \mathcal{K}[\rho_t, \dot{\rho}_t, t] &= \frac{1}{2}\langle \dot{\rho}_t, \dot{\rho}_t \rangle^{W_2}_{T_{\rho_t}}, \qquad \mathcal{U}[\rho_t, t] = -\frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2_{T^{W_2}_{\rho_t}} \end{aligned}$$

matches the solution to the SB *problem in Eq. (50),* $\mathcal{S} = \mathcal{S}_{SB}(\{\mu_{0,1}\}) = \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\})$ *up to a constant* $c(\{\mu_{0,1}\})$ *wrt* ρ_t.

Further, $\mathcal{S}$ *is the solution to the (dual) optimization*

$$\mathcal{S} = \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\|\nabla\Phi_t\|^2 + \frac{\sigma^2}{2}\Delta\Phi_t \right) \rho_t dx_t dt. \tag{57}$$

Thus, we obtain a dual objective for the SB problem, or WLF in Eq. (56), which is linear in ρ_t.

Proof. We consider the following reparameterization [30]

$$s_t = \Phi_t - \frac{\sigma^2}{2} \log \rho_t, \qquad \nabla s_t = \nabla \Phi_t - \frac{\sigma^2}{2} \nabla \log \rho_t. \tag{58}$$

Note that s_t is the drift for the continuity equation in Eq. (53), $\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_t)$. Via the above reparameterization, we see that $\nabla \Phi_t$ corresponds to the drift in the Fokker-Planck dynamics $\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla \Phi_t) + \frac{\sigma^2}{2} \nabla \cdot (\rho_t \nabla \log \rho_t) = -\nabla \cdot (\rho_t \nabla \Phi_t) + \frac{\sigma^2}{2} \Delta \rho_t$.

Wasserstein Lagrangian Dual Objective after Reparameterization: Starting from the dual objective in Eq. (55), we perform the reparameterization in Eq. (58), $s_t = \Phi_t - \frac{\sigma^2}{2} \log \rho_t$,

$$\begin{aligned} \mathcal{S}_{\mathcal{L}} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} & \int \Phi_1 \mu_1 dx_1 - \frac{\sigma^2}{2} \int \log \rho_1 \, \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 + \frac{\sigma^2}{2} \int \log \rho_0 \, \mu_0 dx_0 \\ & - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{\partial}{\partial t} \left(\frac{\sigma^2}{2} \log \rho_t \right) + \frac{1}{2} \left\langle \nabla \Phi_t - \frac{\sigma^2}{2} \nabla \log \rho_t, \nabla \Phi_t - \frac{\sigma^2}{2} \nabla \log \rho_t \right\rangle - \frac{\sigma^4}{8} \| \nabla \log \rho_t \|^2 \right) \rho_t \, dx_t dt \end{aligned} \tag{59}$$

Noting that the $\int \frac{\sigma^2}{2} (\frac{\partial}{\partial t} \log \rho_t) \, \rho_t dx_t$ cancels since $\frac{\partial}{\partial t} \int \rho_t dx_t = 0$, we simplify to obtain

$$\begin{aligned} \mathcal{S}_{\mathcal{L}} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} & \int \Phi_1 \mu_1 dx_1 - \frac{\sigma^2}{2} \int \log \rho_1 \, \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 + \frac{\sigma^2}{2} \int \log \rho_0 \mu_0 dx_0 \\ & - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \big\| \nabla \Phi_t \big\|^2 - \frac{\sigma^2}{2} \big\langle \nabla \Phi_t, \nabla \log \rho_t \big\rangle \right) \rho_t dx_t dt \end{aligned}$$

where the Hamiltonian now matches Eq. 7 in Léger & Li [30]. Taking $\nabla \log \rho_t = \frac{1}{\rho_t} \nabla \rho_t$ and integrating by parts, the final term becomes

$$\begin{aligned} \mathcal{S}_{\mathcal{L}} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} & \int \Phi_1 \mu_1 dx_1 - \frac{\sigma^2}{2} \int \log \rho_1 \, \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 + \frac{\sigma^2}{2} \int \log \rho_0 \mu_0 dx_0 \\ & - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \big\| \nabla \Phi_t \big\|^2 + \frac{\sigma^2}{2} \Delta \Phi_t \right) \rho_t dx_t dt \end{aligned}$$

