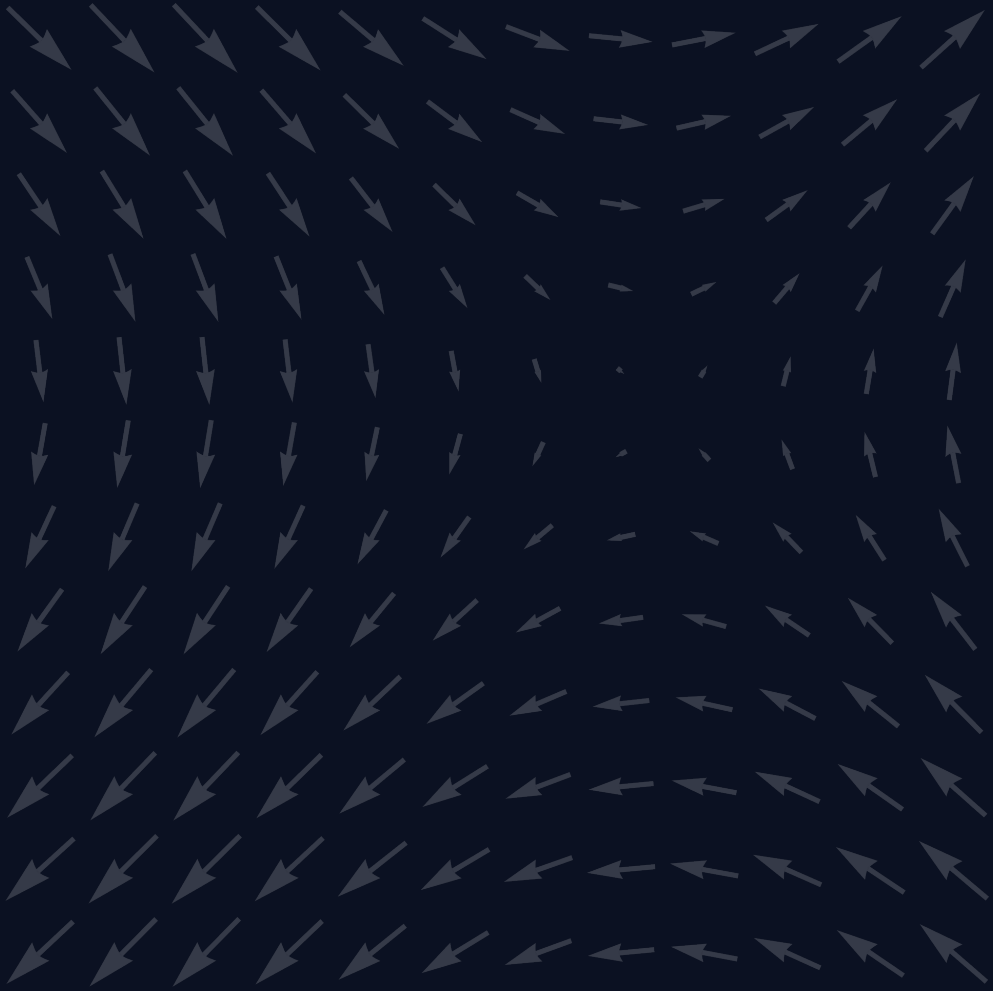
Neural Optimal Transport for Dynamical Systems

Methods and Applications in Biomedicine

Charlotte Bunne



Diss.-No. ETH 29XXX

CHARLOTTE BUNNE

NEURAL OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

METHODS AND APPLICATIONS IN BIOMEDICINE

DISS. ETH NO. ?

NEURAL OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

METHODS AND APPLICATIONS IN BIOMEDICINE

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

CHARLOTTE BUNNE
M. sc. ETH Zurich

born on 29 August 1995

accepted on the recommendation of

Prof. Dr. Andreas Krause, examiner
Prof. Dr. Marco Cuturi, co-examiner
Prof. Dr. Lucas Pelkmans, co-examiner
Prof. Dr. Jure Leskovec, co-examiner

2023

Charlotte Bunne: *Neural Optimal Transport for Dynamical Systems,* © 2023

DOI: 10.3929/ethz-a-

Und was in schwankender Erscheinung schwebt,
Befestiget mit dauernden Gedanken.
— Johann Wolfgang von Goethe, *Faust I* (1808)

ABSTRACT

English abstract here.

ZUSAMMENFASSUNG

Deutsche Zusammenfassung hier.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors Andreas Krause and Marco Cuturi. ... I am very grateful to Anne Carpenter and Shantanu Singh, my mentors at the Broad Institute of MIT and Harvard. ... I will be forever grateful for this. Further, I would also like to thank ... and Jure Leskovec for being on my thesis committee, and for providing me with invaluable feedback. Aviv, in particular, ...

The thesis would not be the same without the longstanding collaboration with Lucas Pelkmans and Gabriele Gut. ...

I am profoundly indebted to other co-authors with whom I worked on many projects throughout the past years, in particular, Gunnar Rätsch, Stefan Stark, Ya-Ping Hsieh, Vignesh Ram Somnath, Frederike Lübeck, Matteo Pariset, Valentin De Bortoli, Octavian Ganea, Laetitia Meng-Papaxanthos, and Philippe Schwaller. Thank you especially to Ya-Ping Hsieh and Stefan Stark for hour-long discussions on various projects. ...

Further, I am very grateful for my first academic mentors, Stefanie Jegelka, David Alvarez Melis, Roland Eils, Thomas Höfer, and Lisa Buchauer. ... I would not have started my scientific journey without the Life-Science Lab of the German Cancer Research Center. It is what sparked my interest in biology and engineering. It ...

Thanks to all members of the Learning and Adaptive Systems group for creating an excellent research environment. In particular, my office mates Parnian Kassraie, Lars Lorch, and Jonas Rothfuss. Thank you, Rita Klute, for entangling the jungle of bureaucracy and never getting tired of my administrative requests.

I am very grateful to my parents Nele and Egon and siblings Kaspar, Henriette, and Frieder for there endless love, advice, and support. Lastly, I always wondered ... Pol, I cannot put in words what you mean to me.

CONTENTS

NOTATION

Σ_d	Probability simplex of size d

ACRONYMS

BDT Black-Derman-Toy.
BM Brownian motion.

DDPM denoising diffusion probabilistic model.

GSB Gaussian Schrödinger bridge.

OT optimal transport.
OU Ornstein-Uhlenbeck.

SB Schrödinger bridge.
SDE stochastic differential equation.
SMLD score matching with Langevin dynamics.

VESDE variance exploding SDE.
VPSDE variance preserving SDE.

1

INTRODUCTION

...

— ..., ... (...)

(Bunne et al., 2021) optimal transport (OT)

OPTIMAL TRANSPORT. For two probability measures μ, ν in $\mathcal{P}(\mathbb{R}^d)$, their squared 2-Wasserstein distance is

$$W_2^2(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \iint \|x-y\|_2^2 \gamma(dx,dy), \tag{1.1}$$

where $\Gamma(\mu,\nu)$ is the set of couplings on $\mathbb{R}^d \times \mathbb{R}^d$ with respective marginals μ, ν. When instantiated on finite discrete measures, such as $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, this problem translates to a linear program, which can be regularized using an entropy term (Cuturi, 2013; Peyré and Cuturi, 2019). For $\varepsilon \geq 0$, set

$$W_\varepsilon(\mu,\nu) := \min_{\mathbf{P}\in U(a,b)} \langle \mathbf{P}, [\|x_i - y_j\|^2]_{ij} \rangle - \varepsilon H(\mathbf{P}), \tag{1.2}$$

where $H(\mathbf{P}) := -\sum_{ij} \mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1)$ and the polytope $U(a,b)$ is the set of $n \times m$ matrices $\{\mathbf{P} \in \mathbb{R}_+^{n\times m}, \mathbf{P}\mathbf{1}_m = a, \mathbf{P}^\top \mathbf{1}_n = b\}$. Notice that the definition above reduces to the usual (squared) 2-Wasserstein distance when $\varepsilon = 0$. Setting $\varepsilon > 0$ yields a faster and differentiable proxy to approximate W_0, but introduces a bias, since $W_\varepsilon(\mu,\mu) \neq 0$ in general. In the rest of this work, we therefore use the *Sinkhorn divergence* (Ramdas et al., 2017; **?**; Salimans et al., 2018; Feydy et al., 2019) as a valid non-negative discrepancy,

$$\overline{W}_\varepsilon(\mu,\nu) := W_\varepsilon(\mu,\nu) - \frac{1}{2}\left(W_\varepsilon(\mu,\mu) + W_\varepsilon(\nu,\nu)\right). \tag{1.3}$$

OT AND CONVEXITY. An alternative formulation for OT is given by the Monge (1781) problem

$$W_2^2(\mu,\nu) = \inf_{T:T_\#\mu=\nu} \int_{\mathcal{X}} ||x - T(x)||^2 d\mu(x) \tag{1.4}$$

where # is the push-forward operator, and the optimal solution $T^\star$ is known as the Monge map between μ and ν. The Brenier theorem 1987 states that if μ has a density, the Monge map $T^\star$ between μ and ν can be recovered as the gradient of a unique (up to constants) convex function ψ whose gradient pushes forward μ to ν. Namely, if $\psi : \mathbb{R}^d \to \mathbb{R}$ is convex and $(\nabla\psi)_{\#}\mu = \nu$, then $T^\star(x) = \nabla\psi(x)$ and

$$W_2^2(\mu,\nu) = \int_{\mathcal{X}} ||x - \nabla\psi(x)||^2 d\mu(x)\,. \tag{1.5}$$

CONVEX NEURAL ARCHITECTURES. Input convex neural networks are neural networks $\psi_\theta(x)$ with specific constraints on the architecture and parameters θ, such that their output is a convex function of some (or all) elements of the input x (Amos et al., 2017). We consider in this work *fully* input convex neural networks (ICNNs), such that the output is a convex function of the entire input x. A typical ICNN is a L-layer, fully connected network such that, for $l = 0, \ldots, L-1$:

$$z_{l+1} = a_l(W_l^x x + W_l^z z_l + b_l) \text{ and } \psi_\theta(x) = z_L, \tag{1.6}$$

where by convention, z_0 and W_0^z are 0, a_l are convex non-decreasing (non-linear) activation functions, $\theta = \{b_l, W_l^z, W_l^x\}_{l=0}^{L-1}$ are the weights and biases of the neural network, with weight matrices W_l^z associated to latent representations z that have non-negative entries. Since Amos et al. (2017)'s work, convex neural architectures have been further extended and shown to capture relevant models despite these constraints (Amos et al., 2017; **?**; Huang et al., 2021). In particular, Chen et al. (2019) provide a theoretical analysis that any convex function over a convex domain can be approximated in sup norm by an ICNN.

2

OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

...

— ..., ... (...)

3

DYNAMICAL PROCESSES IN BIOMEDICINE

The results suggest a helical structure (which must be very closely packed) containing probably 2, 3 or 4 coaxial nucleic acid chains per helical unit and having the phosphate groups near the outside.

— Rosalind Franklin, *Report* (1952)

Part I

STATIC NEURAL OPTIMAL TRANSPORT

4

...

Tout va par degré dans la nature, et rien par saut, et cette regle à l'égard des changements est une partie de ma loi de la continuité.
— Gottfried Wilhelm Leibniz, *Nouveaux essais sur l'entendement humain* (1765)

5

...

...

— ..., ... (...)

Part II

DYNAMIC NEURAL OPTIMAL TRANSPORT

6

...

Living matter evades the decay to equilibrium.
— Erwin Schrödinger, *What is Life?* (1944)

The Schrödinger bridge (SB) (**?**Chen et al., 2021), alternatively known as the *dynamic* entropy-regularized optimal transport (OT), has recently received significant attention from the machine learning community. In contrast to the classical *static* OT where one seeks a coupling between measures that minimizes the average cost (Villani, 2009; **?**), the goal of SBs is to find the optimal *stochastic processes* that evolve a given measure into another. As such, SBs are particularly suitable for learning complex continuous-time systems, and have been successfully applied to a wide range of applications such as sampling (**??**), generative modeling (Chen et al., 2022; De Bortoli et al., 2021; **?**), molecular biology (Holdijk et al., 2022), and mean-field games (Liu et al., 2022).

Despite of these impressive achievements, a common limitation of the existing works is that the SBs are typically solved in a purely numerical fashion. In sharp contrast, it is well-known that many important OT problems for *Gaussian* measures admit *closed-form* solutions, and the advantages of such solutions are numerous: they have inspired new learning methods (**???**), they can serve as the ground truth for evaluating numerical schemes (**?**), and they have lead to the discovery of a new geometry that is both rich in theory and application (**?**).

