A Computational Framework for Solving Wasserstein Lagrangian Flows

Kirill Neklyudov [*] [1] [2] **Rob Brekelmans** [*] [3] **Alexander Tong** [1] [2] **Lazar Atanackovic** [3] [4]
Qiang Liu [5] **Alireza Makhzani** [3] [4]

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. We propose a novel deep learning based framework approaching all of these problems from a unified perspective. Leveraging the dual formulation of the Lagrangians, 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 challenging problem arising throughout the natural sciences (Hashimoto et al., 2016; Lavenant et al., 2021). A particularly important application is analysis of single-cell RNA-sequencing data (Schiebinger et al., 2019; Schiebinger, 2021; Saelens et al., 2019), which provides a heterogeneous snapshot of a cell population at a high resolution, allowing high-throughput observation over tens of thousands of genes (Macosko et al., 2015). 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.

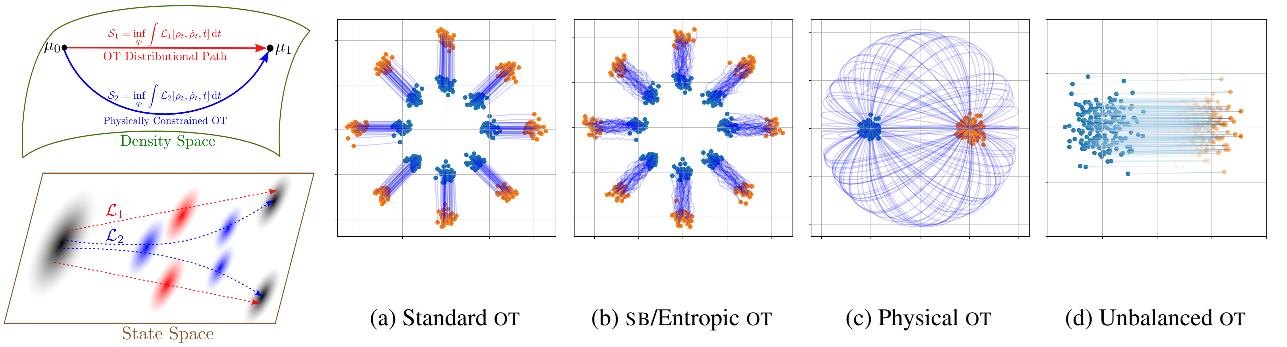
However, when inferring trajectories over cell distributions, there exist multiple cell dynamics that yield the same population marginals. This presents an ill-posed problem, which highlights the need for trajectory inference methods that can flexibly incorporate different types of prior information on the cell dynamics. Commonly, such prior information is specified via posing a variational problem on the space of marginal distributions, where previous works on measure-valued splines (Chen et al., 2018; Benamou et al., 2019; Chewi et al., 2021; Clancy & Suarez, 2022; Chen et al., 2023) are examples which seek to 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 densities that can flexibly incorporate various types of prior information. We consider Lagrangians of the form $\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \mathcal{K}[\rho_t, \dot{\rho}_t] - \mathcal{U}[\rho_t, t]$, referring to the first term as a *kinetic energy* and the second as a *potential energy*. Our methods can solve a diverse set 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, such as the expectation of a physical potential that encodes 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.

[*] Joint first authorship [1] Université de Montréal [2] Mila - Quebec AI Institute [3] Vector Institute [4] University of Toronto [5] University of Texas at Austin. Correspondence to: k.necludov@gmail.com , {rob.brekelmans, makhzani}@vectorinstitute.ai .

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(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 particular 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 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}.

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. (2023).

We refer to the optimal dynamics specified by these choices as *Wasserstein Lagrangian Flows* and propose a single computational approach to solve a class of such problems, including (i) *standard Wasserstein-2 OT* (Ex. 4.1, Benamou & Brenier (2000); Villani (2009)), (ii) *entropy regularized OT* or Schrödinger Bridge (Ex. 4.4, Léonard (2013); Chen et al. (2021b), (iii) *physically constrained OT* (Ex. 4.3, Tong et al. (2020); Koshizuka & Sato (2022)), and (iv) *unbalanced OT* (Ex. 4.2, Chizat et al. (2018a)) (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 exploring different Lagrangians can improve results in single-cell RNA-sequencing applications. We compare our approach 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 as a cost-minimization problem over joint distributions or 'couplings' $\pi \in \Pi(\mu_0, \mu_1) = \{\pi(x_0, x_1) \,|\, \pi(x_0) = \mu_0(x_0),\ \pi(x_1) = \mu_1(x_1)\}$, i.e.

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

The dynamical formulation of Benamou & Brenier (2000) 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 (2021) Lemma 4.1.1), and we have

$$\begin{aligned} \frac{1}{2} W_2(\mu_0, \mu_1)^2 = \inf_{\rho_t} \inf_{v_t} \int_0^1 \int \frac{1}{2} \|v_t\|^2 \, \rho_t \,\mathrm{d}x\,\mathrm{d}t \\ \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t), \quad \rho_0 = \mu_0, \ \ \rho_1 = \mu_1, \end{aligned} \tag{2}$$

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 (2001); Ambrosio et al. (2008), see also Figalli & Glaudo (2021) 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. (2008) characterizing the tangent space $T^{W_2}_\rho \mathcal{P}_2 = \{\dot{\rho} \,|\, \int \dot{\rho}\,\mathrm{d}x = 0\}$ via the continuity equation,

$$T^{W_2}_\rho \mathcal{P}_2(\mathcal{X}) = \{\dot{\rho} \,|\, \dot{\rho} = -\nabla \cdot (\rho \nabla s)\}. \tag{3}$$

We also write the cotangent space as $T^{*W_2}_\rho \mathcal{P}_2(\mathcal{X}) = \{[s] \,|\, s \in \mathcal{C}^\infty(\mathcal{X})\}$, 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

$$\langle \dot{\mu}_t, \dot{\rho}_t \rangle^{W_2}_{T_\rho} = \langle s_{\dot{\mu}_t}, s_{\dot{\rho}_t} \rangle^{W_2}_{T^*_\rho} = \int \langle \nabla s_{\dot{\mu}_t}, \nabla s_{\dot{\rho}_t} \rangle \rho \ \mathrm{d}x. \tag{4}$$

2.2. Wasserstein Fisher-Rao Geometry

Building from the dynamical formulation in Eq. (2), Chizat et al. (2018a;b); Kondratyev et al. (2016); Liero et al. (2016; 2018) consider additional terms allowing for birth and death of particles or teleportation of probability mass. In particular, we extend the continuity equation with a 'growth' term

$g_t : \mathcal{X} \to \mathbb{R}$ whose square is regularized in the cost,

$$\frac{1}{2} WFR_\lambda(\mu_0,\mu_1)^2 = \inf_{\rho_t} \inf_{v_t, g_t} \int_0^1 \int \Big(\frac{1}{2}\|v_t\|^2 + \frac{\lambda}{2} g_t^2\Big)\rho_t \, \mathrm{d}x \, \mathrm{d}t,$$
$$\text{s.t.} \quad \dot\rho_t = -\nabla\cdot(\rho_t v_t) + \lambda \rho_t g_t, \;\; \rho_0 = \mu_0, \;\; \rho_1 = \mu_1. \tag{5}$$

We call this the Wasserstein Fisher-Rao (WFR) distance, since considering *only* the growth terms recovers the non-parametric Fisher-Rao metric (Chizat et al., 2018a; Bauer et al., 2016). We also refer to Eq. (5) as an *unbalanced* OT problem on the space of unnormalized densities $\mathcal{M}(\mathcal{X})$, since the growth terms need not preserve normalization $\int \dot\rho_t \, \mathrm{d}x = \int \lambda g_t \rho_t \, \mathrm{d}x \neq 0$ without further modifications (see e.g. Lu et al. (2019)).

Kondratyev et al. (2016) 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

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

$$\begin{aligned} \langle \dot\mu_t, \dot\rho_t \rangle_{T_\rho}^{WFR_\lambda} &= \langle s_{\dot\mu_t}, s_{\dot\rho_t} \rangle_{T^*_\rho}^{WFR_\lambda} \\ &= \int \big(\langle \nabla s_{\dot\mu_t}, \nabla s_{\dot\rho_t}\rangle + \lambda \, s_{\dot\mu_t} s_{\dot\rho_t}\big)\rho \, \mathrm{d}x. \end{aligned} \tag{7}$$

2.3. Action Matching

Finally, Action Matching (AM) (Neklyudov et al., 2023) 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. (2023) optimize the objective

$$\begin{aligned} \mathcal{A}[\mu_t] = \sup_{s_t} &\int s_1 \mu_1 \, \mathrm{d}x - \int s_0 \mu_0 \, \mathrm{d}x \\ &- \int_0^1 \int \Big(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2\Big)\mu_t \, \mathrm{d}x \, \mathrm{d}t, \end{aligned} \tag{8}$$

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

$$\begin{aligned} \mathcal{A}[\mu_t] &= \int_0^1 \frac{1}{2} \langle \dot\mu_t, \dot\mu_t \rangle_{T_{\mu_t}}^{W_2} \mathrm{d}t = \int_0^1 \frac{1}{2} \langle s_{\dot\mu_t}, s_{\dot\mu_t} \rangle_{T^*_{\mu_t}}^{W_2} \mathrm{d}t \\ &= \int_0^1 \int \frac{1}{2} \|\nabla s_t\|^2 \mu_t \, \mathrm{d}x \, \mathrm{d}t. \end{aligned} \tag{9}$$

Our proposed framework 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 $\mathcal{P}(\mathcal{X})$.[1] 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 (2009) Ch. 7).

We instead 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 energy functionals 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 global information about the distribution of particles or cells (as in the SB problem, Ex. 4.4).

