Memo for the 2021 SIGGRAPH Course
Computational Optimal Transport

Gabriel Peyré
CNRS & ENS

Marco Cuturi
ENSAE and Google Brain

Justin Solomon
MIT

September 13, 2021

Abstract

These notes are intended to be used as a memo to recap the main definitions and results for our 2021 SIGGRAPH course. More details regarding the algorithms and their applications can be found in the book [21], while more theoretical aspects are treated in details in [23].

Contents

1 Optimal Matching between Point Clouds

1.1 Monge Problem between Discrete points

Matching problem Given a cost matrix $(\mathrm{C}_{i,j})_{i\in[\![n]\!],j\in[\![m]\!]}$, assuming $n = m$, the optimal assignment problem seeks for a bijection σ in the set $\mathrm{Perm}(n)$ of permutations of n elements solving

$$\min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n}\sum_{i=1}^{n} \mathrm{C}_{i,\sigma(i)}. \tag{1}$$

One could naively evaluate the cost function above using all permutations in the set $\mathrm{Perm}(n)$. However, that set has size $n!$, which is gigantic even for small n. In general the optimal σ is non-unique.

1D case If the cost is of the form $\mathrm{C}_{i,j} = h(x_i - y_j)$, where $h : \mathbb{R} \to \mathbb{R}^+$ is convex (for instance $\mathrm{C}_{i,j} = |x_i - y_j|^p$ for $p \geqslant 1$), one has that an optimal σ necessarily defines an increasing map $x_i \mapsto x_{\sigma(i)}$, i.e.

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

Indeed, if this property is violated, i.e. there exists (i,j) such that $(x_i - y_j)(x_{\sigma(i)} - y_{\sigma(j)}) < 0$, then one can defines a permutation $\tilde{\sigma}$ by swapping the match, i.e. $\tilde{\sigma}(i) = \sigma(j)$ and $\tilde{\sigma}(j) = \sigma(i)$, with a better cost

$$\sum_i h(x_i - y_{\tilde{\sigma}(i)}) \leqslant \sum_i h(x_i - y_{\sigma(i)}),$$

because

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

So the algorithm to compute an optimal transport (actually all optimal transport) is to sort the points, i.e. find some pair of permutations σ_X, σ_Y such that

$$x_{\sigma_X(1)} \leqslant \sigma_{\sigma_X(2)} \leqslant \ldots \quad \text{and} \quad y_{\sigma_Y(1)} \leqslant \sigma_{\sigma_Y(2)} \leqslant \ldots$$

and then an optimal match is mapping $x_{\sigma_X(k)} \mapsto y_{\sigma_Y(k)}$, i.e. an optimal transport is $\sigma = \sigma_Y \circ \sigma_X^{-1}$. The total computational cost is thus $O(n\log(n))$ using for instance quicksort algorithm. Note that if $\varphi : \mathbb{R} \to \mathbb{R}$ is an increasing map, with a change of variable, one can apply this technique to cost of the form $h(|\varphi(x) - \varphi(y)|)$. A typical application is grayscale histogram equalization of the luminance of images.

Note that is h is concave instead of being convex, then the behavior is totally different, and the optimal match actually rather exchange the positions, and in this case there exists an $O(n^2)$ algorithm.

1.2 Matching Algorithms

There exists efficient algorithms to solve the optimal matching problems. The most well known are the hungarian and the auction algorithm, which runs in $O(n^3)$ operations. Their derivation and analysis is however very much simplified by introducing the Kantorovitch relaxation and its associated dual problem. A typical application of these methods is the equalization of the color palette between images, which corresponds to a 3-D optimal transport.

2 Monge Problem between Measures

2.1 Measures

Histograms We will interchangeably the term histogram or probability vector for any element $\mathrm{a} \in \Sigma_n$ that belongs to the probability simplex

$$\Sigma_n \overset{\text{def.}}{=} \left\{ \mathrm{a} \in \mathbb{R}_+^n \;;\; \sum_{i=1}^n \mathrm{a}_i = 1 \right\}.$$

Discrete measure, empirical measure A discrete measure with weights a and locations $x_1, \ldots, x_n \in \mathcal{X}$ reads

$$\alpha = \sum_{i=1}^n \mathrm{a}_i \delta_{x_i} \tag{2}$$

where δ_x is the Dirac at position x, intuitively a unit of mass which is infinitely concentrated at location x. Such as measure describes a probability measure if, additionally, $\mathrm{a} \in \Sigma_n$, and more generally a positive measure if each of the "weights" described in vector a is positive itself. An "empirical" probability distribution is uniform on a point cloud, i.e. $\mathrm{a} = \frac{1}{n}\sum_i \delta_{x_i}$. In practice, it many application is useful to be able to manipulate both the positions x_i ("Lagrangian" discretization) and the weights a_i ("Eulerian" discretization). Lagrangian modification is usually more powerful (because it leads to adaptive discretization) but it breaks the convexity of most problems.

General measures We consider Borel measures $\alpha \in \mathcal{M}(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$, i.e. one can compute $\alpha(A)$ for any Borel set A (which can be obtained by applying countable union, countable intersection, and relative complement to open sets). The measure should be finite, i.e. have a finite value on compact set. A Dirac measure δ_x is then define as $\delta_x(A) = 1$ is $x \in A$ and 0 otherwise, and this extend by linearity for discrete measures of the form (2) as

$$\alpha(A) = \sum_{x_i \in A} \mathrm{a}_i$$

We denote $\mathcal{M}_+(\mathcal{X})$ the subset of all positive measures on $\mathcal{X}$, i.e. $\alpha(A) \geqslant 0$ (and $\alpha(\mathcal{X}) < +\infty$ for the measure to be finite). The set of probability measures is denoted $\mathcal{M}_+^1(\mathcal{X})$, which means that any $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ is positive, and that $\alpha(\mathcal{X}) = 1$.

Radon measures Using Lebesgue integration, a Borel measure can be used to compute integral of measurable functions (i.e. such that level sets $\{x \;;\; f(x) < t\}$ are Borel sets), and we denote this pairing as

$$\langle f, \alpha \rangle \overset{\text{def.}}{=} \int f(x) \mathrm{d}\alpha(x).$$

Integration of such a measurable f against a discrete measure α computes a sum

$$\int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) = \sum_{i=1}^n \mathrm{a}_i f(x_i).$$

This can be in particular applied to the subspace of continuous functions which are measurable. Integration against a finite measure on a compact space thus defines a continuous linear form $f \mapsto \int f \mathrm{d}\alpha$ on the Banach space of continuous functions $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$, indeed $|\int f \mathrm{d}\alpha| \leqslant \|f\|_\infty |\alpha(\mathcal{X})|$. On compact spaces, the converse is true, namely that any continuous linear form $\ell : f \mapsto \ell(f)$ on $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ is represented as an integral against a measure $\ell(f) = \int f \mathrm{d}\alpha$. This is the Riesz-Markov-Kakutani representation theorem, which is often stated that Borel measures can be identified to Radon measures. Radon measures are thus in some sense "less regular" than functions, but more regular than distributions (which are dual to smooth

functions). For instance, the derivative of a Dirac is not a measure. This duality pairing $\langle f, \alpha \rangle$ between continuous function and measures will be crucial to develop duality theory for the convex optimization problem we will consider later.

The associated norm, which is the norm of the linear form ℓ, is the so-called total variation norm

$$\|\alpha\|_{TV} = \|\ell\|_{\mathcal{C}(\mathcal{X})\to\mathbb{R}} = \sup_{f\in\mathcal{C}(\mathcal{X})} \{\langle f, \alpha\rangle \; ; \; \|f\|_\infty \leqslant 1\}.$$

(note that one can remove the $|\cdot|$ in the right hand side, and such a quantity is often called a "dual norm"). One can in fact show that this TV norm is the total mass of the absolute value measure $|\alpha|$. The space $(\mathcal{M}(\mathcal{X}), \|\cdot\|_{TV})$ is a Banach space, which is the dual of $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$.

Recall that the absolute value of a measure is defined as

$$|\alpha|(A) = \sup_{A=\cup_i B_i} \sum_i |\alpha(B_i)|$$

so that for instance if $\alpha = \sum_i \mathrm{a}_i \delta_{x_i}$, $|\alpha| = \sum_i |\mathrm{a}_i| \delta_{x_i}$ and if $\mathrm{d}\alpha(x) = \rho \mathrm{d}x$ for a positif reference measure $\mathrm{d}x$, then $\mathrm{d}|\alpha|(x) = |\rho(x)|\mathrm{d}x$.

Relative densities A measure α which is a weighting of another reference one $\mathrm{d}x$ is said to have a density, which is denoted $\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x$ (on $\mathbb{R}^d$ $\mathrm{d}x$ is often the Lebesgue measure), often also denoted $\rho_\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}x}$, which means that

$$\forall\, h \in \mathcal{C}(\mathbb{R}^d), \quad \int_{\mathbb{R}^d} h(x)\mathrm{d}\alpha(x) = \int_{\mathbb{R}^d} h(x)\rho_\alpha(x)\mathrm{d}x.$$

Probabilistic interpretation Radon probability measures can also be viewed as representing the distributions of random variables. A random variable X on $\mathcal{X}$ is actually a map $X : \Omega \to \mathcal{X}$ from some abstract (often un-specified) probabized space $(\Omega, \mathbb{P})$, and its distribution is the Radon measure $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ such that $\mathbb{P}(X \in A) = \alpha(A) = \int_A \mathrm{d}\alpha(x)$.

2.2 Push Forward

For some continuous map $T : \mathcal{X} \to \mathcal{Y}$, we define the pushforward operator $T_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$. For a Dirac mass, one has $T_\sharp \delta_x = \delta_{T(x)}$, and this formula is extended to arbitrary measure by linearity. In some sense, moving from T to $T_\sharp$ is a way to linearize any map at the prize of moving from a (possibly) finite dimensional space $\mathcal{X}$ to the infinite dimensional space $\mathcal{M}(\mathcal{X})$, and this idea is central to many convex relaxation method, most notably Lasserre's relaxation. For discrete measures (2), the pushforward operation consists simply in moving the positions of all the points in the support of the measure

$$T_\sharp \alpha \overset{\text{def.}}{=} \sum_i \mathrm{a}_i \delta_{T(x_i)}.$$

For more general measures, for instance for those with a density, the notion of push-forward plays a fundamental to describe spatial modifications of probability measures. The formal definition reads as follow.

Definition 1 (Push-forward)**.** *For $T : \mathcal{X} \to \mathcal{Y}$, the push forward measure $\beta = T_\sharp \alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ satisfies*

$$\forall\, h \in \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y)\mathrm{d}\beta(y) = \int_{\mathcal{X}} h(T(x))\mathrm{d}\alpha(x). \tag{3}$$

Equivalently, for any measurable set $B \subset \mathcal{Y}$, one has

$$\beta(B) = \alpha(\{x \in \mathcal{X} \; ; \; T(x) \in B\}). \tag{4}$$

Note that $T_\sharp$ preserves positivity and total mass, so that if $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ then $T_\sharp \alpha \in \mathcal{M}_+^1(\mathcal{Y})$.

Remark 1 (Push-forward for densities). Explicitly doing the change of variable $x = T(x)$, so that $\mathrm{d}x = |\det(T'(x))|\mathrm{d}y$ in formula (3) for measures with densities $(\rho_\alpha, \rho_\beta)$ on $\mathbb{R}^d$ (assuming T is smooth and a bijection), one has for all $h \in \mathcal{C}(\mathcal{Y})$

$$\int_{\mathcal{Y}} h(y)\rho_\beta(y)\mathrm{d}y = \int_{\mathcal{Y}} h(y)\mathrm{d}\beta(y) = \int_{\mathcal{X}} h(T(x))\mathrm{d}\alpha(x) = \int_{\mathcal{X}} h(T(x))\rho_\alpha(x)\mathrm{d}x$$
$$= \int_{\mathcal{Y}} h(y)\rho_\alpha(T^{-1}y)\frac{\mathrm{d}y}{|\det(T'(T^{-1}y))|},$$

which shows that

$$\rho_\beta(y) = \rho_\alpha(T^{-1}y)\frac{1}{|\det(T'(T^{-1}y))|}.$$

Since T is a diffeomorphism, one obtains equivalently

$$\rho_\alpha(x) = |\det(T'(x))|\rho_\beta(T(x)) \tag{5}$$

where $T'(x) \in \mathbb{R}^{d\times d}$ is the Jacobian matrix of T (the matrix formed by taking the gradient of each coordinate of T). This implies, denoting $y = T(x)$

$$|\det(T'(x))| = \frac{\rho_\alpha(x)}{\rho_\beta(y)}.$$

Remark 2 (Probabilistic interpretation). A random variable X, equivalently, is the push-forward of $\mathbb{P}$ by X, $\alpha = X_\sharp\mathbb{P}$. Applying another push-forward $\beta = T_\sharp\alpha$ for $T : \mathcal{X} \to \mathcal{Y}$, following (3), is equivalent to defining another random variable $Y = T(X) : \omega \in \Omega \to T(X(\omega)) \in Y$, so that β is the distribution of Y. Drawing a random sample y from Y is thus simply achieved by computing $y = T(x)$ where x is drawn from X.

2.3 Monge's Formulation

Monge problem. Monge problem (1) is extended to the setting of two arbitrary probability measures (α, β) on two spaces $(\mathcal{X}, \mathcal{Y})$ as finding a map $T : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$\inf_T \left\{ \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\alpha(x) \;;\; T_\sharp\alpha = \beta \right\}. \tag{6}$$

The constraint $T_\sharp\alpha = \beta$ means that T pushes forward the mass of α to β, and makes use of the push-forward operator (3).