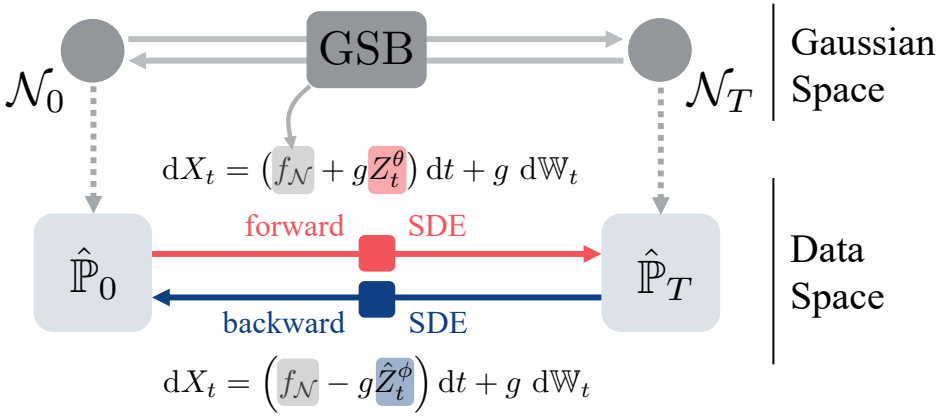


FIGURE 6.1: Solving the SB problem between $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ is notoriously difficult because it requires learning the time-dependent drifts of two SDEs that respect the desired marginals, and a random initialization for these drifts is usually extremely far from satisfying that constraint. We propose a data-dependent procedure that relies first on Gaussian approximations of the data measures, which provide a closed-form drift $f_{\mathcal{N}}$ in (6.29) (the GSB). We show that this facilitates the training of forward/backward drifts $\hat{Z}_t^\theta, \hat{Z}_t^\phi$.

The goal of our paper is to continue this pursuit of closed-form solutions and thereby extending these advantages to SB-based learning methods. For an overview of the method, see Fig. 8.1. To this end, we make the following **contributions**:

1. As our central result, we derive the closed-form expressions for Gaussian Schrödinger bridges (GSBs), i.e.,SBs between Gaussian measures. This is a challenging task for which all existing techniques fail, and thus we need to resort to a number of new ideas from entropic OT, Riemannian geometry, and generator theory; see **??**.
2. We extend the deep connection between geometry and Gaussian OT to Gaussian Schrödinger bridges. In particular, our results can be seen as a vast generalization of the classical Bures-Wasserstein geodesics between Gaussian measures (**??**), which is the foundation of many computational methods (**???**).
3. Via a simple Gaussian approximation on real *single-cell genomics* data, we numerically demonstrate that many benefits of the closed-form expressions in static OT immediately carry over to SB-based learning methods: We report improved numerical stability and tuning insensitivity when trained on benchmark datasets, which ultimately lead to an overall better performance.

6.1 PRELIMINARIES ON GAUSSIAN OPTIMAL TRANSPORT PROBLEMS

Throughout this paper, let $\xi \sim \mathcal{N}(\mu, \Sigma)$ and $\xi' \sim \mathcal{N}(\mu', \Sigma')$ denote two given Gaussian random variables. By abusing the notation, we will continue to denote the measures of these Gaussians by ξ and ξ', respectively. We will also denote by $\Pi(\xi, \xi')$ the set of all their couplings.

6.1.1 *Static Gaussian Optimal Transport*

The *static* entropy-regularized OT between Gaussians refers to the following minimization problem (**?**):

$$\min_{\pi \in \Pi(\xi, \xi')} \int \|x - x'\|^2 \mathrm{d}\pi(x, x') + 2\sigma^2 D_{\mathrm{KL}} \pi \xi \otimes \xi', \tag{6.1}$$

where $\xi \otimes \xi'$ denotes the product measure of ξ and ξ', and $\sigma \geq 0$ is a regularization parameter. When $\sigma = 0$, (6.1) reduces to the classical 2-Wasserstein distance between ξ and ξ' (Villani, 2009), whose closed-form

solution is classical (Dowson and Landau, 1982; **?**). The case for general σ is more involved, and an analytical expression was only recently found (**????**): Setting

$$D_\sigma := (4\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}} + \sigma^4 I)^{\frac{1}{2}}, \quad C_\sigma := \frac{1}{2}(\Sigma^{\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 I), \tag{6.2}$$

then the solution $\pi^\star$ to (6.1) is itself a Gaussian:

$$\pi^\star \sim \mathcal{N}\left(\begin{bmatrix}\mu \\ \mu'\end{bmatrix}, \begin{bmatrix}\Sigma & C_\sigma \\ C_\sigma^\top & \Sigma'\end{bmatrix}\right). \tag{6.3}$$

6.1.2 *Dynamic Gaussian Optimal Transport*

In the literature, (6.1) is commonly referred to as the *static* OT formulation, since it merely asks *where* the mass should be transported to (i.e., $\pi(x, x')$ dictates how much mass at x should be transported to x'). In contrast, the more general problem of *dynamic* Gaussian OT seeks to answer *how* the mass the should be transported:

$$\min_{\rho_0=\xi,\ \rho_1=\xi'} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|v_t\|^2 + \frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2 \,\mathrm{d}t\right]. \tag{6.4}$$

Here, the minimization is taken over all pairs (ρ_t, v_t) where ρ_t is an absolutely continuous curve of measures (Ambrosio et al., 2006), and $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is such that the continuity equation holds:

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t v_t), \tag{6.5}$$

where $(\nabla_x \cdot v_t)(x) := \sum_{i=1}^d \frac{\partial}{\partial x_i} v_t^i(x)$ denotes the divergence operator with respect to the x variable. It can be shown that, if $\rho_t^\star$ is the optimal curve for (6.4), then the joint distribution of the end marginals $(\rho_0^\star, \rho_1^\star)$ coincides with (6.3), hence the interpretation of $\rho_t^\star$ as the optimal *trajectory* in the space of measures (**??**Chen et al., 2021; **?**).

To our knowledge, the only work that has partially addressed the closed-form solution of (6.4) is **?**, whose results are nonetheless insufficient to cover important applications such as generative modeling. In **??**, we will derive a vast generalization of the results in **?** and provide a detailed comparison in **??**–**??**.

6.2 THE GAUSSIAN SCHRÖDINGER BRIDGE PROBLEM AND ANALYSIS OVERVIEW

The purpose of this section is to introduce the core objectives in our paper, the Gaussian Schrödinger bridges, and establish their connection to the Gaussian OT problems in **??**. To help the reader navigate our somewhat technical proofs in **??**–**??**, we illustrate in section 6.2.2 the high-level challenges as well as our new techniques for solving Gaussian Schrödinger bridges.

6.2.1 *Schrödinger Bridges as Dynamic Entropy-Regularized Optimal Transport*

Let ν, ν' be two given measures and let $\mathbb{Q}_t$ be an arbitrary stochastic process. In its most generic form, the Schrödinger bridge refers to the following constrained KL-minimization problem over all stochastic processes $\mathbb{P}_t$ (**?**Chen et al., 2021):

$$\min_{\mathbb{P}_0=\nu,\, \mathbb{P}_1=\nu'} D_{\mathrm{KL}}\mathbb{P}_t\mathbb{Q}_t. \tag{6.6}$$

In practice, ν and ν' typically arise as the (empirical) *marginal* distributions of a complicated continuous-time dynamics observed at the starting and end times, and $\mathbb{Q}_t$ is a "prior process" representing our belief of the dynamics before observing any data. The solution $\mathbb{P}_t^\star$ to (6.6) is thus interpreted as the best dynamics that conforms to the prior belief $\mathbb{Q}_t$ while respecting the data marginals ($\mathbb{P}_0^\star = \nu, \mathbb{P}_1^\star = \nu'$).

In this paper, we will consider a general class of $\mathbb{Q}_t$'s that includes most existing processes in the machine learning applications of SBs. Specifically, with some initial condition Y_0, we will take $\mathbb{Q}_t$ to be the measure of the linear stochastic differential equation (SDE):

$$\mathrm{d}Y_t = (c_t Y_t + \alpha_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t := f_t\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t. \tag{6.7}$$

Here, $c_t : \mathbb{R}^+ \to \mathbb{R}$, $\alpha_t : \mathbb{R}^+ \to \mathbb{R}^d$, and $g_t : \mathbb{R}^+ \to \mathbb{R}^+$ are smooth functions. In this case, SBs can be seen as generalized dynamical OT between two (not necessarily Gaussian) measures:

Theorem 1 *Consider the Schrödinger bridge problem with Y_t as the reference process:*

$$\min_{\mathbb{P}_0=\nu,\, \mathbb{P}_1=\nu'} D_{\mathrm{KL}}\mathbb{P}_t Y_t. \tag{6.8}$$

Then (6.8) is equivalent to

$$\inf_{(\rho_t, v_t)} \mathbb{E}\left[\int_0^1 \frac{\|v_t\|^2}{2g_t^2} + \frac{g_t^2}{8}\|\nabla \log \rho_t\|^2 - \frac{1}{2}\langle f_t, \nabla \log \rho_t\rangle \, \mathrm{d}t\right] \tag{6.9}$$

where the infimum is taken all pairs (ρ_t, v_t) *such that* $\rho_0 = \nu, \rho_1 = \nu'$, ρ_t *absolutely continuous, and*

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + v_t)). \tag{6.10}$$

The proof of theorem 1, which we defer to **??**, is a straightforward extension of the argument in (**???**) which establishes the equivalence when Y_t is a reversible Brownian motion, i.e., $f_t \equiv 0, g_t \equiv \sigma$, and Y_0 follows the Lebesgue measure.[1]

6.2.2 *The Gaussian Schrödinger Bridge Problem*

The central goal of our paper is to derive the closed-form solution of SBs when the marginal constraints are Gaussians $\xi \sim \mathcal{N}(\mu, \Sigma)$, $\xi' \sim \mathcal{N}(\mu', \Sigma')$. Namely, we are interested in the following class of the SBs, termed Gaussian Schrödinger bridges:

$$\min_{\mathbb{P}_0 = \xi,\, \mathbb{P}_1 = \xi'} D_{\mathrm{KL}} \mathbb{P}_t Y_t. \tag{GSB}$$

To emphasize the dependence on the reference SDE, we will sometimes call (GSB) the Y_t-GSB.

Technical challenges; related work. In order to analyze (GSB), we first notice that the objective in (6.9) becomes $\sigma^{-2}\,\mathbb{E}\left[\int_0^1 \frac{1}{2}\|v_t\|^2 + \frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2 \, \mathrm{d}t\right]$ for $\sigma \mathbb{W}_t$-GSBs. Up to a constant factor, this is simply (6.4), so theorem 1 reduces to the well-known fact that $\sigma \mathbb{W}_t$-GSBs are a reformulation of the dynamic Gaussian OT (**???**).

At first sight, this might suggest that one can extend existing tools in Gaussian OT to analyze GSBs. Unfortunately, the major difficulty of tackling GSBs is that these existing tools are fundamentally insufficient for the generalized objective (6.9). To be more precise, there exist three prominent frameworks for studying Gaussian OT problems:

- **Convex analysis:** An extremely fruitful observation in the field is that many Gaussian OT instances can be reduced to a *convex* program, for

1 The reversible Brownian motion is a technical construct to simplify the computations. For our purpose, one can think of $Y_0 \sim \xi$ instead of the Lebesgue measure, and our results still hold verbatim.

which one can import various convex techniques such as KKT or fixed-point arguments. This is the case for static Gaussian OT (6.1), both when $\sigma = 0$ (Dowson and Landau, 1982; **?**; **?**) and $\sigma > 0$ (**?**). Furthermore, in the case of $\sigma = 0$, the solution to the dynamic formulation (6.4) can be recovered from the static one via a simple linear interpolation (McCann, 1997).

- **Ad hoc computations:** When $\sigma > 0$ in (6.4), the problem is no longer reducible to a convex program (**?**Chen et al., 2021). In this case, the only technique we are aware of is the ad hoc approach of (**?**), which manages to find a closed form for (6.4) (and thus $\sigma\mathbb{W}_t$-GSBs) through a series of brute-force computations.
- **Control theory:** On a related note, in a series of papers, **??**Chen et al. (2019) exploit the deep connection between $\sigma\mathbb{W}_t$-GSBs and control theory to study the *existence* and *uniqueness* of the solutions. Although a variety of new optimality conditions are derived in these works, they are all expressed in terms differential equations with coupled initial conditions, and it is unclear whether solving these differential equations is an easier task than (GSB) itself. In particular, no closed-form, even for $\sigma\mathbb{W}_t$-GSBs, can be found therein.

By theorem 1, GSBs are more general than (6.4) and thus irreducible to convex programs, so there is no hope for the convex route. As for ad hoc computations, the time-dependent f_t and g_t terms in (6.9) present a serious obstruction for generalizing the approach of **?** to Y_t-GSBs when $f_t \neq 0$ or g_t is not constant; this is exemplified by the convoluted expressions in our theorem 3, which hopefully will convince the reader that they are beyond any ad hoc guess. Finally, the control-theoretic view has so far fallen short of producing closed-form solutions even for $\sigma\mathbb{W}_t$-GSBs, so it is essentially irrelevant for our purpose.

To conclude, in order to find an analytic expression for general GSBs, we will need drastically different techniques.

OUR APPROACH To overcome the aforementioned challenges, in **??**, we will first develop a principled framework for analyzing the closed-form expressions of $\sigma\mathbb{W}_t$-GSBs, i.e., (6.4). Unlike the ad hoc approach of **?** which is very specific to Brownian motions, our analysis reveals the general role played by the *Lyapunov operator* (see (6.14)) on covariance matrices, thereby essentially reducing the solutions of GSBs to solving a matrix equation. This route is enabled via yet another equivalent formulation of (6.4), namely the action minimization problem on the *Bures-Wasserstein geometry*, which has

recently emerged as a rich source for inspiring new computational methods (**???**). In **??**, we show how the insight gained from our geometric framework in **??** can be easily adapted to GSBs with general reference processes, which ultimately leads to the full resolution of (GSB).

