Optimal Transport in Learning, Control, and Dynamical Systems

Charlotte Bunne

ICML Tutorial 2023

Optimal transport (OT) theory (Santambrogio, 2015; Villani, 2003, 2009) is a core element of the machine learning toolbox and has become within a few years the go-to framework to analyze, model, and solve an ever-increasing variety of tasks involving probability measures. This is best exemplified by its increasing importance to fitting generative models, where the goal is to learn a map (Arjovsky et al., 2017; Genevay et al., 2018; Salimans et al., 2018), or more generally a diffusion (Song et al., 2021; De Bortoli et al., 2021) to morph a simple measure (e.g., Gaussian) onto a data distribution of interest (e.g., images). This is also apparent in the many applications that use OT to align probability measures that have since arisen, e.g., to transfer label knowledge between datasets (Flamary et al., 2016; Singh and Jaggi, 2020), to analyze sampling schemes (Dalalyan, 2017), or study population trajectories (Schiebinger et al., 2019; Bunne et al., 2023b).

In this tutorial, we primarily cast light on the static and dynamic formulations of optimal transport, and simultaneously establish their theoretical nexus by recalling its mathematical history from Monge and Kantorovich to modern Fields Medal winners Villani, Figalli, and Abel and Wolf Prize recipient Caffarelli in order to provide a solid foundation for the discussion ahead.

Contents

Notation

Σ_n	probability simplex of size n.
(μ, ν)	measures defined on spaces $(\mathcal{X}, \mathcal{Y})$.
(u, v)	histograms in the simplices $\Sigma_n \times \Sigma_m$.
$p_\mu = \frac{d\mu}{dx}$	density with respect to the Lebesgues measure.
$(\mu = \sum_i u_i \delta_{x_i}$, $\nu = \sum_j v_j \delta_{y_j})$	discrete measures defined on spaces $(x_1, \dots, x_n \in \mathcal{X}, y_1, \dots, y_m \in \mathcal{Y})$.
$\mathbf{1}$	matrix of $\mathbb{R}^{n \times m}$ with all entries identically set to 1.
Id	identity map.
$c(x, y)$	ground cost, with associated pairwise cost matrix $(c(x_i, y_j))_{ij}$ evaluated on the support of μ, ν.
$\|\cdot\|_2^2$	squared Euclidean norm.
$T_\sharp$	pushforward operator.
$T : \mathcal{X} \times \mathcal{Y}$	Monge map, typically such that $T_\sharp \mu = \nu$.
π	coupling measure between μ and ν, for discrete measures $\pi = \sum_{ij} \mathbf{P}_{ij} \delta_{(x_i, y_j)}$.
$\Pi(\mu, \nu)$	set of couplings, for discrete measures $U(u, v)$.
$\mathrm{supp}(\pi)$	support of π.
$W(\mu, \nu)$	Wasserstein distance between measures μ and ν.
$H(\pi)$	entropy of coupling π.
ε	regularization strength of the entropy regularization.
(f, g)	dual potentials.
f^*	Legendre transform of function f.
$f^\star$	optimum of function f.
φ	convex potential.
$(\mu_t)_{t=0}^T$	dynamic measures with $\mu_{t=0} = \mu_0$ and $\mu_{t=T} = \mu_T$.
v	speed in the dynamic Benamou-Brenier or control in the stochastic optimal control formulation.
Δ	Laplace operator.
τ	step size.
$\langle \cdot, \cdot \rangle$	Euclidean dot-product between vectors.
D_{KL}	Kullback-Leibler divergence.
σ	noise level.
$\mathbb{P}_t$	stochastic process with $t \in [0, 1]$.
$\mathbb{W}_t$	standard Wiener process.
$\mathbb{Q}_t$	reference process, e.g., Brownian motion.
$Z_t, \hat{Z}_t$	time-indexed smooth vector fields indicating the forward and backward policy.
θ and ϕ	parameters of neural networks.
ℓ	loss function.

1 Static Optimal Transport

Optimal transport takes dual roles as it induces a mathematically well-characterized distance measure between distributions as well as provides a geometry-based approach to realize mappings between two probability distributions. In this section, we introduce the mathematical foundations of the **static** OT problem. Further, we provide an extended analysis of the Monge map, which gives an actionable way to transform from one probability distribution to another. We conclude with a complete proof of the celebrated Brenier theorem. This quintessential result and its particularization to translation-invariant costs lay the foundation of the flurry of neural approaches proposed in the literature. This includes modeling Monge maps as gradients of convex functions parameterized through input convex neural networks (ICNNs) (Amos et al., 2017; Huang et al., 2021; Makkuva et al., 2020; Korotin et al., 2021b; Lübeck et al., 2022; Bunne et al., 2022a), i.e., approaches that are a direct consequence of the Brenier theorem, regularizers (Uscidda and Cuturi, 2023), amortized optimization (Amos, 2023; Amos et al., 2023), or entropic maps (Pooladian and Niles-Weed, 2021; Pooladian et al., 2023b; Divol et al., 2022; Cuturi et al., 2023).

1.1 Monge Problem

In the 18th century "Mémoire sur la théorie des déblais et des remblais", Gaspard Monge sets out to solve what is now known as the Monge problem, posing a seemingly simple, yet fundamentally complex question: Given two quantities of mass located at two different sites, what is the most efficient way to transport one into the other? In more formal terms, provided with two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, here restricted to measures supported on $\mathbb{R}^d$, Monge's initial approach was to find a map T that pushes one mass onto the other in a way that minimizes the total cost of transport. Given a measurable const function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the Monge problem then reads

$$T^\star := \arg \inf_{T\sharp\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x)) d\mu(x) \,, \tag{1}$$

where $T_\sharp$ defines the pushforward operator that "moves" an entire probability measure on $\mathcal{X}$ towards a new probability measure on $\mathcal{Y}$, i.e., $T_\sharp$ "pushes forward" each elementary mass of a measure μ on $\mathcal{X}$ by applying the map T to obtain then an elementary mass in $\mathcal{Y}$. For two discrete measures $\mu = \sum_{i=1}^n u_i \delta_{x_i}, \nu = \sum_{j=1}^m v_j \delta_{y_j}$, it seeks a transport map $T : \mathcal{X} \to \mathcal{Y}$ associating each source point x_i to a target point y_j (see Fig. 1a for the discrete and Fig. 1b for the continuous setting). The existence of $T^\star$ is guaranteed under fairly general conditions (Santambrogio, 2015, Theorem 1.22), which require that μ and ν have finite ℓ_2 norm, and that μ puts no mass on $(d-1)$ surfaces of class $\mathcal{C}_2$, i.e., the family of continuous functions that have both a continuous first and a continuous second derivative.

1.2 Kantorovich Relaxation

It was not until the 20th century, however, that the concept found a more tractable development. In 1942, Leonid Kantorovich provided a relaxation to this non-convex and difficult-to-solve problem. Instead of the deterministic matching proposed by Monge,

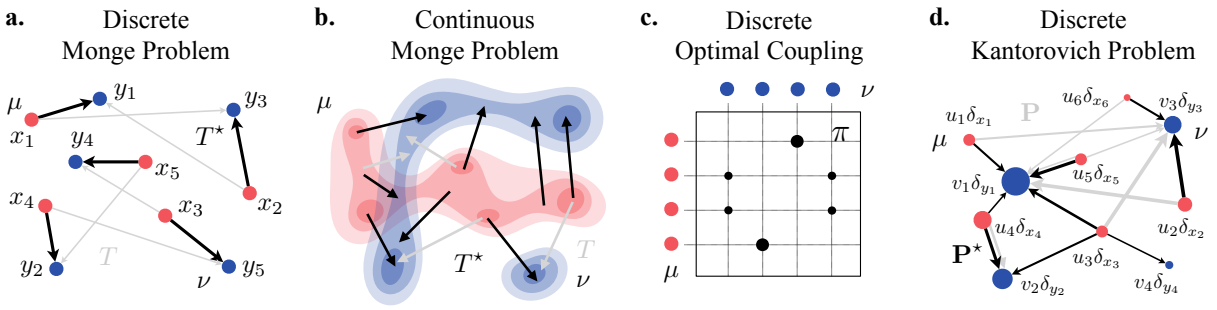

Figure 1: Overview on different formulations of the static OT problem for discrete and continuous measures. Monge map for **a.** discrete and **b.** continuous measures μ, ν. The optimal map $T^\star$ minimizes (1). **c.** Optimal coupling π (2) for discrete measures μ and ν. **d.** Mass splitting principle of the Kantorovich relaxation for discrete measures μ and ν of the optimal transport plan $\mathbf{P}^\star$ and a non-optimal plan $\mathbf{P}$. Figure adapted from Peyré and Cuturi (2019).