Finally, we consider adding terms $c(\{\mu_{0,1}\}) = \frac{\sigma^2}{2} \int \log \mu_1 \, \mu_1 dx_1 - \frac{\sigma^2}{2} \int \log \mu_0 \, \mu_0 dx_0$ which are constant with respect to $\rho_{0,1}$,

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\}) = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} & \int \Phi_1 \mu_1 dx_1 + \frac{\sigma_1^2}{2} \int (\log \mu_1 - \log \rho_1) \mu_1 dx_1 \\ & - \int \Phi_0 \mu_0 dx_0 - \frac{\sigma_0^2}{2} \int (\log \mu_0 - \log \rho_0) \mu_0 dx_0 \\ & - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \big\| \nabla \Phi_t \big\|^2 + \frac{\sigma_t^2}{2} \Delta \Phi_t \right) \rho_t dx_t dt \end{aligned} \tag{60}$$

Finally, the endpoint terms vanish for $\rho_t \in \Gamma(\mu_0, \mu_1)$ satisfying the endpoint constraints,

$$\begin{aligned} & \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\}) \\ & = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \big\| \nabla \Phi_t \big\|^2 + \frac{\sigma^2}{2} \Delta \Phi_t \right) \rho_t dx_t dt \end{aligned} \tag{61}$$

which matches the dual in Eq. (57). We now show that this is also the dual for the SB problem.

Schrödinger Bridge Dual Objective: Consider the optimization in Eq. (50) (here, t may be time-dependent)

$$\mathcal{S}_{SB}(\{\mu_{0,1}\}) = \inf_{\rho_t, v_t} \int_0^1 \frac{1}{2} \|v_t\|^2 \rho_t dx_t \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) + \frac{\sigma_t^2}{2} \nabla \cdot (\rho_t \, \nabla \log \rho_t), \; \rho_0 = \mu_0, \;\; \rho_1 = \mu_1 \tag{62}$$

We treat the optimization over ρ_t as an optimization over a vector space of functions, which is later constrained be normalized via the $\rho_0 = \mu_0, \rho_1 = \mu_1$ constraints and continuity equation (which preserves normalization). It is also constrained to be nonnegative, but we omit explicit constraints for simplicity of notation. The optimization over v_t is also over a vector space of functions. See App. C for additional discussion.

Given these considerations, we may now introduce Lagrange multipliers λ_0, λ_1 to enforce the endpoint constraints and Φ_t to enforce the dynamics constraint,

$$\begin{aligned} \mathcal{S}_{SB}(\{\mu_{0,1}\}) = \inf_{\rho_t, v_t} \sup_{\Phi_t, \lambda_{0,1}} \int_0^1 \frac{1}{2}\|v_t\|^2 \rho_t dx_t + \int \Phi_t \Big(\dot{\rho}_t + \nabla \cdot (\rho_t v_t) - \frac{\sigma_t^2}{2} \nabla \cdot (\rho_t \nabla \log \rho_t) \Big) dx_t \quad (63) \\ + \int \lambda_1(\rho_1 - \mu_1) dx_1 + \int \lambda_0(\rho_0 - \mu_0) dx_0 \\ = \inf_{\rho_t, v_t} \sup_{\Phi_t, \lambda_{0,1}} \int_0^1 \frac{1}{2}\|v_t\|^2 \rho_t dx_t + \int \Phi_1 \rho_1 dx_1 - \int \Phi_0 \rho_0 dx_0 - \int_0^1 \int \frac{\partial \Phi_t}{\partial t} \rho_t dx_t dt \quad (64) \\ - \int_0^1 \int \Big\langle \nabla \Phi_t, v_t - \frac{\sigma_t^2}{2} \nabla \log \rho_t \Big\rangle \rho_t dx_t dt + \int \lambda_1(\rho_1 - \mu_1) dx_1 + \int \lambda_0(\rho_0 - \mu_0) dx_0 \end{aligned}$$

Note that we can freely we can swap the order of the optimizations since the SB optimization in Eq. (62) is convex in ρ_t, v_t, while the dual optimization is linear in Φ_t, λ.