6.3 The Bures-Wasserstein geometry of $\sigma\mathbb{W}_t$-Gaussian Schrödinger bridges

This section illustrates the simple geometric intuition that underlies the somewhat technical proof of our main result (cf. theorem 3). After briefly reviewing the action minimization problems on Euclidean spaces in section 6.3.1, we present the main observation in section 6.3.2: $\sigma\mathbb{W}_t$-GSBs are but action minimization problems on the Bures-Wasserstein manifolds, which can be tackled by following a standard routine in physics.

6.3.1 *A Brief Review on Action Minimization Problems*

Consider the following *action minimization* problem with fixed endpoints $x, x' \in \mathbb{R}^d$:

$$\min_{x(0)=x, x(1)=x'} \int_0^1 \frac{1}{2}\|\dot{x}(t)\|^2 - U(x(t))\,\mathrm{d}t, \tag{6.11}$$

where the minimum is taken over all piecewise smooth curves. A celebrated result in physics asserts that the optimal curve for (6.11) satisfies the *Euler-Lagrange* equation:

$$\ddot{x}(t) = -\nabla U(x(t)), \quad x(0) = x, \quad x(1) = x'. \tag{6.12}$$

In particular, when $U \equiv 0$, (6.12) reduces to $\ddot{x} \equiv 0$, i.e., $x(t)$ is a straight line connecting x and x'.

More generally, one can consider (6.11) on any *Riemannian manifold*, provided that the Euclidean norm $\|\cdot\|$ in (6.11) is replaced by the corresponding Riemannian norm. In this case, the Euler-Lagrange equation (6.12) still holds, with $\ddot{x}$ and ∇U replaced with their Riemannian counterparts (Villani, 2009).

6.3.2 *$\sigma\mathbb{W}_t$-GSBs as Action Minimization Problems*

We begin with the following simple observation. Based on the seminal work by **?**, **?** show that SBs between two arbitrary measures can be formally

understood as an action minimization problem of the form (6.11) on an *infinite*-dimensional manifold. Since we have restricted the measures in (GSB) to be Gaussian, and since Gaussian measures are uniquely determined by their means and covariances, **?** strongly suggests a *finite*-dimensional geometric interpretation of $\sigma\mathbb{W}_t$-GSBs. The main result in this section, theorem 2 below, makes this link precise.

The proper geometry we need is the *Bures-Wasserstein manifold* (**??**) defined as follows. Consider the space of covariance matrices (i.e., symmetric positive definite matrices) of dimension d, which we denote by $\mathbb{S}^d_{++}$, and consider its natural tangent space as the space of symmetric matrices:

$$\mathcal{T}_\Sigma \mathbb{S}^d_{++} := \{U \in \mathbb{R}^{d\times d} : U^\top = U\}. \tag{6.13}$$

A notion that will play a pivotal role is the so-called *Lyapunov operator*: For any $\Sigma \in \mathbb{S}^d_{++}$ and $U \in \mathcal{T}_\Sigma \mathbb{S}^d_{++}$, we define $\mathcal{L}_\Sigma[U]$ to be the symmetric solution to the equation

$$A: \quad \Sigma A + A\Sigma = U. \tag{6.14}$$

It is shown in **?** that the Lyapunov operator defines a geometry on $\mathbb{S}^d_{++}$, known as the *Bures-Wasserstein geometry*: For any two tangent vectors $U, V \in \mathcal{T}_\Sigma \mathbb{S}^d_{++}$, the operation

$$\langle U, V\rangle_\Sigma := \frac{1}{2} \operatorname{tr} \mathcal{L}_\Sigma[U] V \tag{6.15}$$

satisfies all the axioms of the Riemannian metric; additional background on the Bures-Wasserstein geometry can be found in **??**.

We are now ready to state the main result of the section. Let $\|\cdot\|_\Sigma$ be the induced norm of $\langle\cdot,\cdot\rangle_\Sigma$. Fix $\sigma > 0$ and let $\mathbb{W}_t$ be a reversible Brownian motion. Consider the following special case of (GSB):

$$\min_{\mathbb{P}_0=\mathcal{N}(0,\Sigma),\, \mathbb{P}_1=\mathcal{N}(0,\Sigma')} D_{\mathrm{KL}} \mathbb{P}_t \sigma\mathbb{W}_t. \tag{6.16}$$

Then we have:

Theorem 2 *The minimizer of* (6.16) *(and hence* (6.4)*) coincides with the solution of the action minimization problem:*

$$\min_{[0]=\Sigma,\Sigma_1=\Sigma'} \int_0^1 \frac{1}{2} \|\dot\Sigma_t\|^2 - \mathcal{U}() \, \mathrm{d}t \tag{6.17}$$

where $\mathcal{U}() := -\frac{\sigma^4}{8}\operatorname{tr}\Sigma_t^{-1}$ *and the minimum is taken over all piecewise smooth curves in* $\mathbb{S}^d_{++}$*. In particular, the minimizer of* (6.16) *solves the Euler-Lagrange equation in the Bures-Wasserstein geometry:*

$$\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t = -\operatorname{grad}\mathcal{U}(), \quad [0] = \Sigma, \quad [1] = \Sigma', \tag{6.18}$$

where $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ *denotes the Riemannian acceleration and* grad *the Riemannian gradient in the Bures-Wasserstein sense.*

AN IMPORTANT IMPLICATION As alluded to in **??**, the solution curve to (6.4) or (6.16) is not new; it is derived in **?** via a strenuous and rather unenlightening calculation:

$$\Sigma_t := \bar{t}^2\Sigma + t^2\Sigma' + t\cdot\bar{t}\Big(C_\sigma + C_\sigma^\top + \sigma^2 I\Big). \tag{6.19}$$

Here, $\bar{t} := 1 - t$ and C_σ is defined in (6.2). However, the interpretation of (6.19) as the minimizer of (6.17) is new and suggests a principled avenue towards the closed-form solution of $\sigma\mathbb{W}_t$-GSBs: solve the Euler-Lagrange equation (6.18). Inspecting the formulas for $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ and $\operatorname{grad}\mathcal{U}()$ (see (**??**) and (**??**)), one can further reduce (6.18) to computing the Lyapunov operator $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]$, which presents the bottleneck in the proof of theorem 2 as there is, in general, no closed form for the matrix equation (6.14). To this end, our main contribution is the following technical Lemma:

Lemma 1 *Define the matrix* $\tilde{S}_t$ *to be:*

$$\tilde{S}_t := t\Sigma' + \bar{t}C_\sigma - \bar{t}\Sigma - tC_\sigma^\top + \frac{\sigma^2}{2}(\bar{t} - t)I. \tag{6.20}$$

Then $\tilde{S}_t^\top\Sigma_t^{-1}$ *is symmetric.*

Armed with lemma 1, it is straightforward to verify that $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] = \tilde{S}_t^\top\Sigma_t^{-1}$, i.e., $\tilde{S}_t^\top\Sigma_t^{-1}$ is symmetric and satisfies:

$$\tilde{S}_t^\top\Sigma_t^{-1}\cdot\Sigma_t^{-1} + \Sigma_t^{-1}\cdot\Sigma_t^{-1}\tilde{S}_t = \tilde{S}_t^\top + \tilde{S}_t = \dot{\Sigma}_t \tag{6.21}$$

which is more or less equivalent to the original Euler-Lagrange equation (6.18); we defer the details to **??**.

To conclude, in contrast to the purely technical approach of **?**, our theorem 2 provides a geometric and conceptually clean solution for $\sigma\mathbb{W}_t$-GSBs: Compute the Lyapunov operator $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]$ via verifying the symmetry of the matrix in lemma 1. *It turns out that this technique can be readily extended to general GSBs*, and therefore serves as the foundation for the proof of our main result; see **??**.

REMARK It is interesting to note that the matrix $\tilde{S}_t$ in (6.20) is itself *not* symmetric. Other consequences of theorem 2 that might be of independent interest can be found in **??**. We also note that, when $\sigma = 0$, the solution to (6.17) is simply the Wasserstein geodesic between Gaussian measures, whose formula is well-known (Dowson and Landau, 1982; **?**). However, as explained in **??**, the case of $\sigma > 0$ requires a completely different analysis since, unlike when $\sigma = 0$, it is not reducible to a convex program. This leads to the significantly more involved proofs of theorem 2 and of (6.19) in **?**.

6.4 CLOSED-FORM SOLUTIONS OF GENERAL GAUSSIAN SCHRÖDINGER BRIDGES

We now present the closed-form solutions of general GSBs.

6.4.1 *Linear Stochastic Differential Equations*

We need the following background knowledge on the linear SDE Y_t. Let $\tau_t := \exp\left(\int_0^t c_s \, \mathrm{d}s\right)$. Then the solution to (6.7) is (**?**):

$$Y_t = \tau_t \left(Y_0 + \int_0^t \tau_s^{-1} \alpha_s \, \mathrm{d}s + \int_0^t \tau_s^{-1} g_s \, \mathrm{d}\mathbb{W}_s \right). \tag{6.22}$$

Another crucial fact in our analysis is that Y_t is a *Gaussian process given* Y_0, and is thus characterized by the first two moments. Using the independent increments of $\mathbb{W}_t$ and Itô's isometry (**?**), we compute:

$$\mathbb{E}[Y_t \,|\, Y_0] = \tau_t \left(Y_0 + \int_0^t \tau_s^{-1} \alpha_s \, \mathrm{d}s \right) =: \eta(t) \tag{6.23}$$

and, for any $t' \geq t$,

$$\begin{aligned} &\mathbb{E}\left[(Y_t - \eta(t)) \left(Y_{t'} - \eta(t')\right)^\top \,\middle|\, Y_0 \right] \\ &\qquad = \left(\tau_t \tau_{t'} \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s \right) I =: \kappa(t, t') I. \end{aligned} \tag{6.24}$$

6.4.2 *Main Result*

We now present the main result of our paper. With the important application of diffusion-based models in mind, we will not only derive solution curves as in (6.19) but also their SDE representations.

SDE with $\alpha_t \equiv 0$	Setting	$\kappa(t,t')$	$\sigma_\star^2$	r_t	$\bar{r}_t$	ρ_t	$\zeta(t)$
BM	$c_t{\equiv}0$ $g_t{\equiv}\omega{\in}\mathbb{R}^+$	$\omega^2 t$	ω^2	t	$1-t$	t	0
VESDE	$c_t{\equiv}0$ $g_t{=}\sqrt{\mathrm{q}(t)}$	$\mathrm{q}(t)$	$\mathrm{q}(1)$	$\frac{\mathrm{q}(t)}{\mathrm{q}(1)}$	$1-\frac{\mathrm{q}(t)}{\mathrm{q}(1)}$	$\frac{\mathrm{q}(t)}{\mathrm{q}(1)}$	0
VPSDE	$-2c_t{=}g_t^2$	$\tau_{t'}\left(\tau_t^{-1}{-}\tau_t\right)$	$\tau_1^{-1}{-}\tau_1$	$\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}$	$\tau_1\left(\frac{\tau_t}{\tau_1}-\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)$	$\frac{\tau_t^{-1}\left(\tau_t^{-1}-\tau_t\right)}{\tau_1^{-1}\left(\tau_1^{-1}-\tau_1\right)}$	0
sub–VPSDE	$\frac{g_t^2}{-2c_t}{=}1{-}\tau_t^4$	$\tau_t\tau_{t'}\left(\tau_t^{-1}{-}\tau_t\right)^2$	$\tau_1\left(\tau_1^{-1}{-}\tau_1\right)^2$	$\frac{\tau_t}{\tau_1}\cdot\left(\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)^2$	$\tau_t\left(1-\left(\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)^2\right)$	$\left(\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)^2$	0
SDE with $\alpha_t \not\equiv 0$	**Setting**	$\kappa(t,t')$	$\sigma_\star^2$	r_t	$\bar{r}_t$	ρ_t	$\zeta(t)$
OU/Vasicek	$c_t{\equiv}{-}\lambda{\in}\mathbb{R}$ $\alpha_t{\equiv}\mathbf{v}{\in}\mathbb{R}^d$ $g_t{\equiv}\omega{\in}\mathbb{R}^+$	$\frac{\omega^2 e^{-\lambda t'}\sinh\lambda t}{\lambda}$	$\frac{\omega^2\sinh\lambda}{\lambda}$	$\frac{\sinh\lambda t}{\sinh\lambda}$	$\sinh\lambda t\coth\lambda t$ $-\sinh\lambda t\coth\lambda$	$e^{-\lambda(1-t)}$ $\cdot\frac{\sinh\lambda t}{\sinh\lambda}$	$\frac{\mathbf{v}}{\lambda}\left(1-e^{-\lambda t}\right)$
α_t-BDT	$c_t{\equiv}0$ $g_t{\equiv}\omega{\in}\mathbb{R}^+$	$\omega^2 t$	$\omega^2 1$	t	$1-t$	t	$\int_0^t \alpha_s\,\mathrm{d}s$

Table 6.1: Examples of reference SDEs and the corresponding solutions of GSBs. All relevant functions in the Table are either introduced in section 6.4.1 or (6.25).

Let $\xi = \mathcal{N}(\mu_0, \Sigma_0)$ and $\xi' = \mathcal{N}(\mu_1, \Sigma_1)$ be two arbitrary Gaussian distributions in (GSB), and let D_σ, C_σ be as defined in (6.2).

