OPTIMAL TRANSPORT

Contents

Why study optimal transport ? The main motivation studying optimal transport in statistics is the notion of Wasserstein distance between probability measures on a compact metric space X:

- The Wasserstein distances W_p represent faithfully the geometry of the underlying space: $x \in X \mapsto \delta_x \in \mathcal{P}(X)$ is an isometry. This means that unlike many notions of distances between functions/divergences between probability measures (E.g relative entropy),
- Application: inverse problems, Wasserstein GANs
- Application: statistics over the space of probability measures, e.g. geodesics barycenters, k-means, PCA...
- Application: PDE / particle systems

References. Introduction to optimal transport, with applications to PDE and/or calculus of variations can be found in books by Villani [26] and Santambrogio [18]. Villani's second book [27] concentrates on the application of optimal transport to geometric questions (e.g. synthetic definition of Ricci curvature), but its first chapters might be useful. We also mention Gigli, Ambrosio and Savaré [3] for the study of gradient flows with respect to the Monge-Kantorovich/Wasserstein metric.

Notation. In the following, we assume that X is a compact metric space, and we denote $\mathcal{C}^0(X)$ the space of continuous functions over X endowed with the norm of uniform convergence $\|\varphi\|_\infty = \sup_{x\in X} |\varphi(x)|$. We denote $\mathcal{M}(X)$ the space of Radon measures on X, which we identify with the continuous dual of $\mathcal{C}^0(X)$. We will denote $\langle \mu | \varphi \rangle = \int \varphi \mathrm{d}\mu$. We define

$$\mathcal{M}^+(X) := \{\mu \in \mathcal{M}(X) \mid \forall \varphi \in \mathcal{C}^0(X), \varphi \geqslant 0 \Longrightarrow \langle \mu | \varphi \rangle \geqslant 0\}$$

$$\mathcal{P}(X) := \{\mu \in \mathcal{M}^+(X) \mid \langle \mu | 1 \rangle = 1\}$$

The support of a measure μ is denoted $\mathrm{spt}(\mu)$.

The dual space is endowed with the total variation norm

$$\|\mu_n\|_{\mathrm{TV}} = \sup_{\varphi \in \mathcal{C}^0(X), \|\varphi\|_\infty \leqslant 1} \langle \mu | \varphi \rangle.$$

However, the topology that we will consider by default on $\mathcal{M}^0(X)$ is the weak$*$ topology. We recall for instance that a sequence $(\mu_n)_{n\geqslant 0}$ of measures converges weak$*$ to μ if and only if

$$\forall \varphi, \lim_{n\to+\infty} \langle \mu_n | \varphi \rangle = \langle \mu | \varphi \rangle.$$

We note that thanks to the Banach-Alaoglu theorem, any *bounded* sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathcal{M}(X)$ admits a weak$*$ converging subsequence. This applies in particular to any sequence in $\mathcal{P}(X)$: the space of probability measures is weak$*$ sequentially compact (and even compact).

1. The problems of Monge and Kantorovich

1.1. Monge's problem.

Definition 1 (Push-forward and transport map)**.** Let X, Y be compact metric spaces, $\mu \in \mathcal{M}(X)$ and let $T : X \to Y$ be a measurable map. The *push-forward* of μ by T is the measure $T_\# \mu$ on Y defined by

$$\forall B \subseteq Y,\ \mathrm{T}_\#\mu(B) = \mu(T^{-1}(B)).$$

or equivalently if the following change-of-variable formula holds for all test function $\varphi \in \mathcal{C}^0(Y)$:

$$\int_Y \varphi(y) \mathrm{d}\nu(y) = \int_X \varphi(T(x)) \mathrm{d}\mu(x).$$

A measurable map $T : X \to Y$ such that $T_\#\mu = \nu$ is also called a *transport map* between μ and ν.

Example 1. If $Y = \{y_1, \dots, y_n\}$, then $T_\#\mu = \sum_{1\leqslant i\leqslant n} \mu(T^{-1}(\{y_i\}))\delta_{y_i}$.

Definition 2 (Monge's problem)**.** Consider two metric spaces X, Y, two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a *cost function* $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$. *Monge's problem* is the following optimization problem

$$(\mathrm{MP}) := \inf \left\{ \int_X c(x, T(x)) \mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_{\#}\mu = \nu \right\} \tag{1.1}$$

This problem exhibits several difficulties, one of which is that both the constraint ($T_{\#}\mu = \nu$) and the functional are non-convex.

Example 2. There might exist no transport map between μ and ν. For instance, consider $\mu = \delta_x$ for some $x \in X$. Then, $\mathrm{T}_{\#}\mu(B) = \mu(T^{-1}(B)) = \delta_{T(x)}$. In particular, if ν is not a Dirac mass, then there exists no transport map between μ and ν.

1.2. **Kantorovich's problem.**

Definition 3 (Transport plan)**.** Let X, Y be two metric spaces and $\mu \in \mathcal{M}^+(X)$ and $\nu \in \mathcal{M}^+(Y)$ be two non-negative measures. A *transport plan* between μ and ν is a non-negative measure γ on the product space $X \times Y$ whose marginals are μ and ν. The set of transport plans is denoted

$$\Gamma(\mu, \nu) = \{\gamma \in \mathcal{M}_+(X \times Y) \mid \Pi_{X\#}\gamma = \mu,\ \Pi_{Y\#}\gamma = \nu\},$$

where $\Pi_X : X \times Y \to X$ and $\Pi_Y : X \times Y \to Y$ are the projection maps. Note that $\Gamma(\mu, \nu)$ is a convex set, and that it is non-empty if and only if μ and ν have the same total mass, i.e. $\mu(X) = \nu(Y)$.

Definition 4 (Kantorovich's problem)**.** Given two metric spaces X, Y, two non-negative measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a continuous *cost function* $c \in \mathcal{C}^0(X \times Y)$, *Kantorovich's problem* is the following optimization problem

$$(\mathrm{KP}) := \inf \{\langle c|\gamma\rangle \mid \gamma \in \Gamma(\mu, \nu)\} \tag{1.2}$$

We will denote $\mathcal{T}_c$ the associated *transport cost*

$$\begin{aligned} \mathcal{T}_c : \mathcal{M}^+(X) \times \mathcal{M}^+(Y) &\to \mathbb{R} \cup \{+\infty\} \\ (\mu, \nu) &\mapsto \inf\{\langle c|\gamma\rangle \mid \gamma \in \Gamma(\mu, \nu)\}. \end{aligned} \tag{1.3}$$

Note that by convention, the infimum over the empty set is $+\infty$, so that $\mathcal{T}_c(\mu, \nu) = +\infty$ if $\mu(X) \neq \nu(Y)$.

Remark 1. The infimum in Kantorovich's problem is less than the infimum in Monge's problem. Indeed, consider a transport map satisfying $T_{\#}\mu = \nu$ and the associated transport plan $\gamma_T = (\mathrm{id}, T)_{\#}\mu$. Then, by the change-of-variable formula one has

$$\langle c|\gamma_T\rangle \leqslant \int_{X\times Y} c(x, y) \mathrm{d}(id, T)_{\#}\mu(x, y) = \int_X c(x, T(x)) \mathrm{d}\mu,$$

thus proving the claim.

Example 3 (Finite support). Assume that $X = Y = \{1, \dots, N\}$ and that μ, ν are the uniform probability measures over X and Y. Then, Monge's problem can be rewritten as a minimization problem over the set of bijections between the two sets X and Y:

$$\min \left\{ \frac{1}{N} \sum_{1 \leqslant i \leqslant N} c(i, \sigma(i)) \mid \sigma \in \mathfrak{S}_N \right\}.$$

In Kantorovich's relaxation, the set of transport plans $\Gamma(\mu,\nu)$ agrees with the set of bi-stochastic matrices :

$$\gamma \in \Gamma(\mu,\nu) \Longleftrightarrow \gamma \geqslant 0, \sum_i \gamma(i,j) = 1/N = \sum_j \gamma(i,j).$$

By Birkhoff's theorem, any extremal bi-stochastic matrix is induced by a permutation. This shows that, in this case, the solution to Monge's and Kantorovich's problems agree.

Theorem 1 (Existence of solutions to (KP))**.** *Let X, Y be compact metric spaces and let $c \in \mathcal{C}^0(X \times Y)$. Then for any measures $(\mu,\nu) \in \mathcal{M}_+(X)\times\mathcal{M}_+(Y)$ with equal total mass, Kantorovich's problem* (KP) *admits a minimizer. Moreover, the transport cost $\mathcal{T}_c$ is a convex and weak$*$ lower semicontinuous functional on $\mathcal{M}_+(X) \times \mathcal{M}_+(Y)$.*

Proposition 2. *Let X, Y be compact metric spaces and let $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ be sequences of non-negative measures on X and Y with same total mass. Assume that these sequence weak$*$ converge to $\mu \in \mathcal{M}^+(X)$ and $\nu \in \mathcal{M}^+(Y)$ respectively. Then, any sequence of transport plans $\gamma_n \in \Gamma(\mu_n,\nu_n)$ admits a subsequence converging to some $\gamma \in \Gamma(\mu,\nu)$.*

In particular, the previous proposition implies that $\Gamma(\mu,\nu)$ is compact.

Proof. Since $\mu_n \geqslant 0$, one has $\|\mu_n\|_{\mathrm{TV}} = \langle \mu_n | 1 \rangle$, which converges to $\|\mu\|_{\mathrm{TV}}$ by weak$*$ convergence. Thus the sequence (μ_n) is bounded. Since

$$\|\gamma_n\|_{\mathrm{TV}} = \langle \gamma_n | 1 \rangle = \langle \Pi_{\#\gamma_n} | 1 \rangle = \langle \mu_n | 1 \rangle,$$

the sequence $(\gamma_n)_{n\in\mathbb{N}}$ is also bounded. By Banach-Alaoglu's theorem, it admits a weak$*$ converging subsequence. Relabeling if necessary, we therefore assume that γ_n converges weak$*$ to some $\gamma \in \mathcal{M}(X \times Y)$. Then,

$$\forall \varphi \in \mathcal{C}^0(X \times Y) \text{ s.t. } \varphi \geqslant 0, \langle \gamma | \varphi \rangle = \lim_{n\to+\infty} \langle \gamma_n | \varphi \rangle \geqslant 0$$

so that γ is a non-negative measures. Given $\varphi \in \mathcal{C}^0(X)$ and $\hat{\varphi}(x,y) := \varphi(x)$, using $\Pi_{X\#}\gamma_n = \mu_n$ we get $\langle \varphi | \mu_n \rangle = \langle \varphi | \Pi_{X\#}\gamma_n \rangle = \langle \hat{\varphi} | \gamma_n \rangle$. Taking the limit as $n \to +\infty$, we deduce that $\langle \varphi | \mu \rangle = \langle \hat{\varphi} | \gamma \rangle$ for all φ, implying that $\Pi_{X\#}\gamma = \mu$. Similarly, we prove that $\Pi_{Y\#}\gamma = \nu$, proving that $\gamma \in \Gamma(\mu,\nu)$. □

Proof of theorem 1. We first note that the function $\gamma \mapsto \langle c | \gamma \rangle$ is linear and continuous on $\mathcal{M}(X \times Y)$. Second, we note that if $\mu(X) = \nu(Y)$, the set $\Gamma(\mu,\nu)$ is non-empty as it contains a suitably rescaled product of μ and ν. The previous lemma shows that the set $\Gamma(\mu,\nu)$ is weak$*$ compact, so that $\langle c | \gamma \rangle$ attains its minimum on this set. This shows existence of at least one solution to (KP).

To prove that $\mathcal{T}_c$ is lower semicontinous, we consider converging sequences $(\mu_n), (\nu_n)$ in $\mathcal{M}_+(X)$ and $\mathcal{M}_+(Y)$ respectively. with weak$*$ limits μ and ν. Without loss of generality, we assume that μ_n and ν_n have the same total mass (if not, $\mathcal{T}_c(\mu_n,\nu_n) = +\infty$). For each n we consider $\gamma_n \in \Gamma(\mu_n,\nu_n)$ the optimal transport plan. Using the previous proposition, we assume taking a subsequence if necessary that γ_n converges to some $\gamma \in \Gamma(\mu,\nu)$. Then,

$$\mathcal{T}_c(\mu,\nu) \leqslant \langle c | \gamma \rangle = \lim_{n\to+\infty} \langle c | \gamma_n \rangle = \lim_{n\to+\infty} \mathcal{T}_c(\mu_n,\nu_n).$$

□

2. One-dimensional optimal transport

Definition 5 (Monotone set)**.** A subset S of $\mathbb{R} \times \mathbb{R}$ is called *monotone* if

$$\forall (x, y), (x', y') \in S, (x' - x) \cdot (y' - y) \geqslant 0.$$

Definition 6 (Submodular cost)**.** A cost function $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called *strictly submodular* if for every $x_0 < x_1$, the function $y \mapsto c(x_1, y) - c(x_0, y)$ is decreasing.

Theorem 3. *Let μ, ν be probability measures supported in $X = Y = [a, b] \subseteq \mathbb{R}$, and let c be a continuous and strictly submodular cost on $X \times Y$. Then, there exists a unique optimal transport plan $\gamma \in \Gamma(\mu, \nu)$, which is also the unique transport plan with monotone support.*

Proof. **Step 1.** We first establish that any optimal transport plan between μ and ν must be monotone. Consider a transport plan $\gamma \in \Gamma(\mu, \nu)$ and consider (x_0, y_0) and (x_1, y_1) in $\mathrm{spt}(\gamma)$. Since we want to prove that $(x_0 - x_1)(y_0 - y_1) \leqslant 0$, we may assume that $x_1 \neq x_0$ and $y_1 \neq y_0$. By continuity of the cost, for any $\delta > 0$ there exists $r > 0$ such that:

$$\mathrm{B}((x_0, y_0), r) \cap \mathrm{B}((x_1, y_1), r) \neq \emptyset$$

$$\forall a, b \in \{x_1, x_0, y_1, y_0\}, \forall (x, y) \in \mathrm{B}((a, b), r),\ |c(x, y) - c(a, b)| \leqslant \delta$$

Since (x_0, y_0) and (x_1, y_1) both belong to the support of γ, there must exist non-negative measures $\gamma_0 \leqslant \gamma$ and $\gamma_1 \leqslant \gamma$ with equal positive mass ε and such that $\mathrm{spt}(\gamma_i) \subseteq \mathrm{B}((x_i, y_i), r)$. Consider the marginals $\mu_i = \pi_{X\#}\gamma_i$ and $\nu_i = \pi_{Y\#}\gamma_i$, and take any coupling σ_0 (resp. σ_1) between μ_0 and ν_1 (resp. μ_1 and ν_0). Then, one can check that the measure

$$\sigma = \gamma - \gamma_0 - \gamma_1 + \sigma_0 + \sigma_1$$

is a transport plan between μ and ν (the non-negativity comes from $\gamma_i \leqslant \gamma$ and $\mathrm{spt}(\gamma_1) \cap \mathrm{spt}(\gamma_0) = \emptyset$). Using the optimality of γ one gets

$$\begin{aligned} 0 \leqslant F(\sigma) - F(\gamma) &= F(\sigma_0) - F(\gamma_0) + F(\sigma_1) - F(\gamma_1) \\ &= \int_{\mathrm{B}(x_0,r)\times\mathrm{B}(y_1,r)} c \mathrm{d}\sigma_0 + \int_{\mathrm{B}(x_1,r)\times\mathrm{B}(y_0,r)} c \mathrm{d}\sigma_1 \\ &\quad - \int_{\mathrm{B}(x_0,r)\times\mathrm{B}(y_0,r)} c \mathrm{d}\gamma_0 - \int_{\mathrm{B}(x_1,r)\times\mathrm{B}(y_1,r)} c \mathrm{d}\gamma_1 \\ &\leqslant \varepsilon \cdot (c(x_0, y_1) + c(x_1, y_0) - c(x_0, y_0) - c(x_1, y_1) + 4\delta) \end{aligned}$$

Since this holds for all $\delta > 0$ small enough, we deduce that

$$c(x_0, y_0) + c(x_1, y_1) \leqslant c(x_0, y_1) + c(x_1, y_0).$$

Assume without loss of generality that $x_0 < x_1$. Then,

$$c(x_1, y_1) - c(x_0, y_1) \leqslant c(x_1, y_0) - c(x_1, y_0),$$

thus implying by submodularity (the function $y \mapsto c(x_1, y_1) - c(x_0, y_1)$ is decreasing) that $y_0 \leqslant y_1$.

Step 2. We show that there exists at most one monotone transport plan between μ and ν. Recall that a probability measure γ on $\mathbb{R}^2$ is uniquely

defined from the values $\gamma((-\infty,a]\times(-\infty,b])$ for any $a,b\in\mathbb{R}$. This follows from the fact that such sets generate the Borel σ-algebra. Consider $A=(-\infty,a]\times(b,+\infty)$ and $B=(a,+\infty)\times(-\infty,b]$. Then, by monotonicity of $\mathrm{spt}(\gamma)$ one cannot have $\gamma(A)>0$ and $\gamma(B)>0$ at the same time. Hence,

$$\begin{aligned}\gamma((-\infty,a]\times(-\infty,b]) &= \min(\gamma(((-\infty,a]\times(-\infty,b])\cup A),\\ &\qquad \gamma(((-\infty,a]\times(-\infty,b])\cup B))\\ &= \min(\mu((-\infty,a]),\nu((-\infty,b])).\end{aligned}$$

This shows that $\gamma((-\infty,a]\times((-\infty,b])$ is uniquely defined from μ,ν, so that γ is unique. □

2.1. **Quantile function and one-dimensional Wasserstein spaces.**

Definition 7 (Cdf and quantile function)**.** Let μ be a probability measure on $\mathbb{R}$. The *cumulative distribution function* $F_\mu:\mathbb{R}\to[0,1]$ and the *inverse cumulative distribution function* $T_\mu:[0,1]\to\mathbb{R}$ are defined by:

$$F_\mu(x)=\mu((-\infty,x]) \qquad T_\mu(m)=\inf\left\{x\in\mathbb{R}\mid F_\mu(x)\geqslant m\right\}.$$

The function T_μ will also be called the *quantile function.*

In the following, we assume that X is a segment of $\mathbb{R}$.