For empirical measure with same number $n = m$ of points, one retrieves the optimal matching problem. Indeed, this corresponds to the setting of empirical measures $\alpha = \sum_i \delta_{x_i}$ and $\beta = \sum_i \delta_{y_i}$. In this case, $T_\sharp\alpha = \beta$ necessarily implies that σ is one-to-one, $T : x_i \mapsto x_{\sigma(i)}$, so that

$$\int_{\mathcal{X}} c(x, T(x))\mathrm{d}\alpha(x) = \sum_i c(x_i, x_{\sigma(i)}).$$

In general, an optimal map T solving (6) might fail to exist. In fact, the constraint set $T_\sharp\alpha = \beta$, which is the case for instance if $\alpha = \delta_x$ and β is not a single Dirac. Even if the constraint set is not empty the infimum might not be reached, the most celebrated example being the case of α being distributed uniformly on a single segment and β being distributed on two segments on the two sides.

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

$$\tilde{\mathcal{W}}_p^p(\alpha, \beta) \overset{\text{def.}}{=} \inf_T \left\{ \mathcal{E}_\alpha(T) \overset{\text{def.}}{=} \int_{\mathcal{X}} d(x, T(x))^p\mathrm{d}\alpha(x) \;;\; T_\sharp\alpha = \beta \right\}. \tag{7}$$

If the constraint set is empty, then we set $\tilde{\mathcal{W}}_p^p(\alpha, \beta) = +\infty$. The following proposition shows that quantity defines a distance.

Proposition 1. *$\tilde{\mathcal{W}}$ is a distance.*

Proof. If $\tilde{\mathcal{W}}_p^p(\alpha,\beta)=0$ then necessarily the optimal map is Id on the support of α and $\beta=\alpha$. Let us prove that $\tilde{\mathcal{W}}_p(\alpha,\beta)\leqslant\tilde{\mathcal{W}}_p(\alpha,\gamma)+\tilde{\mathcal{W}}_p(\gamma,\beta)$. If $\tilde{\mathcal{W}}_p(\alpha,\beta)=+\infty$, then either $\tilde{\mathcal{W}}_p(\alpha,\gamma)=+\infty$ or $\tilde{\mathcal{W}}_p(\gamma,\beta)=+\infty$, because otherwise we consider two maps (S,T) such that $S_\sharp\alpha=\gamma$ and $T_\sharp\gamma=\beta$ and then $(T\circ S)_\sharp\alpha=\beta$ so that $\tilde{\mathcal{W}}_p^p(\alpha,\beta)\leqslant\mathcal{E}_\alpha(S\circ T)<+\infty$. So necessarily $\tilde{\mathcal{W}}_p^p(\alpha,\beta)<+\infty$ and we can restrict our attention to the cases where $\tilde{\mathcal{W}}_p^p(\alpha,\gamma)<+\infty$ and $\tilde{\mathcal{W}}_p^p(\gamma,\beta)<+\infty$ because otherwise the inequality is trivial. For any $\varepsilon>0$, we consider ε-minimizer $S_\sharp\alpha=\gamma$ and $T_\sharp\gamma=\beta$ such that

$$E_\alpha(S)^{\frac{1}{p}}\leqslant\tilde{\mathcal{W}}_p(\alpha,\gamma)+\varepsilon\quad\text{and}\quad E_\gamma(T)^{\frac{1}{p}}\leqslant\tilde{\mathcal{W}}_p(\gamma,\beta)+\varepsilon.$$

Now we have that $(T\circ S)_\sharp\alpha=\gamma$, so that one has, using sub-optimality of this map and the triangular inequality

$$\mathcal{W}_p(\alpha,\gamma)\leqslant\int d(x,T(S(x)))^p\mathrm{d}\alpha(x)^{\frac{1}{p}}\leqslant\int(d(x,S(x))+d(S(x),T(S(x))))^p\mathrm{d}\alpha(x)^{\frac{1}{p}}.$$

The using Minkowski inequality for the L^p spaces with measure α

$$\|f+g\|_{L^p(\alpha)}\leqslant\|f\|_{L^p(\alpha)}+\|g\|_{L^p(\alpha)}$$

and with $f(x)\triangleq d(x,S(x))$ and $g(x)\triangleq d(S(x),T(S(x)))$ one has

$$\mathcal{W}_p(\alpha,\gamma)\leqslant\int d(x,S(x))^p\mathrm{d}\alpha(x)^{\frac{1}{p}}+\int d(S(x),T(S(x)))^p\mathrm{d}\alpha(x)^{\frac{1}{p}}\leqslant\mathcal{W}_p(\alpha,\beta)+\mathcal{W}_p(\beta,\gamma)+2\varepsilon.$$

Letting $\varepsilon\to0$ gives the result. □

2.4 Existence and Uniqueness of the Monge Map

Brenier's theorem. The following celebrated theorem of [6] ensures that in $\mathbb{R}^d$ for $p=2$, if at least one of the two inputs measures has a density, then Kantorovitch and Monge problems are equivalent.

Theorem 1 (Brenier)**.** *In the case $\mathcal{X}=\mathcal{Y}=\mathbb{R}^d$ and $c(x,y)=\|x-y\|^2$, if α has a density with respect to the Lebesgue measure, then there exists a unique optimal Monge map T. This map is characterized by being the unique gradient of a convex function $T=\nabla\varphi$ such that $(\nabla\varphi)_\sharp\alpha=\beta$.*

Its proof requires to study the relaxed Kantorovitch problems and its dual, so we defer it to later (Section 5.3).

Brenier's theorem, stating that an optimal transport map must be the gradient of a convex function, should be examined under the light that a convex function is a natural generalization of the notion of increasing functions in dimension more than one. For instance, the gradient of a convex function is a monotone gradient field in the sense

$$\forall\,(x,x')\in\mathbb{R}^d\times\mathbb{R}^d,\quad\langle\nabla\varphi(x)-\nabla\varphi(x'),\,x-x'\rangle\geqslant0.$$

Note however that in dimension larger than 1, not all monotone fields are gradient of convex function. For instance, a rotation is monotone but can never be an optimal transport because a gradient field Ax defined by a linear map A is necessarily obtained by a symmetric matrix A. Indeed, such a linear field must be associated to a quadratic form $\varphi(x)=\langle Bx,\,x\rangle/2$ and hence $A=\nabla\varphi=(B+B^\top)/2$. Optimal transport can thus plays an important role to define quantile functions in arbitrary dimensions, which in turn is useful for applications to quantile regression problems [8].

Note also that this theorem can be extended in many directions. The condition that α has a density can be weakened to the condition that it does not give mass to "small sets" having Hausdorff dimension smaller

than $d-1$ (e.g. hypersurfaces). One can also consider costs of the form $c(x,y) = h(x-y)$ where h is a strictly convex smooth function, for for instance $c(x,y) = \|x-y\|^p$ with $1 < p < +\infty$.

Note that Brenier's theorem provides existence and uniqueness, but in general, the map T can be very irregular. Indeed, φ is in general non-smooth, but it is in fact convex and Lipschitz, so that $\nabla\varphi$ is actually well defined α-almost everywhere. Ensuring T to be smooth actually requires the target β to be regular, and more precisely its support must be convex.

If α does not have a density, then T might fail to exists and it should be replaced by a set-valued function included in $\partial\varphi$ which is now the sub-differential of a convex function, which might have singularity on a non-zero measure set. This means that T can "split" the mass by mapping to several locations $T(x) \subset \partial\varphi$. Actually, the condition that $T(x) \subset \partial\varphi(x)$ and $T_\sharp\alpha = \beta$ implies that the multi-map T defines a solution of Kantorovitch problem that will be studied later.

Monge-Ampère equation. For measures with densities, using (5), one obtains that φ is the unique (up to the addition of a constant) convex function which solves the following Monge-AmpÃ¨re-type equation

$$\det(\partial^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x) \tag{8}$$

where $\partial^2\varphi(x) \in \mathbb{R}^{d\times d}$ is the hessian of φ. The convexity constraint forces $\det(\partial^2\varphi(x)) \geqslant 0$ and is necessary for this equation to have a solution and be well-posed. The Monge-Ampère operator $\det(\partial^2\varphi(x))$ can be understood as a non-linear degenerate Laplacian. In the limit of small displacements, one can consider $\varphi(x) = \|x\|^2/2 + \varepsilon\psi$ so that $\nabla\varphi = \mathrm{Id} + \varepsilon\nabla\psi$, one indeed recovers the Laplacian Δ as a linearization since for smooth maps

$$\det(\partial^2\varphi(x)) = 1 + \varepsilon\Delta\psi(x) + o(\varepsilon),$$

where we used the fact that $\det(\mathrm{Id} + \varepsilon A) = 1 + \varepsilon\,\mathrm{tr}(A) + o(\varepsilon)$.

OT in 1-D. For a measure α on $\mathbb{R}$, we introduce the cumulative function

$$\forall\, x \in \mathbb{R}, \quad \mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha, \tag{9}$$

which is a function $\mathcal{C}_\alpha : \mathbb{R} \to [0,1]$. Its pseudo-inverse $\mathcal{C}_\alpha^{-1} : [0,1] \to \mathbb{R}\cup\{-\infty\}$

$$\forall\, r \in [0,1], \quad \mathcal{C}_\alpha^{-1}(r) = \min_x \left\{x \in \mathbb{R}\cup\{-\infty\}\ ;\ \mathcal{C}_\alpha(x) \geqslant r\right\}.$$

That function is also called the quantile function of α. The following proposition shows that these defines push-forward toward the uniform distribution $\mathcal{U}$ on $[0,1]$.

Proposition 2. *One has $(\mathcal{C}_\alpha)_\sharp^{-1}\mathcal{U} = \alpha$, where $\mathcal{U}$ is the uniform distribution in $[0,1]$. If α has a density, then $(\mathcal{C}_\alpha)_\sharp\alpha = \mathcal{U}$.*

Proof. For simplicity, we assume α has a strictly positive density, so that $\mathcal{C}_\alpha$ is a strictly increasing continuous function. Denoting $\gamma \overset{\text{def.}}{=} (\mathcal{C}_\alpha)_\sharp^{-1}\mathcal{U}$ we aim at proving $\gamma = \alpha$, which is equivalent to $\mathcal{C}_\gamma = \mathcal{C}_\alpha$. One has

$$\mathcal{C}_\gamma(x) = \int_{-\infty}^{x} \mathrm{d}\gamma = \int_{\mathbb{R}} 1_{]-\infty,x]}\mathrm{d}((\mathcal{C}_\alpha^{-1})_\sharp\mathcal{U}) = \int_0^1 1_{]-\infty,x]}(\mathcal{C}_\alpha^{-1}(z))\mathrm{d}z = \int_0^1 1_{[0,\mathcal{C}_\alpha(x)]}(z)\mathrm{d}z = \mathcal{C}_\alpha(x)$$

where we use the fact that

$$-\infty \leqslant \mathcal{C}_\alpha^{-1}(z) \leqslant x \quad \Longleftrightarrow 0 \leqslant z \leqslant \mathcal{C}_\alpha(x).$$

□

If α has a density, this shows that the map

$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha \tag{10}$$

satisfies $T_\sharp \alpha = \beta$.

For the cost $c(x,y) = |x = y|^2$, since this T is increasing (hence the gradient of a convex function since we are in 1-D), by Brenier's theorem, T is the solution to Monge problem (at least if we impose that α has a density, otherwise it might lead to a solution of Kantorovitch problem by properly defining the pseudo-inverse). This closed form formula is also optimal for any cost of the form $h(|x-y|)$ for increasing h. For discrete measures, one cannot apply directly this reasoning (because α does not have a density), but if the measure are uniform on the same number of Dirac masses, then this approach is actually equivalent to the sorting formula.

Plugging this optimal map into the definition of the "Wasserstein" distance (we will see later that this quantity defines a distance), so that for any $p \geqslant 1$, one has

$$\mathcal{W}_p(\alpha,\beta)^p = \int_{\mathbb{R}} |x - \mathcal{C}_\beta^{-1}(\mathcal{C}_\alpha(x))| \mathrm{d}\alpha(x) = \int_0^1 |\mathcal{C}_\alpha^{-1}(r) - \mathcal{C}_\beta^{-1}(r)|^p \mathrm{d}r = \|\mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1}\|_{L^p([0,1])}^p. \tag{11}$$

This formula is still valid for any measure (one can for instance approximate α by a measure with density). This formula means that through the map $\alpha \mapsto \mathcal{C}_\alpha^{-1}$, the Wasserstein distance is isometric to a linear space equipped with the L^p norm. For $p=2$, the Wasserstein distance for measures on the real line is thus a Hilbertian metric. This makes the geometry of 1-D optimal transport very simple, but also very different from its geometry in higher dimensions, which is not Hilbertian.

For $p=1$, one even has the simpler formula. Indeed, the previous formula is nothing more than the area between the two graphs of the copula, which can thus be computed by exchanging the role of the two axis, so that

$$\mathcal{W}_1(\alpha,\beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x = \int_{\mathbb{R}} \left| \int_{-\infty}^x \mathrm{d}(\alpha-\beta) \right| \mathrm{d}x. \tag{12}$$

which shows that $\mathcal{W}_1$ is a norm (see paragraph 6.3 for the generalization to arbitrary dimensions).

It is possible to define other type of norm which behave similarly (i.e. metrize the convergence in law), for instance $\|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^p(\mathbb{R})}$ define respectively the Wasserstein, Cramer (i.e. Sobolev) and Kolmogorov-Smirnov norms for $p = 1, 2, \infty$.

