A **crash** introduction to Optimal Transport

Matherials (Inria-Paris) and Cermics (ENPC)

Luca Nenna

January, 2024

Contents

Program

17/01/2024 Monge and Kantorovich problems, discrete and continuous cases, existences results.

18/01/2024 Dual problem, optimality conditions, optimal transport maps.

24/01/2024 Entropic optimal transport and Sinkhorn algorithm.

02/02/2024 A glimpse of multi-marginal OT and applications.

1 Some motivations for studying optimal transport.

- Variational principles for (real) Monge-Ampère equations occurring in geometry (e.g. Gaussian curvature prescription) or optics.

- Wasserstein/Monge-Kantorovich distance between clouds of particles μ, ν on e.g. $\mathbb{R}^d$: how much kinetic energy does one require to move a distribution of particles described by μ to ν ?
$\longrightarrow$ interpretation of some parabolic PDEs as Wasserstein gradient flows, construction of (weak) solutions, numerics, e.g.

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho v) = 0 \\ v = -\nabla \log \rho \end{cases} \quad \text{or} \quad \begin{cases} \partial_t \rho + \operatorname{div}(\rho v) = 0 \\ v = -\nabla p - \nabla V \\ p(1-\rho) = 0 \\ p \geqslant 0, \rho \leqslant 1 \end{cases}$$

(synthetic notion of Ricci curvature for metric spaces), machine learning, inverse problems, etc.
- Quantum physics: electronic configuration in molecules and atoms.
- Economics : μ is the distribution of men and ν the distribution of women: how can we match men and women such that everyone has an happy marriage?
- Imaging, Game theory, Mean Field Games, Fluid Dynamics, Cosmology: **Optimal Transport is everywhere!**

References.

Introduction to optimal transport, with applications to PDE and/or calculus of variations can be found in books by Villani [20] and Santambrogio [18]. Villani's second book [21] concentrates on the application of optimal transport to geometric questions (e.g. synthetic definition of Ricci curvature). We also mention Gigli, Ambrosio and Savaré [1] for the study of gradient flows with respect to the Monge-Kantorovich/Wasserstein metric. On the Economics side we refer the interested reader to [10] and for the applications in data sciences we suggest [15].

2 The problems of Monge and Kantorovich

Let us start by giving some notations/remarks/definitions.
Discrete measures: discrete measure with weights $\mathbf{a}$ and locations $x_1, \cdots, x_n \in X \subset \mathbb{R}^n$ reads

$$\mu = \sum_{I=1}^{n} \mathbf{a}_i \delta_{x_i},$$

where δ_{x_i} is the Dirac at position x_i. Such a measure describes a probability measure if, additionally, $\mathbf{a} \in \Sigma_n := \{\mathbf{a} \in \mathbb{R}^n_+ \mid \sum_{i=1}^n \mathbf{a}_i = 1\}$ and a more generally positive measure if all the elements of the vector $\mathbf{a}$ are nonnegative.
General measures: Let be X a compact subset of $\mathbb{R}^n$; we denote by $\mathcal{P}(X)$ the set of probability measures on X, by $\mathcal{M}_+(X)$ the set of positive measures on X.
Absolutely continuous measures: a measure μ which is a weighting of another reference one $\mathrm{d}x$ is said to have a density, which is denoted $\mathrm{d}\mu = \overline{\mu}\mathrm{d}x$ (in the following we always assume that $\mathrm{d}x$ is the Lebesgue measure), that is,

$$\forall f \in \mathcal{C}(X), \int_X f(x)\mathrm{d}\mu(x) = \int_X f(x)\overline{\mu}(x)\mathrm{d}x.$$

Definition 2.1 (Push-forward)**.** Given $X, Y \subset \mathbb{R}^n$, for $T : X \to Y$, the push-forward measure $\nu = T_\sharp \mu \in \mathcal{M}_+(Y)$ of some $\mu \in \mathcal{M}_+(X)$ satisfies

$$\forall f \in \mathcal{C}(Y), \int_Y f(y) \mathrm{d}\nu(y) = \int_X f(T(x)) \mathrm{d}\mu(x).$$

Note that $T_\sharp$ preserves positivity and total mass, that is if $\mu \in \mathcal{P}(X)$ then $T_\sharp \mu \in \mathcal{P}(Y)$.

Example 2.2. If μ is a discrete measure then

$$T_\sharp \mu := \sum_i \mathbf{a}_i \delta_{T(x_i)}.$$

Example 2.3 (Push-forward for densities)**.** Explicitly doing the change of variable $y = T(x)$ for measures with densities $\overline{\mu}, \overline{\nu}$ (assuming T is a $\mathcal{C}^1$ diffeomorphism), one has for all $f \in \mathcal{C}(Y)$

$$\int_Y f(y)\overline{\nu}(y)\mathrm{d}y = \int_X f(T(x))\overline{\nu}(T(x)) \det(\mathrm{D}T(x))\mathrm{d}x = \int_X f(T(x))\overline{\mu}(x)\mathrm{d}x.$$

Hence,

$$\overline{\mu}(x) = \overline{\nu}(T(x)) \det(\mathrm{D}T(x)).$$

2.1 The matching problem

Definition 2.4 (Matching problem)**.** Given a cost matrix $C \in \mathbb{R}^n \times \mathbb{R}^n$ (we are assuming that the two measures μ and ν are supported on the same number of Diracs with weights equal to $1/n$) the optimal assignment problem seeks for a bijection σ in the set of permutations of n elements $\mathfrak{S}_n$ solving

$$\min_{\sigma \in \mathfrak{S}_n} \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}. \tag{2.1}$$

One can naively evaluate the cost function above by using all permutations in the set $\mathfrak{S}_n$. However, that set has size $n!$, which is gigantic even for small n!!!. In general an optimal σ is not unique.

Let us consider now a cost of the form $C_{ij} = h(x_i - y_j)$ where $h : \mathbb{R} \to \mathbb{R}_+$ is strictly convex, one has that an optimal σ will satisfy the following inequality: given $(x_i, y_{\sigma(i)})$ and $(x_j, y_{\sigma(j)})$ then

$$h(x_i - y_{\sigma(i)}) + h(x_j - y_{\sigma(j)}) \leqslant h(x_i - y_{\sigma(j)}) + h(x_j - y_{\sigma(i)}).$$

Otherwise it would be more efficient to move mass from x_i to $y_{\sigma(j)}$ and x_j to $y_{\sigma(i)}$. The above inequality and the strict convexity of h imply that the optimal σ defines an increasing map, that is,

$$\forall (i,j)\ (x_i - x_j)(y_{\sigma(i)} - y_{\sigma(j)}) \geqslant 0.$$

Thus, the algorithm to compute an optimal transport, i.e. the optimal permutation σ, is to sort the points: find some pair of permutations σ_X, σ_Y such that

$$x_{\sigma_X(1)} \leqslant x_{\sigma_X(2)} \leqslant \cdots \text{ and } y_{\sigma_Y(1)} \leqslant y_{\sigma_Y(2)} \leqslant \cdots$$

and then an optimal matching is to send $x_{\sigma_X(k)}$ to $y_{\sigma_Y(k)}$, that is, the optimal permutation is given by $\sigma = \sigma_Y^{-1} \circ \sigma_X$.

2.2 Monge problem

Definition 2.5 (Monge problem)**.** Consider $X, Y \subseteq \mathbb{R}^n$, 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

$$\mathcal{M}_c(\mu, \nu) := \inf \left\{ \int_X c(x, T(x)) \mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_\# \mu = \nu \right\} \tag{2.2}$$

This problem exhibits several difficulties, one of which is that both the constraint ($T_\# \mu = \nu$) and the functional are non-convex. For empirical measure with the same number $n = m$ of points, one retrieves the optimal matching problem.

Example 2.6. There might exist no transport map between μ and ν. For instance, consider $\mu = \delta_x$ for some $x \in X$. Then, $\mathrm{T}_\# \mu = \delta_{T(x)}$. In particular, if ν is not a single Dirac then there exists no transport map between μ and ν.

In the special case in which $c(x, y) = d^p(x, y)$ where d is a distance, we denote

$$\mathcal{W}_p(\mu, \nu) := \left(\inf \left\{ \int_X d^p(x, T(x)) \mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_\# \mu = \nu \right\} \right)^{1/p}.$$

If the constraint set is empty, then we set $\mathcal{W}_p^p = +\infty$. In particular $\mathcal{W}_p$ defines a distance between probability measures!