Definition 8 (Wasserstein distance)**.** The *Wasserstein distance* of exponent $p\geqslant 1$ between two probability measures $\mu,\nu\in\mathcal{P}(X)$ is defined by

$$\mathrm{W}_p^p(\mu,\nu)=\min_{\gamma\in\Gamma(\mu,\nu)}\int\|x-y\|^p\,\mathrm{d}\gamma(x,y).$$

Proposition 4 (Quantile functions and Wasserstein distance)**.** *Let μ,ν be two probability measure on a segment $X\subseteq\mathbb{R}$. Then,*

(i) T_μ is a transport map between the Lebesgue measure $\lambda_{[0,1]}$ and μ
(ii) $\gamma_{\mu\to\nu}=(T_\mu,T_\nu)_\#\lambda_{[0,1]}$ is the unique monotone transport plan between μ and ν;
(iii) for all $p\geqslant 1$, $\mathrm{W}_p(\mu,\nu)=\|T_\mu-T_\nu\|_{\mathrm{L}^p([0,1])}$.

Example 4 (Translation). If ν is obtained by translating ν by a constant $v\in\mathbb{R}$, then $T_\nu=T_\mu+v$ so that $\mathrm{W}_p(\mu,\nu)=\|T_\mu-T_\nu\|_{\mathrm{L}^p([0,1])}=|v|$.

Example 5 (Discrete measures). If $\mu=\frac{1}{N}\sum_{i=1}^N\delta_{x_i}$ and the sequence $(x_i)_{1\leqslant i\leqslant N}$ is increasing, then the quantile function satisfies

$$T_\mu|_{\left[\frac{i-1}{n},\frac{i}{n}\right]}=x_i.$$

In particular, if $\nu=\frac{1}{N}\sum_{i=1}^N\delta_{y_i}$, where the sequence $y_{1\leqslant i\leqslant N}$ is also increasing,

$$\mathrm{W}_p(\mu,\nu)^p=\frac{1}{N}\sum_i\|x_i-y_i\|^p.$$

Proof. (i) Let $\hat{\mu} = T_{\mu\#}\lambda_{[0,1]}$. Then,

$$\begin{aligned} F_{\hat{\mu}}(x) &= \hat{\mu}((-\infty, x]) \\ &= \lambda(T_\mu^{-1}(-\infty, x]) \\ &= \lambda(\{m \in [0,1], T_\mu(m) \leqslant x\} \\ &= \lambda(\{m \in [0,1], F_\mu(x) \geqslant m\}) \\ &= F_\mu(x). \end{aligned}$$

were we used the equivalence $T_\mu(m) \leqslant x$ iff $F_\mu(x) \geqslant m$. This shows that $\hat{\mu} = \mu$.
(ii) Denote $\gamma := \gamma_{\mu\to\nu}$. We note first that $\Pi_{X\#}\gamma = \Pi_X \circ (T_\mu, T_\nu)_\#\lambda_{[0,1]} = \mu$, and similarly $\Pi_{Y\#}\gamma = \nu$. Thus, γ is a transport plan between μ and ν. In addition, γ is supported on the set $S := \{(T_\mu(m), T_\nu(m)) \mid m \in [0,1]\}$. Given two couples $(x_i, y_i) \in S$, there exists $m_i \in [0,1]$ such that $x_i = T_\mu(m_i)$ and $y_i = T_\nu(m_i)$. Without loss of generality, assume that $m_0 \leqslant m_1$. Then, $T_\mu(m_0) \leqslant T_\mu(m_1)$ and $T_\nu(m_0) \leqslant T_\nu(m_1)$ so that

$$(x_1 - x_0)(y_1 - y_0) \geqslant 0,$$

implying that S is monotone.
(iii) Theorem 3 proves that a solution to the optimal transport problem is given between μ and ν for the convex cost $c(x,y) = \|x-y\|^p$ is given by the monotone plan, i.e.

$$\begin{aligned} \min_{\gamma\in\Gamma(\mu,\nu)} \int \|x-y\|^p \,\mathrm{d}\gamma(x,y) &= \int \|x-y\|^p \,\mathrm{d}\gamma_{\mu\to\nu}(x,y) \\ &= \int \|x-y\|^p \,\mathrm{d}(T_\mu, T_\nu)_{\#\lambda_{[0,1]}}(x,y) \\ &= \int_0^1 \|T_\mu(m) - T_\nu(m)\|^p \,\mathrm{d}m \\ &= \|T_\mu - T_\nu\|^p_{\mathrm{L}^p([0,1])} \end{aligned}$$

□

Proposition 5 (Properties of the 1D Wasserstein spaces)**.** *The following properties hold for any segment $X \subseteq \mathbb{R}$ and any $p \geqslant 1$:*

(i) W_p *is a distance on* $\mathcal{P}(X)$
(ii) W_p *metrizes weak* convergence on* $\mathcal{P}(X)$*, i.e. for any sequence* (μ_n) *in* $\mathcal{P}(X)$ *and any* $\mu \in \mathcal{P}(X)$*,*

$$\lim_{n\to+\infty} \mathrm{W}_p(\mu_n, \mu) = 0 \iff \forall \varphi \in \mathcal{C}^0(X), \lim_{n\to+\infty} \langle \mu_n | \varphi \rangle = \langle \mu | \varphi \rangle.$$

(iii) the application $\mu \mapsto T_\mu$ *mapping a probability measure to its inverse cdf is an isometric embedding of* $(\mathcal{P}(X), \mathrm{W}_p(X))$ *into* $\mathrm{L}^p([0,1])$.

Proof. (i) We note that $\mathrm{W}_p(\mu,\nu) = 0$ implies that $T_\mu = T_\nu$ a.e., so that $\mu = T_{\mu\#}\lambda_{[0,1]} = T_{\nu\#}\lambda_{[0,1]} = nu$. The symmetry is immediate, and the triangle inequality for W_p follows from the triangle inequality in $\mathrm{L}^p([0,1])$.

(ii) Assume first that $\mathrm{W}_p(\mu_n, \mu) = \|T_{\mu_n} - T_\mu\|_{\mathrm{L}^p([0,1])}$ converges to zero as $n \to +\infty$. Then, $\|T_{\mu_n} - T_\mu\|_{\mathrm{L}^1([0,1])}$ also converges to zero as $n \to +\infty$. Let

$f : X \to \mathbb{R}$ be L-Lipschitz. Then,

$$\begin{aligned}|\langle f|\mu_n - \mu\rangle| &= \left|\int_0^1 f(T_{\mu_n}(m)) - f(T_\mu(m))\mathrm{d}m\right| \\ &\leqslant L\int_0^1 \|T_{\mu_n}(m) - T_\mu(m)\|\,\mathrm{d}m \\ &= L\,\mathrm{W}_1(\mu_n,\mu) \xrightarrow{n\to+\infty} 0\end{aligned}$$

Since continuous functions on X can be uniformly approximated by Lipschitz functions, we get weak$*$ convergence.

Conversely, assume that μ_n converges weakly to μ. The non-decreasing map T_μ is continuous on $[0,1]\setminus Z$, where Z is at most countable. It is standard that for any $x \notin Z$, $T_{\mu_n}(x)$ converges to $T_\mu(x)$ as $n \to +\infty$, i.e. T_{μ_n} converges a.e. to T_μ. Since in addition T_{μ_n} is bounded, we deduce that convergence holds in $\mathrm{L}^p([0,1])$ for any $p \geqslant 1$. □

Definition 9 (Geodesic)**.** Let (E,d) be a metric space. A constant speed geodesic between two points $x_0, x_1 \in E$ is a continuous curve $x : [0,1] \to E$ such that for every $s,t \in [0,1]$, $d(x_s,x_t) = |s-t|\,d(x_0,x_1)$.

Proposition 6. *Let X be a segment of $\mathbb{R}$ and let $\mu_0,\mu_1 \in \mathcal{P}(X)$. Define*

$$\mu_t := T_{t\#}\lambda_{[0,1]},\ \textit{where } T_t = (1-t)T_{\mu_0} + tT_{\mu_1}$$

Then, the curve μ_t is a constant speed geodesic between μ_0 and μ_1 in the space $(\mathcal{P}(X),\mathrm{W}_p)$, for any exponent $p \geqslant 1$. In particular, this space is a geodesic space, meaning that any $\mu_0,\mu_1 \in \mathcal{P}_p(X)$ can be joined by (at least one) constant speed geodesic.

Proof. First note that if $0 \leqslant s \leqslant t \leqslant 1$,

$$\mathrm{W}_p(\mu_0,\mu_1) \leqslant \mathrm{W}_p(\mu_0,\mu_s) + \mathrm{W}_p(\mu_s,\mu_t) + \mathrm{W}_p(\mu_t,\mu_1),$$

so that it suffices to prove the inequality $\mathrm{W}_p(\mu_s,\mu_t) \leqslant |t-s|\,\mathrm{W}_p(\mu_0,\mu_1)$ for all $0 \leqslant s \leqslant t \leqslant 1$ to get equality. The inequality is easily checked by taking $\gamma_{st} := (T_s,T_t)_\#\lambda_{[0,1]} \in \Gamma(\mu_s,\mu_t)$, so that

$$\begin{aligned}\mathrm{W}_p(\mu_s,\mu_t)^p &\leqslant \int \|T_s(m) - T_t(m)\|^p\,\mathrm{d}m \\ &= \int \|(1-s)T_0(m) + sT_1(m) - ((1-t)T_0(m) + tT_1(m))\|^p\,\mathrm{d}m \\ &= \int \|(t-s)(T_0(m) - T_t(m))\|^p\,\mathrm{d}m = (t-s)^p\,\mathrm{W}_p(\mu,\nu)^p \qquad \square\end{aligned}$$

Remark 2 (Barycenters). We can also consider barycenters in the Wasserstein, at least in the case $p = 2$ and on a segment X. The weighted barycenter of probability measures $\mu_0,\dots,\mu_k \in \mathcal{P}(X)$ with weights $\alpha_1,\dots,\alpha_k > 0$ is the unique minimizer of

$$\min_{\mu\in\mathcal{P}(X)} \sum_{1\leqslant i\leqslant k} \alpha_k\,\mathrm{W}_2^2(\mu_k,\mu).$$

The quantile function of the barycenter μ therefore solves the following minimization problem

$$T_\mu \in \arg\min_T \sum_{1\leqslant i\leqslant k} \alpha_k \left\|T_{\mu_k} - T\right\|^2_{\mathrm{L}^2([0,1])},$$

so that T_μ is simply a weighted average of the T_{μ_k}:

$$T_\mu = \frac{1}{\sum_k \alpha_k} \sum_{1\leqslant i\leqslant k} \alpha_k T_{\mu_k}.$$

The barycenter is finally recovered thanks to the formula $\mu = T_{\mu\#}\lambda_{[0,1]}$, i.e.

$$\mu = \left(\frac{1}{\sum_k \alpha_k} \sum_{1\leqslant i\leqslant k} \alpha_k T_{\mu_k}\right)_{\#} \lambda_{[0,1]}.$$

3. Kantorovich duality

3.1. **Derivation of the dual problem.** The primal Kantorovich problem (KP) can be reformulated by introducing Lagrange multipliers for the constraints. Namely, we use that for any $\gamma \in \mathcal{M}^+(X\times Y)$,

$$\sup_{\varphi\in\mathcal{C}^0(X)} -\langle \varphi\otimes 1|\gamma\rangle + \langle\varphi|\mu\rangle = \begin{cases} 0 & \text{if } \Pi_{X\#}\gamma = \mu \\ +\infty & \text{if not} \end{cases}$$

$$\sup_{\varphi\in\mathcal{C}^0(X)} -\langle 1\otimes \psi|\gamma\rangle + \langle\psi|\mu\rangle = \begin{cases} 0 & \text{if } \Pi_{X\#}\gamma = \mu \\ +\infty & \text{if not} \end{cases}$$

to deduce that for any $\gamma \in \mathcal{M}_+(X\times Y)$,

$$\sup_{\varphi\in\mathcal{C}^0(X),\psi\in\mathcal{C}^0(Y)} \langle\varphi|\mu\rangle + \langle\psi|\nu\rangle - \langle\varphi\oplus\psi|\gamma\rangle = \begin{cases} 0 & \text{if } \gamma\in\Gamma(\mu,\nu) \\ +\infty & \text{if not.} \end{cases}$$

This leads to the following formulation of the Kantorovich problem

$$(\mathrm{KP}) = \inf_{\gamma\in\mathcal{M}^+(X\times Y)} \sup_{(\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y)} \langle c - (\varphi\oplus\psi)|\gamma\rangle + \langle\varphi|\mu\rangle - \langle\psi|\nu\rangle$$

Kantorovich dual problem is simply obtained by inverting the infimum and the supremum:

$$(\mathrm{KD}) := \sup_{\varphi,\psi} \inf_{\gamma\geqslant 0} \langle c - (\varphi\oplus\psi)|\gamma\rangle + \langle\varphi|\mu\rangle - \langle\psi|\nu\rangle.$$

Note that we will often omit the assumptions that $\gamma\in\mathcal{M}(X\times Y)$ and φ,ψ are continuous, when the context is clear. The dual problem can further be simplified by remarking that

$$\inf_{\gamma\geqslant 0} \langle c - \varphi\oplus\psi|\gamma\rangle = \begin{cases} 0 & \text{if } \varphi\oplus\psi\leqslant c \\ -\infty & \text{if not.} \end{cases}$$

Definition 10 (Kantorovich's dual problem)**.** Given $\mu\in\mathcal{P}(X)$ and $\nu\in\mathcal{P}(Y)$ with X,Y compact metric spaces and $c\in\mathcal{C}^0(X\times Y)$, we define Kantorovich's dual problem by

$$(\mathrm{KD}) = \sup\left\{\int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu \mid (\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y), \varphi\oplus\psi\leqslant c\right\} \tag{3.4}$$

Proposition 7. *Weak duality holds, i.e.* (KP) $\geqslant$ (KD).

Proof. Given $(\varphi, \psi, \gamma) \in \mathcal{C}^0(X) \times \mathcal{C}^0(Y) \times \Gamma(\mu, \nu)$ satisfying the constraint $\varphi \oplus \psi \leqslant c$, one has

$$\langle \varphi | \mu \rangle + \langle \psi | \nu \rangle = \langle \varphi \oplus \psi | \gamma \rangle \leqslant \langle c | \gamma \rangle,$$

where we used $\gamma \in \Gamma(\mu, \nu)$ to get the equality and $\varphi \oplus \psi \leqslant c$ to get the inequality. As a conclusion,

$$(\mathrm{KD}) = \sup_{\varphi \oplus \psi \leqslant c} \langle \varphi | \mu \rangle - \langle \psi | \nu \rangle \leqslant \min_{\gamma \in \Gamma(\mu,\nu)} \langle c | \gamma \rangle = (\mathrm{KP}) \qquad \square$$

Remark 3. As often, the Lagrange multipliers (or Kantorovich potentials) φ, ψ have an economic interpretation as prices. For instance, imagine that μ is the distribution of sand available at quarries, and ν describes the amount of sand required by construction work. Then, (KP) can be interpreted as finding the cheapest way of transporting the sand from μ to ν for a construction company. Imagine that this company wants to externalize the transport, by paying a loading coast $\varphi(x)$ at a point x (in a quarry) and an unloading coast $\psi(y)$ at a point y (at a construction place). Then, the constraint $\varphi(x)+\psi(y) \leqslant c(x, y)$ translates the fact that the construction company would not externalize if its cost is higher than the cost of transporting the sand by itself. Then, Kantorovich's dual problem (KD) describes the problem of a transporting company: maximizing its revenue $\int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$ under the constraint $\varphi \oplus \psi \leqslant c$ imposed by the construction company. The economic interpretation of the strong duality (KP) = (KD) is that in this setting, externalization has exactly the same cost as doing the transport by oneself.

The questions that we will address now are the following:

- When does strong duality ((KP) = (KD)) hold ?
- When is the supremum in Kantorovich's dual problem attained ?
- What does Kantorovich's duality imply about Monge's problem, stability of optimal transport maps/plans, numerics, etc ?

3.2. **Strong duality.** We prove strong duality using a strategy recently proposed by Savaré and Sodini [19], which relies only the Fenchel-Moreau theorem from convex analysis. In addition to the transport cost functional,

$$\mathcal{T}_c : \mathcal{M}(X) \times \mathcal{M}(Y) \to \mathbb{R} \cup \{+\infty\}$$
$$(\mu, \nu) \mapsto \begin{cases} \inf\{\langle c | \gamma \rangle \mid \gamma \in \Gamma(\mu, \nu)\} & \text{if } \mu \geqslant 0, \nu \geqslant 0, \text{ and } \mu(X) = \nu(Y) \\ +\infty & \text{otherwise} \end{cases} \tag{3.5}$$

we will consider the following, non-convex and very singular functional, which encodes the cost of transport between Dirac masses with the same weight:

$$F_c : \mathcal{M}(X) \times \mathcal{M}(Y) \to \mathbb{R} \cup \{+\infty\}$$
$$(\mu, \nu) \mapsto \begin{cases} mc(x, y) & \text{if } \mu = m\delta_x, \nu = m\delta_y \text{ and } m \geqslant 0 \\ +\infty & \text{otherwise} \end{cases} \tag{3.6}$$

Theorem 8 (Savaré and Sodini)**.** $\mathcal{T}_c = F_c^{**}$

Corollary 9 (Strong duality in Kantorovich's problem)**.** (KP) = (KD).

The proof ot these results rely on the Fenchel-Moreau theorem from convex analysis. To state this theorem, we need to define the convex and convex biconjugate of a function on a topological vector space.

Definition 11 (Convex conjugate)**.** Let E be a topological vector space. The *convex conjugate* of a function $F : E \to \mathbb{R} \cup \{+\infty\}$ is the function F^* on the dual space E^* defined by

$$F^*(x^*) = \sup_{x \in E} \langle x^* | x \rangle - F(x).$$

The *biconjugate* of F is then defined as $F^{**} : E \to \mathbb{R} \cup \{+\infty\}$ by

$$F^{**}(x) = \sup_{x^* \in E^*} \langle x^* | x \rangle - F^*(x^*).$$

It is quite easy to see that F^* and F^{**} are convex and lower semicontinuous, as suprema of continuous affine functions. Fenchel-Moreau's theorem show that F^{**} is in fact the *lower semicontinuous convex envelope* of F, i.e. the largest lsc convex function that lies below F.

Theorem 10 (Fenchel-Moreau)**.** *Let E be a locally convex and separated topological vector space and let $F : E \to \mathbb{R} \cup \{+\infty\}$. Then F^{**} is the lsc convex envelope of F, i.e. the largest lsc convex function that lies below F. In particular, $F = F^{**}$ if and only if F is convex and lower semicontinuous.*

Proof. Let G be the lsc convex envelope of F. We first prove that $F^{**} \leqslant G$. Given any point $x \in E$, the definition of F^* as a supremum gives $F^*(x^*) \geqslant \langle x^* | x \rangle - F(x)$. Thus,

$$F^{**}(x) = \sup_{x^* \in E^*} \langle x^* | x \rangle - F^*(x^*) \leqslant \sup_{x^* \in E^*} \langle x^* | x \rangle - (\langle x^* | x \rangle - F(x)) = F(x).$$

This shows that the lsc convex function F^{**} lies below F, so that F^{**} lies below the lsc convex envelope of F.