3.1. Wasserstein Lagrangian and Hamiltonian Flows

We consider Lagrangian action functionals on the space of densities, defined in terms of a *kinetic energy* $\mathcal{K}[\rho_t, \dot\rho_t]$, 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,

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

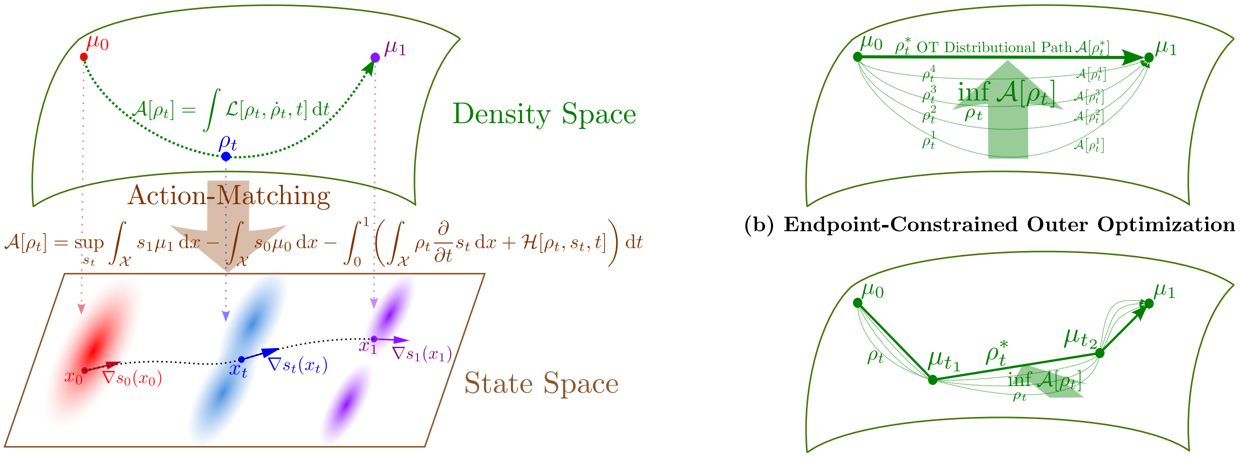
Throughout, we will assume $\mathcal{L}[\rho_t, \dot\rho_t, t]$ is lower semi-continuous (lsc) and strictly convex in $\dot\rho_t$.

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 $\mathcal{A}_\mathcal{L}[\rho_t] = \int_0^1 \mathcal{L}[\rho_t, \dot\rho_t, t] \, \mathrm{d}t$ as the time-integral of the Lagrangian and seek the action-minimizing curve subject to the constraints

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

where $\Gamma(\{\mu_{t_i}\}) = \{\rho_t : [0,1] \to \mathcal{P}(\mathcal{X}) \mid \rho_0 = \mu_0, \; \rho_1 = \mu_1, \; \rho_{t_i} = \mu_{t_i} \;\; (\forall\, 1 \le i \le M-2)\}$ indicates the set of curves matching the given constraints. We proceed writing the value as $\mathcal{S}$, leaving the dependence on $\mathcal{L}$ and μ_{t_i} implicit, but note $\mathcal{S}(\mu_{0,1})$ for $M = 2$ as an important special case.

[1] 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
$\mathcal{A}[\rho_t] = \int \mathcal{L}[\rho_t, \dot\rho_t, t]\,dt$
ρ_t
Density Space
Action-Matching
$\mathcal{A}[\rho_t] = \sup_{s_t} \int_{\mathcal{X}} s_1\mu_1\,dx - \int_{\mathcal{X}} s_0\mu_0\,dx - \int_0^1 \left(\int_{\mathcal{X}} \rho_t \frac{\partial}{\partial t} s_t\,dx + \mathcal{H}[\rho_t, s_t, t]\right) dt$
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 velocities 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.

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 $s_t[\cdot] : \dot\rho_t \mapsto \int s_t \dot\rho_t\,dx$ via the canonical duality bracket. We define the *Hamiltonian* $\mathcal{H}[\rho_t, s_t, t]$ via the Legendre transform

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

where the sign of $\mathcal{U}[\rho_t, t]$ changes and $\mathcal{K}^*[\rho_t, s_t]$ translates the kinetic energy to the dual space. A primary example is when $\mathcal{K}[\rho_t, \dot\rho_t] = \frac{1}{2}\langle \dot\rho_t, \dot\rho_t \rangle_{T_{\rho_t}}$ is given by a Riemannian metric in the tangent space (such as for W_2 or WFR_λ), then $\mathcal{K}^*[\rho_t, s_t] = \frac{1}{2}\langle s_t, s_t \rangle_{T^*_{\rho_t}}$ 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 be written using the Legendre transform,

$$\mathcal{L}[\rho_t, \dot\rho_t, t] = \sup_{s_t \in T^*_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t\,dx - \mathcal{H}[\rho_t, s_t, t]. \tag{13}$$

The following theorem forms the basis for our computational approach, and can be derived using this 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

$$\begin{aligned}\mathcal{S} = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1\mu_1\,dx - \int s_0\mu_0\,dx \\ - \int_0^1 \left(\int \frac{\partial s_t}{\partial t}\rho_t\,dx + \mathcal{H}[\rho_t, s_t, t]\right) dt\end{aligned} \tag{14}$$

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).

The action $\mathcal{A}_{\mathcal{L}}[\rho_t]$ of a given curve is the solution to the inner optimization,

$$\begin{aligned}\mathcal{A}_{\mathcal{L}}[\rho_t] = \sup_{s_t} \int s_1\mu_1\,dx - \int s_0\mu_0\,dx \\ - \int_0^1 \left(\int \frac{\partial s_t}{\partial t}\rho_t\,dx + \mathcal{H}[\rho_t, s_t, t]\right) dt.\end{aligned} \tag{15}$$

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 discussion.

Finally, the solution to Eq. (11) can also be expressed in a Hamiltonian perspective, where the resulting *Wasserstein Hamiltonian flows* (Chow et al., 2020) satisfy optimality conditions $\frac{\partial \rho_t}{\partial t} = \frac{\delta}{\delta s_t}\mathcal{H}[\rho_t, s_t, t]$ and $\frac{\partial s_t}{\partial t} = -\frac{\delta}{\delta \rho_t}\mathcal{H}[\rho_t, s_t, t]$.

To further analyze Thm. 1 and set the stage for our computational approach in Sec. 3.2, we consider the two optimizations in Eq. (14) as (i) ***evaluating*** the action functional $\mathcal{A}_{\mathcal{L}}[\rho_t]$ given a 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. (15) to the AM objective in Eq. (8) for $\mathcal{H}[\rho_t, s_t, t] = \int \frac{1}{2}\|\nabla s_t\|^2 \rho_t \, \mathrm{d}x$, 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. (15).

Neklyudov et al. (2023) 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 enforces $\rho_t \in \Gamma(\{\mu_{t_i}\})$ 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. (15), 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 Objectives

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 exists a function $\mathbb{H}(x, s_t, t)$ s.t.*

$$\mathcal{H}[\rho_t, s_t, t] := \int \mathbb{H}(x, s_t, t) \rho_t(x) \, \mathrm{d}x \,. \tag{16}$$

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. [2]