Theorem 3 *Denote by $\mathbb{P}_t$ the solution to Gaussian Schrödinger bridges (GSB). Set*

$$
\begin{aligned}
&r_t := \frac{\kappa(t,1)}{\kappa(1,1)}, \quad \bar{r}_t := \tau_t - r_t\tau_1, \quad \sigma_\star := \sqrt{\tau_1^{-1}\kappa(1,1)},\\
&\zeta(t) := \tau_t \int_0^t \tau_s^{-1}\alpha_s\,\mathrm{d}s, \rho_t := \frac{\int_0^t \tau_s^{-2} g_s^2\,\mathrm{d}s}{\int_0^1 \tau_s^{-2} g_s^2\,\mathrm{d}s},\\
&P_t := \dot{r}_t(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}), \quad Q_t := -\dot{\bar{r}}_t(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}),\\
&S_t := P_t - Q_t^\top + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big] I.
\end{aligned} \tag{6.25}
$$

Then the following holds:

1. *The solution $\mathbb{P}_t$ is a Markov Gaussian process whose marginal variable $X_t \sim \mathcal{N}(\mu_t, \Sigma_t)$, where*

$$\mu_t := \bar{r}_t\mu_0 + r_t\mu_1 + \zeta(t) - r_t\zeta(1), \tag{6.26}$$

$$\Sigma_t := \bar{r}_t^2\Sigma_0 + r_t^2\Sigma_1 + r_t\bar{r}_t\left(C_{\sigma_\star} + C_{\sigma_\star}^\top\right) + \kappa(t,t)(1-\rho_t)I. \tag{6.27}$$

2. *X_t admits a closed-form representation as the SDE:*

$$\mathrm{d}X_t = f_{\mathcal{N}}(t, X_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbf{W}_t \tag{6.28}$$

where

$$f_{\mathcal{N}}(t,x) := S_t^\top \Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t. \tag{6.29}$$

Moreover, the matrix $S_t^\top \Sigma_t^{-1}$ is symmetric.

As in theorem 2, the key step in the proof of theorem 3 is to recognize the symmetry of the matrix $S_t^\top \Sigma_t^{-1}$ where S_t, defined in (6.25), simply becomes the $\tilde{S}_t$ in lemma 1 (up to an additive factor of $\frac{\sigma^2 \bar{t}}{2} I$) for $\sigma \mathbb{W}_t$-GSBs. Although this can be directly verified via generalizing lemma 1, the computation becomes quite tedious, so our proof of theorem 3 will follow a slightly different route. In any case, given the symmetry of $S_t^\top \Sigma_t^{-1}$, the proof simply boils down to a series of straightforward calculations; see **??**.

CLOSED FORMS FOR CONDITIONAL DISTRIBUTIONS In many practical applications such as generative modeling, a requirement to employ the SDE representation of GSBs in (6.28) is that its *conditional distributions* given the initial points can be computed efficiently. As an immediate corollary of theorem 3, we obtain the following closed-form expressions for these conditional distributions.

Corollary 1 *Let $X_t \sim \mathbb{P}_t$ be the the solution to (GSB). Then the conditional distribution of X_t given end points has a simple solution: $X_t | X_0 = x_0 \sim \mathcal{N}\left(\mu_{t|0}, \Sigma_{t|0}\right)$, where*

$$\mu_{t|0} = \bar{r}_t x_0 + r_t\left(\mu_1 + C_{\sigma_\star}^\top \Sigma_0^{-1}(x_0 - \mu_0)\right) + \zeta(t) - r_t \zeta(1), \tag{6.30}$$

$$\Sigma_{t|0} = r_t^2\left(\Sigma_1 - C_\sigma^\top \Sigma_0^{-1} C_\sigma\right) + \kappa(t,t)(1-\rho_t) I. \tag{6.31}$$

Similarly, $X_t | X_1 = x_1 \sim \mathcal{N}\left(\mu_{t|1}, \Sigma_{t|1}\right)$, where

$$\mu_{t|1} = r_t x_1 + \bar{r}_t\left(\mu_0 + C_{\sigma_\star} \Sigma_1^{-1}(x_1 - \mu_1)\right) + \zeta(t) - r_t \zeta(1), \tag{6.32}$$

$$\Sigma_{t|1} = \bar{r}_t^2\left(\Sigma_0 - C_\sigma \Sigma_1^{-1} C_\sigma^\top\right) + \kappa(t,t)(1-\rho_t) I. \tag{6.33}$$

Examples of GSBs. Our framework captures most popular reference SDEs in the machine learning literature as well as other mathematical models in financial engineering; see table 6.1. A non-exhaustive list includes:

- The basic Brownian motion (BM) and the Ornstein-Uhlenbeck (OU) processes, both widely adopted as the reference process for SB-based models (**?**De Bortoli et al., 2021; Lavenant et al., 2021; Vargas et al., 2021;

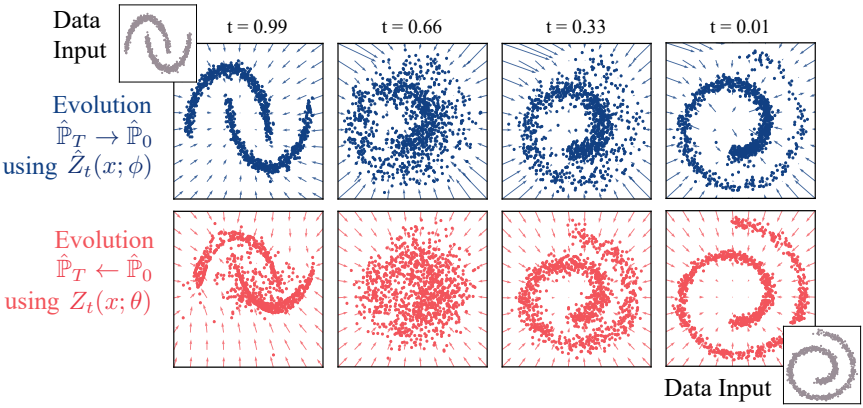


FIGURE 6.2: Illustration of the time-dependent drifts learned by GSBFLOW with VE SDE for two toy marginal distributions. *Top.* Evolution of $\hat{\mathbb{P}}_1$ (moons) $\to \hat{\mathbb{P}}_0$ (spiral) via backward policy $\hat{Z}_t^{\phi}(x)$. *Bottom.* Evolution of $\hat{\mathbb{P}}_0$ (spiral) $\to \hat{\mathbb{P}}_1$ (moons) via forward policy $Z_t^{\theta}(x)$.

?). We also remark that, even though (6.27) is known for BM (**?**), what is crucial in these applications is the SDE presentation (6.28), which is new even for BM.

- The variance exploding SDEs (VESDEs), which underlies the training of score matching with Langevin dynamics for diffusion-based generative modeling (**??**Song et al., 2021).
- The variance preserving SDEs (VPSDEs), which can be seen as the continuous limit of denoising diffusion probabilistic models (**??**Song et al., 2021), another important class of algorithms for diffusion-based generative modeling.
- The *sub-VPSDEs* proposed by (Song et al., 2021), which are motivated by reducing the variance of VPSDEs.
- Several important SDEs in financial engineering, such as the *Vasicek model* (which generalizes OU processes) and the *constant volatility* α_t*-Black-Derman-Toy (BDT) model* (**?**).

Method	Tasks	
	Wasserstein Loss $W_\varepsilon \downarrow$	
	?	Schiebinger et al. (2019)
Song et al. (2021)		
VESDE	20.83 ± 0.18	40.81 ± 0.42
sub-VPSDE	$\mathbf{19.96 \pm 0.58}$	48.15 ± 3.38
GSBFLOW (*ours*)		
VESDE	25.18 ± 0.10	$\mathbf{27.85 \pm 0.68}$

TABLE 6.2: Evaluation of predictive performance w.r.t. the entropy-regularized Wasserstein distance W_ε (Cuturi, 2013) of GSBFLOW and baselines on generating different single-cell datasets (using 3 runs).

6.5 EMPIRICAL EVALUATION

The purpose of our experiments is to demonstrate that, by leveraging moment information, GSBFLOW is significantly more stable compared to other SB-based objectives, especially when moving beyond the *generative* setting where $\hat{\mathbb{P}}_1$ is a simple Gaussian. Indeed, while performing competitively in the generative setting ($\mathcal{N}_0 \to \hat{\mathbb{P}}_1$), our method *outperforms* when modeling the evolution of two complex distributions ($\hat{\mathbb{P}}_0 \to \hat{\mathbb{P}}_1$), the most general and ambitious setting to estimate a bridge. This is demonstrated on synthetic data as well as a task from molecular biology concerned with modeling the dynamics of cellular systems, i.e., single-cell genomics (**?**Frangieh et al., 2021; **?**).

6.5.1 *Synthetic Dynamics*

Before conducting the single-cell genomics experiments, we first test GSBFLOW on a synthetic setting. Our first task involves recovering the stochastic evolution of two-dimensional synthetic data containing two interleaving half circles ($\hat{\mathbb{P}}_1$) into a spiral ($\hat{\mathbb{P}}_0$). fig. 6.2 shows the trajectories learned by GSBFLOW based on the VESDE (see table 6.1 and **??**).

While it is sufficient to parameterize only a single policy ($\hat{Z}_t^\phi(x)$) in generative modeling, the task of learning to evolve $\hat{\mathbb{P}}_0$ into $\hat{\mathbb{P}}_1$ requires one to recover *both* vector fields $\hat{Z}_t^\phi(x)$ and $Z_t^\theta(x)$. As demonstrated in

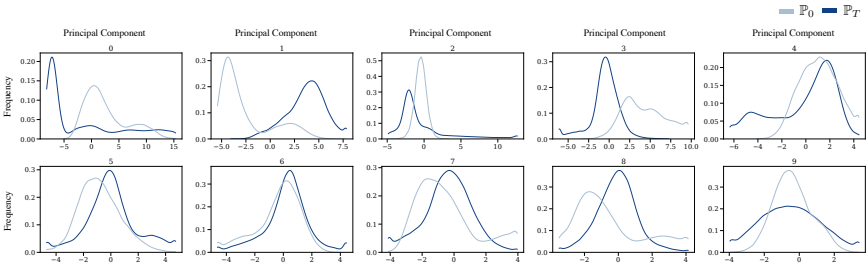


FIGURE 6.3: The expression levels of the first 10 principal components from the dataset by Schiebinger et al. (2019).

fig. 6.2, GSBFLOW is able to successfully learn both policies $Z_t^\theta(x)$ and $\hat{Z}_t^\phi(x)$ and reliably recovers the corresponding targets of the forward and backward evolution. While initializing the reference process through the closed-form SB between the Gaussian approximations of both synthetic datasets provides good results, the power of GSBFLOW becomes evident in more complex applications which we tackle next.

6.5.2 *Single-Cell Dynamics*

Modern single-cell profiling technologies are able to provide rich feature representations (e.g., gene expression) of *individual* cells at any development state. A crucial issue that arises with such profiling methods is their destructive nature: Measuring a cell requires destroying it and thus a cell cannot be measured twice. As a result, independent samples are collected at each snapshot, with no access to ground-truth single-cell trajectories throughout time, resulting in challenging, *unaligned*, datasets. Recovering cellular dynamics from such unaligned snapshots, i.e., $\hat{\mathbb{P}}_0$ to $\hat{\mathbb{P}}_1$, has, however, extremely important scientific and biomedical relevance (**?**). For example, it determines our understanding on how and why tumor cells evade cancer therapies (Frangieh et al., 2021) or unveils mechanisms of cell differentiation and development (Schiebinger et al., 2019). Following related work, in particular previous methods based on optimal transport (Schiebinger et al., 2019; Bunne et al., 2021, 2022; Tong et al., 2020), the task is thus to learn the stochastic process that described the evolution of single cells from $\hat{\mathbb{P}}_0$ to $\hat{\mathbb{P}}_1$.

6.5.2.1 *Experimental Setup*

SINGLE-CELL GENOMICS VIA Let us consider the evolution of a gene, for which we can collect the empirical distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ of its expression levels at the times $t = 0, 1$ (Schiebinger et al., 2019; **?**). Our goal is to two-fold:

1. To solve the **generative modeling** problem, i.e., to generate $\hat{\mathbb{P}}_0$ or $\hat{\mathbb{P}}_1$ from a standard Gaussian noise, and
2. to **evolve** $\mathbb{P}_0 \to \mathbb{P}_1$ or $\mathbb{P}_1 \to \mathbb{P}_0$, i.e., to recover a stochastic process $\mathbb{P}_t$ satisfying $\mathbb{P}_0 = \hat{\mathbb{P}}_0, \mathbb{P}_1 = \hat{\mathbb{P}}_1$.

Although there are numerous algorithms for generative modeling, to our knowledge, the only framework that can simultaneously solve both tasks is the SB-based scheme recently proposed in (Chen et al., 2022). In order to apply this framework, one has to choose a prior process Y_t, which is taken by the authors to be the high-performing VESDE and sub-VPSDE. These SB-based methods, as well as several standard generative modeling algorithms (**??**Song et al., 2021; **?**; **?**,**?**) for the first task, constitute strong baselines for our experiments.