To prove that $F^{**} \geqslant G$, we use the following representation of G as the maximum of continuous affine functions that lie below F:

$$G(x) = \sup \left\{ \langle x^* | x \rangle + \alpha \mid (x^*, \alpha) \in X^* \times \mathbb{R} \text{ s.t. } \langle x^* | \cdot \rangle + \alpha \leqslant F \right\}.$$

We now choose some affine function defined by $(x^*, \alpha) \in E^* \times \mathbb{R}$ and lying below F, i.e. such that $F \geqslant \langle x^* | \cdot \rangle + \alpha$. Then,

$$F^*(x^*) \leqslant \sup_{x \in X} \langle x^* | x \rangle - F(x) \leqslant \sup_{x \in X} \langle x^* | x \rangle - (\langle x^* | x \rangle + \alpha) = -\alpha.$$

This implies that $F^{**}(x) \geqslant \langle x^* | x \rangle - F^*(x) \geqslant \langle x^* | x \rangle + \alpha$. In other words, F^{**} is larger than any affine function that lies below F, i.e. $F^{**} \geqslant G$. □

Proof of Theorem 8. We need to compute the convex conjugate and biconjucate of the functional F_c. This functional is defined on the space $\mathcal{M}(X) \times \mathcal{M}(Y)$ endowed with the product of the weak$*$-topologies, making it a locally convex and separated topological vector space. By definition of the weak$*$

topology, $\mathcal{M}(X)^* = \mathcal{C}^0(X)$, so that we may identify $(\mathcal{M}(X)\times\mathcal{M}(X))^*$ with $\mathcal{C}^0(X)\times\mathcal{C}^0(Y)$. We have

$$\begin{aligned} F_c^*(\varphi,\psi) &= \sup_{\mu,\nu} \langle\mu|\varphi\rangle + \langle\nu|\psi\rangle - F(\mu,\nu) \\ &= \sup_{x,y\in X, m\geqslant 0} m(\langle\delta_x|\varphi\rangle + \langle\delta_y|\psi\rangle - c(x,y)) \\ &= \sup_{x,y\in X, m\geqslant 0} m(\varphi(x)+\psi(y)-c(x,y)) \\ &= \begin{cases} 0 & \text{if } \varphi\oplus\psi\leqslant c \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

Therefore, the biconjugate of F_c is given by

$$\begin{aligned} F_c^{**}(\mu,\nu) &= \sup_{\varphi,\psi} \langle\mu|\varphi\rangle + \langle\nu|\psi\rangle - F_c^*(\varphi,\psi) \\ &= \sup_{\varphi\oplus\psi\leqslant c} \langle\mu|\varphi\rangle + \langle\nu|\psi\rangle = (\mathrm{KD}). \end{aligned}$$

Recall that F_c^{**} is the largest lsc convex function that lie below F_c. Since $\mathcal{T}_c$ is lsc convex and also lies below F_c, we deduce that $(\mathrm{KD}) = F_c^{**} \geqslant \mathcal{T}_c = (\mathrm{KP})$. Since we already know (by weak duality) that $(\mathrm{KP}) \geqslant (\mathrm{KD})$, we deduce strong duality ($(\mathrm{KP}) = (\mathrm{KD})$) and $F_c^{**} = \mathcal{T}_c$. □

3.3. **Existence of solution for the dual problem.** Kantorovich's dual problem (KD) consists in maximizing a concave (actually linear) functional under linear inequality constraints. It can also also easily be turned into an unconstrained minimization problem. The idea is quite simple: given a certain $\psi \in \mathcal{C}^0(Y)$, one wishes to select φ on X which is as large as possible (to maximize the term $\langle\varphi|\mu\rangle$ in (KD)) while satisfying the constraint $\varphi\oplus\psi\leqslant c$. This constraint can be rewritten as

$$\forall x\in X,\ \varphi(x) \leqslant \min_{y\in Y} c(x,y) - \psi(y).$$

The largest function φ satisfying it is $\varphi(x) = \min_{y\in Y} c(x,y) - \psi(y)$. Thus,

$$\begin{aligned} (\mathrm{KP}) &= \sup_{\varphi\oplus\psi\leqslant c} \langle\varphi|\mu\rangle + \langle\psi|\nu\rangle \\ &= \sup_{\psi\in\mathcal{C}^0(Y)} \int_X \left(\min_{y\in Y} c(x,y)-\psi(y)\right) \mathrm{d}\mu(x) + \int \psi(y)\mathrm{d}\nu(y). \end{aligned}$$

This idea is at the basis of many algorithms to solve discrete instances of optimal transport, but also useful in theory. It also suggests to introduce the notion of c-transform.

Definition 12 (c-Transform, c-Concavity)**.** The c-transform (resp. $\bar{c}$-transform) of a function $\psi : Y\to\mathbb{R}\cup\{+\infty\}$ (resp. $\varphi : X\to\mathbb{R}\cup\{+\infty\}$) is

$$\psi^c : x\in X \mapsto \min_{y\in Y} c(x,y) - \psi(y) \tag{3.7}$$

$$\varphi^{\bar{c}} : y\in Y \mapsto \min_{x\in X} c(x,y) - \varphi(x) \tag{3.8}$$

A function φ on X is called *c-concave* if $\varphi = \psi^c$ for some $\psi\in\mathcal{C}^0(Y)$. Similarly, a function ψ on Y is called *$\bar{c}$-concave* if $\psi = \varphi^{\bar{c}}$ for some $\varphi\in\mathcal{C}^0(X)$.

Thanks to this notion of c-transform, one can reformulate the dual problem (KD) as an unconstrained maximization problem:

$$(\mathrm{KD}) = \sup_{\psi \in \mathcal{C}^0(Y)} \int_X \psi^c \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu. \tag{3.9}$$

Theorem 11 (Existence of dual potentials)**.** *The dual Kantorovich problem* (KD) *admits a maximizer. Moreover, for any $x_0 \in X$ there exists a maximizer of the form (φ, ψ), such that $\varphi = \psi^c$ and $\psi = \varphi^{\overline{c}}$, and satisfying $\varphi(x_0) = 0$.*

The existence of maximizers follows from the fact that a c-concave/$\overline{c}$-convex function has the same modulus of continuity as c.

Definition 13 (Modulus of continuity)**.** A real-valued function f on a metric space (Z, d_Z) has modulus of continuity $\omega : \mathbb{R}^+ \to \mathbb{R}$ if ω satisfies $\lim_{t\to 0} \omega(t) = 0$ and if for every $z, z' \in Z$, $|f(z) - f(z')| \leqslant \omega(\mathrm{d}_Z(z, z'))$.

Lemma 12 (Properties of c-transforms)**.** *Let $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ be a modulus of continuity for $c \in \mathcal{C}^0(X \times Y)$ for the distance*

$$\mathrm{d}_{X\times Y}((x, y), (x', y')) = \mathrm{d}_X(x, x') + \mathrm{d}_Y(y, y').$$

Then for every $\varphi \in \mathcal{C}^0(X)$ and every $\psi \in \mathcal{C}^0(Y)$,

- *$\varphi^{\overline{c}}$ and ψ^c also admits ω as modulus of continuity.*
- *$\psi^{c\overline{c}} \geqslant \psi$ and $\psi^{c\overline{c}c} = \psi^c$.*
- *$\varphi^{\overline{c}c} \geqslant \varphi$ and $\varphi^{\overline{c}c\overline{c}} = \varphi^{\overline{c}}$.*

Proof. (i) Let $\psi \in \mathcal{C}^0(Y)$ and let x be a point in X. By compactness, there exists a point y_x in Y realizing the minimum in the definition of ψ^c. Then, for every $x' \in X$,

$$\begin{aligned}
\psi^c(x') &= \min_{y\in Y} c(x', y) - \psi(y) \\
&\leqslant c(x', y_x) - \psi(y_x) = \psi^c(x) + c(x', y_x) - c(x, y_x) \\
&\leqslant \psi^c(x) + \omega(\mathrm{d}_X(x, x')).
\end{aligned}$$

Exchanging the role of x and x' we get $|\psi^c(x') - \psi^c(x)| \leqslant \omega(\mathrm{d}_X(x, x'))$ as desired. The proof that $\varphi^{\overline{c}}$ has the ω as modulus of continuity is similar.
(ii) By definition, of the c and $\overline{c}$-transforms, one has

$$\psi^{c\overline{c}}(y) = \min_{x\in X} \left(c(x, y) - \min_{\tilde{y}\in Y} c(x, \tilde{y}) - \psi(\tilde{y}) \right).$$

Taking $\tilde{y} = y$, one gets $\psi^{c\overline{c}}(y) \geqslant \psi(y)$. Again, by definition, we have

$$\psi^{c\overline{c}c}(x) = \min_{y\in Y} \left(c(x, y) - \min_{\tilde{x}\in X} \left(c(\tilde{x}, y) - \min_{\tilde{y}\in Y} c(\tilde{x}, \tilde{y}) - \psi(\tilde{y}) \right) \right).$$

By taking $\tilde{x} = x$, one gets $\psi^{c\overline{c}c}(x) \geqslant \psi^c(x)$, while taking $\tilde{y} = y$ gives us $\psi^{c\overline{c}c}(x) \leqslant \psi^c(x)$. The claim (iii) is proven similarly. □

Proof of Theorem 11. Let $(\varphi_n, \psi_n)_{n\in\mathbb{N}}$ be a maximizing sequence for (KD), i.e. $\varphi_n \oplus \psi_n \leqslant c$ and $\lim_{n\to+\infty} \langle \varphi_n | \mu \rangle + \langle \psi_n | \nu \rangle = (\mathrm{KD})$. Define $\hat{\varphi}_n = \psi_n^c$ and $\hat{\psi}_n = \hat{\varphi_n}^{\overline{c}}$. Then $\hat{\varphi}_n \oplus \hat{\psi}_n \leqslant c$, $\varphi_n \leqslant \hat{\varphi}_n$ and $\psi_n \leqslant \hat{\psi}_n$, which implies

$$\langle \varphi_n | \mu \rangle + \langle \psi_n | \nu \rangle \leqslant \langle \hat{\varphi}_n | \mu \rangle + \langle \hat{\psi}_n | \nu \rangle.$$

Thus, the sequence $(\hat{\varphi}_n, \hat{\psi}_n)_{n\in\mathbb{N}}$ is also a maximizing sequence for (KD). We note at this point that it is possible to assume that $\hat{\varphi}_n(x_0) = 0$ for all n, where x_0 is a given point in X. Indeed, if this is not the case, we may replace the original sequence $(\hat{\varphi}_n, \hat{\psi}_n)_{n\in\mathbb{N}}$ by $(\hat{\varphi}_n - \hat{\varphi}_n(x_0), \hat{\psi}_n + \hat{\varphi}_n(x_0))_{n\in\mathbb{N}}$, which is also admissible and has the same dual value.

We now prove that the sequence $(\hat{\varphi}_n, \hat{\psi}_n)$ admits a converging subsequence. By Lemma 12, the sequences $(\hat{\varphi}_n)_n$ and $(\hat{\psi}_n)_n$ are equicontinuous. Since $\hat{\varphi}_n(x_0) = 0$, we deduce from uniform continuity that the sequence $(\hat{\varphi}_n)_{n\in\mathbb{N}}$ is uniformly bounded. Then, using

$$\hat{\psi}_n(y) = \hat{\varphi}_n^{\overline{c}}(y) = \max_{x\in X} c(x,y) - \hat{\varphi}_n(x),$$

we deduce that $\|\hat{\psi}_n\|_\infty \leqslant \|c\|_\infty + \|\hat{\varphi}_n\|_\infty$ so that $(\hat{\varphi}_n)_{n\in\mathbb{N}}$ is also uniformly bounded. By Arzelà-Ascoli's theorem, both sequences therefore admit converging subsequences. The limit potentials are then maximizers for (KD) because the functional which is maximized in (KD) is continuous. □

3.4. **Stability of optimal transport plans.**

Proposition 13 (Support of OT plans)**.** *Let $(\varphi, \psi) \in \mathcal{C}^0(X) \times \mathcal{C}^0(Y)$ be admissible for the problem* (KD)*, i.e. $\varphi \oplus \psi \leqslant c$, and let $\gamma \in \Gamma(\mu, \nu)$ be a transport plan. Then the two assertions are equivalent*

- *γ is an optimal transport plan and (φ, ψ) is a maximizer in* (KD)
- $\mathrm{spt}(\gamma) \subseteq \{(x, y) \in X \times Y \mid \varphi(x) \oplus \psi(y) = c(x, y)\}$.

Proof. Using first the admissibility of (φ, ψ) and then $\gamma \in \Gamma(\mu, \nu)$,

$$0 \leqslant \langle c|\gamma\rangle - \langle \varphi \oplus \psi|\gamma\rangle = \langle c|\gamma\rangle - (\langle \varphi|\mu\rangle + \langle \psi|\nu\rangle$$

We see that the last term vanishes if and only if γ minimizes (KP) and (φ, ψ) maximizes (KD) (and if strong duality, (KP) = (KD), holds). But this term also vanishes if and only if the first inequality is an equality. Since $\varphi \oplus \psi \leqslant c$, this is equivalent to $c - \varphi \oplus \psi = 0$ γ-almost everywhere. □

Because of this proposition, one can think of the dual Kantorovich potentials, the prices in the economic interpretation of OT, as an "optimality certificate" for an optimal transport plan (i.e. a way to convince someone that you actually found the optimum). This leads to the following stability theorem for optimal transport maps.

Theorem 14 (Stability of OT plans)**.** *Let X, Y be compact metric spaces and let $c \in \mathcal{C}^0(X \times Y)$. Consider $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ converging weakly to μ and ν respectively.*

- *If $\gamma_k \in \Gamma(\mu_k, \nu_k)$ is optimal then, up to subsequences, (γ_k) converges weakly to an optimal transport plan $\gamma \in \Gamma(\mu, \nu)$.*
- *Let (φ_k, ψ_k) be optimal Kantorovich potentials in the dual problem between μ_k and ν_k, satisfying $\psi_k = \varphi_k^{\overline{c}}$, $\varphi_k = \psi_k^c$ and $\varphi_k(x_0) = 0$ for some $x_0 \in X$. Then, up to subsequences, the sequence (φ_k, ψ_k) converges uniformly to a maximizing pair (φ, ψ) for* (KD) *also satisfying $\varphi = \psi^c$ and $\psi = \varphi^{\overline{c}}$.*

We will use the following lemma about the convergence of the supports of weak$*$ converging measures.

Lemma 15. *If a sequence of non-negative measures $(\mu_n)_{n\in\mathbb{N}}$ weak*-converges to μ, then any point x in $\mathrm{spt}(\mu)$ is the limit as $n \to +\infty$ of points x_n in $\mathrm{spt}(\mu_n)$.*

Proof of Theorem 14. As c-concave functions, φ_k and ψ_k have the same modulus of continuity as the cost function c (see Lemma 12), and they are uniformly bounded (using $\varphi_k(x_0) = 0$). Using Arzelà-Ascoli theorem, we can therefore assume that up to subsequences, (φ_k) (resp. (ψ_k)) converges to some φ (resp ψ) uniformly. Then, one easily sees that $\varphi \oplus \psi \leqslant c$ so that (φ, ψ) are admissible for the limit dual problem (KD). By Proposition 2, we can assume, taking subsequences if necessary, that the sequence $\gamma_k \in \Gamma(\mu_k, \nu_k)$ converges to some $\gamma \in \Gamma(\mu, \nu)$.

By Proposition 13, we see that γ_k is supported on the set

$$S_k = \{(x, y) \in X \times Y \mid \varphi_k(x) + \psi_k(y) = c(x, y)\}.$$

Moreover, by Lemma 15, every pair $(x, y) \in \mathrm{spt}(\gamma)$ can be approximated by a sequence of pairs $(x_k, y_k) \in \mathrm{spt}(\gamma_k)$ i.e. $\lim_{k\to\infty}(x_k, y_k) = (x, y)$. Since γ_k is supported on S_k one has $c(x_k, y_k) = \varphi_k(x_k) + \psi_k(x_k)$. This gives at the limit $c(x, y) = \varphi(x) + \psi(y)$. We have just shown that for every point pair (x, y) in $\mathrm{spt}(\gamma)$, $c(x, y) = \varphi(x) + \psi(y)$ where φ, ψ is admissible. Applying Proposition 13 again, this shows that γ and (φ, ψ) are optimal for their respective problems. □

4. Kantorovich's functional

4.1. **Kantorovich's functional.** As already mentioned in (3.9), the Kantorovich's dual problem (KD) can be expressed as an unconstrained maximization problem, involving the c-transform.

Definition 14. The Kantorovitch functional is defined on $\mathcal{C}^0(Y)$ by

$$\mathcal{K}_\mu(\psi) = \int_X \psi^c \mathrm{d}\mu \tag{4.10}$$

The Kantorovitch dual problem therefore amounts to maximizing the Kantorovitch functional plus a linear term:

$$(\mathrm{KD}) = \max_{\psi\in\mathcal{C}^0(Y)} \mathcal{K}_\mu(\psi) + \langle\psi|\nu\rangle.$$

It is quite easy to see that $\mathcal{K}_\mu$ is concave, recalling the definition of the c-transform as a minimum. If (φ, ψ) are maximizers in the Kantorovich's dual problem (KD) between μ and ν, then ψ is a maximizer of $\mathcal{K}_\mu + \langle\cdot|\nu\rangle$.

This subsection is devoted to the computation of the superdifferential of Kantorovich's functional, in particular when the source measure μ is absolutely continuous. This computation will be used to establish existence of solutions to Monge's problem (following Brenier and Gangbo-McCann) and to construct and study algorithms for (semi-)discretized optimal transport.

Definition 15 (Response map)**.** Given a potential $\psi \in \mathcal{C}^0(Y)$, we call *response map* the set-valued map $\hat{T}_\psi$ defined by

$$\hat{T}_\psi(x) = \arg\min_{y\in Y} c(x, y) - \psi(y) = \{y \in Y \mid c(x, y) - \psi(y) = \psi^c(x)\}.$$

Remark 4 (Construction of optimal transports). One can easily sees that the graph of $\hat{T}_\psi$ is

$$\mathrm{Graph}(\hat{T}_\psi) = \{(x,y) \in X \times Y \mid \psi^c(x) + \psi(y) = c(x,y)\}.$$

We note that if ψ is a maximizer of $\mathcal{K}_\mu + \langle \cdot | \nu \rangle$, then (ψ^c, ψ) is a maximizer of (KD). By proposition Proposition 13, we see that the set of optimal transport plans between μ and ν is equal to

$$\{\gamma \in \Gamma(\mu,\nu) \mid \mathrm{spt}(\gamma) \subseteq \mathrm{Graph}(\hat{T}_\psi)\}, \tag{4.11}$$

making it a priori possible to recover a solution to the primal problem from a maximizer of the $\mathcal{K}_\mu + \langle \cdot | \nu \rangle$.