Swapping the order of the optimizations and eliminating ρ_0 and ρ_1 implies $\lambda_1 = \Phi_1$ and $\lambda_0 = \Phi_0$, while eliminating v_t implies $v_t = \nabla \Phi_t$. Finally, we obtain

$$\begin{aligned} \mathcal{S}_{SB}(\{\mu_{0,1}\}) &= \sup_{\Phi_t} \inf_{\rho_t} \int \Phi_1 \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 - \int_0^1 \Big(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\|\Phi_t\|^2 - \frac{\sigma_t^2}{2} \langle \nabla \Phi_t, \nabla \log \rho_t \rangle \Big) \rho_t dx_t \\ &= \inf_{\rho_t} \sup_{\Phi_t} \int \Phi_1 \mu_1 dx_1 - \int \Phi_0 \mu_0 dx_0 - \int_0^1 \Big(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\|\Phi_t\|^2 + \frac{\sigma_t^2}{2} \Delta \Phi_t \Big) \rho_t dx_t \quad (65) \end{aligned}$$

where we swap the order of optimization again in the second line. This matches the dual in Eq. (60) for $\mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\})$ if $\frac{\sigma^2}{2}$ is independent of time, albeit without the endpoint constraints. However, we have shown above that the optimal $\lambda_0^* = \Phi_0^*$, $\lambda_1^* = \Phi_1^*$ will indeed enforce the endpoint constraints. This is the desired result in Proposition 3. □

Example B.2 (**Schrödinger Bridge with Time-Dependent Diffusion Coefficient**)**.** To incorporate a time-dependent diffusion coefficient for the classical SB problem, we modify the potential energy with an additional term

$$\mathcal{U}[\rho_t, t] = -\frac{\sigma_t^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t dx_t + \int \Big(\frac{\partial}{\partial t} \frac{\sigma_t^2}{2} \Big) \log \rho_t \, \rho_t dx_t \tag{66}$$

This potential energy term is chosen carefully to cancel with the term appearing after reparameterization using $s_t = \Phi_t - \frac{\sigma_t^2}{2} \log \rho_t$ in Eq. (59). In this case,

$$\begin{aligned} \int \frac{\partial s_t}{\partial t} \rho_t dx_t &= \int \Big(\frac{\partial \Phi_t}{\partial t} - \frac{\partial}{\partial t} \Big(\frac{\sigma_t^2}{2} \log \rho_t \Big) \Big) \rho_t \, dx_t \quad (67) \\ &= \int \Big(\frac{\partial \Phi_t}{\partial t} - \Big(\frac{\partial}{\partial t} \frac{\sigma_t^2}{2} \Big) \log \rho_t + \frac{\sigma_t^2}{2} \Big(\frac{\partial}{\partial t} \log \rho_t \Big) \Big) \, \rho_t dx_t \quad (68) \\ &= \int \Big(\frac{\partial \Phi_t}{\partial t} - \Big(\frac{\partial}{\partial t} \frac{\sigma_t^2}{2} \Big) \log \rho_t \Big) \, \rho_t dx_t \quad (69) \end{aligned}$$

where the score term cancels as before. The additional potential energy term is chosen to cancel the remaining term. All other derivations proceed as above, which yields an identical dual objective

$$\mathcal{S}_{SB} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} \int \Phi_1 d\mu_1 - \int \Phi_0 d\mu_0 - \int_0^1 \int \Bigg(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \big\|\nabla \Phi_t\big\|^2 + \frac{\sigma_t^2}{2} \Delta \Phi_t \Bigg) \rho_t dx_t dt$$