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 $\{t^i\}_{i=1}^n \sim \textsc{uniform}[0, 1]$
    $x_t^i = (1 - t^i)x_0^i + t^i x_1^i + t^i(1 - t^i)\textsc{nnet}(t^i, x_0^i, x_1^i; \eta)$
    $\textsc{grad}_\eta = -\nabla_\eta \frac{1}{n} \sum_i^n \big[\frac{\partial s_{t^i}}{\partial t^i}(x_t^i(\eta), \theta) +$
            $+ \mathbb{H}\big(x_t^i(\eta), s_{t^i}(x_t^i(\eta), \theta), t^i\big)\big]$
    for Wasserstein gradient steps do
      $x_t^i \leftarrow x_t^i + \alpha \cdot t^i(1 - t^i)\nabla_x \big[\frac{\partial s_{t^i}}{\partial t^i}(x_t^i, \theta)$
            $+ \mathbb{H}\big(x_t^i, s_{t^i}(x_t^i, \theta), t^i\big)\big]$
    end for
    $\textsc{grad}_\theta = \nabla_\theta \frac{1}{n} \sum_i^n \big[s_1(x_1^i, \theta) - s_0(x_0^i, \theta) - \frac{\partial s_{t^i}(x_t^i, \theta)}{\partial t^i}$
            $- \mathbb{H}\big(x_t^i, s_{t^i}(x_t^i, \theta), t^i\big)\big]$
    update parameters $\eta$ using gradient descent with $\textsc{grad}_\eta$
    update parameters $\theta$ using gradient ascent with $\textsc{grad}_\theta$
  end for
output cotangent vectors $s_t(x, \theta)$
```

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

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

$$\begin{aligned} &x_t = (1 - t)x_0 + t x_1 + t(1 - t)\textsc{nnet}(t, x_0, x_1; \eta), \\ &x_0 \sim \mu_0, \qquad x_1 \sim \mu_1 \,. \end{aligned} \tag{17}$$

Notably, this preserves the endpoint marginals μ_0, μ_1. For multiple constraints, we can modify our sampling procedure

[2] 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. (14) may require explicit modeling of the density ρ_t corresponding to a given set of particles x_t (e.g. see Pfau et al. (2020)).

to interpolate between two intermediate dataset marginals, with neural network parameters η shared across timesteps

$$x_t = \tilde{\beta}_t x_{t_i} + \hat{\beta}_t x_{t_{i+1}} + \Big(1 - \tilde{\beta}_t^2 - \hat{\beta}_t^2\Big)\text{NNET}(t, x_{t_i}, x_{t_{i+1}}; \eta),$$
$$\tilde{\beta}_t = \frac{t_{i+1} - t}{t_{i+1} - t_i}, \quad \hat{\beta}_t = \frac{t - t_i}{t_{i+1} - t_i}\,. \tag{18}$$

For linearizable dual objectives as in Eq. (14) and Def. 3.1, we optimize

$$\min_{\eta}\max_{\theta} \int s_1(x,\theta)\mu_1\,\mathrm{d}x - \int s_0(x,\theta)\mu_0\,\mathrm{d}x \tag{19}$$
$$-\int_0^1\int\Big(\frac{\partial s_t}{\partial t}(x,\theta) + \mathbb{H}\big(x, s_t(x,\theta), t\big)\Big)\rho_t(x,\eta)\,\mathrm{d}x\,\mathrm{d}t,$$

where the optimization w.r.t. η is performed via the reparameterization trick.

Wasserstein Gradient Approach An alternative to parametrizing the distributional path ρ_t is to perform minimization of Eq. (19) via the Wasserstein gradient flow, i.e. the samples x_t from the initial path ρ_t are updated by

$$x_t' = x_t + \alpha\cdot t(1-t)\nabla_x\left[\frac{\partial s_t}{\partial t}(x_t,\theta) + \mathbb{H}\big(x_t, s_t(x_t,\theta), t\big)\right],$$

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 $\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t<0.5]; \eta)$ such that Eq. (17) 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, 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 do 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 $\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \frac{1}{2}\langle\dot{\rho}_t, \dot{\rho}_t\rangle^{W_2}_{T_{\rho_t}} = \mathcal{H}[\rho_t, s_{\dot{\rho}_t}, t] = \frac{1}{2}\int\|\nabla s_{\dot{\rho}_t}\|^2\rho_t\,\mathrm{d}x$. 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 objective for solving OT problem with quadratic cost is

$$\mathcal{S}_{OT} = \inf_{\rho_t\in\Gamma(\mu_0,\mu_1)}\sup_{s_t}\int s_1\mu_1\,\mathrm{d}x - \int s_0\mu_0\,\mathrm{d}x \tag{20}$$
$$-\int_0^1\int\Big(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2\Big)\rho_t\,\mathrm{d}x\,\mathrm{d}t,$$

where the Hamiltonian optimality conditions $\frac{\partial\rho_t}{\partial t} = \frac{\delta}{\delta s_t}\mathcal{H}[\rho_t, s_t, t]$ and $\frac{\partial s_t}{\partial t} = -\frac{\delta}{\delta\rho_t}\mathcal{H}[\rho_t, s_t, t]$ (Chow et al., 2020) recover the well-known characterization of the Wasserstein geodesics via the continuity and Hamilton-Jacobi equations (Benamou & Brenier, 2000),

$$\dot{\rho}_t = -\nabla\cdot(\rho_t\nabla s_t) \qquad \frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 = 0. \tag{21}$$

It is well known that optimal transport plans (or Wasserstein-2 geodesics) are 'straight-paths' in the Euclidean space (Villani, 2009). For the flow induced by a vector field ∇s_t, we calculate the acceleration, or second time derivative, as

$$\ddot{X}_t = \nabla\left[\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2\right] = 0, \tag{22}$$

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. (21).

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 (Schiebinger et al., 2019; Lu et al., 2019). Viewing the dynamical formulation of WFR in Eq. (5) as a Lagrangian optimization,

$$\frac{1}{2}WFR_\lambda(\mu_0,\mu_1)^2 = \inf_{\rho_t\in\Gamma(\mu_0,\mu_1)}\int_0^1\mathcal{L}[\rho_t, \dot{\rho}_t, t]\,\mathrm{d}t \tag{23}$$
$$= \int_0^1\frac{1}{2}\,\langle\dot{\rho}_t, \dot{\rho}_t\rangle^{WFR_\lambda}_{T_{\rho_t}}\,\mathrm{d}t \;\text{ s.t. }\; \rho_0 = \mu_0, \rho_1 = \mu_1.$$

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

$$\mathcal{S}_{uOT} = \inf_{\rho_t\in\Gamma(\mu_0,\mu_1)}\sup_{s_t}\int s_1\mu_1\,\mathrm{d}x - \int s_0\mu_0\,\mathrm{d}x \tag{24}$$
$$-\int_0^1\int\Big(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 + \frac{\lambda}{2}s_t^2\Big)\rho_t\,\mathrm{d}x\,\mathrm{d}t,$$

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, 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)	$0.822 \pm 3.0\text{e-}4$	37.821 ± 0.010	47.268 ± 0.017	44.013 ± 0.010	54.253 ± 0.012
OT-CFM (Tong et al., 2023b)	$\mathbf{0.790 \pm 0.068}$	38.756 ± 0.398	47.576 ± 6.622	45.393 ± 0.416	54.814 ± 5.858
I-CFM (Tong et al., 2023b)	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)	$0.816 \pm 7.7\text{e-}4$	39.240 ± 0.068	47.788 ± 0.111	46.177 ± 0.083	55.716 ± 0.058
[SF]^2 M-Geo (Tong et al., 2023a)	0.879 ± 0.148	38.524 ± 0.293	$\mathbf{44.795 \pm 1.911}$	44.498 ± 0.416	$\mathbf{52.203 \pm 1.957}$
[SF]^2 M-Exact (Tong et al., 2023a)	$\mathbf{0.793 \pm 0.066}$	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)	$\mathbf{0.634 \pm 0.001}$	$\mathbf{34.160 \pm 0.041}$	$\mathbf{36.131 \pm 0.023}$	$\mathbf{41.084 \pm 0.043}$	$\mathbf{45.231 \pm 0.010}$

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

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 $\mathcal{U}[\rho_t, t] = \int V_t(x)\rho_t \, \mathrm{d}x$ (Tong et al., 2020; Koshizuka & Sato, 2022; Pooladian et al., 2023b), 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

$$\mathcal{S}_{pOT} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 \mu_1 \, \mathrm{d}x - \int s_0 \mu_0 \, \mathrm{d}x \qquad (25)$$
$$- \int_0^1 \int \left(\frac{\partial s_t(x)}{\partial t} + \frac{1}{2}\|\nabla s_t(x)\|^2 + V_t(x) \right) \rho_t \, \mathrm{d}x \, \mathrm{d}t,$$

with the optimality conditions

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

As in Eq. (22), the latter condition implies that the acceleration is given by the gradient of the spatial potential $V_t(x)$. 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 (Schiebinger et al., 2019), 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 (Mikami, 2008; Léonard, 2013; Chen et al., 2021b) is given by

$$\mathcal{S}_{SB} = \inf_{\rho_t, v_t} \int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t \, \mathrm{d}x \, \mathrm{d}t \qquad (26)$$
$$\text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) + \frac{\sigma^2}{2}\Delta\rho_t, \quad \rho_0 = \mu_0, \quad \rho_1 = \mu_1.$$

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

$$\mathcal{U}[\rho_t, t] = -\frac{\sigma^4}{8} \int \big\|\nabla \log \rho_t\big\|^2 \rho_t \, \mathrm{d}x, \qquad (27)$$

which arises from the negative entropy $\mathcal{F}[\rho_t] = \int (\log \rho_t - 1)\ \rho_t \, \mathrm{d}x$ via the W_2 gradient $\nabla \frac{\delta}{\delta \rho_t}\mathcal{F}[\rho_t] = \nabla \log \rho_t$ (Figalli & Glaudo, 2021). 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 (Song et al., 2020). With derivations in App. B.2, the dual objective becomes

$$\mathcal{S}_{SB} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 \, \mathrm{d}x - \int \Phi_0 \mu_0 \, \mathrm{d}x \quad (28)$$
$$- \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\big\|\nabla \Phi_t\big\|^2 + \frac{\sigma^2}{2}\Delta \Phi_t \right) \rho_t \, \mathrm{d}x \, \mathrm{d}t.$$

5. Experiments [3]

We apply our methods for trajectory inference of single-cell RNA sequencing data, including the Embryoid body (**EB**) dataset (Moon et al., 2019), CITE-seq (**Cite**) and Multiome (**Multi**) datasets (Burkhardt et al., 2022), and melanoma treatment dataset of (Bunne et al., 2021; Pariset et al., 2023).