OUR CHOICE OF Y_t; THE GSBFLOW Instead of directly diving into the numerical solution of SBs as in Chen et al. (2022), we first empirically verify that the distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ in single-cell genomics are typically close to *non-standard* Gaussian distributions: See fig. 6.3 for the canonical dataset (Schiebinger et al., 2019) and **??** in **??** for the same phenomenon on another standard benchmark (**?**).

Since the solutions of SBs are Lipschitz in terms of $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ (**?**), a reasonable approximation to the original SB objective is to replace $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ by Gaussians with matching moments. This results in a GSB problem which can be solved in closed form by our theorem 3. Intuitively, if we denote an existing prior process by Y_t and the solution of its corresponding GSB by X_t, then X_t presents a more appealing prior process than Y_t since it carries the moment information of $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$, whereas Y_t is completely data-oblivious.

Motivated by these observations, we propose a simple modification of the framework in Chen et al. (2022): Replace the prior process Y_t by its GSB approximation and keep everything else the same. The resulting scheme, which we term the GSBFLOW, learns a pair of forward $Z_t^\theta(x)$ and backward parametrized drifts $\hat{Z}_t^\phi(x)$ that progressively transport samples from $\hat{\mathbb{P}}_0 \to \hat{\mathbb{P}}_1$ and $\hat{\mathbb{P}}_1 \to \hat{\mathbb{P}}_0$, respectively. The full algorithm is presented in **??** for completeness.

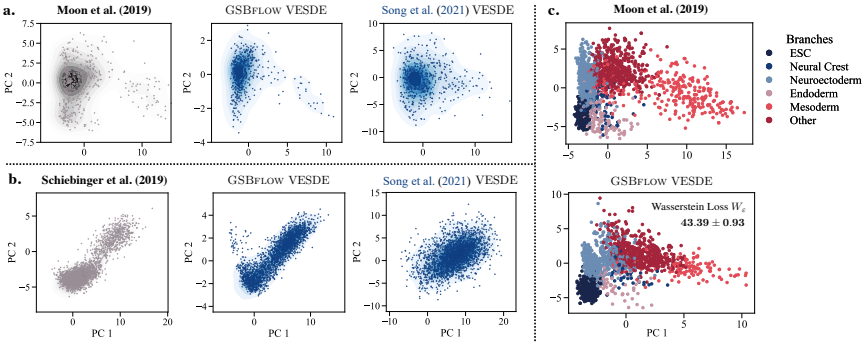


FIGURE 6.4: **a.-b.** Visual evaluation of the ability of our method to model the **generation** of data from **a. ?** and **b.** Schiebinger et al. (2019). Density plots are visualized in 2D PCA space and show generated data points using either GSBFLOW (our method) or the procedure in Song et al. (2021). **c.** Evaluation of GSBFLOW's ability to model the entire **evolution** of a developmental process of **?**, visualized by the data and GSBFLOW predictions colored by the lineage branch class.

6.5.2.2 *Results*

We investigate the ability of GSBFLOW to generate cell populations $\hat{\mathbb{P}}_1$ from noise $\mathcal{N}_0$ ($\mathcal{N}_0 \to \hat{\mathbb{P}}_1$, Fig. 6.4a, b) on the the canonical datasets (**?**Schiebinger et al., 2019); as well as to predict the dynamics of single-cell genomics ($\hat{\mathbb{P}}_0 \to \hat{\mathbb{P}}_1$, Fig. 6.4c) (**?**), i.e., the inference of cell populations $\hat{\mathbb{P}}_1$ resulting from the developmental process of an initial cell population $\hat{\mathbb{P}}_0$, with the goal of learning individual dynamics, identify ancestor and descendant cells. Details on datasets and experimental design can be found in **??**–**??**. The evaluation is conducted on the first 20 or 30 components of the PCA space of the > 1500 highly differentiable genes (see **??**–**??**).

We evaluate the quality of the generated cellular states through the entropy-regularized Wasserstein distance W_ε (see table 6.2) and by visualizing the first two principal components (PC), see fig. 6.4a, b. GSBFLOW performs competitively on reconstructing embryoid body differentiation landscapes (**?**), and outperforms score-based generative models baselines on the iPSC reprogramming task (Schiebinger et al., 2019) as quantified by W_ε between data and predictions. Further, we analyze GSBFLOW's ability to predict the temporal evolution of embryoid body differentiation (**?**), where cells measured at day 1 to 3 serve as samples of $\hat{\mathbb{P}}_0$, while $\hat{\mathbb{P}}_1$ is constructed

from samples between day 12 to 27. As no ground truth trajectories are available in the data, we compare the predicted evolution to the data and compare how well the heterogeneity of lineage (fig. 6.4c, upper panel) or sublineage branches (**??**a) is captured. fig. 6.4c (lower panel) and **??**b thereby closely resemble the data (see W_ε in fig. 6.4c) and thus demonstrate GSBFLOW's ability to learn cell differentiation into various lineages and to capture biological heterogeneity on a more macroscopic level.

6.6 CONCLUSION

We derive closed-form solutions of GSBs, an important class of dynamic OT problems. Our technique originates from a deep connection between Gaussian OT and the Bures-Wasserstein geometry, which we generalize to the case of general SB problems. Numerically, we demonstrate that our new closed forms inspire a simple modification of existing SB-based numerical schemes, which can however lead to significantly improved performance.

Limitation of our framework. In a broader context, we hope our results can serve as the inspiration for more learning algorithms, much like how existing closed-form solutions of Gaussian OT problems have contributed to the machine learning community. We thus acknowledge a severe limitation of our closed-form solutions: These formulas require matrix inversions, which might face scalability issues for high-dimensional data. In addition, existing matrix inversion algorithms are typically extremely sensitive to the condition number, and thus our formulas are not as useful for ill-conditioned data. Lifting these constraints to facilitate further applications, such as to image datasets, is an important future work.

7

...

...

— ..., ... (...)

8

...

Ein solches mathematisch-definierbares System ist überhaupt nicht die Wirklichkeit selbst, sondern nur ein Schema, welches zur Beschreibung der Wirklichkeit dienen kann.

— Andrey Kolmogorov, *Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung* (1931)

MODELING PARTICLE DYNAMICS AS A JKO SCHEME. In this paper, we draw inspiration from both approaches above—the intuition from the recent NF literature that flows should mimic an optimal transport (OT as prior), and be able, through training, to predict future configurations (OT as a loss)—to propose a causal model for population dynamics. Our approach relies on a powerful hammer: the Jordan-Kinderlehrer-Otto (JKO) flow (Jordan et al., 1998), widely regarded as one of the most influential mathematical breakthroughs in recent history. While the JKO flow was initially introduced as an alternative method to solve the Fokker-Planck partial differential equation (PDE), its flexibility can be showcased to handle more complex PDEs (Santambrogio, 2017, §4.7), or even describe the gradient flows of non-differentiable energies that have no PDE representation. On a purely mechanical level, a JKO step is to measures what the proximal step (Combettes and Pesquet, 2011) is to vectors: In a JKO step, particles move to decrease collectively an *energy* (a real-valued function defined on measures), yet remain close (in Wasserstein sense) to the previous configuration. Our goal in this paper is to treat JKO steps as parameterized modules, and fit their parameter (the energy function) so that its outputs agree repeatedly over time with observed data. This approach presents several challenges: While numerical approaches to solve JKO steps have been proposed in low dimensional settings (Burger et al., 2010; Carrillo et al., 2021; Peyré, 2015; Benamou et al., 2016a), scaling it to higher dimensions is an open problem. Moreover, minimizing a loss involving a JKO step w.r.t. energy requires not only solving the JKO problem, but also computing the (transpose) Jacobian of its output w.r.t. energy parameters.

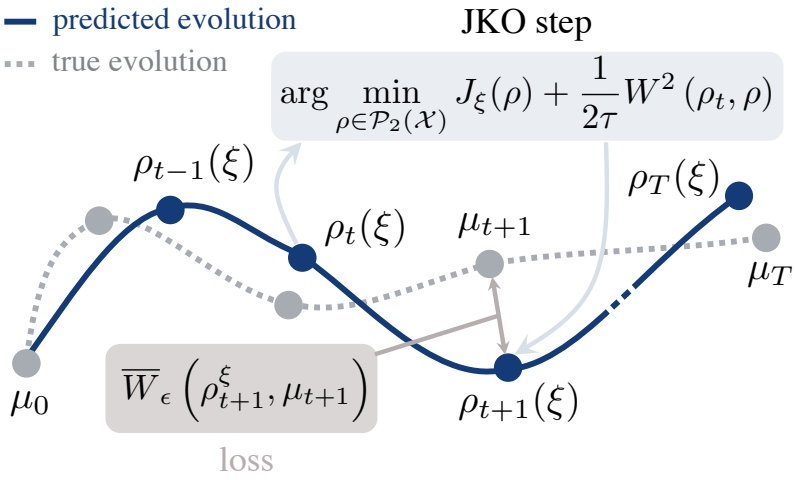


FIGURE 8.1: Given an observed trajectory $(\mu_0,\dots,\mu_T)$ of point clouds (gray), we seek parameters ξ for the energy J_ξ such that the predictions $\rho_1,\dots,\rho_T$ (blue) following a JKO flow from $\rho_0=\mu_0$ are close the observed trajectory (gray), by minimizing (as a function of ξ) the sum of Wasserstein distances between ρ_{t+1}, the JKO step from ρ_{t-1} using J_ξ, and data μ_{t+1}.

CONTRIBUTIONS. Our contributions are two-fold. First, we propose a method, given an input configuration and an energy function, to compute JKO steps using input convex neural networks (ICNN) (Amos et al., 2017; **?**) (see also concurrent works that have proposed similar approaches (Alvarez-Melis et al., 2021; Mokrov et al., 2021)). Second, we view the JKO step as an inner layer, a JKONET module parameterized by an energy function, which is tasked with moving the particles of an input configuration along an OT flow (the gradient of an optimal ICNN), trading off a lower energy with proximity to the previous configuration. We propose to estimate the parameters of the energy by minimizing a fitting loss computed between the outputs of the JKONET module (the prediction) and the ground truth displacements, as illustrated in Figure 8.1. We demonstrate JKONET's range of applications by applying in on synthetic potential- and trajectory-based population dynamics, as well as developmental trajectories of human embryonic stem cells based on single-cell genomics data.

8.1 BACKGROUND

JKO FLOWS. In their seminal paper, Jordan et al. (1998) study diffusion processes under the lens of the OT metric (see also Ambrosio et al., 2006) and introduce a scheme that is now known as the JKO flow: Starting with ρ_0, and given a real-valued energy function $J : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ driving the evolution of the system, they define iteratively for $t \geq 0$, :

$$\rho_{t+1} = \arg\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} J(\rho) + \frac{1}{2\tau} W^2(\rho, \rho_t)\,, \tag{8.1}$$

where τ is a time step parameter. These successive minimization problems result in a sequence of probability measures in $\mathcal{P}(\mathbb{R}^d)$. The JKO flow can thus be seen as the analogy of the usual proximal descent scheme, tailored for probability measures (Santambrogio, 2015, p.285). Jordan et al. (1998) show that as step size $\tau \to 0$, and for a specific energy J that is the sum of a linear term and the negentropy, the measures describing the JKO flow recover solutions to a Fokker-Planck equation. In this work, following in the footsteps of more general applications of the JKO scheme (Santambrogio, 2017, §4.8), we model dynamics without necessarily having in mind PDE solutions in mind, to interpret instead the JKO step as a more general parametric type of dynamic for probability measures, exclusively parameterized by the energy J itself.

8.2 PROXIMAL OPTIMAL TRANSPORT MODEL

Given T discrete measures $\mu_0, \ldots, \mu_T$ describing the time evolution of a population, we posit that such an evolution follows a JKO flow for the free energy functional J, and assume that energy does not change throughout the dynamic. We parameterize the energy J as a neural network with parameters ξ, and fit ξ so that the JKO flow model matches the observed data.

Fitting parameter ξ with a reconstruction loss requires, using the chain rule, being able to differentiate the JKO step's output w.r.t. ξ (see Fig. 8.1), and more precisely provide a way to apply that transpose Jacobian to an arbitrary vector when using reverse-mode differentiation. To achieve this, we introduce a novel approach to numerically solve JKO flows using ICNNs (§ 8.2.1), resulting in a bilevel optimization problem targeting the energy J_ξ (§ 8.2.2).

8.2.1 *Reformulation of JKO Flows via ICNNs*

Given a starting condition ρ_t and energy functional J_ξ, the JKO step consists in producing a new measure ρ_{t+1} implicitly defined as the minimizer of (8.1). Solving directly (8.1) on the space of measures, involves substantial computational costs. Different numerical schemes have been developed, e.g., based notably on Eulerian discretization of measures (Carrillo et al., 2021; Benamou et al., 2016b), and/or entropy-regularized optimal transport (Peyré, 2015). However, these methods are limited to small dimensions since the cost of discretizing such spaces grows exponentially. Except for the Eulerian approach proposed in (Peyré, 2015), obtained as the fixed point of a Sinkhorn type iteration, the differentiation would also prove extremely challenging as a function of the energy parameter ξ.