OT on 1-D Gaussians We first consider the case where $\alpha = \mathcal{N}(m_\alpha, s_\alpha^2)$ and $\beta = \mathcal{N}(m_\beta, s_\beta^2)$ are two Gaussians in $\mathbb{R}$. Then one verifies that

$$T(x) = \frac{s_\beta}{s_\alpha}(x - m_\alpha) + m_\beta$$

satisfies $T_\sharp \alpha = \beta$, furthermore it is the the derivative of the convex function

$$\varphi(x) = \frac{s_\beta}{2 s_\alpha}(x - m_\alpha)^2 + m_\beta x,$$

so that according to Brenier's theorem, for the cost $c(x-y) = (x-y)^2$, T is the unique optimal transport, and the associated Monge distance is, after some computation

$$\tilde{\mathcal{W}}_2^2(\alpha,\beta) = \int_{\mathbb{R}} \left(\frac{s_\beta}{s_\alpha}(x - m_\alpha) + m_\beta - x \right)^2 \mathrm{d}\alpha(x) = (m_\alpha - m_\beta)^2 + (s_\alpha - s_\beta)^2.$$

This formula still holds for Dirac masses, i.e. if $s_\alpha = 0$ or $s_\beta = 0$. The OT geometry of Gaussians is thus the Euclidean distance on the half plane $(m,s) \in \mathbb{R} \times \mathbb{R}_+$. This should be contrasted with the geometry of KL, where singular Gaussians (for which $s=0$) are infinitely distant.

OT on Gaussians If $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$ are two Gaussians in $\mathbb{R}^d$, we now look for an affine map

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha). \tag{13}$$

This map is the gradient of the convex function $\varphi(x) = \langle \mathbf{m}_\beta, x\rangle + \langle A(x - \mathbf{m}_\alpha), x - \mathbf{m}_\alpha\rangle/2$ if and only if A is a symmetric positive matrix.

Proposition 3. *One has $T_\sharp \alpha = \beta$ if and only if*

$$A\mathbf{\Sigma}_\alpha A = \mathbf{\Sigma}_\beta. \tag{14}$$

Proof. Indeed, one simply has to notice that the change of variables formula (5) is satisfied since

$$\begin{aligned}
\rho_\beta(T(x)) &= \det(2\pi\mathbf{\Sigma}_\beta)^{-\frac{1}{2}} \exp(-\langle T(x) - \mathbf{m}_\beta, \mathbf{\Sigma}_\beta^{-1}(T(x) - \mathbf{m}_\beta)\rangle) \\
&= \det(2\pi\mathbf{\Sigma}_\beta)^{-\frac{1}{2}} \exp(-\langle x - \mathbf{m}_\alpha, A^\top \mathbf{\Sigma}_\beta^{-1} A(x - \mathbf{m}_\alpha)\rangle) \\
&= \det(2\pi\mathbf{\Sigma}_\beta)^{-\frac{1}{2}} \exp(-\langle x - \mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha^{-1}(x - \mathbf{m}_\alpha)\rangle),
\end{aligned}$$

and since T is a linear map we have that

$$|\det T'(x)| = \det A = \left(\frac{\det \mathbf{\Sigma}_\beta}{\det \mathbf{\Sigma}_\alpha}\right)^{\frac{1}{2}}$$

and we therefore recover $\rho_\alpha = |\det T'|\rho_\beta$ meaning $T_\sharp\alpha = \beta$. □

Equation (14) is a quadratic equation on A. Using the square root of positive matrices, which is uniquely defined, one has

$$\mathbf{\Sigma}_\alpha^{\frac{1}{2}} \mathbf{\Sigma}_\beta \mathbf{\Sigma}_\alpha^{\frac{1}{2}} = \mathbf{\Sigma}_\alpha^{\frac{1}{2}} A \mathbf{\Sigma}_\alpha A \mathbf{\Sigma}_\alpha^{\frac{1}{2}} = (\mathbf{\Sigma}_\alpha^{\frac{1}{2}} A \mathbf{\Sigma}_\alpha^{\frac{1}{2}})^2,$$

so that this equation has a unique solution, given by

$$A = \mathbf{\Sigma}_\alpha^{-\frac{1}{2}} \left(\mathbf{\Sigma}_\alpha^{\frac{1}{2}} \mathbf{\Sigma}_\beta \mathbf{\Sigma}_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}} \mathbf{\Sigma}_\alpha^{-\frac{1}{2}} = A^\top.$$

Using Brenier's theorem [6], we conclude that T is optimal.

With additional calculations involving first and second order moments of ρ_α, we obtain that the transport cost of that map is

$$\tilde{\mathcal{W}}_2^2(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \tag{15}$$

where $\mathcal{B}$ is the so-called Bures' metric [7] between positive definite matrices (see also [10]),

$$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \overset{\text{def.}}{=} \operatorname{tr}\left(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - 2(\mathbf{\Sigma}_\alpha^{1/2} \mathbf{\Sigma}_\beta \mathbf{\Sigma}_\alpha^{1/2})^{1/2}\right), \tag{16}$$

where $\mathbf{\Sigma}^{1/2}$ is the matrix square root. One can show that $\mathcal{B}$ is a distance on covariance matrices, and that $\mathcal{B}^2$ is convex with respect to both its arguments. In the case where $\mathbf{\Sigma}_\alpha = \operatorname{diag}(r_i)_i$ and $\mathbf{\Sigma}_\beta = \operatorname{diag}(s_i)_i$ are diagonals, the Bures metric is the Hellinger distance

$$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta) = \|\sqrt{r} - \sqrt{s}\|_2.$$

3 Kantorovitch Relaxation

3.1 Discrete Relaxation

Monge discrete matching problem is problematic because it cannot be applied when $n \neq m$. One needs to take into account masses $(\mathrm{a}_i, \mathrm{b}_j)$ to handle this more general situation. Monge continuous formulation (6) using push-forward is also problematic because it can be the case that there is no transport map T such that $T_\sharp \alpha = \beta$, for instance when α is made of a single Dirac to be mapped to several Dirac. Associated to this, it is not symmetric with respect to exchange of α and β (one can map two Diracs to a single one, but not the other way). Also, these are non-convex optimization problem which are not simple to solve numerically.

The key idea of [17] is to relax the deterministic nature of transportation, namely the fact that a source point x_i can only be assigned to another, or transported to one and one location $T(x_i)$ only. Kantorovich proposes instead that the mass at any point x_i be potentially dispatched across several locations. Kantorovich moves away from the idea that mass transportation should be "deterministic" to consider instead a "probabilistic" (or "fuzzy") transportation, which allows what is commonly known now as "mass splitting" from a source towards several targets. This flexibility is encoded using, in place of a permutation σ or a map T, a coupling matrix $\mathrm{P} \in \mathbb{R}_+^{n \times m}$, where $\mathrm{P}_{i,j}$ describes the amount of mass flowing from bin i (or point x_i) towards bin j (or point x_j), x_i towards y_j in the formalism of discrete measures $\alpha = \sum_i \mathrm{a}_i \delta_{x_i}$, $\beta = \sum_j \mathrm{b}_j \delta_{y_j}$. Admissible couplings are only constrained to satisfy the conservation of mass

$$\mathrm{U}(\mathrm{a}, \mathrm{b}) \overset{\text{def.}}{=} \left\{ \mathrm{P} \in \mathbb{R}_+^{n \times m} \;;\; \mathrm{P}\mathbb{1}_m = \mathrm{a} \quad \text{and} \quad \mathrm{P}^\top \mathbb{1}_n = \mathrm{b} \right\}, \tag{17}$$

where we used the following matrix-vector notation

$$\mathrm{P}\mathbb{1}_m = \left(\sum_j \mathrm{P}_{i,j} \right)_i \in \mathbb{R}^n \quad \text{and} \quad \mathrm{P}^\top \mathbb{1}_n = \left(\sum_i \mathrm{P}_{i,j} \right)_j \in \mathbb{R}^m.$$

The set of matrices $\mathrm{U}(\mathrm{a}, \mathrm{b})$ is bounded, defined by $n+m$ equality constraints, and therefore a convex polytope (the convex hull of a finite set of matrices).

Additionally, whereas the Monge formulation is intrinsically asymmetric, Kantorovich's relaxed formulation is always symmetric, in the sense that a coupling P is in $\mathrm{U}(\mathrm{a}, \mathrm{b})$ if and only if $\mathrm{P}^\top$ is in $\mathrm{U}(\mathrm{b}, \mathrm{a})$.

Kantorovich's optimal transport problem now reads

$$\mathrm{L}_\mathrm{C}(\mathrm{a}, \mathrm{b}) \overset{\text{def.}}{=} \min_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{C}, \mathrm{P} \rangle \overset{\text{def.}}{=} \sum_{i,j} \mathrm{C}_{i,j} \mathrm{P}_{i,j}. \tag{18}$$

This is a linear program, and as is usually the case with such programs, its solutions are not necessarily unique.

Note however that there are always sparse optimal transport plan, in the sense that there is always a solution P to (18) with $n + m - 1$ solution. This is because there are $n + m$ linear constraint, but the rank of the constraint is actually $n + m - 1$ (there is a single redundancy among the constraint, which stems from the fact that both $\mathrm{P}\mathbb{1}_m = \mathbb{1}_n$ and $\mathrm{P}^\top \mathbb{1}_n = \mathbb{1}_m$, then $\sum_{i,j} \mathrm{P}_{i,j} = 1$.

Linear programming algorithms. The reference algorithms to solve (18) are network simplexes. There exists instances of this method which scale like $O(n^3 \log n)$. Alternative include interior points, which are usually inferior on this particular type of linear program.

1-D cases. In 1-D, if $c(x, y) = |x - y|^p$ on $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ with $p \geqslant 1$, then an optimal transport map is given by an increasing map. So as explained in (1.1), the case $n = m$ and $\mathrm{a}_i = \mathrm{b}_j = \frac{1}{n}$ is solved in $O(n \log(n))$ operations. In the general case, an optimal coupling matrix P can be computed similarly in $O(n \log(n) + m \log(m))$ by sorting the points and then sweeping the mass in a single pass from left to right .

Permutation Matrices as Couplings We restrict our attention to the special case $n = m$ and $\mathrm{a}_i = \mathrm{b}_i = 1$ (up to a scaling by $1/n$, these are thus probability measures). In this case one can solve Monge optimal matching problem (1), and it is convenient to re-write it using permutation matrices. For a permutation $\sigma \in \mathrm{Perm}(n)$, we write P_σ for the corresponding permutation matrix,

$$\forall\, (i,j) \in [\![n]\!]^2, \quad (\mathrm{P}_\sigma)_{i,j} = \begin{cases} 1 & \text{if} \quad j = \sigma_i, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

We denote the set of permutation matrices as

$$\mathcal{P}_n \overset{\text{def.}}{=} \{\mathrm{P}_\sigma \;;\; \sigma \in \mathrm{Perm}(n)\},$$

which is a discrete, hence non-convex, set. One has

$$\langle \mathrm{C},\, \mathrm{P}_\sigma \rangle = \sum_{i=1}^n \mathrm{C}_{i,\sigma_i}$$

so that (1) is equivalent to the non-convex optimization problem

$$\min_{\mathrm{P} \in \mathcal{P}_n} \langle \mathrm{C},\, \mathrm{P} \rangle.$$

In contrast, one has that $\mathrm{U}(\mathrm{a}, \mathrm{b}) = \mathcal{B}_n$ is equal to the convex set of bistochastic matrices

$$\mathcal{B}_n \overset{\text{def.}}{=} \left\{ \mathrm{P} \in \mathbb{R}_+^{n \times n} \;;\; \mathrm{P}\mathbb{1}_n = \mathrm{P}^\top \mathbb{1}_n = \mathbb{1}_n \right\}$$

so that Kantorovitch problem reads

$$\min_{\mathrm{P} \in \mathcal{B}_n} \langle \mathrm{C},\, \mathrm{P} \rangle.$$

The set of permutation matrices is strictly included in the set of bistochastic matrices, and more precisely

$$\mathcal{P}_n = \mathcal{B}_n \cap \{0,1\}^{n \times n}.$$

This shows that one has the following obvious relation between the cost of Monge and Kantorovitch problem

$$\min_{\mathrm{P} \in \mathcal{B}_n} \langle \mathrm{C},\, \mathrm{P} \rangle \leqslant \min_{\mathrm{P} \in \mathcal{P}_n} \langle \mathrm{C},\, \mathrm{P} \rangle.$$

We will now show that there is in fact an equality between these two costs, so that both problems are in some sense equivalent.

For this, we will make a detour through more general linear optimization problem of the form $\min_{\mathrm{P} \in \mathcal{C}} \langle \mathrm{C},\, \mathrm{P} \rangle$ for some compact convex set $\mathcal{C}$. We firs introduce the notion of extremal point, which are intuitively the vertices of $\mathcal{C}$

$$\mathrm{Extr}(\mathcal{C}) \overset{\text{def.}}{=} \left\{ \mathrm{P} \;;\; \forall\, (Q,R) \in \mathcal{C}^2, \mathrm{P} = \frac{Q+R}{2} \Rightarrow Q = R \right\}.$$

So to show that $\mathrm{P} \notin \mathrm{Extr}(\mathcal{C})$ is suffices to split P as $\mathrm{P} = \frac{Q+R}{2}$ with $Q \neq R$ and $(Q,R) \in \mathcal{C}^2$. We will assume the following fundamental result.

Proposition 4. *If $\mathcal{C}$ is compact, then* $\mathrm{Extr}(\mathcal{C}) \neq 0$.

The fact that $\mathcal{C}$ is compact is crucial, for instance the set $\left\{(x,y) \in \mathbb{R}_+^2 \;;\; xy \geqslant 1\right\}$ has no extremal point.

We can now use this result to show the following fundamental result, namely that there is always a solution to a linear program which is an extremal point. Note that of course the set of solution (which is non-empty because one minimizes a continuous function on a compact) might not be a singleton.

Proposition 5. *If $\mathcal{C}$ is compact, then*

$$Extr(\mathcal{C}) \cap \left(\operatorname*{argmin}_{\mathrm{P} \in \mathcal{C}} \langle \mathrm{C}, \mathrm{P} \rangle \right) \neq \emptyset.$$

Proof. One consider $\mathcal{S} \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathrm{P} \in \mathcal{C}} \langle \mathrm{C}, \mathrm{P} \rangle$. We first note that $\mathcal{S}$ is convex (as always for an argmin) and compact, because $\mathcal{C}$ is compact and the objective function is continuous, so that $\mathrm{Extr}(\mathcal{S}) \neq \emptyset$. We will show that $\mathrm{Extr}(\mathcal{S}) \subset \mathrm{Extr}(\mathcal{C})$. □

The following theorem states that the extremal points of bistochastic matrices are the permutation matrices. It implies as a corollary that the cost of Monge and Kantorovitch are the same, and that they share a common solution.