Proposition 16. *Let X, Y be compact metric spaces and let $c \in \mathcal{C}^0(X \times Y)$. Then, for all measure $\mu \in \mathcal{P}(X)$ and any $\psi \in \mathcal{C}^0(Y)$, one has*

$$\partial^+ \mathcal{K}_\mu(\psi) = \left\{ -\nu \mid \exists \gamma \in \Gamma(\mu,\nu) \text{ s.t. } \mathrm{spt}(\gamma) \subseteq \mathrm{Graph}(\hat{T}_\psi) \right\}.$$

Proof. Let $\psi \in \mathcal{C}^0(Y)$ and let $\nu \in (\mathcal{C}^0(Y))^* = \mathcal{M}(Y)$. Assume that $-\nu$ belongs to $\partial^+ \mathcal{K}_\mu(\psi)$. Then,

$$\forall \psi' \in \mathcal{C}^0(Y), \mathcal{K}_\mu(\psi') \leqslant \mathcal{K}_\nu(\psi) - \langle \psi' - \psi | \nu \rangle,$$

which is equivalent to

$$\forall \psi' \in \mathcal{C}^0(Y), \langle (\psi')^c | \mu \rangle + \langle \psi' | \nu \rangle \leqslant \langle \psi^c | \mu \rangle + \langle \psi | \nu \rangle,$$

so that (ψ^c, ψ) is a maximizer of the dual Kantorovich problem between μ and ν. By strong Kantorovich duality ($\mathcal{T}_c(\mu,\nu) = (\mathrm{KD})$), this implies that ν is non-negative, with same mass as μ, and that $\langle \psi^c | \mu \rangle + \langle \psi | \nu \rangle = \mathcal{T}_c(\mu,\nu)$. Let $\gamma \in \Gamma(\mu,\nu)$ be an optimal transport plan between μ and ν for the cost c. Then, by Proposition 13, we see that $\psi^c \oplus \psi = c$ on $\mathrm{spt}(\gamma)$ as desired.

Conversely, if a measure ν is such that there exists $\gamma \in \Gamma(\mu,\nu)$ supported on $\psi^c \oplus \psi = c$, we get using $(\psi')^c \oplus \psi \leqslant c$

$$\begin{aligned}
\mathcal{K}_\mu(\psi') = \langle (\psi')^c | \mu \rangle &= \langle (\psi')^c \oplus \psi' | \gamma \rangle - \langle \psi' | \nu \rangle \\
&\leqslant \langle c | \gamma \rangle - \langle \psi | \nu \rangle \\
&= \langle \psi^c \oplus \psi | \gamma \rangle - \langle \psi' | \nu \rangle \\
&= \mathcal{K}_\mu(\psi) + \langle \psi' - \psi | - \nu \rangle,
\end{aligned}$$

thus proving that $-\nu \in \partial^+ \mathcal{K}_\mu(\psi)$. □

4.2. **Solution of Monge's problem.** We now use Proposition 16 to prove the existence of optimal transport maps when the source measure is absolutely continuous on a compact subset of $\mathbb{R}^d$ and when the cost function satisfies a *twist condition*. This result is due to Brenier [9] in the case of the quadratic cost, that is $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d$, and Gangbo-McCann in the general case of twisted costs [14]. The question is to determine conditions under which the response map $\hat{T}_\psi$ is single-valued μ-almost everywhere.

Definition 16 (Twisted cost)**.** Let Ω_X, Ω_Y be open subsets of $\mathbb{R}^d$, and let $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$. The cost function c satisfies the *twist condition* if

$$\forall x_0 \in \Omega_X, \quad \text{the map } y \in \Omega_Y \mapsto \nabla_x c(x_0, y) \in \mathbb{R}^d \text{ is injective,} \tag{4.12}$$

where $\nabla_x c(x_0, y)$ is the gradient of $x \mapsto c(\cdot, y)$ at $x = x_0$.

Proposition 17. *Let Ω_X, Ω_Y be open subsets of $\mathbb{R}^d$, let $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$ be a cost satisfying the twist condition* (4.12)*, and let X, Y be compact subsets of Ω_X and Ω_Y. Then, for Lebesgue-almost every $x \in X$, the response map is a singleton:*

$$\hat{T}_\psi(x) = \arg\min_{y \in Y} c(x,y) - \psi(y) =: \{T_\psi(x)\}.$$

In particular, if $\mu \in \mathcal{P}(X)$ is absolutely continuous, then

$$\nabla \mathcal{K}_\mu(\psi) = -T_{\psi\#}\mu.$$

Proof. Define $\varphi = \psi^c$, i.e. $\varphi(x) = \min_{y \in Y} c(x,y) - \psi(y)$. If the minimum in the definition of the response map is not unique, there exists two distinct points y_0, y_1 in $\hat{T}_\psi(x)$. For any $i \in \{0,1\}$, we have

$$\varphi(x') = \min_{y \in Y} c(x,y) - \psi(y) \leqslant c(x', y_i) - \psi(y_i),$$

with equality at $x' = x$. Since $\nabla c(x', y_1) \neq \nabla c(x', y_0)$ by injectivity of $y \mapsto \nabla c(x', y)$, we see that φ is not differentiable at x.

Using $c \in \mathrm{Lip}(X \times Y)$, we get that φ is Lipschitz. Rademacher's theorem then implies that φ is differentiable on a set B with full Lebesgue measure in X. By the previous paragraph, we obtain that $\hat{T}_\psi$ is a singleton at any point of B. We conclude with the next lemma. □

Lemma 18. *Let $\mu \in \mathcal{P}(X)$ and let $\hat{T} : X \to Y$ be a set-valued map such that $\hat{T}(x) = \{T(x)\}$ for μ-almost every x. Then, there exists only one transport plan $\gamma \in \Gamma(\mu,\nu)$ satisfying $\mathrm{spt}(\gamma) \subseteq \gamma(\mathrm{Graph}(\hat{T}))$. This transport plan is induced by the map T, i.e. $\gamma = (\mathrm{id}, T)_\#\gamma$.*

Proof. By definition of $\gamma_T = (\mathrm{id}, T)_\#\gamma$ one has $\gamma_T(A \times B) = \mu(T^{-1}(B) \cap A)$ for all Borel sets $A \subseteq X$ and $B \subseteq Y$. On the other hand, consider the set $X' \subseteq X$ of points such that $\hat{T}(x) = \{T(x)\}$, so that $X \setminus X'$ is μ-negligible by assumption. Then,

$$\begin{aligned}
\gamma(A \times B) &= \gamma((A \cap X') \times B) \\
&= \gamma(\{(x,y) \mid x \in A \cap X', \text{ and } y \in B\}) \\
&= \gamma(\{(x,y) \mid x \in A \cap X', y \in B \text{ and } y = T(x)\}) \\
&= \gamma(\{(x,y) \mid x \in A \cap X' \cap T^{-1}(B), y = T(x)\} \\
&= \mu(A \cap X' \cap T^{-1}(B)) \\
&= \mu(A \cap T^{-1}(X))
\end{aligned}$$

thus proving the claim. □

Theorem 19 (Gangbo-McCann [14])**.** *Let Ω_X, Ω_Y be open subsets of $\mathbb{R}^d$ and let $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$ be a cost satisfying the twist condition* (4.12)*. Given compact subsets X and Y of Ω_X and Ω_Y and two probability measures $(\mu,\nu) \in \mathcal{P}^{\mathrm{ac}}(X) \times \mathcal{P}(Y)$. Then, there exists $\psi \in \mathcal{C}^0(Y)$ such that the unique optimal tranport map between μ and ν is induced by T_ψ.*

Proof of Theorem 19. Let ψ be a maximizer of $\mathcal{K}_\mu + \langle \nu | \cdot \rangle$. By equation (4.11), the set of optimal transport plans is $\{\gamma \in \Gamma(\mu,\nu) \mid \mathrm{spt}(\gamma) \subseteq \mathrm{Graph}(\hat{T}_\psi)\}$.

Combining Proposition 17 and Lemma 18, we deduce that the unique element of this set is $\gamma = (\mathrm{id}, T_\psi)_\# \mu$. □

Corollary 20 (Brenier [9]). *Let X, Y be two compact subsets of $\mathbb{R}^d$, let $c(x,y) = \|x-y\|^2$ and let $(\mu,\nu) \in \mathcal{P}^{\mathrm{ac}}(X) \times \mathcal{P}(Y)$. Then, there exists $\varphi : \mathbb{R}^d \to \mathbb{R}$ convex such that $\nabla\varphi_\#\mu = \nu$ and the unique optimal transport plan between μ and ν is induced by the map $T = \nabla\varphi$.*

Proof. We need to compute the response map associated to the maximizer ψ of $\mathcal{K}_\mu + \langle\cdot|\nu\rangle$ for the quadratic cost:

$$\begin{aligned} T_\psi(x) &= \arg\min_y \|x-y\|^2 - \psi(y) \\ &= \arg\min_y \|y\|^2 - 2\langle x|y\rangle - \psi(y) \\ &= \arg\max_y \langle x|y\rangle - \frac{1}{2}(\|y\|^2 - \psi(y)). \end{aligned}$$

Recalling the definition of the convex conjugate, one can see at once that $T_\psi = \nabla u$ where $u = \left(\frac{1}{2}(\|\cdot\|^2 - \psi)\right)^*$. □

Remark 5 (Monge-Kantorovich quantiles). Given a fixed probability density ρ on a compact domain of $\mathbb{R}^d$, e.g. $\rho \equiv 1$ on $[0,1]^d$, and any compactly supported $\nu \in \mathcal{P}(\mathbb{R}^d)$, one can denote T_ν the quadratic optimal transport map between ρ and ν. In dimension $d = 1$, one recovers the quantile function. In higher dimension, there is no canonical definition of a quantile function, but T_ν was proposed as a challenger under the name "Monge-Kantorovich quantile" by Chernozhukov, Galichon, Hallin, Henry in [11]. Being the gradient of a convex function, the Monge-Kantorovich quantile is monotone, i.e.

$$\text{for a.e. } x, y \in \mathrm{spt}(\nu), \langle T_\nu(x) - T_\nu(y) | x - y\rangle \geqslant 0.$$

This notion can be used to define multivariate notions of ranks and depth.

4.3. **Semi-discrete optimal transport.** Our working assumptions for the remainder of this section are the following:

- Ω_X, Ω_Y are two open subsets of $\mathbb{R}^d$. The cost function c belongs to $\mathcal{C}^1(\Omega_X \times \Omega_Y)$ and satisfies the twist condition (4.12).
- the source measure ρ is absolutely continuous with respect to the Lebesgue measure and is supported in a compact subset X of Ω_X.
- the target space Y is finite so that $\nu \in \mathcal{P}(Y)$ can be written under the form $\nu = \sum_{y\in Y} \nu_y \delta_y$. For simplicity, we assume that $\min_y \nu_y > 0$.

Note that by an abuse of notation, we will often conflate ρ with its density with respect to the Lebesgue measure.

Definition 17 (Laguerre tessellation)**.** The Laguerre tessellation associated to a set of prices $\psi : Y \to \mathbb{R}$ is a decomposition of the space into *Laguerre cells* defined by

$$\mathrm{Lag}_y(\psi) := \{x \in \Omega_X \mid \forall z \in Y, c(x,y) - \psi(y) \leqslant c(x,z) - \psi(z)\}. \tag{4.13}$$

When $\psi \equiv 0$, the Laguerre cells are called Voronoi cells. The Voronoi cell of the point $y \in Y$ is denoted $\mathrm{Vor}_y(\psi)$.

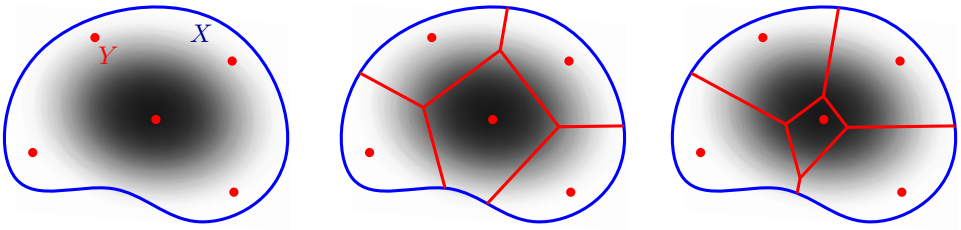

FIGURE 1. (Left) The domain X (with boundary in blue) is endowed with a probability density pictured in grayscale representing the density of population in a city. The set Y (in red) represents the location of bakeries. Here, $X, Y \subseteq \mathbb{R}^2$ and $c(x,y) = |x-y|^2$ (Middle) The Voronoi tessellation induced by the bakeries (Right) The Laguerre tessellation: the price of bread the bakery near the center of X is higher than at the other bakeries, effectively shrinking its Laguerre cell.

Remark 6 (Response map). Let $\psi \in \mathbb{R}^Y$. The response map T_ψ is constant on the interior of the Laguerre cells (and undefined on their boundary) by:

$$\forall y \in Y,\ T_\psi|\operatorname{Lag}_y = y.$$

In particular,

$$T_{\psi\#}\rho = \sum_{y\in Y} G_y(\psi)\delta_y,\ \text{where } G_y(\psi) = \rho(\operatorname{Lag}_y). \tag{4.14}$$

Theorem 21 (Aurenhammer, Hoffman, Aronov)**.** *Under the assumptions of this paragraph, the Kantorovich functional $\mathcal{K}_\mu$ is $\mathcal{C}^1$-smooth on $\mathbb{R}^Y$. Its gradient is given by*

$$\nabla\mathcal{K}_\rho(\psi) = -\sum_{y\in Y} \rho(\operatorname{Lag}_y(\psi))\delta_y \tag{4.15}$$

In particular $\psi \in \mathbb{R}^Y$ maximizes $\mathcal{K}_\rho + \langle\cdot|\nu\rangle$, where $\nu \in \mathcal{P}(Y)$, if and only if

$$\forall y \in Y,\ \rho(\operatorname{Lag}_y(\psi)) = \nu(\{y\}).$$

The only new statement in this theorem, compared to Proposition 17 is that $\mathcal{K}_\mu$ is $\mathcal{C}^1$. This is proven as point (iv) of the following lemma. In what follows, we will denote R the oscillation of the cost function:

$$R := \max_{X\times Y} c - \min_{X\times Y} c, \tag{4.16}$$

Lemma 22. *Assume c is twisted (Def. 16) and $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$. Then,*

(i) $\forall y \in Y$, the map $t \mapsto G_y(\psi + t\mathbf{1}_y)$ is non-decreasing,
(ii) $\forall y \neq z \in Y$, the map $t \mapsto G_y(\psi + t\mathbf{1}_z)$ is non-increasing,
(iii) if $\psi \in \mathbb{R}^Y$ is such that $G_{y_0}(\psi) > 0$, then $\psi(y_0) \leqslant \min_Y \psi + R$,
(iv) for all $y \in Y$, the function G_y is continuous.

Proof. The properties (i), (ii) are straightforward consequences of the definition of Laguerre cells. To prove (iii), take ψ such that $G_{y_0}(\psi) > 0$, implying in particular that the Laguerre cell $\operatorname{Lag}_{y_0}(\psi)$ is non-empty and contains a

point $x \in X$. Then, by definition of the cell one has for all $y \in Y \setminus \{y_0\}$, $c(x,y_0) + \psi(y_0) \leqslant c(x,y) + \psi(y)$, thus showing that $\psi(y_0) \leqslant \min_Y \psi + R$.

It remains to establish that each of the maps G_y is continuous. For this purpose, we consider a sequence $(\psi_n)_{n\in\mathbb{N}}$ converging to some ψ_∞. We first note that thanks to the Twist hypothesis, the set S defined by

$$\begin{aligned} S &= \{x \in X \mid \exists y \neq z \in Y \text{ s.t. } c(x,y) - \psi(y) = c(x,y) - \psi(z)\} \\ &\subseteq \bigcup_{y\in Y, z\in Y\setminus\{y\}} \{x \in X \mid c(x,y) - \psi(y) = c(x,y) - \psi(z)\}. \end{aligned}$$

is included in a finite union of $(d-1)$-dimensional submanifolds, which are all Lebesgue-negligible. Thus, S is also ρ-negligible. Defining $\chi = \mathbf{1}_{\mathrm{Lag}_y(\psi)}$ and $\chi_n = \mathbf{1}_{\mathrm{Lag}_y(\psi_n)}$, we have

$$G_y(\psi_n) = \int \chi_n \mathrm{d}\rho, \text{ and } G(\psi) = \int \chi \mathrm{d}\rho.$$

To prove that $\lim_{n\to+\infty} G_y(\psi_n) = G_y(\psi)$ it suffices to establish that χ_n converges to χ on $X \setminus S$, which is straightforward (because the inequalities defining the set $X \setminus S$ are strict), and to apply Lebesgue's dominated convergence theorem. □

4.4. **Oliker–Prussner's algorithm.** Oliker-Prussner's algorithm for solving $G(\psi) = \nu$ is described in Algorithm 1, and bears strong resemblance with Bertsekas' auction algorithm for the assignement problem [6, 7]. In particular, the values of ψ are evolved in a monotonic way.

Algorithm 1 Oliker-Prussner algorithm

Input: A tolerence parameter $\delta > 0$.
Initialization: Fix some $y_0 \in Y$ once for all. Set

$$\psi^{(0)}(y) := \begin{cases} 0 & \text{if } y = y_0 \\ R & \text{if not.} \end{cases}$$

While: $\exists y \in Y \setminus \{y_0\}$ such that $G_y(\psi^{(k)})) \leqslant \nu_y - \frac{\delta}{N}$
Step 1: Compute

$$t_y = \min\{t \geqslant 0 \mid G_y(\psi^{(k)} + t\mathbf{1}_y) \geqslant \nu_y\}. \tag{4.17}$$

Step 2: Set $\psi^{(k+1)} = \psi^{(k)} + t\mathbf{1}_y$.
Output: A vector $\psi^{(k)}$ that satisfies $\max_y \left\|G_y(\psi^{(k)}) - \nu(\{y\})\right\|_\infty \leqslant \delta$.