To reach scalability and differentiability, we build upon the approach outlined in Benamou et al. (2016b) to reformulate the JKO scheme as a problem solved over convex functions, rather than on measures ρ. Effectively, this is equivalent to making a change of variables in (8.1): Introduce a (variable) convex function ψ, and replace the variable ρ by the variable $\nabla\psi_\# \rho_t$. Writing

$$\mathcal{E}_J(\rho, \nu) := J(\rho) + \frac{1}{2\tau} W_2^2(\rho, \nu), \tag{8.2}$$

this identity states that, assuming μ and ν being absolutely continuous w.r.t. Lebesgue measure that

$$\min_\rho \mathcal{E}_J(\rho, \nu) = \min_{\psi \text{ convex}} \mathcal{F}_J(\psi, \nu) := \mathcal{E}_J(\nabla\psi_\# \nu, \nu)\,,$$

simplifying the Wasserstein term in (8.2), using the assumption that ψ is convex and Brenier's theorem (§ **??**):

$$\mathcal{F}_J(\psi, \nu) = J(\nabla\psi_\# \nu) + \frac{1}{2\tau} \int \|x - \nabla\psi(x)\|^2 d\nu(x) \tag{8.3}$$

We pick an ICNN architecture to optimize over a restricted family of convex functions, $\{\psi_\theta\}$, and define, starting from $\rho_0(\xi) := \mu_0$, the recursive sequence for $t \geq 0$,

$$\rho_{t+1}(\xi) := \nabla\psi_{\theta^\star(\xi, \rho_t(\xi))\,\#}\, \rho_t(\xi)\,, \tag{8.4}$$

with $\theta^\star(\xi, \rho_t)$ defined implicitly using ξ and any ν as

$$\theta^\star(\xi, \nu) := \arg\min_\theta \mathcal{F}_J(\psi_\theta, \nu) \tag{8.5}$$

STRONG CONVEXITY OF ψ_θ. The strong convexity and smoothness of a potential ψ impacts the regularity of the corresponding OT map $\nabla\psi$ (Caffarelli, 2000; Figalli, 2010), since one can show that for a ℓ-strongly convex, L-smooth ψ one has (Paty et al., 2020) that

$$\ell\|x-y\| \leq \|\nabla\psi(x) - \nabla\psi(y)\| \leq L\|x-y\|.$$

While it is more difficult to enforce the L-smoothness of a neural network, and more generally its Lipschitz constants (Scaman and Virmaux, 2018) it is easy to enforce its strong convexity, by simply adding a term $\ell\|x\|^2/2$ to the corresponding potential, or a residual rescaled term ℓx to the output $\nabla\psi(x)$. This approach can be used to enforce that the push-forward of the gradient of an ICNN does not collapse to a single point, maintaining spatial diversity.

8.2.2 *Learning the Free Energy Functional*

The energy function $J_\xi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ can be any parameterized function taking a measures as an input. Since our model assumes that the observed dynamic is parameterized entirely by that energy (and the initial observation ρ_0), the more complex this dynamic, the more complex one would expect the energy J_ξ to be. We focus in this first attempt on linear functions in the space of measures, that is expectations over ρ of a vector-input neural network E_ξ

$$J_\xi(\rho) := \int E_\xi(x) d\rho(x), \tag{8.6}$$

where $E_\xi : \mathbb{R}^d \to \mathbb{R}$ is a multi-layer perceptron (MLP).

Algorithm 1 JKONET Algorithm.

Input: Dataset $\mathcal{D} = \{\{\mu_t^0\}_{t=0}^T, \ldots, \{\mu_t^N\}_{t=0}^T\}$ of N population trajectories, ξ^0 energy parameter initialization, θ^0 ICNN parameter initialization, learning rates lr_θ and lr_ξ, step τ, regularizer ε, tolerance α, `TeacherForcing` flag.

Output: Free energy J_ξ explaining underlying population dynamics of snapshot data.

$\xi \leftarrow \xi^0$
for $\{\mu_t\}_{t=0}^T \in \mathcal{D}$ **do**
  for $t \leftarrow 0$ ***to*** $T-1$ **do**
    $\theta \leftarrow \theta^0$
    if *TeacherForcing* **then**
      $\nu \leftarrow \mu_t$
    else
      $\nu \leftarrow \rho_t(\xi)$
    while $\frac{\sum_i \|\nabla_{\theta_i} \mathcal{F}_{J_\xi}(\theta)\|_2}{\sum_i count(\theta_i)} \geq \alpha$ **do**
      $\theta \leftarrow \theta - \mathrm{lr}_\theta \times \nabla_\theta \mathcal{F}_{J_\xi, \nu}(\theta)$
    $\rho_{t+1}(\xi) \leftarrow \nabla \psi_{\theta\#} \nu$
  $\xi \leftarrow \xi - \mathrm{lr}_\xi \times \nabla_\xi \overline{W}_\varepsilon(\rho_{t+1}(\xi), \mu_{t+1})$
return J_ξ

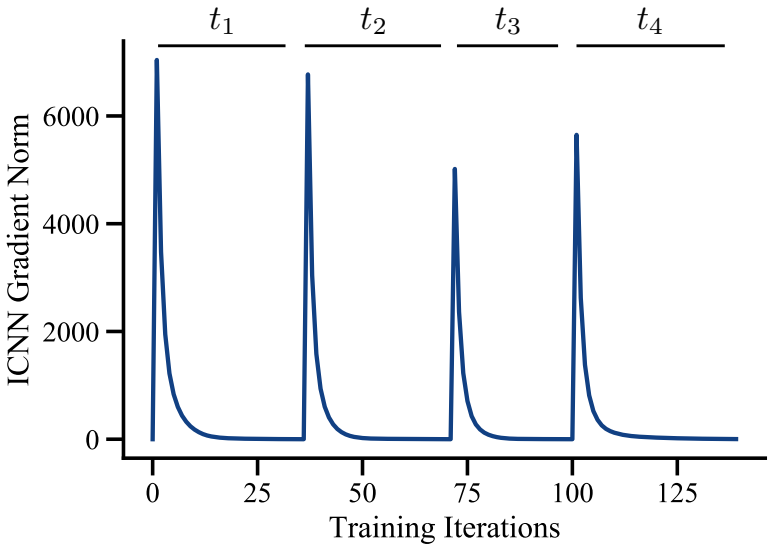


FIGURE 8.2: Optimization of the ICNN used in JKO steps. The bumps correspond to a change in the outer iteration, the smooth decrease in between

Inferring nonlinear energies accounting for population growth and decline, as well as interactions between points, using the formalism of (De Bie et al., 2019), transformers (Vaswani et al., 2017) or set pooling methods (Edwards and Storkey, 2017; Zaheer et al., 2017), is an exciting direction for future work.

To address slow convergence and instabilities for dynamics with many snapshots, we use teacher forcing (Williams and Zipser, 1989) to learn J_ξ through time. In those settings, during training, J_ξ uses the ground truth as input instead of predictions from the previous time step. At test time, we do not use teacher forcing.

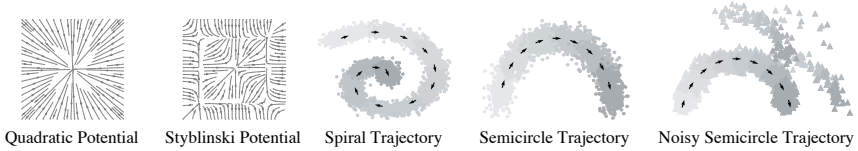
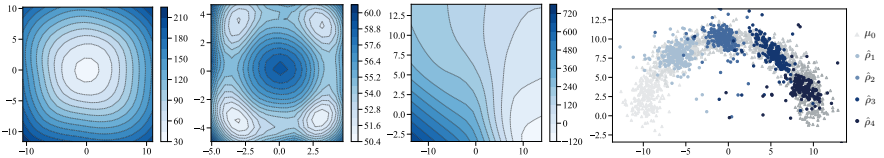


FIGURE 8.3: Overview on different tasks including trajectory- and potential-based dynamics.



(a) Quadratic Potential.

(b) Styblinski Potential.

(c) Semicircle Trajectory *with* teacher forcing.

(d) Predicted Population Evolution on Semicircle Trajectory.

FIGURE 8.4: **Results of JKOnet on Potential- and Trajectory-based Dyanamics.** (a)-(c) Contour plots of the energy functionals J_ξ of JKONET on potential- and trajectory-based population dynamics in different training settings (i.e., trained with or without teacher forcing § 8.2.2), color gradients depict the magnitude of J_ξ. (d) Predicted population snapshots $(\hat{\rho}_1, \ldots, \hat{\rho}_4)$ (blue) and data trajectory $(\mu_0, \ldots, \mu_4)$ (gray).

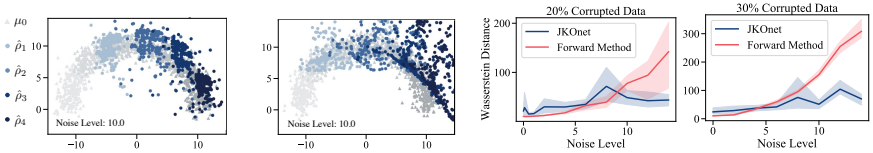
8.2.3 *Bilevel Formulation of* JKONET

Learning the free energy functional J_ξ while solving each JKO step via an ICNN results in a challenging bilevel optimization problem. At each time

step, the predicted dynamics are compared to the ground truth trajectory $(\mu_0, \mu_1, \ldots, \mu_T)$ with a Sinkhorn loss (1.3),

$$\begin{aligned} &\min_{\xi} \sum_{t=0}^{T-1} \overline{W}_\varepsilon(\rho_{t+1}(\xi), \mu_{t+1}), \\ &\text{s.t. } \rho_0(\xi) := \mu_0, \\ &\qquad \rho_{t+1}(\xi) := \nabla\psi_{\theta^\star} \# \rho_t(\xi)\,, \\ &\qquad \theta^\star := \arg\min_{\theta} \mathcal{F}_{J_\xi}(\psi_\theta, \rho_t(\xi)) \end{aligned} \tag{8.7}$$

The dependence of the Sinkhorn divergence losses in (8.7) on ξ only appears in the fact that the predictions $\rho_{t+1}(\xi)$ are themselves implicitly defined as solving a JKO step parameterized with the energy J_ξ. Learning J_ξ through the exclusive supervision of data observations requires therefore to differentiate the arg-minimum of a JKO problem, down therefore through to the lower-level optimization of the ICNN. We achieve this by implementing a differentiable double loop in `JAX`, differentiating first the Sinkhorn divergence using the `OTT`[1] package (Cuturi et al., 2022), and then backpropagating through the ICNN optimization by unrolling Adam steps (Kingma and Ba, 2014; Metz et al., 2017; Lorraine et al., 2020).



(a) JKONET on 30% corrupted data.

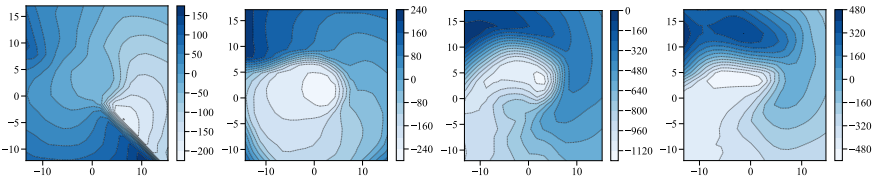
(b) Forward method on 30% corrupted data.

(c) W_ϵ (1.2) vs. noise level on 20% corrupted data.

(d) W_ϵ (1.2) vs. noise level on 30% corrupted data.

FIGURE 8.5: Comparison between JKONET and the forward method in settings of increasing noise on corrupted data on the semicircle trajectory task.

1 github.com/ott-jax/ott



(a) Forward method, *with* teacher forcing.

(b) Forward method, *no* teacher forcing.

(c) JKONET, *with* teacher forcing.

(d) JKONET, *no* teacher forcing.

FIGURE 8.6: Comparison between energy functionals J_ξ of the spiral trajectory task (see 8.3) between the forward method and JKONET, trained with or without teacher forcing § 8.2.2). When using teacher forcing, the forward method overfits a gap on the lower-right corner of the spiral, outputting a highly irregular energy. When taking into account the entire trajectory recursively, the Forward method does better overall, but is unable to recover an energy as precise as that returned by JKONET.

INNER LOOP TERMINATION. A question that arises when defining $\rho_{t+1}(\xi)$ lies in the budget of gradient steps needed or allowed to optimize the parameters θ of the ICNN, before taking a new gradient step on ξ in the outer loss. A straightforward approach in `JAX` (Bradbury et al., 2018) would be to use a preset number of iterations with a `for` loop (`jax.lax.scan`). We do observe, however, that the number of iterations needed to converge in relevant scenarios can vary significantly with the ICNN architecture and/or the hardness of the underlying task. We propose to use instead a differentiable fixed-point loop to solve each JKO step up to a desired convergence threshold. We measure convergence of the optimization of the ICNN via the average norm of the gradient of the JKO objective w.r.t. the ICNN parameters θ, i.e., $\sum_i \|\nabla_{\theta_i} \mathcal{F}_{J_\xi}(\theta_i, \xi)\|_2 / \sum_i \text{count}(\theta_i)$. We observe that this approach is robust across datasets and architectures of the ICNN. An exemplary training curve for the ICNNs updated successively along a time sequence is shown in Figure 8.2.