Theorem 2 (Birkhoff and von Neumann)**.** *One has $Extr(\mathcal{B}_n) = \mathcal{P}_n$.*

Proof. We first show the simplest inclusion $\mathcal{P}_n \subset \mathrm{Extr}(\mathcal{B}_n)$. Indeed it follows from the fact that $\mathrm{Extr}([0,1]) = \{0,1\}$. Take $\mathrm{P} \in \mathcal{P}_n$, if $\mathrm{P} = (Q+R)/2$ with $Q_{i,j}, R_{i,j} \in [0,1]$, since $\mathrm{P}_{i,j} \in \{0,1\}$ then necessarily $Q_{i,j} = R_{i,j} \in \{0,1\}$.

Now we show $\mathrm{Extr}(\mathcal{B}_n) \subset \mathcal{P}_n$ by showing that $\mathcal{P}_n^c \subset \mathrm{Extr}(\mathcal{B}_n)^c$ where the complementary are computed inside the larger set $\mathcal{B}_n$. So picking $\mathrm{P} \in \mathcal{B}_n \backslash \mathcal{P}_n$, we need to split $\mathrm{P} = (Q+R)/2$ where Q, R are distinct bistochastic matrices. P defines a partite graph linking two sets of n vertices. This graph is composed of isolated edge when $\mathrm{P}_{i,j} = 1$ and connected edges corresponding to $0 < \mathrm{P}_{i,j} < 1$. If i is such a connected vertex on the left (similarly for j on the right), because $\sum_j \mathrm{P}_{i,j} = 1$, there is necessarily at least two edges (i, j_1) and (i, j_2) emating from it (similarely on the right there are at least two converging edges (i_1, j) and (i_2, j)). This means that by following these connexions, one necessarily can extract a cycle (if not, one could alway extend it by the previous remarks) of the form

$$(i_1, j_1, i_2, j_2, \ldots, i_p, j_p), \quad \text{i.e.} \quad i_{p+1} = i_1.$$

We assume this cycle is the shortest one among all this (finite) ensemble of cycle. Along this cycle, the left-right and right-left edges satisfy

$$0 < \mathrm{P}_{i_s, j_s}, \mathrm{P}_{i_{s+1}, j_s} < 1.$$

The $(i_s)_s$ and $(j_s)_s$ are also all distincts because the cycle is the shortest. Lets pick

$$\varepsilon \overset{\text{def.}}{=} \min_{0 \leqslant s \leqslant p} \{\mathrm{P}_{i_s, j_s}, \mathrm{P}_{j_s, i_{s+1}}, 1 - \mathrm{P}_{i_s, j_s}, 1 - P_{j_s, i_{s+1}}\}$$

so that $0 < \varepsilon < 1$.We split the graph in two set of edges, left-right and right-left

$$\mathcal{A} \overset{\text{def.}}{=} \{(i_s, j_s)\}_{s=1}^p \quad \text{and} \quad \mathcal{B} \overset{\text{def.}}{=} \{(j_s, i_{s+1})\}_{s=1}^p.$$

We define then two matrices as

$$Q_{i,j} \overset{\text{def.}}{=} \begin{cases} \mathrm{P}_{i,j} & \text{if} \quad (i,j) \notin \mathcal{A} \cup \mathcal{B}, \\ \mathrm{P}_{i,j} + \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{A}, \\ \mathrm{P}_{i,j} - \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{B}, \end{cases} \quad \text{and} \quad R_{i,j} \overset{\text{def.}}{=} \begin{cases} \mathrm{P}_{i,j} & \text{if} \quad (i,j) \notin \mathcal{A} \cup \mathcal{B}, \\ \mathrm{P}_{i,j} - \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{A}, \\ \mathrm{P}_{i,j} + \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{B}, \end{cases}.$$

Because of the choice of ε, one has $0 \leqslant Q_{i,j}, R_{i,j} \leqslant 1$. Because each left-right edge in $\mathcal{A}$ is associated to a right-left edge in $\mathcal{B}$, (and the other way) the sum constraint on the row (and on the column) is maintain, so that $U, V \in \mathcal{B}_n$. Finally, note that $\mathrm{P} = (P+Q)/2$. □

By putting together Proposition 5 and Theorem 2, one obtains that for the discrete optimal problem with empirical measures, Monge and Kantoritch problems are equivalent.

Corollary 1 (Kantorovich for matching). *If $m = n$ and $\mathrm{a} = \mathrm{b} = \mathbb{1}_n$, then there exists an optimal solution for Problem (18) $\mathrm{P}_{\sigma^\star}$, which is a permutation matrix associated to an optimal permutation $\sigma^\star \in \mathrm{Perm}(n)$ for Problem (1).*

The following proposition shows that these problems result in fact in the same optimum, namely that one can always find a permutation matrix that minimizes Kantorovich's problem (18) between two uniform measures $\mathrm{a} = \mathrm{b} = \mathbb{1}_n/n$, which shows that the Kantorovich relaxation is *tight* when considered on assignment problems.

Remark 3 (General case). For general input measure, one does not have equivalence between Monge and Kantorovitch problems (since the Monge constraint is in general empty). But the support of the optimal coupling P still enjoys some strong regularity, in particular, it defines a cycle-free bipartite graph. This implies in particular that the resulting P matrix is sparse, for instance one can show that there are always solutions with less than $n + m - 1$ non-zero elements.

3.2 Relaxation for Arbitrary Measures

Continuous couplings. The definition of $\mathcal{L}_c$ in (18) is extended to arbitrary measures by considering couplings $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ which are joint distributions over the product space. The marginal constraint $\mathrm{P}\mathbb{1}_m = \mathrm{a}, \mathrm{P}\mathbb{1}_n = \mathrm{b}$ must be replaced by "integrated" versions, which are written $\pi_1 = \alpha$ and $\pi_2 = \beta$, where $(\pi_1, \pi_2) \in \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{Y})$ are the two marginals. They are defined as $\pi_1 \overset{\text{def.}}{=} P_{1\sharp}\pi$ and $\pi_2 \overset{\text{def.}}{=} P_{2\sharp}\pi$ the two marginals of π, which are defined using push-forward by the projectors $P_1(x, y) = x$ and $P_2(x, y) = y$.

A heuristic way to understand the marginal constraint $\pi_1 = \alpha$ and $\pi_2 = \beta$, which mimics the discrete case where one sums along the rows and columns is to write

$$\int_{\mathcal{Y}} \mathrm{d}\pi(x, y) = \mathrm{d}\alpha(x) \quad \text{and} \quad \int_{\mathcal{X}} \mathrm{d}\pi(x, y) = \mathrm{d}\beta(y),$$

and the mathematically rigorous way to write this, which corresponds to the change of variables formula, is

$$\forall\, (f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} f(x)\mathrm{d}\pi(x, y) = \int_{\mathcal{X}} f \mathrm{d}\alpha \quad \text{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}\pi(x, y) = \int_{\mathcal{Y}} g\mathrm{d}\beta.$$

Using (4), these marginal constraints are also equivalent to imposing that $\pi(A \times \mathcal{Y}) = \alpha(A)$ and $\pi(\mathcal{X} \times B) = \beta(B)$ for sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$.

Replacing continuous functions by indicator function, one can also rephrase this conservation of mass constraint as

$$\forall\, (A, B) \in \mathcal{X} \times \mathcal{Y}, \quad \pi(A \times \mathcal{Y}) = \alpha(A) \quad \text{and} \quad \pi(\mathcal{X} \times B) = \beta(B).$$

In the general case, the mass conservation constraint (17) should thus rewritten as a marginal constraint on joint probability distributions

$$\mathcal{U}(\alpha, \beta) \overset{\text{def.}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \;;\; \pi_1 = \alpha \quad \text{and} \quad \pi_2 = \beta \right\}. \tag{20}$$

The discrete case, when $\alpha = \sum_i \mathrm{a}_i \delta_{x_i}$, $\beta = \sum_j \mathrm{a}_j \delta_{x_j}$, the constraint $\pi_1 = \alpha$ and $\pi_2 = \beta$ necessarily imposes that π is discrete, supported on the set $\{(x_i, y_j)\}_{i,j}$, and thus has the form $\pi = \sum_{i,j} \mathrm{P}_{i,j} \delta_{(x_i, y_j)}$. The discrete formulation is thus a special case (and not some sort of approximation) of the continuous formulation.

The set $\mathcal{U}(\alpha, \beta)$ is always non-empty because it contains at least the tensor product coupling $\alpha \otimes \beta$ defined by $\mathrm{d}(\alpha \otimes \beta)(x, y) = \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$ i.e.

$$\forall\, h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x, y)\mathrm{d}(\alpha \otimes \beta)(x, y) = \int_{\mathcal{X}} \Big(\int_{\mathcal{Y}} h(x, y)\mathrm{d}\beta(y) \Big) \mathrm{d}\alpha(x) = \int_{\mathcal{X}} \Big(\int_{\mathcal{Y}} h(x, y)\mathrm{d}\alpha(x) \Big) \mathrm{d}\beta(y).$$

Indeed, $(\alpha \otimes \beta)_1 = \alpha$ since

$$\forall\, f \in \mathcal{C}(\mathcal{X}), \quad \int_{\mathcal{X}} f(x)\mathrm{d}(\alpha \otimes \beta)_1(x) = \int_{\mathcal{X}\times\mathcal{Y}} f(x) d\alpha(x)\mathrm{d}\beta(y) = \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) \int_{\mathcal{Y}} \mathrm{d}\beta = \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x)$$

because $\int_{\mathcal{Y}} \mathrm{d}\beta = 1$.

A very different (concentrated) type of coupling is defined when there exists a map $T : \mathcal{X} \to \mathcal{Y}$ such that $T_\sharp \alpha = \beta$ (i.e. the constraint set of Monge's problem (6)Â is non-empty). In this case, one has that $\pi = (\mathrm{Id}, T)_\sharp \alpha \in \mathcal{U}(\alpha, \beta)$. This coupling is defined though the integrated definition of push-forward as

$$\forall\, h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\mathrm{d}\pi(x,y) = \int_{\mathcal{X}} h(x, T(x))\mathrm{d}\alpha.$$

In particular, applying this formula to $h(x,y) = f(x)$ or $h(x,y) = g(y)$ shows that $\pi_1 = \alpha$ and $\pi_2 = \beta$.

Continuous Kantorovitch problem. The Kantorovich problem (18) is then generalized as

$$\mathcal{L}_c(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y). \tag{21}$$

This is an infinite-dimensional linear program over a space of measures.

On compact domain $(\mathcal{X}, \mathcal{Y})$, (21) always has a solution, because using the weak-* topology (so called weak topology of measures), the set of measure is compact, and a linear function with a continuous $c(x,y)$ is weak-* continuous. And the set of constraint is non empty, taking $\alpha \otimes \beta$. On non compact domain, one needs to impose moment condition on α and β.

Probabilistic interpretation. If we denote $X \sim \alpha$ the fact that the law of a random vector X is the probability distribution α, then the marginal constraint appearing in (21) is simply that π is the law of a couple (X, Y) and that its coordinates X and Y have laws α and β. The coupling π encodes the statistical dependency between X and Y. For instance, $\pi = \alpha \otimes \beta$ means that X and Y are independent, and it unlikely that such a coupling is optimal. Indeed as stated by Brenier's theorem, optimal coupling for a square Euclidean loss on contrary describe totally dependent variable, which corresponds to a coupling of the form $\pi = (\mathrm{Id}, T)_\sharp \alpha$ in which case $Y = T(X)$ where $T : \mathcal{X} \to \mathcal{Y}$ is a measurable map.

With this remark, problem (21) reads equivalently

$$\mathcal{L}_c(\alpha, \beta) = \min_{X\times\alpha, Y\sim\beta} \mathbb{E}(c(X,Y)). \tag{22}$$

Monge-Kantorovitch equivalence. The proof of Brenier theorem 1 (detailed in Section 5.3, Remark 7) to prove the existence of a Monge map actually studies Kantorovitch relaxation (and makes use of duality), and proves that this relaxation is tight in the sense that it has the same cost as Monge problem.

Indeed, it shows that the support of an optimal π is contained in the subdifferential $\partial\varphi$ of a convex function φ, which in general is a set-valued mapping . When α does not have a density, then φ is non-smooth and non-smooth points where $\alpha(\{x\}) > 0$ leads to mass splitting, for instance moving δ_0 to $(\delta_{-1} + \delta{+}1)/2$ can be achieved using $\varphi(x) = |x|$.

If α has a density, then this φ is differentiable α-almost everywhere and we denote $T = \nabla\varphi$ the unique optimal transport (which is a valid definition almost everywhere and one can use any value at point of non differentiability), then the coupling

$$\pi = (\mathrm{Id}, T)_\sharp \alpha \quad \text{i.e.} \quad \forall\, h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h\mathrm{d}\pi = \int_{\mathcal{X}} h(x, T(x))\mathrm{d}\alpha(x)$$

is optimal. In term of random vector, denoting (X, Y) a random vector with law π, it means that any such optimal random vector satisfies $Y = T(X)$ where $X \sim \alpha$ (and of course $T(X) \sim \beta$ by the marginal constraint).

This key result is similar to Birkoff-von-Neumann Theorem 1 in the sense that it provides conditions ensuring the equivalence between Monge and Kantorovitch problems (note however that Birkoff-von-Neumann does not implies uniqueness). Note however that the settings are radically difference (one is fully discrete while the other requires the sources to be "continuous", i.e. to have a density).