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

[3] The code reproducing the experiments is available at https://github.com/necludov/wl-mechanics

Table 2. Results for train/test splits of 5-dim PCA on EB dataset, with the setting and baseline results taken from Koshizuka & Sato (2022, 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 (Li et al., 2020)	0.69	0.91	0.85	0.81	0.82
TrajectoryNet (Tong et al., 2020)	0.73	1.06	0.90	1.01	0.93
IPF (GP) (Vargas et al., 2021)	0.70	1.04	0.94	0.98	0.92
IPF (NN) (De Bortoli et al., 2021)	0.73	0.89	0.84	0.83	0.82
SB-FBSDE (Chen et al., 2021a)	0.56	0.80	1.00	1.00	0.84
NLSB (Koshizuka & Sato, 2022)	0.68	0.84	0.81	0.79	0.78
OT-CFM (Tong et al., 2023b)	0.78	0.76	0.77	0.75	0.77
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

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) to learn 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 (Tong et al., 2020) 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. (2023a;b) (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 (Tong et al., 2023b), 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 (Flamary et al., 2021). 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 which encode various prior information. WLF-SB (ours) and [SF]^2 M-Geo and -Exact (Tong et al., 2023a) 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 notably improve performance, using either W_2 or WFR kinetic energy.

Table 3. Results in the setting of Pariset et al. (2023, 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 (Chen et al., 2021a)	1.86e-2	6.23
uDSB (no growth) (Pariset et al., 2023)	1.86e-2	6.27
uDSB (w/growth) (Pariset et al., 2023)	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 further SB baselines, we consider the setting of Koshizuka & Sato (2022, 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 $\mu_{t_{i-1}}$. In Table 2, we find that WLF-SB outperforms several SB methods 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. (2023, 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. (2023).

6. Related Work

Wasserstein Hamiltonian Flows Chow et al. (2020) 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. (2020); Wu et al. (2023) emphasize solving the Cauchy problem suggested by the Hamiltonian perspective (see also Taghvaei & Mehta (2019); Wang & Li (2022) for accelerated flows). Our approach recovers the Hamiltonian flow (ρ_t, s_t) in the cotangent bundle at optimality, but does so by solving a variational problem. Finally, Conforti (2019); Gentil et al. (2020); Conforti & Pavon (2018) provide theoretical analysis related to Lagrangian flows on the Wasserstein space.

Flow Matching and Diffusion SB Methods Flow Matching methods (Liu, 2022; Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023; Tong et al., 2023b;a) learn a marginal vector field corresponding to a mixture-of-bridges process parameterized by a coupling and interpolating bridge (Shi et al., 2023). When samples from the endpoint marginals are coupled via an OT plan, Flow Matching solves a dynamical optimal transport problem (Pooladian et al., 2023a). Rectified Flow obtains couplings using ODE simulation with the goal of straight-path trajectories for generative modeling (Liu, 2022; Liu et al., 2022b), which is extended to SDEs in bridge matching methods (Peluchetti, 2022; 2023; Shi et al., 2023). Diffusion Schrödinger Bridge (DSB) methods (De Bortoli et al., 2021; Chen et al., 2021a) 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. (2023). Finally, Liu et al. (2022a; 2023) 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 (Amos et al., 2017) provide an efficient approach to static OT (Makkuva et al., 2020; Korotin et al., 2021; Bunne et al., 2021; 2022) but are limited to the Euclidean cost. Several works extend to other costs using static (Fan et al., 2022; Pooladian et al., 2023b; Uscidda & Cuturi, 2023) or dynamical formulations (Liu et al., 2021; Koshizuka & Sato, 2022). The most general way to define a transport cost is via a Lagrangian action in the state-space (Villani (2009) Ch. 7). While we focus on lifted Lagrangians in the density space, our framework 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. As an application, 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 the observed data, resulting in significant improvement in several benchmarks. We expect our approach can be extended to other natural science domains such as quantum mechanics and social sciences by incorporating new prior information for learning the underlying dynamics.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *International Conference on Learning Representations*, 2022.

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint 2303.08797*, 2023.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

Arnol'd, V. I. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013.

Bauer, M., Bruveris, M., and Michor, P. W. Uniqueness of the fisher–rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Benamou, J.-D., Gallouët, T. O., and Vialard, F.-X. Second-order models for optimal transport and cubic splines on the wasserstein space. *Foundations of Computational Mathematics*, 19:1113–1143, 2019.

Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, pp. 2021–12, 2021.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Burkhardt, D., Bloom, J., Cannoodt, R., Luecken, M. D., Krishnaswamy, S., Lance, C., Pisco, A. O., and Theis, F. J. Multimodal single-cell integration across time, individuals, and batches. In *NeurIPS Competitions*, 2022.

Chen, T., Liu, G.-H., and Theodorou, E. Likelihood training of schrödinger bridge using forward-backward sdes theory. In *International Conference on Learning Representations*, 2021a.

Chen, T., Liu, G.-H., Tao, M., and Theodorou, E. A. Deep momentum multi-marginal Schrödinger bridge. *arXiv preprint arXiv:2303.01751*, 2023.

Chen, Y., Conforti, G., and Georgiou, T. T. Measure-valued spline curves: An optimal transport viewpoint. *SIAM Journal on Mathematical Analysis*, 50(6):5947–5968, 2018.

Chen, Y., Georgiou, T. T., and Pavon, M. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge. *Siam Review*, 63(2):249–313, 2021b.

Chewi, S., Clancy, J., Le Gouic, T., Rigollet, P., Stepaniants, G., and Stromme, A. Fast and smooth interpolation on wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pp. 3061–3069. PMLR, 2021.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018a.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018b.

Chow, S.-N., Li, W., and Zhou, H. Wasserstein hamiltonian flows. *Journal of Differential Equations*, 268(3):1205–1219, 2020.

Clancy, J. and Suarez, F. Wasserstein-fisher-rao splines. *arXiv preprint arXiv:2203.15728*, 2022.

Conforti, G. A second order equation for schrödinger bridges with applications to the hot gas experiment and entropic transportation cost. *Probability Theory and Related Fields*, 174(1):1–47, 2019.

Conforti, G. and Pavon, M. Extremal flows in wasserstein space. *Journal of mathematical physics*, 59(6), 2018.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Fan, J., Liu, S., Ma, S., Chen, Y., and Zhou, H.-M. Scalable computation of monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

Figalli, A. and Glaudo, F. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. 2021.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

Gentil, I., Léonard, C., and Ripani, L. Dynamical aspects of the generalized schrödinger problem via otto calculus–a heuristic point of view. *Revista matemática iberoamericana*, 36(4):1071–1112, 2020.

Hashimoto, T., Gifford, D., and Jaakkola, T. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pp. 2417–2426. PMLR, 2016.

Kondratyev, S., Monsaingeon, L., and Vorotnikov, D. A new optimal transport distance on the space of finite radon measures. *Advances in Differential Equations*, 21(11/12): 1117–1164, 2016.

Korotin, A., Li, L., Genevay, A., Solomon, J. M., Filippov, A., and Burnaev, E. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021.

Koshizuka, T. and Sato, I. Neural lagrangian Schrödinger bridge: Diffusion modeling for population dynamics. In *The Eleventh International Conference on Learning Representations*, 2022.

Lavenant, H., Zhang, S., Kim, Y.-H., and Schiebinger, G. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

Léger, F. and Li, W. Hopf–cole transformation via generalized schrödinger bridge problem. *Journal of Differential Equations*, 274:788–827, 2021.

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

Li, X., Wong, T.-K. L., Chen, R. T., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Liero, M., Mielke, A., and Savaré, G. Optimal transport in competition with reaction: The Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *International Conference on Learning Representations*, 2022.