Kantorovich considered probabilistic correspondences that allow for the transportation of mass from a single source point to various target points (mass splitting), resulting in the problem formulation

$$W(\mu,\nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \iint c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y, \tag{2}$$

where $\Pi(\mu,\nu) \coloneqq \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : P_{\mathcal{X}\sharp}\pi = \mu \text{ and } P_{\mathcal{Y}\sharp}\pi = \nu\}$ is the set of couplings on $\mathbb{R}^d \times \mathbb{R}^d$ with respective marginals μ, ν. Here, $P_{\mathcal{X}\sharp}$ and $P_{\mathcal{Y}\sharp}$ are the pushforward by the projections $P_{\mathcal{X}}(x,y) = x$ and $P_{\mathcal{Y}}(x,y) = y$ Given the optimal transport coupling π, the resulting distance $W(\mu,\nu)$ between μ and ν is known as the Wasserstein distance. A visualization of the discrete setting is provided in Fig. 1c.

For his work, Kantorovich received the Nobel Prize in economics. The connection of OT to basic questions in economy becomes clear when interpreting μ as a density of resource units, and ν a density of factories, where the coupling π denotes the optimal transportation plan of distributing resources to factories.

Despite its elegance, the Wasserstein distance (2) presents a computationally challenging optimization problem. A partial remedy proposed by Cuturi (2013) is to solve regularized optimal transportation problems for an approximate solution. One example of an effective regularization is entropy regularization: For $\varepsilon \geq 0$, set

$$W_\varepsilon(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \iint c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y \; - \varepsilon H(\pi), \tag{3}$$

where $H(\pi) \coloneqq -\iint \pi(x,y)\log\pi(x,y)\mathrm{d}x\mathrm{d}y$ is the entropy of coupling π. Notice that the definition above reduces to the usual Wasserstein distance (2) when $\varepsilon = 0$. When instantiated on finite discrete measures, such as $\mu = \sum_{i=1}^n u_i\delta_{x_i}$ and $\nu = \sum_{j=1}^m v_j\delta_{y_j}$, (2) translates to a linear program

$$W_\varepsilon(\mu,\nu) \coloneqq \min_{\mathbf{P}\in U(\mu,\nu)} \langle \mathbf{P}, [\|x_i - y_j\|^2]_{ij}\rangle \; - \varepsilon H(\mathbf{P}), \tag{4}$$

where $H(\mathbf{P}) \coloneqq -\sum_{ij}\mathbf{P}_{ij}(\log\mathbf{P}_{ij} - 1)$ and the polytope $U(\mu,\nu)$ is the set of $n \times m$ matrices $\{\mathbf{P} \in \mathbb{R}_+^{n\times m}, \mathbf{P}\mathbf{1}_m = \mu, \mathbf{P}^T\mathbf{1}_n = \nu\}$. Regularization with an entropy term results in a significantly more efficient optimization (Cuturi, 2013) and differentiability

w.r.t. the inputs. As a consequence, 4 is commonly used as a loss function or evaluation metric in machine learning applications, e.g., for structured prediction (Frogner et al., 2015; Janati et al., 2020) or generative model fitting (Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018). While setting $\varepsilon > 0$ yields a faster and differentiable proxy to approximate W_0, it introduces a bias, since $W_\varepsilon(\mu, \mu) \neq 0$ in general.

1.3 Kantorovich Duality

The Kantorovich formulation (2) is a *convex* problem on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and thus admits a dual formulation introduced by Kantorovich (1942), i.e., a constrained concave maximization problem defined as

$$W(\mu, \nu) \coloneqq \sup_{(f,g)\in\Phi_c} \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}v, \tag{5}$$

where the set of admissible dual potentials is given by $\Phi_c \coloneqq \{(f, g) \in L^1(\mu) \times L^1(\nu) : f(x) + g(y) \leq c(x, y), \forall (x, y)\, d\mu \otimes d\nu \text{ a.e.}\}$. (f, g) is thus a pair of continuous functions, often referred to as *Kantorovich potentials*. An informal interpretation of (5) was provided by Caffarelli (2003), revisiting the connection of OT to economics: A logistics company is concerned with transporting products from each resource unit x to a factory y. The transportation company charges $f(x)$ for loading resources at point x and $g(y)$ for unloading it at destination y but is constrained to charge $f(x) + g(y) \leq c(x, y)$. In order to arrange prizes f and g that increase profit, they thus maximize objective (5).

The Kantorovich duality (5) is a core pillar of optimal transport, powerful due to its generality and computationally attractive as it is easier to store two functions (f, g) than an entire coupling π. In the following, we will introduce the concept of c-transforms, a useful machinery to reduce (5) even further into an optimization problem over only one instead of two dual potentials.

Definition 1 (c-transform)**.** *c-transforms (also called c-conjugate functions) are generalizations of the Legendre transform from convex analysis defined as*

$$\forall y \in \mathcal{Y}, \quad f^c(y) \coloneqq \inf_{x\in\mathcal{X}} c(x, y) - f(x)\,. \tag{c-transform}$$

The definition of f^c is also often referred to as a "Hopf-Lax formula". Similarly to the c-transform of $f : \mathcal{X} \to \mathbb{R}$, we can define the $\bar{c}$-transform of $g : \mathcal{Y} \to \bar{\mathbb{R}}$ by

$$\forall x \in \mathcal{X}, \quad g^{\bar{c}}(y) \coloneqq \inf_{y\in\mathcal{Y}} c(x, y) - g(y)\,,$$

where $\bar{c}(y, x) = c(x, y)$.

Remark 1. *As well-known in convex analysis (Rockafellar, 1970), the c-transform is a generalization of the Legendre transform. More precisely, for function $f : \mathbb{R}^d \to \mathbb{R}$ its Legendre transform is defined as*

$$f^*(y) \coloneqq \sup_{x\in\mathbb{R}^d} \langle y, x\rangle - f(x) \tag{Legendre transform}$$

The c-transform corresponds to this notion by considering $c(x, y) = \langle x, y\rangle$ (up to a change of sign).

Further, f is a $\bar{c}$-*concave* function if there exists a g such that $f = g^{\bar{c}}$, and analogously, a function g is said to be c-*concave* if there is a function f such that $g = f^c$. When $\mathcal{X} = \mathcal{Y}$ and c is symmetric, no distinction between c and $\bar{c}$ is necessary.

Using the concept of c-transforms, we can reduce (5) to a single potential: Assume we keep dual potential f fixed and given the constraint of the dual formulation (5)

$$\begin{aligned} f(x) + g(y) &\leq c(x, y) \\ g(y) &\leq c(x, y) - f(x)\,, \end{aligned}$$

we can see that the "best" potential g is given by the c-transform of f

$$g(y) \leq \inf_x c(x, y) - f(x).$$

Then, doing an alternate optimization on either f or g, we replace the dual potentials (f, g) with (f, f^c), and then $(f^{c\bar{c}}, f^c)$, whilst preserving the constraints and increasing the value of the integrals of (5). Although one could continue this alternate optimization further, the invariance property $f^{c\bar{c}c} = f^c$ for any f shows that one can only "improve" once the dual potential using c-transforms, resulting in the semi-dual formulation of optimal transport

$$f^\star := \arg \max_{f\ c\text{-concave}} \int f \mathrm{d}\mu + \int f^c \mathrm{d}\nu\,, \tag{6}$$

where $f^\star$ is the optimal dual function and c-concave.

1.4 Geometry of Optimal Transport

Following Gangbo and McCann (1996), $f^\star$ can be linked to the optimal transport map $T^\star$ via the following result:

Theorem 1 (Gangbo-McCann Theorem)**.** *Given a cost function c, the relationship between the optimal transport map $T^\star : \mathcal{X} \to \mathcal{X}$ and the c-concave function $f^\star$ denoting the optimal dual potential is given by the expression*

$$T^\star(x) = \nabla_x c(x, \cdot)^{-1} \circ \nabla f^\star(x)\,. \tag{7}$$

Thus, map T depends explicitly on the gradient of the cost, or rather on its inverse map (Gangbo and McCann, 1995).

Following Gangbo and McCann (1996) and considering translation-invariant costs[1] generated by a convex potential $h : \mathbb{R}^d \to \mathbb{R}$, i.e., c(x, y)=h(x-y), this reduces to

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)\,, \tag{8}$$

where h^* is the Legendre transform of h given by

$$\forall z, \quad h^*(z) := \sup_x \langle x, z\rangle - h(x)\,. \tag{9}$$

[1] A cost is translation-invariant if $c(x, y) = h(x - y)$ for $h(z) = h(-z)$.