3.3 Metric Properties

OT defines a distance. An important feature of OT is that it defines a distance between histograms and probability measures as soon as the cost matrix satisfies certain suitable properties. Indeed, OT can be understood as a canonical way to lift a ground distance between points to a distance between histogram or measures. The proof of this result rely on a "gluing lemma", which we first prove in the discrete case.

Lemma 1 (Discrete gluing lemma)**.** *Given* $(\mathrm{a},\mathrm{b},\mathrm{c}) \in \Sigma_n \times \Sigma_p \times \Sigma_m$ *Let* $\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})$ *and* $\mathrm{Q} \in \mathrm{U}(\mathrm{b},\mathrm{c})$*. Then there exists at least a 3-D tensor coupling* $\mathrm{S} \in \mathbb{R}_+^{n\times p\times m}$ *such that the 2-D marginals satisfies*

$$\sum_k \mathrm{S}_{i,j,k} = \mathrm{P}_{i,j} \quad \textit{and} \quad \sum_i \mathrm{S}_{i,j,k} = \mathrm{Q}_{j,k}.$$

Note that this implies that the three 1-D marginals of S *are* $(\mathrm{a},\mathrm{b},\mathrm{c})$*.*

Proof. One verifies that

$$\mathrm{S}_{i,j,k} = \begin{cases} \frac{\mathrm{P}_{i,j}\mathrm{Q}_{j,k}}{\mathrm{b}_j} & \text{if} \quad \mathrm{b}_j \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{23}$$

is acceptable. Indeed, if $\mathrm{b}_j \neq 0$

$$\sum_k \mathrm{S}_{i,j,k} = \sum_k \frac{\mathrm{P}_{i,j}\mathrm{Q}_{j,k}}{\mathrm{b}_j} = \frac{\mathrm{P}_{i,j}}{\mathrm{b}_j}(\mathrm{Q}\mathbb{1}_m)_j = \frac{\mathrm{P}_{i,j}}{\mathrm{b}_j}\mathrm{b}_j.$$

If $\mathrm{b}_j = 0$, then necessarily $\mathrm{P}_{i,j} = 0$ and $\sum_k \mathrm{S}_{i,j,k} = 0 = \mathrm{P}_{i,j}$. □

Proposition 6. *We suppose* $n = m$*, and that for some* $p \geqslant 1$*,* $\mathrm{C} = \mathrm{D}^p = (\mathrm{D}_{i,j}^p)_{i,j} \in \mathbb{R}^{n\times n}$ *where* $\mathrm{D} \in \mathbb{R}_+^{n\times n}$ *is a distance on* $[\![n]\!]$*,* i.e.

1. $\mathrm{D} \in \mathbb{R}_+^{n\times n}$ *is symmetric;*
2. $\mathrm{D}_{i,j} = 0$ *if and only if* $i = j$*;*
3. $\forall\, (i,j,k) \in [\![n]\!]^3, \mathrm{D}_{i,k} \leqslant \mathrm{D}_{i,j} + \mathrm{D}_{j,k}$*.*

Then

$$\mathrm{W}_p(\mathrm{a},\mathrm{b}) \overset{\text{def.}}{=} \mathrm{L}_{\mathrm{D}^p}(\mathrm{a},\mathrm{b})^{1/p} \tag{24}$$

(note that W_p *depends on* D*) defines the p-Wasserstein distance on* Σ_n*,* i.e. W_p *is symmetric, positive,* $\mathrm{W}_p(\mathrm{a},\mathrm{b}) = 0$ *if and only if* $\mathrm{a} = \mathrm{b}$*, and it satisfies the triangle inequality*

$$\forall\, \mathrm{a},\mathrm{b},\mathrm{c} \in \Sigma_n, \quad \mathrm{W}_p(\mathrm{a},\mathrm{c}) \leqslant \mathrm{W}_p(\mathrm{a},\mathrm{b}) + \mathrm{W}_p(\mathrm{b},\mathrm{c}).$$

Proof. For the symmetry, since D^p is symmetric, we use the fact that if $\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})$ is optimal for $\mathrm{W}_p(\mathrm{a},\mathrm{b})$, then $\mathrm{P}^\top \in \mathrm{U}(\mathrm{b},\mathrm{a})$ is optimal for $\mathrm{W}_p(\mathrm{b},\mathrm{a})$. For the definiteness, since $\mathrm{C} = \mathrm{D}^p$ has a null diagonal, $\mathrm{W}_p(\mathrm{a},\mathrm{a}) = 0$, with corresponding optimal transport matrix $\mathrm{P}^\star = \mathrm{diag}(\mathrm{a})$; by the positivity of all off-diagonal elements of D^p, $\mathrm{W}_p(\mathrm{a},\mathrm{b}) > 0$ whenever $\mathrm{a} \neq \mathrm{b}$ (because in this case, an admissible coupling necessarily has a non-zero element outside the diagonal).

To prove the triangle inequality of Wasserstein distances for arbitrary measures, we consider $\mathrm{a},\mathrm{b},\mathrm{c} \in \Sigma_n$, and let P and Q be two optimal solutions of the transport problems between a and b, and b and c respectively. We use the gluing Lemma 1 which defines $\mathrm{S} \in \mathbb{R}_+^{n^3}$ with marginals $\sum_k \mathrm{S}_{\cdot,\cdot,k} = \mathrm{P}$ and $\sum_i \mathrm{S}_{i,\cdot,\cdot} = \mathrm{Q}$. We define $\mathrm{R} = \sum_j \mathrm{S}_{\cdot,j,\cdot}$, which is an element of $\mathrm{U}(\mathrm{a},\mathrm{c})$. Note that if one assumes $\mathrm{b} > 0$ then $\mathrm{R} = \mathrm{P}\,\mathrm{diag}(1/\mathrm{b})\mathrm{Q}$.

The triangle inequality follows from

$$\begin{aligned}
\mathrm{W}_p(\mathrm{a},\mathrm{c}) &= \Big(\min_{\tilde{\mathrm{R}} \in \mathrm{U}(\mathrm{a},\mathrm{c})} \langle \tilde{\mathrm{R}},\, \mathrm{D}^p \rangle \Big)^{1/p} \leqslant \langle \mathrm{R},\, \mathrm{D}^p \rangle^{1/p} \\
&= \Big(\sum_{i,j,k} \mathrm{D}_{ik}^p \sum_j \mathrm{S}_{i,j,k} \Big)^{1/p} \leqslant \Big(\sum_{i,j,k} \Big(\mathrm{D}_{ij} + \mathrm{D}_{j,k} \Big)^p \mathrm{S}_{i,j,k} \Big)^{1/p} \\
&\leqslant \Big(\sum_{i,j,k} \mathrm{D}_{ij}^p \mathrm{S}_{i,j,k} \Big)^{1/p} + \Big(\sum_{i,j,k} \mathrm{D}_{j,k}^p \mathrm{S}_{i,j,k} \Big)^{1/p} \\
&= \Big(\sum_{i,j} \mathrm{D}_{i,j}^p \sum_k \mathrm{S}_{i,j,k} \Big)^{1/p} + \Big(\sum_{j,k} \mathrm{D}_{j,k}^p \sum_i \mathrm{S}_{i,j,k} \Big)^{1/p} \\
&= \Big(\sum_{i,j} \mathrm{D}_{i,j}^p \mathrm{P}_{i,j} \Big)^{1/p} + \Big(\sum_{j,k} \mathrm{D}_{j,k}^p \mathrm{Q}_{j,k} \Big)^{1/p} = \mathrm{W}_p(\mathrm{a},\mathrm{b}) + \mathrm{W}_p(\mathrm{b},\mathrm{b}).
\end{aligned}$$

The first inequality is due to the sub-optimality of S, the second is the usual triangle inequality for elements in D, and the third comes from Minkowski's inequality. □

Proposition 6 generalizes from histogram to arbitrary measures that need not be discrete. For this, one needs the following general gluing lemma.

Lemma 2 (Gluing lemma)**.** *Let $(\alpha,\beta,\gamma) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y}) \times \mathcal{M}_+^1(\mathcal{Y})$ where $(\mathcal{X},\mathcal{Y},\mathcal{Z})$ are three polish spaces (i.e. separable topological space which can be metrized using a distance which makes it a complete metric space). Given $\pi \in \mathcal{U}(\alpha,\beta)$ and $\xi \in \mathcal{U}(\beta,\gamma)$, then there exists at least a tensor coupling measure $\sigma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ such that*

$$(P_{\mathcal{X},\mathcal{Y}})_\sharp \sigma = \pi \quad \text{and} \quad (P_{\mathcal{Y},\mathcal{Z}})_\sharp \sigma = \xi$$

where we denoted the projector $P_{\mathcal{X},\mathcal{Y}}(x,y,z) = (x,y)$ and $P_{\mathcal{Y},\mathcal{Z}}(x,y,z) = (y,z)$.

Proof. The proof of this fundamental result is involved since it requires using the disintegration of measure (which corresponds to conditional probabilities). The disintegration of measures is applicable because the spaces are polish. We disintegrate π and ξ against β to obtain two families $(\pi_y)_{y\in\mathcal{Y}}$ and $(\xi_y)_{y\in\mathcal{Y}}$ of probability distributions on $\mathcal{X}$ and $\mathcal{Z}$. These families are defined by the fact that

$$\forall\, h \in \mathcal{C}(\mathcal{X}\times\mathcal{Y}), \quad \int_{\mathcal{Y}} \Big(\int_{\mathcal{X}} h(x,y) \mathrm{d}\pi_y(x) \Big) \mathrm{d}\beta(y) = \int h(x,y) \mathrm{d}\pi(x,y).$$

and similarly for ξ. When $\beta = \sum_i \mathrm{b}_j \delta_{y_j}$ and $\pi = \sum_{i,j} \mathrm{P}_{i,j} \delta_{y_j}$, then this conditional distribution is defined on the support of β as $\pi_{y_j} = \sum_i \frac{\mathrm{P}_{i,j}}{\mathrm{b}_j} \delta_{x_i}$ (and similarly for ξ). Then one defines the glued measure informally "$\sigma(x,y,z) = \pi_y(x)\xi_y(z)\beta(y)$", which formally reads

$$\forall\, g \in \mathcal{C}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}), \quad \int g(x,y,z) \mathrm{d}\sigma(x,y,z) = \int g(x,y,z) \mathrm{d}\pi_y(x) \mathrm{d}\xi_y(z) \mathrm{d}\beta(y).$$

For discrete measures, this matches the definition (23), since $\sigma = \sum_{i,j,k} \mathrm{S}_{i,j,k} \delta_{x_i,y_j,z_k}$ where

$$\mathrm{S}_{i,j,k} = \frac{\mathrm{P}_{i,j}}{\mathrm{b}_j} \frac{\mathrm{Q}_{j,k}}{\mathrm{b}_j} \mathrm{b}_j.$$

□

Using this gluing lemma, we can now construct the Wasserstein distance in the general setting of arbitrary distributions on a Polish space.

Proposition 7. *We assume $\mathcal{X} = \mathcal{Y}$, and that for some $p \geqslant 1$, $c(x,y) = d(x,y)^p$ where d is a distance on $\mathcal{X}$,* i.e.

(i) $d(x,y) = d(y,x) \geqslant 0$;

(ii) $d(x,y) = 0$ *if and only if* $x = y$;

(ii) $\forall\,(x,y,z) \in \mathcal{X}^3, d(x,z) \leqslant d(x,y) + d(y,z)$.

Then

$$\mathcal{W}_p(\alpha,\beta) \overset{\text{def.}}{=} \mathcal{L}_{d^p}(\alpha,\beta)^{1/p} \tag{25}$$

(note that $\mathcal{W}_p$ depends on d) defines the p-Wasserstein distance on $\mathcal{X}$, i.e. *$\mathcal{W}_p$ is symmetric, positive, $\mathcal{W}_p(\alpha,\beta) = 0$ if and only if $\alpha = \beta$, and it satisfies the triangle inequality*

$$\forall\,(\alpha,\beta,\gamma) \in \mathcal{M}_+^1(\mathcal{X})^3, \quad \mathcal{W}_p(\alpha,\gamma) \leqslant \mathcal{W}_p(\alpha,\beta) + \mathcal{W}_p(\beta,\gamma).$$

Proof. The symmetry follows from the fact that since d is symmetric, if $\pi(x,y)$ is optimal for $\mathcal{L}_{d^p}(\alpha,\beta)$, then $\pi(y,x) \in \mathcal{U}(\beta,\alpha)$ is optimal for $\mathcal{L}_{d^p}(\beta,\alpha)$. If $\mathcal{L}_{d^p}(\alpha,\beta) = 0$, then necessarily an optimal coupling π is supported on the diagonal $\Delta \overset{\text{def.}}{=} \{(x,x)\}_x \subset \mathcal{X}^2$. We denote $\lambda(x)$ the corresponding measure on the diagonal, i.e. such that $\int h(x,y)\mathrm{d}\pi(x,y) = \int h(x,x)\mathrm{d}\lambda(x)$. Then since $\pi \in \mathcal{U}(\alpha,\beta)$ necessarily $\lambda = \alpha$ and $\lambda = \beta$ so that $\alpha = \beta$.