REVERSE-MODE DIFFERENTIATION. The Jacobian $\partial \rho_{t+1} / \partial \xi$ arising when computing the gradient $\nabla_\xi \overline{W}_\varepsilon(\rho_{t+1}(\xi), \mu_{t+1})$ is obtained by unrolling the while loop above. The gradient term of the Sinkhorn divergence w.r.t the first argument is given by the Danskin envelope theorem (Danskin, 1967).

Method	Prediction Loss (W_ε)			
	Day 6 to 9	Day 12 to 15	Day 18 to 21	Day 24 to 27
One-Step Ahead				
Forward Method	0.187 ± 0.001	0.162 ± 0.010	0.185 ± 0.020	0.203 ± 0.004
JKONET	$\mathbf{0.133 \pm 0.020}$	$\mathbf{0.133 \pm 0.008}$	$\mathbf{0.172 \pm 0.0130}$	$\mathbf{0.169 \pm 0.004}$
All-Steps Ahead				
Forward Method	0.225 ± 0.023	0.160 ± 0.001	0.171 ± 0.016	0.183 ± 0.007
JKONET	$\mathbf{0.148 \pm 0.015}$	$\mathbf{0.144 \pm 0.013}$	$\mathbf{0.154 \pm 0.024}$	$\mathbf{0.138 \pm 0.034}$

TABLE 8.1: Evaluation of predictive performance w.r.t. the entropy-regularized Wasserstein distance W_ε (1.2) of JKONET and the forward method on the embryoid body scRNA-seq data per time step (using 3 runs).

SETTING τ IN (**??**). In usual JKO applications, τ needs to be tuned manually. In this work, the energy J_ξ is not fixed, but trained to fit data. Since we put no constraints on the scaling of J_ξ, τ can be set to 1 without loss of generality, as the parameter ξ will automatically adjust so that the scale of J_ξ induces steps of a relevant length to fit data. This only holds (as with a usual JKO step) if the trajectories are sampled regularly. For irregularly spaced time series, τ can be adapted at train and test time to the spacing of timestamps (shorter steps requiring larger τ).

8.3 EVALUATION

In the following, we evaluate our method empirically on a variety of tasks. This includes recovering synthetic potential- and trajectory-based population dynamics (see Fig. 8.3), as well as the evolution of high-dimensional single-cell populations during a developmental process.

8.3.1 *Synthetic Population Dynamics*

ENERGY-DRIVEN TRAJECTORIES. The first task involves evolutions of partial differential equations with known potential. We hereby consider both convex (e.g., the quadratic function $J(x) = \|x\|_2^2$) and nonconvex potentials (e.g., Styblinski function) (see Fig. 8.3). These two-dimensional synthetic flows are generated using the Euler-Maruyama method (Kloeden and Platen, 1992). For details, see § **??**. To recover the true potential via JKONET, we parameterize both energy J_ξ and ICNN ψ_θ with linear layers ($\epsilon = 1.0$, $\tau = 1.0$, § **??**). More details on the architectures can be found

in § **??**. Figure 8.4a-b demonstrate JKONET's ability to recover convex and nonconvex potentials via energy J_ξ.

ARBITRARY TRAJECTORIES. As a sanity check, we evaluate if JKONET can recover an energy functional J_ξ from trajectories that are not necessarily arising from the gradient of an energy. Here, a 2-dimensional Gaussian moves along a predefined trajectory with nonconstant speed. For details on the data generation, see § **??**. We consider a line, a spiral, and movement along a semicircle (Fig. 8.3). As visible in Figure 8.4c (5 snapshots), Figure **??**b (2 snapshots), and Figure 8.6c-d (10 snapshots), JKONET learns energy functionals J_ξ that can then model the ground truth trajectories. These trajectory-based dynamics are learned using the strong convexity regularizer ($\ell = 0.8$, see § 8.2.1).

COMPARISON TO FORWARD METHODS. Instead of parameterizing the next iteration $\rho_{t+1}(\xi)$ as we do in the JKONET formulation (8.1), the *forward* scheme states that the prediction at time $t+1$, η_{t+1}, can be obtained as $(\nabla F_\xi)_\# \eta_t(\xi)$, where F_ξ is any arbitrary neural network, as considered in Hashimoto et al. (2016), namely $\eta_0 := \mu_0$ and subsequently $\eta_{t+1}(\xi) := (\nabla F_\xi)_\# \eta_t(\xi)$. Although OT still plays an important role in that paper, since the potential F is estimated by minimizing a Sinkhorn loss $\overline{W}_\varepsilon(\eta_{t+1}, \mu_{t+1})$, as we do in (8.7), the forward displacement operator $(\nabla F_\xi)_\#$ has no spatial regularity. Because of that, we observe that the forward method can get more easily trapped in local minima, and, in particular, overfits the training data (see § **??**) as shown by a substantial decrease in performance in the presence of noise. We demonstrate this in different scenarios: First, we compare the robustness of both JKONET and the forward method to noise. For this, we corrupt 20% or 30% of the training data on the example of the semicircle trajectory with different levels of noise (see Fig. 8.3). We insist that noise is only added at training time, as random shifts on both feature dimensions, while we test on the original semicircle trajectory. In low noise regimes, where train and test data are similar, the forward method overfits and performs marginally better than JKONET (see Fig. 8.5c,d). As noise increases, the performance of the forward method deteriorates (Fig. 8.5b), while JKONET, constrained to move points with OT maps, is robust (Fig. 8.5a).

In a second experiment, we evaluate the capacity of JKONET and the forward method to extrapolate and generalize the learned trajectories, e.g., when vertically translating a line during test time (Fig. **??**). Due to

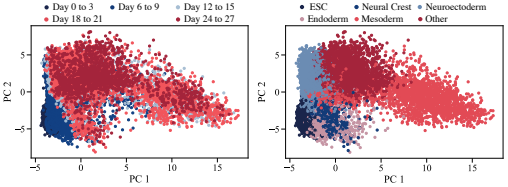
the less constrained energy, the *forward* method perfectly resembles the seen trajectory during training, but fails to extrapolate to shifted test data (Table **??** in § **??**).

Lastly, we compare the resulting energy functionals F_ξ and J_ξ of the forward method and JKONET, respectively, on the spiral trajectory (see Fig. 8.6). When learning long and complex population dynamics, teacher forcing improves training (see additional results in Fig. **??**c-d as well as Fig. 8.4c-d). While facilitating training of the forward method in some settings, it likewise results in wrong energy functionals F_ξ (Fig. 8.6a). JKONET, on the other hand, is able to globally learn the energy functional J_ξ, despite being only exposed to a one-step history of snapshots during training with teacher forcing (see Fig. 8.6c).

8.3.2 *Single-Cell Population Dynamics*

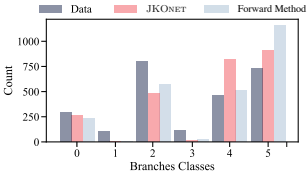
We investigate the ability of JKONET to predict the evolution of cellular and molecular processes through time. The advent of single cell profiling technologies has enabled the generation of high-resolution single-cell data, making it possible to profile individual cells at different states in the development. A key difficulty in learning the evolution of cell populations is that a cell is (usually) destroyed during a measurement. Thus, although one is able to collect features at the level of individual cells, the same cell cannot be measured twice. Instead, we collect independent samples at each snapshot, resulting in *unaligned* distributions across snapshots, without access to ground-truth single-cell trajectories. The goal of learning individual dynamics is to identify ancestor and descendant cells, and get a better understanding of biological differentiation or reprogramming mechanisms.

We apply JKONET to embryoid body single-cell RNA sequencing (scRNA-seq) data (Moon et al., 2019), describing the differentiation of human embryonic stem cells grown as embryoid bodies into diverse cell lineages over a period of 27 days. During this time, cells are collected at 5 different snapshots (day 1 to 3, day 6 to 9, day 12 to 15, day 18 to 21, day 24 to 27) and measured via scRNA-seq (resulting in 15,150 cells). For details on the dataset and data preprocessing see § **??**. We run JKONET as well as the baseline on the first 20 components of a principal component analysis (PCA) of the 4000 highly differentiable genes (see Fig. **??**). We split the dataset into train and test data ($\sim 15\%$) and parameterize both energy J_ξ and ICNN ψ_θ with linear layers ($\epsilon = 1.0$, $\tau = 1.0$, § **??**).

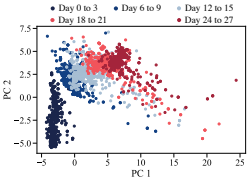


(a) PCA embedding of the embryoid body scRNA-seq data colored by the snapshot time.

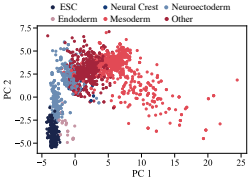
(b) PCA embedding of the embryoid body scRNA-seq data colored by the lineage branch class.



(c) Distribution of cell lineage branch classes in the data or predicted by JKONET or the forward method.



(d) PCA embedding of JKONET predictions colored by the snapshot time.



(e) PCA embedding of JKONET predictions colored by the lineage branch class.

Method	Cell Lineage Classification	
	ℓ_1	H^2
One-Step Ahead		
Forward Method	132.27 ± 5.00	0.026 ± 0.002
JKONET	$\mathbf{88.80 \pm 0.57}$	$\mathbf{0.016 \pm 0.001}$
All-Steps Ahead		
Forward Method	$\mathbf{185.47 \pm 12.18}$	0.033 ± 0.002
JKONET	215.60 ± 12.53	0.034 ± 0.004

TABLE 8.3: Evaluation of cell lineage branch classification performance of JKONET and the forward method on the embryoid body scRNA-seq data based on the ℓ_1-distance of the histograms and the Hellinger distance H^2 (8.8) of the predicted branch class distributions (using 3 runs).

CAPTURING SPATIO-TEMPORAL DYNAMICS. Given the samples from the cell population at day 1 to 3 (μ_0), JKONET learns the underlying spatio-temporal dynamics giving rise to the developmental evolution of embryonic stem cells. As no ground truth trajectories are available in the data, we use distributional distances, i.e., the entropy-regularized Wasserstein distance W_ε (1.2) (Flamary et al., 2021), to measure the correctness of the predictions at each time step. We hereby measure the W_ε discrepancy between data and predictions for one-step ahead as well as inference of the entire evolution (all-steps ahead) for each time step t_i, see results in Table 8.1. JKONET outperforms the forward method in terms of W_ε (1.2) distance for both one-step ahead and all-steps ahead predictions for all time steps. The performance of both methods is relatively stable even until day 24 to 27, i.e., the W_ε distance does not significantly grow for future snapshots. We

further visualize the first two principal components of the entire dataset (Fig. 8.7a) and of JKOnet's predictions on the test dataset ($\sim$ 500 cells per snapshot, Fig. 8.7d). Visualization of predictions of the forward method can be found in the Appendix (Fig. **??**a).

CAPTURING BIOLOGICAL HETEROGENEITY. Besides measuring the ability of JKOnet to model and predict the spatio-temporal dynamics of embryonic stem cells, we would like to guarantee, at a more macroscopic level, that JKOnet is also able to learn the cell's differentiation into various cell lineages. Embryoid bodies differentiation covers key aspects of early embryogenesis and thus captures the development of embryonic stem cells (ESC) into the mesoderm, endoderm, neuroectoderm, neural crest and others.

Following Moon et al. (2019, Fig. 6, Suppl. Note 4), we compute lineage branch classes (Fig. **??**c) for all cells based on an initial k-means clustering ($k = 30$) in a 10-dimensional embedding space using PHATE, a non-linear dimensionality reduction method capturing a denoised representation of both local and global structure of a dataset (Fig. **??**b). For details, see § **??**. We then train a k-nearest neighbor (k-NN) classifier ($k = 5$) to infer the lineage branch class based on a 20-dimensional PCA embedding of a cell (classes: ESC: 0, neural crest: 1, neuroectoderm: 2, endoderm: 3, mesoderm: 4, other: 5).

We analyze the captured lineage branch heterogeneity of the population predicted by JKOnet and the forward method by estimating the lineage branch class of each cell using the trained k-NN classifier. The predicted populations colored by the estimated lineage branch as well as the data with the true lineage branch labels are visualized in Figure 8.7e and Figure 8.7b, respectively. The corresponding predicted and true distributions of lineage branch classes are shown in Figure 8.7c. To quantify how well JKOnet and the forward method capture different cell lineage branches, we compute the ℓ_1 distance between the predicted and true histograms as well as the Hellinger distance

$$H^2(a,b) = \frac{1}{2}\sum_{i=1}^{k}\left(\sqrt{a_i/\|a\|_1} - \sqrt{b_i/\|b\|_1}\right)^2 \tag{8.8}$$

between both true and predicted class discrete distributions a and b. Figure 8.7c and Table 8.3 demonstrate that both, JKOnet and the forward method, capture most lineage branches during the differentiation of embryonic stem cells. Both methods, however, have difficulties recovering cells of

the neural crest (class 1) and the endoderm (class 3), lineage branches which are scarcely represented in the original data. The analysis further suggests that both methods reduce in performance w.r.t. biological heterogeneity when predicting the entire trajectory (all-steps ahead), instead of inferring the next snapshot only (one-step ahead).