Theorem 23 (Oliker-Prussner). *Assume that the cost $c \in \mathcal{C}^2(\Omega_X \times \Omega_Y)$ is twisted (Def. 16) and that $\rho \in \mathcal{P}^{\mathrm{ac}}(X) \cap \mathrm{L}^\infty(X)$. Then,*

- *Oliker-Prussner's algorithm terminates in a finite number of steps.*
- *Furthermore, at the final step k, one has*

$$\max_{y\in Y} \left| G_i(\psi^{(k)}) - \nu_i \right| \leqslant \delta.$$

Proof of Theorem 23.
Step 1 *(Correctness)* When Algorithm 1 terminates with $\psi := \psi^{(k)}$, one has

for any $y \neq y_0$, $\rho(\mathrm{Lag}_y(\psi)) \leqslant \nu_y$. When it stops, it also means that one has $\rho(\mathrm{Lag}_y(\psi)) \geqslant \nu_y - \frac{\delta}{N}$. Then, as desired, we get

$$\rho(\mathrm{Lag}_{y_0}(\psi)) = 1 - \sum_{y \neq y_0} \rho(\mathrm{Lag}_{y_0}(\psi)) \in [\nu_{y_0}, \nu_{y_0} + \delta].$$

Step 2 *(A priori bound on ψ_k)* By construction one has $\rho(\mathrm{Lag}_y(\psi^{(k)})) \leqslant \nu_y$, which also imply that

$$\rho(\mathrm{Lag}_{y_0}(\psi^{(k)})) = 1 - \sum_{y \in Y \setminus \{y_0\}} \rho(\mathrm{Lag}_y(\psi^{(k)})) \geqslant \nu_{y_0} > 0.$$

By Proposition 22–(iii), we get $0 = \psi^k(y_0) \leqslant \min_Y \psi^{(k)} + R$. Since the price of y_0 is never changed, $\psi^{(k)}(y_0) = 0$ and $R \geqslant \psi^{(k)} \geqslant -R$.

Step 3 *(Minimum decrease and termination)* Since by Lemma 22–(iv) G_y is continuous, it admits a continuity modulus on the compact set $[-R, R]^Y$, i.e. a function $\omega_y : \mathbb{R} \to \mathbb{R}$ such that $\lim_{t \to 0} \omega_y(t) = 0$ and such that

$$\forall \psi, \psi' \in [-R, R]^Y, \left| G_y(\psi) - G_y(\psi') \right| \leqslant \left\| \psi - \psi' \right\|_\infty .$$

In the second step of the algorithm, when $\psi^{(k)}$ is updated one has $G_y(\psi^{(k)} - t_y \mathbf{1}_y) \geqslant G_y(\psi^{(k)}) + \frac{\delta}{N}$. Using the uniform continuity of G_y, we have

$$\frac{\delta}{N} \leqslant \left| G_y(\psi^{(k)} - t_y \mathbf{1}_y) - G_y(\psi^{(k)}) \right| \leqslant \omega(t_y),$$

implying that there exists $\tau > 0$ such that $t_y \geqslant \tau$. Since for any k, $\psi_k(y) \in [-R, R]$, the number of times k_y the price of a point $y \in Y$ has been updated is bounded: $k_y \leqslant 2R/\tau$. Thus, the algorithm terminates in finite time. □

Remark 7 (Quadratic cost). For the cost $c(x, y) = \|x - y\|^2$, but also in more general cases (see e.g. [16]), one can show that G is Lipschitz, with constant larger than CN. In this case, the number of iterations is of the algorithm is bounded by $\mathrm{O}(N^3)$.

5. Entropy-regularized optimal transport

5.1. **Primal problem.** We start from the primal formulation of the optimal transport problem, but instead of imposing the non-negativity constraints $\gamma \geqslant 0$, we add a term to the transport cost, which promotes (minus) the entropy of the transport plan and acts as a barrier for the non-negativity constraint. The entropy of a measure $\mu \in \mathcal{M}(X)$ on a compact metric space X with respect to a probability measure ω on X is defined by

$$H(\mu \mid \omega) = \begin{cases} \int h(\rho) \mathrm{d}\omega & \text{if } \mathrm{d}\mu = \rho \mathrm{d}\omega \\ +\infty & \text{otherwise} , \end{cases}$$
$$\text{where } h(r) = \begin{cases} r(\log r - 1) & \text{if } r > 0, \\ 0 & \text{if } r = 0, \\ +\infty & \text{if } r < 0. \end{cases} \tag{5.18}$$

The regularized optimal transport problem is then defined as

$$(\mathrm{KP}^\varepsilon) := \inf_{\gamma \in \Gamma(\mu, \nu)} \langle c | \gamma \rangle + \varepsilon H(\gamma \mid \mu \otimes \nu). \tag{5.19}$$

We will rely on the following dual representation of entropy.

Proposition 24 (Donsker-Varadhan)**.** *Let Z be a compact space, and let $\omega \in \mathcal{M}_+(Z)$ be finite. Then, for any measure $\mu \in \mathcal{M}(Z)$,*

$$H(\mu \mid \omega) = \sup_{f \in \mathcal{C}^0(Z)} \langle f|\mu\rangle - \langle e^f|\omega\rangle. \tag{5.20}$$

In particular, $\mu \mapsto H(\mu \mid \omega)$ is convex and weak lsc. In addition:*

(i) the supremum in (5.20) is attained at $f \in \mathcal{C}^0(Z)$ if and only if e^f is the density of μ with respect to ω.
(ii) the restriction of $\mu \mapsto H(\mu \mid \omega)$ to the set of absolutely continuous measures with respect to ω is strictly convex.

Remark 8 (Finite entropy implies non-negativity). We can prove thanks to (5.20) that if $\mu \notin \mathcal{M}^+(Z)$, then $H(\mu \mid \omega) = +\infty$. Indeed, if $\langle \mu|g\rangle < 0$ for some continuous function $g \geqslant 0$, one can check by taking $f = -\lambda g$ that

$$H(\mu \mid \gamma) \geqslant \lambda \underbrace{\langle \mu| - g\rangle}_{>0} - \langle \underbrace{e^{\lambda g}}_{\leqslant 1}|\omega\rangle \xrightarrow{\lambda \to +\infty} +\infty.$$

This means that the regularized optimal transport problem can be equivalently written by removing non-negativity constraint $\gamma \geqslant 0$:

$$(\mathrm{KP}^\varepsilon) = \inf_{\gamma \in \mathcal{M}(X\times Y) | \Pi_{X\#}\gamma = \mu, \Pi_{Y\#}\gamma = \nu} \langle c|\gamma\rangle + \varepsilon H(\gamma \mid \mu \otimes \nu).$$

Proof. Note that for $r > 0$, $h'(r) = \ln(r)$ for $r > 0$. The convex conjugate of h is therefore given by

$$h^*(s) = \sup_{r>0} rs - h(r) = e^r.$$

The Fenchel-Young inequality reads $h^*(s) + h(r) \geqslant rs$ with equality if and only if $r = e^s$. Assume that μ has density ρ with respect to ω. Then,

$$\begin{aligned} H(\mu \mid \omega) &= \int h(\rho(x))\mathrm{d}\omega(x) \\ &= \int h^{**}(\rho(x))\mathrm{d}\omega(x) \\ &= \int \sup_s s\rho(x) - h^*(\rho(x))\mathrm{d}\omega(x) \end{aligned}$$

In particular, for any bounded measurable function f we have

$$H(\mu \mid \omega) \geqslant \langle f|\rho\omega\rangle - \langle e^f|\omega\rangle = \langle f|\mu\rangle - \langle e^f|\omega\rangle,$$

with equality if $f = e^\rho$ a.e. □

Proposition 25. *The regularized optimal transport problem admits a unique solution. Moreover, the density of γ with respect to $\mu \otimes \nu$ is positive a.e.*

Remark 9 (No transport maps). In this entropy regularized setting, one cannot expect to find an optimal transport map, since minimizers of the regularized optimal transport problem are supported on the whole support of the product $\mu \otimes \nu$.

Remark 10 (Barrier). The main ingredient of the previous proposition is that the slope of $h : r \mapsto r \ln r$ is $+\infty$ at $r = 0$, which forbids the density of γ with respect to $\mu \otimes \nu$ to vanish on sets of positive measure. A stronger effect could be obtained by using a penalization of the form $\varepsilon G(\gamma \mid \mu \otimes \nu)$ instead of $\varepsilon H(\gamma \mid \mu \otimes \nu)$ where

$$G(\mu \mid \omega) = \begin{cases} \int g(\rho) \mathrm{d}\mu \otimes \nu & \text{if } \mathrm{d}\mu = \rho \mathrm{d}\omega \\ +\infty & \text{otherwise ,} \end{cases} \tag{5.21}$$

where

$$g(r) = \begin{cases} -\log r & \text{if } r > 0, \\ +\infty & \text{if } r \leqslant 0. \end{cases}$$

This barrier is stronger, as it forbis $r = 0$. When X and Y are finite, this choice is related to the interior point method for solving the optimal transport problem, where one would solve subsequent problems of the form

$$\min_{\gamma \in \Gamma(\mu,\nu)} \langle c|\gamma\rangle + \varepsilon_k H(\gamma \mid \mu \otimes \nu)$$

for a sequence of parameters ε_k converging to zero.

Proof. Existence follows from lower semi-continuity of the functional and compactness of $\Gamma(\mu, \nu)$, while uniqueness follows from the strict convexity.

Let γ^* be the optimizer of $(\mathrm{KP}^\varepsilon)$, and let ρ be the density of γ^* with respect to $\mu \otimes \nu$. We will prove by contradiction that the set $Z := \{(x, y) \mid \rho = 0\}$ satisfies $\rho(Z) = 0$. For this purpose, we define a new transport plan γ^t between μ and ν by setting $\gamma^t = (1 - t)\gamma^* + t\mu \otimes \nu$. The density of γ^t with respect to $\mu \otimes \nu$ is $\rho^t = (1 - t)\rho + t$. We give an upper bound on the energy of γ^t. We first observe that by convexity of $h(r) = r(\ln r - 1)$, we have

$$\begin{aligned} \int_{X\times Y\setminus Z} h(\rho^t) \mathrm{d}\mu \otimes \nu &\leqslant (1 - t) \int_{X\times Y\setminus Z} h(\rho^t) \mathrm{d}\mu \otimes \nu + t \int_{X\times Y\setminus Z} h(1) \mathrm{d}\mu \otimes \nu \\ &= (1 - t)H(\gamma^t \mid \mu \otimes \nu) - t \cdot \mu \otimes \nu(X \times Y \setminus Z). \end{aligned}$$

On the other hand, on Z we have $\rho^t = t$, so that

$$\int_{X\times Y\setminus Z} h(\rho^t) \mathrm{d}\mu \otimes \nu = t(\ln(t) - 1) \cdot \mu \otimes \nu(Z).$$

Finally, we note that $\langle c|\gamma^t\rangle = \langle c|\gamma^*\rangle + t(\langle \mu \otimes \nu - \gamma^*|c\rangle$. Summing these equalities and inequalities, we get

$$\langle c|\gamma^t\rangle + \varepsilon H(\gamma^t \mid \mu \otimes \nu) \leqslant \langle c|\gamma^*\rangle + \varepsilon H(\gamma^* \mid \mu \otimes \nu) + t(C + \ln(t) \cdot \mu \otimes \nu(Z)).$$

Taking t small enough, one get a contradiction on the optimality of γ^*, unless the set Z has zero $\mu \otimes \nu$–measure. □

5.2. **Dual problem.** The dual problem is constructed, as before, by introducing Lagrange multipliers $\varphi \in \mathcal{C}^0(X)$ and $\psi \in \mathcal{C}^0(Y)$ for the constraints $\Pi_{X\#}\gamma = \mu$ and $\Pi_{Y\#}\gamma = \nu$, and also dualizing the entropy using the Donsker-Varadhan formula. We have

$$\begin{aligned} (\mathrm{KP}^\varepsilon) &= \inf_{\gamma | \Pi_{X\#}\gamma = \mu \text{ and } \Pi_{X\#}\gamma} \langle c|\gamma\rangle + \varepsilon H(\gamma \mid \mu \otimes \nu) \\ &= \inf_{\gamma} \sup_{\varphi,\psi,f} \langle c - \varphi \oplus \psi|\gamma\rangle + \langle \varphi|\mu\rangle + \langle \psi|\nu\rangle + \varepsilon(\langle f|\gamma\rangle - \langle e^f|\mu \otimes \nu\rangle) \end{aligned}$$

The dual problem is constructed by inverting the infimum and the supremum:

$$(\mathrm{KD}^\varepsilon) = \sup_{\varphi,\psi,f} \inf_{\gamma} \langle c - \varphi\oplus\psi + \varepsilon f|\gamma\rangle + \langle\varphi|\mu\rangle + \langle\psi|\nu\rangle - \varepsilon\langle e^f|\mu\otimes\nu\rangle)$$

One notices that the infimum is $-\infty$ unless $c-\varphi\oplus\psi+\varepsilon f = 0$, i.e. $f = \frac{\varphi\oplus\psi-c}{\varepsilon}$. This gives us the following dual formulation

$$(\mathrm{KD}^\varepsilon) = \sup_{\varphi\in\mathcal{C}^0(X),\psi\in\mathcal{C}^0(Y)} \mathcal{K}^\varepsilon(\varphi,\psi)$$

with

$$\mathcal{K}^\varepsilon(\varphi,\psi) = \langle\varphi|\mu\rangle + \langle\psi|\nu\rangle - \varepsilon\langle e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}|\mu\otimes\nu\rangle,$$

which is a concave maximization problem.

Remark 11 (Penalization of $\varphi\oplus\psi\leqslant c$). The dual of the entropy-regularized $(\mathrm{KD}^\varepsilon)$ resembles the dual of the standard optimal transport problem, but where the hard constraint $\varphi\oplus\psi\leqslant c$ is replaced by a soft penalization: for small values of ε, $e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}$ is small only $\varphi\oplus\psi-c$ is not much larger than zero.

Lemma 26 (Weak duality)**.** *For any potentials $(\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y)$ and any transport plan $\gamma\in\Gamma(\mu,\nu)$, one has*

$$\mathcal{K}^\varepsilon(\varphi,\psi) \geqslant \langle c|\gamma\rangle + \varepsilon H(\gamma\mid\mu\otimes\nu),$$

with equality if $\gamma = e^{\frac{\varphi+\psi-c}{\varepsilon}}\mu\otimes\nu$. In particular, weak duality $(\mathrm{KP}^\varepsilon)\geqslant(\mathrm{KD}^\varepsilon)$ holds.

Proof. Denote $f = \frac{\varphi+\psi-c}{\varepsilon}$. Then,

$$\begin{aligned}\langle\varphi|\mu\rangle + \langle\psi|\nu\rangle - \varepsilon\langle e^{\frac{\varphi+\psi-c}{\varepsilon}}|\mu\otimes\nu\rangle &= \langle c|\gamma\rangle + \varepsilon\langle f|\gamma\rangle - \varepsilon\langle e^f|\mu\otimes\nu\rangle\\ &\geqslant \langle c|\gamma\rangle + \varepsilon H(\gamma\mid\mu\otimes\nu),\end{aligned}$$

with equality if and only if the density of γ with respect to $\mu\otimes\nu$ is e^f. □

Lemma 27 (Optimality condition)**.** *The gradients of $\mathcal{K}^\varepsilon$ are given by:*

$$\nabla_\varphi\mathcal{K}^\varepsilon(\varphi,\psi) = \mu - \Pi_{X\#}e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}\mu\otimes\nu$$

$$\nabla_\psi\mathcal{K}^\varepsilon(\varphi,\psi) = \nu - \Pi_{Y\#}e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}\mu\otimes\nu$$

Proof. We compute the first gradient, the second being similar. Let $(\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y)$ and let $v\in\mathcal{C}^0(X)$. Then,

$$\frac{1}{t}(\mathcal{K}^\varepsilon(\varphi+tv,\psi) - \mathcal{K}^\varepsilon(\varphi,\psi)) = \langle v|\mu\rangle - \frac{\varepsilon}{t}\langle e^{\frac{(\varphi+tv)\oplus\psi-c}{\varepsilon}} - e^{\frac{(\varphi)\oplus\psi-c}{\varepsilon}}|\mu\otimes\nu\rangle.$$

Taking the limit as $t\to 0$, we get

$$\begin{aligned}\langle\nabla\mathcal{K}^\varepsilon(\varphi,\psi)|\rangle &= \langle v|\mu\rangle - \langle ve^{\frac{\varphi\oplus\psi-c}{\varepsilon}}|\mu\otimes\nu\rangle\\ &= \langle v|\mu - \Pi_{X\#}e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}\mu\otimes\nu\rangle.\end{aligned}$$

□

Remark 12 (Existence of a maximizer to $(\mathrm{KD}^\varepsilon)$ implies strong duality). If the dual problem admits a maximizer $(\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y)$, then the optimality conditions read $\Pi_{X\#}\gamma = \mu$ and $\Pi_{Y\#}\gamma = \nu$, where

$$\gamma = e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}\mu\otimes\nu.$$

Thus, by Lemma 26, we see that γ is a minimizer for the primal problem, and that strong duality holds.