For the triangle inequality, we consider optimal couplings $\pi \in \mathcal{U}(\alpha,\beta)$ and $\xi \in \mathcal{U}(\beta,\gamma)$ and we glue them according to the Lemma 2. We define the composition of the two couplings (π,ξ) as $\rho \overset{\text{def.}}{=} (P_{\mathcal{X},\mathcal{Z}})_\sharp\sigma$. Note that if π and ξ are coupling induced by two Monge maps $T_\mathcal{X}(x)$ and $T_\mathcal{Y}(y)$, then ρ is itself induced by the Monge map $T_\mathcal{Y} \circ T_\mathcal{X}$, so that this notion of composition of coupling generalizes the composition of maps. The triangular inequality follows from

$$\begin{aligned}
\mathcal{W}_p(\alpha,\gamma) &\leqslant \Big(\int_{\mathcal{X}\times\mathcal{Z}} d(x,z)^p \mathrm{d}\rho(x,z)\Big)^{1/p} = \Big(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} d(x,z)^p \mathrm{d}\sigma(x,y,z)\Big)^{1/p} \\
&\leqslant \Big(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} (d(x,y)+d(y,z))^p \mathrm{d}\sigma(x,y,z)\Big)^{1/p} \\
&\leqslant \Big(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} d(x,y)^p \mathrm{d}\sigma(x,y,z)\Big)^{1/p} + \Big(\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} d(y,z)^p \mathrm{d}\sigma(x,y,z)\Big)^{1/p} \\
&= \Big(\int_{\mathcal{X}\times\mathcal{Y}} d(x,y)^p \mathrm{d}\pi(x,y,z)\Big)^{1/p} + \Big(\int_{\mathcal{Y}\times\mathcal{Z}} d(y,z)^p \mathrm{d}\xi(y,z)\Big)^{1/p} = \mathcal{W}_p(\alpha,\beta) + \mathcal{W}_p(\beta,\gamma).
\end{aligned}$$

□

This distance $\mathcal{W}_p$ defined though Kantorovitch problem (25) should be contrasted with the distance $\tilde{\mathcal{W}}$ obtained using Monge's problem (7). Kantorovitch distance is always finite, while Monge's one might be infinite if the constraint set $\{T\ ;\ T_\sharp\alpha = \beta\}$ is empty. In fact, one can show that as soon as this constraint set is non-empty, and even if no optimal T exists, then one has $\mathcal{W}_p = \tilde{\mathcal{W}}_p$, which is a non-trivial result. Kantorovitch distance should thus be seen as a (convex) relaxation of Monge's distance, which behave in a much nicer way, as we will explore next (it is continuous with respect to the convergence in law topology.

Convergence in law topology. Let us first note that on a bounded metric space, all $\mathcal{W}_p$ distance defines the same topology (although they are not equivalent, the notion of converging sequence is the same).

Proposition 8. *One has for $p \leqslant q$*

$$\mathcal{W}_p(\alpha,\beta) \leqslant \mathcal{W}_q(\alpha,\beta) \leqslant \mathrm{diam}(\mathcal{X})^{\frac{q-p}{q}}\,\mathcal{W}_p(\alpha,\beta)^{\frac{q}{p}}$$

where $\mathrm{diam}(\mathcal{X}) \triangleq \sup_{x,y} d(x,y)$.

Proof. The left inequality follows from Jensen inequality, $\varphi(\int c(x,y)\mathrm{d}\pi(x,y)) \leqslant \int \varphi(c(x,y))\mathrm{d}\pi(x,y)$, applied to any probability distribution π and to the convex function $\varphi(r) = r^{q/p}$ to $c(x,y) = \|x-y\|^p$, so that one gets

$$\left(\int \|x-y\|^p \mathrm{d}\pi(x,y)\right)^{\frac{q}{p}} \leqslant \int \|x-y\|^q \mathrm{d}\pi(x,y).$$

The right inequality follows from

$$\|x-y\|^q \leqslant \mathrm{diam}(\mathcal{X})^{q-p}\|x-y\|^p.$$

□

The Wasserstein distance $\mathcal{W}_p$ has many important properties, the most important one being that it is a weak distance, *i.e.* it allows to compare singular distributions (for instance discrete ones) and to quantify spatial shift between the supports of the distributions. This corresponds to the notion of weak* convergence.

Definition 2 (Weak* topology)**.** *$(\alpha_k)_k$ converges weakly* to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any continuous function $f \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} f\mathrm{d}\alpha_k \to \int_{\mathcal{X}} f\mathrm{d}\alpha$.*

In term of random vectors, if $X_n \sim \alpha_n$ and $X \sim \alpha$ (not necessarily defined on the same probability space), the weak* convergence corresponds to the convergence in law of X_n toward X.

Remark 4 (Central limit theorem)*.* The central limit theorem states that if $(X_1, \ldots, X_n)$ are i.i.d. distribution with finite second order moments, assuming for instance $\mathbb{E}(X_i X_i^\top) = \mathrm{Id}$, the rescaled average $Z_n \triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ converges in law toward a Gaussian $\mathcal{N}(0, \mathrm{Id})$. This means that the measure α_n representing the law of Z_n converges weak* toward the measure α of the centered normalized Gaussian.

Definition 3 (Strong topology)**.** *The simplest distance on Radon measures is the total variation norm, which is the dual norm of the L^∞ norm on $\mathcal{C}(\mathcal{X})$ and whose topology is often called the "strong" topology*

$$\|\alpha-\beta\|_{\mathrm{TV}} \overset{\text{def.}}{=} \sup_{\|f\|_\infty \leqslant 1} \int f\mathrm{d}(\alpha-\beta) = |\alpha-\beta|(\mathcal{X})$$

where $|\alpha-\beta|(\mathcal{X})$ is the mass of the absolute value of the difference measure. When $\alpha-\beta = \rho\mathrm{d}x$ has a density, then $\|\alpha-\beta\|_{\mathrm{TV}} = \int |\rho(x)|\mathrm{d}x = \|\rho\|_{L^1(\mathrm{d}x)}$ is the L^1 norm associated to $\mathrm{d}x$. When $\alpha-\beta = \sum_i u_i\delta_{z_i}$ is discrete, then $\|\alpha-\beta\|_{\mathrm{TV}} = \sum_i |u_i| = \|u\|_{\ell^1}$ is the discrete ℓ^1 norm.

The following proposition shows that the TV norm can be seen as a Wasserstein distance, but for a "degenerate" 0/1 metric.

Proposition 9. *Denoting d the 0/1 distance such that $d(x,x) = 0$ and $d(x,y) = 1$ if $x \neq y$, then*

$$\mathcal{W}_p(\alpha,\beta)^p = \frac{1}{2}\|\alpha-\beta\|_{\mathrm{TV}}.$$

Proof. For the sake of simplicity, we do the proof for discrete measures with weights (a, b) and without loss of generality assume they have the same support $(x_i)_i$ and we denote $\mathrm{D} \triangleq (d(x_i,x_j))_{i,j}$ which is 0 on the diagonal and one outside. Also since $d^p = d$ we consider $p = 1$. We denote $\mathrm{c}_i = \min(\mathrm{a}_i, \mathrm{b}_i)$. By conservation of mass, for $\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})$, $\mathrm{P}_{i,i} \leqslant \mathrm{c}_i$, thus

$$\mathcal{W}_1(\alpha,\beta) = \inf_{\mathrm{P}\mathbb{1}=\mathrm{a}, \mathrm{P}^\top\mathbb{1}=\mathrm{b}} \langle \mathrm{P}, \mathrm{D}\rangle = \sum_{i\neq j} \mathrm{P}_{i,j} = 1 - \sum_i \mathrm{P}_{i,i} = 1 - \sum_i \mathrm{c}_i.$$

We need to show that this bound is tight, namely to construct $\hat{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})$ such that $\mathrm{diag}(\hat{P}) = \mathrm{c}$. Let

$$\bar{\mathrm{a}} \triangleq \mathrm{a} - \mathrm{c} = (\mathrm{a}-\mathrm{b})_+ \geqslant 0 \quad \text{and} \quad \bar{\mathrm{b}} \triangleq \mathrm{b} - \mathrm{c} = (\mathrm{b}-\mathrm{a})_+ \geqslant 0$$

One has

$$\frac{\bar{a}\otimes\bar{b}}{\langle\bar{a},\,\mathbb{1}\rangle}\in \mathrm{U}(\bar{\mathrm{a}},\bar{\mathrm{b}})$$

and we remark that $\langle\bar{a},\,\mathbb{1}\rangle = \langle\bar{b},\,\mathbb{1}\rangle = 1-\langle \mathrm{c},\,\mathbb{1}\rangle$. Thus denoting

$$\hat{\mathrm{P}} \triangleq \mathrm{diag}(\mathrm{c}) + \frac{\bar{a}\otimes\bar{b}}{\langle\bar{a},\,\mathbb{1}\rangle}\in \mathrm{U}(\bar{\mathrm{a}},\bar{\mathrm{b}}) \geqslant 0$$

satisfies

$$\mathrm{P}\mathbb{1} = \mathrm{c}+\bar{\mathrm{a}} = \mathrm{a} \quad \text{and} \quad \mathrm{P}^\top\mathbb{1} = \mathrm{c}+\bar{\mathrm{b}} = \mathrm{b}$$

so that $\hat{P}\in \mathrm{U}(\mathrm{a},\mathrm{b})$ is a coupling so that $\mathrm{diag}(\hat{P}) = \mathrm{diag}(\mathrm{c})$ since $\mathrm{diag}(\bar{a}\otimes\bar{b}) = 0$. We thus conclude that

$$\mathrm{W}_1(\mathrm{a},\mathrm{b}) = \langle \mathrm{D},\,\hat{\mathrm{P}}\rangle = \sum_{i,j}\frac{\mathrm{a}_i\mathrm{b}_j}{\langle\bar{a},\,\mathbb{1}\rangle} = \sum_i \bar{\mathrm{a}}_i = \sum_i \bar{\mathrm{b}}_i == \frac{1}{2}\sum_i \bar{\mathrm{a}}_i+\bar{\mathrm{b}}_i = \frac{1}{2}\|\mathrm{a}-\mathrm{b}\|_{\mathrm{TV}}.$$

□

In the special case of Diracs, having $\int f\mathrm{d}\delta_{x_n} = f(x_n)\to\int f\mathrm{d}\delta_x = f(x)$ for any continuous f is equivalent to $x_n\to x$. One can then contrast the strong topology with the Wasserstein distance, if $x_n\neq x$,

$$\|\delta_{x_n}-\delta_x\|_{\mathrm{TV}} = 2 \quad \text{and} \quad \mathcal{W}_p(\delta_{x_n},\delta_x) = d(x_n,x).$$

This shows that for the strong topology, Diracs never converge, while they do converge for the Wasserstein distance. In fact it is a powerful property of the Wasserstein distance, which is regular with respect to the weak* topology, and metrizes it.

Proposition 10. *If $\mathcal{X}$ is compact, $\alpha_k\rightharpoonup\alpha$ if and only if $\mathcal{W}_p(\alpha_k,\alpha)\to 0$.*

The proof of this proposition requires the use of duality, and is delayed to later, see Proposition 2. On non-compact spaces, one needs also to impose the convergence of the moments up to order p. Note that there exists alternative distances which also metrize weak convergence. The simplest one are Hilbertian kernel norms, which are detailed in Section 6.4.

Another example of such a weak convergence is the fact that on $\mathcal{X}=\mathbb{R}$

$$\frac{1}{n}\sum_{k=1}^n \delta_{k/n}\rightharpoonup \mathcal{U}_{[0,1]}$$

(convergence toward the uniform measure on $[0,1]$), which comes from the convergence of Riemann sums

$$\forall\, f\in\mathcal{C}(\mathbb{R}),\quad \frac{1}{n}\sum_{k=1}^n f(k/n)\longrightarrow\int_0^1 f(x)\mathrm{d}x.$$

In contrary, one has that for all n, since the two measure are mutually singular

$$\|\frac{1}{n}\sum_{k=1}^n \delta_{k/n}-\mathcal{U}_{[0,1]}\|_{\mathrm{TV}} = \|\frac{1}{n}\sum_{k=1}^n \delta_{k/n}\|_{\mathrm{TV}}+\|\mathcal{U}_{[0,1]}\|_{\mathrm{TV}} = 2$$

so that there is no strong convergence.

The following proposition show that

On discrete space, the strong and the weak topology coincide, and the following proposition relates the TV and Wasserstein distance together.

Proposition 11. *One has*

$$\frac{d_{\min}}{2}\|\alpha-\beta\|_{\mathrm{TV}} \leqslant \mathcal{W}_1(\alpha,\beta) \leqslant \frac{d_{\max}}{2}\|\alpha-\beta\|_{\mathrm{TV}} \quad \textit{where} \quad \begin{cases} d_{\min} \overset{\text{def.}}{=} \inf_{x\neq y} d(x,y) \\ d_{\max} \overset{\text{def.}}{=} \sup_{x,y} d(x,y) \end{cases}$$

Proof. We denote $d_0(x,y)$ the distance such that $d_0(x,x)=0$ and $d_0(x,y)=1$ for $x\neq y$. One has

$$d_{\min} d_0(x,y) \leqslant d(x,y) \leqslant d_{\max} d_0(x,y)$$

so that integrating this against any $\pi \in \mathcal{U}(\alpha,\beta)$ and taking the minimum among those π gives the result using Proposition (9). □

This bound is sharp, as this can be observed by taking $\alpha=\delta_x$ and $\beta=\delta_y$, in which case the bound simply reads, if $x\neq y$

$$d_{\min} \leqslant d(x,y) \leqslant d_{\max}.$$

This shows that the ratio between the two distances can blows as $d_{\max}/d_{\min}$ increases, and on non discrete space, if $d_{\min}=0$, then the two distance are not equivalent, which is inline with the fact that the strong and the weak topology do not coincide.

Remark 5 (Berry-Esseen theorem). The Wasserstein distance is a natural candidate to quantify the convergence in law in the central limit theorem (Remark 4). To obtain rates, one needs further assumption on the random vector, and Berry-Esseen theorem ensure that if $\mathbb{E}(\|X_i\|^3)<+\infty$, then $\mathcal{W}_p(\alpha_k,\alpha)=O(1/\sqrt{n})$.

Applications and implications Applications for having a geometric distance : barycenters, shape registration loss functions, density fitting. The typical setup is to fit a parametric measure $\theta \mapsto \alpha_\theta$ to an (empirical) measure β by minimizing the function $\theta \mapsto \mathcal{W}_p(\alpha_\theta,\beta)$.