8.4 CONCLUSION

We proposed JKONET, a model to infer and predict the evolution of population dynamics using a proximal optimal transport scheme, the JKO flow. JKONET solves local JKO steps using ICNNs and learns the energy that parameterizes these steps by fitting JKO flow predictions to observed trajectories using a fully differentiable bilevel optimization problem. We validate its effectiveness through experiments on synthetic potential- and trajectory-based population dynamics, and observe that it is far more robust to noise than a more direct Forward approach. We use JKONET to infer the developmental trajectories of human embryonic stem cells captured via high-dimensional and time-resolved single-cell RNAseq. Our analysis also shows that JKONET captures diverse cell fates during the incremental differentiation of embryonic cells into multiple lineage branches. Using proximal optimal transport to model real complex population dynamics thus makes for an exciting avenue of future work. Extensions could include modeling higher-order interactions among population particles in the energy function, e.g., cell-cell communication.

9

CONCLUSION AND FUTURE DIRECTIONS

It's odd the way life works, the way it mutates and wanders, the way one thing becomes another.
— Siri Hustvedt, *What I Loved* (2003)

In this work we propose CellOT, a framework to model single-cell perturbation responses from unpaired treated and untreated cell states using neural optimal transport. By adequately modeling the nature of the problem through the lens of optimal transport, CellOT determines how perturbations affect cellular properties, reconstructs the most likely trajectory single cells take upon perturbation, and subsequently assists in a better understanding of driving factors of cell fate decision and cellular evasion mechanisms. CellOT builds on the recent successes of optimal transport applications in single-cell biology (Schiebinger et al., 2019; Lavenant et al., 2021), by introducing a fully parameterized transport map that can be applied to incoming unseen samples. Previous methods (Jacob et al., 2018; Yang and Uhler, 2019; Prasad et al., 2020) rely on an unconstrained parameterization of the *primal* optimal transport map. However, the unconstrained nature of these models makes robust optimization challenging and results in reduced performance (Makkuva et al., 2020, Table 1). Instead, we learn the transformation of unperturbed to perturbed cell states through the *dual* optimal transport problem, parameterized via a pair of neural networks constrained to be convex (Makkuva et al., 2020). These constraints are important inductive biases that facilitate learning and result in a reliable and easy-to-train framework, as evidenced by the consistently strong performance of CellOT on several problems without the need for extensive hyperparameter tuning (see Online Methods).

CellOT infers the highly complex and nonlinear evolution of cell populations in response to perturbations without making strong simplifying assumptions on the nature of these dynamics. Unlike current approaches comprising autoencoder-based baselines (Lopez et al., 2018; Lotfollahi et al., 2019; Yang et al., 2020), CellOT does not necessarily rely on learning meaningful low-dimensional embeddings in which perturbations are modeled as linear shifts. We confirm this advantage through experiments on single-cell responses to different drugs in cancer cell lines obtained with RNA-seq

and spatially resolved 4i measurements, where CellOT consistently outperforms (Fig. **??** and **??**). Our evaluations went beyond the often-used average treatment effect and correlation analysis across all cells; we analyzed marginals and computed MMD scores, a strong measure of how well predicted and observed distributions match.

Using CellOT to perform cell-state-aware drug profiling enables us to quantify perturbation effects as a function of the underlying heterogeneity of the studied system, in our cases a co-culture of two melanoma cell lines with different sensitivities to drug treatments. In doing so, we *sharpen* the response profiles of the measured drugs and reveal cell-state-specific responses of multiple signaling pathway in relation to treatment history of the cell line donor. We find the signaling activity associated to the MEK and PI3k pathways to decouple in cells pre-exposed to MEK inhibitors, a known adaptation mechanism for therapy evasion in melanoma cells (Kun et al., 2021). This *pathway rewiring* is associated to alteration in the molecular feedback structure of cells from effectors to receptors (Kun et al., 2021; Turke et al., 2012). Thus, combining CellOT with a larger set of combination treatments, multiplexed imaging, and cellular systems reflective of disease adaptations may help us to elucidate the molecular mechanisms of signaling pathway evolution in the context of cancer therapy.

We further analyze how well the learned maps generalize beyond samples used for training (o.o.s. setting) and to different sample compositions (o.o.d. setting). In Fig. **??**, we therefore test CellOT's ability to to predict treatment responses in unseen lupus patients, infer developmental trajectories on stem cells of lower potency, and translate innate immune responses across patients. In all cases, CellOT's accuracy and precision are superior to current state-of-the-art methods (Fig. **??**). Moreover, the predicted cell states after perturbation are still very close to the actually observed cell states. We consider these results as particularly promising, as it illustrates that accurate o.o.s. and o.o.d. predictions are indeed possible.

The ability to make predictions out-of-distribution, such as on unseen patients, is, however, only feasible if a) similar samples have been observed in the unperturbed setting, and b) the training set contains cases that are similar not only in their unperturbed state but also their perturbation response. An analysis of glioblastoma patients treated with Panobinostat (Zhao et al., 2021) (see **??**a-c) indeed confirms this restriction: CellOT and the baselines are able to predict treatment outcomes for those patients that are similar to other patients in both unperturbed state as well as perturbation effect (see Fig. **??**f), but fail to capture perturbation effects for

patients that exhibit unique responses (see **??**g). This limitation is important to consider when applying CellOT in o.o.d. settings. To overcome such problems, larger cohorts, additional meta-information, and methodological extensions are required. Bunne et al. (2022) partially address this issue by deriving a neural optimal transport scheme that can be conditioned on a context, e.g., patient meta-data, when predicting perturbation responses.

We also observe that the predictive performance for CellOT drops when perturbations are too strong, i.e., the cell distributions before and after perturbations are very different (see Fig. **??**j); a similar drop is observed for the other methods (see **??**). The principle underlying the optimal transport theory is ideally suited for acute cellular perturbations during which single cells do not redistribute entirely and randomly in multidimensional measurement space, but typically only in a few dimensions, such that the overall correlation structure is preserved. While this modeling hypothesis is satisfied when perturbation responses are observed via regularly and frequently sampled snapshots, molecular transitions cannot be reconstructed when perturbation responses have progressed too far. For particularly strong or complicated perturbations, cellular multiplex profiles might change too drastically, violating OT assumptions and making it challenging to reconstruct the alignments between unperturbed and perturbed populations based on the *minimal effort* principle. In such settings, additional information is likely needed, for instance, a model of the underlying biology or models that integrate observations of multiple smaller time steps.

Despite the stochastic nature of cell fate decisions and the fact that cellular dynamics are intrinsically noisy (Wilkinson, 2009), CellOT models cell responses as deterministic trajectories. Approaches treating cell fate decisions as probabilistic events have previously allowed estimation of the full dynamical model to a greater extent than their deterministic counterparts (Bergen et al., 2020). By connecting OT and stochastic difference equations, recent work (Bunne et al., 2023; Somnath et al., 2023) can build up on CellOT to account for biological heteroscedasticity, at the cost of added model complexity and other simplifying assumptions.

Despite having provided a proof-of-concept of the capacity of CellOT to model various chemical perturbations for different data modalities through an in-depth analysis of the nature of the learned mapping as well as a demonstration of its versatility in a broad class of applications, CellOT's generalization capacity has been evaluated on relatively small datasets. Crucially, large cohorts comprised of patients with different molecular profiles, such as cancer patients with various underlying genetics, could

result in strongly heterogeneous treatment responses. It is evident that approaches addressing these challenges could readily exploit the upcoming availability of large-scale patient cohort studies. The use of neural optimal transport to learn single-cell drug responses makes thus for an exciting avenue for future work, including its use to improve our understanding of cell therapies, study drug responses from patient samples, and better account for cell-to-cell variability in large-scale drug design efforts.

BIBLIOGRAPHY

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks. *arXiv Preprint arXiv:2106.00774*, 2021.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

Brandon Amos, Lei Xu, and J Zico Kolter. Input Convex Neural Networks. In *International Conference on Machine Learning (ICML)*, volume 34, 2017.

Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented lagrangian approach to wasserstein gradient flows and applications. *ESAIM: Proceedings and surveys*, 54:1–17, 2016a.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge–Ampére operator. *Numerische Mathematik*, 134(3), 2016b.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12), 2020.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305, 1987.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Ratsch. Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *bioRxiv*, 2021.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised Training of Conditional Monge Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Charlotte Bunne, Ya-Ping Hsieh, Marci Cuturi, and Andreas Krause. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Martin Burger, José A. Carrillo, and Marie-Therese Wolfram. A mixed finite element method for nonlinear diffusion equations. *Kinetic & Related Models*, 3(1), 2010.

Luis A Caffarelli. Monotonicity Properties of Optimal Transportation and the FKG and Related Inequalities. *Communications in Mathematical Physics*, 214(3), 2000.

Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal Dual Methods for Wasserstein Gradient Flows. *Foundations of Computational Mathematics*, 2021.

Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations (ICLR)*, 2022.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal Control Via Neural Networks: A Convex Approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Review*, 63(2), 2021.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv Preprint arXiv:2201.12324*, 2022.

John M Danskin. *The Theory of Max-Min and its Applications to Weapons Allocation Problems*, volume 5. Springer, 1967.

Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic Deep Networks. In *International Conference on Machine Learning (ICML)*, volume 36, 2019.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3), 1982.

Harrison Edwards and Amos Storkey. Towards a Neural Statistician. In *International Conference on Learning Representations (ICLR)*, volume 5, 2017.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

Alessio Figalli. The Optimal Partial Transport Problem. *Archive for Rational Mechanics and Analysis*, 195(2), 2010.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22, 2021.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3):332–341, 2021.

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning Population-Level Diffusions with Generative Recurrent Networks. In *International Conference on Machine Learning (ICML)*, volume 33, 2016.

Lars Holdijk, Yuanqi Du, Ferry Hooft, Priyank Jaini, Bernd Ensing, and Max Welling. Path integral stochastic optimal control for sampling transition paths. *arXiv preprint arXiv:2207.02149*, 2022.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Leygonie Jacob, Jennifer She, Amjad Almahairi, Sai Rajeswar, and Aaron Courville. W2GAN: Recovering an Optimal Transport Map with a GAN. In *arXiv Preprint*, 2018.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1998.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

Peter E Kloeden and Eckhard Platen. Stochastic Differential Equations. In *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

E Kun, YTM Tsang, CW Ng, DM Gershenson, and KK Wong. MEK inhibitor resistance mechanisms and recent developments in combination trials. *Cancer Treatment Reviews*, 92:102137, 2021.

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

Guan-Horng Liu, Tianrong Chen, Oswin So, and Evangelos A Theodorou. Deep Generalized Schrödinger Bridge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12), 2018.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing Millions of Hyperparameters by Implicit Differentiation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8), 2019.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning (ICML)*, volume 37, 2020.

Robert J McCann. A Convexity Principle for Interacting Gases. *Advances in Mathematics*, 128(1), 1997.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-Scale Wasserstein Gradient Flows. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

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

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Gabriel Peyré. Entropic Approximation of Wasserstein Gradient Flows. *SIAM Journal on Imaging Sciences*, 8(4), 2015.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

Neha Prasad, Karren Yang, and Caroline Uhler. Optimal Transport using GANs for Lineage Tracing. *arXiv preprint arXiv:2007.12098*, 2020.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, 2017.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs Using Optimal Transport. In *International Conference on Learning Representations (ICLR)*, 2018.

Filippo Santambrogio. Optimal Transport for Applied Mathematicians. *Birkhäuser, NY*, 55(58-63):94, 2015.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1), 2017.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4), 2019.

Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned Diffusion Schrödinger Bridges. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics. In *International Conference on Machine Learning (ICML)*, 2020.

Alexa B Turke, Youngchul Song, Carlotta Costa, Rebecca Cook, Carlos L Arteaga, John M Asara, and Jeffrey A Engelman. MEK inhibition leads to PI3K/AKT activation by relieving a negative feedback on ERBB receptors. *Cancer Research*, 72(13), 2012.

Francisco Vargas, Pierre Thodoroff, Neil D Lawrence, and Austen Lamacraft. Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9), 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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

Darren J Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2), 2009.

Ronald J Williams and David Zipser. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2), 1989.

Karren D Yang and Caroline Uhler. Scalable Unbalanced Optimal Transport using Generative Adversarial Networks. *International Conference on Learning Representations (ICLR)*, 2019.

Karren Dai Yang, Karthik Damodaran, Saradha Venkatachalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS Computational Biology*, 16(4), 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Wenting Zhao, Athanassios Dovas, Eleonora Francesca Spinazzi, Hanna Mendes Levitin, Matei Alexandru Banu, Pavan Upadhyayula, Tejaswi Sudhakar, Tamara Marie, Marc L Otten, Michael B Sisti, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Medicine*, 13(1), 2021.

CURRICULUM VITAE

PERSONAL DATA

Name	Charlotte Bunne
Date of Birth	August 29, 1995
Place of Birth	Karlsruhe, Germany
Citizen of	Germany

EDUCATION

2022-2023	Broad Institute of MIT and Harvard, Cambridge (MA), USA Visiting Graduate Student
2016 – 2019	Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland *Final Degree:* Master of Science
2018-2019	Massachusetts Institute of Technology (MIT), Cambridge (MA), USA Visiting Student
2013 – 2016	Heidelberg University Heidelberg, Germany *Final Degree:* Bachelor of Science

EMPLOYMENT

2022	Research Intern *Apple*, Paris, France
2020	Research Intern *Google Research*, Zürich, Switzerland
2017	Research Intern *IBM Research*, Zürich, Switzerland

PUBLICATIONS

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised Training of Conditional Monge Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A

APPENDIX

Here be dragons.