Lemma 28 (Uniqueness of maximizer up to a constant)**.** *If (φ^*, ψ^*) is a maximizer of $(\mathrm{KD}^\varepsilon)$, then for any other maximizer (φ, ψ) of $(\mathrm{KD}^\varepsilon)$, there exists a constant C such that*

$$\varphi = \varphi^* + C \ \mu\text{-a.e.}, \qquad \psi = \psi^* - C \ \nu\text{-a.e.}.$$

Proof. Let φ, ψ be another maximizer of $(\mathrm{KD}^\varepsilon)$, and let

$$\varphi' = \frac{1}{2}\varphi + \frac{1}{2}\varphi^*, \quad \psi' = \frac{1}{2}\psi + \frac{1}{2}\psi^*.$$

Then, by optimality of (φ, ψ) and (φ^*, ψ^*), we have

$$\begin{aligned} 0 &\geqslant \mathcal{K}^\varepsilon(\varphi', \psi') - \frac{1}{2}\mathcal{K}^\varepsilon(\varphi', \psi') - \frac{1}{2}\mathcal{K}^\varepsilon(\varphi*, \psi*) \\ &= -\int \left(e^{\frac{\varphi' \oplus \psi' - c}{\varepsilon}} - \frac{1}{2} e^{\frac{\varphi^* \oplus \psi^* - c}{\varepsilon}} - e^{\frac{1}{2}\frac{\varphi \oplus \psi - c}{\varepsilon}} \right) \mathrm{d}\mu \otimes \nu. \end{aligned}$$

By strong convexity of $t \mapsto e^t$, this is possible if and only if $\varphi' \oplus \psi' = \varphi^* \oplus \psi^*$ $\mu \otimes \nu$–almost everywhere. Now, choose $x^* \in \mathrm{spt}\mu$, and define $C = \langle \varphi - \varphi^* | \mu \rangle$. Then, expanding the square in the following expression and using Fubini's theorem, we obtain

$$\begin{aligned} 0 &= \int (\varphi^* \oplus \psi^* - \varphi \oplus \psi)^2 \mathrm{d}\mu \otimes \nu \\ &= \int (\varphi^*(x) - \varphi(x) - C + \psi^*(y) - \psi(y) + C)^2 \mathrm{d}(\mu \otimes \nu) \\ &= \int (\varphi^*(x) - \varphi(x) - C)^2 \mathrm{d}\mu(x) + \int (\psi^*(y) - \psi(y) + C)^2 \mathrm{d}\nu(y) \end{aligned}$$ □

5.3. **Existence of a solution to the dual.** We now prove the existence of a solution to the dual problem. As in optimal transport the trick is to prove that the maximum can be taken over a compact subset of $\mathcal{C}^0(X) \times \mathcal{C}^0(Y)$, where the potentials are uniformly continuous. This is obtained by taking the maximum with respect to one of the two variables only. For instance, let $\psi \in \mathcal{C}^0(Y)$. Then, the maximum of $\mathcal{K}^\varepsilon(\cdot, \psi)$ is attained for some φ satisfying

$$\nabla_\varphi \mathcal{K}^\varepsilon(\varphi, \psi) = 0 = \mu - \Pi_{X\#} e^{\frac{\varphi \oplus \psi - c}{\varepsilon}} \mu \otimes \nu.$$

A sufficient condition is that for μ-almost every $x \in X$,

$$1 = \int_Y e^{\varphi(x) + \psi(y) - c(x,y)} \mathrm{d}\nu(y) = e^{\frac{\varphi(x)}{\varepsilon}} \langle e^{\frac{\psi - c(x,\cdot)}{\varepsilon}} | \nu \rangle.$$

Definition 18 ((c,ε)-Transform)**.** We define the (c,ε)–transform of $\psi \in \mathcal{C}^0(Y)$ and the $(\overline{c},\varepsilon)$–transform of $\varphi \in \mathcal{C}^0(X)$ by

$$\begin{aligned} \psi^{c,\varepsilon}(x) &= -\varepsilon \log \left(\langle e^{\frac{\psi - c(x,\cdot)}{\varepsilon}} | \nu \rangle \right) \\ \varphi^{\overline{c},\varepsilon}(y) &= -\varepsilon \log \left(\langle e^{\frac{\varphi - c(\cdot,y)}{\varepsilon}} | \mu \rangle \right) \end{aligned} \tag{5.22}$$

Remark 13 (Convergence to the c-transform as $\varepsilon \to 0$). Bounding the term in the exponential in the integral defining $\psi^{c,\varepsilon}$ from below, one clearly sees

$$\psi^{c,\varepsilon}(x) \leqslant \min_{y \in \mathrm{spt}(\nu)} c(x,y) - \psi(y). \tag{5.23}$$

On the other hand, by definition of the support of ν and by continuity of $c(x,y)-\psi(y)$, for any $\eta>0$ there exists a measurable set $A\subseteq \mathrm{spt}(A)$ with $\nu(A)>0$ and such that

$$\forall z\in A,\ c(x,z)-\psi(z)\leqslant \min_{y\in\mathrm{spt}(\nu)} c(x,y)-\psi(y)+\eta=\eta$$

Then,

$$\begin{aligned}\psi^{c,\varepsilon}(x)&\geqslant -\varepsilon\log\left(\int_A e^{\frac{\psi(z)-c(x,z)}{\varepsilon}}\mathrm{d}\nu(z)\right)\\&\geqslant -\varepsilon\log\left(\int_A e^{\frac{\min_{y\in\mathrm{spt}(\nu)}c(x,y)-\psi(y)+\eta}{\varepsilon}}\mathrm{d}\nu(z)\right)\\&\geqslant \min_{y\in\mathrm{spt}(\nu)} c(x,y)-\psi(y)+\eta-\varepsilon\log\nu(A)\end{aligned}$$

Thus, $\liminf_{\varepsilon\to 0}\psi^{c,\varepsilon}(x)\geqslant \min_{y\in\mathrm{spt}(\nu)}c(x,y)-\psi(y)+\eta$. Since this holds for all $\eta>0$, we deduce with (5.23) that if $\mathrm{spt}(\nu)=Y$, then

$$\lim_{\varepsilon\to 0}\psi^{c,\varepsilon}(x)=\psi^c(x).$$

Lemma 29 (Modulus of continuity)**.** *For any $(\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y)$, the transforms $\psi^{c,\varepsilon}$ and $\varphi^{\bar{c},\varepsilon}$ have the same modulus of continuity as the cost c.*

Proof. We only prove this property for $\psi^{c,\varepsilon}$, denoting ω_c the continuity modulus of the cost c:

$$\begin{aligned}\psi^{c,\varepsilon}(x')-\psi^{c,\varepsilon}(x)&=\varepsilon\left(\log\left(\langle e^{\frac{\psi-c(x,\cdot)}{\varepsilon}}|\nu\rangle\right)-\log\left(\langle e^{\frac{\psi-c(x',\cdot)}{\varepsilon}}|\nu\rangle\right)\right)\\&=\varepsilon\left(\log\left(\langle e^{\frac{\psi-c(x',\cdot)}{\varepsilon}}e^{\frac{c(x',\cdot)-c(x,\cdot)}{\varepsilon}}|\nu\rangle\right)-\log\left(\langle e^{\frac{\psi-c(x',\cdot)}{\varepsilon}}|\nu\rangle\right)\right)\\&\leqslant\varepsilon\left(\log\left(\langle e^{\frac{\psi-c(x',\cdot)}{\varepsilon}}e^{\frac{\omega_c(\mathrm{d}_X(x,x'))}{\varepsilon}}|\nu\rangle\right)-\log\left(\langle e^{\frac{\psi-c(x',\cdot)}{\varepsilon}}|\nu\rangle\right)\right)\\&\leqslant\omega_c(\mathrm{d}_X(x,x')).\end{aligned}$$

□

Corollary 30 (Existence of solution to $(\mathrm{KD}^\varepsilon)$)**.** *The supremum in the definition of $(\mathrm{KD}^\varepsilon)$ is attained for a couple $(\varphi,\psi)\in\mathcal{C}^0(X)\times\mathcal{C}^0(Y)$ such that*

- *φ,ψ have the same continuity modulus as c,*
- *$\langle\psi|\nu\rangle=0$*

Then, $(\mathrm{KP}^\varepsilon)=(\mathrm{KD}^\varepsilon)$ and the unique solution to $(\mathrm{KP}^\varepsilon)$ is given by

$$\gamma=e^{\frac{\varphi\oplus\psi-c}{\varepsilon}}\mu\otimes\nu.$$

Proof. We note that by definition of the (c,ε) and $(\bar{c},\varepsilon)$–transforms,

$$\begin{aligned}\sup_{\varphi,\psi}\mathcal{K}^\varepsilon(\varphi,\psi)&=\sup_{\psi}\mathcal{K}^\varepsilon(\psi^{\bar{c},\varepsilon},\psi)\\&=\sup_{\psi}\mathcal{K}^\varepsilon(\psi^{\bar{c},\varepsilon},(\psi^{\bar{c},\varepsilon})^{c,\varepsilon})\\&=\sup_{\psi}\mathcal{K}^\varepsilon(((\psi^{\bar{c},\varepsilon})^{c,\varepsilon})^{\bar{c},\varepsilon},(\psi^{\bar{c},\varepsilon})^{c,\varepsilon})\\&=\sup_{\psi\in\mathcal{C}^{0,\omega_c}(X)}\mathcal{K}^\varepsilon(\psi^{\bar{c},\varepsilon},\psi),\end{aligned}$$

where $\mathcal{C}^{0,\omega}(X)$ denotes the space of continuous functions with continuity modulus ω. Since for any constant $C \in \mathbb{R}$, one has $\mathcal{K}^\varepsilon(\varphi + C, \psi - C) = \mathcal{K}^\varepsilon(\varphi, \psi)$, we may impose without loss of generality that $\langle \psi | \nu \rangle = 0$ in the optimization problem. Thus,

$$(\mathrm{KD}^\varepsilon) = \sup_{\psi \in \mathcal{C}^{0,\omega_c}(Y) | \langle \psi | \nu \rangle = 0} \mathcal{K}^\varepsilon(\psi^{\overline{c},\varepsilon}, \psi).$$

Since ψ belongs to $\mathcal{C}^{0,\omega_c}(Y)$, we have

$$\mathrm{osc}(\psi) := \max_Y \psi - \min_Y \psi \leqslant \mathrm{osc}(c) \leqslant 2 \|c\|_\infty .$$

Using in addition that $\langle \psi | \nu \rangle = 0$, we get $\|\psi\|_\infty \leqslant 2 \|c\|_\infty$. This shows that the set

$$\{\psi \in \mathcal{C}^{0,\omega_c}(Y) \mid \langle \psi | \nu \rangle = 0\}$$

is a compact subset of $\mathcal{C}^0(Y)$. Finally, we check that $\psi \mapsto \mathcal{K}^\varepsilon(\psi^{\overline{c},\varepsilon}, \psi)$ is continuous on this set, and we conclude by Arzelà-Ascoli's theorem that the maximum in $(\mathrm{KD}^\varepsilon)$ is attained. □

5.4. **Sinkhorn algorithm as block-coordinate ascent.** We study in this section the algorithm that consists in computing a maximizer to the dual problem $(\mathrm{KD}^\varepsilon)$ by optimizing the functional $\mathcal{K}^\varepsilon$ alternatively in φ and ψ. The iterations are defined by

$$\begin{cases} \varphi^{(k+1)} = (\psi^{(k)})^{c,\varepsilon} \\ \psi^{(k+1)} = (\varphi^{(k+1)})^{\overline{c},\varepsilon}. \end{cases} \tag{5.24}$$

or equivalently $\psi^{(k+1)} = S(\psi^{(k)})$ where

$$S(\psi) = (\psi^{c,\varepsilon})^{\overline{c},\varepsilon}. \tag{5.25}$$

Remark 14 (Fixed point). Assume that (φ, ψ) is a fixed point of the algorithm, i.e. $\varphi = \psi^{c,\varepsilon}$ and $\psi = \varphi^{\overline{c},\varepsilon}$, and denote $\gamma = e^{\frac{\varphi \oplus \psi - c}{\varepsilon}} \mu \otimes \nu$. Thus,

$$\max_{\hat{\varphi}} \mathcal{K}^\varepsilon(\hat{\varphi}, \psi) = \mathcal{K}^\varepsilon(\varphi, \psi).$$

The first-order optimality condition for this problem, $\nabla_\varphi \mathcal{K}^\varepsilon(\varphi, \psi) = 0$, implies that $\Pi_{X\#}\gamma = \mu$. Similarly, we get $\Pi_{Y\#}\gamma = \nu$, showing by Lemma 26 that (φ, ψ) maximizes $(\mathrm{KD}^\varepsilon)$ and γ minimizes $(\mathrm{KP}^\varepsilon)$.

Remark 15 (Relation to matrix factorization). Algorithm (5.24) is in fact a reformulation, using a logarithmic change of variable, of Sinkhorn's algorithm for finding a factorization of non-negative matrices [22]. Let $X = \{x_1, \ldots, x_N\}$, $Y = \{y_1, \ldots, y_M\}$, $c_{ij} = c(x_i, y_j)$, $\mu = \sum_i \mu_i \delta_{x_i}$ and $\nu = \sum_j \nu_j \delta_{y_j}$. Then, by the discussion of the previous paragraph, $\gamma = \sum_{i,j} \gamma_{ij} \delta_{ij}$ is a solution to the entropy-regularized optimal transport problem between μ and ν if there exists $\varphi \in \mathbb{R}^N$ and $\psi \in \mathbb{R}^N$ such that

$$\gamma_{ij} = e^{\frac{\varphi_i + \psi_j - c_{ij}}{\varepsilon}}$$

$$\text{s.t.} \quad \begin{cases} \forall i \in \{1, \ldots, N\}, \ \sum_{1 \leqslant j \leqslant N} \gamma_{ij} = \mu_i \\ \forall j \in \{1, \ldots, N\}, \ \sum_{1 \leqslant i \leqslant N} \gamma_{ij} = \nu_j. \end{cases}$$

Denote $K_{ij} = e^{-\frac{c_{ij}}{\varepsilon}}$. The iterates of Sinkhorn's algorithm are

$$\begin{cases} \varphi_i^{k+1} = -\varepsilon \log\left(\sum_j e^{\frac{\psi_j^k - c_{ij}}{\varepsilon}} \nu_j\right) \\ \psi_j^{k+1} = -\varepsilon \log\left(\sum_i e^{\frac{\varphi_i^{k+1} - c_{ij}}{\varepsilon}} \mu_i\right) \end{cases} \tag{5.26}$$

One may also record the transport plan γ^k induced by φ^k and ψ^k:

$$\gamma_{ij}^k = e^{\frac{\varphi_i^k + \psi_j^k - c_{ij}}{\varepsilon}} \mu_i \nu_j$$

Denoting $u_i^k = e^{\frac{\varphi^k}{\varepsilon}} \mu_i$, $v_i^k = e^{\frac{\varphi^k}{\varepsilon}} \nu_i$ and $K_{ij} = e^{\frac{-c_{ij}}{\varepsilon}}$, we may even simplify the iterations further:

$$\begin{cases} u_i^{k+1} = \mu_i / (K v^k)_i \\ v_j^{k+1} = \nu_j / (K^T u^{k+1})_j \\ \gamma^k = \operatorname{diag}(v^k) K \operatorname{diag}(u^k), \end{cases} \tag{5.27}$$

where $\operatorname{diag}(x)$ is the square diagonal matrix with entries x_i. It is also possible to drop the variables u, v and write the iterations purely in term of γ. In practice, this is not advised because of memory requirements: the memory to store u and v is $N + M$ while the memory to store γ is NM. In addition, the use of the variables u and v instead of φ, ψ is not advised, because the iteration (5.27) is less stable numerically than the formula (5.26) for small values of ε. In particular, for (5.26), one may use robust implementation of the LogSumExp function provided in most machine learning frameworks.

The following two properties are very similar to some properties holding for the standard c-transform. In the following, we denote $\|\cdot\|_{o,\infty}$ the pseudo-norm of uniform convergence up to addition of a constant:

$$\|f\|_{o,\infty} = \inf_{a \in \mathbb{R}} \|f + a\|_\infty = \frac{1}{2}(\sup f - \inf f).$$

This pseudo-norm will be very useful to state convergence results for Sinkhorn's algorithm for solving the regularized optimal transport problem. We first note that the Sinkhorn map is 1-Lipschitz with respect to this norm.

Proposition 31. *Let $\psi, \overline{\psi} \in \mathbb{R}^Y$. Then,*

(i) for $a \in \mathbb{R}$, $(\psi + a)^{c,\varepsilon} = \psi^{c,\varepsilon} + a$.

(ii) $\left\|\psi^{c,\varepsilon} - \overline{\psi}^{c,\varepsilon}\right\|_{\infty,o} \leqslant \left\|\psi - \overline{\psi}\right\|_{\infty,o}$.

Similar properties hold for the map $\varphi \in \mathbb{R}^X \mapsto \varphi^{\bar{c},\varepsilon}$.

Proof. (i) follows immediately from the definition

(ii) We first show that the map is 1-Lipschitz with respect to the norm of uniform convergence:

$$\begin{aligned} &\psi^{c,\varepsilon}(x) - \overline{\psi}^{c,\varepsilon}(x) \\ &\quad = \varepsilon \log\left(\langle e^{\frac{\overline{\psi} - c(x,\cdot)}{\varepsilon}} | \nu \rangle\right) - \varepsilon \log\left(\langle e^{\frac{\psi - c(x,\cdot)}{\varepsilon}} | \nu \rangle\right) \\ &\quad = \varepsilon \log\left(\langle e^{\frac{\psi - c(x,\cdot)}{\varepsilon}} e^{\frac{\overline{\psi} - \psi}{\varepsilon}} | \nu \rangle\right) - \varepsilon \log\left(\langle e^{\frac{\psi - c(x,\cdot)}{\varepsilon}} | \nu \rangle\right) \leqslant \left\|\psi - \overline{\psi}\right\|_\infty \end{aligned}$$

Taking the maximum over x leads to $\|\psi^{c,\varepsilon} - \overline{\psi}^{c,\varepsilon}\| \leqslant \|\psi - \overline{\psi}_\infty\|$. The same inequality with $\|\cdot\|_{o,\infty}$ will follow easily using (i) and the definition of the norm $\|\cdot\|_{\infty,o}$ as a minimum. □

5.5. **Linear convergence of Sinkhorn's algorithm.** In order to prove convergence, we need to strengthen the 1-Lipschitz estimation from Proposition 31. This allows to apply Picard's fixed point theorem to get the contraction of the Sinkhorn iteration (5.25). The proof we present in this chapter has been first introduced in course notes of Vialard [25].

Theorem 32 (Convergence of Sinkhorn, [25]). *The map S is a contraction for $\|\cdot\|_{o,\infty}$. More precisely,*

$$\left\|S(\psi^0) - S(\psi^1)\right\|_{o,\infty} \leqslant \left(1 - e^{-2\frac{\|c\|_{o,\infty}}{\varepsilon}}\right) \left\|\psi^0 - \psi^1\right\|_{o,\infty}.$$

In particular, the iterates $(\varphi^{(k)}, \psi^{(k)})$ of Sinkhorn's algorithm (5.24) converge with linear rate to the unique (up to constant) maximizer the regularized dual problem (KP^ε).

Remark 16 (Other convergence proofs). The convergence of Sinkhorn's algorithm is usually proven (e.g. in [23]) using a theorem of Birkhoff [8]. We refer to the recent book by Peyré and Cuturi [17] for this point of view. Other convergence proofs exist, see for instance Berman [5] (in the continuous case), and Altschuler, Weed and Rigollet [1], or Carlier [] and Nutz [] for proofs relying on the strong concavity of $\mathcal{K}^\varepsilon$.

Remark 17 (Convergence speed). This theorem shows that the Sinkhorn algorithm converges with linear speed, but the contraction constant has a bad dependency in ε. Denoting $C = \|c\|_{o,\infty}$, to get an error of $\eta > 0$, the number of iterations must satisfy

$$(1 - e^{-2C/\varepsilon})^k \lesssim \eta$$
$$\text{i.e. } k \gtrsim e^{2C/\varepsilon} \log(1/\eta),$$

where the second inequality holds for small values of ε. This bad dependency in ε seems to be a practical obstacle to choosing a very small smoothing parameter. This calls for scaling techniques, as for the auction's algorithm, and was considered by Schmitzer [20, 21].