Proposition 2.7. *$\mathcal{W}_p$ is a distance.*

Proof. If $\mathcal{W}_p^p(\mu, \nu) = 0$ then the optimal map is the identity Id which means that $\mu = \nu$. We have now to prove the triangle inequality

$$\mathcal{W}_(\mu, \nu) \leqslant \mathcal{W}_p(\mu, \eta) + \mathcal{W}_p(\eta, \nu).$$

If $\mathcal{W}_p^p(\mu, \nu) = +\infty$, then either $\mathcal{W}_p^p(\mu, \eta) = +\infty$ or $\mathcal{W}_p^p(\eta, \nu) = +\infty$. Indeed, consider two maps S, T such that $S_\sharp \mu = \eta$ and $T_\sharp \eta = \nu$ then $(T \circ S)_\sharp \mu = \nu$ and we have $\mathcal{W}_p^p(\mu, \nu) \leqslant \int_X d^p(x, T \circ S(x)) \mathrm{d}\mu(x) < +\infty$. So consider $\mathcal{W}_p^p(\mu, \nu) < +\infty$ and restrict our attention to the case in which $\mathcal{W}_p^p(\mu, \eta) < +\infty$ and $\mathcal{W}_p^p(\eta, \nu) < +\infty$, otherwise the inequality is trivial. For any $\varepsilon > 0$, we consider $\varepsilon-$minimizers S and T such that

$$\left(\int_X d^p(x, S(x)) \mathrm{d}\mu(x) \right)^{1/p} \leqslant \mathcal{W}_p(\mu, \eta) + \varepsilon \text{ and } \left(\int_X d^p(x, T(x)) \mathrm{d}\eta(x) \right)^{1/p} \leqslant \mathcal{W}_p(\eta, \nu) + \varepsilon.$$

Take the map $T \circ S$, then we have

$$\mathcal{W}_p(\mu, \nu) \leqslant \left(\int_X d^p(x, T \circ S(x)) \mathrm{d}\mu(x) \right)^{1/p} \leqslant \left(\int_X (d(x, S(x)) + d(S(x), T \circ S(x)))^p \mathrm{d}\mu(x) \right)^{1/p},$$

And by using the Minkowski inequality we obtain

$$\mathcal{W}_p(\mu, \nu) \leqslant \left(\int_X d^p(x, S(x)) \mathrm{d}\mu(x) \right)^{1/p} + \left(\int_X d^p(S(x), T \circ S(x)) \mathrm{d}\mu(x) \right)^{1/p}.$$

Thus

$$\mathcal{W}_p(\mu, \nu) \leqslant \mathcal{W}_p(\mu, \eta) + \mathcal{W}_p(\eta, \nu) + 2\varepsilon,$$

and by letting $\varepsilon \to 0$ we have the desired inequality. $\square$

We consider now the 1−dimensional case: for a measure μ on $\mathbb{R}$ we define the cumulative function

$$\forall x \in \mathbb{R},\ F_\mu(x) := \int_{-\infty}^{x} \mathrm{d}\mu(x),$$

which is a function $F_\mu : \mathbb{R} \to [0,1]$ and its pseudo-inverse $F_\mu^{-1} : [0,1] \to \mathbb{R} \cup \{-\infty\}$ is given by

$$\forall s \in [0,1],\ F_\mu^{-1} = \min_x \{x \in \mathbb{R} \mid F_\mu(x) \geqslant s\}.$$

If μ has a density, one can prove that for a strictly convex h, such that $h(x-y) = c(x,y)$, the optimal transport map is given by $T = F_\nu^{-1} \circ F_\mu$. Notice that if $c(x,y) = d^p(x,y)$ with $p \geqslant 1$, one has

$$\mathcal{W}_p^p(\mu,\nu) = \int_X |x - F_\nu^{-1} \circ F_\mu(x)|^p \mathrm{d}\mu(x) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)|^p \mathrm{d}s = ||F_\mu^{-1} - F_\nu^{-1}||_{L^p([0,1])}.$$

This formula shows that, through the map $\mu \mapsto F_\mu^{-1}$, the Wasserstein distance is isometric to a linear space equipped with the L^p norm!

2.3 Kantorovich problem

Definition 2.8 (Marginals)**.** The *marginals* of a measure γ on a product space $X \times Y$ are the measures $\pi_{X\#}\gamma$ and $\pi_{Y\#}\gamma$, where $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ are their projection maps, that is

$$\forall (f,g) \in \mathcal{C}(X) \times \mathcal{C}(y),\ \int_{X\times Y} f(x)\mathrm{d}\gamma(x,y) = \int_X f(x)\mathrm{d}\mu(x) \text{ and } \int_{X\times Y} g(y)\mathrm{d}\gamma(x,y) = \int_Y g(y)\mathrm{d}\nu(y).$$

Definition 2.9 (Transport plan)**.** A transport plan between two probabily measures μ, ν on X and Y is a probability measure γ on the product space $X \times Y$ whose marginals are μ and ν. The space of transport plans is denoted $\Pi(\mu,\nu)$, i.e.

$$\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(X \times Y) \mid \pi_{X\#}\gamma = \mu, \quad \pi_{Y\#}\gamma = \nu\}.$$

Note that $\Pi(\mu,\nu)$ is a convex set.

Example 2.10 (Tensor product)**.** Note that the set $\Pi(\mu,\nu)$ of transport plans is never empty, as it contains the measure $\mu \otimes \nu$.

Example 2.11 (Transport plan associated with a map)**.** Let T be a transport map between μ and ν, and define $\gamma_T = (id, T)_\#\mu$. Then, γ_T is a transport plan between μ and ν.

Definition 2.12 (Kantorovich problem)**.** Consider two compact subsets X, Y of $\mathbb{R}^n$ two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a continuous *cost function* $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$. *Kantorovich's problem* is the following optimization problem

$$\mathcal{K}_c(\mu,\nu) := \inf \left\{ \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) \mid \gamma \in \Pi(\mu,\nu) \right\}. \tag{2.3}$$

Remark 2.13. The infimum in Kantorovich problem is less than the infimum in Monge problem. Indeed, consider a transport map satisfying $T_\#\mu = \nu$ and the associated transport plan γ_T. Then, by the change of variable one has

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

thus proving the claim.

Definition 2.14 (Support)**.** Let Ω be a separable metric space. The *support* of a non-negative measure μ is the smallest closed set on which μ is concentrated

$$\operatorname{spt}(\mu) := \bigcap \{A \subseteq \Omega \mid A \text{ closed and } \mu(X \setminus A) = 0\}.$$

A point x belongs to $\operatorname{spt}(\mu)$ iff for every $r > 0$ one has $\mu(\mathrm{B}(x, r)) > 0$.

Theorem 2.15 (Existence)**.** *Let X, Y be two compact subspaces, and $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ be a continuous cost function. Then Kantorovich's problem admits a minimizer.*

Proof. Define $\mathcal{F}(\gamma) := \int c\mathrm{d}\gamma$, then is l.s.c. for the narrow convergence. We just need to show that the set $\Pi(\mu, \nu)$ is compact for narrow topology. Take a sequence $\gamma_n \in \Pi(\mu, \nu)$, since they are probability measures then they are bounded in the dual of $\mathcal{C}(X \times Y)$. Hence, usual weak-$\star$ compactness in dual spaces guarantees the existence of a converging subsequence $\gamma_{n_k} \rightharpoonup \gamma \in \mathcal{P}(X \times Y)$. We need to check that $\gamma \in \Pi(\mu, \nu)$. Fix $\varphi \in \mathcal{C}(X)$, then $\int \varphi(x)\mathrm{d}\gamma_{n_k} = \int \varphi \mathrm{d}\mu$ and by passing to the limit we have $\int \varphi(x)\mathrm{d}\gamma = \int \varphi \mathrm{d}\mu$. This shows that $\pi_{X\#}\gamma = \mu$. The same may be done for π_Y, which concludes the proof. $\square$

The main question is to establish the equality between the infimum in Monge problem and the minimum in Kantorovich problem. Then the following result holds.

Theorem 2.16. *Let $X = Y$ be a compact subset of $\mathbb{R}^d$, $c \in \mathcal{C}(X \times Y)$ and $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. Assume that μ is atomless. Then,*

$$\inf \mathcal{M}_c(\mu, \nu) = \min \mathcal{K}_c(\mu, \nu).$$

3 The dual problem

We now focus on duality theory without enter into details. We firstly find a formal dual problem by exchanging $\inf - \sup$. Let us writing down the constraint $\gamma \in \Pi(\mu, \nu)$ as follows: if $\gamma \in \mathcal{M}_+(X \times Y)$ (we remind that X, Y are compact spaces) we have

$$\Psi := \sup_{\varphi, \psi} \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu - \int_{X \times Y} (\varphi(x) + \psi(y))\mathrm{d}\gamma = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu, \nu), \\ +\infty & \text{otherwise,} \end{cases}$$

where the supremum is taken on $\mathcal{C}_b(X) \times \mathcal{C}_b(Y)$. Thus we can now remove the constraint on γ in $\mathcal{K}_c(\mu, \nu)$

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c\mathrm{d}\gamma + \Psi$$

and by interchanging sup and inf we get

$$\sup_{\varphi, \psi} \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu + \inf_{\gamma \in \mathcal{M}^+(X \times Y)} \int_{X \times Y} (c(x, y) - \varphi(x) - \psi(y))\mathrm{d}\gamma.$$

One can now rewrite the inf in γ as constraint on φ and ψ as

$$\inf_{\gamma \in \mathcal{M}^+(X \times Y)} \int_{X \times Y} (c - \varphi \oplus \psi)\mathrm{d}\gamma = \begin{cases} 0 & \text{if } \varphi \oplus \psi \leqslant c \text{ on } X \times Y \\ -\infty & \text{otherwise} \end{cases},$$

where $\varphi \oplus \psi(x, y) := \varphi(x) + \psi(y)$.