Proof. Santambrogio (2015, Theorem 1.39) proves that the solutions of (2) and (5) are equivalent, i.e.,

$$\int_{\mathcal{X}} f \mathrm{d}\mu + \int_{\mathcal{Y}} f^c \, \mathrm{d}\nu = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) + f^c(y)) \, \mathrm{d}\pi(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y) \, .$$

Then, a point (x_0, y_0) in the coupling π, i.e., $(x_0, y_0) \in \mathrm{supp}(\pi)$, necessarily satisfies the constraint of the dual problem (5)

$$\pi^\star (x_0, y_0) > 0 \Leftrightarrow f^\star (x_0) + g^\star (y_0) = c (x_0, y_0) \, .$$

Replacing g by the c-transform of f, we have

$$\begin{aligned} &\Leftrightarrow f^\star (x_0) + f^{c\star} (y_0) = c (x_0, y_0) \\ &\Leftrightarrow f^{c\star} (y_0) = c (x_0, y_0) - f^\star (x_0) \, . \end{aligned}$$

Yet, by definition of the c-transform, $f^{c\star}$ is given by

$$\Leftrightarrow f^{c\star} (y_0) = \inf_z c (z, y_0) - f^\star (z) \, .$$

Thus, x_0 is a minimizer of the above expression and $\nabla_{x_0} c (x_0, y_0) = \nabla f^\star (x_0)$. Therefore, after inversion, we have $y_0 = \nabla_x c(x, \cdot)^{-1} \circ \nabla f^\star(x)$ and

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x) \, .$$

Applying this result to $c(x, y) = h(x - y)$, we get

$$\begin{aligned} \nabla_x c(x,y) &= \nabla h(x-y) \\ \nabla_x c(x,\cdot) &= \nabla h(x-\cdot) \\ \nabla_x c(x,\cdot)^{-1} &= x - (\nabla h)^{-1}(\cdot) \, . \end{aligned}$$

Note that $(\nabla h)^{-1}$ is equivalent to ∇h^* with the convex conjugate h^* and thus,

$$\nabla_x c(x,\cdot)^{-1} = x - \nabla h^*(\cdot) \, .$$

□

Alternative formulations that relate the Kantorovich setting (with general costs) (5) to that of Monge (1) were proposed by Rüschendorf (1991a,b); Caffarelli (1996).

The case of the squared Euclidean cost[2] $c(x, y) = \frac{1}{2}\|x - y\|^2$ in $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ deserves a special attention. Taking advantage of the particular form of the quadratic cost function, we can expand the constraint of (5) such that

$$f(x) + g(y) \leq \frac{1}{2}\|x-y\|_2^2 \Longleftrightarrow \left[\frac{1}{2}\|x\|_2^2 - f(x)\right] + \left[\frac{1}{2}\|y\|_2^2 - g(y)\right] \geq \langle x, y\rangle$$

and subsequently reparameterize $\varphi(x) \coloneqq \frac{1}{2}\|x\|_2^2 - f(x)$ and $\psi(y) \coloneqq \frac{1}{2}\|y\|_2^2 - g(y)$. Mirroring the same logic as for the c-transform, we derive the semi-dual in the Euclidean setting

$$\inf_{\varphi \text{ convex}} \int \varphi \mathrm{d}\mu + \int \varphi^* \mathrm{d}\nu \, . \tag{10}$$

[2]For elegance, we consider $c(x, y) = \frac{1}{2}\|x - y\|^2$ instead of $c(x, y) = \|x - y\|^2$.

Following the double convexification trick as outlined in Villani (2003, Lemma 2.10), we see that applying the Legendre transform twice yields function pair $(\varphi^{**}, \varphi^{*})$. As each of them is defined as the supremum of a family of linear functions, the result is an optimization problem over two convex lower semi-continuous (l.s.c.) functions.

Similarly, for the special case of the squared Euclidean distance the Gangbo-McCann Theorem (8) with $h = \frac{1}{2}\|\cdot\|_2^2$ implies that $\nabla h = \nabla h^* = \mathrm{Id}$, and thus

$$T(x) = x - \nabla f(x) = \nabla\left(\frac{1}{2}\|x\|_2^2 - f(x)\right)(x) = \nabla\varphi(x),$$

where we again reparameterize $\varphi(x) \coloneqq \frac{1}{2}\|x\|_2^2 - f(x)$ and $\varphi(x) = \frac{1}{2}\|x\|_2^2 - f(x)$ can be shown to be convex.

This connection presents a well-known fact that has been investigated first by Brenier (1987, 1991), establishing for the special case of the Euclidean distance the equivalence of the Monge (1) and Kantorovich formulation (2), the uniqueness of the optimal coupling π, and instantiating that there must exist a unique (up to the addition of a constant) potential $\varphi^\star : \mathbb{R}^d \to \mathbb{R}$ such that $T^\star = \nabla\varphi^\star$. This theorem has far-reaching implications: When seeking optimal transport maps, it is sufficient, to restrict the computational effort to seek a "good" convex potential φ, such that its gradient pushes μ towards ν. Let us state the celebrated Brenier theorem (1987) in more formal terms:

Theorem 2 (Brenier Theorem). *In the setting where both $\mathcal{X}$ and $\mathcal{Y}$ are equal to $\mathbb{R}^d$, and the cost function $c(x, y) = \|x - y\|^2$ is employed, and at least one of the two input measures μ possesses a density p_μ in relation to the Lebesgue measure, then there exists a unique optimal solution π in the Kantorovich formulation (2). This solution is exclusively supported on the graph $(x, T(x))$ of Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$. In other terms, we can express π as $(\mathrm{Id}, T)_\sharp\alpha$, meaning that for any function h belonging to the set $\forall h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$, the following equality holds*

$$\int_{\mathcal{X}\times\mathcal{Y}} h(x, y)\mathrm{d}\pi(x, y) = \int_{\mathcal{X}} h(x, T(x))\mathrm{d}\mu(x).$$

Moreover, this map T is uniquely determined by the gradient of a convex function φ, denoted as $T(x) = \nabla\varphi(x)$. The function φ is the unique convex function, up to an additional constant, for which $(\nabla\varphi)_\sharp\mu = \nu$.

Corollary 1. *Under the assumption of the Brenier Theorem, $\nabla\varphi$ is the unique solution to the Monge transportation problem (1), i.e.,*

$$\int_{\mathbb{R}^d} \|x - \nabla\varphi(x)\|^2 \mathrm{d}\mu(x) = \inf_{T_\sharp\mu=\nu} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\mu(x). \tag{11}$$

The Brenier Theorem has been exploited to propose neural OT solvers (Taghvaei and Jalali, 2019; Makkuva et al., 2020; Korotin et al., 2021a; Bunne et al., 2022b; Alvarez-Melis et al., 2022; Mokrov et al., 2021; Amos, 2023), proving its essential nature in multiple instances and modern developments of optimal transport. Further, it presents an elegant way to solve the Monge problem in a geometric sense and has profound implications for the dynamic version of the problem, which we will study next.

2 Dynamic Optimal Transport

We have hitherto engaged with the *static* optimal transport problem, establishing a solid foundation upon which to build more desirable dynamic formulations. In fact,

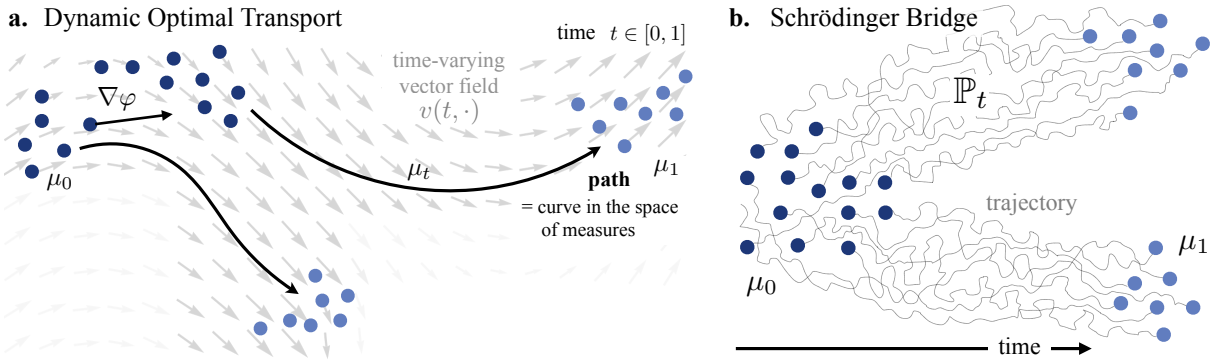

Figure 2: Overview on different formulations of the dynamic OT problem. a. We can model the evolution of a measure μ_t over time as minimal path on a time-varying vector field $v(t, \cdot)$ or according to the gradient of a convex potential $\nabla\varphi$. **b.** Alternatively, taking a stochastic perspective, we can study the dynamic formulation of the entropy-regularized OT problem (3) and find a stochastic process $\mathbb{P}_t$ that describes the particle dynamics from μ_0 to μ_1.

the roots of these dynamic formulations are embedded within the static OT framework: As posited by Benamou and Brenier (2000), the dynamic formulation "was already implicitly contained in the original problem addressed by Monge", where "eliminating the time variable was just a clever way of reducing the dimension of the problem." When reintroducing time in the dynamic version, the optimal transport map becomes a time-dependent flow capable of describing the evolution of a measure over time.

