A dynamic perspective of optimal transport

Bernhard Schmitzer

Göttingen, June 2025

1 Preface

In this chapter we explore some aspects of dynamic optimal transport, in particular for the Wasserstein-2 distance on $\mathbb{R}^d$, with and without entropic regularization. The unregularized setting is treated in Section 2. We put a special emphasis on the primal-dual structure of the dynamic Benamou–Brenier formula and explore how the dynamic dual potential directs the movement of mass particles via its gradient in a drift equation and how this dynamic dual potential is linked to the static Kantorovich dual potentials via the Hopf–Lax formula. We obtain this formula directly via convex duality of suitable intermediate multi-marginal problems.

In Section 3 we add entropic regularization. While this destroys the metric structure, it gives rise to a dynamic interpretation via the notion of Schrödinger bridges. We give some intuition for the origin of this interpretation and then study the behaviour of these bridges in more detail, once more via an auxiliary multi-marginal formulation, which leads to a system of drift-diffusion equations. This then gives rise to an entropic variant of the Benamou–Brenier formula.

For accessibility, we make an effort to use only rather basic tools from convex analysis on compact spaces and avoid more advanced tools such as measures on paths or stochastic calculus. References to more in-depth treatments are given throughout the text. While all results in this chapter are well known and covered in the literature, we hope that our exploration in this chapter via basic tools and the discussion of various equivalent reformulations will help to broaden the readers' understanding of dynamic optimal transport.

Remark 1.1 (Notation)**.** Throughout this chapter, (X, d) is a compact metric space. We denote by $\mathrm{C}(X)$ the Banach space of continuous functions on X, equipped with the sup-norm. Its topological dual can be identified with the space of Radon measures $\mathcal{M}(X)$, equipped with the total variation norm. The weak* topology on $\mathcal{M}(X)$ is the one induced by the pairing with $\mathrm{C}(X)$. Probability measures are noted by $\mathcal{P}(X)$, non-negative measures by $\mathcal{M}_+(X)$. $\langle \cdot, \cdot \rangle$ denotes duality pairing, the implied paired spaces are always clear from context.

2 Optimal transport

2.1 Kantorovich formulation

Definition 2.1 (Kantorovich formulation of optimal transport)**.** Let (X, d) be a compact metric space, $c \in \mathrm{C}(X \times X)$, and $\mu, \nu \in \mathcal{P}(X)$. Then the Kantorovich optimal transport problem between measures μ and ν for cost function c is given by

$$C(\mu, \nu) := \inf \left\{ \int_{X \times X} c(x, y) \, \mathrm{d}\gamma(x, y) \,\middle|\, \gamma \in \Gamma(\mu, \nu) \right\} \tag{2.1}$$

where $\Gamma(\mu, \nu)$ is the set of *transport plans* or *couplings* between μ and ν, given by

$$\Gamma(\mu, \nu) := \{\gamma \in \mathcal{M}_+(X \times X) \,|\, \mathrm{P}_1\gamma = \mu, \, \mathrm{P}_2\gamma = \nu\} \tag{2.2}$$

and $\mathrm{P}_i\gamma = \mathrm{p}_{i\#}\gamma$ where $\mathrm{p}_i(x_1, x_2) = x_i$ is the projection onto the i-th coordinate.

Proposition 2.2. Minimizers in (2.1) exist.

Proof. The objective $\gamma \mapsto \langle c, \gamma \rangle$ is continuous. The set $\Gamma(\mu, \nu)$ is weak* closed and compact (by Banach–Alaoglu or Prokhorov's theorems). The same arguments can be generalized to lower-semicontinuous c and Polish X, see [51, Theorem 4.1]. □

(2.1) is a prototypical linear program and convex duality is an essential tool for its analysis. We will establish convex duality results via the Fenchel–Rockafellar theorem between topologically paired spaces. Two vector spaces U and U^* with locally convex topologies are said to be topologically paired if the bounded linear functionals of each can be identified with elements of the other. The pairing between these spaces is a bilinear form $\langle \cdot, \cdot \rangle : U \times U^* \to \mathbb{R}$. Our example of interest is the space of continuous functions $\mathrm{C}(X)$ with the sup-norm topology paired with the Radon measures $\mathcal{M}(X)$ with the weak* topology.

Theorem 2.3 (Fenchel–Rockafellar [45])**.** Let (U, U^*) and (V, V^*) be two couples of topologically paired spaces. Let $A : U \to V$ be a bounded linear operator and $A^* : V^* \to U^*$ its adjoint. Let F and G be lower-semicontinuous proper convex functions on U and V respectively, with values in $\mathbb{R} \cup \{\infty\}$. If there exists an $x \in U$ such that $F(-x) < \infty$ and G is continuous at Ax, then

$$\sup_{x \in U} -F(-x) - G(Ax) = \min_{y \in V^*} F^*(A^*y) + G^*(y) \tag{2.3}$$

and the minimum is attained. Moreover, there exists a maximizer $x \in U$ if and only if there is some $y \in V^*$ such that $Ax \in \partial G^*(y)$ and $A^*y \in \partial F(-x)$.

Note that the two optimality conditions at the end of the statement are equivalent to $y \in \partial G(Ax)$ and $x \in -\partial F^*(A^*y)$, respectively.

We are now ready to state the duality result. A more general duality statement for non-compact X and lower-semicontinuous cost c can be found in [51, Theorem 5.10].

Proposition 2.4 (Dual Kantorovich problem)**.**

$$C(\mu,\nu) = \sup\left\{ \int_X \phi \,\mathrm{d}\mu + \int_X \psi \,\mathrm{d}\nu \,\middle|\, \phi,\psi \in \mathrm{C}(X),\ \phi \oplus \psi \leq c \right\} \tag{2.4}$$

where $\phi \oplus \psi$ denotes the function $(x,y) \mapsto \phi(x) + \psi(y)$ and the inequality $\phi \oplus \psi \leq c$ is to be enforced on all of $X \times X$. Minimizers in (2.1) exist.

Proof. This follows from Theorem 2.3 as follows: Let $U = \mathrm{C}(X) \times \mathrm{C}(X)$, $V = \mathrm{C}(X \times X)$, $A : U \to V$, $(\phi,\psi) \mapsto \phi \oplus \psi$, $F : (\phi,\psi) \mapsto \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$ and $G : \eta \mapsto 0$ if $\eta \leq c$ and $+\infty$ otherwise. Then (2.4) has the form $\sup_{(\phi,\psi)\in U} -F(-(\phi,\psi)) - G(A(\phi,\psi))$. F is finite everywhere and G is continuous at any function η that is strictly smaller than c, e.g. at $\eta : (x,y) \mapsto C$ for $C := \min_{(x',y')\in X\times X} c(x',y') - 1$ and $\eta = A(\phi,\phi)$ for $\phi : x \mapsto C/2$. Therefore, (2.4) is dual to $\min_{\gamma \in V^*} F^*(A^*\gamma) + G^*(\gamma)$. One finds that $A^*\gamma = (\mathrm{P}_1\gamma, \mathrm{P}_2\gamma)$, $F^*(\rho,\sigma) = 0$ if $\rho = \mu$ and $\sigma = \nu$ and $+\infty$ otherwise, and $G^*(\gamma) = \int c \mathrm{d}\gamma$ if $\gamma \geq 0$ and $+\infty$ otherwise, which means that the dual problem is equal to (2.1). □

Proposition 2.4 does not directly imply existence of dual maximizers (ϕ,ψ). Fortunately, their existence can be established with some simple explicit arguments that will also be helpful later on. Consider the dual problem (2.4) and assume for now that $\psi \in \mathrm{C}(X)$ is fixed. What is the best possible choice for the remaining supremum over ϕ? Formally, ignoring the constraint that ϕ must be continuous, this partial optimization problem can be solved point-wise for each value $\phi(x)$, by setting $\phi(x)$ such that the first of the constraints $\phi(x) \leq c(x,y) - \psi(y)$ for some $y \in X$ becomes active. This motivates the following definition.

Definition 2.5 (c-transform)**.** For a cost function $c \in \mathrm{C}(X \times X)$ and a potential $\psi \in \mathrm{C}(X)$ the $c-$transform of ψ is defined as the function ψ^c given by

$$\psi^c : x \mapsto \inf_{y\in X} c(x,y) - \psi(y).$$

In the same vein we define the $\bar{c}$-transform as

$$\psi^{\bar{c}} : y \mapsto \inf_{x\in X} c(x,y) - \psi(x).$$

Of course, when c is symmetric, both transformations are identical. For continuous c and ψ it turns out that ψ^c is indeed continuous and therefore a maximizer for (2.4) with respect to ϕ for fixed ψ. In fact, ψ^c is even more regular, which is the key for the following existence proof.

Proposition 2.6. Maximizing (ϕ,ψ) for (2.4) exist.

Proof. We sketch here the key steps of the proof. More details can be found in [48, Section 1.2].

As discussed above, for fixed ψ the maximizing ϕ is given by $\phi = \psi^c$. Conversely, for fixed ϕ, the maximizing ψ is given by $\psi = \phi^{\bar{c}}$. In addition, it is easy to verify that

$((\psi^c)^{\bar{c}})^c = \psi^c$. Consequently, we can restrict ourselves to maximizing sequences $(\phi_n, \psi_n)_n$ where the potentials are c-transforms of each other, i.e. $\phi_n = \psi_n^c$ and $\psi_n = \phi_n^{\bar{c}}$.

Next, observe that ψ^c inherits the modulus of continuity of c: Since c is continuous on a compact domain, there is a continuous function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ with $\omega(0) = 0$ such that $|c(x,y) - c(x',y)| \leq \omega(d(x,x'))$ (and likewise for the second argument). Assume now that $\phi = \psi^c$. Then find

$$\phi(x) \leq c(x,y) - \psi(y) \leq c(x',y) - \psi(y) + \omega(d(x,x')).$$

Taking now the infimum over y on the right-hand side one obtains $\phi(x) \leq \phi(x') + \omega(d(x,x'))$. By the symmetric argument we obtain $|\phi(x) - \phi(x')| \leq \omega(d(x,x'))$. The argument for ψ is identical.

Therefore, the maximizing sequence $(\phi_n, \psi_n)_n$ is equicontinuous. The objective (2.4) is invariant under applying constant shifts of the form $(\psi, \psi) \mapsto (\phi + \lambda, \psi - \lambda$ for $\lambda \in \mathbb{R}$. Thus we may assume that $\phi_n(x_0) = 0$ for some fixed $x_0 \in X$, which then implies that $(\phi_n, \psi_n)_n$ is also equi-bounded. Then, by the Arzelà–Ascoli theorem (e.g. [46, Thm. 11.28]) there exists a converging subsequence with limit (ϕ, ψ). The objective $\langle \phi, \mu \rangle + \langle \psi, \nu \rangle$ is continuous, and the admissible set imposed by the constraint $\phi \oplus \psi \leq c$ is closed, hence the limit is a maximizer. □

Definition 2.7 (Contact set)**.** Let $(\phi, \psi) \in \mathrm{C}(X)^2$ such that

$$\phi(x) + \psi(y) \leq c(x,y) \quad \text{for all} \quad (x,y) \in X \times X.$$

We call the pairs $(x,y) \in X \times X$ where one has equality the *contact set* of (ϕ, ψ).

Proposition 2.8. $\gamma \in \Gamma(\mu, \nu)$ and $(\phi, \psi) \in \mathrm{C}(X)^2$ with $\phi \oplus \psi \leq c$ are primal and dual optimal in (2.1) and (2.4) if and only if $\phi \oplus \psi = c$ γ-almost everywhere, i.e. γ is concentrated on the contact set of (ϕ, ψ).

Proof. Consider the primal-dual gap between (2.1) and (2.4) for admissible candidates:

$$0 \leq \int_{X \times X} c \mathrm{d}\gamma - \int_X \phi \mathrm{d}\mu - \int_X \psi \mathrm{d}\nu = \int_{X \times X} (c - \phi \oplus \psi) \mathrm{d}\gamma \tag{2.5}$$

The integrand $c - \phi \oplus \psi$ on the right-hand side is non-negative and thus the integral equals zero if and only if $\phi \oplus \psi = c$ γ-almost everywhere. □

Remark 2.9 (Primal-dual optimality conditions via Fenchel–Rockafellar duality)**.** The primal-dual optimality condition (2.8) can also be obtained via the condition given in Theorem 2.3. In the following we reuse the conventions established in the proof of Proposition 2.4.

In this case $F^*(A^*\gamma) = F^*((\mathrm{P}_1\gamma, \mathrm{P}_2\gamma)) = 0$ if $(\mathrm{P}_1\gamma, \mathrm{P}_2\gamma) = (\mu, \nu)$ and $+\infty$ otherwise, i.e. $\partial F^*(A^*\gamma) = \mathrm{C}(X) \times \mathrm{C}(X)$ if $(\mathrm{P}_1\gamma, \mathrm{P}_2\gamma) = (\mu, \nu)$ and $\emptyset$ otherwise. Therefore the condition $(\phi, \psi) \in -\partial F^*(A^*\gamma)$ is equivalent to the constraint that γ has the correct marginals (but not necessarily that γ is non-negative).

If $\gamma \in \partial G(A(\phi,\psi))$ then one must have that

$$G(A(\phi,\psi)+\eta) \geq G(A(\phi,\psi)) + \int_{X\times X} \eta \mathrm{d}\gamma$$

for all $\eta \in \mathrm{C}(X \times X)$, which implies that $\gamma \geq 0$ and $A(\phi,\psi) = c$ γ-almost everywhere.

Remark 2.10 (Relaxation of dual function space)**.** The inequality (2.5) holds for any $\phi \in \mathrm{L}^1(\mu)$, $\psi \in \mathrm{L}^1(\nu)$ with $\phi \oplus \psi \leq c$, so the inequality (2.4) $\leq$ (2.1) still holds when the function spaces for ϕ and ψ are relaxed to $\mathrm{L}^1(\mu)$ and $\mathrm{L}^1(\nu)$ (but we still impose the inequality constraint for all $(x,y) \in X^2$).

2.2 Wasserstein distance

A particularly important instance of the optimal transport problem (2.1) is when the ground transport cost function $c \in \mathrm{C}(X \times X)$ is chosen to be a power $p \in [1,\infty)$ of a metric on the base space X. This induces the celebrated Wasserstein distances. We focus here on the choice $p = 2$ and restrict ourselves to compact metric spaces. Of course, other p and non-compact X can also be considered. A more complete treatment including some valuable historical context and bibliographical notes are given in [51, Chapter 6]. We add a factor $1/2$ in the following definition for notational convenience in the remainder of this chapter.

Definition 2.11 (Wasserstein distance)**.** Let (X,d) be a compact metric space. Then the Wasserstein distance between two measures $\mu,\nu \in \mathcal{P}(X)$ is given by

$$\mathrm{W}(\mu,\nu) := \inf\left\{ \int_{X\times X} \tfrac{1}{2} d(x,y)^2 \,\mathrm{d}\gamma(x,y) \middle| \gamma \in \Gamma(\mu,\nu) \right\}^{1/2}. \tag{2.6}$$

Theorem 2.12. W is a metric on the set $\mathcal{P}(X)$ and metrizes the weak* topology.

A more general statement for non-compact spaces is shown in [51, Chapter 6], where a suitable notion of weak convergence must be considered and the convergence (or boundedness) of the second moment (or more generally, the p-th moment) must be verified in addition to the value of W. On compact domains this moment condition can be ignored. As preparation for later it will be instructive to recall the proof for the triangle inequality here, based on the gluing lemma.

Lemma 2.13 (Gluing lemma)**.** Let $\mu_1,\mu_2,\mu_3 \in \mathcal{P}(X)$ and let $\gamma_{12} \in \Gamma(\mu_1,\mu_2)$, $\gamma_{23} \in \Gamma(\mu_2,\mu_3)$. Then there is some $\gamma \in \mathcal{P}(X^3)$ such that $\mathrm{P}_{12}\gamma = \gamma_{12}$ and $\mathrm{P}_{23}\gamma = \gamma_{23}$. Here P_{12} and P_{23} denote the projection operators onto the joint $(1,2)$-marginal or $(2,3)$-marginal of γ.

Proof. This can be proved by an explicit construction using disintegration [3, Theorem 5.3.1], which formalizes the concept of conditional probabilities. For instance, there is

a family of probability measures $(\gamma_{12,y})_{y\in X}$, unique for μ_2-almost all y, such that for a continuous test function $\psi \in \mathrm{C}(X \times X)$ one has

$$\int_{X\times X} \psi(x,y)\mathrm{d}\gamma_{12}(x,y) = \int_X \left[\int_X \psi(x,y)\,\mathrm{d}\gamma_{12,y}(x)\right]\mathrm{d}\mu_2(y).$$

We call $(\gamma_{12,y})_{y\in X}$ the disintegration of γ_{12} with respect to its second marginal (which is μ_2). If $(\boldsymbol{x},\boldsymbol{y})$ is a pair of X-valued random variables with joint law γ_{12}, then $\gamma_{12,y}$ is the law of $\boldsymbol{x}$ when conditioned on $\boldsymbol{y} = y$ (in a suitable almost everywhere sense).

Likewise, we note by $(\gamma_{23,y})_y$ the disintegration of γ_{23} with respect to its first marginal (which is μ_2). Let then $\gamma \in \mathcal{P}(X^3)$ be the measure that is characterized by

$$\int_{X\times X\times X} \psi(x,y,z)\,\mathrm{d}\gamma(x,y,z) = \int_X \left[\int_X \psi(x,y,z)\,\mathrm{d}\gamma_{12,y}(x)\,\mathrm{d}\gamma_{23,y}(z)\right]\mathrm{d}\mu_2(y)$$

for $\psi \in \mathrm{C}(X^3)$. By choosing ψ that are constant with respect to the first or third argument one can then verify that γ satisfies the above requirements on marginals. □

Proof of the triangle inequality. Let $\mu_1,\mu_2,\mu_3 \in \mathcal{P}(X)$ and let $\gamma_{12} \in \Gamma(\mu_1,\mu_2)$, $\gamma_{23} \in \Gamma(\mu_2,\mu_3)$. Let $\gamma \in \mathcal{P}(X^3)$, as provided by the gluing lemma and set $\gamma_{13} := \mathrm{P}_{13}\gamma$ to be the joint $(1,3)$-marginal. Then one finds that $\gamma_{13} \in \Gamma(\mu_1,\mu_3)$ and therefore

$$\begin{aligned}
\mathrm{W}(\mu_1,\mu_3) &\le \left(\int_{X\times X} \tfrac{1}{2}d(x,z)^2\mathrm{d}\gamma_{13}(x,z)\right)^{1/2}\\
&= \left(\int_{X\times X\times X} \tfrac{1}{2}d(x,z)^2\mathrm{d}\gamma(x,y,z)\right)^{1/2}\\
&\le \left(\int_{X\times X\times X} \tfrac{1}{2}[d(x,y)+d(y,z)]^2\mathrm{d}\gamma(x,y,z)\right)^{1/2}\\
&\le \left(\int_{X\times X\times X} \tfrac{1}{2}d(x,y)^2\mathrm{d}\gamma(x,y,z)\right)^{1/2} + \left(\int_{X\times X\times X} \tfrac{1}{2}d(y,z)^2\mathrm{d}\gamma(x,y,z)\right)^{1/2}\\
&= \left(\int_{X\times X\times X} \tfrac{1}{2}d(x,y)^2\mathrm{d}\gamma_{12}(x,y)\right)^{1/2} + \left(\int_{X\times X\times X} \tfrac{1}{2}d(y,z)^2\mathrm{d}\gamma_{23}(y,z)\right)^{1/2}.
\end{aligned}$$

Here we used the triangle inequality on d in the second inequality and the Minkowski inequality in $\mathrm{L}^2(X^3,\gamma)$ in the third inequality. The claim then follows by taking the infimum over γ_{12} and γ_{23} over $\Gamma(\mu_1,\mu_2)$ and $\Gamma(\mu_2,\mu_3)$, respectively. □

Theorem 2.14 (Brenier)**.** Let X be a compact, convex subset of $\mathbb{R}^d$, let $\mu,\nu \in \mathcal{P}(X)$ and assume $\mu \ll \mathcal{L}$. Then the optimal transport plan for the Wasserstein distance from μ to ν is unique and supported on the graph of an optimal transport *map* $T : X \to X$. The map T is given by

$$T = \mathrm{id} - \nabla\phi = \nabla\left(\tfrac{1}{2}|\cdot|^2 - \phi\right) \tag{2.7}$$

where ϕ is a maximizer of the dual problem of (2.6), i.e. for (2.4) with $c = \frac{1}{2}d^2$. In particular, all such maximizers are differentiable for μ-almost every $x \in X$ and these gradients agree μ-almost everywhere for all maximizers. The function $\frac{1}{2}|\cdot|^2 - \phi$ is convex.