Liu, G., Lipman, Y., Nickel, M., Karrer, B., Theodorou, E. A., and Chen, R. T. Generalized schrödinger bridge matching, 9 2023.

Liu, G.-H., Chen, T., So, O., and Theodorou, E. Deep generalized schrödinger bridge. *Advances in Neural Information Processing Systems*, 35:9374–9388, 2022a.

Liu, Q. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

Liu, S., Ma, S., Chen, Y., Zha, H., and Zhou, H. Learning high dimensional wasserstein geodesics. *arXiv preprint arXiv:2102.02992*, 2021.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *International Conference on Learning Representations*, 2022b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, Y., Lu, J., and Nolen, J. Accelerating langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

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

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

Neklyudov, K., Brekelmans, R., Severo, D., and Makhzani, A. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, 2023.

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

Pariset, M., Hsieh, Y.-P., Bunne, C., Krause, A., and De Bortoli, V. Unbalanced diffusion Schrödinger bridge. *arXiv preprint arXiv:2306.09099*, 2023.

Peluchetti, S. Non-denoising forward-time diffusions, 2022. URL https://openreview.net/forum?id=oVfIKuhqfC.

Peluchetti, S. Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling. *arXiv preprint arXiv:2304.00917*, 2023.

Pfau, D., Spencer, J. S., Matthews, A. G., and Foulkes, W. M. C. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Physical Review Research*, 2(3):033429, 2020.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. Multisample flow matching: Straightening flows with minibatch couplings. *International Conference on Machine Learning*, 2023a.

Pooladian, A.-A., Domingo-Enrich, C., Chen, R. T., and Amos, B. Neural optimal transport with lagrangian costs. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023b.

Ronneberger, O., Fischer, P., and Brox, T. 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.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

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

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

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. Diffusion Schrödinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Taghvaei, A. and Mehta, P. Accelerated flow for probability distributions. In *International conference on machine learning*, pp. 6076–6085. PMLR, 2019.

Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Tong, A., Malkin, N., Fatras, K., Atanackovic, L., Zhang, Y., Huguet, G., Wolf, G., and Bengio, Y. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023a.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023b.

Uscidda, T. and Cuturi, M. The monge gap: A regularizer to learn all transport maps. *arXiv preprint arXiv:2302.04953*, 2023.

Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wang, Y. and Li, W. Accelerated information gradient flow. *Journal of Scientific Computing*, 90:1–47, 2022.

Wu, H., Liu, S., Ye, X., and Zhou, H. 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. (14) 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*

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

is equivalent to the following dual

$$\begin{aligned} \mathcal{S} = \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} & \int s_1 \mu_1 \, \mathrm{d}x - \int s_0 \mu_0 \, \mathrm{d}x \\ & - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, \mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \end{aligned} \tag{14}$$

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).

The action $\mathcal{A}_{\mathcal{L}}[\rho_t]$ of a given curve is the solution to the inner optimization,

$$\begin{aligned} \mathcal{A}_{\mathcal{L}}[\rho_t] = \sup_{s_t} & \int s_1 \mu_1 \, \mathrm{d}x - \int s_0 \mu_0 \, \mathrm{d}x \\ & - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, \mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t. \end{aligned} \tag{15}$$

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

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{\dot{\rho}_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, \mathrm{d}x - \mathcal{L}[\rho_t, \dot{\rho}_t, t] \tag{29}$$

$$\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \sup_{s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}} \int s_t \dot{\rho}_t \, \mathrm{d}x - \mathcal{H}[\rho_t, s_t, t] \tag{30}$$

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. (29) and integration by parts in time in step (i),

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) &= \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t]\, \mathrm{d}t \quad \text{s.t.} \quad \rho_0 = \mu_0, \qquad \rho_1 = \mu_1 \\ &= \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t]\, \mathrm{d}t \\ &= \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}} \int_0^1 \left(\int s_t \dot{\rho}_t \, \mathrm{d}x - \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \\ &\overset{(i)}{=} \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 \rho_1 \, \mathrm{d}x - \int s_0 \rho_0 \, \mathrm{d}x - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, \mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \\ &\overset{(ii)}{=} \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 \mu_1 \, \mathrm{d}x - \int s_0 \mu_0 \, \mathrm{d}x - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \, \mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \end{aligned} \tag{31}$$

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 $\dot\rho_t$,

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) &= \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot\rho_t, t]\,\mathrm{d}t \quad \text{s.t.} \quad \rho_{t_i} = \mu_{t_i} \;\; (\forall\ 0 \le i \le M-1) \\ &= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \int_0^1 \mathcal{L}[\rho_t, \dot\rho_t, t]\,\mathrm{d}t \end{aligned} \tag{32}$$

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

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) &= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t \in \mathcal{T}^*_{\rho_t}\mathcal{P}} \int_0^1 \left(\int s_t \dot\rho_t \,\mathrm{d}x - \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \\ &= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1 \rho_1 \,\mathrm{d}x - \int s_0 \rho_0 \,\mathrm{d}x - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \,\mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \\ &= \inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1 \mu_1 \,\mathrm{d}x - \int s_0 \mu_0 \,\mathrm{d}x - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \,\mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \end{aligned} \tag{33}$$

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 We confirm that $\mathcal{S}_{\mathcal{L}}(\{\mu_{t_i}\}_{i=0}^{M-1}) = \sum_{i=0}^{M-2} \mathcal{S}_{\mathcal{L}}(\{\mu_{t_i,t_{i+1}}\})$ matches the sum of piecewise optimizations, which is a result of the fact that the action functional $\mathcal{A}_{\mathcal{L}}[\rho_t]$ is simply an integration over time. To confirm this, consider the piecewise dual objective for action-minimization problems between $\{\mu_{0,t_1}\}$ and $\{\mu_{t_1,1}\}$ $(M = 3)$,

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{0,t_1}\}) + \mathcal{S}_{\mathcal{L}}(\{\mu_{t_1,1}\}) = &\inf_{\rho_t \in \Gamma(\{\mu_0, \mu_{t_1}\})} \sup_{s_t} \int s_{t_1} \mu_{t_1} \,\mathrm{d}x - \int s_0 \mu_0 \,\mathrm{d}x + \int_0^{t_1} \left(\int \frac{\partial s_t}{\partial t} \rho_t \,\mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \quad (34) \\ &+ \inf_{\rho_t \in \Gamma(\{\mu_{t_1}, \mu_1\})} \sup_{s_t} \int s_1 \mu_1 \,\mathrm{d}x - \int s_{t_1} \mu_{t_1} \,\mathrm{d}x + \int_{t_1}^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \,\mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \\ = &\inf_{\rho_t \in \Gamma(\{\mu_{t_i}\})} \sup_{s_t} \int s_1 \mu_1 \,\mathrm{d}x - \int s_0 \mu_0 \,\mathrm{d}x - \int_0^1 \left(\int \frac{\partial s_t}{\partial t} \rho_t \,\mathrm{d}x + \mathcal{H}[\rho_t, s_t, t] \right) \mathrm{d}t \quad (35) \end{aligned}$$

After telescoping cancellation and taking the union of the constraints, we obtain the same objective as in Eq. (33).

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. (14), 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

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{\dot\rho_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t \,\mathrm{d}x - \mathcal{L}[\rho_t, \dot\rho_t, t] = \sup_{\dot\rho_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t \,\mathrm{d}x - \mathcal{K}[\rho_t, \dot\rho_t] + \mathcal{U}[\rho_t, t] \tag{36}$$