B.3 Schrödinger Equation

Example B.3 (Schrödinger Equation)**.** Intriguingly, we obtain the Schrödinger Equation via a simple change of sign in the potential energy $\mathcal{U}[\rho_t, t] = \frac{\sigma_t^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t dx_t$ compared to Eq. (51)

or, in other words, an imaginary weighting $i\sigma_t$ of the gradient norm of the Shannon entropy,

$$\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \frac{1}{2}\langle \dot{\rho}_t, \dot{\rho}_t \rangle^{W_2}_{T_{\rho_t}} - \int \left[\frac{1}{8}\|\nabla \log \rho_t\|^2 + V_t(x_t)\right] \rho_t \, dx_t \tag{70}$$

This Lagrangian corresponds to a Hamiltonian $\mathcal{H}[\rho_t, s_t, t] = \frac{1}{2}\langle s_t, s_t \rangle^{W_2}_{T^*_{\rho_t}} + \int \left[\frac{1}{8}\|\nabla \log \rho_t\|^2 + V_t(x_t)\right] \rho_t \, dx_t$, which leads to the dual objective

$$\begin{aligned} \mathcal{S}_{SE} = \; & \sup_{s_t} \inf_{\rho_t} \int s_1 d\mu_1 - \int s_0 d\mu_0 \\ & - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 + \frac{1}{8}\|\nabla \log \rho_t\|^2 + V_t(x_t) \right) \rho_t dx_t dt. \end{aligned} \tag{71}$$

Unlike the Schrödinger Bridge problem, the Hopf-Cole transform does not linearize the dual objective in density. Thus, we cannot approximate the dual using only the Monte Carlo estimate.

The first-order optimality conditions for Eq. (71) are

$$\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_t), \quad \frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 = \frac{1}{8}\|\nabla \log \rho_t\|^2 + \frac{1}{4}\Delta \log \rho_t - V_t(x_t) \tag{72}$$

Note, that Eq. (72) is the Madelung transform of the Schrödinger equation, i.e. for the equation

$$\frac{\partial}{\partial t}\psi_t(x) = -i\hat{H}\psi_t(x), \quad \text{where} \quad \hat{H} = -\frac{1}{2}\Delta + V_t(x), \tag{73}$$

the wave function $\psi_t(x)$ can be written in terms $\psi_t(x) = \sqrt{\rho_t(x)} \exp(i s_t(x))$. Then the real and imaginary part of the Schrödinger equation yield Eq. (72).

C Lagrange Multiplier Approach

Our Thm. 1 is framed completely in the abstract space of densities and the Legendre transform between functionals of $\dot{\rho}_t \in \mathcal{T}_{\rho_t}\mathcal{P}$ and $s_{\dot{\rho}_t} \in \mathcal{T}^*_{\rho_t}\mathcal{P}$. We contrast this approach with optimizations such as the Benamou-Brenier formulation in Eq. (2), which are formulated in terms of the state space dynamics such as the continuity equation $\dot{\rho}_t = -\nabla \cdot (\rho_t v_t)$. In this appendix, we claim that the latter approaches require a potential energy $\mathcal{U}[\rho_t, t]$ which is concave or linear in ρ_t. We restrict attention to continuity equation dynamics in this section, although similar reasoning holds with growth terms.

In particular, consider optimizing ρ_t, v_t over a topological vector space of functions. The notable difference here is that $\rho_t : \mathcal{X} \to \mathbb{R}$ is a function, which we later constrain to be a normalized probability density using $\rho_0 = \mu_0, \rho_1 = \mu_1$, the continuity equation $\dot{\rho}_t = -\nabla \cdot (\rho_t v_t)$ (which preserves normalization), and nonnegativity constraints. Omitting the latter for simplicity of notation, we consider the W_2 kinetic energy with an arbitrary potential energy,

$$\mathcal{S} = \inf_{\rho_t, v_t} \int_0^1 \int L(x_t, v_t, t)\rho_t dx_t dt - \int_0^1 \mathcal{U}[\rho_t, t] dt \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) \quad \rho_0 = \mu_0, \; \rho_1 = \mu_1 \tag{74}$$