Definition 3.1 (Dual problem)**.** Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a *cost function* $c \in \mathcal{C}(X \times Y)$. The *dual problem* is the following optimization problem

$$\mathcal{D}_c(\mu,\nu) := \sup\left\{\int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu \mid \varphi \in \mathcal{C}_b(X),\ \psi \in \mathcal{C}_b(Y),\ \varphi \oplus \psi \leqslant c\right\} \tag{3.4}$$

Remark 3.2. One trivially has the weak duality inequality $\mathcal{K}_c(\mu,\nu) \geqslant \mathcal{D}_c(\mu,\nu)$. Indeed, denoting

$$L(\gamma,\varphi,\psi) = \int_{X\times Y}(c-\varphi\oplus\psi)\mathrm{d}\gamma) + \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu,$$

one has for any $(\varphi,\psi,\gamma) \in \mathcal{C}_b(X) \times \mathcal{C}_b(Y) \times \mathcal{M}^+(X\times Y)$,

$$\inf_{\tilde{\gamma}\geqslant 0} L(\tilde{\gamma},\varphi,\psi) \leqslant L(\gamma,\varphi,\psi) \leqslant \sup_{\tilde{\varphi},\tilde{\psi}} L(\gamma,\tilde{\varphi},\tilde{\psi})$$

Taking the supremum with respect to (φ,ψ) on the left and the infimum with respect to γ on the right gives $\inf \mathcal{K}_c(\mu,\nu) \geqslant \sup \mathcal{D}_c(\mu,\nu)$. When $\sup \mathcal{D}_c(\mu,\nu) = \inf \mathcal{K}_c(\mu,\nu)$, one talks of *strong duality*. Note that this is independent of whether the infimum and the supremum are attained.

Remark 3.3. 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, $\mathcal{K}_c(\mu,\nu)$ 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 $\mathcal{D}_c(\mu,\nu)$ 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 $\mathcal{K}_c(\mu,\nu) = \mathcal{D}_c(\mu,\nu)$ is that, in this setting, externalization has exactly the same cost as doing the transport by oneself.

Existence

We now focus on the existence of a pair (ψ,ψ) which solves $\mathcal{D}_c(\mu,\nu)$.

Definition 3.4 (c-transform and $\overline{c}$-transform)**.** Given a function $f : x \to \overline{\mathbb{R}}$, we define its c-transform $f^c : Y \to \overline{\mathbb{R}}$ by

$$f^c(y) = \inf_{x\in X} c(x,y) - f(x).$$

We also define the $\overline{c}$-transform of $g : Y \to \overline{\mathbb{R}}$ by

$$g^{\overline{c}}(x) = \inf_{y\in Y} c(x,y) - g(y).$$

We also say that a function ψ on Y is $\overline{c}$-concave if there exists f such that $\psi = f^c$. Notice now that if c is continuous on a compact set, and hence uniformly continuous, then there exists am increasing function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ with $\omega(0) = 0$ such that

$$|c(x,y) - c(x',y')| \leqslant \omega(d_X(x,x') + d_Y(y,y')).$$

If we consider f^c we have that $f^c(y) = \inf_x \tilde{f}_x(y)$ with $\tilde{f}_x(y) = c(x,y) - f(x)$, and the functions $\tilde{f}_x$ satisfy $|\tilde{f}_x(y) - \tilde{f}_x(y)| \leqslant \omega(d_Y(y,y'))$. This implies that f^c actually shares the same continuity modulus of c. It is now quite easy to see that given an admissible pair (φ,ψ) in $\mathcal{D}_c(\mu,\nu)$, one can always replace it with (φ,φ^c) and then $(\varphi^{c\overline{c}},\varphi^c)$ and the constraints are preserved and the integrals increased. The underlying idea of these transformations is actually to improve a maximizing sequence to get a uniform bound on its continuity.

Theorem 3.5. *Suppose that X and Y are compact and $c \in \mathcal{C}(X \times Y)$. Then there exists a pair $(\varphi^{c\overline{c}},\varphi^c)$ which solves $\mathcal{D}_c(\mu,\nu)$.*

Proof. Let us first denote by $\mathcal{J}(\varphi,\psi)$ the following functional

$$J(\varphi,\psi) = \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu,$$

then it is clear that for every constant λ we have $\mathcal{J}(\varphi - \lambda, \psi + \lambda) = \mathcal{J}(\varphi,\psi)$. Given now a maximising sequence (φ_n,ψ_n) we can improve it by means of the c- and $\overline{c}$-transform obtaining a new one $(\varphi_n^{c\overline{c}},\varphi_n^c)$. Notice that by the consideration above the sequences $\varphi_n^{c\overline{c}}$ and φ_n^c are uniformly equicontinuous. Since φ_n^c is continuous on a compact set we can always subtract its minimum and assume that $\min_Y \varphi_n^c = 0$. This implies that the sequence φ_n^c is also equibounded as $0 \leqslant \varphi_n^c \leqslant \omega(\mathrm{diam}(Y))$. We also deduce uniform bounds on $\varphi_n^{c\overline{c}}$ as $\varphi_n^{c\overline{c}} = \inf_Y c(x,y) - \varphi^c(y)$. This let us apply Ascoli-Arzela's theorem and extract two uniformly converging subsequences $\varphi_{n_k}^{c\overline{c}} \to \overline{\varphi}$ and $\varphi_{n_k}^c \to \overline{\psi}$ where the pair $(\overline{\varphi},\overline{\psi})$ satisfies the inequality constraint. Moreover, since $(\varphi_n^{c\overline{c}},\varphi_n^c)$ is a maximising sequence we get that the pair $(\overline{\varphi},\overline{\psi})$ is optimal. now one can apply again the c- and $\overline{c}$-transforms obtaining an optimal pair of the form $(\overline{\varphi}^{c\overline{c}},\overline{\varphi}^c)$. □

3.1 Duality via the discrete case

The case of discrete optimal transport

We start with the case of finite discrete probability measures, which is important because:

- It often comes up in applications (e.g. optimal matching in economy);
- Numerical methods for the continuous case often resort to discretization;
- It is a convenient way to study the general case, through density arguments.

Proposition 3.6 (Duality, discrete case)**.** *If μ and ν are finitely supported, then $\mathcal{D}_c(\mu,\nu) = \mathcal{K}_c(\mu,\nu)$.*

Proof. Let us write $\mu = \sum_{i=1}^m \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$ where all μ_i and ν_j are strictly positive. Consider the linear program

$$\mathcal{L}_c(\mu,\nu) := \min\Big\{ \sum_{i,j} c(x_i,y_j)\gamma_{i,j} \mid \gamma_{i,j} \geqslant 0,\ \sum_j \gamma_{i,j} = \mu_i,\ \sum_i \gamma_{i,j} = \nu_j \Big\}.$$

which admits a solution that we denote γ. By linear programming duality (which is standard in the finite dimensional case, see e.g. [4, Sec. 5.2] or [16, Sec. 37.3]), we have strong duality

$$\mathcal{L}_c(\mu,\nu) = \max\Big\{ \sum_i \varphi_i\mu_i + \sum_j \psi_j\nu_j \mid \varphi_i + \psi_j \leqslant c(x_i,y_j) \Big\}$$

and at optimality $\gamma_{i,j}(c_{i,j} - \varphi_i - \psi_j) = 0$ (the complementary slackness in Karush-Kuhn-Tucker theorem). Let us now build a pair (φ, ψ) of functions which is feasible for the dual problem and that takes the value (φ_i, ψ_j) at (x_i, y_j). For this purpose, we introduce

$$\psi(y) = \begin{cases} \psi_i & \text{if } y = y_i, \\ -\infty & \text{otherwise}, \end{cases}$$

and let $\varphi = \psi^{\bar{c}} \in \mathcal{C}(X)$. For $i_0 \in [n]$, there exists $j_0 \in [n]$ such that $\gamma_{i_0,j_0} > 0$ and thus, by complementary slackness, $\varphi_{i_0} + \psi_{j_0} = c(x_{i_0}, y_{j_0})$ and thus

$$\varphi(x_{i_0}) = \inf_{y \in Y} \Big(c(x_{i_0}, y) - \psi(y) \Big) = \min_{j \in [n]} \Big(c(x_{i_0}, y_j) - \psi_j \Big) = c(x_{i_0}, y_{j_0}) - \psi_{j_0} = \varphi_{i_0}.$$

Similarly, one can show that $\varphi^c(y_j) = \psi_j$ for all $j \in [n]$. Finally, we define $\gamma = \sum_{i,j} \gamma_{i,j} \delta_{(x_i, y_j)} \in \Pi(\mu, \nu)$. Since we have built admissible primal γ and dual (φ, ψ) variables for which the primal and dual objective agree, this concludes the proof. □

Density of discrete measures

In order the prove the general case, we will use the density of discrete measures for the weak topology and a stability property of optimal dual and primal solutions.