Remark 2.15. There are many versions of this theorem with varying levels of generality for the domain X, the measure μ, and the cost function, see for instance [12, 27, 42, 50, 51, 48]. The above simple version is a special case of [48, Theorem 1.17].

Definition 2.16 (Geodesic space)**.** We say a metric space (X, d) is geodesic, if for any two $x, y \in X$ there exists a curve $Z : [0,1] \to X$ with $Z(0) = x$ and $Z(1) = y$, such that $d(Z(s), Z(t)) = |s-t| \cdot d(x, y)$ for all $s, t \in [0,1]$. For given x, y we call the corresponding curve Z a *(constant speed) geodesic* from x to y.

Example 2.17. Let X be a convex subset of $\mathbb{R}^d$ with $d(x, y) = |x - y|$. For given $x, y \in X$, the unique constant speed geodesic between them is given by $Z(x, y, \cdot) : [0,1] \ni t \mapsto (1-t) \cdot x + t \cdot y$.

Theorem 2.18 (Wasserstein space inherits geodesic property)**.** Assume that (X, d) is a geodesic space and that there is a measurable selection of constant speed geodesics, i.e. there is a measurable map $Z : [0,1] \times X \times X \to X$, such that for fixed $(x, y) \in X \times X$, the map $t \mapsto Z(t, x, y)$ is a constant speed geodesic from x to y. Then $(\mathcal{P}(X), \mathrm{W})$ is a geodesic space. For $\mu, \nu \in \mathcal{P}(X)$ a constant speed geodesic is given by

$$[0,1] \ni t \mapsto \rho_t := Z(t, \cdot, \cdot)_{\#}\gamma$$

where γ is an optimal transport plan for $\mathrm{W}(\mu, \nu)$ in (2.6).

Proof. The proof uses similar ideas as that for the triangle inequality of W in Theorem 2.12. We need to show that $\mathrm{W}(\rho_s, \rho_t) = |s-t| \cdot \mathrm{W}(\mu, \nu)$ for $s, t \in [0,1]$. W.l.o.g. assume $s < t$. For this we will first construct admissible transport plans $\gamma_{s,t} \in \Gamma(\rho_s, \rho_t)$ as follows. ρ_s was constructed by observing that mass of γ at (x, y) should at time s be at position $Z(s, x, y)$. At time t it should have moved to $Z(t, x, y)$. From this we intuit that

$$\gamma_{s,t} := \hat{Z}(s, t, \cdot, \cdot)_{\#}\gamma$$

for

$$\hat{Z} : [0,1] \times [0,1] \times X \times X \to X \times X, \qquad (s, t, x, y) \mapsto (Z(s, x, y), Z(t, x, y))$$

is a reasonable transport plan. Indeed, it is easy to verify that $\gamma_{s,t} \in \Gamma(\rho_s, \rho_t)$. For its induced cost we obtain

$$\begin{aligned}
\mathrm{W}(\rho_s, \rho_t)^2 &\leq \int_{X \times X} \tfrac{1}{2} d(v, w)^2 \mathrm{d}\gamma_{s,t}(v, w) \\
&= \int_{X \times X} \tfrac{1}{2} d\big(Z(s, x, y), Z(t, x, y)\big)^2 \mathrm{d}\gamma(x, y) \\
&= \int_{X \times X} \tfrac{1}{2} \cdot |s-t|^2 \cdot d(x, y)^2 \mathrm{d}\gamma(x, y) \\
&= |s-t|^2 \cdot \mathrm{W}(\mu, \nu)^2
\end{aligned} \tag{2.8}$$

where we used the constant speed geodesic property of $Z(\cdot, x, y)$ in the second equality. We still need to show that the first inequality is actually an equality. For this we use (2.8) on the time pairs $(0, s)$, (s, t), and $(t, 1)$ to obtain

$$\mathrm{W}(\mu, \rho_s) + \mathrm{W}(\rho_s, \rho_t) + \mathrm{W}(\rho_t, \nu) \leq \mathrm{W}(\mu, \nu) \cdot (|0 - s| + |s - t| + |t - 1|) = \mathrm{W}(\mu, \nu).$$

The triangle inequality on $(\mathcal{P}(X), \mathrm{W})$ yields the reverse inequality and therefore (2.8) must actually be an equality. □

Example 2.19. Returning to Example 2.17, we find that the map Z is clearly measurable as it is continuous. Therefore, in a Wasserstein geodesic, mass particles move with constant speed on a straight line from their initial to their target location.

Example 2.20 (Geodesics in the setting of Brenier's theorem)**.** Consider the setting of Brenier's theorem (Theorem 2.14). If an optimal plan $\gamma \in \Gamma(\mu, \nu)$ for $\mathrm{W}(\mu, \nu)$ is induced by some map $T = \mathrm{id} - \nabla\phi$, then the induced geodesic in the sense of Theorem 2.18 is given by

$$\rho_t := T_{t\#}\mu \quad \text{with} \quad T_t = (1 - t) \cdot \mathrm{id} + t \cdot T = \mathrm{id} - t \cdot \nabla\phi. \tag{2.9}$$

μ-almost everywhere, mass particles initially located at x, will move to $T(x)$ along the straight line between the points with velocity $T(x) - x = -\nabla\phi(x)$. $-\nabla\phi(x)$ should therefore be interpreted as the Lagrangian velocity field of the moving particles and this Lagrangian velocity is constant in time for particles in a Wasserstein geodesic.

We have just seen that once an optimal transport plan γ is known, a whole geodesic with respect to W can be constructed and we know the position of any mass particle at all times. In fact, more knowledge about these intermediate positions can be extracted. We will now show that mass particles do not collide at intermediate times.

Proposition 2.21. Consider the setting of Example 2.19. Let $(\phi, \psi) \in \mathrm{C}(X)^2$ such that $\phi \oplus \psi \leq c$ on X^2 and let $t \in (0, 1)$. The map $(x, y) \mapsto Z(t, x, y) = (1 - t) \cdot x + t \cdot y$ is injective on the contact set of (ϕ, ψ).

Proof. Assume that there exists $z \in Z$ such that $z = Z(t, x, y) = Z(t, x', y')$ where (x, y) and (x', y') are two points in the contact set of (ϕ, ψ). One obtains

$$|x - y|^2 = \tfrac{|x-z|^2}{t} + \tfrac{|z-y|^2}{1-t}$$

and likewise for (x', y') (cf. Lemma 2.45). An explicit geometric calculation then yields that

$$\tfrac{1}{2}|x - y|^2 + \tfrac{1}{2}|x' - y'|^2 \geq \tfrac{1}{2}|x - y'|^2 + \tfrac{1}{2}|x' - y|^2$$

with equality if and only if $(x, y) = (x', y')$. One then obtains the chain of inequalities

$$\begin{aligned}[\phi(x) + \psi(y)] + [\phi(x') + \psi(y')] &= \tfrac{1}{2}|x - y|^2 + \tfrac{1}{2}|x' - y'|^2 \\ &\geq \tfrac{1}{2}|x - y'|^2 + \tfrac{1}{2}|x' - y|^2 \geq [\phi(x) + \psi(y')] + [\phi(x') + \psi(y)]\end{aligned}$$

where the first inequality must now be an equality and therefore $(x, y) = (x', y')$. □

The following statement is then an immediate consequence of the primal-dual optimality condition Proposition 2.8 and Proposition 2.21.

Corollary 2.22 (Mass particles do not collide at intermediate times)**.** Let $t \in (0,1)$. For an optimal transport plan γ the map $Z(t,\cdot,\cdot)$ is injective γ-almost everywhere.

Remark 2.23. When restricting to the setting of Brenier's theorem, this implies that the map $T_t := \mathrm{id} - t \cdot \nabla\phi$ as introduced in Example 2.20 is invertible for $t \in (0,1)$. This can also be deduced from the fact that $\frac{1}{2}|\cdot|^2 - \phi$ is convex by Brenier's theorem and thus for $t \in (0,1)$ the map $\frac{1}{2}|\cdot|^2 - t \cdot \phi$ is strictly convex, i.e. its gradient (which exists μ-almost everywhere) is invertible.

Another corollary of Proposition 2.21 is that for $0 < t < s < 1$ the intermediate transport plans $\gamma_{s,t} \in \Gamma(\rho_s, \rho_t)$ constructed in the proof of Theorem 2.18 are deterministic in the setting of Example 2.19, i.e. they are supported on the graph of a map, obtained by composing the inverse of $Z(s,\cdot,\cdot)$ with $Z(t,\cdot,\cdot)$.

A deeper exploration of these ideas is given in [51, Chapter 8] on the *Monge–Mather shortening principle*, which includes local spatial Lipschitz regularity of the optimal transport from intermediate times.

2.3 Benamou–Brenier formula

2.3.1 Definition and basic properties

The Kantorovich formulation of the Wasserstein distance, (2.6) can be seen as a Lagrangian description: particles are identified by their initial and final location; which imply their location at intermediate times (and thus the whole Wasserstein geodesic) via Theorem 2.18. One can also adopt a Eulerian point of view: What is the distribution of mass at some intermediate time $t \in (0,1)$? And what is the momentary velocity of a mass particle that passes through position $z \in X$ at time $t \in (0,1)$? As it turns out, it is also possible to find Wasserstein geodesics in the Eulerian picture via the celebrated Benamou–Brenier formula [9].

Remark 2.24 (Intuitive motivation)**.** Consider the setting of Example 2.20. Recall that the intermediate distribution of particles at some time $t \in (0,1)$ is given by $\rho_t = T_{t\#}\mu$ and since the Lagrangian velocity field $-\nabla\phi$ is constant in time, the intermediate Eulerian velocity field is given by $v_t := -\nabla\phi \circ T_t^{-1}$ (recall that by Remark 2.23 T_t is indeed invertible).

Let now $\psi \in \mathrm{C}^1([0,1] \times X)$. Then a simple computation (similar to the proof of Proposition 2.34) yields that

$$\int_0^1 \int_X (\partial_t \psi)\, \mathrm{d}\rho_t \mathrm{d}t + \int_0^1 \int_X \nabla\psi \cdot v_t \mathrm{d}\rho_t \mathrm{d}t = \int_X \psi(1,\cdot)\, \mathrm{d}\nu - \int_X \psi(0,\cdot)\, \mathrm{d}\mu.$$

By formally using integration by parts on the derivatives acting on ψ we say that the pair of curves $(\rho_t, v_t)_t$ solves the continuity equation

$$\partial_t \rho_t + \nabla \cdot (v_t \cdot \rho_t) = 0$$

in a weak sense with temporal boundary conditions $\rho_0 = \mu$, $\rho_1 = \nu$.

So the Wasserstein geodesic with the corresponding velocity field constructed by the Kantorovich optimal transport plan (or the Brenier map) is a solution $(\rho_t, v_t)_t$ to this equation. In this section we will show that among all such solutions it is the one which minimizes the action functional

$$(\rho_t, v_t)_t \mapsto \frac{1}{2}\int_0^1 \int_X |v_t(x)|^2 \mathrm{d}\rho_t \,\mathrm{d}t. \tag{2.10}$$

In the following we will formulate this minimization problem rigorously by using two transformations. First, we switch from a velocity field v_t to a momentum measure $\omega_t := v_t \cdot \rho_t$. One then recovers v_t as density $\frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t}$. This change of variables turns the continuity equation into a linear PDE constraint and the action functional into a convex function. Second, instead of considering curves $(\rho_t, \omega_t)_t$ in $\mathcal{M}(X)^{1+d}$, we only consider two measures $(\rho, \omega) \in \mathcal{M}([0,1] \times X)^{1+d}$ over space and time. Intuitively, (ρ_t, ω_t) can be recovered by considering the disintegration of (ρ, ω) with respect to the time-axis at t. The validity of this change of variables and the disintegration is established in Proposition 2.30.

We will now introduce rigorous definitions for the concepts introduced in Remark 2.24, starting with the continuity equation for measures on space and time.

Definition 2.25 (Distributional continuity equation)**.** Let $\mu, \nu \in \mathcal{P}(X)$. A pair $(\rho, \omega) \in \mathcal{M}([0,1] \times X) \times \mathcal{M}([0,1] \times X)^d$ is said to solve the distributional continuity equation with temporal boundary conditions μ and ν if

$$\int_{[0,1]\times X} (\partial_t \psi)\,\mathrm{d}\rho + \int_{[0,1]\times X} \nabla\psi \cdot \mathrm{d}\omega = \int_X \psi(1,\cdot)\,\mathrm{d}\nu - \int_X \psi(0,\cdot)\,\mathrm{d}\mu \tag{2.11}$$

for all $\psi \in C^1([0,1] \times X)$. We denote the set of solutions by $\mathcal{CE}(\mu, \nu)$.

Next, we prepare a rigorous definition of the action functional (2.10).

Remark 2.26 (1-homogeneous functions, integral functionals, and conjugation)**.** A function $f : \mathbb{R}^n \to (-\infty, \infty]$ is 1-homogeneous if

$$f(\lambda \cdot s) = \lambda \cdot f(s)$$

for all $\lambda \geq 0$, $s \in \mathbb{R}^n$ (with the convention $0 \cdot \infty = 0$ for $f(s) = \infty$ and $\lambda = 0$).

In the following, assume that f only takes values in $[0, \infty]$ and that f is convex and lower-semicontinuous. Let $\mu \in \mathcal{M}(X)^n$ and let $\sigma \in \mathcal{M}_+(X)$ such that $\mu \ll \sigma$. Then one can consider the following integral functional:

$$F(\mu, \sigma) := \int_X f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}\right) \mathrm{d}\sigma$$

Due to the 1-homogeneity of f the integral does not actually depend on the choice of σ as long as $\mu \ll \sigma$, i.e. F can be interpreted as being solely a function of μ. Under the

above assumptions on f, F is convex and weak* lower-semicontinuous [2, Proposition 2.37 and Theorem 2.38]. Note that one can also consider spatially varying f that also explicitly depend on the position x in the integral.

As f is 1-homogeneous, its Fenchel–Legendre conjugate f^* is the indicator function of a closed, convex set $C \subset \mathbb{R}^n$, i.e. $f^* = \iota_C$ and in fact $C = \partial f(0)$. The conjugate of F is then the functional

$$F^* : \mathrm{C}(X) \to \mathbb{R} \cup \{\infty\}, \qquad \eta \mapsto \begin{cases} 0 & \text{if } \eta(x) \in \partial f(0) \text{ for all } x \in X, \\ +\infty & \text{else.} \end{cases}$$

Definition 2.27 (Action functional)**.** Let

$$\Phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}, \qquad (r, w) \mapsto \begin{cases} \frac{|w|^2}{2r} & \text{if } r > 0, \\ 0 & \text{if } (r, w) = (0, 0), \\ +\infty & \text{else.} \end{cases}$$

Φ is convex and lower semi-continuous (jointly in (r, w)), since its sub-level sets are closed and convex. Moreover, Φ is 1-homogeneous. The action is then given by

$$\mathcal{A}(\rho, \omega) = \int_{[0,1]\times X} \Phi\left(\tfrac{\mathrm{d}(\rho,\omega)}{\mathrm{d}\sigma}\right) \mathrm{d}\sigma(t, x)$$

where σ is any measure in $\mathcal{M}_+([0, 1] \times X)$ such that $(\rho, \omega) \ll \sigma$. Due to Remark 2.26 the definition does not depend on the choice of σ.

Remark 2.28. If $\omega \ll \rho$ so that in particular $\omega = v \cdot \rho$ for some velocity field $v \in \mathrm{L}^1([0, 1] \times X, \mathbb{R}^d)$, then one can choose $\sigma = \rho$ for the evaluation of $\mathcal{A}(\rho, \omega)$ and obtains

$$\mathcal{A}(\rho, \omega) = \frac{1}{2} \int_{[0,1]\times X} |v(t, x)|^2 \mathrm{d}\rho(t, x)$$

in line with Remark 2.24.

Now we have gathered the ingredients for the rigorous definition of the Benamou–Brenier formula.

Definition 2.29 (Benamou–Brenier formula)**.**

$$W_{\mathrm{BB}}(\mu, \nu)^2 := \inf \{\mathcal{A}(\rho, \omega) | (\rho, \omega) \in \mathcal{CE}(\mu, \nu)\} \tag{2.12}$$

In (2.12) ρ and ω were only defined as measures on $[0, 1] \times X$, which does not necessarily imply a well-defined notion of mass distribution at intermediate times $t \in (0, 1)$. Fortunately, for pairs $(\rho, \omega) \in \mathcal{CE}(\mu, \nu)$ with $\mathcal{A}(\rho, \omega) < \infty$, one can show that both measures can be decomposed into 'time-slices' such that for Lebesgue-almost every time t the spatial arrangement of mass is fixed. Moreover, ω is actually dominated by ρ and thus it can be written as $\omega = v \cdot \rho$ for a suitable velocity field v. This justifies both transformations sketched in Remark 2.24.

Proposition 2.30 (Time-disintegration of ρ and ω, and velocity field)**.** Let $(\rho, \omega) \in \mathcal{CE}(\mu, \nu)$ such that $\mathcal{A}(\rho, \omega) < \infty$. Then the following holds:

(i) $\rho \in \mathcal{M}_+([0,1] \times X)$, $\|\rho\|_{\mathcal{M}([0,1]\times X)} = 1$.

(ii) $\mathrm{T}_\# \rho = \mathcal{L} \llcorner [0,1]$ where $\mathrm{T} : [0,1] \times X \to [0,1]$, $(t,x) \mapsto t$ and $\mathcal{L} \llcorner [0,1]$ denotes the Lebesgue measure on $[0,1]$. There is a Lebesgue-a.e. unique measurable family of measures $(\rho_t)_{t\in[0,1]}$ with $\rho_t \in \mathcal{P}(X)$ such that for any $\psi \in \mathrm{C}([0,1] \times X)$ one has

$$\int_{[0,1]\times X} \psi(t,x) \,\mathrm{d}\rho(t,x) = \int_{[0,1]} \left[\int_X \psi(t,x) \,\mathrm{d}\rho_t(x) \right] \mathrm{d}t \,. \tag{2.13}$$

(iii) $\omega \ll \rho$. This implies that ω can also be decomposed into time-slices and that it can be written as $\omega = v \cdot \rho$ for some $v \in \mathrm{L}^1(\rho, \mathbb{R}^d)$. $v(t,\cdot)$ is then the Eulerian velocity field of particles at time $t \in [0,1]$.