Since we are now optimizing ρ_t over a vector space, we introduce Lagrange multipliers $\lambda_{0,1}$ to enforce the endpoint constraints and s_t to enforce the continuity equation. Integrating by parts in t and x, we have

$$\begin{aligned} \mathcal{S} = & \inf_{\rho_t, v_t} \sup_{\lambda_{0,1}, s_t} \int_0^1 \int L(x_t, v_t, t)\rho_t dx_t dt - \int_0^1 \mathcal{U}[\rho_t, t] dt + \int_0^1 \int s_t \dot{\rho}_t dx_t + \int_0^1 \int s_t \, \nabla \cdot (\rho_t v_t) dx_t dt \\ & \qquad + \int \lambda_0 (\rho_0 - \mu_0) dx_0 + \int \lambda_1 (\rho_1 - \mu_1) dx_1 \end{aligned} \tag{75}$$

$$\begin{aligned} = & \inf_{\rho_t, v_t} \sup_{\lambda_{0,1}, s_t} \int_0^1 \int L(x_t, v_t, t)\rho_t dx_t dt - \int_0^1 \mathcal{U}[\rho_t, t] dt - \int_0^1 \int \frac{\partial s_t}{\partial t} \rho_t dx_t - \int_0^1 \int \langle \nabla s_t, v_t \rangle \rho_t \, dx_t dt \\ & \qquad + \int \lambda_1 \rho_1 dx_1 - \int \lambda_0 \rho_0 dx_0 + \int \lambda_0 \rho_0 dx_0 - \int \lambda_0 \mu_0 \, dx_0 + \int \lambda_1 \rho_1 \, dx_1 - \int \lambda_1 \mu_1 \, dx_1 \end{aligned} \tag{76}$$

To make further progress by swapping the order of the optimizations, we require that Eq. (76) is convex in ρ_t, v_t and concave in $\lambda_{0,1}, s_t$. However, to facilitate this, we require that $\mathcal{U}[\rho_t, t]$ is concave in ρ_t, which is an additional constraint which was not necessary in the proof of Thm. 1.

By swapping the order of optimization to eliminate ρ_0, ρ_1 and v_t, we obtain the optimality conditions

$$\lambda_0 = s_0,\ \lambda_1 = s_1 \qquad\qquad v_t = \nabla_p H(x_t, \nabla s_t, t) \tag{77}$$

where the gradient is with respect to the second argument. Swapping the order of optimizations again, the dual becomes

$$\mathcal{S} = \inf_{\rho_t} \sup_{s_t} \int s_1 \mu_1 \, dx_1 - \int s_0 \mu_0 \, dx_0 - \int_0^1 \left(\int \Big(\frac{\partial s_t}{\partial t} + H(x_t, \nabla s_t, t) \Big) \rho_t dx_t + \mathcal{U}[\rho_t, t] \right) dt.$$

which is analogous to Eq. (13) in Thm. 1 for the W_2 kinetic energy. While the dual above does not explicitly enforce the endpoint marginals on ρ_t, the conditions $\lambda_0^* = s_0^*,\ \lambda_1^* = s_1^*$ serve to enforce the constraint at optimality.

D Expressivity of Parameterization

Proposition 2. *For any absolutely-continuous distributional path $\rho_t : [0,1] \mapsto \mathcal{P}_2(\mathcal{X})$ on the W_2 manifold, there exists a function $\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta)$ such that Eq. (16) samples from ρ_t.*

Proof. For every absolutely-continuous distributional path ρ_t, we have a unique gradient flow $\nabla s_t^*(x_t)$ satisfying the continuity equation (Ambrosio et al. [3] Thm. 8.3.1),

$$\dot{\rho_t} = -\nabla \cdot (\rho_t \nabla s_t^*(x_t)) \, . \tag{78}$$

Consider the function

$$\varphi_t(x_0, x_1) = \begin{cases} x_0 + \int_0^t \nabla s_\tau^*(x_\tau) d\tau, & t \le 1/2 \, , \\ x_1 + \int_1^t \nabla s_\tau^*(x_\tau) d\tau, & t > 1/2 \, , \end{cases} \tag{79}$$

which integrates the ODE $dx/dt = \nabla s_t^*(x_t)$ forward starting from x_0 for $t \le 1/2$, and integrates the same ODE backwards starting from x_1 otherwise.