Lemma 3.7 (Density of discrete measures)**.** *Let X be a compact space and $\mu \in \mathcal{P}(X)$. Then, there exists a sequence of finitely supported probability measures weakly converging to μ.*

Proof. By compactness, for any $\epsilon > 0$, there exists N points $x_1, \ldots, x_n$ such that $X \subset \bigcup_i B(x_i, \epsilon)$. We introduce the partition $K_1, \ldots, K_n$ of X defined recursively by $K_i = B(x_i, \epsilon) \setminus K_1 \cup \cdots \cup K_{i-1}$ and

$$\mu_\epsilon := \sum_{i=1}^n \mu(K_i) \delta_{x_i}.$$

To prove weak convergence of μ_ϵ to μ as $\epsilon \to 0$, take $\varphi \in \mathcal{C}(X)$. By compactness of X, φ admits a modulus of continuity ω, i.e. an increasing function satisfying $\lim_{t\to 0} \omega(t) = 0$ and $|\varphi(x) - \varphi(y)| \leqslant \omega(\mathrm{dist}(x, y))$. Using that $\mathrm{diam}(K_i) \leqslant \epsilon$, we get

$$\Big| \int \varphi \mathrm{d}\mu - \int \varphi \mathrm{d}\mu_\epsilon \Big| \leqslant \sum_{i=1}^n \int_{K_i} |\varphi(x) - \varphi(x_i)| \mathrm{d}\mu(x) \leqslant \omega(\epsilon).$$

We deduce that μ_ϵ weakly converges to μ (remember that for measures on a compact space, narrow, weak and weak* topologies are the same). □

Note that we even have weak density in $\mathcal{P}(X)$ of empirical measures, that is measures of the form $\frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ for $n \in \mathbb{N}^*$ and $x_i \in X$. Indeed, take $x_1, \ldots, x_n$ independent random variables with distribution μ. Then the uniform law of large numbers (a.k.a. Varadarajan's theorem) states that $\frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ weakly converges to μ with probability 1.

Strong duality for the general case

Theorem 3.8 (Duality, general case)**.** *Let X, Y be compact metric spaces and $c \in \mathcal{C}(X \times Y)$. Then $\mathcal{K}_c(\mu, \nu) = \mathcal{D}_c(\mu, \nu)(\mu, \nu)$.*

Proof. By Lemma 3.7, there exists a sequence $\mu_k \in \mathcal{P}(X)$ (resp. $\nu_k \in \mathcal{P}(Y)$) of finitely supported measures which converge weakly to μ (resp. ν). By Proposition 3.6 and its proof, there exists for all k, γ_k and (φ_k, φ_k^c) with φ_k c-concave which are optimal primal-dual solutions to $\mathcal{K}_c(\mu_k, \nu_k)$ and such that γ_k is supported on the set

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

Adding a constant if necessary, we can also assume that $\varphi_k(x_0) = 0$ for some point $x_0 \in X$. As in the previous lecture, we see that $\{\varphi_k\}$ and $\{\varphi_k^c\}$ are uniformly continuous and bounded so that by Ascoli-Arzelà theorem converge uniformly to some (φ, ψ) up to a subsequence. We easily have that $\varphi \oplus \psi \leqslant c$, so (φ, ψ) is feasible for the dual problem (in fact uniform convergence implies that $\psi = \varphi^c$, although we will not use this fact here).

By weak compactness of $\mathcal{P}(X \times Y)$, we can assume that the sequence γ_k weakly converges to $\gamma \in \Pi(\mu, \nu)$. Moreover, by Lemma 3.9, every pair $(x, y) \in \mathrm{spt}(\gamma)$ can be approximated by a sequence of pairs $(x_k, y_k) \in \mathrm{spt}(\gamma_k)$ with $\lim_{k\to\infty}(x_k, y_k) = (x, y)$. One has $c(x_k, y_k) = \varphi_k(x_k) + \varphi_k^c(y_k)$, which gives at the limit $c(x, y) = \varphi(x) + \psi(y)$. Thus we have

$$\mathcal{K}_c(\mu, \nu) \leqslant \int c\mathrm{d}\gamma = \int \Big(\varphi(x) + \psi(y)\Big)\mathrm{d}\gamma(x, y) = \int \varphi\mathrm{d}\mu + \int \psi\mathrm{d}\nu \leqslant \mathcal{D}_c(\mu, \nu)$$

Since we already know that $\mathcal{D}_c(\mu, \nu) \leqslant \mathcal{K}_c(\mu, \nu)$ this is sufficient to conclude. □

Lemma 3.9. *If μ_n converges weakly to μ, then for any point $x \in \mathrm{spt}(\mu)$ there exists a sequence $x_n \in \mathrm{spt}(\mu_n)$ converging to x.*

Proof. Consider $x \in \mathrm{spt}(\mu)$. For any $k \in \mathbb{N}$, consider the function $\varphi_k(z) = \max\{0, 1 - k\mathrm{dist}(x, z)\}$ which is continuous. Then

$$\lim_{n\to\infty} \int \varphi_k \mathrm{d}\mu_n = \varphi_k \mathrm{d}\mu > 0.$$

Thus, there exists n_k such that for any $n \geqslant n_k$, $\int \varphi_k \mathrm{d}\mu_n > 0$. This implies the existence of a sequence $(x_n^{(k)})_n \in X$ such that $x_n^{(k)} \in \mathrm{spt}(\mu_n)$ and $\mathrm{dist}(x_n^{(k)}, x) \leqslant 1/k$ for $n \geqslant n_k$. By a diagonal argument, we build the sequence $x_n = x_n^{k_n}$ where $k_n = \max\{k \mid k = 0 \text{ or } n \geqslant n_k\}$. Since by construction $k_n \to \infty$, we have $x_n \to x$. □

3.2 Optimality conditions and transport maps

Let us write down three important properties that follow from our previous results. First, remark that the proof of Theorem 3.8 can be used to prove the following stability property (the modifications are left as an exercise).

Proposition 3.10 (Stability)**.** *Let X, Y be compact metric spaces. 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 and $(c_k)_{k\in\mathbb{N}}$ in $\mathcal{C}(X \times Y)$ converging uniformly to c.*

- *If γ_k is a minimizer for $\mathcal{K}_{c_k}(\mu_k, \nu_k)$ then, up to subsequences, (γ_k) converges weakly to a minimizer for $\mathcal{K}_c(\mu, \nu)$.*
- *Let $(\varphi_k, \varphi_k^{c_k})$ be a maximizer for $\mathcal{D}_{c_k}(\mu_k, \nu_k)$ and be such that φ_k is c_k-concave and $\varphi_k(x_0) = 0$. Then, up to subsequences, $(\varphi_k, \varphi_k^{c_k})$ converges uniformly to (φ, φ^c) a maximizer for $\mathcal{D}_c(\mu, \nu)$ with φ c-concave satisfying $\varphi(x_0) = 0$.*

Let us emphasize on the optimality conditions, which are just a continuous version of complementary slackness.

Proposition 3.11 (Optimality conditions)**.** *For $\gamma \in \Pi(\mu, \nu)$ and $(\varphi, \psi) \in \mathcal{C}(X) \times \mathcal{C}(Y)$ satisfying $\varphi \oplus \psi \leqslant c$, the following are equivalent:*

(i) $\varphi(x) + \psi(y) = c(x, y)$ *holds γ-almost everywhere.*

(ii) γ *is a minimizer of* $\mathcal{K}_c(\mu, \nu)$, (φ, ψ) *is a maximizer of* $\mathcal{D}_c(\mu, \nu)$.

Proof. Assuming (i), we have

$$\mathcal{K}_c(\mu, \nu) \leqslant \int c \mathrm{d}\gamma = \int \Big(\varphi(x) + \psi(y)\Big) \mathrm{d}\gamma(x, y) = \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu \leqslant \mathcal{D}_c(\mu, \nu)$$

Since we already know that $\mathcal{D}_c(\mu, \nu) \leqslant \mathcal{K}_c(\mu, \nu)$, this implies (ii). To show $(ii) \Rightarrow (i)$, notice that Theorem 3.8 and (ii) imply

$$0 = \int c(x, y) \mathrm{d}\gamma(x, y) - \int \varphi(x) + \psi(y) \mathrm{d}\gamma(x, y) = \int \Big(c(x, y) - \varphi(x) - \psi(y)\Big) \mathrm{d}\gamma(x, y).$$

Since the last integrand is nonnegative, it must vanish γ-almost everywhere. □

Another useful notion attached to optimal transport solutions is that of cyclical monotonicity.

Definition 3.12 (Cyclical monotonicity)**.** A set $S \subset X \times Y$ is said c-cyclically monotone if for any $n \in \mathbb{N}^*$ and $(x_i, y_i)_{i=1}^n \in S^n$, it holds

$$\sum_{i=1}^n c(x_i, y_i) \leqslant \sum_{i=1}^n c(x_i, y_{i+1}) \tag{3.5}$$

with the convention $y_{n+1} = y_1$.

Note that Eq. (3.5) is equivalent to requiring $\sum_{i=1}^n c(x_i, y_i) \leqslant \sum_{i=1}^n c(x_i, y_{\sigma(i)})$ for any permutation σ of $\{1, \dots, n\}$, since one can chose the ordering freely when selecting the n points $(x_i, y_i)_{i=1}^n \in S^n$.

Proposition 3.13. *Let X, Y be compact metric spaces, $c \in \mathcal{C}(X \times Y)$ and $\gamma \in \Pi(\mu, \nu)$ an optimal transport plan between $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then* $\mathrm{spt}(\gamma)$ *is c-cyclically monotone.*

This result is rather direct in the discrete case and can also be proved without duality in the general case but our duality results lead to a straighforward proof.