In this section, we will cover several perspectives and frameworks of the **dynamic** OT problem: As mentioned earlier, the Brenier theorem forms a critical bridge that connects the static and dynamic formulation, perpetuated in the Monge-Ampère equation. Further, Benamou and Brenier (2000) introduce how the dynamic point of view offers an alternate and intuitive interpretation of optimal transport with links to fluid dynamics. The resulting framework surprisingly leads to a convex optimization problem that can be parameterized through continuous normalizing flows (NF) (Tong et al., 2020; Chen et al., 2018) or flow matching frameworks (Lipman et al., 2023; Liu et al., 2022b; Pooladian et al., 2023a; Albergo et al., 2023). We further highlight the connections of OT to partial differential equations (PDEs) such as Fokker-Planck-like equations through the Jordan, Kinderlehrer, and Otto scheme. Lastly, moving beyond PDEs and taking a stochastic control perspective, we will introduce the notion of the Schrödinger bridge (SB) problem.

2.1 Monge-Ampère Equation

As a direct consequence of the Brenier Theorem, if $T(x) = \nabla\varphi(x)$, φ being smooth and strictly convex, and μ and ν absolutely continuous with densities p_μ and p_ν, we can express $T_\sharp\mu = \nu$ in a nonlinear PDE form. More concretely, as a consequence of a simple change-of-variable computation, φ is a solution of the Monge-Ampère equation that reads

$$\det\left(\partial^2\varphi(x)\right) p_\nu(\nabla\varphi(x)) = p_\mu(x), \tag{12}$$

where $\partial^2\varphi(x) \in \mathbb{R}^{d\times d}$ is the Hessian of φ, describing the continuous evolution from μ to ν. First studied by Monge in 1781 and later by Ampère in 1819, this nonlinear

partial differential equation arises in several problems from analysis to geometry, for example, in the Weyl and Minkowski problems in differential geometry of surfaces. The regularity of the solutions of (12), with implications on regularity results of the optimal transport map T, has been subject of a series of works by Caffarelli in the 1990s, for which he was awarded the Abel Prize in 2023, as well as more recently by Figalli, recognized with the Fields Medal in 2018.

2.2 Benamou-Brenier Formulation

Avoiding solving (12) directly, Benamou and Brenier (2000) introduce an alternative numerical framework by connecting the optimal mass transfer problem to continuum mechanic frameworks. Deviating from the previous notation of (μ, ν), in the following sections we study the dynamic problem via the evolution from measure μ_0 at time $t = 0$ to μ_1 at $t = 1$. In the setting $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ with the squared Euclidean cost $c(x, y) = \|x - y\|^2$, the solution of (2) then coincides with finding the minimal path $(\mu_t)_{t=0}^1$, or more concretely, a curve in the space of measures, minimizing a total length. Such path μ_t can be described through a time-varying vector field $v(t, \cdot)$ which moves particles around, satisfying the continuity equation in fluid dynamics or conservation of mass formula

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v) = 0, \qquad \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1 \,, \tag{13}$$

where the vector field $v(t, \cdot)$ denotes the speed and $\mu_t v(t, \cdot) = J_t$ corresponds to the momentum. Reformulating the optimal transportation problem in a differential way, an "Eulerian" formulation inspired by fluid mechanics, will be crucial for the subsequent study of dynamical problems. Every curve μ_t describing the evolution of the measure over time can be interpreted as the fluid flow along a family of vector fields. We are searching for the vector field $v(t, \cdot)$ that (i.) satisfies the conservation of mass (13), and (ii.) minimizes the kinetic energy of the path. The infinitesimal length of such a vector field can be computed via

$$\|v\|_{\ell^2(\mu_t)} = \left(\int_{\mathbb{R}^d} \|v(t, x)\|^2 \, \mathrm{d}\mu_t(x) \right)^{1/2} . \tag{14}$$

resulting, in the case of $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$, in the minimal-path reformulation of (2)

$$\begin{aligned} W(\mu_0, \mu_1) = \inf_{(\mu_t, v)} & \int_0^1 \int_{\mathbb{R}^n} \frac{1}{2} \|v(t, x)\|^2 d\mu_t(x) dt \\ & \frac{\partial \mu_t}{\partial t} + \nabla \cdot (v \mu_t) = 0 \\ & \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1 \,. \end{aligned} \tag{15}$$

Thus, path μ_t describes the time-evolving density of a set of particles moving continuously with velocity $v(t, \cdot)$. Taking the perspective of fluid dynamics, (14) can also be interpreted as the *kinetic energy* of the particles. The Benamou-Brenier formulation (15) then selects the vector field v that minimizes the total efforts or the total kinetic energy one has to spend in order to move particles around according to the vector field v.

A particularly important case occurs when there exists an optimal Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$ with $T_\sharp \mu_0 = \mu_1$ (see Brenier Theorem): The solution of the time-dependent OT problem (15) then coincides with McCann's displacement interpolation between two measures. Reciting the Brenier Theorem, with $T = \nabla\varphi$, μ_t is equal to McCann's interpolation between μ_0 and μ_1 given by

$$\mu_t = [(1-t)I + t\nabla\varphi]_\sharp \mu_0 = [(1-t)I + tT]_\sharp \mu_0 \,. \tag{16}$$

Despite their simplicity, this concept possesses remarkable applications beyond the realm of optimal transport (Bonneel et al., 2011). In particular, its interpretation as a geodesic formula in Riemannian geometry is discussed in Otto (2001) and serves as a pivotal link to the subsequent discussion.

2.3 Jordan-Kinderlehrer-Otto Flows

The time-dependent Benamou-Brenier formulation (15) not only provides us with a more complete description of optimal transport but also the discovery that the resulting path $(\mu_t)_{t=0}^1$ may be seen as a constant-speed geodesic interpolating between population μ_0 and μ_1 in the space of measures, i.e., a Wasserstein geodesic. When studying dynamic processes in biomedicine, however, phenomena such as cellular differentiation in developmental processes, tissue formation, or cell migration involve intricate spatiotemporal dynamics that cannot be adequately captured by solely studying the interpolation between two measures μ_0 and μ_1. Instead, many phenomena in biology and physics can be modeled through an energy functional J such that the minimization of J describes the observed dynamics of the studied system —a concept known as gradient flows. At their core, gradient flows provide a powerful framework for understanding the evolution of functions or systems toward an optimal state through the direction of the steepest descent of a function J. More concretely, gradient flows capture the intuitive idea of objects moving in a direction that decreases their energy, seeking a state of minimum potential or maximum stability. In the following, we will study gradient flows in the Euclidean setting before considering generalizations to arbitrary measures that allow studying the evolution of populations over time.

Euclidean case. Consider the evolution of a vector x over time in Euclidean space. Provided with a smooth functional J, this can be realized through the standard gradient descent (forward) scheme

$$x_{t+1} \coloneqq x_t - \tau \nabla J\left(x_t\right),$$

where τ is the step size. The resulting sequence $x_0, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T$ then describes the trajectory of a single particle x over time. For non-smooth functions, one can resort to a proximal scheme, i.e.,

$$x_{t+1} \coloneqq \operatorname{Prox}_{\tau J}^{\|\cdot\|}\left(x_t\right) \coloneqq \operatorname*{argmin}_x \frac{1}{2\tau}\left\|x - x_t\right\|^2 + J(x).$$

The proximal scheme can thus be seen as a *backward* Euler discretization of the gradient flow.

Table 2: Equivalence between gradient flows and PDEs where the gradient flow of flow functional $J(\mu_t)$ in Wasserstein space satisfies the corresponding PDE (Alvarez-Melis et al., 2022; Villani, 2003). The function $f : \mathbb{R} \to \mathbb{R}$ is convex and superlinear and $V, W : \mathcal{X} \to \mathbb{R}$ are convex and sufficiently smooth.