Clearly, for $t \le 1/2$ the designed function serves as a push-forward map for the samples $x_0 \sim \rho_0$, and produces samples from ρ_t by Eq. (78). The same applies for $t > 1/2$. Thus, φ_t samples from the correct marginals, i.e.

$$\int \delta(x_t - \varphi_t(x_0, x_1)) \rho_0(x_0) \rho_1(x_1) dx_0 dx_1 = \rho_t(x_t), \ \ \forall t \in [0,1]. \tag{80}$$

We now show that $\varphi_t(x_0, x_1)$ can be expressed using the parameterization in Eq. (16), which constructs x_t as

$$x_t = (1-t)x_0 + t x_1 + t(1-t)\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta), \ \ x_0 \sim \mu_0, \ \ x_1 \sim \mu_1. \tag{81}$$

Then taking the function $\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta)$ as follows

$$\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta) = \begin{cases} \frac{1}{1-t} \Big(x_0 - x_1 + \frac{1}{t} \int_0^t \nabla s_\tau^*(x_\tau) d\tau \Big), & t \le 1/2 \, , \\ \frac{1}{t} \Big(x_1 - x_0 + \frac{1}{1-t} \int_1^t \nabla s_\tau^*(x_\tau) d\tau \Big), & t > 1/2 \, , \end{cases} \tag{82}$$

we have

$$(1-t)x_0 + t x_1 + t(1-t)\text{NNET}^*(t, x_0, x_1; \mathbb{1}[t < 0.5]; \eta) = \varphi_t(x_0, x_1), \tag{83}$$

which samples from the correct marginals by construction. □

E Details of Experiments

E.1 Single-cell Experiments

We consider low dimensional (Table 2) and high dimensional (Table 1) single-cell experiments following the experimental setups in Tong et al. [63, 62]. The Embryoid body (**EB**) dataset Moon et al. [46] and the CITE-seq (**Cite**) and Multiome (**Multi**) datasets [11] are repurposed and preprocessed by Tong et al. [63, 62] for the task of trajectory inference.

The **EB** dataset is a scRNA-seq dataset of human embryonic stem cells used to observe differentiation of cell lineages [46]. It contains approximately 16,000 cells (examples) after filtering, of which the first 100 principle components over the feature space (gene space) are used. For the low dimensional (5-dim) experiments, we consider only the first 5 principle components. The **EB** dataset comprises a collection of 5 timepoints sampled over a period of 30 days.

The **Cite** and **Multi** datasets are taken from the Multimodal Single-cell Integration challenge at NeurIPS 2022 [11]. Both datasets contain single-cell measurements from CD4+ hematopoietic stem and progenitor cells (HSPCs) for 1000 highly variables genes and over 4 timepoints collected on days 2, 3, 4, and 7. We use the **Cite** and **Multi** datasets for both low dimensional (5-dim) and high dimensional (50-dim, 100-dim) experiments. We use 100 computed principle components for the 100-dim experiments, then select the first 50 and first 5 principle components for the 50-dim and 5-dim experiments, respectively. Further details regarding the raw dataset can be found at the competition website. [4]

For all experiments, we train k independent models over k partitions of the single-cell datasets. The training data partition is determined by a left out intermediary timepoint. We then average test performance over the k independent model predictions computed on the respective left-out marginals. For experiments using the **EB** dataset, we train 3 independent models using marginals from timepoint partitions $[1, 3, 4, 5]$, $[1, 2, 4, 5]$, $[1, 2, 3, 5]$ and evaluate each model using the respective left-out marginals at timepoints $[2]$, $[3]$, $[4]$. Likewise, for experiments using **Cite** and **Multi** datasets, we train 2 independent models using marginals from timepoint partitions $[2, 4, 7]$, $[2, 3, 7]$ and evaluate each model using the respective left-out marginals at timepoints $[3]$, $[4]$.

For both $s_t(x, \theta)$ and $\rho_t(x, \eta)$, we consider Multi-Layer Perceptron (MLP) architectures and a common optimizer [41]. For detailed description of the architectures and hyperparameters we refer the reader to the code supplemented.