Proof. Let $(x_i, y_i)_{i=1}^n$ be n points in $\mathrm{spt}(\gamma)$. By Prop. 3.11, we know that there exists (φ, ψ) such that $\varphi(x_i) + \psi(y_j) \leqslant c(x_i, y_j)$ for all i, j and such that $\varphi(x_i) + \psi(y_i) = c(x_i, y_i)$ for all i. Thus

$$\sum_i c(x_i, y_{i+1}) - \sum_i c(x_i, y_i) \geqslant \sum_i \Big(\varphi(x_i) + \psi(y_{i+1})\Big) - \sum_i \Big(\varphi(x_i) + \psi(y_i)\Big) = 0.$$

□

Remark 3.14. The cautious reader might have noticed that Prop. 3.11 only guarantees that $\gamma\{(x, y) \in X \times Y \,;\, \varphi(x) + \psi(y) < c(x, y)\} = 0$ (*) while we used a different property. But (*) and the continuity of c, φ and ψ implies that if $\varphi(x) + \psi(y) < c(x, y)$ then there exists a nonempty open ball around (x, y) with 0 mass under γ, i.e. $(x, y) \notin \mathrm{spt}(\gamma)$ thus $\varphi(x) + \psi(y) = c(x, y)$ for all $(x, y) \in \mathrm{spt}(\gamma)$ (which is the property use above).

Remark 3.15 (see Thm 5.10 [21])**.** A stronger property in fact holds: any c-cyclically monotonous set is contained in a set of the form $\{(x,y) \in X \times Y\ ; \varphi(x) + \varphi^c(y) = c(x,y)\}$ for some c-concave function φ. This implies that any $\gamma \in \Pi(\mu,\nu)$ such that $\mathrm{spt}(\gamma)$ is c-cyclically monotone is optimal.

We recall the following characterization of solutions to Monge's problem.

Lemma 3.16. *Let $\gamma \in \Pi(\mu,\nu)$ and $T : X \to Y$ measurable be such that $\gamma(\{(x,y) \in X \times Y \mid T(x) \neq y\}) = 0$. Then, $\gamma = \gamma_T := (\mathrm{id}, T)_{\#}\mu$.*

If γ is a minimizer for $\mathcal{K}_c(\mu,\nu)$ and (φ, φ^c) is a maximizer for $\mathcal{D}_c(\mu,\nu)$, we know that $\varphi \oplus \varphi^c = c$ γ-almost everywhere. To build a solution to Monge's problem, it is therefore sufficient to show that the set $\{\varphi \oplus \varphi^c = c\}$ is contained in the graph of a function. This will be possible for the following class of costs:

Definition 3.17 (Twisted cost)**.** A cost function $c \in \mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^d)$ is said to satisfy the *twist condition* if

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

where $\nabla_x c(x_0,y)$ denotes the gradient of $x \mapsto c(\cdot, y)$ at $x = x_0$. Given $x, v \in \mathbb{R}^d$, we denote $y_c(x_0, v)$ the unique point such that $\nabla_x c(x_0, y_c(x_0,v)) = v$.

Theorem 3.18. *Let $c \in \mathrm{C}^1(\mathbb{R}^d \times \mathbb{R}^d)$ be a twisted cost, let $X, Y \subset \mathbb{R}^d$ be compact subsets and $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Assume that μ is absolutely continuous with respect to the Lebesgue measure. Then, there exists a c-concave function φ that is differentiable almost everywhere such that $\nu = T_{\#}\mu$ where $T(x) = y_c(x, \nabla\varphi(x))$. Moreover, the only optimal transport plan between μ and ν is γ_T.*

Proof. Enlarging X if necessary, we may assume that $\mathrm{spt}(\mu)$ is contained in the interior of X. First note that by compactness of $X \times Y$ and since c is $\mathcal{C}^1$, the cost c is Lipschitz continuous on $X \times Y$. Take (φ, φ^c) a maximizing pair for $\mathcal{D}_c(\mu,\nu)$ with φ c-concave. Since $\varphi(x) = \min_{y \in Y} c(x,y) - \varphi^c(y)$ we see that φ is Lipschitz. By Rademacher's theorem[1], φ is thus differentiable Lebesgue almost everywhere and, since μ is assumed absolutely continuous, it is differentiable on a set $B \subset \mathrm{spt}(\mu)$ with $\mu(B) = 1$.

Consider an optimal transport plan $\gamma \in \Pi(\mu,\nu)$. For every pair of points $(x_0, y_0) \in \mathrm{spt}(\gamma) \cap (B \times Y)$, we have

$$\varphi^c(y_0) \leqslant c(x, y_0) - \varphi(x),\ \forall x \in X$$

with equality at $x = x_0$, so that x_0 minimizes the function $x \mapsto c(x, y_0) - \varphi(x)$. Since $x_0 \in \mathrm{spt}(\mu)$ and x_0 belongs to the interior of X, one necessarily has $\nabla\varphi(x_0) = \nabla_x c(x_0, y_0)$. Then, by the twist condition, one necessarily has $y_0 = y_c(x_0, \nabla\varphi(x_0))$. This shows that any optimal transport plan γ is supported on the graph of the map $T : x \in B \mapsto y_c(x_0, \nabla\varphi(x_0))$, and $\gamma = \gamma_T$ by Lemma 3.16. ☐

4 Back to discret Optimal Transport

We now consider the optimal transport problems between probability measures on two finite sets X and Y with, for simplicity, both of cardinality N and we set

$$\mu = \sum_{x \in X} \mu_x \delta_x \qquad \nu = \sum_{y \in Y} \nu_y \delta_y.$$

[1] https://en.wikipedia.org/wiki/Rademacher%27s_theorem

Definition 4.1 (Discrete OT)**.** The discrete Optimal transport problem between two given measures μ and ν and a given cost function $c : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ is the following minimization problem

$$\inf \left\{ \sum_{x \in X} \sum_{y \in Y} \gamma_{xy} c(x, y) \mid \gamma \in \Pi(\mu, \nu) \right\}, \tag{4.6}$$

where the set of admissible couplings is now defined as

$$\Pi(\mu, \nu) := \{\gamma \in X \times Y \mid \gamma_{xy} \geqslant 0, \sum_{y \in Y} \gamma_{xy} = \mu_x \; \forall x \in X, \sum_{x \in X} \gamma_{xy} = \nu_y \; \forall y \in Y\}.$$

Unfortunately, this linear programming problem has complexity $O(N^3)$ which actually means that it is infeasible for large N. A way to overcome this difficulty is by means of the **Entropic Regularization** which provides an approximation of Optimal Transport with lower computational complexity and easy implementation.

References: Entropic regularisation of Optimal Transport is a very active research field. We refer the interested reader to [2, 7, 12, 15, 8] and the citations therein. We also remark that these notes are inspired by the graduate class on Numerical Optimal Transport given by F.-X. Vialard [19] as well as the one given by M. Nutz [14].

5 The Entropic Optimal Transport

5.1 The discrete case

We start from the primal formulation of the optimal transport problem, but instead of imposing the constraints $\gamma_{xy} \geqslant 0$, we add a term $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$, involving the (opposite of the) entropy

$$e(r) = \begin{cases} r(\log r - 1) & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ +\infty & \text{if } r < 0 \end{cases}$$

More precisely, given a parameter $\varepsilon > 0$ we consider

$$\boxed{\mathcal{K}_c^\varepsilon(\mu, \nu) = \inf \left\{ \langle \gamma | c \rangle + \varepsilon \, \mathrm{Ent}(\gamma) \mid \gamma \in X \times Y, \sum_{y \in Y} \gamma_{xy} = \mu_x, \sum_{x \in X} \gamma_{xy} = \nu_y \right\}, \quad (5.7)}$$

where $\langle \gamma | c \rangle = \sum_{x,y} \gamma_{xy} c(x, y)$ and $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$.

Theorem 5.1. *The problem $\mathcal{K}_c^\varepsilon(\mu, \nu)$ has a unique solution $\gamma^\star$, which belongs to $\Pi(\mu, \nu)$. Moreover, if $\min(\min_{x \in X} \mu_x, \min_{y \in Y} \nu_y) > 0$ then*

$$\gamma_{x,y} > 0 \; \forall (x, y) \in X \times Y.$$

Before introducing the duality, it is important to state the following convergence result in ε.

Theorem 5.2 (Convergence in ε)**.** *The unique solution γ_ε to* (5.7) *converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is*

$$\gamma_\varepsilon \xrightarrow[\varepsilon \to 0]{} \operatorname{argmin}\left\{\operatorname{Ent}(\gamma) \mid \gamma \in \Pi(\mu,\nu),\ \langle \gamma | c\rangle = \mathcal{K}_c^0(\mu,\nu)\right\}, \tag{5.8}$$

where $\mathcal{K}_c^0(\mu,\nu)$ denotes the unregularized problem.