Proof. First claim: If $\rho \notin \mathcal{M}_+([0,1] \times X)$ then $\frac{\mathrm{d}\rho}{\mathrm{d}\sigma} < 0$ for a set that is not σ-negligible, for every $\sigma \in \mathcal{M}_+([0,1] \times X)$ with $\rho \ll \sigma$ and thus $\mathcal{A}(\rho, \omega) = \infty$. Using the test function $\psi(t,x) = t$ in the definition of $\mathcal{CE}(\mu, \nu)$ we get

$$\int_{[0,1]\times X} \mathrm{d}\rho(t,x) = \int_X \mathrm{d}\nu(x) = 1 \,.$$

This is the total variation norm of ρ since ρ is non-negative.

Second claim: Let $f \in \mathrm{C}([0,1])$ and let

$$F(t) := \int_0^t f(s) \,\mathrm{d}s \text{ for } t \in [0,1].$$

Also let $\mathrm{T} : [0,1] \times X \to [0,1]$, $(t,x) \mapsto t$ be the projection to the time coordinate. By construction $\partial_t F = f$ and $F \circ \mathrm{T} \in \mathrm{C}^1([0,1] \times X)$. So from the continuity equation we know that

$$\begin{aligned} \int_{[0,1]} f \,\mathrm{d}(\mathrm{T}_\# \rho) &= \int_{[0,1]\times X} (\partial_t F) \circ \mathrm{T} \,\mathrm{d}\rho \\ &= \int_X (F \circ \mathrm{T})(1,\cdot) \,\mathrm{d}\nu - \int_X (F \circ \mathrm{T})(0,\cdot) \,\mathrm{d}\mu = F(1) - F(0) \,. \end{aligned}$$

So the integral against $\mathrm{T}_\# \rho$ coincides with the Lebesgue measure for all continuous test functions, hence $\mathrm{T}_\# \rho = \mathcal{L} \llcorner [0,1]$.

Third claim: Let σ be some reference measure with $(\rho, \omega) \ll \sigma$. If $\omega \not\ll \rho$ then there must be a set $A \subset [0,1] \times X$ with $\sigma(A) > 0$ where $\frac{\mathrm{d}\rho}{\mathrm{d}\sigma} = 0$ but $\frac{\mathrm{d}\omega}{\mathrm{d}\sigma} \neq 0$ and consequently $\Phi(\frac{\mathrm{d}\rho}{\mathrm{d}\sigma}, \frac{\mathrm{d}\omega}{\mathrm{d}\sigma}) = \infty$ and thus $\mathcal{A}(\rho, \omega) = \infty$. □

With these basic regularity properties it is then easy to establish existence of minimizers.

Proposition 2.31. Minimizers of the Benamou–Brenier formulation (2.12) exist.

Proof. If $W_{\mathrm{BB}}(\mu,\nu)=\infty$, one merely needs to show that $\mathcal{CE}(\mu,\nu)$ is non-empty. This will be established in the proof of Proposition 2.34. So assume from now on that $W_{\mathrm{BB}}(\mu,\nu)<\infty$.

Let $(\rho_n,\omega_n)_n$ be a minimizing sequence, which implies $(\rho_n,\omega_n)\in\mathcal{CE}(\mu,\nu)$ for all n and that $\mathcal{A}(\rho_n,\omega_n)\leq C$ for some $C<\infty$. By Proposition 2.30 $\rho_n\geq 0$, $\|\rho_n\|_{\mathcal{M}([0,1]\times X)}=1$ and $\omega_n\ll\rho_n$. Therefore we can pick $\sigma=\rho_n$ as reference measure in the definition of the action $\mathcal{A}(\rho_n,\omega_n)$ and obtain

$$\mathcal{A}(\rho_n,\omega_n)=\int_{[0,1]\times X}\Phi\left(\tfrac{\mathrm{d}\rho_n}{\mathrm{d}\rho_n},\tfrac{\mathrm{d}\omega_n}{\mathrm{d}\rho_n}\right)\mathrm{d}\rho_n=\frac{1}{2}\int_{[0,1]\times X}\left|\tfrac{\mathrm{d}\omega_n}{\mathrm{d}\rho_n}\right|^2\mathrm{d}\rho_n.$$

With this we can bound the total variation norm of ω_n:

$$\begin{aligned}\|\omega_n\|_{\mathcal{M}([0,1]\times X)^d}=\int_{[0,1]\times X}\left|\tfrac{\mathrm{d}\omega_n}{\mathrm{d}\rho_n}\right|\mathrm{d}\rho_n&\leq\left(\int_{[0,1]\times X}\left|\tfrac{\mathrm{d}\omega_n}{\mathrm{d}\rho_n}\right|^2\mathrm{d}\rho_n\right)^{1/2}\\&=(2\,\mathcal{A}(\omega_n,\rho_n))^{1/2}\leq(2C)^{1/2}\end{aligned}$$

where we used Jensen's inequality. Therefore, the sequence is uniformly bounded in norm and thus, by the Banach–Alaoglu theorem it must have a convergent subsequence with limit (ρ,ω). Since $\mathcal{CE}(\mu,\nu)$ is weak$*$-closed, one has $(\rho,\omega)\in\mathcal{CE}(\mu,\nu)$. Moreover, since Φ is convex, lower semi-continuous and 1-homogeneous, the functional $\mathcal{A}$ is lower semi-continuous (Remark 2.26). Therefore, (ρ,ω) is a minimizer of W_{BB}. □

Remark 2.32 (Absolutely continuous curves in Wasserstein space)**.** It turns out that one can establish much more regularity than in Proposition 2.30. When $\mathcal{A}(\rho,\omega)<\infty$ for $(\rho,\omega)\in\mathcal{CE}(\mu,\nu)$, it is possible to show that the family $(\rho_t)_t$ can be chosen such that the path $[0,1]\ni t\mapsto\rho_t$ is continuous in $(\mathcal{P}(X),\mathrm{W})$, in fact *absolutely continuous*. It then becomes meaningful to consider continuity equations without prescribed temporal boundary conditions. (This is achieved by enforcing (2.11) only for C^1 test functions ψ that are compactly supported in $(0,1)\times X$.) Conversely, one finds that if a curve $t\mapsto\rho_t$ is absolutely continuous, then there is some ω such that (ρ,ω) solve the continuity equation and $\mathcal{A}(\rho,\omega)<\infty$. We refer to [48, Section 5.3] for a more detailed discussion of this topic.

Remark 2.33 (Applications and generalizations)**.** The functional $\mathcal{A}$ can be used for much more than for merely finding geodesics in Wasserstein space. It can be used to model problems where the curve ρ_t interacts with other things at intermediate times, such as measurement constraints. An example for dynamic medical imaging where $\mathcal{A}$ acts as temporal regularizer is developed in [49, 41]. It is also a natural starting point for generalizing the Wasserstein distance, for instance to *unbalanced transport* between measures of unequal mass [8, 32, 37, 22, 23] or to vector-valued measures [14, 19].

2.3.2 Equivalence to Kantorovich formulation

Proposition 2.34. The Benamou–Brenier formula provides a lower bound for the Kantorovich formulation, i.e. (2.12) $\leq$ (2.6).

For the proof we will use the following small Lemma, which ca be found, for instance in [13, Lemma 3.15]. It is a corollary of the disintegration theorem and Jensen's inequality.

Lemma 2.35. Let A, B be measurable spaces, $T : A \to B$ measurable, $\mu \in \mathcal{M}(A)^n$, $\nu \in \mathcal{M}_+(A)$, $\mu \ll \nu$ (which implies $T_\#\mu \ll T_\#\nu$), and $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex. Then

$$\int_B f\left(\frac{\mathrm{d}T_\#\mu}{\mathrm{d}T_\#\nu}\right) \mathrm{d}T_\#\nu \leq \int_A f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu.$$

Proof of Proposition 2.34. The strategy is to construct admissible candidates for (2.12) from admissible candidates for (2.6) that have a potentially lower objective value. Let $\gamma \in \Gamma(\mu, \nu)$ and let $Z : X \times X \times [0,1]$, $Z(x, y, t) = (1-t) \cdot x + t \cdot y$ be the map that parametrizes all pairwise constant speed geodesics. Let $v : X \times X \to \mathbb{R}^d$ be given by $v(x, y) = y - x$. Then for $t \in [0, 1]$ introduce the measures $\rho_t := Z(\cdot, \cdot, t)_\#\gamma$ and $\omega_t := Z(\cdot, \cdot, t)_\#(v \cdot \gamma)$. By Theorem 2.18, the map $t \mapsto \rho_t$ describes a Wasserstein geodesic from μ to ν (if γ is minimal) and ω_t is the Eulerian momentum field, which is constructed from the Lagrangian momentum field $v \cdot \gamma$, by putting the mass to the appropriate intermediate location at intermediate time t in the analogous way as for ρ_t. Observe that $\partial_t Z(x, y, t) = v(x, y)$.

From the family $(\rho_t)_t$ we can then construct a measure ρ in $\mathcal{P}([0,1] \times X)$ via the characterization

$$\int_{[0,1]\times X} \psi \mathrm{d}\rho := \int_0^1 \left[\int_X \psi(t, \cdot) \mathrm{d}\rho_t\right] \mathrm{d}t$$

for all $\psi \in \mathrm{C}([0,1] \times X)$. This is essentially the reversal of the decomposition (2.13). Analogously, we construct ω from the family $(\omega_t)_t$.

We find that the pair (ρ, ω) is in $\mathcal{CE}(\mu, \nu)$, since for $\psi \in \mathrm{C}^1([0,1] \times X)$ one has

$$\begin{aligned}
&\int_{[0,1]\times X} \partial_t\psi \mathrm{d}\rho + \int_{[0,1]\times X} \nabla\psi \cdot \mathrm{d}\omega \\
&= \int_0^1 \left[\int_X \partial_t\psi(t, \cdot) \mathrm{d}\rho_t + \int_X \nabla\psi(t, \cdot) \cdot \mathrm{d}\omega_t\right] \mathrm{d}t \\
&= \int_0^1 \left[\int_{X\times X} [\partial_t\psi(t, Z(x, y, t)) + \nabla\psi(t, Z(x, y, t)) \cdot v(x, y)] \,\mathrm{d}\gamma(x, y)\right] \mathrm{d}t \\
&= \int_0^1 \left[\int_{X\times X} \left[\frac{\mathrm{d}}{\mathrm{d}t}\psi(t, Z(x, y, t))\right] \mathrm{d}\gamma(x, y)\right] \mathrm{d}t \\
&= \int_{X\times X} [\psi(1, Z(x, y, 1)) - \psi(0, Z(x, y, 0))] \,\mathrm{d}\gamma(x, y) \\
&= \int_X \psi(1, \cdot) \mathrm{d}\nu - \int_X \psi(0, \cdot) \mathrm{d}\mu.
\end{aligned}$$

And finally, as claimed, we find that $\mathcal{A}(\rho,\omega) \leq \frac{1}{2}\int |x-y|^2 \mathrm{d}\gamma(x,y)$. For this we use that $v\gamma \ll \gamma$ implies that $\omega \ll \rho$ and therefore we can use ρ as reference measure σ in the definition of $\mathcal{A}$. We also use Lemma 2.35 on the map $Z(\cdot,\cdot,t)$. One finds:

$$\begin{aligned}\mathcal{A}(\rho,\omega) &= \int_{[0,1]\times X} \Phi\left(\tfrac{\mathrm{d}\rho}{\mathrm{d}\rho}, \tfrac{\mathrm{d}\omega}{\mathrm{d}\rho}\right) \mathrm{d}\rho = \int_0^1 \left[\int_X \Phi\left(\tfrac{\mathrm{d}\rho_t}{\mathrm{d}\rho_t}, \tfrac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t}\right) \mathrm{d}\rho_t\right] \mathrm{d}t \\ &\leq \int_0^1 \left[\int_{X\times X} \Phi(1,v)\,\mathrm{d}\gamma\right] \mathrm{d}t = \tfrac{1}{2}\int |x-y|^2 \mathrm{d}\gamma(x,y). \qquad \square\end{aligned}$$

Remark 2.36 (Reverse inequality)**.** The proof for the reverse inequality (and thus the equivalence of the definitions (2.6) and (2.12) for the Wasserstein distance) is more involved. The basic idea is to consider an admissible pair $(\rho,\omega) \in \mathcal{CE}(\mu,\nu)$ with $\mathcal{A}(\rho,\omega) < \infty$ and to construct a corresponding transport plan $\gamma \in \Gamma(\mu,\nu)$ with a potentially lower objective value.

Intuitively, this plan γ is constructed by tracing the trajectories of mass particles as described by the Eulerian velocity field $v = \frac{\mathrm{d}\omega}{\mathrm{d}\rho}$ to obtain once more a Lagrangian description. Formally this is done by solving the initial value problems

$$\partial_t T(t,x) = v(t,T(t,x)), \qquad T(0,x) = x \tag{2.14}$$

for $t \in [0,1]$ and μ-almost all $x \in X$. However, the fact that $v \in \mathrm{L}^1(\rho;\mathbb{R}^d)$ provides far too little regularity for this ODE to be well-posed and problems may arise from the fact that μ contains atoms, and that their mass may need to be split for the optimal transport. This can be remedied by smoothing arguments, as outlined in [50, Theorem 8.1] or [48, Section 5.3.3]. We skip these arguments here and assume for simplicity that (2.14) has a unique solution and that $\mu \ll \mathcal{L}$. Using a calculation analogous to that in the proof of Proposition 2.34 for the continuity equation, one can then show that $\rho_t = T(t,\cdot)_\# \mu$ for almost all t, and in fact, ρ_t can be chosen such that this holds for all t, see Remark 2.32. Therefore in particular $\nu = T(1,\cdot)_\#\mu$ and thus the plan $\gamma := (\mathrm{id}, T(1,\cdot))_\#\mu$ is admissible, i.e. $\gamma \in \Gamma(\mu,\nu)$. One can then show that the Kantorovich objective for γ is bounded from above by the Benamou–Brenier objective along the following lines:

$$\begin{aligned}\int_{X\times X} \tfrac{1}{2}|x-y|^2 \mathrm{d}\gamma(x,y) &= \int_X \tfrac{1}{2}|T(1,x) - T(0,x)|^2 \mathrm{d}\mu(x) \\ &\leq \int_X \int_0^1 \tfrac{1}{2}|\partial_t T(t,x)|^2 \mathrm{d}\mu(x) \\ &= \int_X \int_0^1 \tfrac{1}{2}|v(t,x)|^2 \mathrm{d}T(t,\cdot)_\#\mu(x)\end{aligned}$$

2.3.3 Duality

Now we consider a dual formulation of (2.12). Applying Theorem 2.3 to (2.12) yields the following problem. We will subsequently analyze its structure and interpretation in more depth.

Proposition 2.37 (Dual Benamou–Brenier formula)**.**

$$W_{\mathrm{BB}}(\mu,\nu) = \sup\left\{ \int_X \Psi(1,\cdot)\,\mathrm{d}\nu - \int_X \Psi(0,\cdot)\,\mathrm{d}\mu \,\middle|\, \Psi \in \mathrm{C}^1([0,1]\times X), \right.$$
$$\left. \partial_t\Psi + \frac{1}{2}|\nabla\Psi|^2 \le 0 \right\} \quad (2.15)$$

Proof. We will proceed analogously to the proof of Proposition 2.4 and write the supremum in (2.15) in the form of (2.3). For this choose $U = \mathrm{C}^1([0,1]\times X)$ and $V = \mathrm{C}([0,1]\times X)^{1+d}$, $A : \Psi \mapsto (\partial_t\Psi, \nabla\Psi)$, $F(\Psi) = \int_X \Psi(1,\cdot)\mathrm{d}\nu - \int_X \Psi(0,\cdot)\mathrm{d}\mu$ and

$$G : \mathrm{C}([0,1]\times X)\times \mathrm{C}([0,1]\times X)^d \ni (\xi,\zeta) \mapsto \begin{cases} 0 & \text{if } \xi + \frac{1}{2}|\zeta|^2 \le 0, \\ +\infty & \text{else.} \end{cases}$$

F is finite for all $\Psi \in U$. Choosing $\Psi(t,x) := -t$ yields a candidate such that G is continuous at $A\Psi$.

The dual problem is then an optimization over $V^* = \mathcal{M}([0,1]\times X)^{1+d}$ and we denote the candidates as pairs (ρ,ω) as above. Observing that

$$\Phi^*(a,b) = \begin{cases} 0 & \text{if } a + |b|^2/2 \le 0, \\ +\infty & \text{else.} \end{cases}$$

one has that $G^*(\rho,\omega) = \mathcal{A}(\rho,\omega)$ (see Remark 2.26). Since $F(\Psi)$ is linear, it can be written as $\langle l, \Psi\rangle_{\mathrm{C}^{1*},\mathrm{C}^1}$ for some suitable $l \in \mathrm{C}^{1*}$ characterized by

$$\langle l, \Psi\rangle_{\mathrm{C}^{1*},\mathrm{C}^1} = \int_X \Psi(1,\cdot)\mathrm{d}\nu - \int_X \Psi(0,\cdot)\mathrm{d}\mu$$

and F^* is then consequently the indicator function of that single element set $\{l\}$, i.e. $F^*(l') = 0$ if $l' = l$ and $+\infty$ otherwise. From the definition of A we find that

$$\langle A^*(\rho,\omega), \Psi\rangle_{\mathrm{C}^{1*},\mathrm{C}^1} = \langle (\rho,\omega), A\Psi\rangle_{\mathcal{M},\mathrm{C}} = \int_{[0,1]\times X} \partial_t\Psi \mathrm{d}\rho + \int_{[0,1]\times X} \nabla\Psi\cdot\mathrm{d}\omega$$

and thus $F^*(A^*(\rho,\omega)) = 0$ if and only if $(\rho,\omega) \in \mathcal{CE}(\mu,\nu)$ (cf. (2.11)). □

Proposition 2.38 (Primal-dual optimality conditions for the Benamou–Brenier formulation)**.** A pair $(\rho,\omega) \in \mathcal{M}([0,1]\times X)^{1+d}$ and $\Psi \in \mathrm{C}^1([0,1]\times X)$ is primal-dual optimal for (2.12) and (2.15) if and only if

$$\partial_t\rho + \nabla\cdot\omega = 0 \quad \text{with temporal boundary conditions } \rho(0,\cdot) = \mu,\ \rho(1,\cdot) = \nu$$

in the distributional sense of (2.11),

$$\rho \ge 0, \qquad \omega = \nabla\Psi\cdot\rho,$$

and

$$\partial_t\Psi + \tfrac{1}{2}|\nabla\Psi|^2 \le 0 \text{ with equality } \rho\text{-almost everywhere.}$$

Proof. The primal-dual optimality conditions for (2.12) ad (2.15) implied by Theorem 2.3 are (using the conventions of the proof of Proposition 2.37) $\Psi \in -\partial F^*(A^*(\rho,\omega))$ and $(\rho,\omega) \in \partial G(A\Psi)$. The former implies $(\rho,\omega) \in \mathcal{CE}(\mu,\nu)$ (argument similar as in Remark 2.9).