4 Sinkhorn

4.1 Entropic Regularization for Discrete Measures

Entropic Regularization for Discrete Measures. The idea of the entropic regularization of optimal transport is to use the Shannon-Boltzmann entropy

$$\mathrm{H}(\mathrm{P}) \triangleq -\sum_{i,j} \mathrm{P}_{i,j}\log(\mathrm{P}_{i,j}),$$

with the convention $0\log(0)$ as a regularizing function to obtain approximate solutions to the original transport problem (18)

$$\mathrm{L}_{\mathrm{C}}^{\varepsilon}(\mathrm{a},\mathrm{b}) \overset{\text{def.}}{=} \min_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{P},\, \mathrm{C}\rangle - \varepsilon \mathrm{H}(\mathrm{P}). \tag{26}$$

This is a strictly convex optimization problem. Indeed, the function $-\mathrm{H}$ is strongly convex, because its hessian is $-\partial^2\mathrm{H}(\mathrm{P})=\mathrm{diag}(1/\mathrm{P}_{i,j})$ and $\mathrm{P}_{i,j}\leqslant 1$.

Smoothing effect. Since the objective is a ε-strongly convex function, problem 26 has a unique optimal solution. This smoothing, beyond providing uniqueness, actually leads to $\mathrm{L}_{\mathrm{C}}^{\varepsilon}(\mathrm{a},\mathrm{b})$ being a smooth function of a,b and C. The effect of the entropy is to act as a barrier function for the positivity constraint. As we will show next, this forces the solution P to be strictly positive on the support of $\mathrm{a}\otimes\mathrm{b}$.

One has the following convergence property.

Proposition 12 (Convergence with ε). *The unique solution P_ε of (26) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$\mathrm{P}_\varepsilon \overset{\varepsilon\to 0}{\longrightarrow} \operatorname*{argmin}_{\mathrm{P}} \{-\mathrm{H}(\mathrm{P}) \,;\, \mathrm{P}\in \mathrm{U}(\mathrm{a},\mathrm{b}), \langle \mathrm{P},\, \mathrm{C}\rangle = \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b})\} \tag{27}$$

so that in particular

$$\mathrm{L}^\varepsilon_\mathrm{C}(\mathrm{a},\mathrm{b}) \overset{\varepsilon\to 0}{\longrightarrow} \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b}).$$

One has

$$\mathrm{P}_\varepsilon \overset{\varepsilon\to\infty}{\longrightarrow} \mathrm{a}\otimes\mathrm{b}. \tag{28}$$

Proof. **Case $\varepsilon\to 0$.** We consider a sequence $(\varepsilon_\ell)_\ell$ such that $\varepsilon_\ell\to 0$ and $\varepsilon_\ell>0$. We denote P_ℓ the solution of (26) for $\varepsilon=\varepsilon_\ell$. Since $\mathrm{U}(\mathrm{a},\mathrm{b})$ is bounded, we can extract a sequence (that we do not relabel for sake of simplicity) such that $\mathrm{P}_\ell\to\mathrm{P}^\star$. Since $\mathrm{U}(\mathrm{a},\mathrm{b})$ is closed, $\mathrm{P}^\star\in\mathrm{U}(\mathrm{a},\mathrm{b})$. We consider any P such that $\langle \mathrm{C},\,\mathrm{P}\rangle = \mathrm{L}_\mathrm{C}(\mathrm{a},\mathrm{b})$. By optimality of P and P_ℓ for their respective optimization problems (for $\varepsilon=0$ and $\varepsilon=\varepsilon_\ell$), one has

$$0 \leqslant \langle \mathrm{C},\,\mathrm{P}_\ell\rangle - \langle \mathrm{C},\,\mathrm{P}\rangle \leqslant \varepsilon_\ell(\mathrm{KL}(\mathrm{P}_\ell|\mathrm{a}\otimes\mathrm{b}) - \mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b})). \tag{29}$$

Since KL is continuous, taking the limit $\ell\to+\infty$ in this expression shows that $\langle \mathrm{C},\,\mathrm{P}^\star\rangle = \langle \mathrm{C},\,\mathrm{P}\rangle$ so that $\mathrm{P}^\star$ is a feasible point of (27). Furthermore, dividing by ε_ℓ in (29) and taking the limit shows that $\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) \leqslant \mathrm{KL}(\mathrm{P}^\star|\mathrm{a}\otimes\mathrm{b})$, which shows that $\mathrm{P}^\star$ is a solution of (27). Since the solution $\mathrm{P}^\star_0$ to this program is unique by strict convexity of $\mathrm{KL}(\cdot|\mathrm{a}\otimes\mathrm{b})$, one has $\mathrm{P}^\star=\mathrm{P}^\star_0$, and the whole sequence is converging.

Case $\varepsilon\to+\infty$. Evaluating at $\mathrm{a}\otimes\mathrm{b}$ (which belongs to the constraint set $\mathrm{U}(\mathrm{a},\mathrm{b})$) the energy, one has

$$\langle \mathrm{C},\,\mathrm{P}_\varepsilon\rangle + \varepsilon\mathrm{KL}(\mathrm{P}_\varepsilon|\mathrm{a}\otimes\mathrm{b}) \leqslant \langle \mathrm{C},\,\mathrm{a}\otimes\mathrm{b}\rangle + \varepsilon\times 0$$

and since $\langle \mathrm{C},\,\mathrm{P}_\varepsilon\rangle\geqslant 0$, this leads to

$$\mathrm{KL}(\mathrm{P}_\varepsilon|\mathrm{a}\otimes\mathrm{b}) \leqslant \varepsilon^{-1}\langle \mathrm{C},\,\mathrm{a}\otimes\mathrm{b}\rangle \leqslant \frac{\|\mathrm{C}\|_\infty}{\varepsilon}$$

so that $\mathrm{KL}(\mathrm{P}_\varepsilon|\mathrm{a}\otimes\mathrm{b})\to 0$ and thus $\mathrm{P}_\varepsilon\to\mathrm{a}\otimes\mathrm{b}$ since KL is a valid divergence. □

4.2 Sinkhorn's Algorithm

The following proposition shows that the solution of (26) has a specific form, which can be parameterized using $n+m$ variables. That parameterization is therefore essentially dual, in the sense that a coupling P in $\mathrm{U}(\mathrm{a},\mathrm{b})$ has nm variables but $n+m$ constraints.

Proposition 13. *P is the unique solution to (26) if and only if there exists $(\mathrm{u},\mathrm{v})\in\mathbb{R}_+^n\times\mathbb{R}_+^m$ such that*

$$\forall\,(i,j)\in[\![n]\!]\times[\![m]\!],\quad \mathrm{P}_{i,j} = \mathrm{u}_i\mathrm{K}_{i,j}\mathrm{v}_j \quad \textit{where} \quad \mathrm{K}_{i,j}\overset{\text{def.}}{=} e^{-\frac{-\mathrm{C}_{i,j}}{\varepsilon}}, \tag{30}$$

and $\mathrm{P}\in\mathcal{U}(\mathrm{a},\mathrm{b})$.

Proof. Without loss of generality, we assume $\mathrm{a}_i,\mathrm{b}_j>0$ (otherwise, we can set the corresponding u_i or v_j to 0).

The first thing to prove is that if $\mathrm{P}^\star$ is the solution (which is unique by strict convexity of the entropy) then $\mathrm{P}^\star_{i,j}>0$ for all (i,j). Indeed, if $\mathrm{P}^\star_{i,j}=0$, then we can consider $\mathrm{P}_t=(1-t)\mathrm{P}^\star+t\mathrm{a}\otimes\mathrm{b}$, which satisfies the marginal constraint for $t\in[0,1]$. One then check that, denoting $\mathcal{E}(\mathrm{P})\overset{\text{def.}}{=}\langle \mathrm{P},\,\mathrm{C}\rangle+\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b})$ the objective function, and $f(t)\overset{\text{def.}}{=}\mathcal{E}(\mathrm{P}_t)$, then $f'(0)=-\infty$, so that for t small enough, $\mathcal{E}(\mathrm{P}_t)<\mathcal{E}(\mathrm{P}^\star)$ which is a contradiction.

We can thus ignore the positivity constraint when introducing two dual variables $\mathrm{f}\in\mathbb{R}^n,\mathrm{g}\in\mathbb{R}^m$ for each marginal constraint, so that the Lagrangian of (26) reads

$$\mathcal{E}(\mathrm{P},\mathrm{f},\mathrm{g}) = \langle \mathrm{P},\,\mathrm{C}\rangle + \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) + \langle \mathrm{f},\,\mathrm{a}-\mathrm{P}\mathbb{1}_m\rangle + \langle \mathrm{g},\,\mathrm{b}-\mathrm{P}^\top\mathbb{1}_n\rangle.$$

Considering first order conditions (where we ignore the positivity constraint as explained above), we have

$$\frac{\partial \mathcal{E}(\mathrm{P},\mathrm{f},\mathrm{g})}{\partial \mathrm{P}_{i,j}} = \mathrm{C}_{i,j} + \varepsilon \log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{a}_i \mathrm{b}_j}\right) - \mathrm{f}_i - \mathrm{g}_j = 0.$$

which results, for an optimal P coupling to the regularized problem, in the expression $\mathrm{P}_{i,j} = \mathrm{a}_i \mathrm{b}_j e^{\frac{\mathrm{f}_i + \mathrm{g}_j - \mathrm{C}_{i,j}}{\varepsilon}}$ which can be rewritten in the form provided in the proposition using non-negative vectors $\mathrm{u} \stackrel{\text{def.}}{=} (\mathrm{a}_i e^{\mathrm{f}_i/\varepsilon})_i$ and $\mathrm{v} \stackrel{\text{def.}}{=} (\mathrm{b}_j e^{\mathrm{g}_j/\varepsilon})_j$. □

The factorization of the optimal solution exhibited in Equation (30) can be conveniently rewritten in matrix form as $\mathrm{P} = \mathrm{diag}(\mathrm{u})\mathrm{K}\,\mathrm{diag}(\mathrm{v})$. u, v must therefore satisfy the following non-linear equations which correspond to the mass conservation constraints inherent to $\mathrm{U}(\mathrm{a}, \mathrm{b})$,

$$\mathrm{diag}(\mathrm{u})\mathrm{K}\,\mathrm{diag}(\mathrm{v})\mathbb{1}_m = \mathrm{a}, \quad \text{and} \quad \mathrm{diag}(\mathrm{v})\mathrm{K}^\top \mathrm{diag}(\mathrm{u})\mathbb{1}_n = \mathrm{b}, \tag{31}$$

These two equations can be further simplified, since $\mathrm{diag}(\mathrm{v})\mathbb{1}_m$ is v, and the multiplication of $\mathrm{diag}(\mathrm{u})$ times Kv is

$$\mathrm{u} \odot (\mathrm{Kv}) = \mathrm{a} \quad \text{and} \quad \mathrm{v} \odot (\mathrm{K}^\top \mathrm{u}) = \mathrm{b} \tag{32}$$

where $\odot$ corresponds to entry-wise multiplication of vectors. That problem is known in the numerical analysis community as the matrix scaling problem (see [20] and references therein). An intuitive way to try to solve these equations is to solve them iteratively, by modifying first u so that it satisfies the left-hand side of Equation (32) and then v to satisfy its right-hand side. These two updates define Sinkhorn's algorithm

$$\mathrm{u}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathrm{a}}{\mathrm{Kv}^{(\ell)}} \quad \text{and} \quad \mathrm{v}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathrm{b}}{\mathrm{K}^\top \mathrm{u}^{(\ell+1)}}, \tag{33}$$

initialized with an arbitrary positive vector, for instance $\mathrm{v}^{(0)} = \mathbb{1}_m$. The division operator used above between two vectors is to be understood entry-wise. Note that a different initialization will likely lead to a different solution for u, v, since u, v are only defined up to a multiplicative constant (if u, v satisfy (31) then so do $\lambda \mathrm{u}, \mathrm{v}/\lambda$ for any $\lambda > 0$). It turns out however that these iterations converge, as we detail next.

A chief advantage, beside its simplicity, of Sinkhorn's algorithm is that the only computationnaly expensive step are matrix-vector multiplication by the Gibbs kernel, so that its complexity scales likes Knm where K is the number of Sinkhorn iteration, which can be kept polynomially in $1/\varepsilon$ if one is interested in reaching an accuracy ε on the (unregularized) transportation cost. Note however that in many situation, one is not interested in reaching high accuracy, because targeted application success is often only remotely connected to the ability to solve an optimal transport problem (but rather only being able to compare in a geometrically faithful way distribution), so that K is usually quite small. This should be contrasted with interior point methods, which also operate by introducing a barrier function of the form $-\sum_i \log(\mathrm{P}_{i,j})$. These algorithm have typically a complexity of the order $O(n^6 \log(|\varepsilon|))$.

The second crucial aspect of Sinkhorn is that matrix-vector multiplication streams extremely well on GPU. Even better, if one is interested in computing many OT problem with a fixed cost matrix C, one can replace many matrix-vector multiplication by matrix-matrix multiplication, so that the computation gain is enormous.

4.3 Reformulation using relative entropy

A convenient tool to re-formulate and "normalize" this discrete entropy (which is crucial to formulate a continuous version of the problem bellow) is to consider the relative entropy, also called Kullback-Leibler divergence, which is defined as

$$\mathrm{KL}(\mathrm{P}|\mathrm{Q}) \stackrel{\text{def.}}{=} \sum_{i,j} \mathrm{P}_{i,j} \log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{Q}_{i,j}}\right) - \mathrm{P}_{i,j} + \mathrm{Q}_{i,j}. \tag{34}$$

with the convention $0\log(0) = 0$ and $\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = +\infty$ if there exists some (i,j) such that $\mathrm{Q}_{i,j} = 0$ but $\mathrm{P}_{i,j} \neq 0$. The Shannon-Boltzmann neg-entropy is obtained when considering $\mathrm{Q} = \mathbb{1}_{n\times m}$, i.e.

$$\mathrm{H}(\mathrm{P}) = -\mathrm{KL}(\mathrm{P}|\mathbb{1}_{n\times m}).$$

KL is a particular instance (and actually the unique case) of both a φ-divergence and a Bregman divergence. This unique property is at the heart of the fact that this regularization leads to elegant algorithms and a tractable mathematical analysis. One thus has $\mathrm{KL}(\mathrm{P}|\mathrm{Q}) \geqslant 0$ and $\mathrm{KL}(\mathrm{P}|\mathrm{Q}) = 0$ if and only if $\mathrm{P} = \mathrm{Q}$.