E.2 Single-step Image Generation via Optimal Transport

Learning the vector field that corresponds to the optimal transport map between some prior distribution (e.g. Gaussian) and the target data allows to generate data samples evaluating the vector field only once. Indeed, the optimality condition (Hamilton-Jacobi equation) for the dynamical optimal transport yields

$$\ddot{X}_t = \nabla \left[\frac{\partial s_t(x_t)}{\partial t} + \frac{1}{2} \|\nabla s_t(x_t)\|^2 \right] = 0 \,, \tag{84}$$

hence, the acceleration along every trajectory is zero. This implies that the learned vector field can be trivially integrated, i.e.

$$X_1 = X_0 + \nabla s_0(X_0) \,. \tag{85}$$

Thus, X_1 is generated with a single evaluation of $\nabla s_0(\cdot)$.

For the image generation experiments, we follow common practices of training the diffusion models [60], i.e. the vector field model $s_t(x, \theta)$ uses the U-net architecture [54] with the time embedding and hyperparameters from [60]. For the distribution path model $\rho_t(x, \eta)$, we found that the U-net architectures works best as well. For detailed description of the architectures and hyperparameters we refer the reader to the code supplemented.

[4] https://www.kaggle.com/competitions/open-problems-multimodal/data

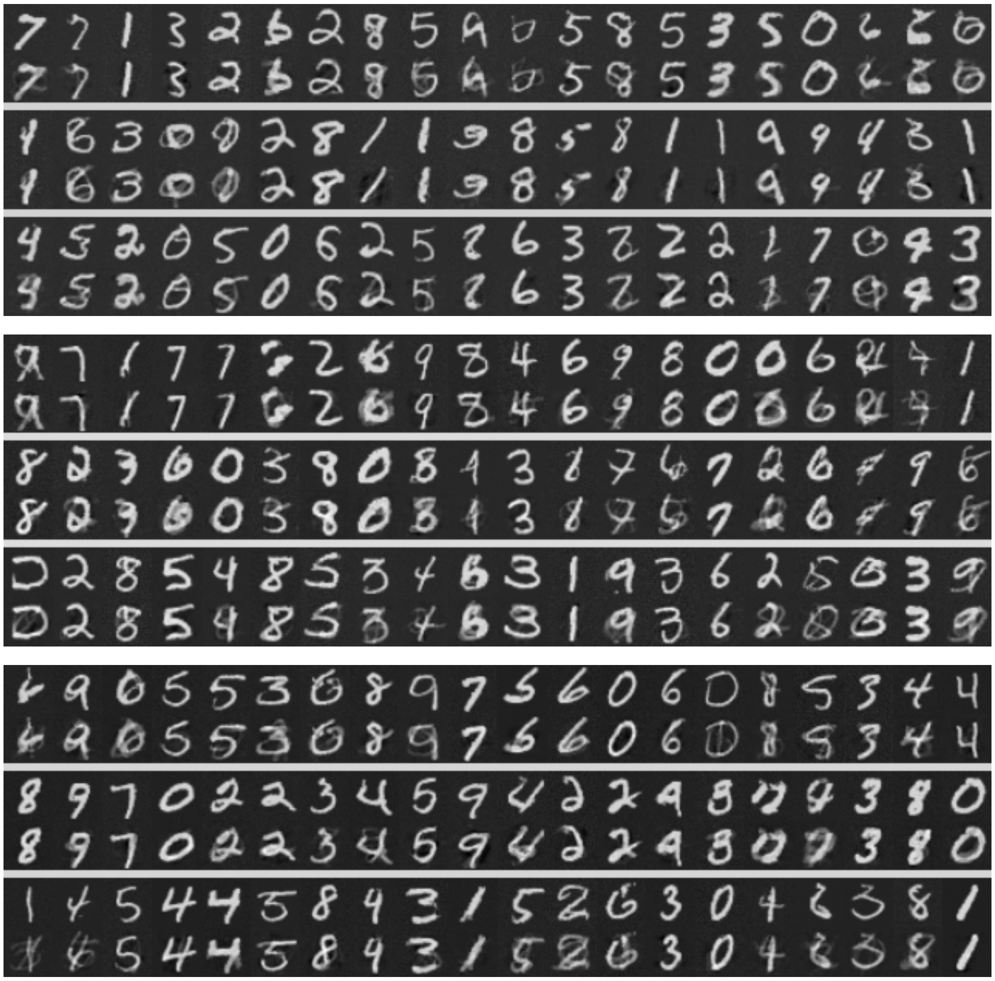


Figure 3: MNIST 32x32 image generation. Every top row is the integration of the corresponding ODE via Dormand-Prince's $5/4$ method, which makes 108 function evaluations. Every bottom row corresponds to single function evaluation approximation.