The latter implies first that $\partial G(A\Psi) \neq \emptyset$, i.e. $G(A\Psi) < \infty$, which means $G(A\Psi) = 0$, which implies that Ψ satisfies the inequality constraint $\partial_t\Psi + \frac{1}{2}|\nabla\Psi|^2 \leq 0$. Then using the equivalent condition (cf. Fenchel–Young inequality)

$$G^*(\rho,\omega) = \langle(\rho,\omega), A\Psi\rangle - G(A\Psi)$$

we conclude that $G^*(\rho,\omega) < \infty$, i.e. $\rho \geq 0$, $\omega = v \cdot \rho$ for some density velocity field v (as in the proof of Proposition 2.30) and therefore the above equality becomes

$$G^*(\rho,\omega) = \int_{[0,1]\times X} \Phi(1,v)\mathrm{d}\rho = \int_{[0,1]\times X} [1 \cdot \partial_t\Psi + v \cdot \nabla\Psi]\mathrm{d}\rho$$

which implies that $(\partial_t\Psi, \nabla\Psi) \in \partial\Phi(1,v)$ ρ-almost everywhere, which finally means that ρ-almost everywhere $\partial_t\Psi + \frac{1}{2}|\nabla\Psi|^2 \leq 0$ is actually an equality and $v = \nabla\Psi$. □

Remark 2.39 (Gradient structure of optimal velocity fields)**.** Intuitively it makes sense that the optimal velocity field $v = \nabla\Psi$ in Proposition 2.38 is a gradient: if it were not so, one could (at least formally) subtract the divergence-free (with respect to ρ) component of v (akin to the Helmholtz decomposition) to reduce the objective $\frac{1}{2}\int |v|^2\mathrm{d}\rho$ while preserving the continuity equation. This will result in a gradient vector field. In fact, one may interpret Brenier's theorem as a non-linear version of the Helmholtz decomposition [12, Section 1.3], and this intuition is the mechanism behind the celebrated Angenent–Haker–Tannenbaum scheme [4].

Remark 2.40 ((Non-)existence of dual maximizers)**.** At this point we should start to suspect that continuously differentiable dual maximizers of (2.15) are unlikely to exist in general: We have seen that the initial velocity field of the particles (which equals their Lagrangian velocity field), given by Brenier's theorem (Theorem 2.14, see also Example 2.20) will in general be defined only μ-almost everywhere (and only if μ has no atoms), and we could not establish sufficient regularity for v in (2.14) to integrate the flow. We will revisit the question of dual regularity in Section 2.4.3.

Remark 2.41 (Equivalence with Kantorovich dual problem)**.** Given the equivalence between the primal Kantorovich and Benamou–Brenier formulations (Section 2.3.2) one might wonder about the equivalence between the dual problems (2.4) (for $c = \frac{1}{2}d^2$) and (2.15). By comparing the respective objective functions it seems tempting to seek for an equivalence between (ϕ,ψ) in (2.4) and $(-\Psi(0,\cdot), \Psi(1,\cdot))$ in (2.15). This intuition is furthered by the observation that the initial velocity field is given by $-\nabla\phi$, consistent with the optimality condition that $v = \nabla\Psi$ (Proposition 2.38).

However, at this point it seems unclear how the respective constraints $\phi \oplus \psi \leq c$ and $\partial_t\Psi + \frac{1}{2}|\nabla\psi|^2 \leq 0$ can be in correspondence with each other. This connection can be made via the celebrated Hopf–Lax formula (see for instance [25]). We will naturally

stumble upon this formula by means of convex analysis via an auxiliary multi-marginal formulation in Section 2.4.2 and show its connection to the Benamou–Brenier formulation in Section 2.4.3.

Remark 2.42 (Connection to Burger's equation and geodesic equation for W)**.** In Example 2.19 we established that mass particles move with constant speed along straight lines in a Wasserstein geodesic. In the special case of Brenier's theorem (Example 2.20) this constant velocity field was shown to be $v_0 = -\nabla\phi$ (in a Lagrangian frame, where we identify particles by their initial position). At a formal level, the Eulerian velocity field should then be a velocity field that *propagates itself*, i.e. one should have

$$\partial_t v + (\nabla v) \cdot v = 0 \tag{2.16}$$

with initial condition $v(0,\cdot) = v_0$. This is known as (inviscid) Burger's equation, which is known for its tendency to develop shocks [25, Section 3.4.1].

From Brenier's theorem we know that $\frac{1}{2}|\cdot|^2 - \phi$ is convex, which implies that $\frac{1}{2}|\cdot|^2 - t\cdot\phi$ is strictly convex for $t \in [0,1)$ and therefore that $\mathrm{id} + t \cdot v$ is invertible for $t \in [0,1)$ (Remark 2.23). Hence, no shocks develop at intermediate times along a Wasserstein geodesic. However, shocks may develop at $t = 1$, or when one tries to extrapolate a Wasserstein geodesic beyond the interval $[0,1]$. This is consistent with Proposition 2.21.

From Proposition 2.38 we learn that the Eulerian velocity field $v_t = \nabla\Psi(t,\cdot)$ is formally given as a gradient at all intermediate times where Ψ satisfies

$$\partial_t \Psi + \tfrac{1}{2}|\nabla\Psi|^2 = 0 \quad \rho\text{-almost everywhere.} \tag{2.17}$$

Formally applying a spatial gradient to (2.17) one indeed obtains (2.16). Thus, (2.17) can be seen as reformulation of (2.16) to explicitly incorporate the gradient property of v. If for a given μ and some $\Psi(0,\cdot) \in \mathrm{C}^1(X)$ there would be a classical solution of (2.17) on some non-empty time interval $[0,t]$, the corresponding velocity field would induce a path of measures via the continuity equation, which would then indeed be a Wasserstein geodesic. Therefore, (2.17) has been dubbed the *geodesic equation* for W [39].

2.4 Multi-marginal formulation

2.4.1 Primal problem

Multi-marginal optimal transport is a versatile tool with many applications, such as modeling in economics, fluid dynamics, and density function theory, see [48, Section 1.7.4] for some references. In this section we merely use it in the context of dynamic optimal transport, similar to [11]. Our main motivation is to serve as an intermediate step between the Kantorovich and the Benamou–Brenier formulations of optimal transport, to provide a better understanding of the dual formulation of the latter, and to prepare similar constructions for the entropic setting in Section 3. From a modeling perspective this reformulation is also interesting since it allows to couple the movement of mass to other observations and constraints at intermediate times (Remark 2.46).

Definition 2.43 (Multi-marginal cost function)**.** Let X be a compact, convex subset of $\mathbb{R}^d$, and let $c(x,y) = \frac{1}{2}|x-y|^2$. For a natural number $N \geq 2$ and time points $t_0 = 0 < t_1 < t_2 < \ldots < t_N = 1$, the multi-marginal cost function is given by

$$c_{\mathrm{MM}} : X^{N+1} \mapsto \mathbb{R}, \qquad (x_0, \ldots, x_N) \mapsto \sum_{i=1}^{N} \tfrac{1}{2(t_i - t_{i-1})} |x_i - x_{i-1}|^2. \tag{2.18}$$

We will occasionally denote tuples $(x_0, \ldots, x_N) \in X^{N+1}$ more compactly as $\vec{x}$.

We now show that the Kantorovich formulation of W, (2.6), can be rewritten as a multi-marginal transport problem with cost c_{MM}, where only the first and last marginal are prescribed.

Proposition 2.44 (Multi-marginal formulation of W)**.** Problem (2.6) is equivalent to the problem

$$\inf \left\{ \int_{X^{N+1}} c_{\mathrm{MM}}(\vec{x}) \, \mathrm{d}\gamma_{\mathrm{MM}}(\vec{x}) \middle| \gamma_{\mathrm{MM}} \in \mathcal{P}(X^{N+1}),\, \mathrm{P}_0 \gamma_{\mathrm{MM}} = \mu, \mathrm{P}_N \gamma_{\mathrm{MM}} = \nu \right\} \tag{2.19}$$

in the following sense: First, their minimal values are equal. Moreover, if γ_{MM} is a minimizer of (2.19), then $P_{0,N}\gamma_{\mathrm{MM}}$ is a minimizer of (2.6). Here $P_{0,N}$ denotes the push-forward by the map $X^{N+1} \to X^2$, $(x_0, \ldots, x_N) \mapsto (x_0, x_N)$, i.e. $P_{0,N}\gamma_{\mathrm{MM}}$ extracts the joint $(0,N)$-th marginal of γ_{MM}. Conversely, if γ is a minimizer of (2.6) then $F_{N\#}\gamma$ is a minimizer of (2.19) where

$$F_N : X^2 \to X^{N+1}, \qquad (x_0, x_N) \mapsto (x_0, x_1, \ldots, x_{N-1}, x_N)$$

with $x_i := (1-t_i) \cdot x_0 + t_i \cdot x_N$ for $i \in \{1, \ldots, N-1\}$.

The proof hinges on the following lemma, which can be proved by a simple explicit calculation.

Lemma 2.45. For $(x_0, x_1) \in X^2$ one has

$$c(x_0, x_1) = \min \{ c_{\mathrm{MM}}(x_0, \ldots, x_N) | x_1, \ldots, x_{N-1} \in X \}$$

and the unique minimizer on the r.h.s. is given by $x_i = (1-t_i) \cdot x_0 + t_i \cdot x_N$ for $i \in \{1, \ldots, N-1\}$.

Proof of Proposition 2.44. Let γ_{MM} be admissible in (2.19). Then $\gamma := P_{0,N}\gamma_{\mathrm{MM}}$ clearly lies in $\Gamma(\mu,\nu)$ and is therefore admissible in (2.6) and by Lemma 2.45 one has

$$\int_{X^2} c \, \mathrm{d}\gamma = \int_{X^N+1} c \circ p_{0,N} \, \mathrm{d}\gamma_{\mathrm{MM}} \leq \int_{X^{N+1}} c_{\mathrm{MM}} \, \mathrm{d}\gamma_{\mathrm{MM}}, \tag{2.20}$$

where $p_{0,N} : (x_0, \ldots, x_N) \mapsto (x_0, x_N)$. Thus (2.6) $\leq$ (2.19). Conversely, let $\gamma \in \Gamma(\mu,\nu)$ be an admissible candidate for (2.6) and set $\gamma_{\mathrm{MM}} := F_{N\#}\gamma$, which is then again admissible for (2.19) with cost

$$\int_{X^{N+1}} c_{\mathrm{MM}} \, \mathrm{d}\gamma_{\mathrm{MM}} = \int_{X^2} c_{\mathrm{MM}} \circ F_M \, \mathrm{d}\gamma = \int_{X^2} c \, \mathrm{d}\gamma \tag{2.21}$$

where the second equality is again a consequence of Lemma 2.45. Therefore (2.19) $\leq$ (2.6). Consequently, both infima are the same, and minimizers for one can be constructed from the other as in the proof. In particular existence of minimizers for (2.19) follows from existence in (2.6) (which is in turn a consequence of Proposition 2.4). □

Remark 2.46 (Interpretation of (2.6) and (2.19))**.** In the Kantorovich formulation of W, (2.6), the plan γ encodes the distribution of initial and final locations of mass particles. Via Theorem 2.18 it is implied that in a dynamic interpretation these particles travel on a constant speed straight line from their initial to the final location during transport. This perspective is further developed through the equivalence with the Benamou–Brenier formulation (Section 2.3.2).

In contrast, a plan γ_{MM} in (2.19) can be seen as a more detailed description of the itinerary of the mass particles, where a sequence of locations $(x_0, \dots, x_N)$ at times $(t_0 = 0, \dots, t_N = 1)$ is specified, including intermediate times $(t_i)_{i=1}^{N-1}$. Of course, in hindsight, this description is redundant, as shown by Proposition 2.44. However, first, it provides a more gradual path from the static Kantorovich formulation to the dynamic Benamou–Brenier formula (see the remainder of this section, in particular Section 2.4.3), and second, it allows to model more general dynamic processes, where the particles interact with the environment at intermediate times, e.g. through volume constraints or measurement information, see for instance [11, 30, 34]. Note that all the references actually use entropic optimal transport. However, in [11] it appears merely as a numerical tool, whereas in the other two references entropy is actually part of the data model. We will study this in more detail in Section 3.2.

The dynamic description implied by (2.19) is Lagrangian, since the measure γ_{MM} keeps track of all individual particles and their trajectories. The following equivalent reformulation (stated without proof) has a more Eulerian flavour, as it keeps track of the intermediate mass distributions and their next steps. It is also more tractable numerically, see Remark 2.48.

Proposition 2.47. Problems (2.6) and (2.19) are equivalent to the following problem

$$\inf \left\{ \sum_{i=0}^{N-1} \int_{X^2} \frac{|x-y|^2}{2(t_{i+1}-t_i)} \,\mathrm{d}\gamma_i(x,y) \,\middle|\, (\rho_0, \dots, \rho_N) \in \mathcal{P}(X),\ \rho_0 = \mu,\ \rho_N = \nu, \right. \\ \left. \gamma_i \in \Gamma(\rho_i, \rho_{i+1}) \text{ for } i \in \{0, \dots, N-1\} \right\} \quad (2.22)$$

in the sense that the minimal values are identical and minimizers can be constructed from each for the other in a similar spirit as in Theorem 2.18 and Proposition 2.44.

For minimizers of (2.22), the transport plans γ_i are optimal for the Wasserstein optimal transport problem $\mathrm{W}(\rho_i, \rho_{i+1})$, (2.6), between their marginals.

Remark 2.48 (Markov property)**.** The equivalence between (2.19) and (2.22) holds since the cost c_{MM} is in fact only a sum of functions involving adjacent time steps. Thus, not the full joint distribution γ_{MM} is relevant, but merely the collection of pairwise distributions

of all successive time step pairs, which is captured in (2.22). We may therefore assume, for instance, that γ_{MM} is Markov in the following sense: Let γ_{MM} be the joint law of a tuple $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ of X-valued random variables. Then $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$ are independent when conditioned on some $\boldsymbol{x}_j$ for a triple $i < j < k$.

At first glance, in (2.19) there might also exist minimizers that are not Markov. However, Corollary 2.22 implies that this is not the case, since all particles at some intermediate position $z \in X$ at some intermediate time $t_j \in (0,1)$ must have the same initial and final positions. Therefore the positions $\boldsymbol{x}_i$, $\boldsymbol{x}_k$ at earlier or later times become concentrated on single points when conditioned on $\boldsymbol{x}_j = z$ at time t_j and are thus independent.

In the language of [30] this Markov property means that c_{MM} corresponds to a chain graph structure and it is one of the avenues to make high-dimensional multi-marginal transport problems numerically tractable [11, 10, 30, 1, 7]. This Markov property becomes also relevant in the presence of entropic regularization (Section 3.1) where it is again essential for numerical tractability (see some of the references above) and also crucial for the dynamic interpretation and modeling (Remark 3.23).

Remark 2.49 (Limit $N \to \infty$)**.** Of course, formulations (2.19) and (2.22) are not fully dynamic in the sense that they are still time-discrete and we implicitly rely on the fact that the γ_i are pairwise optimal plans and then interpolate the gaps with Theorem 2.18. It appears natural to consider the limit $N \to \infty$. In (2.19) in this limit the tuples $(x_0, \dots, x_N)$ will have to be replaced by a suitable class of paths $x : [0,1] \to X$ where $x(t)$ gives the position of a mass particle at arbitrary times $t \in [0,1]$, and γ_{MM} will be replaced by a measure over such paths.

The proper regularity class of paths is determined by considering where the corresponding limit of the cost function c_{MM} is well-behaved, i.e. we need that for every admissible path x the supremum of

$$\sum_{i=1}^{N} (t_i - t_{i-1}) \cdot \frac{|x(t_i) - x(t_{i+1})|^2}{(t_i - t_{i-1})^2}$$

over N and intermediate time points $(t_i)_i$ is bounded. This can be seen as finite-difference approximation of the integral $\int_0^1 |\partial_t x(t)|^2 \mathrm{d}t$ and therefore naturally leads to the space $\mathrm{H}^1([0,1], X)$ when X is a subset of $\mathbb{R}^d$, or to the class of *absolutely continuous curves*, denoted by $\mathrm{AC}([0,1], X)$, if X is a more general metric space. So the natural limit for γ_{MM} will be a measure on paths over X, concentrated on $\mathrm{AC}([0,1], X)$ and the mass distribution at time t can be extracted by the push-forward under the evaluation map

$$\mathrm{ev}_t : \mathrm{AC}([0,1], X) \to X, \qquad x \mapsto x(t).$$

In (2.22) the natural limit for all intermediate $(\rho_i)_i$ will be a curve $[0,1] \to \mathcal{P}(X)$ and the intermediate γ_i will play the role of the momentum measure ω, as in the Benamou–Brenier formula (2.12).

Recall that in Remark 2.32 we mentioned the notion of absolutely continuous paths of measures in $(\mathcal{P}(X), \mathrm{W})$, which we now denote as $\mathrm{AC}([0,1], \mathcal{P}(X))$. It was shown in

[38] that the two notions are essentially equivalent: For a measure γ_{MM} on absolutely continuous paths the curve $t \mapsto \rho_t := \mathrm{ev}_{t\#}\gamma_{\mathrm{MM}}$ of time marginals lies in $\mathrm{AC}([0,1], \mathcal{P}(X))$. Conversely, if $t \mapsto \rho_t$ lies in $\mathrm{AC}([0,1], \mathcal{P}(X))$, then there is a γ_{MM} concentrated on $\mathrm{AC}([0,1], X)$ such that $\rho_t = \mathrm{ev}_{t\#}\gamma_{\mathrm{MM}}$.

Measures on paths as a Lagrangian description of dynamic optimal transport are a fundamental and popular tool, see for instance [15, 35].

We do not consider this limit here to avoid the related technicalities. By choosing N and the time positions of the intermediate marginals flexibly, and via the equivalences of (2.6), (2.19) and (2.22) we can obtain all results of interest in this chapter.

2.4.2 Dual problem

Now we consider a dual perspective on the multi-marginal problems considered in the previous section. This will eventually help us to better understand the dual formulation of the Benamou–Brenier formulation.

Proposition 2.50 (Dual multi-marginal problem)**.** A dual problem for (2.19) is given by

$$\sup\left\{\int_X \phi \,\mathrm{d}\mu + \int_X \psi \,\mathrm{d}\nu \,\middle|\, \phi, \psi \in \mathrm{C}(X), \right. \\ \left. \phi(x_0) + \psi(x_N) \le c_{\mathrm{MM}}(x_0, \dots, x_N) \,\forall\, (x_0, \dots, x_N) \in X^{N+1} \right\} \tag{2.23}$$

and (2.23) is equivalent to (2.4) (for cost $c = \frac{1}{2}d^2$) in the sense that the optimal values and the set of maximizers are identical.

Proof. The form of (2.23) can be obtained via Fenchel–Rockafellar duality from (2.19) in almost the exact same way as (2.4) was obtained from (2.1).

The form of (2.23) and equivalence with (2.4) can be obtained from Lemma 2.45 by observing that the constraint

$$\phi(x_0) + \psi(x_N) \le c_{\mathrm{MM}}(x_0, \dots, x_N) \,\forall\, (x_0, \dots, x_N) \in X^{N+1}$$

is equivalent to

$$\phi(x_0) + \psi(x_N) \le \inf_{(x_1,\dots,x_{N-1}) \in X^{N-1}} c_{\mathrm{MM}}(x_0, \dots, x_N) \,\forall\, (x_0, x_N) \in X^2. \qquad \square$$

To gain some intuition for the next steps, consider the above problem for $N = 2$. Then the constraint can be written as

$$\psi(x_2) \le \frac{|x_0 - x_1|^2}{2t_1} + \frac{|x_1 - x_2|^2}{2(1-t_1)} - \phi(x_0).$$

We can take the infimum over x_0 to obtain

$$\psi(x_2) \le \frac{|x_1 - x_2|^2}{2(1-t_1)} + \psi_1(x_1) \quad \text{with} \quad \psi_1(x_1) := \inf_{x_0 \in X} \frac{|x_0 - x_1|^2}{2t_1} - \phi(x_0).$$

We see that by introducing suitable auxiliary dual potentials for intermediate times, we can localize the constraints in time. This motivates the following definition.