Since the potential energy does not depend on $\dot\rho_t$, it will not affect any later derivations in this section.

We focus on *kinetic energies* $\mathcal{K}[\rho_t, \dot\rho_t]$ which are linear in the density (see Def. 3.1). We consider two primary examples, the WFR_λ metric $\mathcal{K}[\rho_t, \dot\rho_t]$ 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 (2009) Ch. 7, Ex. B.1 below),

$$\mathcal{K}[\rho_t, \dot\rho_t] = \inf_{v_t, g_t} \int \Big(\underbrace{L(x, v_t) + \frac{\lambda}{2} g_t(x)^2}_{=:K(x, v_t, g_t)}\Big) \rho_t \,\mathrm{d}x, \qquad \text{s.t} \qquad \dot\rho_t = -\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t \tag{37}$$

The kinetic energy above captures three important special cases:

- W_2 kinetic energy for $L(x, v_t) = \frac{1}{2}\|v_t(x)\|^2$ and $\lambda = 0$
- Optimal transport with Lagrangian costs (App. B.1.2)
- Wasserstein Fisher-Rao (WFR_λ) with $L(x, v_t) = \frac{1}{2}\|v_t(x)\|^2$ (App. B.1.1)

We proceed to simplify from Eq. (36) using the kinetic energy in Eq. (37). We let $K(x, v_t, g_t) \coloneqq L(x, v_t) + \frac{\lambda}{2} g_t(x)^2$ for simplicity of notation and to call attention to the fact that these terms do not change in the derivations leading to Eq. (39).

$$\begin{aligned}
\mathcal{H}[\rho_t, s_t, t] &= \sup_{\dot\rho_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t \,\mathrm{d}x - \mathcal{K}[\rho_t, \dot\rho_t] + \mathcal{U}[\rho_t, t] && (38)\\
&= \sup_{\dot\rho_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t \,\mathrm{d}x - \Big(\inf_{(v_t, g_t)} \int K(x, v_t, g_t)\, \rho_t \,\mathrm{d}x\Big) + \mathcal{U}[\rho_t, t] \quad \text{s.t. } \dot\rho_t = -\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t\\
&= \sup_{\dot\rho_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \sup_{(v_t, g_t)} \int s_t \dot\rho_t \,\mathrm{d}x - \int K(x, v_t, g_t)\, \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t] \quad \text{s.t. } \dot\rho_t = -\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t\\
&= \sup_{(v_t, g_t)} \sup_{\dot\rho_t \in \mathcal{T}_{\rho_t}\mathcal{P}} \int s_t \dot\rho_t \,\mathrm{d}x - \int K(x, v_t, g_t)\, \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t] \quad \text{s.t. } \dot\rho_t = -\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t\\
&= \sup_{(v_t, g_t)} \int s_t\big(-\nabla \cdot (\rho_t v_t) + \lambda \rho_t g_t\big) \,\mathrm{d}x - \int K(x, v_t, g_t)\, \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t]
\end{aligned}$$

where, in the fourth line, we can always swap the order of the supremum and, in the last line, we use the form of the constraint to reparameterize the optimization over $\dot\rho_t$ in terms of (v_t, g_t).

Integrating by parts, we have

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{(v_t, g_t)} \int \big(\langle \nabla s_t, v_t \rangle \rho_t + \lambda \rho_t s_t g_t\big) \,\mathrm{d}x - \int K(x, v_t, g_t)\, \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t]. \tag{39}$$

B.1.1. Wasserstein Fisher-Rao and W_2

For $K(x, v_t, g_t) = \frac{1}{2}\|v_t(x)\|^2 + \frac{\lambda}{2} g_t(x)^2$, we proceed from Eq. (39),

$$\mathcal{H}[\rho_t, s_t, t] = \sup_{(v_t, g_t)} \int \Big(\langle \nabla s_t, v_t \rangle \rho_t + \lambda \rho_t s_t g_t\Big) \,\mathrm{d}x - \int \Big(\frac{1}{2}\|v_t\|^2 + \frac{\lambda}{2} g_t^2\Big) \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t] \tag{40}$$

Eliminating v_t and g_t implies

$$v_t = \nabla s_t \qquad\qquad g_t = s_t \tag{41}$$

where $v_t = \nabla s_t$ also holds for the W_2 case with $\lambda = 0$. Substituting into Eq. (40), we obtain a Hamiltonian with a dual kinetic energy $\mathcal{K}^*[\rho_t, \dot\rho_t]$ below that is linear in ρ_t and matches the metric expressed in the cotangent space $\frac{1}{2}\langle s_t, s_t \rangle^{WFR_\lambda}_{T_{\rho_t}}$,

$$\mathcal{H}[\rho_t, s_t, t] = \int \Big(\frac{1}{2}\|\nabla s_t\|^2 + \frac{\lambda}{2} s_t^2\Big) \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t] = \frac{1}{2}\langle s_t, s_t \rangle^{WFR_\lambda}_{T_{\rho_t}} + \mathcal{U}[\rho_t, t]. \tag{42}$$

We make a similar conclusion for the W_2 metric with $\lambda = 0$, where the dual kinetic energy is $\mathcal{K}^*[\rho_t, \dot\rho_t] = \frac{1}{2}\langle s_t, s_t \rangle^{W_2}_{T_{\rho_t}} = \frac{1}{2}\int \|\nabla s_t\|^2 \rho_t \,\mathrm{d}x$. Note, the kinetic energy could also use ground-space Lagrangian costs along with the FR growth term.

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

We first provide background on Lagrangian actions in the ground space and their use to define optimal transport costs, before continuing the above derivations to calculate the density-space Hamiltonian in Eq. (49) - (51) below.

As in Villani (2009) Thm. 7.21, we can consider using the cost associated with action-minimizing curves $\gamma^*(x_0, x_1)$ 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 $\mathcal{K}[\rho_t, \dot{\rho}_t] = \int L(x, v_t)\rho_t \,\mathrm{d}x$ as in Eq. (37). 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 (Arnol'd, 2013). 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 $L(\gamma_t, \dot{\gamma}_t)$ along the curve to define the *action* as the time integral $A(\gamma) = \int_0^1 L(\gamma_t, \dot{\gamma}_t)\,\mathrm{d}t$. Given two endpoints $x_0, x_1 \in \mathcal{X}$, we consider minimizing the action along all curves with the appropriate endpoints $\gamma \in \Pi(x_0, x_1)$,

$$c_L(x_0, x_1) = \inf_{\gamma \in \Pi(x_0,x_1)} A(\gamma) = \inf_{\gamma_t} \int_0^1 L(\gamma_t, \dot{\gamma}_t)\,\mathrm{d}t \quad \text{s.t. } \gamma_0 = x_0,\ \gamma_1 = x_1 \tag{43}$$

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

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

$$H(\gamma_t, p_t) = \sup_{\dot{\gamma}_t} \langle \dot{\gamma}_t, p_t \rangle - L(\gamma_t, \dot{\gamma}_t) \tag{44}$$

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

$$\dot{\gamma}_t = \frac{\partial}{\partial p_t} H(\gamma_t, p_t) \qquad\qquad \dot{p}_t = -\frac{\partial}{\partial \gamma_t} H(\gamma_t, p_t). \tag{45}$$

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 $L(\gamma_t, \dot{\gamma}_t)$ in the ground space $\gamma_t : [0,1] \to \mathcal{X}$, as in Eq. (43) (Villani (2009) 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 $\gamma^*(x_0, x_1)$. For Tonelli Lagrangians, which respect conditions such as convexity in v_t and superlinearity, we have (Schachter, 2017, Def. 3.3.1, Thm. 3.5.1)

$$W_L(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0,\mu_1)} \int\int c_L(x_0, x_1)\pi(x_0, x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1 \quad \text{s.t.} \quad \pi(x_0) = \mu_0(x_0), \quad \pi(x_1) = \mu_1(x_1) \tag{46}$$

$$= \inf_{X:[0,1]\times\mathcal{X}\to\mathcal{X}} \int_0^1 \int L\Big(X(t,x), \dot{X}(t,x)\Big)\mu_0(x)\,\mathrm{d}x\,\mathrm{d}t \quad \text{s.t.} \quad X(1,\cdot)_{\#}\mu_0 = \mu_1 \tag{47}$$

where $X(\cdot, x)$ is a curve for every x. This optimization may also be viewed as (Schachter, 2017, Def. 3.4.1)

$$W_L(\mu_0, \mu_1) = \inf_{\rho_t} \inf_{v_t} \int_0^1 \int L(x, v_t)\rho_t\,\mathrm{d}x\,\mathrm{d}t \quad \text{s.t. } \dot{\rho}_t = -\nabla\cdot(\rho_t v_t),\ \rho_0 = \mu_0,\ \rho_1 = \mu_1. \tag{48}$$

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 $\mathcal{K}[\rho_t, \dot{\rho}_t] = \int L(x, v_t)\rho_t\,\mathrm{d}x$ (see below). Of course, the Benamou-Brenier dynamical formulation of W_2-OT in Eq. (2) may be viewed as a special case with $L(x, v_t) = \frac{1}{2}\|v_t\|^2$.

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

$$\begin{aligned}\mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) &= \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \int_0^1 \mathcal{K}[\rho_t, \dot{\rho}_t] - \mathcal{U}[\rho_t, t]\,\mathrm{d}t \qquad (49)\\ &= \inf_{\rho_t} \int_0^1 \left(\int L(x, v_t)\rho_t \,\mathrm{d}x - \mathcal{U}[\rho_t, t] \right) \mathrm{d}t \quad \text{s.t.} \quad \rho_0 = \mu_0, \qquad \rho_1 = \mu_1\end{aligned}$$

We proceed to derive the Wasserstein Hamiltonian, where the the tangent space is represented using the continuity equation as in Eq. (37), 48).

Continuing from Eq. (39) with $\lambda = 0$ (no growth dynamics), we have

$$\begin{aligned}\mathcal{H}[\rho_t, s_t, t] &= \sup_{v_t} \int \langle \nabla s_t, v_t \rangle \rho_t \,\mathrm{d}x - \mathcal{K}[\rho_t, \dot{\rho}_t] + \mathcal{U}[\rho_t, t].\\ &= \sup_{v_t} \int \langle \nabla s_t, v_t \rangle \rho_t \,\mathrm{d}x - \int L(x, v_t)\rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t]\\ &= \int \left(\sup_{v_t} \langle \nabla s_t, v_t \rangle - L(x, v_t) \right) \rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t] \qquad (50)\end{aligned}$$

which is simply a Legendre transform between velocity and momentum variables in the ground space (Eq. (44)). We can finally write,