Proof. Consider a sequence $(\varepsilon_k)_k$ such that $\varepsilon_k \to 0$ and $\varepsilon_k > 0$ and denote γ_k the solution to (5.7) with $\varepsilon = \varepsilon_k$. Since $\Pi(\mu,\nu)$ is bounded and close we can extract a converging subsequence $\gamma_k \to \gamma^\star \in \Pi(\mu,\nu)$. Take now any optimal γ for the unregularized problem then by optimality of γ_k and γ one has

$$0 \leqslant \langle \gamma_k | c\rangle - \langle \gamma | c\rangle \leqslant \varepsilon_k(\operatorname{Ent}(\gamma) - \operatorname{Ent}(\gamma_k)). \tag{5.9}$$

Since $\operatorname{Ent}(\cdot)$ is continuous, by taking the limit $k \to +\infty$ in (5.9) we get $\langle \gamma^\star | c\rangle = \langle \gamma | c\rangle$. Furthermore, dividing by ε_k and taking the limit we obtain that $\operatorname{Ent}(\gamma) \geqslant \operatorname{Ent}(\gamma^\star)$ showing that $\gamma^\star$ is a solution to the minimization problem in (5.8). By strict convexity of Ent the optimization problem (5.8) has a unique solution and the whole sequence is converging to $\gamma^\star$. □

We want now to derive formally the dual problem. For this purpose we introduce the Lagrangian associated to (5.7)

$$\begin{aligned}\mathcal{L}(\gamma,\varphi,\psi) := \sum_{x,y} \gamma_{xy} c(x,y) + \varepsilon e(\gamma_{xy}) + \sum_{x\in X} \varphi(x)\left(\mu_x - \sum_{y\in Y}\gamma_{xy}\right) \\ + \sum_{y\in Y}\psi(y)\left(\nu_y - \sum_{y\in Y}\gamma_{xy}\right),\end{aligned} \tag{5.10}$$

where $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ are the Lagrange multipliers. Then,

$$\mathcal{K}_c^\varepsilon(\mu,\nu) = \inf_\gamma \sup_{\varphi,\psi} \mathcal{L}(\gamma,\varphi,\psi),$$

and the dual problem is obtained by interchanging the infimum and the supremum :

$$\begin{aligned}\mathcal{D}_c^\varepsilon(\mu,\nu) = \sup_{\varphi,\psi}\min_\gamma \sum_{x,y} \gamma_{xy}(c(x,y) - \psi(y) - \varphi(x) + \varepsilon(\log(\gamma_{xy}) - 1)) + \\ \sum_{x\in X}\varphi(x)\mu_x + \sum_{y\in Y}\psi(y)\nu_y.\end{aligned} \tag{5.11}$$

Taking the derivative with respect to γ_{xy}, we find that for a given φ,ψ, the optimal γ must satisfy:

$$\begin{aligned} c(x,y) - \psi(y) - \varphi(x) + \varepsilon \log(\gamma_{xy}) = 0 \\ \text{i.e. } \gamma_{xy} = \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}\right)\end{aligned} \tag{5.12}$$

Putting these values in the definition of $\mathcal{D}_c^\varepsilon(\mu,\nu)$ gives

$$\mathcal{D}_c^\varepsilon(\mu,\nu) = \sup_{\varphi,\psi} \Phi_\varepsilon(\varphi,\psi) \text{ with} \tag{5.13}$$

$$\Phi_\varepsilon(\varphi,\psi) := \sum_{x\in X} \varphi(x)\mu_x + \sum_{y\in Y} \psi(y)\nu_y - \sum_{x,y} \varepsilon \exp\Big(\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}\Big)$$

Note that thanks to the relation (5.12), one can recover a solution to the primal problem from the dual one. This is true because, unlike the original linear programming formulation of the optimal transport problem, the regularized problem (5.7) is smooth and strictly convex. The following duality result holds

Theorem 5.3 (Strong duality). *Strong duality holds and the maximum in the dual problem is attained, that is $\exists\varphi,\psi$ such that*

$$\mathcal{K}_c^\varepsilon(\mu,\nu) = \mathcal{D}_c^\varepsilon(\mu,\nu) = \Phi_\varepsilon(\varphi,\psi).$$

Corollary 5.4. *If (φ,ψ) is the solution to* (5.13)*, then the solution $\gamma^\star$ to* (5.7) *is given by*

$$\gamma_{x,y} = \exp\Big(\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}\Big)$$

Notice now that the optimal coupling γ can be written as

$$\gamma_{x,y} = D_\varphi e^{\frac{-c(x,y)}{\varepsilon}} D_\psi,$$

where D_φ and D_ψ are the diagonal matrices associated to $e^{\varphi/\varepsilon}$ and $e^{\psi/\varepsilon}$, respectively. The problem is now similar to a matrix scaling problem

Definition 5.5 (Matrix scaling problem). Let $K \in \mathbb{R}^{N\times N}$ be a matrix with positive coefficients. Find D_ψ and D_ψ positive diagonal matrices in $K \in \mathbb{R}^{N\times N}$ such that $D_\varphi K D_\psi$ is doubly stochastic, that is sum along each row and each column is equal to 1.

Remark 5.6. Uniqueness fails since if (D_φ, D_ψ) is a solution then so is $(cD_\varphi, \frac{1}{c}D_\psi)$ for every $c \in \mathbb{R}_+$.

The matrix scaling problem can be easily solved by using an iterative algorithm, known as Sinkhorn-Knopp algorithm, which simply alternates updating D_φ and D_ψ in order to match the marginal constraints (a vector $\mathbf{1}_N$ of ones in this simple case).

Algorithm 1 Sinkhorn-Knopp algorithm for the matrix scaling problem

```
function Sinkhorn-Knopp(K)
    D_φ^0 ← 1_N, D_ψ^0 ← 1_N
    for 0 ⩽ k < k_max do
        D_φ^{k+1} ← 1_N./(K D_ψ^k)
        D_ψ^{k+1} ← 1_N./(K^T D_φ^{k+1})
    end for
end function
```

where ./ stand for the element-wise division. Denoting by $(K_\varepsilon)_{x,y} = e^{\frac{-c(x,y)}{\varepsilon}}$ the algorithm takes the form 2 for the regularized optimal transport problem.

Algorithm 2 Sinkhorn-Knopp algorithm for the regularised optimal transport problem

function SINKHORN-KNOPP(K_ε, μ, ν)
  $D_\varphi^0 \leftarrow \mathbf{1}_X,\ D_\psi^0 \leftarrow \mathbf{1}_Y$
  for $0 \leqslant k < k_{\max}$ **do**
    $D_\varphi^{k+1} \leftarrow \mu./(KD_\psi^k)$
    $D_\psi^{k+1} \leftarrow \nu./(K^T D_\varphi^{k+1})$
  end for
end function

Notice that one can recast the regularized OT in the framework of bistochastic matrix scaling by replacing the kernel $e^{\frac{-c(x,y)}{\varepsilon}}$ with $(K_\varepsilon)_{x,y} = \operatorname{diag}(\mu)e^{\frac{-c(x,y)}{\varepsilon}}\operatorname{diag}(\nu)$, where $\operatorname{diag}(\mu)$ ($\operatorname{diag}(\nu)$) denotes the diagonal matrix with the vector μ (ν) as main diagonal. In this case the problem (5.7) can be re-written as

$$\mathcal{K}_c^\varepsilon(\mu,\nu) = \inf\left\{ \langle\gamma|c\rangle + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma \in X\times Y,\ \sum_{y\in Y}\gamma_{xy} = \mu_x,\ \sum_{x\in X}\gamma_{xy} = \nu_y \right\}, \tag{5.14}$$

where $\mathcal{H}(\rho|\mu) := \sum_x \rho_x(\log(\frac{\rho_x}{\mu_x}) - 1)$ is the relative entropy or the Kullback-Leibler divergence.

> **Good to know:** one can easily recast the regularized OT in the continuous framework as follows
>
> $$\mathcal{K}_c^\varepsilon(\mu,\nu) = \inf\left\{ \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma\in\Pi(\mu,\nu)\right\}, \tag{5.15}$$
>
> where
>
> $$\mathcal{H}(\rho|\pi) = \begin{cases} \displaystyle\int_{X\times Y}\Big(\log\Big(\frac{\mathrm{d}\rho(x,y)}{\mathrm{d}\pi(x,y)}\Big) - 1\Big)\mathrm{d}\rho(x,y), & \text{if } \rho \ll \pi \\ +\infty, & \text{otherwise,}\end{cases}$$
>
> and the marginals μ, ν are probability measures on the compact metric spaces X and Y, respectively. This problem is often referred to as the *static Schrödinger problem* [12] since it was initially considered by Schrödinger in statistical physics. Once again, under mild assumptions on the cost functions, one can prove that the regularized problem converges to original one as $\varepsilon \to 0$; see [6, 11].

6 The convergence of Sinkhorn: the discrete setting

We focus now on the global convergence analysis of the Sinkhorn algorithm in the discrete setting by using the *Hilbert* projective metric on $\mathbb{R}^n_{+,\star}$ (positive vectors).