Proposition 2.51 (Dynamic duals)**.** Let (ϕ, ψ) be a pair of dual maximizers of (2.4) that satisfy $\phi = \psi^c$, $\psi = \phi^c$. We then introduce dynamic dual functions $\Phi, \Psi : [0,1] \times X \to \mathbb{R}$ as follows:

$$\Phi(t,x) := \inf_{y \in X} \frac{|x-y|^2}{2 \cdot (1-t)} - \psi(y) \qquad \text{for } t \in [0,1), \qquad \Phi(1,x) := -\psi(x), \tag{2.24a}$$

$$\Psi(t,x) := \inf_{y \in X} \frac{|x-y|^2}{2 \cdot t} - \phi(y) \qquad \text{for } t \in (0,1], \qquad \Psi(0,x) := -\phi(x). \tag{2.24b}$$

Then Φ and Ψ are Lipschitz continuous with respect to both arguments and they satisfy the recursive definitions

$$\Phi(s,x) = \inf_{y \in X} \frac{|x-y|^2}{2 \cdot (t-s)} + \Phi(t,y), \qquad \Psi(t,x) = \inf_{y \in X} \frac{|x-y|^2}{2 \cdot (t-s)} + \Psi(s,y) \tag{2.25}$$

for $s, t \in [0,1]$ with $s < t$.

Expressions (2.24) and (2.25) are special cases of the *Hopf–Lax formula* (see for instance [25]) which provide suitable generalized solutions to evolution equations as they appear in the constraint of the dual Benamou–Brenier formula, (2.15). In Section 2.4.3 we will make this relation more explicit.

Proof. The recursive definition (2.25) can be verified by using the original definitions for (2.24) for Φ and Ψ on the right-hand sides and then invoking Lemma 2.45.

The proof of Lipschitz continuity with respect to x is closely related to the arguments how dual potentials inherit regularity from the cost function via the c-transform, as used in the proof of Proposition 2.6. First, note that since X is bounded, c is Lipschitz continuous on $X \times X$ and therefore $\phi = \psi^c$ and $\psi = \phi^c$ are both Lipschitz continuous. In the following we denote a suitable Lipschitz constant by L. One can then extend ϕ and ψ to $\mathbb{R}^d$ via $\psi(x) := \sup_{y \in X} -L|x-y| + \psi(y)$ for all $x \in \mathbb{R}^d \setminus X$ (and the same formula for ϕ). These global extensions will still have Lipschitz constant L and satisfy $\psi(x) \leq \psi(p_X(x))$ where p_X denotes the projection onto the compact and convex set X. Therefore, when $\tilde{c} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $(x,y) \mapsto \tilde{c}(x,y)$ is an increasing function of $|x-y|$, one finds for $x \in X$ that

$$\inf_{y \in X} \tilde{c}(x,y) - \psi(y) = \inf_{y \in \mathbb{R}^d} \tilde{c}(x,y) - \psi(y).$$

As a consequence, Φ and Ψ inherit the Lipschitz constant L from ϕ and ψ with respect to the spatial argument. Indeed one has for $t \in [0,1)$ and $x, x' \in X$, $y \in \mathbb{R}^d$,

$$\Phi(t,x) \leq \frac{|x-y|^2}{2 \cdot (1-t)} - \psi(y) \leq \frac{|x'-(y+x'-x)|^2}{2 \cdot (1-t)} - \psi(y+x'-x) + L\,|x'-x|.$$

Taking now the infimum over $y \in \mathbb{R}^d$ on the right-hand side yields $\Phi(t,x) \leq \Phi(t,x') + L\,|x - x'|$ and thus by the symmetric argument on obtains the Lipschitz bound with respect to x. Lipschitz continuity of $\Phi(1,\cdot) = -\psi$ follows directly from the definition. Arguments for Ψ are identical.

For the Lipschitz bound with respect to t observe first that (2.25) implies $\Phi(s,x) \leq \Phi(t,x)$ for $0 \leq s < t \leq 1$. Using the spatial Lipschitz regularity one then finds

$$\begin{aligned}\Phi(s,x) &= \inf_{y\in X} \frac{|x-y|^2}{2\cdot(t-s)} + \Phi(t,y) \geq \inf_{y\in X} \frac{|x-y|^2}{2\cdot(t-s)} + \Phi(t,x) - L|x-y| \\ &\geq \Phi(t,x) - \frac{L^2(t-s)}{2}.\end{aligned}$$

□

The following proposition provides additional structure on the dynamic duals Φ and Ψ.

Proposition 2.52. Consider the setting of Proposition 2.51 and let $t \in (0,1)$. For any $z \in X$ one has

$$\Phi(t,z) + \Psi(t,z) \geq 0$$

with equality if and only if z can be written as $z = (1-t)\cdot x + t\cdot y$ for some (x,y) in the contact set of (ϕ,ψ) (Definition 2.7). In this case the pair (x,y) is unique (in the contact set) and one has

$$\Phi(t,z) = \frac{|z-x|^2}{2\cdot(1-t)} - \psi(x), \qquad \Psi(t,z) = \frac{|z-y|^2}{2\cdot t} - \phi(y),$$

i.e. x and y are minimizers in (2.24) for $\Phi(t,z)$ and $\Psi(t,z)$ and in fact are the unique minimizers.

Proof. Using the definitions (2.24) we find

$$\Phi(t,z) + \Psi(t,z) = \inf_{(x,y)\in X^2} \frac{|x-z|^2}{2\cdot(1-t)} + \frac{|z-y|^2}{2\cdot t} - \phi(x) - \psi(y) \tag{2.26}$$

$$\geq \inf_{(x,y)\in X^2} \frac{|x-y|^2}{2} - \phi(x) - \psi(y) \geq 0 \tag{2.27}$$

where the first inequality is a consequence of Lemma 2.45.

Now let (x,y) be in the contact set and set $z = (1-t)\cdot x + t\cdot y$. Then we observe that the candidate (x,y) yields the value zero in both infima above and thus indeed attains both infima. Conversely, if $\Phi(t,z) + \Psi(t,z) = 0$, then both inequalities must be equalities and the infimal values must both equal zero. Let (x,y) be a minimizer of the first infimum (existence implied by compactness and continuity). For the infimum to be zero, one must have $z = (1-t)\cdot x + t\cdot y$ and (x,y) must be in the contact set. Uniqueness of this (x,y) follows from Proposition 2.21. □

In (2.26) we see that the dynamic duals take on a particularly simple form on suitable straight lines that correspond to the flow of the optimal transport velocity field (cf. Remark 2.42). This is related to the *method of characteristics*, which plays an important role for the Hopf–Lax formula.

We can now rewrite (2.23) with auxiliary potentials and constraints that are local in time. The resulting problem starts to resemble the dual Benamou–Brenier formula (2.15).

Proposition 2.53. Problem (2.23) has the same optimal value as

$$\sup\left\{\int_X \psi_N \,\mathrm{d}\nu - \int_X \psi_0 \,\mathrm{d}\mu \,\middle|\, (\psi_0,\dots,\psi_N)\in \mathrm{C}(X)^{N+1},\right.$$
$$\left.\psi_{i+1}(x)-\psi_i(y)\le \frac{|x-y|^2}{2(t_{i+1}-t_i)}\,\forall\, x,y\in X,\, i\in\{0,\dots,N-1\}\right\}. \quad (2.28)$$

Given maximizers (ϕ,ψ) for (2.23) that satisfy $\phi=\psi^c$, $\psi=\phi^c$ (such maximizers exist, as shown in Proposition 2.6), a maximizer for (2.28) is given constructing first the function Ψ from (ϕ,ψ) via (2.24) and then setting $\psi_i=\Psi(t_i,\cdot)$ for $i\in\{0,\dots,N\}$. For a maximizer $(\psi_0,\dots,\psi_N)$ in (2.28), a maximizer for (2.23) is given by setting $(\phi,\psi):=(-\psi_0,\psi_N)$.

Proof. Consider the claimed construction of a maximizer for (2.28) form one of (2.23). By (2.25) the tuple $(\psi_0,\dots,\psi_N)$ indeed satisfies the inequality constraints in (2.28). And as by construction $\psi_0=\Psi(0,\cdot)=-\phi$, they also yield the same objective in (2.28) as (ϕ,ψ) yield in (2.23). So (2.28) $\ge$ (2.23).

Now consider the converse construction. Summing the constraints in (2.28) over i we obtain that

$$\psi_N(x_N)-\psi_0(x_0)\le\sum_{i=0}^{N-1}\frac{|x_i-x_{i+1}|^2}{2(t_{i+1}-t_i)}$$

for all $(x_0,\dots,x_N)\in X^{N+1}$. Taking now the infimum over $(x_1,\dots,x_{N-1})$ and using Lemma 2.45 we obtain that $(\phi,\psi):=(-\psi_0,\psi_N)$ is admissible in (2.23) and again it has the same objective value. This yields the reverse inequality and therefore equality with the confirmation the the two above constructions convert maximizers of one problem into maximizers of the other. □

Corollary 2.54. Given maximizers $(\psi_0,\dots,\psi_N)$ of (2.28), the pair $(-\psi_i,\psi_{i+1})$ for $i\in\{0,\dots,N-1\}$, is a maximizer for the dual problem of one time-step of (2.22), given by

$$\inf\left\{\int_{X^2}\frac{|x-y|^2}{2(t_{i+1}-t_i)}\,\mathrm{d}\gamma_i(x,y)\,\middle|\,\gamma_i\in\Gamma(\rho_i,\rho_{i+1})\right\}$$

for fixed (ρ_i,ρ_{i+1}) which are taken from a primal minimizer in (2.22), i.e. they maximize

$$\sup\left\{\int_X \psi_{i+1}\,\mathrm{d}\rho_{i+1}-\int_X\psi_i\,\mathrm{d}\rho_i\,\middle|\,\psi_i,\psi_{i+1}\in\mathrm{C}(X),\right.$$
$$\left.\psi_{i+1}(y)-\psi_i(x)\le\frac{|x-y|^2}{2(t_{i+1}-t_i)}\,\forall x,y\in X\right\}. \quad (2.29)$$

Proof. For simplicity we only sketch the proof for minimizers in (2.22) and maximizers in (2.28) that were constructed from minimizers in (2.6) and maximizers in (2.23), but the proof can easily be extended to general optimizers. First, the pair $(-\psi_i, \psi_{i+1})$ satisfies the constraint in (2.29), as it is also part of (2.28). By virtue of Proposition 2.52, constructing ψ_i and ψ_{i+1} via Ψ yields that the contact set of $(-\psi_i, \psi_{i+1})$ with respect to the cost function $\frac{|x-y|^2}{2(t_{i+1}-t_i)}$ is given by the image of the contact set of (ϕ, ψ) under the map $\hat{Z}(t_i, t_{i+1}, \cdot, \cdot)$ (cf. Theorem 2.18). On the primal side, the optimal plan γ_i constructed from γ via the push-forward with the map $\hat{Z}(t_i, t_{i+1}, x, y)$ is concentrated on this contact set, hence their respective optimality is provided by Proposition 2.8. □

2.4.3 Connection to dual Benamou–Brenier formula

Assume now for simplicity that optimal $(\rho_i)_i$ in Proposition 2.47 are absolutely continuous with respect to the Lebesgue measure, so that we may apply Brenier's theorem. Then by Corollary 2.54 the Lagrangian velocity field for each ρ_i in the time interval $[t_i, t_{i+1}]$ is given by $\nabla\psi_i = \nabla\Psi(t_i, \cdot)$ (here the factor $1/(t_{i+1} - t_i)$ in the cost function cancels with the fact that the time interval is shorter), which is consistent with Proposition 2.38. This suggests that the connection of the static Kantorovich duals of (2.4) and the dynamic potential of the Benamou–Brenier formulation (2.15) may be obtained via the time-interpolation of Proposition 2.51.

Clearly, yet another equivalent dual formulation of (2.23) and (2.28) is given by

$$\sup\left\{\int_X \Psi(1,\cdot)\,\mathrm{d}\nu - \int_X \Psi(0,\cdot)\,\mathrm{d}\mu \,\middle|\, \Psi \in \mathrm{C}([0,1]\times X),\right.$$
$$\left.\Psi(s,x) - \Psi(t,y) \le \frac{|x-y|^2}{2(t-s)}\,\forall\, x, y \in X,\ s,t \in [0,s] \text{ with } s < t\right\}. \quad (2.30)$$

We now show that this is a natural relaxation of the dual Benamou–Brenier formulation. First, candidates admissible in (2.15) are admissible in (2.30).

Lemma 2.55. If $\Psi \in \mathrm{C}^1([0,1]\times X)$ is admissible for the dual Benamou–Brenier formula (2.15), in particular when $\partial_t\psi + \frac{1}{2}|\nabla\psi|^2 \le 0$, then $\Psi(t,x) \le \frac{1}{2\cdot(t-s)}|x-y|^2 + \Psi(s,y)$ for all $s < t$, x, y.

Proof. Let $z : [s,t] \to X$, $z(r) = \frac{r-s}{t-s}x + \frac{t-r}{t-s}y$, i.e. the constant speed straight line interpolation from y to x, parametrized in the time interval $[s,t]$ with $\partial_r z(r) = (x -$

$y)/(t-s)$. Then

$$\begin{aligned}\psi(t,x) &= \int_s^t \frac{\mathrm{d}}{\mathrm{d}r}\psi(r,z(r))\,\mathrm{d}r + \psi(s,y)\\ &= \int_s^t \left[\partial_t\psi(r,z(r)) + \nabla\psi(r,z(r))\cdot\frac{x-y}{t-s}\right]\mathrm{d}r + \psi(s,y)\\ &\leq \int_s^t \left[-\tfrac{1}{2}|\nabla\psi(r,z(r))|^2 + \nabla\psi(r,z(r))\cdot\frac{x-y}{t-s}\right]\mathrm{d}r + \psi(s,y)\\ &\leq \int_s^t \frac{|x-y|^2}{2(t-s)^2}\mathrm{d}r + \psi(s,y) = \frac{|x-y|^2}{2(t-s)} + \psi(s,y)\end{aligned}$$

where we used the feasibility condition in the third line and the inequality $-\frac{1}{2}|v|^2+v\cdot w \leq \frac{1}{2}|w|^2$ for arbitrary vectors $w, v \in \mathbb{R}^d$ in the last line. □

Conversely, we conclude by showing that if an admissible candidate of (2.30) is differentiable at a given point, it will satisfy the constraint of (2.15) in that point. Note that we may restrict ourselves to consider maximizers in (2.30) that are generated via Proposition 2.51, which are Lipschitz continuous. Hence, they are differentiable almost everywhere on $[0,1]\times X$.

Lemma 2.56. If $\psi : [0,1]\times X \to \mathbb{R}$ is differentiable in $(s,x)\in(0,1)\times X$ and satisfies $\psi(t,y)-\psi(s,x) \leq \frac{|x-y|^2}{2(t-s)}$ for all $t\in(s,1]$ and $y\in X$, then $\partial_t\psi(s,x)+\frac{1}{2}|\nabla\psi(s,x)|^2 \leq 0$.

Proof. We set $t = s+\tau$, $\tau>0$ and $y = x+\delta$ and approximate the finite difference in the inequality $\psi(t,y)-\psi(s,x)\leq\frac{|x-y|^2}{2(t-s)}$ by a linear expansion to obtain

$$\partial_t\psi(s,x)\cdot\tau + \nabla\psi(s,x)\cdot\delta \leq \tfrac{|\delta|^2}{2\tau} + o(\tau) + o(|\delta|).$$

We now choose $\delta = \tau\cdot\nabla\psi(s,x)$ and divide by τ to obtain

$$\partial_t\psi(s,x) + |\nabla\psi(s,x)|^2 \leq \tfrac{|\nabla\psi(s,x)|^2}{2} + o(1)$$

where $o(1)$ is to be understood with respect to the limit $\tau\to 0$ and the claim follows by passing to this limit. □

3 Entropic optimal transport

3.1 Entropic Kantorovich problem

In these notes we will work with the following notion of entropy.

Definition 3.1 (Entropy)**.** For measures $\mu,\lambda\in\mathcal{M}(Z)$ on a compact metric space Z we set the relative entropy of μ with respect to λ as

$$H(\gamma|\lambda) := \begin{cases}\int_X \varphi\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda & \text{if } \gamma,\lambda\geq 0,\ \gamma\ll\lambda,\\ +\infty & \text{else,}\end{cases} \quad \text{with } \varphi:\mathbb{R}_+\to\mathbb{R}_+,\quad s\mapsto s\log(s)-s. \tag{3.1}$$

There are slight variations in the literature about the choice of the integrand φ in (3.1). The common choice $\varphi(s) = s\log(s)$ corresponds to the negative Shannon entropy, another common choice is $\varphi(s) = s\log(s) - s + 1$, which yields the Kullback–Leibler divergence. From the latter we drop the $+1$ to remove a number of constant terms in various dual problems (but which can easily be added back), but keep the -1 as it will yield more convenient primal-dual relations. As long as λ has finite mass (as assumed above), $H(\cdot|\lambda)$ is bounded form below.

Definition 3.2 (Entropic Kantorovich problem)**.** Let (X, d) be a compact metric space, $c \in \mathrm{C}(X \times X)$, and $\mu, \nu \in \mathcal{P}(X)$. Let $\lambda_1, \lambda_2 \in \mathcal{M}_+(X)$ such that $H(\mu|\lambda_1) < \infty$, $H(\nu|\lambda_2) < \infty$ and set $\lambda := \lambda_1 \otimes \lambda_2$. Let $\varepsilon > 0$. Then the entropic optimal transport problem is given by

$$C_\varepsilon(\mu, \nu) := \inf \left\{ \int_{X\times X} c(x, y)\, \mathrm{d}\gamma(x, y) + \varepsilon\, H(\gamma|\lambda) \,\middle|\, \gamma \in \Gamma(\mu, \nu) \right\}. \tag{3.2}$$

One striking advantage of the regularized problem (3.2) compared to (2.1) is that the former can be solved by the celebrated Sinkhorn algorithm. We refer to [43] for an introduction, some historical context, and an overview on variants and modifications. Other advantages and consequences of regularization are briefly discussed in Remark 3.11.

Remark 3.3 (Choice of reference measure)**.** There are various common choices for the reference measure λ in (3.2). The two most common choices for λ are $(\lambda_1, \lambda_2) = (\mu, \nu)$ and $\lambda_1 = \lambda_2 = \mathcal{L} \llcorner X$ in which case $X \subset \mathbb{R}^d$ and the marginals must satisfy $\mu, \nu \ll \mathcal{L}$. For finite X, the counting measure is sometimes used for simplicity.

Proposition 3.4. (3.2) has a unique minimizer.

Proof. By assumption $\gamma = \mu \otimes \nu$ satisfies $H(\gamma|\lambda) < \infty$ and $\int c\mathrm{d}\gamma < \infty$ and the objective is bounded from below, thus the infimum is finite. The objective is weak* lower-semicontinuous and the admissible set is weak* compact, thus a minimizer exists. It is unique by strict convexity of $H(\cdot|\lambda)$. □

Existence of minimizers in more general settings can be found, for instance, in [24, 47], see also [36]. To proceed, we need a better characterization of the minimizers. This can be obtained via duality. We will split this into several steps. First establishing the general form of the dual problem, then existence of dual maximizers (under the assumption $\lambda = \mu \otimes \nu$) and the primal-dual relation, and finally a relaxation of the dual problem.