$$\mathcal{H}[\rho_t, s_t, t] = \int H(x, \nabla s_t)\rho_t \,\mathrm{d}x + \mathcal{U}[\rho_t, t] \qquad (51)$$

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

We leave empirical exploration of various Lagrangian costs for future work, but note that $H(x, \nabla s_t)$ in Eq. (51) must be known or optimized using Eq. (50) 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

$$S_{SB} = \inf_{\rho_t, v_t} \int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t \,\mathrm{d}x\,\mathrm{d}t \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) - \frac{\sigma^2}{2}\Delta\rho_t \quad \rho_0 = \mu_0,\ \rho_1 = \mu_1. \qquad (52)$$

which we will solve using the following (linear in ρ_t) dual objective from Eq. (28)

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

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

$$\mathcal{U}[\rho_t, t] = -\frac{\sigma^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t \,\mathrm{d}x \qquad (53)$$

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

$$\mathcal{L}[\rho_t, \dot{\rho}_t, t] = \frac{1}{2}\langle \dot{\rho}_t, \dot{\rho}_t \rangle_{T_{\rho_t}}^{W_2} + \frac{\sigma^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t \,\mathrm{d}x. \qquad (54)$$

As in Eq. (38)-(40), 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. Integrating by parts,

$$\begin{aligned}\mathcal{H}[\rho_t, s_t, t] &= \sup_{\dot{\rho}_t} \int s_t \dot{\rho}_t \,\mathrm{d}x - \mathcal{L}[\rho_t, \dot{\rho}_t, t] \qquad (55)\\ &= \sup_{v_t} \int \langle \nabla s_t, v_t \rangle \rho_t \,\mathrm{d}x - \frac{1}{2}\int \|v_t\|^2 \rho_t \,\mathrm{d}x - \frac{\sigma_t^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t \,\mathrm{d}x\end{aligned}$$

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

$$\mathcal{H}[\rho_t, s_t, t] = \frac{1}{2} \int \|\nabla s_t\|^2 \rho_t \,\mathrm{d}x - \frac{\sigma^4}{8} \int \|\nabla \log \rho_t\|^2 \rho_t \,\mathrm{d}x. \tag{56}$$

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 (2021) Eq. 8. As in Thm. 1, the dual for the Wasserstein Lagrangian Flow with the Lagrangian in Eq. (54) involves the Hamiltonian in Eq. (56),

$$\mathcal{S}_\mathcal{L} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{s_t} \int s_1 \mu_1 \,\mathrm{d}x - \int s_0 \mu_0 \,\mathrm{d}x - \int_0^1 \int \Big(\frac{\partial s_t}{\partial t} + \frac{1}{2} \|\nabla s_t\|^2 - \frac{\sigma^4}{8} \int \|\nabla \log \rho_t\|^2 \Big) \rho_t \,\mathrm{d}x \,\mathrm{d}t \tag{57}$$

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*

$$\mathcal{S}_\mathcal{L}(\{\mu_{0,1}\}) = \inf_{\rho_t} \int_0^1 \mathcal{L}[\rho_t, \dot{\rho}_t, t] \,\mathrm{d}t \qquad \textit{s.t.} \quad \rho_0 = \mu_0, \qquad \rho_1 = \mu_1 \tag{58}$$

$$\textit{where } \; \mathcal{K}[\rho_t, \dot{\rho}_t] = \frac{1}{2} \langle \dot{\rho}_t, \dot{\rho}_t \rangle^{W_2}_{T_{\rho_t}}, \qquad \mathcal{U}[\rho_t, t] = -\frac{\sigma^4}{8} \|\nabla \log \rho_t\|^2_{T^{W_2}_{\rho_t}}$$

matches the solution to the SB problem in Eq. (52), $\mathcal{S} = \mathcal{S}_{SB}(\{\mu_{0,1}\}) = \mathcal{S}_\mathcal{L}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\})$ up to a constant $c(\{\mu_{0,1}\})$ wrt ρ_t.

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

$$\mathcal{S} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 \,\mathrm{d}x - \int \Phi_0 \mu_0 \,\mathrm{d}x - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2} \big\|\nabla \Phi_t\big\|^2 + \frac{\sigma^2}{2} \Delta \Phi_t \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t. \tag{59}$$

Thus, we obtain a dual objective for the SB problem, or WLF in Eq. (58), which is linear in ρ_t.

Proof. We consider the following reparameterization (Léger & Li, 2021)

$$s_t = \Phi_t - \frac{\sigma^2}{2} \log \rho_t, \qquad\qquad \nabla s_t = \nabla \Phi_t - \frac{\sigma^2}{2} \nabla \log \rho_t. \tag{60}$$

Note that s_t is the drift for the continuity equation in Eq. (55), $\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_t)$. 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. (57), we perform the reparameterization in Eq. (60), $s_t = \Phi_t - \frac{\sigma^2}{2} \log \rho_t$,

$$\begin{aligned} \mathcal{S}_\mathcal{L} = \inf_{\rho_t \in \Gamma(\mu_0, \mu_1)} \sup_{\Phi_t} & \int \Phi_1 \mu_1 \,\mathrm{d}x - \frac{\sigma^2}{2} \int \log \rho_1 \, \mu_1 \,\mathrm{d}x - \int \Phi_0 \mu_0 \,\mathrm{d}x + \frac{\sigma^2}{2} \int \log \rho_0 \, \mu_0 \,\mathrm{d}x \\ & - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{\partial}{\partial t} \left(\frac{\sigma^2}{2} \log \rho_t \right) + \frac{1}{2} \left\langle \nabla \Phi_t - \frac{\sigma^2}{2} \nabla \log \rho_t, \nabla \Phi_t - \frac{\sigma^2}{2} \nabla \log \rho_t \right\rangle - \frac{\sigma^4}{8} \|\nabla \log \rho_t\|^2 \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t \end{aligned} \tag{61}$$

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

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

where the Hamiltonian now matches Eq. 7 in Léger & Li (2021). Taking $\nabla \log \rho_t = \frac{1}{\rho_t}\nabla \rho_t$ and integrating by parts, the final term becomes

$$\mathcal{S}_{\mathcal{L}} = \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 \,\mathrm{d}x - \frac{\sigma^2}{2}\int \log \rho_1 \; \mu_1 \,\mathrm{d}x - \int \Phi_0 \mu_0 \,\mathrm{d}x + \frac{\sigma^2}{2}\int \log \rho_0 \mu_0 \,\mathrm{d}x \\ - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\big\|\nabla \Phi_t\big\|^2 + \frac{\sigma^2}{2}\Delta \Phi_t \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t$$

Finally, we add terms $c(\{\mu_{0,1}\}) = \frac{\sigma^2}{2}\int (\log \mu_1)\; \mu_1 \,\mathrm{d}x - \frac{\sigma^2}{2}\int (\log \mu_0)\; \mu_0 \,\mathrm{d}x$ which are constant with respect to $\rho_{0,1}$,

$$\begin{aligned} \mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\}) = \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \sup_{\Phi_t} &\int \Phi_1 \mu_1 \,\mathrm{d}x + \frac{\sigma_1^2}{2}\int (\log \mu_1 \; - \log \rho_1)\mu_1 \,\mathrm{d}x \\ &- \int \Phi_0 \mu_0 \,\mathrm{d}x - \frac{\sigma_0^2}{2}\int (\log \mu_0 \; - \log \rho_0)\mu_0 \,\mathrm{d}x \\ &- \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\big\|\nabla \Phi_t\big\|^2 + \frac{\sigma_t^2}{2}\Delta \Phi_t \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t \end{aligned} \tag{62}$$

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

$$\begin{aligned} &\mathcal{S}_{\mathcal{L}}(\{\mu_{0,1}\}) + c(\{\mu_{0,1}\}) \\ &= \inf_{\rho_t \in \Gamma(\mu_0,\mu_1)} \sup_{\Phi_t} \int \Phi_1 \mu_1 \,\mathrm{d}x - \int \Phi_0 \mu_0 \,\mathrm{d}x - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\big\|\nabla \Phi_t\big\|^2 + \frac{\sigma^2}{2}\Delta \Phi_t \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t, \end{aligned} \tag{63}$$

which matches the dual in Eq. (59). We now show that this is also the dual for the SB problem.

Schrödinger Bridge Dual Objective: Consider the optimization in Eq. (52) (here, t may be time-dependent)

$$\mathcal{S}_{SB}(\{\mu_{0,1}\}) = \inf_{\rho_t, v_t} \int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t \,\mathrm{d}x \,\mathrm{d}t \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) + \frac{\sigma_t^2}{2}\nabla \cdot (\rho_t \,\nabla \log \rho_t),\ \rho_0 = \mu_0,\ \ \rho_1 = \mu_1 \tag{64}$$

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,

$$\begin{aligned} \mathcal{S}_{SB}(\{\mu_{0,1}\}) = \inf_{\rho_t, v_t} \sup_{\Phi_t, \lambda_{0,1}} &\int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t \,\mathrm{d}x \,\mathrm{d}t + \int_0^1 \int \Phi_t \Big(\dot{\rho}_t + \nabla \cdot (\rho_t v_t) - \frac{\sigma_t^2}{2}\nabla \cdot (\rho_t \,\nabla \log \rho_t) \Big) \,\mathrm{d}x \,\mathrm{d}t \\ &+ \int \lambda_1 (\rho_1 - \mu_1) \,\mathrm{d}x + \int \lambda_0 (\rho_0 - \mu_0) \,\mathrm{d}x \end{aligned} \tag{65}$$

$$\begin{aligned} = \inf_{\rho_t, v_t} \sup_{\Phi_t, \lambda_{0,1}} &\int_0^1 \int \frac{1}{2}\|v_t\|^2 \rho_t \,\mathrm{d}x dt + \int \Phi_1 \rho_1 \,\mathrm{d}x - \int \Phi_0 \rho_0 \,\mathrm{d}x - \int_0^1 \int \frac{\partial \Phi_t}{\partial t}\rho_t \,\mathrm{d}x \,\mathrm{d}t \\ &- \int_0^1 \int \Big\langle \nabla \Phi_t, v_t - \frac{\sigma_t^2}{2}\nabla \log \rho_t \Big\rangle \rho_t \,\mathrm{d}x \,\mathrm{d}t + \int \lambda_1 (\rho_1 - \mu_1) \,\mathrm{d}x + \int \lambda_0 (\rho_0 - \mu_0) \,\mathrm{d}x \end{aligned} \tag{66}$$

Note that we can freely we can swap the order of the optimizations since the SB optimization in Eq. (64) 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