Definition 6.1 (Hilbert projective metric)**.** The *Hilbert* projective metric on $\mathbb{R}^n_{+,\star}$ is defined as

$$\forall (u,v)\in(\mathbb{R}^n_{+,\star})^2,\ d_H(u,v) := ||\log(u) - \log(v)||_V,$$

Where

$$||x||_V = \max_i x_i - \min_i x_i.$$

Before stating the convergence result we need the following fundamental theorem, which shows that a positive matrix is a strict contraction on the cone of positive vector

Theorem 6.2 ([3, 17])**.** *Let $K \in \mathbb{R}^{n\times n}_{+,\star}$, then for $(u,v) \in (\mathbb{R}^n_{+,\star})^2$*

$$d_H(Ku, Kv) \leqslant \lambda(K) d_H(u,v), \tag{6.16}$$

where

$$\lambda(K) = \frac{\sqrt{\eta(K)} - 1}{\sqrt{\eta(K)} + 1} < 1$$

and

$$\eta(K) = \max_{i,j,kl} \frac{K_{ik}K_{jl}}{K_{jk}K_{il}}.$$

We have then the following convergence result (we use the same notations as in 2)

Theorem 6.3 ([9])**.** *One has $(D^k_\varphi, D^k_\psi) \to (D^\star_\varphi, D^\star_\psi)$ and*

$$d_H(D^k_\varphi, D^\star_\varphi) = O(\lambda(K)^{2k}),\ d_H(D^k_\psi, D^\star_\psi) = O(\lambda(K)^{2k}), \tag{6.17}$$

where $D^\star_\varphi, D^\star_\psi$ are the optimal solutions. Moreover,

$$d_H(D^k_\varphi, D^\star_\varphi) \leqslant \frac{d_H(\gamma^k \mathbf{1}_n, \mu)}{1-\lambda(K)^2}, \tag{6.18}$$

$$d_H(D^k_\psi, D^\star_\psi) \leqslant \frac{d_H(\gamma^k \mathbf{1}_n, \nu)}{1-\lambda(K)^2}, \tag{6.19}$$

where $\gamma^k = \mathrm{diag}(D^k_\varphi) K \,\mathrm{diag}(D^k_\psi)$. Last, one has

$$|| \log(\gamma^k) - \log(\gamma^\star)||_\infty \leqslant d_H(D^k_\varphi, D^\star_\varphi) + d_H(D^k_\psi, D^\star_\psi). \tag{6.20}$$

where $\gamma^\star$ is the unique solution to (5.7)*.*

Proof. Notice that for any $(u,v) \in (\mathbb{R}^n_{+,\star})^2$, on has

$$d_H(u,v) = d_H(u/v, \mathbf{1}_n) = d_H(\mathbf{1}_n/u, \mathbf{1}_n/v).$$

This shows that

$$d_H(D^k_\varphi, D^\star_\varphi) = d_H(\frac{\mu}{KD^k_\psi}, \frac{\mu}{KD^\star_\psi}) = d_H(KD^k_\psi, KD^\star_\psi) \leqslant \lambda(K) d_H(D^k_\psi, D^\star_\psi),$$

where we used Theorem 6.2. This shows (6.17). By using triangular inequality we have

$$\begin{aligned} d_H(D^k_\varphi, D^\star_\varphi) &\leqslant d_H(D^{k+1}_\varphi, D^k_\varphi) + d_H(D^{k+1}_\varphi, D^\star_\varphi) \\ &\leqslant d_H(\frac{\mu}{KD^k_\psi}, D^k_\varphi) + \lambda(K) d_H(D^k_\varphi, D^\star_\varphi) \\ &= d_H(\mu, D^k_\varphi \odot (KD^k_\psi)) + \lambda(K)^2 d_H(D^k_\varphi, D^\star_\varphi) \\ &= d_H(\mu, \gamma^k \mathbf{1}_n) + \lambda(K)^2 d_H(D^k_\varphi, D^\star_\varphi), \end{aligned}$$

where $\odot$ denotes the element wise multiplication. (6.19) can be proved in an analogous way. (6.20) is trivial. □

6.1 The convergence of Sinkhorn in the continuous setting

As presented in Lecture 1, the existence of Kantorovich potentials for the standard Optimal Transport problem can be proven by standard compactness arguments. By using similar arguments we show existence for the regularized dual problem (and convergence of Sinkhorn at the same time) in the continuous framework. We firstly recall that a coordinate ascent algorithm on a fucntion of two variables $f(x, y)$ can written as

$$\begin{aligned} y_{k+1} &= \operatorname{argmax}_y f(x_k, y), \\ x_{k+1} &= \operatorname{argmax}_x f(x, y_{k+1}). \end{aligned}$$

The Sinkhorn algorithm is actually a coordinate ascent algorithm: the main idea is indeed to maximize $\Phi_\varepsilon(\varphi, \psi)$ by maximizing alternatively in φ and ψ. From now on we assume for simplicity that $X = Y$ are compact and c is a continuous cost function.

Proposition 6.4. *The dual problem to* (5.15) *reads as*

$$\mathcal{D}_c^\varepsilon(\mu, \nu) = \sup\{\Phi_\varepsilon(\varphi, \psi) \mid \varphi, \psi \in \mathcal{C}_0(X)\}, \tag{6.21}$$

where

$$\begin{aligned} \Phi_\varepsilon(\varphi, \psi) := &\int_X \varphi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y) \\ &- \varepsilon \int_{X\times Y} \exp\Big(\frac{\varphi(x) + \psi(y) - c(x, y)}{\varepsilon}\Big) \mathrm{d}\mu \otimes \mathrm{d}\nu(x, y). \end{aligned}$$

It is strictly concave w.r.t. each argument φ and ψ and strictly concave w.r.t. $\varphi(x)+\psi(y)$. It is also Fréchet differentiable for the $(\mathcal{C}_0, \|\cdot\|_\infty)$ topology. Furthermore, if a maximizer exists it is unique up to a constant, that is $\Phi_\varepsilon(\varphi, \psi) = \Phi_\varepsilon(\varphi + C, \psi - C)$ for every $C \in \mathbb{R}$.

Proof. We leave the proof as an exercice. □

Proposition 6.5. *The maximization of $\Phi_\varepsilon(\varphi, \psi)$ w.r.t. each variable can be made explicit, and the Sinkhorn algorithm can be defined as*

$$\varphi_{k+1}(x) = -\varepsilon \log\left(\int_X \exp\Big(\frac{1}{\varepsilon}(\psi_k(y) - c(x, y)\Big) \mathrm{d}\nu(y)\right) := S_\nu(\psi_k), \tag{6.22}$$

$$\psi_{k+1}(y) = -\varepsilon \log\left(\int_X \exp\Big(\frac{1}{\varepsilon}(\varphi_{k+1}(x) - c(x, y)\Big) \mathrm{d}\mu(x)\right) := S_\mu(\varphi_{k+1}). \tag{6.23}$$

Moreover, the following properties hold

(i) $\Phi_\varepsilon(\varphi_k, \psi_k) \leqslant \Phi_\varepsilon(\varphi_{k+1}, \psi_k) \leqslant \Phi_\varepsilon(\varphi_{k+1}, \psi_{k+1})$*;*

(ii) If $c(x, y)$ is $\omega-$continuous then $\varphi_{k+1}, \psi_{k+1}$ are also $\omega-$continuous ;

(iii) If $\psi_k - C$ ($\varphi_{k+1} - C$) is bounded by M on the support of ν (μ), then so is φ_{k+1} (ψ_{k+1}).

Proof. (6.22) and (6.23) follow by writing the first-order necessary condition which gives us

$$1 - \exp\left(\frac{\varphi(x)}{\varepsilon}\right) \int_Y \exp\Big(-\frac{1}{\varepsilon}(\psi(y) - c(x, y))\Big) \mathrm{d}\nu(y) = 0,\ x - a.e.$$

implying the desired formula (and by symmetry, the same result on S_μ holds). Therefore, $S_\nu(\psi)$ is the unique maximizer of $\varphi \mapsto \Phi_\varepsilon(\varphi, \psi)$.

By definition of ascent on each coordinate, (i) is obtained directly. More generally one can prove that the application S_ν (S_μ) is $\omega-$continuous. Let $x_1, x_2 \in X$ then

$$
\begin{aligned}
&|S_\nu(\psi)(x_1) - S_\nu(\psi)(x_2)| = \varepsilon \log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_2,y)\right)}\mathrm{d}\nu(y)\right) - \varepsilon \log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)\\
&= \varepsilon \log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)+c(x_1,y)-c(x_2,y)\right)}\mathrm{d}\nu(y)\right) - \varepsilon \log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)\\
&\leqslant \varepsilon \log\left(e^{\frac{\omega(d(x_1,x_2))}{\varepsilon}}\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right) - \varepsilon \log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)\\
&= \omega(d(x_1,x_2)).
\end{aligned}
\tag{6.24}
$$

The last point is just a bound on the iterates. ☐

Proposition 6.6. *The sequence* (φ_k, ψ_k) *defined by* (6.22) *and* (6.23) *converges in* $(\mathcal{C}_0, \|\cdot\|_\infty)$ *to the unique (up to a constant) couple of potentials* (φ, ψ) *which maximizes* Φ_ε.