Remark 18 (Implementation). The numerical implementation of Sinkhorn's algorithm is more complicated than it seems:

- In a naive implementation, the computation of the smoothed c-transforms (5.22) has a cost proportional to $\mathrm{Card}(X)\mathrm{Card}(Y)$. This can be alleviated for instance when $X = Y$ are grids and when the cost is a $\|\cdot\|_p$ norm, using fast convolution techniques (see e.g. [24] or [17, Remark 4.17]), or when the cost is the squared geodesic distance on a Riemannian manifold [12, 24].
- The convergence speed can be slow when the supports of the data X, Y are "far" from each other, and when ε is small. This difficulty is cirvumvented using the ε-scaling techniques mentioned above, often combined with multi-scale (coarse-to-fine) strategies, studied in this context by Benamou, Carlier and Nenna [4] and Schmitzer [20].

- Finally, some numerical difficulties (divisions by zero) can occur when ε is small and the potential ψ is far from the solution.

The book of Cuturi and Peyré present these difficulties in more details and explain how to circumvent them [17]. In addition to the works already cited, we refer to the PhD work of Feydy [10, 13], and especially to the implementation of regularized optimal transport in the library GeomLoss[1].

In order to prove this theorem, we will make use of the following elementary lemma, giving an upper bound on the total variation distance between two Gibbs kernels.

Lemma 33. *Let $u_0, u_1 \in \mathcal{C}^0(Y)$ and $\nu \in \mathcal{P}(Y)$. We denote $g_i = e^{u_i}/Z_i\nu$ where $Z_i = \langle e^{u_i}|\nu\rangle$. Then,*

$$\forall v \in \mathcal{C}^0(Y),\ |\langle v|g_1 - g_0\rangle| \leqslant 2(1 - e^{-2\|u_0-u_1\|_{o,\infty}})\,\|v\|_{o,\infty}\,.$$

Proof. Note that by definition the Gibbs kernel g_i does not change if a constant is added to u_i, so that we can assume that

$$\varepsilon := \|u_0 - u_1\|_{o,\infty} = \|u_0 - u_1\|_\infty\,.$$

Using the inequality $u_0 - \varepsilon \leqslant u_1 \leqslant u_0 + \varepsilon$, one easily shows that

$$e^{u_0-\varepsilon} \leqslant e^{u_1} \leqslant e^{u_0+\varepsilon}.$$

Integrating this inequality multiplied by ν, this implies that

$$e^{-\varepsilon}Z_0 \leqslant Z_1 \leqslant e^{\varepsilon}Z_0,\ \text{i.e. } e^{-\varepsilon}\frac{1}{Z_0} \leqslant \frac{1}{Z_1} \leqslant e^{\varepsilon}\frac{1}{Z_0}.$$

Multiplying this last inequality with the first one, we get

$$e^{-2\varepsilon}\frac{e^{u_0}}{Z_0} \leqslant \frac{e^{u_1}}{Z_1} \leqslant e^{2\varepsilon}\frac{e^{u_0}}{Z_0}.$$

Let $v \in \mathcal{C}^0(Y)$ be non-negative. Then,

$$e^{-2\varepsilon}\langle v|g_0\rangle \leqslant \langle v|g_1\rangle \leqslant e^{2\varepsilon}\langle v|g_0\rangle,$$

thus implying

$$|\langle v|g_1 - g_0\rangle| \leqslant (1 - e^{-2\varepsilon})\max(\langle v|g_0\rangle, \langle v|g_1\rangle) \leqslant (1 - e^{-2\varepsilon})\,\|v\|_\infty\,.$$

If v is not positive, we apply the previous inequality to $\hat{v} = v - \min_Y v \geqslant 0$, remarking that $\|\hat{v}\|_\infty = 2\,\|v\|_{\infty,o}$. □

Proof of Theorem 32. Consider $\psi_0, \psi_1 \in \mathcal{C}^0(Y)$. We will first give an upper bound on $\|\psi_1^{c,\varepsilon} - \psi_0^{c,\varepsilon}\|_{o,\infty}$, and to do that we will give an upper bound on

$$A(x, x') = (\psi_1^{c,\varepsilon}(x) - \psi_0^{c,\varepsilon}(x)) - (\psi_1^{c,\varepsilon}(x') - \psi_0^{c,\varepsilon}(x'))$$

which is independent of $x, x' \in X$. For this purpose, we introduce $\psi_t = \psi_0 + tv$ with $v = \psi_1 - \psi_0$, and

$$\begin{aligned} B(t, x, x') &= \psi_t^{c,\varepsilon}(x) - \psi_t^{c,\varepsilon}(x') \\ &= \varepsilon \log\left(\langle e^{\frac{\psi_t - c(x',\cdot)}{\varepsilon}}|\nu\rangle\right) - \varepsilon \log\left(\langle e^{\frac{\psi_t - c(x',\cdot)}{\varepsilon}}|\nu\rangle\right) \end{aligned}$$

[1] `https://www.kernel-operations.io/geomloss/`

Then,

$$\partial_t B(t,x,x') = \langle v | g_{x,t} - g_{x',t} \rangle, \text{ with } g_{x,t} = \frac{e^{\frac{\psi_t - c(x',\cdot)}{\varepsilon}} \nu}{\langle e^{\frac{\psi_t - c(x',\cdot)}{\varepsilon}} | \nu \rangle}.$$

Lemma 33 directly gives us

$$\left|\partial_t B(t,x,x')\right| \leqslant 2(1 - e^{-2\|c(x',\cdot)-c(x,\cdot)\|_\infty}) \, \|v\|_{\infty,o} \, .$$

We therefore get

$$\left|A(x,x')\right| \leqslant \int_0^1 \left|\partial_t B(t,x,x')\right| \leqslant 2(1 - e^{-2\|c\|_{\infty,o}}) \, \|\psi_1 - \psi_0\|_{\infty,o} \, .$$

Taking the supremum over $x, x' \in X$, we obtain

$$\|\psi_1^{c,\varepsilon} - \psi_0^{c,\varepsilon}\|_{o,\infty} = \frac{1}{2} \max_{x,x'} \left|A(x,x')\right| \leqslant \left(1 - e^{-2\frac{\|c\|_{o,\infty}}{\varepsilon}}\right) \|\psi_1 - \psi_0\|_{o,\infty} \, .$$

We conclude the proof of the contraction inequality by remarking that the map $\varphi \mapsto \varphi^{\bar{c},\varepsilon}$ is 1-Lipschitz, thanks to Proposition 31.(ii). □

6. Wasserstein distances

6.1. p-Wasserstein spaces over compact metric spaces.

Definition 19 (Wassertein distance)**.** Let (X, d_X) be a compact metric space and $p \geqslant 1$. The Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{P}(X)$ is defined as

$$\mathrm{W}_p(\mu,\nu) = \left(\min_{\gamma \in \Gamma(\mu,\nu)} \langle c_p | \gamma \rangle \right)^{1/p}, \quad c_p(x,y) := \mathrm{d}_X(x,y)^p \tag{6.28}$$

Theorem 34 (Kantorovich-Rubinstein)**.** *The Wasserstein-1 distances admit the following formulation:*

$$\mathrm{W}_1(\mu,\nu) = \sup \left\{ \langle f | \mu \rangle - \langle f | \nu \rangle \mid f \in \mathcal{C}^0(X), \ \mathrm{Lip}(f) \leqslant 1 \right\}. \tag{6.29}$$

Proof. Note that for $c = \mathrm{d}_X$, $\psi^c(x) = \min_{y \in X} d(x,y) - \psi(y)$ is 1-Lipschitz as a infimum of 1-Lipschitz functions. This implies that the dual problem may be rewritten as

$$\min_{\psi \in \mathcal{C}^0(X) | \mathrm{Lip}(\psi) \leqslant 1} \langle \psi^c | \mu \rangle + \langle \psi | \nu \rangle.$$

If ψ is 1-Lipschitz, then $d(x,y) - \psi(y) \geqslant -\psi(x)$, so that

$$\psi^c(x) = \inf_y d(x,y) - \psi(y) = -\psi(x)$$

Maximal correlation?

Theorem 35 (Properties of W_p)**.** *The following properties hold:*

(i) $\mathrm{W}_1 \leqslant \mathrm{W}_p$ *for all* $p \geqslant 1$,
(ii) W_p *is a distance on* $\mathcal{P}(X)$,
(iii) W_p *metrizes weak convergence.*

Proof. (i) The first claim is a consequence of the Jensen's inequality.
(ii) To prove the second claim, we note that the stability of optimal transport plans (Theorem 14) implies in particular that the Wasserstein distances W_p^p are weak$*$ continuous with respect to their arguments. To establish the triangle inequality, we let $\mu, \nu, \sigma \in \mathcal{P}(X)$ and we consider empirical measures

$$\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_N^i}, \quad \nu_N = \frac{1}{N}\sum_{i=1}^N \delta_{y_N^i}, \quad \sigma_N = \frac{1}{N}\sum_{i=1}^N \delta_{z_N^i}.$$

converging weak$*$ to μ, ν and σ respectively. Without loss of generality, we can reorder the points so that the optimal transport map between μ_N and ν_N is given by $x_N^i \to y_N^i$, and that the optimal transport map beween ν_N and σ_N is $y_N^i \to z_N^i$. Then,

$$\begin{aligned}
\mathrm{W}_p(\mu_N, \sigma_N) &\leqslant \left(\frac{1}{N}\sum_{1\leqslant i\leqslant N} \left\|x_N^i - z_N^i\right\|^p\right)^{1/p} \\
&\leqslant \left(\frac{1}{N}\sum_{1\leqslant i\leqslant N} \left\|x_N^i - y_N^i\right\|^p\right)^{1/p} + \left(\frac{1}{N}\sum_{1\leqslant i\leqslant N} \left\|y_N^i - z_N^i\right\|^p\right)^{1/p} \\
&= \mathrm{W}_p(\mu_N, \nu_N) + \mathrm{W}_p(\nu_N, \sigma_N)
\end{aligned}$$

We conclude by taking the limit $N \to +\infty$.
(iii) Since $\mathrm{W}_1 \leqslant \mathrm{W}_p$, if a sequence (μ_n) converges to μ with respect to W_p, then it also converges to μ with respect to W_1. Kantorovich-Rubinstein's formula then implies that for any function $f \in \mathcal{C}^0(X)$ with $\mathrm{Lip}(f) \leqslant 1$ one has $\lim_{n\to+\infty} \int f \mathrm{d}\mu_n = \int f \mathrm{d}\mu$, thus proving weak$*$ convergence of (μ_n) towards μ as $n \to +\infty$. Conversely, if μ_n converges weak$*$ to μ, then by the weak$*$ continuity of W_p^p we get

$$\lim_{n\to+\infty} \mathrm{W}_p(\mu_n, \mu) = \mathrm{W}_1(\mu, \mu) = 0. \qquad \square$$

Theorem 36 (Subdifferential of W_p^p)**.** *Let $\mu \in \mathcal{P}(X)$. The function $F = \mathrm{W}_p^p(\mu, \cdot)$ is convex and continuous in $\mathcal{P}(X) \times \mathcal{P}(X)$. Its subdifferential is given by*

$$\partial F(\nu) = \left\{\psi \in \mathcal{C}^0(X) \mid \langle \psi^c | \mu\rangle + \langle \psi | \nu\rangle = \mathrm{W}_p^p(\mu, \nu)\right\}.$$

In particular, if the dual problem $\max_\psi \langle \psi^c|\mu\rangle + \langle \psi|\nu\rangle$ has a unique solution ψ up to an additive constant, then for any measure $\nu' \in \mathcal{P}(X)$ one has

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\nu + t(\nu' - \nu))\big|_{t=0} = \langle \psi | \nu' - \nu\rangle.$$

Proof. Let $(\psi^c, \psi) \in \mathcal{C}^0(X) \times \mathcal{C}^0(X)$ be a maximizer of the dual Kantorovich problem. Then, for all measures $\nu' \in \mathcal{P}(X) \times \mathcal{P}(X)$ one has

$$\begin{aligned}
F(\nu') = \mathrm{W}_p^p(\mu, \nu') &\geqslant \langle \psi^c | \mu\rangle + \langle \psi | \nu'\rangle \\
&= \langle \psi^c | \mu\rangle + \langle \psi | \nu\rangle + \langle \psi | \nu' - \nu\rangle \\
&= F(\nu) + \langle \psi | \nu' - \nu\rangle,
\end{aligned}$$

thus showing that ψ belong to $\partial F(\nu)$. To prove the converse, we introduce $\tilde{\mathcal{K}}_\mu(\psi) = -\int \psi^c \mathrm{d}\mu$. Then,

$$\tilde{\mathcal{K}}^*_\mu(\psi) = \sup_{\nu\in\mathcal{P}(X)} \langle\nu|\psi\rangle + \langle\mu|\psi^c\rangle = \mathrm{W}_p^p(\mu,\nu) = F(\nu).$$

By subdifferential calculus, we have

$$\begin{aligned}\psi \in \partial F(\nu) &\iff \nu \in \partial F^*(\psi) = \partial\tilde{\mathcal{K}}_\mu(\psi)\\ &\iff \psi \in \arg\max\,(\mathrm{KD}),\end{aligned}$$

where the last equivalence comes from Proposition 16. □

Remark 19 (Horizontal perturbations in the discrete case). For simplicity, assume that $\mu = \frac{1}{N}\sum_i \delta_{x_i}$ and $\nu = \frac{1}{N}\sum_i \delta_{y_i}$ and that there exists unique optimal transport maps $S : \mu \to \nu$ and $T : \nu \to \nu$ (which are thus inverse of each other). Let ξ be a smooth and compactly supported vector field. Then, for small values of t, the map $(\mathrm{id} + t\xi)\circ S$ is optimal between μ and $\nu_t = (\mathrm{id} + t\xi)_\#\nu$. Thus,

$$\mathrm{W}_p^p(\nu_t,\mu) = \int \|y - (\mathrm{id}+t\xi)\circ T(y)\|^p \,\mathrm{d}\mu(y),$$

directly implying that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{W}_p^p(\nu_t,\mu) &= \int \frac{\mathrm{d}}{\mathrm{d}t}\|y - (\mathrm{id}+t\xi)\circ S(y)\|^p \,\mathrm{d}\mu(y),\\ &= \int p\,\|y - S(y)\|^{p-2}\,\langle \xi\circ S(y)|S(y)-y\rangle \mathrm{d}\mu(y),\\ &= \int p\,\|T(x) - x\|^{p-2}\,\langle \xi(x)|x - T(x)\rangle \mathrm{d}\nu(x)\end{aligned}$$

Letting T be the optimal transport map between μ. More concretely, if we denote

$$\hat{F} : (z_1,\dots,z_N) \in \mathbb{R}^{dN} \mapsto F(\frac{1}{N}\sum_i \delta_{z_i},\nu),$$

then the previous computation shows that

$$\nabla_{z_i}\hat{F}(x_1,\dots,x_N) = \frac{p}{N}\,\|T(x_i) - x_i\|^{p-2}\,(x_i - T(x_i)).$$

6.2. **p-Wasserstein geodesics on $\mathbb{R}^d$.** In this subsection, we provide a short introduction to the geometry of the Wasserstein space on $\mathbb{R}^d$. We refer to [2] for a more complete exposition.

Definition 20 (Geodesic)**.** In a metric space (X,d_X), a curve $\omega : [0,1] \to X$ is called a constant speed geodesic if

$$\forall s,t \in [0,1], \mathrm{d}_X(\omega_s,\omega_t) \leqslant |t-s|\,\mathrm{d}_X(\omega_0,\omega_1).$$

A space is called *geodesic* if any pair of points in X is joined by a geodesic.

Remark 20. Let ω be a constant speed geodesic and assume that $s \leqslant t$. Then, the triangle inequality gives us

$$\begin{aligned}\mathrm{d}_X(\omega_0,\omega_1) &\leqslant \mathrm{d}_X(\omega_0,\omega_s) + \mathrm{d}_X(\omega_s,\omega_t) + \mathrm{d}_X(\omega_t,\omega_1)\\ &\leqslant ((1-t) + (s-t) + (1-t))\mathrm{d}_X(\omega_0,\omega_1)\\ &\leqslant \mathrm{d}_X(\omega_0,\omega_1).\end{aligned}$$

Thus, all inequalities must in fact be equalities, showing in particular that

$$\mathrm{d}_X(\omega_s, \omega_t) = |t-s|\,\mathrm{d}_X(\omega_0, \omega_1).$$

Theorem 37 (Geodesics in W_p). *Let X be a convex subset of $\mathbb{R}^d$, let $\mu_0, \mu_1 \in \mathcal{P}(X)$ and let $\gamma \in \Gamma(\mu, \nu)$ be an optimal transport plan for the cost $c_p(x,y) = \|x-y\|^p$. Then, the curve $t \in [0,1] \mapsto \mu_t \in \mathcal{P}(X)$ is a constant speed geodesic between μ_0 and μ_1, with*

$$\mu_t = P_{t\#}\gamma, \qquad \text{where } P_t : (x,y) \mapsto (1-t)x + ty$$

Moreover, all constant speed geodesics between μ_0 and μ_1 are of this form. In particular, if μ_0 or μ_1 are absolutely continuous with respect to the Lebesgue measure, then the geodesic between μ_0 and μ_1 is unique.

Example 6 (Geodesics when a transport map exists). If there exists an optimal transport map T between μ_0 and μ_1, then the geodesic defined above is $\mu_t = ((1-t)\mathrm{id} + tT)_{\#}\mu_0$. In the discrete case, if

$$\mu_0 = \frac{1}{N}\sum_{1 \leqslant i \leqslant N} \delta_{x_0^i} \text{ and } \mu_1 = \frac{1}{N}\sum_{1 \leqslant i \leqslant N} \delta_{x_0^i}$$

are two empirical measures, and if the points are ordered such that

$$\mathrm{W}_p^p(\mu_0, \mu_1) = \frac{1}{N}\sum_{1 \leqslant i \leqslant N} \left\|x_1^i - x_0^i\right\|^p,$$

a geodesic between μ_0 and μ_1 is given by

$$\mu_t = \frac{1}{N}\sum_{x \in X_0} \delta_{(1-t)x_0^i + tx_1^i}.$$

Thus, μ_t provides an interpolation between the supports of μ_0 and μ_1.