Class	**PDE** $\frac{\partial \mu_t}{\partial t} =$	**Flow Functional** $J(\mu_t) =$
Heat Equation	$\Delta \mu_t$	$\int \mu_t(x) \log \mu_t(x) \mathrm{d}x$
Advection	$\nabla \cdot (\mu_t \nabla V)$	$\int V(x) \mu_t(x) \mathrm{d}x$
Fokker-Planck	$\Delta \mu_t + \nabla \cdot (\mu_t \nabla V)$	$\int \mu_t(x) \log \mu_t(x) \mathrm{d}x + \int V(x) \mu_t(x) \mathrm{d}x$
Porous Media	$\Delta \left(\mu_t^m\right) + \nabla \cdot (\mu_t \nabla V)$	$\frac{1}{m-1} \int \mu_t(x)^m \, \mathrm{d}x + \int V(x) \mu_t(x) \mathrm{d}x$
Advection, Diffusion, and Interaction	$\nabla \cdot \left[\mu_t \left(\nabla f'(\mu_t) + \nabla V + (\nabla W) * \mu_t\right)\right]$	$\int V(x) \mu_t(x) \mathrm{d}x + \int f(\mu_t(x)) \mathrm{d}x$ $+\frac{1}{2} \iint W\left(x - x'\right) \mu_t(x) \mu_t\left(x'\right) \mathrm{d}x \mathrm{d}x'$

Wasserstein case. When studying the evolution of a population or measure μ_t over time, however, we need to resort to optimal transport metrics $W(\cdot, \cdot)$ (2) instead of the ℓ_2-norm $\|\cdot\|^2$. Considering functionals J that take a measure or population as input, a gradient flow of μ w.r.t. to J can be similarly expressed through forward and backward schemes. Assuming $J(\mu) \coloneqq \int E(x) d\mu(x)$, i.e., ignoring particle interaction, the forward scheme reads

$$\mu_{t+1} \coloneqq (I - \tau \nabla E)_\sharp \, \mu_t$$

with the corresponding backward formulation defined as

$$\mu_{t+1} \coloneqq \underset{\mu}{\operatorname{argmin}} \; \frac{1}{2\tau} W\left(\mu, \mu_t\right) + J(\mu). \tag{17}$$

This implicit time stepping is a useful tool to construct continuous flows: In the limit $\tau \to 0$ the resulting sequence $\{\mu_t\}_{t=0}^T$ approximates a continuous flow μ_t, i.e., a path in the Wasserstein space, and can thus be seen as the analogy of the usual proximal descent scheme, tailored for probability measures (Santambrogio, 2015, p.285)

Interest in Wasserstein gradient flows was sparked by the seminal work of Jordan, Kinderlehrer, and Otto (1998) who studied diffusion processes under the lens of the OT metric (see also Ambrosio et al., 2006). For a broad class of potentials J and provided with an initial distribution μ_0, the resulting time-discrete, iterative variational scheme induced by the so-called *Jordan-Kinderlehrer-Otto (JKO) step* (17) reconstructs the evolution of measure μ_t over time. As $\tau \to 0$, the solution of the time-discretized gradient flow converges to the solution of a corresponding PDE, and the resulting evolutions are often referred to as *JKO flows*.

Following Otto (2001) on the calculus of optimal transport (Otto calculus), a large class of partial differential equation may then be viewed as gradient flows on the Wasserstein space (Jordan et al., 1998). For instance, the standard heat equation of physics, i.e.,

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t,$$

with Δ being the spatial Laplacian, can be expressed as a gradient flow of the energy $J(\mu) = \int \mu_t(x) \log \mu_t(x) \mathrm{d}x$, i.e., Gibbs-Boltzmann's famous functional with the physical interpretation of the negative of an entropy. Among further examples displayed in Table 2 (Alvarez-Melis et al., 2022; Villani, 2003), a classic subject of the theory of PDEs

also comprises the linear Fokker-Planck equation

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t + \nabla \cdot (\mu_t \nabla V), \tag{18}$$

that is connected to flow functional

$$J(\mu_t) = \int \mu_t(x) \log \mu_t(x) \mathrm{d}x + \int V(x) \mu_t(x) \mathrm{d}x \,.$$

Here, the first term again represents the negative Gibbs-Boltzmann entropy and the second term plays the role of an energy functional with a smooth potential function $V : \mathbb{R}^d \to \mathbb{R}$.

Thus, JKO flows have found application in inferring the evolution of populations over time, crucial in many scientific disciplines, for instance in biomedicine to reconstruct cellular dynamics from observations (Bunne et al., 2022b; Alvarez-Melis et al., 2022; Mokrov et al., 2021; Benamou et al., 2016). This formulation is particularly interesting for studying dynamical systems in biomedicine, as the exact expression of the PDE corresponding to functional J does not need to be known. Instead, we can propose parameterizations of energy functional J that can be learned from data, an idea explored in Bunne et al. (2022b). While providing a general framework for studying general and complex population dynamics, each step of the JKO scheme (17) is costly as it requires solving a minimization problem involving the Wasserstein distance (2). Beyond introducing learning schemes for functional J, Bunne et al. (2022b) thus introduce novel efficient and differentiable schemes for solving JKO flows.

2.4 Stochastic Control Perspective

Benamou-Brenier motivated the introduction of the dynamic optimal transport problem from the perspective of fluid dynamics. As we shall see, both the OT problem (2) and its regularized version (3) can be viewed as stochastic *optimal control* problems. Control theory at the heart is concerned with finding optimal policies for dynamic systems subject to constraints. Despite wide-ranging progress on both the theory and applications, deploying control theory to large-scale and often unknown systems remains a grand challenge. As we will explore in the following, stochastic optimal control problems to regulate dynamic systems emerge from the theory of optimal transport (Santambrogio, 2015) that provides a geometric variational framework for studying flows of distributions on metric spaces (Chen et al., 2021a). These theoretical concepts build the foundation of recently developed deep learning architectures employed as generative models (Song et al., 2021; De Bortoli et al., 2021) or for studying the evolution of dynamical systems over time (Chen et al., 2022; Bunne et al., 2022b; Vargas et al., 2021). Further, celebrated control principles such as the Pontryagin maximum principle have been emphasized repeatedly in neural ordinary differential equation (ODE) (Chen et al., 2018) and stochastic differential equation (SDE) works (Jia and Benson, 2019).

... on Optimal Transport

Following Chen et al. (2021a,b), we will establish this stochastic control viewpoint by studying the Benamou-Brenier formulation using elementary control considerations.

For this, we consider a system with state distribution $\mathrm{d}X_t = v\,(t, X_t)\,\mathrm{d}t$ and initial state $X_0 \sim \mu_0$. Provided with a time-dependent feedback control $v(t, \cdot)$, the objective of (15) has the following stochastic interpretation

$$\int_0^1 \int_{\mathbb{R}^n} \frac{1}{2}\|v(t,x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t = \mathbb{E}\left\{\int_0^1 \frac{1}{2}\left\|v(t,X_t)\right\|^2 \mathrm{d}t\right\},$$

resulting in the stochastic control formulation of the OT problem

$$\inf_{v\in\mathcal{V}} \mathbb{E}\left\{\int_0^1 \frac{1}{2}\|v(t,X_t)\|^2 \mathrm{d}t\right\} \tag{19}$$

$$\mathrm{d}X_t = v\,(t, X_t)\,\mathrm{d}t \tag{20}$$

$$X_0 \sim \mu_0, \quad X_1 \sim \mu_1.$$

$\mathcal{V}$ here represents the family of admissible state feedback control strategies. Typically, in a density control problem, the objective is to guide a dynamical system from an initial state X_0 characterized by μ_0 to a desired state μ_1 with minimum cost and control that is a member of the set of admissible actions, i.e., $v \in \mathcal{V}$. The above strategy, however, differs from standard optimal control in the added constraint on the terminal state distribution and the absence of a terminal penalty.

... on Regularized Optimal Transport

Similarly, the regularized OT problem (3) can be cast as a stochastic control problem.

$$\inf_{v\in\mathcal{V}} \mathbb{E}\left\{\int_0^1 \frac{1}{2}\|v(t,X_t)\|^2 \mathrm{d}t\right\} \tag{21}$$

$$\mathrm{d}X_t = v(t, X_t)\mathrm{d}t + \sigma \mathrm{d}\mathbb{W}_t \tag{22}$$

$$X_0 \sim \mu_0, \quad X_1 \sim \mu_1,$$

where $\mathbb{W}_t$ denotes a Wiener process, i.e., integrals of white noise. Different from (20), however, (22) is a stochastic diffusion process.

Besides, this problem exhibits a fluid-dynamic interpretation, i.e.,

$$\inf_{(\mu_t, v)} \int_0^1 \int_{\mathbb{R}^n} \frac{1}{2}\|v(t,x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t \tag{23}$$

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v\mu_t) - \frac{\sigma^2}{2}\Delta\mu_t = 0 \tag{24}$$

$$\mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1\,,$$

where Δ denotes the Laplace operator. (24) is the Fokker-Planck equation capturing the state distribution evolution. As $\sigma^2 \to 0$, the solution to this problem converges to the one of the Benamou-Brenier problem (15) (Mikami and Thieullen, 2008). For an extended discussion, see Dai Pra (1991); Mikami (2000, 2002).

2.5 Schrödinger Bridges

Interestingly, Eq. (21) first emerged in a very different setting. In his work "Über die Umkehrung der Naturgesetze" published in 1931, Schrödinger studied the most likely

random evolution between two marginals, i.e., two point clouds of diffusive particles. His Gedankenexperiment is best illustrated through a population of independent and identically distributed particles in $\mathbb{R}^d$ observed at $t = 0$ as the empirical distribution μ_0, and again at $t = 1$ as μ_1. To describe the most likely dynamics of these particles over time, we aim at finding the stochastic process $\mathbb{P}_t$ on $[0, 1]$ such that $\mathbb{P}_0 = \mu_0, \mathbb{P}_1 = \mu_1$.