Proposition 3.5 (Duality for entropic transport, part I)**.**

$$C_\varepsilon(\mu, \nu) = \sup \left\{ \int_X \phi\, \mathrm{d}\mu + \int_X \psi\, \mathrm{d}\nu - \varepsilon \int_{X\times X} \exp\left([\phi \oplus \psi - c]/\varepsilon\right) \mathrm{d}\lambda \,\middle|\, \phi, \psi \in \mathrm{C}(X) \right\} \tag{3.3}$$

Proof. The proof is analogous to that of Proposition 2.4. We choose U, V, F and A in the same way but now set

$$G : \mathrm{C}(X \times X) \to \mathbb{R}, \qquad \eta \mapsto \varepsilon \int_{X\times X} \exp\left([\eta - c]/\varepsilon\right) \mathrm{d}\lambda$$

Now both F and G are finite and continuous everywhere, thus Theorem 2.3 can be applied. Observe that for φ in (3.1) we have $\varphi^*(s) = \exp(s)$ and therefore we obtain that $G^*(\gamma) = H(\gamma|\lambda)$ [44]. □

Proposition 3.6 (Duality for entropic transport, part II)**.** If $(\lambda_1, \lambda_2) = (\mu, \nu)$ then maximizers in (3.3) exist. The unique minimizer of (3.2) takes the form $\gamma = \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \mu \otimes \nu$ where (ϕ, ψ) are (arbitrary) maximizers of (3.3).

The proof is somewhat analogous to that of Proposition 2.6 and involves an entropic version of the c-transform (Definition 2.5), obtained by a formal pointwise maximization of (3.3) with respect to ϕ for fixed ψ (and vice versa). This yields the following definition.

Definition 3.7 (Entropic c-transform)**.** For a cost function $c \in \mathrm{C}(X \times X)$, a positive parameter $\varepsilon > 0$ and a potential $\psi \in \mathrm{C}(X)$ the entropic c-transform of ψ is the function $\psi^{c,\varepsilon}$ given by

$$\psi^{c,\varepsilon} : x \mapsto -\varepsilon \log \left(\int_X \exp\left(\tfrac{\psi - c(x,\cdot)}{\varepsilon} \right) \mathrm{d}\nu \right)$$

and again in analogy the entropic $\bar{c}$-transform as

$$\psi^{\bar{c},\varepsilon} : y \mapsto -\varepsilon \log \left(\int_X \exp\left(\tfrac{\psi - c(\cdot,y)}{\varepsilon} \right) \mathrm{d}\mu \right)$$

Proof of Proposition 3.6. Similar as in the proof of Proposition 2.6, for given $\psi \in \mathrm{C}(X)$, the function $\psi^{c,\varepsilon}$ inherits the modulus of continuity of c (in particular it is in $\mathrm{C}(X)$ and therefore admissible) and the choice $\phi = \psi^{c,\varepsilon}$ maximizes (3.3) for fixed ψ. We may therefore restrict ourselves to maximizing sequences of the form $(\phi_n = \psi_n^{c,\varepsilon}, (\psi_n^{c,\varepsilon})^{\bar{c},\varepsilon})$, which are equicontinuous and by the same argument about constant shifts also equibounded. Once more we can extract a cluster point via the Arzelà–Ascoli theorem, which must be maximizer by continuity of the objective (3.3) (which is now unconstrained).

Let now (ϕ, ψ) be a maximizer of (3.3) and let γ be the unique minimizer in (3.2) (Proposition 3.4). By duality the primal-dual gap must vanish and since $H(\gamma|\mu\otimes\nu) < \infty$, γ must be of the form $\gamma = u \cdot \mu \otimes \nu$ with this we find

$$\begin{aligned} 0 &= \int c\mathrm{d}\gamma + \varepsilon H(\gamma|\mu\otimes\nu) - \int \phi \mathrm{d}\mu - \psi \mathrm{d}\nu + \varepsilon \int \exp([\phi \oplus \psi - c]/\varepsilon) \mathrm{d}\mu \otimes \nu \\ &= \int \left(\varepsilon\varphi(u) + \varepsilon \exp([\phi \oplus \psi - c]/\varepsilon) - [\phi \oplus \psi - c] \cdot u\right) \mathrm{d}\mu \otimes \nu \end{aligned}$$

Recalling that $\varphi^* = \exp$ and using the Fenchel–Young inequality [6] one finds that the integrand is non-negative and zero if and only if $u = \exp([\phi \oplus \psi - c]/\varepsilon)$ $\mu \otimes \nu$-almost everywhere. □

Proposition 3.8 (Duality for entropic transport, part III)**.** The admissible spaces in (3.3) can be relaxed to $(\phi, \psi) \in \mathrm{L}^1(\mu, [-\infty, \infty)) \times \mathrm{L}^1(\nu, [-\infty, \infty))$ without increasing the value of the supremum and maximizers exist in this space. The unique minimizer of (3.2) takes the form $\gamma = \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \lambda$ where (ϕ, ψ) are (arbitrary) maximizers of (3.3) in the relaxed space and we use the convention $\exp(-\infty) = 0$.

Proof. Note that the dual objective (3.3) is well-defined with values in $[-\infty, \infty)$ on the relaxed space, since the first two terms are finite and the third term is bounded from above (as exp is bounded from below) and can possibly take the value $-\infty$. The primal dual gap between (3.2) and (3.3) is given by (see also proof of Proposition 3.6)

$$0 \leq \int c \mathrm{d}\gamma + \varepsilon H(\gamma|\lambda) - \int \phi \mathrm{d}\mu - \psi \mathrm{d}\nu + \varepsilon \int \exp([\phi \oplus \psi - c]/\varepsilon) \mathrm{d}\lambda.$$

Again, by the Fenchel–Young inequality this inequality also holds on the relaxed space, so the supremum is not increased by the relaxation of the admissible space. By assumptions in Definition 3.2 we have $H(\mu|\lambda_1) < \infty$, $H(\nu|\lambda_2) < \infty$ and therefore

$$H(\gamma|\lambda) = H(\gamma|\mu \otimes \nu) - H(\mu|\lambda_1) - H(\nu|\lambda_2).$$

This means that we can reduce the general primal problem (3.2) to the special case of Proposition 3.6 where we assumed $\lambda = \mu \otimes \nu$. In particular both problems have the same minimizer γ. Denote $\mu_\lambda = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda_1}$ and $\nu_\lambda = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda_2}$ (these densities must exist due to the finite entropy assumption). By Proposition 3.6 we find that $\gamma = \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \mu \otimes \nu$, which can be written as $\gamma = \exp([\hat{\phi} \oplus \hat{\psi} - c]/\varepsilon) \cdot \lambda$ for $\hat{\phi} = \phi + \varepsilon \log(\mu_\lambda)$ and $\hat{\psi} = \psi + \varepsilon \log(\nu_\lambda)$ with the convention $\log(0) = -\infty$. Evaluating the primal-dual gap for γ and $(\hat{\phi}, \hat{\psi})$ yields that the latter are dual maximizers. □

Remark 3.9 (Absorption of cost function into reference measure)**.** It is easy to verify that in (3.2) one has

$$\int_{X \times X} c(x, y) \, \mathrm{d}\gamma(x, y) + \varepsilon \, H(\gamma|\lambda) = \varepsilon \, H(\gamma| \exp(-c/\varepsilon) \cdot \lambda),$$

i.e. we can interpret (3.2) as the problem of finding the generalized projection of the measure $\exp(-c/\varepsilon) \cdot \lambda$ onto the set $\Gamma(\mu, \nu)$ with respect to the divergence function H.

For X being a subset of $\mathbb{R}^d$, $c(x, y) := \frac{1}{2}|x - y|^2$, and $\lambda = \mathcal{L} \otimes \mathcal{L}$ (restricted to $X \times X$) one has that $\exp(-c/\varepsilon) \cdot \lambda$ is (proportional to) a Gaussian kernel with isotropic variance ε along each direction. Re-scaling such a kernel by a factor C corresponds to subtracting a constant $\varepsilon \log(C)$ from the cost function c and therefore will not change the minimizer of the corresponding entropic transport problem (3.2) but merely shift the objective value by $-\varepsilon \log(C)$.

Remark 3.10 (Convergence to unregularized optimal transport as $\varepsilon \to 0$)**.** Clearly (3.2) can be seen as a regularized variant of (2.1) and thus the question of convergence of the former to the latter as $\varepsilon \to 0$ is of interest, in particular convergence of minimizers,

minimal value, and likewise for the corresponding dual problems. A good starting point for the literature on this topic are [35, 17] and the references therein.

A simple explicit way of showing Γ-convergence of (3.2) to (2.1) is the *block approximation* introduced in [16] which can easily be adapted to the above setting.

Remark 3.11 (Entropic bias, Sinkhorn divergence, and loss of metric structure)**.** When comparing arbitrary probability measures in $\mathcal{P}(X)$ via (3.2) the most natural choice as reference measure appears to be $\lambda = \mu \otimes \nu$ as in Proposition 3.6, since no fixed choice for λ_1, λ_2 will satisfy the assumptions of Definition 3.2 for all μ, ν.

One drawback of this adaptive choice $\lambda = \mu \otimes \nu$ is that the function C_ε becomes non-convex, due to the added dependency of λ on the arguments (it remains convex separately in each of the two arguments, however).

An advantage is that the optimal dual potentials become more regular (since they can be chosen to be entropic c-transforms of each other, see proof of Proposition 3.6) and thus allow for more robust statistical estimation of $C_\varepsilon(\mu, \nu)$ when only empirical approximations $\hat{\mu}$ and $\hat{\nu}$ of the two measures are available, see for instance [28, 40], and ultimately also estimation of $C(\mu, \nu)$ [21].

An issue of (3.2) is the loss of the metric structure as in (2.6). Both for fixed λ and for $\lambda = \mu \otimes \nu$ one will not have in general that $\mu \in \operatorname{argmin}_{\nu \in \mathcal{P}(X)} C_\varepsilon(\mu, \nu)$. This bias can be removed by using the *Sinkhorn divergence* [26] instead, which is given as

$$S_\varepsilon(\mu, \nu) := C_\varepsilon(\mu, \nu) - \tfrac{1}{2} C_\varepsilon(\mu, \mu) - \tfrac{1}{2} C_\varepsilon(\nu, \nu)$$

where in each of the three terms on the right-hand side λ is chosen as the product of the input measures. Under suitable assumptions on c one can show that $S_\varepsilon(\mu, \nu) \geq 0$ with equality if and only if $\mu = \nu$ and that $\lim_n S_\varepsilon(\mu_n, \mu) = 0$ is equivalent to $(\mu_n)_n$ converging weak* to μ (i.e. S_ε 'metrizes' weak* convergence). However, S_ε is also no longer jointly convex and it does not satisfy the triangle inequality [33, Section 7]. A recipe to construct a metric on $\mathcal{P}(X)$ from S_ε is given in [33].

Despite this loss of metric structure, (3.2) with cost $c = \frac{1}{2} d^2$ on $X \subset \mathbb{R}^d$ is associated with an exciting dynamic perspective related to drift-diffusion equations, which we will explore in the following sections.

3.2 Interlude: diffusion and Schrödinger bridges

3.2.1 Gaussian kernels and the diffusion equation

In this section we collect some basic definitions and properties of Gaussian kernels and the diffusion equation, which will be fundamental later on.

Definition 3.12 (Gaussian kernels)**.** For a constant $\varepsilon > 0$ we introduce the Gaussian kernel with isotropic variance ε along each direction, given by

$$K_\varepsilon : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad (x, y) \mapsto \mathcal{N}_\varepsilon \cdot \exp\left(-\frac{|x - y|^2}{2\varepsilon}\right) \quad \text{with } \mathcal{N}_\varepsilon := (2\pi\varepsilon)^{-d/2} \tag{3.4}$$

where the normalization constant $\mathcal{N}_\varepsilon$ is chosen such that

$$\int_{\mathbb{R}^d} K_\varepsilon(x,y)\mathrm{d}y = 1,$$

i.e. $K_\varepsilon(x,\cdot)$ can be interpreted as probability density of a normal random variable with mean x and isotropic variance ε.

Lemma 3.13 (Convolution of Gaussian kernels)**.** For two constants $\varepsilon_1, \varepsilon_2 > 0$ one has

$$\int_{\mathbb{R}^d} K_{\varepsilon_1}(x,y)\, K_{\varepsilon_2}(y,z)\,\mathrm{d}y = K_{\varepsilon_1+\varepsilon_2}(x,z).$$

This can be verified by a simple (but somewhat tedious) explicit computation. The identity can also be interpreted from the perspective of adding independent normal random variables: Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be independent normal $\mathbb{R}^d$-valued random variables with isotropic covariance matrices $\varepsilon_1 \cdot \mathrm{id}$ and $\varepsilon_2 \cdot \mathrm{id}$, respectively. Then $\boldsymbol{x}_1 + \boldsymbol{x}_2$ is a normal $\mathbb{R}^d$-valued random variable with isotropic covariance matrix $(\varepsilon_1 + \varepsilon_2) \cdot \mathrm{id}$.

Lemma 3.14 (Diffusion equation or heat equation)**.** For $u_0 \in \mathrm{L}^1(\mathbb{R}^d)$, $t > 0$ and $x \in \mathbb{R}^d$ let

$$u(t,x) := \int_{\mathbb{R}} u_0(y)\, K_{\varepsilon t}(x,y)\mathrm{d}y.$$

Then u is infinitely often continuously differentiable with respect to both variables, with derivatives bounded on $(\delta,\infty] \times \mathbb{R}^d$ for any $\delta > 0$ and it solves the diffusion or heat equation

$$\partial_t u(t,x) = \frac{\varepsilon}{2}\Delta u(t,x)$$

for all $(t,x) \in \mathbb{R}_{++} \times \mathbb{R}^d$ with the boundary condition at $t = 0$ in the sense that $\|u(0,\cdot) - u(t,\cdot)\|_{\mathrm{L}^1(\mathbb{R}^d} \to 0$. For times $0 < s < t$ one finds with Lemma 3.13 that

$$u(t,x) = \int_{\mathbb{R}} u(s,y)\, K_{\varepsilon(t-s)}(x,y)\mathrm{d}y.$$

For an introductory treatment of this equation see [25].

Remark 3.15 (Stochastic interpretation)**.** Let the stochastic process $(\boldsymbol{x}_t)_{t\geq 0}$ describe the trajectory of a particle moving in $\mathbb{R}^d$, subjected to Brownian motion with strength ε. That is, $\boldsymbol{x}_t$ is a solution to the stochastic partial differential equation

$$\mathrm{d}\boldsymbol{x}_t = \sqrt{\varepsilon}\mathrm{d}W_t$$

where W_t denotes a standard Wiener process. Let the probability density for the law of $\boldsymbol{x}_0$ be given by u_0. Then $u(t,\cdot)$ will be the distribution of $\boldsymbol{x}_t$ at time $t > 0$, since the density of the distribution of $\boldsymbol{x}_t$, conditioned on $\boldsymbol{x}_0 = x_0$ is given by $K_{\varepsilon t}(x_0,\cdot)$. $u_0(x_0) \cdot K_{\varepsilon t}(x_0,x_t)$ will be the joint density of first observing $\boldsymbol{x}_0$ in x_0 and then $\boldsymbol{x}_t$ at x_t, and more generally $u_0(x_0)\prod_{i=1}^N K_{\varepsilon(t_i-t_{i-1})}(x_{i-1},x_i)$ will be the joint density of observing $\boldsymbol{x}_{t_i}$ in x_i for an increasing tuple $t_0 = 0 < t_1 < \ldots < t_N$. Note that this joint distribution is Markov in the sense that $\boldsymbol{x}_{t_i}$ and $\boldsymbol{x}_{t_k}$ are independent when conditioned on $\boldsymbol{x}_{t_j}$ for some triple $i < j < k$.

The following definition is the equivalent of Definition 2.43 for the entropic setting.

Definition 3.16 (Multi-marginal Gaussian kernel)**.** For a natural number $N \geq 2$ and time points $t_0 = 0 < t_1 < t_2 < \ldots < t_N = 1$, the multi-marginal Gaussian kernel is given by

$$\begin{aligned} &K_{\mathrm{MM}} : (\mathbb{R}^d)^{N+1} \mapsto \mathbb{R}, \\ &(x_0, \ldots, x_N) \mapsto \prod_{i=0}^{N-1} K_{\varepsilon(t_{i+1}-t_i)}(x_i, x_{i+1}) = \prod_{i=0}^{N-1} \mathcal{N}_{\varepsilon(t_{i+1}-t_i)} \exp\left(-\tfrac{|x_{i+1}-x_i|^2}{2\varepsilon(t_{i+1}-t_i)}\right) \end{aligned} \tag{3.5}$$

with normalization factors $\mathcal{N}_{\varepsilon(t_{i+1}-t_i)}$ as defined in (3.4).

And the following is the corresponding equivalent of Lemma 2.45, which is an immediate consequence of Lemma 3.13.

Lemma 3.17. For K_{MM} as in Definition 3.16 one has

$$K_\varepsilon(x_0, x_N) = \int_{(\mathbb{R}^d)^{N-1}} K_{\mathrm{MM}}(x_0, \ldots, x_N) \mathrm{d}x_1 \ldots \mathrm{d}x_{N-1}.$$

3.2.2 Schrödinger bridges

In this section we briefly sketch a thought experiment proposed by Erwin Schrödinger in 1931 which provides an insightful interpretation of the entropic optimal transport problem (3.2) and helps to discern its dynamic structure. A translation of the original paper with historical context is given in [20], see also [36] and [18] and references therein for more literature on this problem.

Consider once more a stochastic process $(\boldsymbol{x}_t)_{t\geq 0}$, describing a particle subjected to Brownian motion of strength ε, as in Remark 3.15. Assume its initial position $\boldsymbol{x}_0$ was fixed to x_0. Then for any $t > 0$, the law of $\boldsymbol{x}_t$ will have the marginal probability density $K_{\varepsilon t}(x_0, \cdot)$. Assume now that we 'observe' the particle at $t = 1$ in position x_1. Based on this knowledge, what do we know about the likely positions of the particle at intermediate times $t \in (0, 1)$? Hypothetically, we could consider a setup where we initialize the experiment many times and observe the position of the particle at time $t = 1$ and discard all instances where the particle is not within an infinitesimally small environment of x_1. A second observer will determine the position of the particle at a fixed agreed-upon intermediate time $t \in (0, 1)$ and only keep those measurements for which we later reported that the particle ended up in the aforementioned small environment. What will be the distribution of the non-discarded observed intermediate positions? What is the law of $\boldsymbol{x}_t$ conditioned on $\boldsymbol{x}_0 = x_0$ and $\boldsymbol{x}_1 = x_1$? Such a process is called a *Brownian bridge* [5].

Using that the density for $\boldsymbol{x}_1 = x_1$ is given by $K_\varepsilon(x_0, x_1)$ and the joint density for $(\boldsymbol{x}_t = x_t, \boldsymbol{x}_1 = x_1)$ is given by $K_{\varepsilon t}(x_0, x_t) \cdot K_{\varepsilon(1-t)}(x_t, x_1)$ (Remark 3.15) a tedious but simple computation yields that the conditional density for $\boldsymbol{x}_t = x_t$ conditioned on

$\boldsymbol{x}_0 = x_0$, $\boldsymbol{x}_1 = x_1$ is given by

$$\begin{aligned} u(x_t|\boldsymbol{x}_0 = x_0, \boldsymbol{x}_1 = x_1) &= \frac{K_{\varepsilon t}(x_0, x_t) \cdot K_{\varepsilon(1-t)}(x_t, x_1)}{K_\varepsilon(x_0, x_1)} \\ &= \frac{\mathcal{N}_{\varepsilon t} \cdot \mathcal{N}_{\varepsilon(1-t)}}{\mathcal{N}_\varepsilon} \exp\left(-\frac{|x_0 - x_t|^2}{2\varepsilon t} - \frac{|x_t - x_1|^2}{2\varepsilon(1-t)} + \frac{|x_0 - x_1|^2}{2\varepsilon}\right) \\ &= \mathcal{N}_{\varepsilon t(1-t)} \exp\left(-\frac{|x_t - [(1-t)x_0 + tx_1]|^2}{2\varepsilon t(1-t)}\right) \qquad (3.6) \\ &= K_{\varepsilon t(1-t)}((1-t)x_0 + tx_1, x_t) \qquad (3.7) \end{aligned}$$

So we find that the conditional $\boldsymbol{x}_t$ is also a normal random variable with mean $(1-t) \cdot x_0 + t \cdot x_1$ and covariance $\varepsilon\, t\, (1-t) \cdot \mathrm{id}$.