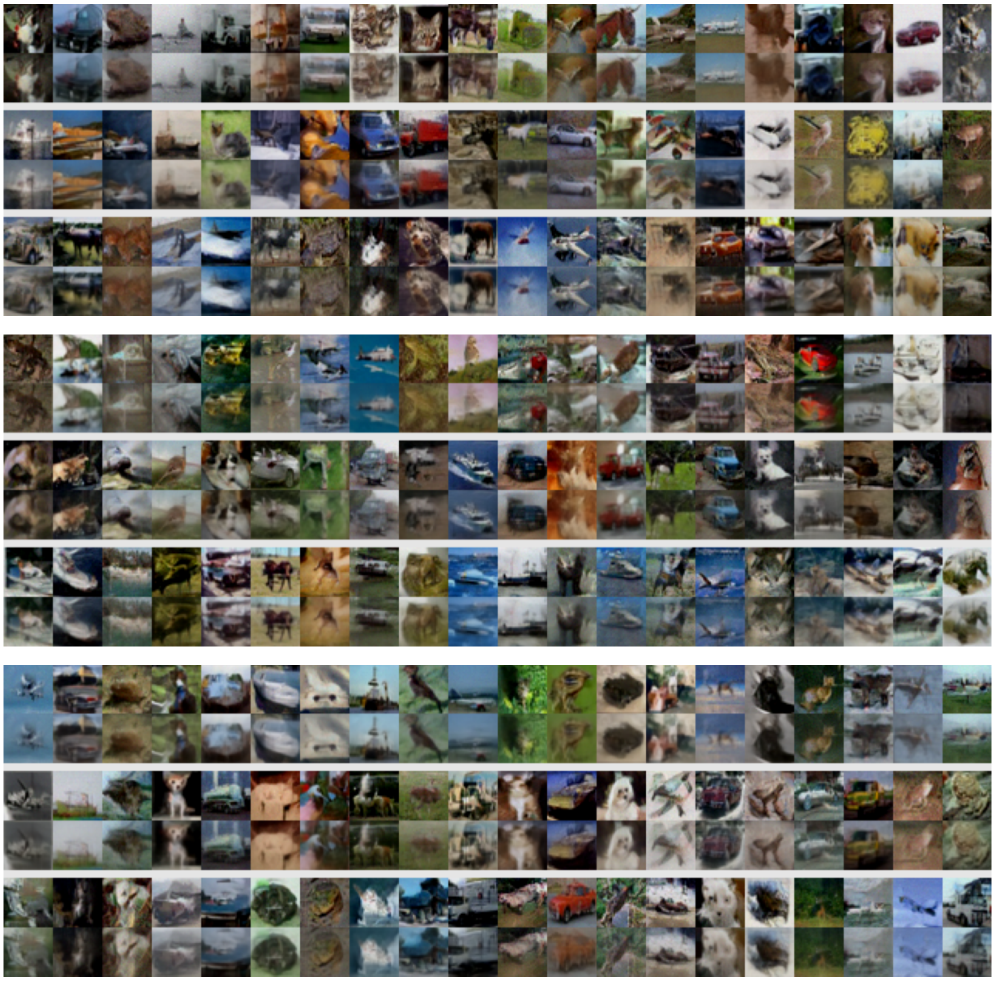


Figure 4: CIFAR-10 image generation. Every top row is the integration of the corresponding ODE via Dormand-Prince's $5/4$ method, which makes 78 function evaluations. Every bottom row corresponds to single function evaluation approximation.