Provided with some prior knowledge of a reference process $\mathbb{Q}_t$, e.g., that the underlying dynamics follow a Brownian motion (BM), we aim to identify the stochastic process $\mathbb{P}_t$ that best describes the particles evolution, i.e., minimizes the overall relative entropy

$$\min_{\mathbb{P}_0=\mu_0,\, \mathbb{P}_1=\mu_1} D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{Q}_t) = \int_{\mathcal{C}[0,1]} \log\left(\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}\right) d\mathbb{P}_t, \tag{25}$$

where $\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}$ denotes the Radon-Nikodym derivative and $\mathcal{C}[0, 1]$ the continuous paths on $\mathbb{R}^d$ over the time interval $[0, 1]$. More concretely, to find $\mathbb{P}_t$, Schrödinger (1931, 1932) considers the objective (25) as the "mostly likely process" that explains the marginal distributions $\mathbb{P}_0, \mathbb{P}_1$ relative to reference process $\mathbb{Q}_t$. This *KL-minimization* problem is thus called the (generalized) Schrödinger bridge. This idea generalizes verbatim to any reference process $\mathbb{Q}_t$. Unfortunately, in most applications, notably biology, we often have little to no prior information about the underlying process $\mathbb{P}_t$ (Liberali et al., 2014).

Recovering the stochastic calculus of variations formulation of the Schrödinger bridge (21) can be achieved via the Girsanov theorem which tells us how stochastic processes behave under changes in measure. The equivalence between both formulations can be then established via

$$\frac{\mathrm{d}\mathbb{P}_t}{\mathrm{d}\mathbb{P}_t^{v=0}} = \exp\left\{\int_0^1 \frac{1}{\sigma} v(t, X_t^v) \cdot \mathrm{d}\mathbb{W}_t + \int_0^1 \frac{1}{2\sigma^2} \|v(t, X_t^v)\|^2 dt\right\}$$

$$\text{and thus } \; D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{P}_t^{v=0}) = \mathbb{E}\left\{\int_0^1 \frac{1}{2\sigma^2} \|v(t, X_t)\|^2 dt\right\},$$

where $\mathbb{P}_t$ and $\mathbb{P}_t^{v=0}$ denote a controlled process (with control v) and an uncontrolled process, i.e., with $v(t, \cdot) = 0$, respectively. In other words, the relative entropy between the stochastic process describing the particle dynamics and the reference process is equal to the control energy (scaled by $\frac{1}{\sigma^2}$).

Optimality criteria. Classical strategies for solving (21) commonly replace the boundary constraint $X_1 \sim \mu_1$ with a penalty or artificial terminal cost, thus transforming (21) to standard stochastic optimal control formulations. The resulting optimality conditions are

$$\frac{\partial \psi}{\partial t} = -\frac{1}{2}\|\nabla \psi\|^2 - \frac{\sigma^2}{2} \Delta \psi \tag{26}$$

$$\frac{\partial \mu_t}{\partial t} = -\nabla \cdot (\mu_t \nabla \psi) + \frac{\sigma^2}{2} \Delta \mu_t \tag{27}$$

with value function $\psi(t, x)$ and μ_t being the associated optimal marginal density. Here, $v \equiv \nabla \psi$ and $\psi(1, \cdot)$ is equivalent to the terminal cost. Further, Eq. (26) is a second-order Hamilton-Jacobi-Bellman equation, while (27) is the continuity equation. After applying the Hopf-Cole transform $(\psi, \mu_t) \to (\Phi, \widehat{\Phi})$,

$$\Phi = \exp(\frac{\psi}{\sigma^2}) \;\text{ and }\; \widehat{\Phi} = \mu_t \exp(\frac{-\psi}{\sigma^2}),$$

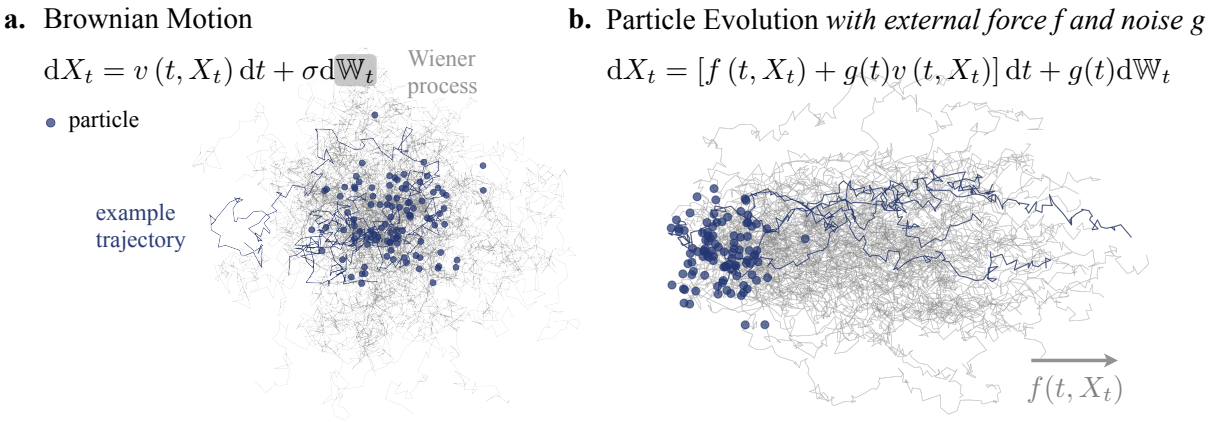

Figure 3: Comparison of different SDE classes. a. Particles evolve according to simple BM in all directions depending on the noise level σ. **b.** A particle evolution with an external speed f, here exemplified through a horizontal drift, and time-dependent noise g. The initial location of the particles is denoted as a blue dot, example trajectories are highlighted by blue lines.

we obtain the SB system associated to the SDE class in (22) given by

$$\begin{aligned} \frac{\partial \Phi}{\partial t} &= -\frac{\sigma^2}{2}\Delta\Phi \quad \text{s.t.} & \Phi(0,\cdot)\widehat{\Phi}(0,\cdot) &= \mu_0, \\ \frac{\partial \widehat{\Phi}}{\partial t} &= \frac{\sigma^2}{2}\Delta\widehat{\Phi} & \Phi(1,\cdot)\widehat{\Phi}(1,\cdot) &= \mu_1. \end{aligned} \tag{28}$$

i.e., a backward Kolmogorov and a Fokker-Planck equation, respectively. The optimal control is then given by $v(t, X_t) = \sigma^2 \nabla \log \Phi(t, x)$.

Generalizations to Other SDE Classes

To describe complex biological processes, however, we need to consider SDE classes comprising nonlinear drifts, affine control, and time-varying diffusion. In the following, let us consider SDEs with an external speed $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$, time-dependent diffusion $g(t) \in \mathbb{R}$, and standard Wiener process $\mathbb{W}_t \in \mathbb{R}^d$.[3] Caluya and Halder (2021); Chen et al. (2022) provide a generalization of the above framework that reads

$$\inf_{v \in \mathcal{V}} \mathbb{E}\left\{\int_0^1 \frac{1}{2}\|v(t, X_t)\|^2 dt\right\} \tag{29}$$

$$\mathrm{d}X_t = [f\,(t, X_t) + g(t)v(t, X_t)]\,\mathrm{d}t + \sigma g(t)\mathrm{d}\mathbb{W}_t \tag{30}$$

$$X_0 \sim \mu_0, \quad X_1 \sim \mu_1,$$

with $g(t)$ being uniformly lower-bounded and $f(t, X_t)$ satisfying Lipschitz conditions with at most linear growth in x. The effect of adding an external force f, here exemplified through a horizontal drift, compared to standard Brownian motion is visualized in Fig. 3.

[3] Hereafter, we will sometimes drop $f \equiv f(t, X_t)$ and $g \equiv g(t)$ for brevity.

Optimality criteria. Again, we recover the optimality criteria via a Hopf-Cole transform of (29) resulting in

$$\begin{aligned} \frac{\partial \Phi}{\partial t} &= -\nabla \Phi^T f - \frac{\sigma^2}{2} g^2 \Delta \Phi \qquad \text{s.t.} \qquad \Phi(0,\cdot)\widehat{\Phi}(0,\cdot) = \mu_0, \\ \frac{\partial \widehat{\Phi}}{\partial t} &= -\nabla \cdot (\widehat{\Phi} f) + \frac{\sigma^2}{2} g^2 \Delta \widehat{\Phi} \qquad\qquad \Phi(1,\cdot)\widehat{\Phi}(1,\cdot) = \mu_1 \end{aligned} \tag{31}$$

with the optimal control $v(t, X_t) = \sigma^2 g(t) \nabla \log \Phi\,(t, X_t)$. The solution in (31) can be expressed through two coupled SDEs of the form (Léonard, 2013)

$$\mathrm{d}X_t = \left[f + g^2 \nabla \log \Phi\,(t, X_t)\right] \mathrm{d}t + g\, \mathrm{d}\mathbb{W}_t, \qquad X_0 \sim \mu_0, \tag{32}$$

$$\mathrm{d}X_t = \left[f - g^2 \nabla \log \widehat{\Phi}\,(t, X_t)\right] \mathrm{d}t + g\, \mathrm{d}\mathbb{W}_t, \qquad X_T \sim \mu_T, \tag{33}$$

where $T = 1$, $\sigma = 1$, and $\nabla \log \Phi\,(t, X_t)$ and $\nabla \log \widehat{\Phi}\,(t, X_t)$ are the optimal forward and backward drifts for the Schrödinger bridge.