Schrödinger considered a generalization of this problem: Assume now that the initial and final positions are no longer fixed, but instead the law of $\boldsymbol{x}_0$ is prescribed to be given by some measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ and the law of $\boldsymbol{x}_1$ is prescribed to be given by some $\nu \in \mathcal{P}(\mathbb{R}^d)$. What will be the law of the intermediate time position $\boldsymbol{x}_t$? This conditional process has been dubbed the *Schrödinger bridge* between μ and ν [18] and it is closely related to the entropic optimal transport problem (3.2). We will examine this question in the following by intuitive considerations on discrete spaces. A more thorough but still rather accessible approach can be found in [31, Section 6].

In this thought experiment we face the difficulty of how to 'condition' $\boldsymbol{x}_1$ on a particular law ν instead of a fixed position. This can be resolved by the following intuitive arguments. We introduce $M \in \mathbb{N}$ independent and identically distributed copies of the process $\boldsymbol{x}_t$, denoted as $(\boldsymbol{x}_t^i)_{t\geq 0, i\in\{1,\dots,M\}}$. We then conduct the experiment many times where we observe the initial and final positions of all particles at times $t = 0$ and $t = 1$ and the second observer records the positions at the fixed intermediate time $t \in (0, 1)$. Afterwards we consider for each run of the experiment the empirical measures

$$\mu^M := \tfrac{1}{M} \sum_{i=1}^M \delta_{\boldsymbol{x}_0^i} \qquad \text{and} \qquad \nu^M := \tfrac{1}{M} \sum_{i=1}^M \delta_{\boldsymbol{x}_1^i}$$

and only keep those realizations where μ^M and ν^M are 'close' in some suitable sense to the prescribed μ and ν. For instance, we could partition $\mathbb{R}^d$ into cells and compare the masses within the cells. What will be the distribution of empirical measures at intermediate times,

$$\rho_t^M := \tfrac{1}{M} \sum_{i=1}^M \delta_{\boldsymbol{x}_t^i},$$

that the second observer reports?

Concretely, let now $X := \{x_1, \dots, x_L\}$ be a finite space (e.g. a finite subset of $\mathbb{R}^d$) and let $\mu := \sum_{l=1}^L \mu_l \cdot \delta_{x_l}$ be a probability measure on X with mass weights given by coefficients μ_l, i.e. by a slight abuse of notation we identify μ with a vector in $\mathbb{R}^L$. Assume that the probability for observing $\boldsymbol{x}_0 = x_l$ equals μ_l (e.g. by 'truncating' the above random

variable $\boldsymbol{x}_0$ from $\mathbb{R}^d$ to X). Then we are interested in the law of the empirical random measure

$$\mu^M := \tfrac{1}{M}\sum_{i=1}^{M}\delta_{\boldsymbol{x}_0^i} = \sum_{l=1}^{L}\tfrac{\boldsymbol{m}_l}{M}\delta_{x_l} = \sum_{l=1}^{L}\boldsymbol{\mu}_l^M\delta_{x_l}.$$

Here $(\boldsymbol{m}_l)_{l=1}^L$ are random variables defined by counting how many of the $\boldsymbol{x}_0^i$ are equal to each x_l and $\boldsymbol{\mu}_l^M$ are the random empirical mass weights. Clearly the $(\boldsymbol{m}_l)_{l=1}^L$ follow a multinomial distribution with probability

$$\mathbb{P}(\boldsymbol{m}_1 = m_1, \dots, \boldsymbol{m}_L = m_L) = \frac{M!}{\prod_{l=1}^L m_l!}\prod_{l=1}^{L}\mu_l^{m_l}.$$

Consider now the regime of very large M. Using Stirling's approximation $\log(n!) \approx n\log(n) - n$ one finds

$$\begin{aligned}-\log(\mathbb{P}(\boldsymbol{m}_1 = m_1, \dots, \boldsymbol{m}_L = m_L)) &\approx M\sum_{l=1}^{L}(m_l/M)\log((m_l/M)/\mu_l)\\ &= M\cdot\mathrm{KL}((m_l/M)_l|\mu).\end{aligned}$$

This simple computation underlines the intimate connection of entropy to random sampling. In conclusion, for very large M, the empirical discrete weights μ^M will be very close to μ with high probability.

Now we expand this to the product space. Let $K = \sum_{i,j=1}^{L} K_{i,j}\,\delta_{(x_i,x_j)} \in \mathcal{P}(X\times X)$ be the discrete joint law of $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ with mass coefficients $K_{i,j}$, i.e. similar to above we identify K with a matrix in $\mathbb{R}^{L\times L}$. Again, let $\boldsymbol{m}_{j,l}$ be the random variable that counts how often we observe $(\boldsymbol{x}_0^i, \boldsymbol{x}_1^i) = (x_j, x_l)$ for $j, l \in \{1, \dots, L\}$. Let $\boldsymbol{k}_{j,l}^M := \boldsymbol{m}_{j,l}/M$ be the discrete empirical weights and we identify the weight matrix $\boldsymbol{k}^M$ with the corresponding empirical measure on $\mathcal{P}(X\times X)$. Then for large M again

$$-\log(\mathbb{P}(\boldsymbol{m}_{j,l} = m_{j,l} \text{ for } j,l \in \{1,\dots,L\}) \approx M\cdot\mathrm{KL}((m_{j,l}/M)_{j,l}|K).$$

Now we conduct this experiment many times. With high probability in most instances $\boldsymbol{k}^M$ will be close to K. However, we now choose to only retain instances where the marginals of $\boldsymbol{k}^M$ are equal to (or very close to) the prescribed measures μ and ν, i.e. we condition on $\boldsymbol{k}^M \in \Gamma(\mu,\nu)$. Then, by the above intuitive arguments, with high probability $\boldsymbol{k}^M$ will be close to the minimizer

$$\operatorname{argmin}\left\{\mathrm{KL}(\gamma|K) | \gamma \in \Gamma(\mu,\nu)\right\}. \tag{3.8}$$

Once this minimizer γ has been determined, which describes the joint law of $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ 'conditioned' on the marginals μ and ν in the above sense, then the law of $\boldsymbol{x}_t$ can be constructed via (3.7). The density of $\boldsymbol{x}_t$ will be given by

$$\sum_{j,l=1}^{L} u(x_t|\boldsymbol{x}_0 = x_j, \boldsymbol{x}_1 = x_l)\,\gamma(\{(x_j,x_l)\}) \tag{3.9}$$

and hence we have established the link between Schrödinger bridges and the entropic optimal transport problem (3.2) (see also Remark 3.9). We will examine this further in Section 3.3.

Remark 3.18 (Lazy gas experiment)**.** In the above thought experiment we have intuitively conditioned a collection of many particles undergoing Brownian motion on a specific marginal distribution ν at time $t = 1$. The expectation for the random measure $\frac{1}{M}\sum_{i=1}^{M} \delta_{\boldsymbol{x}_1^i}$ is given by the evolution of the initial distribution μ under the diffusion equation (Remark 3.15). For large M, deviations from this prediction become increasingly unlikely. When we decrease the amplitude of the Brownian motion by sending the parameter ε to 0, then the expected distribution at $t = 1$ will converge to μ. For small $\varepsilon > 0$, particles will barely move at all. If we still insist on conditioning on some 'exotic' law ν, then the mass particles will try to reach this configuration by moving as little as possible (or at least barely more). This means that the particles will move approximately along an unregularized Wasserstein geodesic. This thought experiment has been dubbed the *lazy gas experiment* in [51, Chapter 16]. The limit $\varepsilon \to 0$ is discussed in [36, Section 6.2], see also [35]. In [51, Chapter 16] the limit case $\varepsilon = 0$ is considered directly and it is discussed how it can reveal information on the curvature of the base space X (in cases where X is a Riemannian manifold).

3.3 Entropic multi-marginal formulation

Remark 3.19 (Lebesgue densities and measures)**.** Throughout this section we will only consider measures that are absolutely continuous with respect to the Lebesgue measure $\mathcal{L}$ on $\mathbb{R}^d$, or its restriction to X which we denote by $\mathcal{L} \llcorner X$, or with respect to the product of the Lebesgue measure on spaces like $(\mathbb{R}^d)^{N+1}$ et cetera. We denote their Lebesgue densities by the same symbol as the measure itself. Conversely, we will also use the symbols K_ε and K_{MM} introduced as Gaussian densities in Section 3.2.1 to refer to the corresponding measures. We will view K_ε as a measure restricted to $X \times X$ and K_{MM} as a measure restricted $X \times (\mathbb{R}^d)^{N-1} \times X$ such that both are finite.

Based on the discussion in Section 3.2.2 we now introduce the following definition.

Definition 3.20 (Static Schrödinger bridge problem)**.**

$$\min \{\varepsilon\, H(\gamma|K_\varepsilon) | \gamma \in \Gamma(\mu, \nu)\} \tag{3.10}$$

Taking into account Proposition 3.8 and Remark 3.9 we obtain that (3.10) has a unique minimizer of the form $\gamma = \exp(\phi \oplus \psi/\varepsilon) \cdot K_\varepsilon$ with (ϕ, ψ) being maximizers of the following dual problem

$$\max \left\{ \int_X \phi \mathrm{d}\mu + \int_X \psi \mathrm{d}\nu - \varepsilon \int_{X\times X} \exp\left(\frac{\phi \oplus \psi}{\varepsilon}\right) \mathrm{d}K_\varepsilon \middle| \phi \in \mathrm{L}^1(\mu), \psi \times \mathrm{L}^1(\nu) \right\}. \tag{3.11}$$

In analogy to (2.19), inspired by Section 3.2.2, to study the distribution of particles at intermediate times, we introduce the following multi-marginal problems.

Proposition 3.21 (Multi-marginal Schrödinger bridge problem)**.** Problems (3.10) and (3.11) are equivalent to problems

$$\min \left\{ \varepsilon\, H(\gamma_{\mathrm{MM}} | K_{\mathrm{MM}}) \middle| \gamma_{\mathrm{MM}} \in \mathcal{P}(X \times (\mathbb{R}^d)^{N-1} \times X), \mathrm{P}_0 \gamma_{\mathrm{MM}} = \mu,\, \mathrm{P}_N \gamma_{\mathrm{MM}} = \nu \right\} \quad (3.12)$$

and

$$\max \left\{ \int_X \phi \mathrm{d}\mu + \int_X \psi \mathrm{d}\nu - \varepsilon \int_{X \times (\mathbb{R}^d)^{N-1} \times X} \exp\left(\frac{\phi(x_0) + \psi(x_N)}{\varepsilon} \right) \mathrm{d}K_{\mathrm{MM}}(x_0, \dots, x_N) \middle| \right.$$
$$\left. \phi \in \mathrm{L}^1(\mu), \psi \in \mathrm{L}^1(\nu) \right\} \quad (3.13)$$

respectively in the following sense: their optimal values are identical, (3.11) and (3.13) have the same maximizers, and minimizers for (3.12) are given by $\gamma_{\mathrm{MM}} = \exp(\phi \oplus \psi / \varepsilon) \cdot K_{\mathrm{MM}}$ where (ϕ, ψ) are maximal in (3.11) or (3.13), in analogy to the minimizers for (3.10).

Proof. Equivalence between (3.11) and (3.13) is immediate from Lemma 3.17. (3.13) $\leq$ (3.12) follows by showing that the corresponding primal-dual gap is non-negative using the Fenchel–Young inequality, in the exact way as done, for instance in the proof of Proposition 3.8. Plugging $\gamma_{\mathrm{MM}} = \exp(\phi \oplus \psi / \varepsilon) \cdot K_{\mathrm{MM}}$ into (3.12) we find that the primal-dual gap vanishes, establishing minimality of γ_{MM}. Uniqueness follows from strict convexity of $H(\cdot | K_{\mathrm{MM}})$.

Instead of showing minimality of γ_{MM} via duality, we can also use an explicit primal argument to obtain the inequality (3.12) $\geq$ (3.10) as follows: Let γ_{MM} be admissible in (3.12) with finite objective, i.e. it will have the form $\gamma_{\mathrm{MM}} = u_{\mathrm{MM}} \cdot K_{\mathrm{MM}}$ for some (relative) density u_{MM}. Then $\gamma := \mathrm{P}_{0,N} \gamma_{\mathrm{MM}}$ will be admissible in (3.10) and it will have the form $\gamma = u \cdot K_\varepsilon$ with

$$u(x_0, x_N) = \int_{(\mathbb{R}^d)^{N-1}} u_{\mathrm{MM}}(x_0, \dots, x_N) \frac{K_{\mathrm{MM}}(x_0, \dots, x_N)}{K_\varepsilon(x_0, x_N)} \mathrm{d}x_1 \dots \mathrm{d}x_{N-1}.$$

Using that $\frac{K_{\mathrm{MM}}(x_0, \dots, x_N)}{K_\varepsilon(x_0, x_N)}$ is a probability density on $(\mathbb{R}^d)^{N-1}$ and that φ (the integrand in (3.1)) is convex one obtains that $H(\gamma | K_\varepsilon) \leq H(\gamma_{\mathrm{MM}} | K_{\mathrm{MM}})$ through Jensen's inequality. □

Given a minimizer γ_{MM} of (3.12) one can evaluate the marginal at intermediate times to obtain a notion of how the measure μ is gradually transformed into ν. Choosing for simplicity $N = 2$ and $t_0 = 0$, $t_1 = t$, $t_2 = 1$ one obtains for (the Lebesgue density of) the

marginal at time t, which we denote by $\rho(t,\cdot)$ that

$$\begin{aligned}
\rho(t,y) &= \int_{X\times X} \gamma_{\mathrm{MM}}(x,y,z)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{X\times X} \exp(\phi(x)/\varepsilon)\exp(\psi(z)/\varepsilon) K_{\mathrm{MM}}(x,y,z)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{X\times X} \exp(\phi(x)/\varepsilon)\exp(\psi(z)/\varepsilon) K_{\varepsilon t}(x,y)\,K_{\varepsilon(1-t)}(y,z)\,\mathrm{d}x\,\mathrm{d}z \qquad (3.14)\\
&= \int_{X\times X} \frac{K_{\varepsilon t}(x,y)\,K_{\varepsilon(1-t)}(y,z)}{K_\varepsilon(x,z)}\,\mathrm{d}\gamma(x,z) \\
&= \int_{X\times X} K_{\varepsilon t(1-t)}(t\cdot x + (1-t)\cdot z, y)\,\mathrm{d}\gamma(x,z) \qquad (3.15)
\end{aligned}$$

for $y\in\mathbb{R}^d$ and γ being a corresponding minimizer of (3.10). Note that (3.15) is the same formula as (3.9). For $N>2$ and any given intermediate marginal $i\in\{1,\dots,N-1\}$ one obtains a similar formula by integrating over all other intermediate axes $j\neq i$.

Remark 3.22 (Applications of (3.12) and limit $N\to\infty$)**.** In analogy to Remark 2.46, formulation (3.12) provides a way to interpret how the measure μ is gradually transformed into the measure ν by looking at marginals of γ_{MM} at intermediate times as in (3.14). These are the time-marginals of the Schrödinger bridge between μ and ν (cf. Section 3.2.2). As before, formulation (3.12) has the advantage over (3.10) that one can include interactions of the measure with its environment at intermediate times, such as in [30, 34].

Analogous to Remark 2.49 one may also consider again the limit $N\to\infty$ such that γ_{MM} becomes a measure on paths. K_{MM} will then essentially turn into the Wiener measure (but without fixing the initial point). We refer to [36] and references therein for more information on this formulation.

Remark 3.23 (Markov property)**.** From (3.14) we can observe that the minimizing γ_{MM} is Markov in the same sense as in the unregularized setting in Remark 2.48. Therefore, it is again possible to restrict to such Markov densities in (3.12). This property remains preserved, then the objective is augmented by additional terms that interact with the intermediate densities and plays a key role for numerical tractability [11, 30, 7, 34].

Equation (3.14) motivates the following definitions, which are the entropic equivalent of the Hopf–Lax formulas (2.24).

Definition 3.24. Let (ϕ,ψ) be maximizers for (3.11) or equivalently (3.13). Then we introduce the auxiliary functions $U,V,\Phi,\Psi:[0,1]\times\mathbb{R}^d\to[-\infty,\infty]$ as follows

$$U(0,\cdot) := \exp(\phi/\varepsilon), \qquad U(t,x) := \int_X U(0,y)\,K_{\varepsilon t}(x,y)\mathrm{d}y \qquad \text{for } t\in(0,1], \qquad (3.16)$$

$$V(1,\cdot) := \exp(\psi/\varepsilon), \qquad V(t,x) := \int_X V(1,y)\,K_{\varepsilon(1-t)}(x,y)\mathrm{d}y \qquad \text{for } t\in[0,1) \qquad (3.17)$$

with the convention $U(0,x) = 0$ if $\phi(0,x) = -\infty$ (and likewise for $V(1,x)$ and $\psi(1,x)$), and

$$\Phi(t,x) := \varepsilon \log(U(t,x)), \qquad \Psi(t,x) := \varepsilon \log(V(t,x)) \tag{3.18}$$

with the convention $\Phi(0,x) = -\infty$ if $\phi(0,x) = -\infty$ for $x \in X$ or $x \in \mathbb{R}^d \setminus X$ (and analogously for $\Psi(1,\cdot)$).

Proposition 3.25. Consider the functions U, V, and Ψ as introduced in Definition 3.24. Let $\rho : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ be given by (3.14) for $t \in (0,1)$ and by the densities of μ and ν for $t \in \{0,1\}$ (which we extend by 0 beyond X). Then on $[0,1] \times \mathbb{R}^d$ one has

$$\rho = U \cdot V, \tag{3.19a}$$

and on $(0,1) \times \mathbb{R}^d$ one has

$$\partial_t U = \tfrac{\varepsilon}{2} \Delta U, \tag{3.19b}$$
$$\partial_t V = -\tfrac{\varepsilon}{2} \Delta V, \tag{3.19c}$$
$$\partial_t \rho = -\nabla[\nabla\Psi \cdot \rho] + \tfrac{\varepsilon}{2} \Delta \rho, \tag{3.19d}$$
$$\partial_t \Psi = -\tfrac{1}{2} |\nabla \Psi|^2 - \tfrac{\varepsilon}{2} \Delta \Psi. \tag{3.19e}$$

Proof. For $t \in (0,1)$ (3.19a) follows by plugging the definitions of U and V into (3.14). For $t \in \{0,1\}$ it follows by plugging the definitions of U and V into the condition that minimizers γ of (3.10) lie in $\Gamma(\mu,\nu)$.

(3.19b) and (3.19c) follow from Lemma 3.14, taking into account that V is convolved 'backwards' in time. (3.19d) and (3.19e) follow from (3.19b) and (3.19c) and exploiting that U and V are infinitely often differentiable and strictly positive on $(0,1) \times \mathbb{R}^d$, and using the definition of Ψ. □

We observe that the density ρ is on the one hand given as product of two solutions to the diffusion equation, one forward and one backward in time. And it is given as solution to a Fokker–Planck equation with a drift potential that satisfies a suitable adjoint equation. These two equations are remarkably similar to the system of equations appearing in Proposition 2.38 up to the additional ε-terms.