Remark 21 (Many geodesics). It is quite easy to construct examples of measures μ_0 and μ_1 such that there exists more than one transport map between μ_0 and μ_1. For instance, take $\mu_0 = \frac{1}{N}\sum_i \delta_{(i/N,0)}$ and $\mu_1 = \frac{1}{N}\sum_i \delta_{(0,i/N)}$. Then, *every* bijection between the supports of μ_0 and μ_1 is optimal for $p = 2$, and therefore there exists an countably infinite number of geodesics between μ_0 and μ_1. In particular, this shows that the space $(\mathcal{P}([0,1]^2, \mathrm{W}_2)$ cannot be embedded isometrically into any Banach space.

Proof of Theorem 37. One can observe that $\gamma_{st} = (P_s, P_t)_{\#\gamma}$ has marginals μ_s and μ_t. In particular,

$$\begin{aligned}
\mathrm{W}_p^p(\mu_s, \mu_t) &\leqslant \int \|x_s - x_t\|^p \,\mathrm{d}\gamma_{st}(x_s, x_t) \\
&= \int \|(1-s)x + sy - (1-t)x + sx\|^p \,\mathrm{d}\gamma(x,y) \\
&= (t-s)^p \int \|x-y\|^o \,\mathrm{d}\gamma(x,y) = (t-s)^o\,\mathrm{W}_p^p(\mu_0, \mu_1),
\end{aligned}$$

thus proving that μ_t is a constant speed geodesic.

Let us now prove that all geodesics are of this form. For every $T \in \mathbb{N}$ and any $i \in \{1, \dots, T+1\}$, denote $\gamma_{i,i+1}^T$ an optimal transport between μ_{t_i} and

$\mu_{t_{i+1}}$, with $t_i = (i-1)/T$. By the gluing lemma recalled below, there exists $\Gamma^T \in \mathcal{P}(X^{T+1})$ whose projection on (X_i, X_{i+1}) agrees with $\gamma^T_{i,i+1}$. Moreover,

$$\begin{aligned}\left(\int \|x_1 - x_{T+1}\|^2 \,\mathrm{d}\Gamma^T(x_1,\dots,x_{T+1})\right)^{1/2} &\leqslant \sum_{j=0}^{T-1} \left(\int \|x_{i+1} - x_i\|^2 \,\mathrm{d}\Gamma^T(x_1,\dots,x_{T+1})\right)^{1/2} \\ &= \sum_{j=1}^{T+1} \mathrm{W}_2(\mu_{t_i}, \mu_{t_i}) \\ &= \mathrm{W}_2(\mu_0,\mu_1)\end{aligned}$$

This implies in particular that $\gamma^T = (\Pi_1, \Pi_{T+1})_\# \Gamma^T$, but also that for Γ^T-almost every $x = (x_1,\dots,x_{T+1})$, the points $x_1,\dots,x_{T+1}$ are aligned, i.e. $x_i = (1-t_i)x_1 + t_i x_{T+1}$. Thus, we see that $\Gamma^T = (P_0, P_{1/T}, \dots, P_1)_\# \gamma^T$ with $P_t(x,y) = (1-t)x + ty$. In particular, we have $\mu_t = P_{t\#}\gamma^T$ for all $t \in \{0/T,\dots,T/T\}$. One can finally check that if γ is a weak$*$-limit of γ, then for all $t \in [0,1]$, one has $\mu_t = P_{t\#}\gamma$. □

Lemma 38 (Gluing)**.** *Let $X_1,\dots,X_N$ be compact metric spaces, and for any $1 \leqslant i \leqslant N-1$ consider a transport plan $\gamma_i \in \Gamma(\mu_i,\mu_{i+1})$. Then, there exists $\gamma \in \mathcal{P}(X_1,\dots,X_N)$ such that for all $i \in \{1,\dots,N-1\}$, $\pi_{i,i+1}\gamma = \gamma_i$, where $\pi_{i,i+1} : X_1 \times \cdots \times X_N \to X_i \times X_{i+1}$ is the projection.*

Proof. See Lemma 5.3.2 and Remark 5.3.3 in [3]. □

6.3. Geodesic convexity with respect to W_2 on $\mathbb{R}^d$.

Definition 21 (Geodesic convexity for sets)**.** A set $S \subseteq \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ is called geodesically convex if for any $\mu_0,\mu_1 \in S$, any W_2–geodesic between μ_0 and μ_1 remains in S.

Example 7 (Geodesically convex subsets of $(\mathcal{P}(X), \mathrm{W}_2)$.). Example of geodesically convex subsets of $\mathcal{P}(X)$ include :

(a) the set obtained by translating and shearing a reference measure μ,
$$\{T_\#\mu \mid T(x) = Ax + b, A \text{ symmetric}, A \geqslant 0\}$$
In particular, the set of Gaussians densities is geodesically convex in $\mathcal{P}(\mathbb{R}^d)$. The restriction of the Wasserstein distance on this set can be computed in near closed-form, and called the Bures-Wasserstein metric.
(b) the set $\mathcal{P}^{\mathrm{ac}}(X)$ of absolutely continuous measures
(c) the set of probability densities whose density is upper bounded by a constant
(d) the set of measures of the form $\mu = \frac{1}{N}\sum_i \delta_{x_i}$ (where the points x_i are not necessary distinct) is convex under *some* geodesics, namely those induced by bijections (cf Example 6.

Proposition 39. *The set $\mathcal{P}^{\mathrm{ac}}(X)$ is geodesically convex. More precisely, given $\mu_0 \in \mathcal{P}^{\mathrm{ac}}(X)$ and $\mu_1 \in \mathcal{P}(X)$, one has $\mu_t \in \mathcal{P}^{\mathrm{ac}}(X)$ for any $t \in [0,1)$.*

Proof. Let $\mu_0 \in \mathcal{P}^{\mathrm{ac}}(X), \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ and $\varphi \in \mathrm{Lip}(X)$ be a convex function so that $\mu_t = ((1-t)\mathrm{id} + t\nabla\varphi)_\#\mu_0$ is the unique Wasserstein geodesic between

μ_0 and μ_1. Define $T_t = (1-t)\mathrm{id} + t\nabla\varphi$. Then, for any $x, y \in \mathrm{spt}(\mu_0)$,

$$\begin{aligned}\langle T_t(x) - T_t(y) | x - y\rangle &= (1-t)\left\|x-y\right\|^2 + t\langle \nabla\varphi(x) - \nabla\varphi(y) | x - y\rangle \\ &\geqslant (1-t)\left\|x-y\right\|^2,\end{aligned}$$

where we used the monotonicity of the gradient of convex functions to get the inequality. In particular, if $x \neq y$ and $t < 1$, then $T_t(x) \neq T_t(y)$ and the inverse map T_t^{-1} is well-defined. Moreover, the same inequality shows that T_t^{-1} is Lipschitz with constant $L = 1/(1-t)$. In addition, T_t^{-1} transports μ_t to μ_0, i.e. $\mu_t(B) = \mu_0(T_t^{-1}(B))$ for any Borel set B. Thus, if N is Lebesgue-negligible, $T_t^{-1}(N)$ is also negligible (by the next lemma), so that $\mu_t(N) = \mu_0(T_t^{-1}(N)) = 0$. This implies that $\mu_t \ll \lambda$. □

Lemma 40. *If N is Lebesgue-negligible, and if S is Lipschitz, then $S(N)$ is Lebesgue-negligible.*

Definition 22 (Geodesic convexity for functions)**.** A function $F : \mathcal{P}^{\mathrm{ac}}(X)$ to $\mathbb{R} \cup \{+\infty\}$ is *geodesically convex* if and only if for any $\mu_0, \mu_1 \in \mathcal{P}^{\mathrm{ac}}(W)$,

$$F(\mu_t) \leqslant (1-t)F(\mu_0) + tF(\mu_1) \tag{6.30}$$

where (μ_t) is the W_2–geodesic. Following McCann, a geodesically convex function is often called displacement convex.

Definition 23 (Internal energy)**.** Let $A : \mathbb{R}^+ \to \mathbb{R} \cup \{+\infty\}$. The *înternal energy* associated to A generalizes Boltzmann's functional. It is defined as

$$E_A : \mu \in \mathcal{P}(X) \mapsto \begin{cases} \int_\Omega A(\rho(x))\mathrm{d}x \text{ if } \mu \ll \lambda \text{ and } \rho := \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \\ +\infty \text{ if not} \end{cases} \tag{6.31}$$

Theorem 41 (McCann)**.** *Let $A : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\}$ be such that*

(i) $A(0) = 0$ and

(ii) $r \mapsto A(r^{-d})r^d$ is convex non-increasing.

Then internal energy E_A is displacement convex on $\mathcal{P}(X)$.

We will call conditions (i) and (ii) *McCann's conditions*. Example of functions A that satisfy such conditions include

- $A(r) = r^q$ for $q > 1$;
- $A(r) = r\log r$;
- $A(r) = -r^m$ for $m \in [1 - 1/d, 1)$.

This theorem is a corollary of the more general result below. Indeed, take $\mu_0 = \mu \in \mathcal{P}^{\mathrm{ac}}(X)$, $\varphi_0 = \frac{1}{2}\left\|\cdot\right\|^2$ and φ_1 a convex function such that $T = \nabla\varphi_1$ is the optimal transport map between μ_0 and μ_1. Then,

$$\mu_t = ((1-t)\nabla\varphi_0 + t\nabla\varphi_1)_\#\mu = ((1-t)\mathrm{id} + tT)_\#\mu_0$$

is the unique Wasserstein geodesic between μ_0 and μ_1.

Theorem 42. *Let $\mu \in \mathcal{P}^{\mathrm{ac}}(X)$ and let $\varphi_0, \varphi_1 \in \mathrm{Lip}(X)$ be two convex functions such that $\nabla\varphi_i(X) \subseteq X$, and let $\varphi_t = (1-t)\varphi_0 + t\varphi_1$. Assume that $A : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies McCann's conditions. Then*

$$t \in [0,1] \mapsto E_A(\nabla\varphi_{t\#}\mu)$$

is convex.

We only prove this theorem when the functions φ_0 and φ_1 are $\mathcal{C}^2$ and uniformly convex, which implies that the gradients $\nabla\varphi_i$ are diffeomorphisms from X to $\nabla\varphi_i(X)$. The proof in the general case can be found in the article of McCann [15] or in Villani's first book [26].

Lemma 43. *Assume that $\mu \in \mathcal{P}^{\mathrm{ac}}(X)$ has density ρ and that $\varphi \in \mathcal{C}^2(X)$ is uniformly convex . Then*

$$E_A(\nabla\varphi_{\#}\mu) = \int_{\mathbb{R}^d} A\left(\frac{\rho(x)}{\det(\mathrm{D}^2\varphi(x))}\right)\det(\mathrm{D}^2\varphi(x))\mathrm{d}x.$$

Proof. Since T is a diffeomorphism, the measure $T_{\#}\mu$ is absolutely continuous with respect to the Lebesgue measure. We denote σ the density of $T_{\#}\mu$, which satisfies

$$\sigma(T(x))\det(\mathrm{D}T(x)) = \rho(x)$$

Moreover, by the change of variable formula $y = T(x)$ and using $\det(\mathrm{D}T(x)) = |\det \mathrm{D}T(x))|$, which follows from the convexity of T, we get

$$\begin{aligned} E_A(\nabla\varphi_{\#}\mu) &= \int A(\sigma(y))\mathrm{d}y \\ &= \int A(\sigma(T(x)))\det(\mathrm{D}T(x))\mathrm{d}x \\ &= \int A\left(\frac{\rho(x)}{\det(\mathrm{D}T(x))}\right)\det(\mathrm{D}T(x))\mathrm{d}x \end{aligned}$$ □

Lemma 44. *The map $M \mapsto \det(M)^{1/d}$ is concave over the set of symmetric positive d-by-d matrices.*

Proof. Recall Hadamard's formula for a symmetric positive matrix M:

$$\det(M) = \min_{e_1,\dots,e_d \text{ orthonormal}} \langle e_1|Me_1\rangle \cdots \langle e_d|Me_d\rangle,$$

where the minimum is taken over orthonormal bases. Given a fixed orthonormal basis $e_1, \dots, e_d$ consider $f(M) = (\langle e_1|Me_1\rangle \cdots \langle e_d|Me_d\rangle)^{1/d}$. Then f is concave over the set of matrices M satisfying $\langle e_i|Me_i\rangle \geqslant 0$ as the composition ofthe geometric mean $(x \in (\mathbb{R}^+)^d \mapsto (x_1 \cdots x_d)^{1/d})$ with linear functions. Then, $\det(\cdot)^{1/d}$ is concave over the set of symmetric positive matrices, as a minimum of concave functions. □

Proof of Theorem 42. If φ_0, φ_1 are $\mathcal{C}^2$ and uniformly convex, the interpolant $\varphi_t := (1-t)\varphi_0 + t\varphi_1$ is also $\mathcal{C}^2$ and uniformly convex. Hence, by Lemma 43,

$$E_A(\nabla\varphi_{t\#}\mu) = \int_X B(D(x,t))\rho(x)\mathrm{d}x,$$

where we have set $B(r) = A(r^{-d})r^d$ and $D(x,t) = (\det(\mathrm{D}^2\varphi_t(x))/\rho(x))^{1/d}$. By Lemma 44, for all $x \in X$, $t \in [0,1] \mapsto D(x,t)$ is concave so that

$$D(x,t) \geqslant (1-t)D(x,0) + tD(x,1).$$

Hence, since B is non-decreasing and convex,

$$B(D(x,t)) \leqslant B((1-t)D(x,0)+tD(x,1)) \leqslant (1-t)B(D(x,0))+tB(D(x,1)).$$

Integrating this inequality gives the desired convexity result. □

Corollary 45 (Brunn-Minkowski's inequality). *Let K_0, K_1 be compact subsets of X, and let $K_t = (1-t)K^0$. Then,*

$$\lambda(K_t)^{1/d} \geqslant (1-t)\lambda(K_0)^{1/d} + t\lambda(K_1)^{1/d}.$$

Proof. Assume that $\lambda(K_0), \lambda(K_1) > 0$. Let $\mu_i = \lambda|_{K_i}/\lambda(K_i)$, let μ_t be the geodesic between μ_0 and μ_1. Then μ_t is absolutely continuous, with density ρ_t, and supported on K_t. The convexity of $A(r) = -r^{1-1/d}$ and Jensen's inequality implies

$$\begin{aligned}\int_{K_t} A(\rho_t(x))\mathrm{d}\lambda(x) &= \lambda(K_t)\int_{K_t} A(\rho_t(x))\frac{\mathrm{d}\lambda(x)}{\lambda(K_t)} \\ &\leqslant \lambda(K_t)A\left(\int_{K_t}\rho_t(x)\frac{\mathrm{d}\lambda(x)}{\lambda(K_t)}\right) \\ &= \lambda(K_t)A(1/\lambda(K_t)) = -\lambda(K_t)^{1/d}\end{aligned}$$

Moreover, for $t = 0$ and $t = 1$ we get

$$\int_{K_i} A(\rho_i(x))\mathrm{d}\lambda(x) = \int_{K_i}\lambda(K_i)A(1/\lambda(K_i)) = -\lambda(K_i)^{1/d}$$

□

References

1. Jason Altschuler, Jonathan Weed, and Philippe Rigollet, *Near-linear time approximation algorithms for optimal transport via sinkhorn iteration*, Advances in Neural Information Processing Systems, 2017, pp. 1964–1974.
2. Luigi Ambrosio and Nicola Gigli, *A user's guide to optimal transport*, Modelling and optimisation of flows on networks, Springer, 2013, pp. 1–155.
3. Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.
4. Jean-David Benamou, Guillaume Carlier, and Luca Nenna, *A numerical method to solve multi-marginal optimal transport problems with coulomb cost*, Splitting Methods in Communication, Imaging, Science, and Engineering, Springer, 2016, pp. 577–601.
5. Robert J. Berman, *The Sinkhorn algorithm, parabolic optimal transport and geometric Monge-Ampère equations*, arXiv preprint arXiv:1712.03082, 2017.
6. D.P. Bertsekas, *A new algorithm for the assignment problem*, Mathematical Programming **21** (1981), no. 1, 152–171.
7. D.P. Bertsekas and J. Eckstein, *Dual coordinate step methods for linear network flow problems*, Mathematical Programming **42** (1988), no. 1, 203–243.
8. Garrett Birkhoff, *Tres observaciones sobre el algebra lineal*, Univ. Nac. Tucuman, Ser. A **5** (1946), 147–154.
9. Yann Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on pure and applied mathematics **44** (1991), no. 4, 375–417.
10. Benjamin Charlier, Jean Feydy, Joan Alexis Glaunes, and Alain Trouvé, *An efficient kernel product for automatic differentiation libraries, with applications to measure transport*, Working version, 2017.
11. Victor Chernozhukov, Alfred Galichon, Marc Hallin, Marc Henry, et al., *Monge–kantorovich depth, quantiles, ranks and signs*, The Annals of Statistics **45** (2017), no. 1, 223–256.
12. Keenan Crane, Clarisse Weischedel, and Max Wardetzky, *Geodesics in heat: A new approach to computing distance based on heat flow*, ACM Transactions on Graphics (TOG) **32** (2013), no. 5, 152.

13. Jean Feydy, Pierre Roussillon, Alain Trouvé, and Pietro Gori, *Fast and scalable optimal transport for brain tractograms*, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 636–644.
14. Wilfrid Gangbo and Robert J McCann, *The geometry of optimal transportation*, Acta Mathematica **177** (1996), no. 2, 113–161.
15. Robert J McCann, *A convexity principle for interacting gases*, Advances in mathematics **128** (1997), no. 1, 153–179.
16. Quentin Merigot and Boris Thibert, *Optimal transport: discretization and algorithms*, Handbook of Numerical Analysis, vol. 22, Elsevier, 2021, pp. 133–212.
17. Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.
18. Filippo Santambrogio, *Optimal transport for applied mathematicians*, Springer, 2015.
19. Giuseppe Savaré and Giacomo E Sodini, *A simple relaxation approach to duality for optimal transport problems in completely regular spaces*, Journal of Convex Analysis **29** (2022), no. 1, 1–12.
20. Bernhard Schmitzer, *A sparse multiscale algorithm for dense optimal transport*, Journal of Mathematical Imaging and Vision **56** (2016), no. 2, 238–259.
21. ———, *Stabilized sparse scaling algorithms for entropy regularized transport problems*, SIAM Journal on Scientific Computing **41** (2019), no. 3, A1443–A1481.
22. Richard Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The annals of mathematical statistics **35** (1964), no. 2, 876–879.
23. Richard Sinkhorn and Paul Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific Journal of Mathematics **21** (1967), no. 2, 343–348.
24. Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas, *Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains*, ACM Transactions on Graphics (TOG) **34** (2015), no. 4, 66.
25. François-Xavier Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*, Lecture, May 2019.
26. Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.
27. ———, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.