Proof. Shifting the potentials by an additive constant, one can replace the optimization set by the couples (φ, ψ) which have uniformly bounded modulus of continuity and such that $\varphi(x_0) = 0$ for a given $x_0 \in X$. Recall that by proposition 6.4 the maximum of Φ is achieved at some couple (φ^*, ψ^*) which is unique up to a constant. Then, by prop. 6.5 (φ_k, ψ_k) are uniformly bounded and have uniformly modulus of continuity and one can extract a converging subsequence to $(\overline{\varphi}, \overline{\psi})$. By continuity of Φ and the monotonicity of the sequence, $\Phi_\varepsilon(\overline{\varphi}, S_\mu(\overline{\varphi})) \leqslant \Phi_\varepsilon(S_\nu \circ S_\mu(\overline{\varphi}), S_\mu(\overline{\varphi})) = \Phi_\varepsilon(\overline{\varphi}, S_\mu(\overline{\varphi}))$. Therefore, the maximizer coordinatewise being unique, one has

$$S_\nu(\overline{\psi}) = \overline{\varphi}, \tag{6.25}$$

$$S_\mu(\overline{\varphi}) = \overline{\psi}. \tag{6.26}$$

These show that $(\overline{\varphi}, \overline{\psi})$ is a critical point for Φ_ε, thus being a maximizer. ☐

The proof of convergence relies on some important properties of the log−sum−exp (LSE) function $\log \int \exp$ which we summarise in the next Lemma. Before that let define the pseudo-norm $\|\cdot\|_{\circ,\infty}$ of uniform convergence as

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

Lemma 6.7. *The LSE function is convex and*

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \|\varphi_1 - \varphi_2\|_{\circ,\infty} . \tag{6.27}$$

Proof. Convexity is easily verified. We can get the $1-$Lipschitz property as follows

$$
\begin{aligned}
|S_\mu(\varphi_1)(x) - S_\mu(\varphi_2)(x)| &= \left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} S_\mu(\varphi_2 + t(\varphi_1 - \varphi_2))\mathrm{d}t\right|\\
&\leqslant \int_0^1 \left|\int_X (\varphi_1 - \varphi_2)\frac{\exp(\frac{1}{\varepsilon}(\varphi_2 + t(\varphi_1 - \varphi_2) - c))}{\int_X \exp(\frac{1}{\varepsilon}(\varphi_2 + t(\varphi_1 - \varphi_2) - c))\mathrm{d}\mu}\mathrm{d}\mu\right|\\
&\leqslant \|\varphi_1 - \varphi_2\|_\infty .
\end{aligned}
$$

Notice that the equality occurs if and only if $\varphi_1 - \varphi_2$ is constant $\mu-$a.e.. In particular we would have $\varphi_1 = \varphi_2 + a$ and $S_\mu(\varphi_1) = S_\mu(\varphi_2) + a$. Thus it is natural to consider the set of continuous functions up to an additive constant $\mathcal{C}(X)/\mathbb{R}$ endowed with the pseudo-norm introduced above. Then, since $S_\mu(\varphi_1 + a) = S_\mu(\varphi_1) + a$ we got the same inequality for the norm $\|\cdot\|_{\circ,\infty}$. $\square$

Lemma 6.8. *Let $u, v \in \mathcal{C}(X)$ and $\mu \in \mathcal{P}(X)$ and denote ν_u and ν_v the Gibbs measures associated to u and v, that is $\mathrm{d}\nu_u = \frac{1}{Z_u} e^u \mathrm{d}\mu$ and $\mathrm{d}\nu_v = \frac{1}{Z_v} e^v \mathrm{d}\mu$, where Z_u and Z_v are the normalizing constants, then*

$$\|\nu_u - \nu_v\|_{L^1} \leqslant 2(1 - e^{-2\|u-v\|_{\circ,\infty}}).$$

Proof. Consider a bounded function g on X and define

$$\eta_g(t) := \int_X g \frac{e^{tv+(1-t)u}}{Z_{t,g}} \mathrm{d}\mu,$$

where $Z_{t,g} = \int_X e^{tv+(1-t)u} \mathrm{d}\mu$. Differentiating we get

$$\eta_g'(t) + \eta_{v-u}(t)\eta_g(t) = \eta_{(v-u)g}(t),$$

and

$$e^{\int_0^t \eta_{v-u}(s)\mathrm{d}s} \eta_g(t) - \eta_g(0) = \int_0^t \eta_{(v-u)g}(s) e^{\int_0^s \eta_{v-u}(r)\mathrm{d}r} \mathrm{d}s.$$

Observe that

$$\begin{aligned} |e^{\int_0^t \eta_{v-u}(s)\mathrm{d}s} \eta_g(t) - \eta_g(0)| &\leqslant \|g\|_\infty \int_0^t \eta_{(u-v)}(s) e^{\int_0^s \eta_{u-v}(r)\mathrm{d}r} \mathrm{d}s \\ &\leqslant \|g\|_\infty \left(e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} - 1 \right). \end{aligned}$$

Interchanging the role of u and v we have two possible cases: $\eta_g(1) \geqslant \eta_g(0) \geqslant 0$ or $\eta_g(1) \geqslant 0 \geqslant \eta_g(0)$. In the first case one has

$$|e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} (\eta_g(t) - \eta_g(0))| \leqslant |e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} \eta_g(t) - \eta_g(0)| \leqslant \|g\|_\infty \left(e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} - 1 \right).$$

In the second case there exists $t_0 \in [0,1]$ such that $\eta_g(t_0) = 0$ and we get

$$\begin{aligned} |\eta_g(1)| &\leqslant \|g\|_\infty \underbrace{\left(1 - e^{\int_{t_0}^1 \eta_{u-v}(s)\mathrm{d}s} \right)}_{:=a_1} \\ |\eta_g(0)| &\leqslant \|g\|_\infty \underbrace{\left(1 - e^{\int_0^{t_0} \eta_{u-v}(s)\mathrm{d}s} \right)}_{:=a_0}. \end{aligned}$$

Thus,

$$|\eta_g(1) - \eta_g(0)| \leqslant |\eta_g(1)| + |\eta_g(0)| \leqslant 2 \|g\|_\infty \max(a_1, a_0)$$

By exploiting the fact that $\eta_{u-v}(t) \leqslant 2\|u - v\|_{\circ,\infty}$ we obtain in both cases that

$$\|\nu_u - \nu_v\| \leqslant 2(1 - e^{-2\|u-v\|_{\circ,\infty}})$$

$\square$

Theorem 6.9. *(Convergence of Sinkhorn) The map $S = S_\nu \circ S_\mu$ is a contraction for $\|\cdot\|_{\circ,\infty}$. In particular the sequence (φ_k, ψ_k) defined by the Sinkhorn algorithm linearly converges to the unique (up to a constant) maximiser of the dual problem.*

Proof. We actually have to prove that

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}. \tag{6.28}$$

Once we have established that S_μ is a contraction then by lemma 6.7 it easily follows that

$$\|S(\varphi_1) - S(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty},$$

which would conclude the proof.
In order to prove (6.28) we start by giving an estimation of the oscillations of S_μ

$$\frac{1}{2}|S_\mu(\varphi_1)(y) - S_\mu(\varphi_2)(y) - S_\mu(\varphi_1)(x) + S_\mu(\varphi_2)(x)| \leqslant \frac{1}{2}\left|\int_0^1 \int_X (\varphi_1 - \varphi_2)(\mathrm{d}\eta_{t,y} - \mathrm{d}\eta_{t,x})dt\right|,$$

where $\mathrm{d}\eta_{t,z} := \frac{1}{Z} e^{\frac{t(\varphi_1-\varphi_2)+\varphi_2-c(z,\cdot)}{\varepsilon}} \mathrm{d}\mu$ where Z is the normalising constant. Since $\mathrm{d}\eta_{t,z}$ is a Gibbs measure we can apply the L^1 bound of lemma 6.8 to estimate $\|\eta_{t,y} - \eta_{t,x}\|_{L^1}$ and get

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}$$

with $\kappa_\mu = (1 - e^{-2\frac{\|c\|_{\circ,\infty}}{\varepsilon}})$. □

Remark 6.10 (Convergence speed)**.** This theorem shows that the Sinkhorn algorithm converges linearly, but notice that the contraction constant has a bad dependency in ε. Denoting $C = \|c\|_{\circ,\infty}$, to get an error of β one needs

$$(1 - e^{-2\frac{C}{\varepsilon}})^k \leqslant \beta$$

that is

$$k \gtrapprox e^{2C/\varepsilon} \log(1/\beta).$$

Remark 6.11. We refer the interested reader to [5, 13] where the convergence of Sinkhorn algorithm in infinite dimension (and generalized also to the multi-marginal case) is treated.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[2] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré, *Iterative bregman projections for regularized transportation problems*, SIAM Journal on Scientific Computing **37** (2015), no. 2, A1111–A1138.

[3] Garrett Birkhoff, *Extensions of jentzsch's theorem*, Transactions of the American Mathematical Society **85** (1957), no. 1, 219–227.

[4] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[5] Guillaume Carlier, *On the linear convergence of the multi-marginal sinkhorn algorithm*, (2021).

[6] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM Journal on Mathematical Analysis **49** (2017), no. 2, 1385–1418.

[7] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *Scaling algorithms for unbalanced transport problems*, arXiv preprint arXiv:1607.05816 (2016).

[8] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems, 2013, pp. 2292–2300.

[9] Joel Franklin and Jens Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its applications **114** (1989), 717–735.

[10] Alfred Galichon, *Optimal transport methods in economics*, Princeton University Press, 2018.

[11] Christian Léonard, *From the Schrödinger problem to the monge-kantorovich problem*, arXiv preprint arXiv:1011.2564 (2010).

[12] ———, *A survey of the Schrödinger problem and some of its connections with optimal transport*, arXiv preprint arXiv:1308.0215 (2013).

[13] Simone Di Marino and Augusto Gerolin, *An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm*, Journal of Scientific Computing **85** (2020), no. 2.

[14] Marcel Nutz, *Introduction to entropic optimal transport*, Lecture notes, Columbia University (2021).

[15] Gabriel Peyré, Marco Cuturi, et al., *Computational optimal transport: With applications to data science*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.

[16] R. Tyrrell Rockafellar, *Convex analysis, volume 28 of Princeton mathematics series*, 1970.

[17] Hans Samelson et al., *On the perron-frobenius theorem.*, The Michigan Mathematical Journal **4** (1957), no. 1, 57–59.

[18] Filippo Santambrogio, *Optimal transport for applied mathematicians*, Springer, 2015.

[19] François-Xavier Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*, (2019).

[20] Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.

[21] ———, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.