2.5.1 Diffusion Schrödinger Bridges

Interestingly, the underlying SDEs (30) coincides with the dynamic systems considered in score-based generative models (SGMs) (Song et al., 2021), an emerging generative model class that has achieved remarkable results in synthesizing high-fidelity data (Song and Ermon, 2019; Kong et al., 2021). It also represents a key connection that has recently fueled the development of diffusion Schrödinger bridges (DSBs) (De Bortoli et al., 2021; Chen et al., 2021b; Bunne et al., 2023a; Liu et al., 2022a). Compared to classical diffusion-based generative models (Daniels et al., 2021; Song et al., 2021), these algorithms allow interpolation between complex distributions. Extended to the Riemannian geometry (Thornton et al., 2022; De Bortoli et al., 2022), it has found applications in molecular dynamics (Holdijk et al., 2022; Somnath et al., 2023), and cell differentiation processes (Vargas et al., 2021; Bunne et al., 2023a; Tong et al., 2023).

To learn and parameterize $\nabla \log \Phi\,(t, X_t)$ in (32) and $\nabla \log \widehat{\Phi}\,(t, X_t)$ in (33) with $Z_t, \hat{Z}_t$: $\mathbb{R}^d \to \mathbb{R}^d$, i.e., two time-indexed smooth *vector fields* called the optimal forward and backward drift, respectively. Note that (33) runs backward in time, i.e., from $1 \to 0$ (Anderson, 1982). Choosing f and g depending on the considered SDE class, the forward and backward policies $Z_t, \hat{Z}_t$ are generally *unknown*. Similar as in score-based generative models (Song et al., 2021; Hyvärinen and Dayan, 2005) which parameterize the score function, in order to *estimate* the resulting SB from data, we learn the forward and backward drift through neural networks (NNs) with parameters θ, ϕ, i.e., $Z_t^\theta(x)$ and $\hat{Z}_t^\phi(x)$. For a visualization of the resulting parameterization, see Fig. 4.

Several estimators and training procedures for the so-called diffusion Schrödinger bridges, i.e., for learning Z_t and $\hat{Z}_t$, have been proposed based on either Gaussian processes (Vargas et al., 2021), dual potentials (Finlay et al., 2020), or neural networks (De Bortoli et al., 2021; Chen et al., 2022). In this thesis, we consider the likelihood training framework by Chen et al. (2022) grounded on forward-backward SDE (FBSDE) theory (Ma and Yong, 1999; Exarchos and Theodorou, 2018). Crucially, these forward-backward SDEs (FBSDEs) can be used to construct the likelihood objectives for SBs that, surprisingly, generalize the ones for SGMs as special cases.

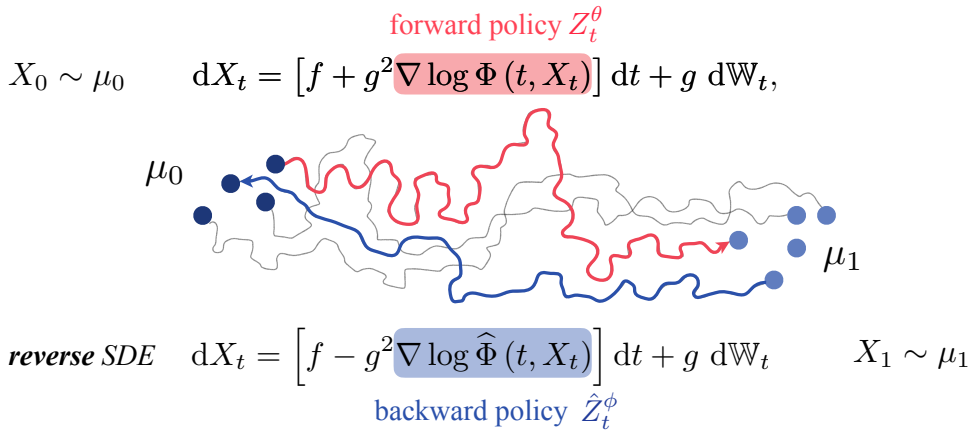

Figure 4: Parameterization of diffusion Schrödinger bridges. The forward SDE with forward policy Z_t steers particles $X_0 \sim \mu_0$ from $t = 0$ to μ_1 at $t = 1$. The reverse SDE runs backward in time. Here, backward policy $\hat{Z}_t$ determines the evolution of particles $X_1 \sim \mu_1$ at $t = 1$ to μ_0 at $t = 0$.

The negative likelihood functions for θ and ϕ are then given by

$$\ell(x_0; \phi) = \int_0^1 \mathbb{E}_{(32)}\left[\frac{1}{2}\|\hat{Z}_t^\phi\|^2 + g\nabla_x \cdot \hat{Z}_t^\phi + \langle Z_t^\theta, \hat{Z}_t^\phi \rangle\, \mathrm{d}t \middle| X_0 = x_0\right], \tag{34a}$$

$$\ell(x_1; \theta) = \int_0^1 \mathbb{E}_{(33)}\left[\frac{1}{2}\|Z_t^\theta\|^2 + g\nabla_x \cdot Z_t^\theta + \langle \hat{Z}_t^\phi, Z_t^\theta \rangle\, \mathrm{d}t \middle| X_1 = x_1\right]. \tag{34b}$$

and serve as loss functions for likelihood-based training of DSBs. Here, $\nabla_x \cdot$ denotes the divergence operator w.r.t. the x variable: For any $v \colon \mathbb{R}^d \to \mathbb{R}^d$, $\nabla_x \cdot v(x) := \sum_{i=1}^d \frac{\partial}{\partial x_i} v_i(x)$.

Unfortunately, such frameworks necessitate a forward-backward learning process known as the iterative proportional fitting (IPF) procedure (Fortet, 1940; Kullback, 1968). As both policies $Z_t, \hat{Z}_t$ are initially unknown and randomly parameterized, training DSBs often results in numerical and scalability issues. Several works have thus studied initialization schemes and methods to improve numerical robustness (Bunne et al., 2023a).

Further, none of these approaches is capable of incorporating *alignment* of the data. This can be seen by inspecting the objective (25), in which the coupling information (x_0^i, x_1^i) is completely lost as only its individual marginals μ_0, μ_1 play a role therein. Thus, several approaches propose an algorithmic framework that solves (32)-(33) in settings where sparse trajectories, or partially aligned data, are available *without* resorting to IPF (Somnath et al., 2023; Shi et al., 2023; Tong et al., 2023; Pooladian et al., 2023a).

Resources

Extended discussions on theoretical properties and numerical considerations can be found in the following books and review articles:

- G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. Foundations and Trends in Machine Learning 11.5-6 (2019)

- C. Villani. Topics in Optimal Transportation. GSM Vol. 58, AMS, 2009.
- F. Santambrogio. Optimal Transport for Applied Mathematicians. Birkhäuser, 2015.
- Y. Chen, T. T. Georgiou, and M. Pavon. Optimal Transport in Systems and Control. Annual Review of Control, Robotics, and Autonomous Systems Vol. 4 (2021).

Further, the above-mentioned methods are implemented in various `Python` libraries. In particular,

- `OTT`: An OT library in `JAX` (Bradbury et al., 2018) by Cuturi et al. (2022).
- `POT`: An OT library in `NumPy` (Bradbury et al., 2018) and `PyTorch` (Paszke et al., 2019) by Flamary et al. (2021).
- `GeomLoss`: An OT library in `PyTorch` (Paszke et al., 2019) by Feydy et al. (2019).

References

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

D. Alvarez-Melis, Y. Schiff, and Y. Mroueh. Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks. *Transactions on Machine Learning Research (TMLR)*, 2022.

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

B. Amos. On amortizing convex conjugates for optimal transport. In *International Conference on Learning Representations (ICLR)*, 2023.

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

B. Amos, S. Cohen, G. Luise, and I. Redko. Meta Optimal Transport. In *International Conference on Machine Learning (ICML)*, 2023.