The next remark (in the spirit of [10, Lemma 3.4]) shows that the Markov property also allows to decompose the objective of (3.12) into a sum of step-wise objectives, similar to (2.22) for the unregularized case.

Remark 3.26 (Entropy for Markov plans)**.** Assume that a density γ_{MM} admissible in (3.12) has the form

$$\gamma_{\mathrm{MM}}(x_0, \dots, x_N) = \mu(x_0) \prod_{i=0}^{N-1} \gamma_{i,i+1}(x_i, x_{i+1})$$

for suitable pairwise densities $\gamma_{i,i+1}$. Then

$$H(\gamma_{\mathrm{MM}}|K_{\mathrm{MM}}) = H(\mu|\mathcal{L}\llcorner X) + \sum_{i=0}^{N-1} \int_{X\times X} \log\left(\frac{\gamma_{i,i+1}(x_i,x_{i+1})}{K_{\varepsilon(t_{i+1}-t_i)}(x_i,x_{i+1})}\right) \gamma_{i,i+1}(x_i,x_{i+1})\,\mathrm{d}x_i\,\mathrm{d}x_{i+1}. \quad (3.20)$$

This is easy to verify by explicit computations. When γ_{MM} is the minimizer, doing a calculation as for (3.14) but for the joint distribution of two intermediate times, one finds that one can choose (ignoring potential regularity issues at $t \in \{0,1\}$)

$$\gamma_{i,i+1}(x_i,x_{i+1}) = \frac{V(t_{i+1},x_{i+1})}{V(t_i,x_i)} \cdot K_{\varepsilon(t_{i+1}-t_i)}(x_i,x_{i+1}).$$

With this (3.20) reduces to

$$\begin{aligned}
H(\gamma_{\mathrm{MM}}|K_{\mathrm{MM}}) &= H(\mu|\mathcal{L}\llcorner X) + \sum_{i=0}^{N-1} \int_{X\times X} \log\left(\frac{V(t_{i+1},x_{i+1})}{V(t_i,x_i)}\right) \gamma_{i,i+1}(x_i,x_{i+1})\,\mathrm{d}x_i\,\mathrm{d}x_{i+1}.\\
&= H(\mu|\mathcal{L}\llcorner X) + \sum_{i=0}^{N-1} \Big[\int_X \log\left(V(t_{i+1},x_{i+1})\right) \rho(t_{i+1},x_{i+1})\,\mathrm{d}x_{i+1}\\
&\qquad - \int_X \log\left(V(t_i,x_i)\right) \rho(t_i,x_i)\,\mathrm{d}x_i\Big]\\
&= H(\mu|\mathcal{L}\llcorner X) + \tfrac{1}{\varepsilon}\sum_{i=0}^{N-1} \Big[\int_X \Psi(t_{i+1},x_{i+1})\,\rho(t_{i+1},x_{i+1})\,\mathrm{d}x_{i+1}\\
&\qquad - \int_X \Psi(t_i,x_i)\,\rho(t_i,x_i)\,\mathrm{d}x_i\Big]\\
&= H(\mu|\mathcal{L}\llcorner X) + \tfrac{1}{\varepsilon}\int_X \Psi(1,\cdot)\mathrm{d}\nu - \tfrac{1}{\varepsilon}\int_X \Psi(0,\cdot)\mathrm{d}\mu. \qquad (3.21)
\end{aligned}$$

Using the relation $\phi(x) = -\Psi(0,\cdot) + \varepsilon\log(\mu(x))$, and $\int_{X\times X} \exp\left(\frac{\phi\oplus\psi}{\varepsilon}\right)\mathrm{d}K_\varepsilon = 1$, this then equals (3.11), thus showing that indeed in this case the primal-dual gap vanishes.

3.4 Entropic Benamou–Brenier formula

3.4.1 Primal and dual formulation

In Section 3.3 we have obtained a Lagrangian dynamic formulation of entropic optimal transport, that provided a way to interpret how measure μ is transformed into ν. In Section 2.3 we studied a Eulerian dynamic formulation for *unregularized* optimal transport. The existence of such a formulation might not be so surprising, since we saw in the unregularized setting that the moving particles do not collide at intermediate times (Corollary 2.22). However, according to (3.15) in the entropic setting, particles extensively move though each other at intermediate times. It may therefore be surprising that

a Eulerian formulation even exists in the regularized setting, introduced in [18] and also studied in [29]. In this section we set up this formulation and sketch the equivalence in Section 3.4.2.

Inspired by (3.19) we now introduce a diffusive version of the distributional continuity equation (Definition 2.25). Formally, these correspond to the PDE

$$\partial_t \rho + \nabla \cdot \omega = \tfrac{\varepsilon}{2}\Delta\rho \tag{3.22}$$

with temporal boundary conditions $\rho(0,\cdot) = \mu$ and $\rho(1,\cdot) = \nu$.

Definition 3.27 (Continuity equation with diffusion)**.** Let μ, $\nu \in \mathcal{P}(X)$, $\varepsilon > 0$. A pair $(\rho, \omega) \in \mathcal{M}([0,1]\times X) \times \mathcal{M}([0,1]\times X)^d$ is said to solve the distributional continuity equation with diffusion and temporal boundary conditions μ and ν if

$$\int_{[0,1]\times\mathbb{R}^d} (\partial_t\psi)\,\mathrm{d}\rho + \int_{[0,1]\times\mathbb{R}^d} \nabla\psi\cdot\mathrm{d}\omega + \tfrac{\varepsilon}{2}\int_{[0,1]\times\mathbb{R}^d} \Delta\psi\mathrm{d}\rho = \int_X \psi(1,\cdot)\,\mathrm{d}\nu - \int_X \psi(0,\cdot)\,\mathrm{d}\mu \tag{3.23}$$

for all $\psi \in C^2([0,1]\times X)$. We denote the set of solutions by $\mathcal{CE}_\varepsilon(\mu,\nu)$.

The entropic primal Benamou–Brenier formula is then obtained by simply replacing $\mathcal{CE}(\mu,\nu)$ with $\mathcal{CE}_\varepsilon(\mu,\nu)$ as admissible set in (2.12).

Definition 3.28 (Entropic Benamou–Brenier formula)**.**

$$C_{\varepsilon,\mathrm{BB}}(\mu,\nu) := \inf\{\mathcal{A}(\rho,\omega) | (\rho,\omega) \in \mathcal{CE}_\varepsilon(\mu,\nu)\} \tag{3.24}$$

Similar to Proposition 2.37 we can now obtain a corresponding dual formulation.

Proposition 3.29 (Dual entropic Benamou–Brenier formula)**.**

$$C_{\varepsilon,\mathrm{BB}}(\mu,\nu) = \sup\left\{ \int_X \Psi(1,\cdot)\,\mathrm{d}\nu - \int_X \Psi(0,\cdot)\,\mathrm{d}\mu \,\middle|\, \Psi \in \mathrm{C}^2([0,1]\times X),\ \partial_t\Psi + \frac{1}{2}|\nabla\Psi|^2 + \tfrac{\varepsilon}{2}\Delta\Psi \le 0 \right\} \tag{3.25}$$

Proof. The proof works in complete analogy to Proposition 2.37 where one now chooses $U = \mathrm{C}^2([0,1]\times X)$ and $A : \Psi \mapsto (\partial_t\Psi + \frac{\varepsilon}{2}\Delta\Psi, \nabla\Psi)$. □

And in complete analogy with Proposition 2.38 one obtains the following primal-dual optimality conditions. Similar to Remark 2.40 we do not insist on existence of dual maximizers here, but we are primarily interested in the formal interaction between primal mass movement and the dynamic dual potential.

Proposition 3.30 (Primal-dual optimality conditions for the entropic Benamou–Brenier formulation)**.** A pair $(\rho,\omega) \in \mathcal{M}([0,1]\times X)^{1+d}$ and $\Psi \in \mathrm{C}^1([0,1]\times X)$ is primal-dual optimal for (3.24) and (3.25) if and only if

$$\partial_t\rho + \nabla\cdot\omega + \tfrac{\varepsilon}{2}\Delta\rho = 0 \quad \text{with temporal boundary conditions } \rho(0,\cdot) = \mu,\ \rho(1,\cdot) = \nu$$

in the distributional sense of (3.23),

$$\rho \geq 0, \qquad \omega = \nabla\Psi \cdot \rho,$$

and

$$\partial_t \Psi + \tfrac{1}{2}|\nabla\Psi|^2 + \tfrac{\varepsilon}{2}\Delta\Psi \leq 0 \text{ with equality } \rho\text{-almost everywhere.}$$

For simpler duality arguments we have restricted the above formulations to the compact domain X. Relaxation to the whole $\mathbb{R}^d$ requires careful approximation arguments that we cannot cover here.

3.4.2 Equivalence with entropic Kantorovich formulation

In this section we now sketch the formal equivalence between formulations (3.24) and (3.25) with their Lagrangian counter parts (3.10) and (3.11). This equivalence was shown in [18] with tools from stochastic analysis. For the sake of brevity the arguments will not be fully rigorous as there are some issues related to the regularity of dual candidates for (3.24) and the compact support assumed in Section 3.4.1 which would require more extensive arguments.

Remark 3.31. Formally one has (3.24) $\leq$ (3.10) $- \varepsilon H(\mu)$.

Sketch of proof. Let (ϕ, ψ) be solutions to (3.3) or (3.13), let ρ be as in (3.14), and let U, V, Φ, and Ψ constructed as in Definition 3.24. Let $\omega := \nabla\Psi \cdot \rho$ be a vector-valued density. Then by (3.19d), the pair (ρ, ω) is a strong solution of (3.22) on $(0,1) \times \mathbb{R}^d$ and therefore by partial integration as distributional solution to (3.23) relaxed to the whole $\mathbb{R}^d$. Therefore, the pair (ρ, ω) is admissible in (3.24). We obtain Plugging (ρ, ω) into the objective of (3.24) one obtains

$$\begin{aligned}
(3.24) &\leq \mathcal{A}(\rho, \omega) \\
&= \int_{[0,1]\times\mathbb{R}^d} \Phi(1, \nabla\Psi)\, \rho(t,x)\, \mathrm{d}x\mathrm{d}t \\
&= \int_{[0,1]\times\mathbb{R}^d} \tfrac{1}{2}|\nabla\Psi|^2\, \rho(t,x)\, \mathrm{d}x\mathrm{d}t \\
&= \int_0^1 \left[\int_{\mathbb{R}^d} \partial_t(\Psi \cdot \rho)\, \mathrm{d}x \right] \mathrm{d}t \\
&= \int_X \Psi(1,\cdot)\rho(1,\cdot)\, \mathrm{d}x - \int_X \Psi(0,\cdot)\rho(0,\cdot)\, \mathrm{d}x \\
&= \int_X \psi \cdot \nu\, \mathrm{d}x + \int_X \phi \cdot \mu\, \mathrm{d}x - \varepsilon \int_X \log(\mu) \cdot \mu\, \mathrm{d}x \\
&= \int_X \psi \cdot \nu\, \mathrm{d}x + \int_X \phi \cdot \mu\, \mathrm{d}x - \varepsilon \int_{X\times X} \exp\left(\frac{\phi \oplus \psi}{\varepsilon}\right) \mathrm{d}K_\varepsilon - \varepsilon H(\mu) \\
&= (3.11) - \varepsilon H(\mu) = (3.10) - \varepsilon H(\mu).
\end{aligned}$$

In this chain of equalities we used Lemma 3.32 below, the relation $\phi(x) = -\Psi(0,\cdot) + \varepsilon \log(\mu(x))$, and that $\int_{X\times X} \exp\left(\frac{\phi\oplus\psi}{\varepsilon}\right) \mathrm{d}K_\varepsilon = 1$. □

Lemma 3.32. In the setting of Definition 3.24 one has for any $t \in (0,1)$ that

$$\int_{\mathbb{R}^d} \partial_t(\Psi \cdot \rho)\,\mathrm{d}x = \int_{\mathbb{R}^d} \tfrac{1}{2}|\nabla\Psi|^2 \cdot \rho\,\mathrm{d}x. \tag{3.26}$$

Proof. The proof is a simple explicit computation, using that on $(0,1)\times\mathbb{R}^d$ Ψ and ρ are differentiable densities that solve (3.19) in a strong sense. One has

$$\begin{aligned}
\int_{\mathbb{R}^d} \partial_t(\Psi \cdot \rho)\,\mathrm{d}x &= \int_{\mathbb{R}^d} [\Psi \cdot (\partial_t\rho) + (\partial_t\Psi)\cdot\rho]\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} [\Psi \cdot (\tfrac{\varepsilon}{2}\Delta\rho - \nabla[\nabla\Psi\cdot\rho]) - (\tfrac{1}{2}|\nabla\Psi|^2 + \tfrac{\varepsilon}{2}\Delta\Psi)\cdot\rho]\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} [\tfrac{\varepsilon}{2}(\Delta\Psi)\cdot\rho + |\nabla\Psi|^2\cdot\rho]) - (\tfrac{1}{2}|\nabla\Psi|^2 + \tfrac{\varepsilon}{2}\Delta\Psi)\cdot\rho]\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \tfrac{1}{2}|\nabla\Psi|^2\cdot\rho\,\mathrm{d}x
\end{aligned}$$

where we used integration by parts in the third equality. □

Note that the expression $\int_{\mathbb{R}^d} \partial_t(\Psi\cdot\rho)\,\mathrm{d}x$ can also be related to the step-wise entropy decomposition in (3.21) where a finite-difference version of this temporal derivative appears.

Remark 3.33. Formally one has (3.25) $\geq$ (3.11) $- \varepsilon H(\mu)$.

Sketch of proof. Consider the same setting as in the sketch for Remark 3.31. Observe that by (3.19e) Ψ is formally admissible in (3.25) (but it will not be differentiable or even continuous at $t \in \{0,1\}$). Recall that $\phi(x) = -\Psi(0,\cdot) + \varepsilon\log(\mu(x))$ and $\int_{X\times X} \exp\left(\frac{\phi\oplus\psi}{\varepsilon}\right)\mathrm{d}K_\varepsilon = 1$. Then

$$(3.11) - \varepsilon H(\mu) = \int_X \Psi(1,\cdot)\,\mathrm{d}\nu - \int_X \Psi(0,\cdot)\,\mathrm{d}\mu \leq (3.25).$$ □

References

[1] Jason M. Altschuler and Enric Boix-Adsera. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *Math. Program.*, 199:1107–1178, 2023.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford mathematical monographs. Oxford University Press, 2000.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. Birkhäuser Boston, 2005.

[4] Sigurd Angenent, Steven Haker, and Allen Tannenbaum. Minimizing flows for the Monge–Kantorovich problem. *SIAM J. Math. Anal.*, 35(1):61–97, 2003.

[5] Fabrice Baudoin. *Diffusion processes and stochastic calculus*. EMS textbooks in mathematics. European Mathematical Society, 2014.

[6] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 1st edition, 2011.

[7] F. Beier, J. von Lindheim, S. Neumayer, and G. Steidl. Unbalanced multi-marginal optimal transport. *J. Math. Imaging Vis.*, 65:394–413, 2022.

[8] J-D. Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM Math. Model. Numer. Anal.*, 37(5):851–868, 2003.

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

[10] Jean-David Benamou, Guillaume Carlier, Simone Di Marino, and Luca Nenna. An entropy minimization approach to second-order variational mean-field games. *Mathematical Models and Methods in Applied Sciences*, pages 1–31, 2019.

[11] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numerische Mathematik*, 142(1):33–54, 2019.

[12] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[13] Tianji Cai, Junyi Cheng, Bernhard Schmitzer, and Matthew Thorpe. The linearized Hellinger–Kantorovich distance. *SIAM J. Imaging Sci.*, 15(1):45–83, 2022.

[14] Eric A. Carlen and Jan Maas. Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance. *Journal of Functional Analysis*, 273(5):1810–1869, 2017.

[15] G. Carlier, C. Jimenez, and F. Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM J. Control Optim*, 47(3):1330–1350, 2008.

[16] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

[17] Guillaume Carlier, Paul Pegon, and Luca Tamanini. Convergence rate of general entropic optimal transport costs. *Calc. Var. Partial Differential Equations*, 62:116, 2023.

[18] Y. Chen, T. T. Georgiou, and M. Pavon. On the relation between optimal transport and Schrödinger bridges: a stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169:671–691, 2016.

[19] Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: A quantum mechanical approach. *IEEE Transactions on Automatic Control*, 63(8):2612–2619, 2018.

[20] R. Chetrite, P. Muratore-Ginanneschi, and K. E. Schwieger. Schrödinger's 1931 paper "on the reversal of the laws of nature" ["Über die umkehrung der Naturgesetze", Sitzungsberichte der preussischen Akademie der Wissenschaften, physikalisch-mathematische Klasse, 8 n9 144–153]. *Eur. Phys. J. H*, 46:28, 2021.

[21] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[22] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Found. Comp. Math.*, 18(1):1–44, 2018.

[23] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

[24] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

[25] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 1999.

[26] 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 *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[27] Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Math.*, 177(2):113–161, 1996.

[28] Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583, 2019.

[29] Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. *Ann. Fac. Sci. Toulouse Math.*, 26(3):569–601, 2017.

[30] Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem. *SIAM Journal on Control and Optimization*, 59(4):2428–2453, 2021.

[31] Peter Koltai, Johannes von Lindheim, Sebastian Neumayer, and Gabriele Steidl. Transfer operators from optimal transport plans for coherent set detection. *Physica D*, 426:132980, 2021.

[32] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Adv. Differential Equations*, 21(11-12):1117–1164, 2016.

[33] Hugo Lavenant, Jonas Luckhardt, Gilles Mordant, Bernhard Schmitzer, and Luca Tamanini. The Riemannian geometry of Sinkhorn divergences. arXiv:2405.04987, 2024.

[34] Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *Ann. Appl. Probab.*, 34(1A), 2024.

[35] Christian Léonard. From the Schrödinger problem to the Monge–Kantorovich problem. *J. Funct. Anal.*, 262(4):1879–1920, 2012.

[36] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst. A*, 34(4):1533–1574, 2014.

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

[38] Stefano Lisini. Characterization of absolutely continuous curves in Wasserstein spaces. *Calc. Var. Partial Differential Equations*, 28(1):85–120, 2007.

[39] John Lott. Some geometric calculations on Wasserstein space. *Comm. Math. Phys.*, 277:423–437, 2008.

[40] Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via Frank–Wolfe algorithm. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

[41] Marco Mauritz, Bernhard Schmitzer, and Benedikt Wirth. A Bayesian model for dynamic mass reconstruction from PET listmode data. *SIAM J. Math. Anal.*, 56(5):5840–5880, 2024.

[42] R. J. McCann. Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.*, 11(3):589–608, 2001.

[43] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*, volume 11 of *Foundations and Trends in Machine Learning*. 2019.

[44] R. T. Rockafellar. Integrals which are convex functionals. *Pacific J. Math.*, 24(3):525–539, 1968.

[45] Ralph Tyrrell Rockafellar. Duality and stability in extremum problems involving convex functions. *Pacific J. Math*, 21(1):167–187, 1967.

[46] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Company, 3rd edition, 1986.

[47] Ludger Rüschendorf and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statistics and Prob. Letters*, 17:369–375, 1993.

[48] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser Boston, 2015.

[49] Bernhard Schmitzer, Klaus P. Schäfers, and Benedikt Wirth. Dynamic cell imaging in PET with optimal transport regularization. *IEEE Trans Med Imaging*, 39(5):1626–1635, 2020.

[50] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 2003.

[51] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.