$$\begin{aligned} \mathcal{S}_{SB}(\{\mu_{0,1}\}) &= \sup_{\Phi_t} \inf_{\rho_t} \int \Phi_1 \mu_1 \,\mathrm{d}x - \int \Phi_0 \mu_0 \,\mathrm{d}x - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\|\Phi_t\|^2 - \frac{\sigma_t^2}{2}\langle \nabla \Phi_t, \nabla \log \rho_t \rangle \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t \\ &= \inf_{\rho_t} \sup_{\Phi_t} \int \Phi_1 \mu_1 \,\mathrm{d}x - \int \Phi_0 \mu_0 \,\mathrm{d}x - \int_0^1 \int \left(\frac{\partial \Phi_t}{\partial t} + \frac{1}{2}\|\Phi_t\|^2 + \frac{\sigma_t^2}{2}\Delta \Phi_t \right) \rho_t \,\mathrm{d}x \,\mathrm{d}t \end{aligned} \tag{67}$$

where we swap the order of optimization again in the second line. This matches the dual in Eq. (62) 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

$$\mathcal{U}[\rho_t, t] = -\frac{\sigma_t^4}{8}\int \|\nabla \log \rho_t\|^2 \rho_t \,\mathrm{d}x + \int \left(\frac{\partial}{\partial t}\frac{\sigma_t^2}{2}\right) \log \rho_t \; \rho_t \,\mathrm{d}x \tag{68}$$

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. (61). In this case,

$$\begin{aligned}
\int \frac{\partial s_t}{\partial t}\rho_t \,\mathrm{d}x &= \int \Big(\frac{\partial \Phi_t}{\partial t} - \frac{\partial}{\partial t}\Big(\frac{\sigma_t^2}{2}\log\rho_t\Big)\Big)\rho_t \,\mathrm{d}x \\
&= \int \left(\frac{\partial \Phi_t}{\partial t} - \Big(\frac{\partial}{\partial t}\frac{\sigma_t^2}{2}\Big)\log\rho_t - \frac{\sigma_t^2}{2}\Big(\frac{\partial}{\partial t}\log\rho_t\Big)\right)\rho_t \,\mathrm{d}x \\
&= \int \left(\frac{\partial \Phi_t}{\partial t} - \Big(\frac{\partial}{\partial t}\frac{\sigma_t^2}{2}\Big)\log\rho_t\right)\rho_t \,\mathrm{d}x
\end{aligned}$$

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

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

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 $\mathcal{U}[\rho_t, t] = \frac{\sigma_t^4}{8}\int \|\nabla \log\rho_t\|^2 \rho_t \,\mathrm{d}x$ compared to Eq. (53) or, in other words, an imaginary weighting $i\sigma_t$ of the gradient norm of the Shannon entropy,

$$\mathcal{L}[\rho_t, \dot\rho_t, t] = \frac{1}{2}\langle\dot\rho_t, \dot\rho_t\rangle^{W_2}_{T_{\rho_t}} - \int \left[\frac{1}{8}\|\nabla\log\rho_t\|^2 + V_t(x)\right]\rho_t \,\mathrm{d}x \tag{69}$$

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

$$\mathcal{S}_{SE} = \sup_{s_t}\inf_{\rho_t} \int s_1\mu_1 \,\mathrm{d}x - \int s_0\mu_0 \,\mathrm{d}x - \int_0^1 \int \left(\frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 + \frac{1}{8}\|\nabla\log\rho_t\|^2 + V_t(x)\right)\rho_t \,\mathrm{d}x\,\mathrm{d}t. \tag{70}$$

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. (70) are

$$\dot\rho_t = -\nabla\cdot(\rho_t\nabla s_t), \quad \frac{\partial s_t}{\partial t} + \frac{1}{2}\|\nabla s_t\|^2 = \frac{1}{8}\|\nabla\log\rho_t\|^2 + \frac{1}{4}\Delta\log\rho_t - V_t(x) \tag{71}$$

Note, that Eq. (71) is the Madelung transform of the Schrödinger equation, i.e. for the equation

$$\frac{\partial}{\partial t}\psi_t(x) = -i\hat{H}\psi_t(x), \quad \text{where} \quad \hat{H} = -\frac{1}{2}\Delta + V_t(x), \tag{72}$$

the wave function $\psi_t(x)$ can be written in terms $\psi_t(x) = \sqrt{\rho_t(x)}\exp(is_t(x))$. Then the real and imaginary part of the Schrödinger equation yield Eq. (71).

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 a (Tonelli) Lagrangian kinetic energy, with an arbitrary potential energy (that is concave in ρ_t),

$$\mathcal{S} = \inf_{\rho_t, v_t} \int_0^1 \int L(x, v_t)\rho_t \, \mathrm{d}x \, \mathrm{d}t - \int_0^1 \mathcal{U}[\rho_t, t] \, \mathrm{d}t \quad \text{s.t. } \dot{\rho}_t = -\nabla \cdot (\rho_t v_t) \quad \rho_0 = \mu_0, \;\; \rho_1 = \mu_1 \tag{73}$$

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

$$\begin{aligned}\mathcal{S} = \inf_{\rho_t, v_t} \sup_{\lambda_{0,1}, s_t} & \int_0^1 \int L(x, v_t)\rho_t \, \mathrm{d}x \, \mathrm{d}t - \int_0^1 \mathcal{U}[\rho_t, t] \, \mathrm{d}t + \int_0^1 \int s_t \dot{\rho}_t \, \mathrm{d}x \, \mathrm{d}t + \int_0^1 \int s_t \; \nabla \cdot (\rho_t v_t) \, \mathrm{d}x \, \mathrm{d}t \\ & + \int \lambda_0(\rho_0 - \mu_0) \, \mathrm{d}x + \int \lambda_1(\rho_1 - \mu_1) \, \mathrm{d}x \end{aligned} \tag{74}$$

$$\begin{aligned}= \inf_{\rho_t, v_t} \sup_{\lambda_{0,1}, s_t} & \int_0^1 \int L(x, v_t)\rho_t \, \mathrm{d}x \, \mathrm{d}t - \int_0^1 \mathcal{U}[\rho_t, t] \, \mathrm{d}t - \int_0^1 \int \frac{\partial s_t}{\partial t} \rho_t \, \mathrm{d}x \, \mathrm{d}t - \int_0^1 \int \langle \nabla s_t, v_t \rangle \rho_t \, \mathrm{d}x \, \mathrm{d}t \\ & + \int s_1 \rho_1 \, \mathrm{d}x - \int s_0 \rho_0 \, \mathrm{d}x + \int \lambda_0 \rho_0 \, \mathrm{d}x - \int \lambda_0 \mu_0 \, \mathrm{d}x + \int \lambda_1 \rho_1 \, \mathrm{d}x - \int \lambda_1 \mu_1 \, \mathrm{d}x \end{aligned} \tag{75}$$

To make further progress by swapping the order of the optimizations, we require that Eq. (75) 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

$$\lambda_0 = s_0, \;\; \lambda_1 = -s_1 \qquad\qquad v_t = \nabla_p H(x, \nabla s_t) \tag{76}$$

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

$$\mathcal{S} = \inf_{\rho_t} \sup_{s_t} \int s_1 \mu_1 \, \mathrm{d}x - \int s_0 \mu_0 \, \mathrm{d}x - \int_0^1 \Big(\int \Big(\frac{\partial s_t}{\partial t} + H(x, \nabla s_t) \Big) \rho_t \, \mathrm{d}x + \mathcal{U}[\rho_t, t] \Big) \, \mathrm{d}t.$$

which is analogous to Eq. (14) in Thm. 1 for the Lagrangian 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. (17) samples from ρ_t.*

Proof. For every absolutely-continuous distributional path ρ_t, we have a unique gradient flow $\nabla s_t^*(x)$ satisfying the continuity equation (Ambrosio et al. (2008) Thm. 8.3.1),

$$\dot{\rho}_t = -\nabla \cdot (\rho_t \nabla s_t^*(x)) \, . \tag{77}$$

Consider the function

$$\varphi_t(x_0, x_1) = \begin{cases} x_0 + \int_0^t \nabla s_\tau^*(x_\tau) \, \mathrm{d}\tau, & t \leq 1/2 \, , \\ x_1 + \int_1^t \nabla s_\tau^*(x_\tau) \, \mathrm{d}\tau, & t > 1/2 \, , \end{cases} \tag{78}$$

which integrates the ODE $\frac{\mathrm{d}x}{\mathrm{d}t} = \nabla s_t^*(x)$ forward starting from x_0 for $t \leq 1/2$, and integrates the same ODE backwards starting from x_1 otherwise.

Clearly, for $t \leq 1/2$ 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. (77). The same applies for $t > 1/2$. Thus, φ_t samples from the correct marginals, i.e.

$$\int\int \delta(x_t - \varphi_t(x_0, x_1))\rho_0(x_0)\rho_1(x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1 = \rho_t(x), \quad \forall t \in [0,1]. \tag{79}$$

We now show that $\varphi_t(x_0, x_1)$ can be expressed using the parameterization in Eq. (17), which constructs x_t as

$$x_t = (1-t)x_0 + tx_1 + t(1-t)\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta), \quad x_0 \sim \mu_0, \quad x_1 \sim \mu_1. \tag{80}$$

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

$$\text{NNET}^*(t, x_0, x_1, \mathbb{1}[t < 0.5]; \eta) = \begin{cases} \frac{1}{1-t}\Big(x_0 - x_1 + \frac{1}{t}\int_0^t \nabla s_\tau^*(x_\tau)\,\mathrm{d}\tau\Big), & t \leq 1/2\,, \\ \frac{1}{t}\Big(x_1 - x_0 + \frac{1}{1-t}\int_1^t \nabla s_\tau^*(x_\tau)\,\mathrm{d}\tau\Big), & t > 1/2\,, \end{cases} \tag{81}$$

we have

$$(1-t)x_0 + tx_1 + t(1-t)\text{NNET}^*(t, x_0, x_1; \mathbb{1}[t < 0.5]; \eta) = \varphi_t(x_0, x_1), \tag{82}$$

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 setups in Tong et al. (2023b;a). The Embryoid body (**EB**) dataset Moon et al. (2019) and the CITE-seq (**Cite**) and Multiome (**Multi**) datasets (Burkhardt et al., 2022) are repurposed and preprocessed by Tong et al. (2023b;a) 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 (Moon et al., 2019). 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 (Burkhardt et al., 2022). 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 (Loshchilov & Hutter, 2017). 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

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

$$\ddot{X}_t = \nabla\left[\frac{\partial s_t(x_t)}{\partial t} + \frac{1}{2}\|\nabla s_t(x_t)\|^2\right] = 0\,, \tag{83}$$

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

$$X_1 = X_0 + \nabla s_0(X_0)\,. \tag{84}$$

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 (Song et al., 2020), i.e. the vector field model $s_t(x,\theta)$ uses the U-net architecture (Ronneberger et al., 2015) with the time embedding and hyperparameters from (Song et al., 2020). 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.

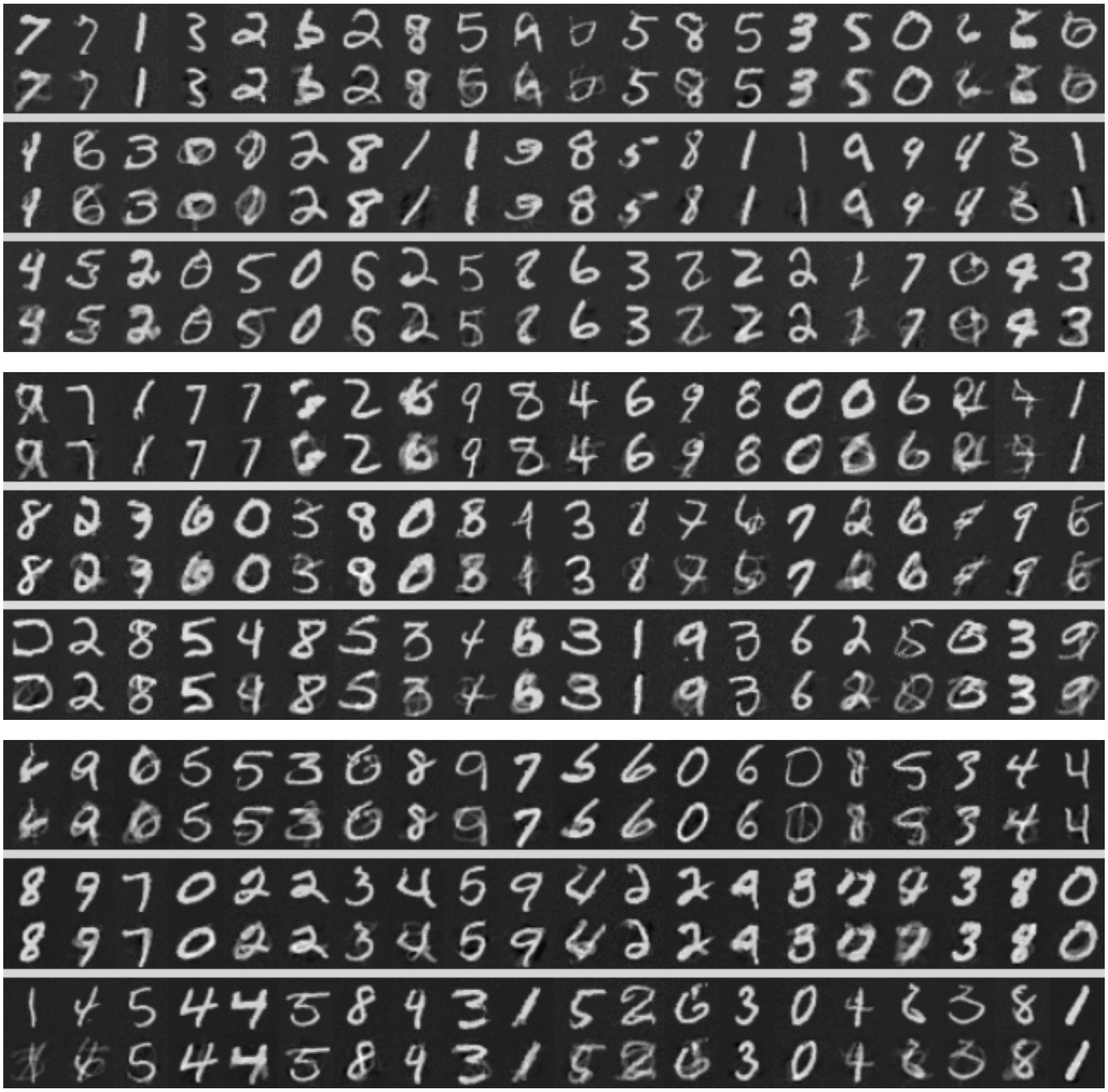


Figure 3. MNIST 32x32 image generation. Every top row is the integration of the corresponding ODE via Dormand-Prince's $5/4$ method, which makes 108 function evaluations. Every bottom row corresponds to single function evaluation approximation.

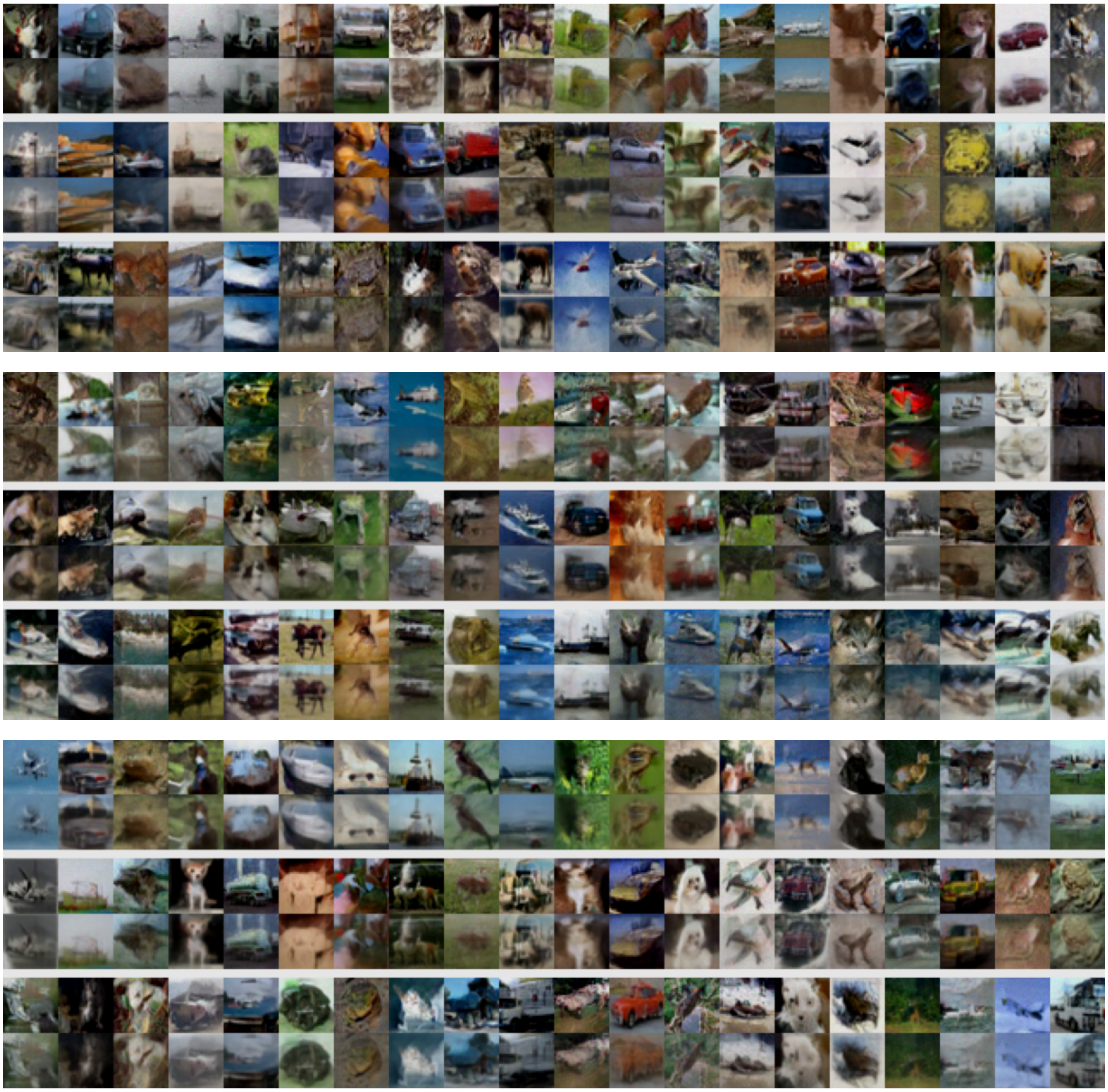


Figure 4. CIFAR-10 image generation. Every top row is the integration of the corresponding ODE via Dormand-Prince's $5/4$ method, which makes 78 function evaluations. Every bottom row corresponds to single function evaluation approximation.