A.-M. Ampère. *Mémoire contenant l'application de la théorie exposée dans le XVII. e Cahier du Journal de l'École polytechnique, à l'intégration des équations aux différentielles partielles du premier et du second ordre*. De l'Imprimerie royale, 1819.

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3), 1982.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, 2017.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3), 2000.

J.-D. Benamou, G. Carlier, and M. Laborde. An augmented lagrangian approach to wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54, 2016.

N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, 2011.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

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

Y. Brenier. Polar Factorization and Monotone Rearrangement of Vector-Valued Functions. *Communications on Pure and Applied Mathematics*, 44(4), 1991.

C. Bunne, A. Krause, and M. Cuturi. Supervised Training of Conditional Monge Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022a.

C. Bunne, L. Meng-Papaxanthos, A. Krause, and M. Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 25, 2022b.

C. Bunne, Y.-P. Hsieh, M. Cuturi, and A. Krause. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206, 2023a.

C. Bunne, S. G. Stark, G. Gut, J. S. del Castillo, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Rätsch. Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *Nature Methods*, 2023b.

L. A. Caffarelli. Interior $W^{2,p}$ estimates for solutions of the Monge-Ampère equation. *Annals of Mathematics*, 1990.

L. A. Caffarelli. Allocation Maps with General Cost Functions. In *Partial Differential Equations and Applications*, volume 177 of Lecture Notes in Pure and Appl. Math. Dekker, 1996.

L. A. Caffarelli. The Monge-Ampère equation and optimal transportation, an elementary review. *Lecture Notes in Mathematics: Optimal Transportation and Applications*, 1813, 2003.

K. F. Caluya and A. Halder. Wasserstein Proximal Algorithms for the Schrödinger Bridge Problem: Density Control with Nonlinear Drift. *IEEE Transactions on Automatic Control*, 67(3), 2021.

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

T. Chen, G.-H. Liu, and E. A. Theodorou. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations (ICLR)*, 2022.

Y. Chen, T. T. Georgiou, and M. Pavon. Optimal Transport in Systems and Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 4, 2021a.

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

J. D. Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9(3), 1951.

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

M. Cuturi, L. Meng-Papaxanthos, Y. Tian, C. Bunne, G. Davis, and O. Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv Preprint arXiv:2201.12324*, 2022. URL https://github.com/ott-jax/ott.

M. Cuturi, M. Klein, and P. Ablin. Monge, Bregman and Occam: Interpretable Optimal Transport in High-Dimensions with Feature-Sparse Maps. In *International Conference on Machine Learning (ICML)*, 2023.

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1), 1991.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2017.

M. Daniels, T. Maunu, and P. Hand. Score-based Generative Neural Networks for Large-Scale Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian Score-Based Generative Modelling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

V. Divol, J. Niles-Weed, and A.-A. Pooladian. Optimal transport map estimation in general function spaces. *arXiv Preprint arXiv:2212.03722*, 2022.

I. Exarchos and E. A. Theodorou. Stochastic optimal control via forward and backward stochastic differential equations and importance sampling. *Automatica*, 87, 2018.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019. URL https://www.kernel-operations.io/geomloss.

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

A. Figalli. *The Monge–Ampère equation and Its Applications*. Zurich Lectures in Advanced Mathematics, 2017.

C. Finlay, A. Gerolin, A. M. Oberman, and A.-A. Pooladian. Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv Preprint arXiv:2006.06033*, 2020.

R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 2016. URL https://pythonot.github.io.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research (JMLR)*, 22, 2021.

R. Fortet. Résolution d'un système déquations de M. Schrödinger. *J. Math. Pure Appl. IX*, 1, 1940.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein Loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

W. Gangbo and R. J. McCann. Optimal maps in Monge's mass transport problem. *Comptes Rendus de l'Academie des Sciences-Serie I-Mathematique*, 321(12), 1995.

W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2), 1996.

A. Genevay, G. Peyré, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

L. Holdijk, Y. Du, F. Hooft, P. Jaini, B. Ensing, and M. Welling. Path Integral Stochastic Optimal Control for Sampling Transition Paths. *arXiv Preprint arXiv:2207.02149*, 2022.

E. Hopf. The Partial Differential Equation $u_t + uu_x = \mu^*_{xx}$. *Communications on Pure and Applied Mathematics*, 3(3), 1950.

C.-W. Huang, R. T. Q. Chen, C. Tsirigotis, and A. Courville. Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

A. Hyvärinen and P. Dayan. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research (JMLR)*, 6(4), 2005.

H. Janati, T. Bazeille, B. Thirion, M. Cuturi, and A. Gramfort. Multi-subject MEG/EEG source imaging with sparse multi-task regression. *NeuroImage*, 220, 2020.

J. Jia and A. R. Benson. Neural Jump Stochastic Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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

L. Kantorovich. On the transfer of masses (in Russian). In *Doklady Akademii Nauk*, volume 37, 1942.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev. Wasserstein-2 Generative Networks. In *International Conference on Learning Representations (ICLR)*, 2021a.

A. Korotin, L. Li, A. Genevay, J. M. Solomon, A. Filippov, and E. Burnaev. Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021b.

S. Kullback. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4), 1968.

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

P. Liberali, B. Snijder, and L. Pelkmans. A hierarchical map of regulatory genetic interactions in membrane trafficking. *Cell*, 157(6), 2014.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

G.-H. Liu, T. Chen, O. So, and E. A. Theodorou. Deep Generalized Schrödinger Bridge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

X. Liu, L. Wu, M. Ye, and Q. Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv Preprint arXiv:2209.03003*, 2022b.

F. Lübeck, C. Bunne, G. Gut, J. S. del Castillo, L. Pelkmans, and D. Alvarez-Melis. Neural Unbalanced Optimal Transport via Cycle-Consistent Semi-Couplings. *arXiv Preprint arXiv:2209.15621*, 2022.

J. Ma and J. Yong. *Forward-Backward Stochastic Differential Equations and their Applications*. Springer Science & Business Media, 1999.

A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning (ICML)*, volume 119, 2020.

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

T. Mikami. Dynamical Systems in the Variational Formulation of the Fokker-Planck Equation by the Wasserstein Metric. *Applied Mathematics and Optimization*, 42, 2000.

T. Mikami. Optimal control for absolutely continuous stochastic processes and the mass transportation problem. *Electronic Communications in Probability*, 2002.

T. Mikami and M. Thieullen. Optimal transportation problem by stochastic optimal control. *SIAM Journal on Control and Optimization*, 47, 2008.

P. Mokrov, A. Korotin, L. Li, A. Genevay, J. Solomon, and E. Burnaev. Large-Scale Wasserstein Gradient Flows. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, 1781.

F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Taylor & Francis*, 2001.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://pytorch.org.

G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *arXiv Preprint arXiv:2109.12004*, 2021.

A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. Chen. Multisample Flow Matching: Straightening Flows with Minibatch Couplings. In *International Conference on Machine Learning (ICML)*, 2023a.

A.-A. Pooladian, V. Divol, and J. Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning (ICML)*, 2023b.

R. T. Rockafellar. Conjugate convex functions in optimal control and the calculus of variations. *Journal of Mathematical Analysis and Applications*, 32(1), 1970.

L. Rüschendorf. Bounds for Distributions with Multivariate Marginals. *Lecture Notes-Monograph Series*, 1991a.

L. Rüschendorf. Fréchet-bounds and their applications. In *Advances in Probability Distributions with Given Marginals*, volume 67 of Mathematics and Its Applications. Springer, 1991b.

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs Using Optimal Transport. In *International Conference on Learning Representations (ICLR)*, 2018.

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

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4), 2019.

E. Schrödinger. *Über die Umkehrung der Naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u. Company, 1931.

E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, 1932.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion Schrödinger Bridge Matching. *arXiv Preprint arXiv:2303.16852*, 2023.

S. P. Singh and M. Jaggi. Model Fusion via Optimal Transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

V. R. Somnath, M. Pariset, Y.-P. Hsieh, M. R. Martinez, A. Krause, and C. Bunne. Aligned Diffusion Schrödinger Bridges. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

Y. Song and S. Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021.

A. Taghvaei and A. Jalali. 2-Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs. *arXiv Preprint arXiv:1902.07197*, 2019.

J. Thornton, M. Hutchinson, E. Mathieu, V. De Bortoli, Y. W. Teh, and A. Doucet. Riemannian Diffusion Schrödinger Bridge. *arXiv Preprint arXiv:2207.03024*, 2022.

A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy. TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics. In *International Conference on Machine Learning (ICML)*, 2020.

A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio. Conditional Flow Matching: Simulation-Free Dynamic Optimal Transport. *arXiv Preprint arXiv:2302.00482*, 2023.

T. Uscidda and M. Cuturi. The Monge Gap: A Regularizer to Learn All Transport Maps. In *International Conference on Machine Learning (ICML)*, 2023.

F. Vargas, P. Thodoroff, N. D. Lawrence, and A. Lamacraft. Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9), 2021.

C. Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2003.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.