For instance, one can use as reference measure the tensor product $\mathrm{a}\otimes\mathrm{b} = (\mathrm{a}_i\mathrm{b}_j)_{i,j}$ and consider

$$\min_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \langle \mathrm{P},\, \mathrm{C}\rangle - \varepsilon\,\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}). \tag{35}$$

But in fact, the choice of normalization (i.e. reference measure), has no importance for the selection of the optimal P since it only affects the objective by a constant, as shown in the following proposition. In particular, (35) and (26) have the same unique solution.

Proposition 14. *For* $\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})$*, one has*

$$\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \mathrm{KL}(\mathrm{P}|\mathrm{a}'\otimes\mathrm{b}') - \mathrm{KL}(\mathrm{a}|\mathrm{a}') - \mathrm{KL}(\mathrm{b}|\mathrm{b}').$$

Proof. This follows from

$$\sum_{i,j}\mathrm{P}_{i,j}\log\frac{\mathrm{P}_{i,j}}{\mathrm{a}_i\mathrm{b}_j} = \sum_{i,j}\mathrm{P}_{i,j}\log\frac{\mathrm{P}_{i,j}}{\mathrm{a}'_i\mathrm{b}'_j} + \sum_{i,j}\mathrm{P}_{i,j}\left(\log\frac{\mathrm{a}'_i}{\mathrm{a}_i} + \log\frac{\mathrm{b}'_j}{\mathrm{b}_j}\right).$$

□

The choice of using the reference measure $\mathrm{a}\otimes\mathrm{b}$ is however important to deal with situation where the support of a and b can change (so that some coordinate of a or b might vanish), and more importantly in the following section which deals with possibly continuous distributions.

4.4 General Formulation

One can consider arbitrary measures by replacing the discrete entropy by the relative entropy with respect to the product measure $\mathrm{d}\alpha\otimes\mathrm{d}\beta(x,y) \overset{\text{def.}}{=} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$, and propose a regularized counterpart to (21) using

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{X\times Y} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) \tag{36}$$

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence (34)

$$\mathrm{KL}(\pi|\xi) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\Big)\mathrm{d}\pi(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} (\mathrm{d}\xi(x,y) - \mathrm{d}\pi(x,y)), \tag{37}$$

and by convention $\mathrm{KL}(\pi|\xi) = +\infty$ if π does not have a density $\frac{\mathrm{d}\pi}{\mathrm{d}\xi}$ with respect to ξ. It is important to realize that the reference measure $\alpha\otimes\beta$ chosen in (36) to define the entropic regularizing term $\mathrm{KL}(\cdot|\alpha\otimes\beta)$ plays no specific role (because Proposition 14 still applies in this general setting), only its support matters. This problem is often referred to as the "static Schrödinger problem", since π is intended to model the most likely coupling between particules of gaz which can be only observed at two different times (it is the so-called lazy gaz model). The parameter ε controls the temperature of the gaz, and particules do not move in deterministic straight line as in optimal transport for the Euclidean cost, but rather according to a stochastic Brownian bridge.

Remark 6 (Probabilistic interpretation). If $(X,Y) \sim \pi$ have marginals $X \sim \alpha$ and $Y \sim \beta$, then $\mathrm{KL}(\pi|\alpha\otimes\beta) = \mathcal{I}(X,Y)$ is the mutual information of the couple, which is 0 if and only if X and Y are independent. The entropic problem (36) is thus equivalent to

$$\min_{(X,Y),X\sim\alpha,Y\sim\beta} \mathbb{E}(c(X,Y)) + \varepsilon\mathcal{I}(X,Y).$$

Using a large ε thus enforces the optimal coupling to describe independent variables, while, according to Brenier's theorem, small ε rather imposes a deterministic dependency between the couple according to a Monge map.

4.5 Convergence of Sinkhorn

This section provides a first overview of convergence proof for Sinkhorn. For the sake of simplicity, this section is written for discrete measures, but the analysis caries over to general measure.

Alternating KL projections. The following proposition explains that the minimized objective is equal to a KL distance toward the Gibbs distribution.

Proposition 15. *One has*

$$\langle \mathrm{P}, \mathrm{C}\rangle + \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{K}) + cst,$$

Proof. The objective is indeed equal to

$$\varepsilon\sum_{i,j}\mathrm{P}_{i,j}\frac{\mathrm{C}_{i,j}}{\varepsilon} + \log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{a}_i\mathrm{b}_j}\right)\mathrm{P}_{i,j} = \varepsilon\sum_{i,j}\mathrm{P}_{i,j}\log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{a}_i\mathrm{b}_j e^{-\mathrm{C}_{i,j}/\varepsilon}}\right) = \varepsilon\sum_{i,j}\mathrm{P}_{i,j}\log\left(\frac{\mathrm{P}_{i,j}}{\mathrm{K}_{i,j}}\right) + \mathrm{cst}.$$

□

This shows that the unique solution P_ε of (26) is a projection onto $\mathrm{U}(\mathrm{a},\mathrm{b})$ of the Gibbs kernel K

$$\mathrm{P}_\varepsilon = \mathrm{Proj}^{\mathrm{KL}}_{\mathrm{U}(\mathrm{a},\mathrm{b})}(\mathrm{K}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathrm{P}\in\mathrm{U}(\mathrm{a},\mathrm{b})} \mathrm{KL}(\mathrm{P}|\mathrm{K}). \tag{38}$$

Denoting

$$\mathcal{C}^1_{\mathrm{a}} \overset{\text{def.}}{=} \{\mathrm{P}\ ;\ \mathrm{P}\mathbb{1}_m = \mathrm{a}\} \quad\text{and}\quad \mathcal{C}^2_{\mathrm{b}} \overset{\text{def.}}{=} \left\{\mathrm{P}\ ;\ \mathrm{P}^\top\mathbb{1}_m = \mathrm{b}\right\}$$

the rows and columns constraints, one has $\mathrm{U}(\mathrm{a},\mathrm{b}) = \mathcal{C}^1_{\mathrm{a}}\cap\mathcal{C}^2_{\mathrm{b}}$. One can use Bregman iterative projections [5]

$$\mathrm{P}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^1_{\mathrm{a}}}(\mathrm{P}^{(\ell)}) \quad\text{and}\quad \mathrm{P}^{(\ell+2)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^2_{\mathrm{b}}}(\mathrm{P}^{(\ell+1)}). \tag{39}$$

Since the sets $\mathcal{C}^1_{\mathrm{a}}$ and $\mathcal{C}^2_{\mathrm{b}}$ are affine, these iterations are known to converge to the solution of (38), see [5].

The two projector are simple to compute since they corresponds to scaling respectively the rows and the columns, as explained in this proposition.

Proposition 16. *One has*

$$\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^1_{\mathrm{a}}}(\mathrm{P}) = \mathrm{diag}\left(\frac{\mathrm{a}}{\mathrm{P}\mathbb{1}_m}\right)\mathrm{P} \quad \textit{and} \quad \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^2_{\mathrm{b}}}(\mathrm{P}) = \mathrm{P}\,\mathrm{diag}\left(\frac{\mathrm{b}}{\mathrm{P}^\top\mathbb{1}_n}\right).$$

Proof. One considers the problem along each row or column vector to impose a fixed sum $s\in\mathbb{R}_+$

$$\min_p \{\mathrm{KL}(p|q)\ ;\ \langle p,\mathbb{1}\rangle = s\}.$$

The Lagrange multiplier for this problem read

$$\log(p/q) + \lambda\mathbb{1} = 0 \quad\Longrightarrow\quad p = uq \quad\text{where}\quad u = e^{-\lambda} > 0.$$

One has $\langle p,\mathbb{1}\rangle = s$ which is equivalent to $\langle uq,\mathbb{1}\rangle = s$ i.e. $u = s/\sum_i q_i$ and hence the desired scaling resulting formula $p = sp/\sum_i q_i$. □

These iterate are equivalent to Sinkhorn iterations (33) since defining

$$\mathrm{P}^{(2\ell)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathrm{u}^{(\ell)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)}),$$

one has

$$\begin{aligned} \mathrm{P}^{(2\ell+1)} &\stackrel{\text{def.}}{=} \mathrm{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)}) \\ \text{and} \quad \mathrm{P}^{(2\ell+2)} &\stackrel{\text{def.}}{=} \mathrm{diag}(\mathrm{u}^{(\ell+1)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell+1)}) \end{aligned}$$

In practice however one should prefer using (33) which only requires manipulating scaling vectors and multiplication against a Gibbs kernel, which can often be accelerated.

Such a convergence analysis using Bregman projection is however of limited interested because it only works in finite dimension. For instance, the linear convergence speed one can obtain with these analyses (because the objective is strongly convex) will degrade with the dimension (and of course also with ε). It is also possible to decay ε during the iterates to improve the speed and rely on multiscale strategies in low dimension.

Convergence for the Hilbert metric As initially explained by [11], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert projective metric on $\mathbb{R}^n_{+,*}$ (positive vectors), defined as

$$\forall\,(\mathrm{u},\mathrm{u}') \in (\mathbb{R}^n_{+,*})^2, \quad d_{\mathcal{H}}(\mathrm{u},\mathrm{u}') \stackrel{\text{def.}}{=} \|\log(\mathrm{u}) - \log(\mathrm{v})\|_V$$

where the variation semi-norm is

$$\|z\|_V = \max(z) - \min(z).$$

One can show that $d_{\mathcal{H}}$ is a distance on the projective cone $\mathbb{R}^n_{+,*}/\sim$, where $\mathrm{u} \sim \mathrm{u}'$ means that $\exists s > 0, \mathrm{u} = s\mathrm{u}'$ (the vector are equal up to rescaling, hence the naming "projective"), and that $(\mathbb{R}^n_{+,*}/\sim, d_{\mathcal{H}})$ is then a complete metric space. It was introduced independently by [4] and [22] to provide a quantitative proof of Perron-Frobenius theorem (convergence of iterations of positive matrices). Sinkhorn should be thought as a non-linear generalization of Perron-Frobenius.

Theorem 3. *Let* $\mathrm{K} \in \mathbb{R}^{n\times m}_{+,*}$*, then for* $(\mathrm{v},\mathrm{v}') \in (\mathbb{R}^m_{+,*})^2$

$$d_{\mathcal{H}}(\mathrm{Kv},\mathrm{Kv}') \leqslant \lambda(\mathrm{K}) d_{\mathcal{H}}(\mathrm{v},\mathrm{v}') \text{ where } \begin{cases} \lambda(\mathrm{K}) \stackrel{\text{def.}}{=} \frac{\sqrt{\eta(\mathrm{K})}-1}{\sqrt{\eta(\mathrm{K})}+1} < 1 \\ \eta(\mathrm{K}) \stackrel{\text{def.}}{=} \max\limits_{i,j,k,\ell} \frac{\mathrm{K}_{i,k}\mathrm{K}_{j,\ell}}{\mathrm{K}_{j,k}\mathrm{K}_{i,\ell}}. \end{cases}$$

The following theorem, proved by [11], makes use of this Theorem 3 to show the linear convergence of Sinkhorn's iterations.

Theorem 4. *One has* $(\mathrm{u}^{(\ell)},\mathrm{v}^{(\ell)}) \to (\mathrm{u}^\star,\mathrm{v}^\star)$ *and*

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell)},\mathrm{u}^\star) = O(\lambda(\mathrm{K})^{2\ell}), \quad d_{\mathcal{H}}(\mathrm{v}^{(\ell)},\mathrm{v}^\star) = O(\lambda(\mathrm{K})^{2\ell}). \tag{40}$$

One also has

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell)},\mathrm{u}^\star) \leqslant \frac{d_{\mathcal{H}}(\mathrm{P}^{(\ell)}\mathbb{1}_m,\mathrm{a})}{1-\lambda(\mathrm{K})} \quad \text{and} \quad d_{\mathcal{H}}(\mathrm{v}^{(\ell)},\mathrm{v}^\star) \leqslant \frac{d_{\mathcal{H}}(\mathrm{P}^{(\ell),\top}\mathbb{1}_n,\mathrm{b})}{1-\lambda(\mathrm{K})}, \tag{41}$$

where we denoted $\mathrm{P}^{(\ell)} \stackrel{\text{def.}}{=} \mathrm{diag}(\mathrm{u}^{(\ell)})\mathrm{K}\,\mathrm{diag}(\mathrm{v}^{(\ell)})$*. Lastly, one has*

$$\|\log(\mathrm{P}^{(\ell)}) - \log(\mathrm{P}^\star)\|_\infty \leqslant d_{\mathcal{H}}(\mathrm{u}^{(\ell)},\mathrm{u}^\star) + d_{\mathcal{H}}(\mathrm{v}^{(\ell)},\mathrm{v}^\star) \tag{42}$$

where $\mathrm{P}^\star$ *is the unique solution of* (26)*.*

Proof. One notice that for any $(\mathrm{v},\mathrm{v}')\in(\mathbb{R}^m_{+,*})^2$, one has

$$d_{\mathcal{H}}(\mathrm{v},\mathrm{v}')=d_{\mathcal{H}}(\mathrm{v}/\mathrm{v}',\mathbb{1}_m)=d_{\mathcal{H}}(\mathbb{1}_m/\mathrm{v},\mathbb{1}_m/\mathrm{v}'),$$

since indeed $d_{\mathcal{H}}(\mathrm{a}/\mathrm{v},\mathrm{a}/\mathrm{v}')=d_{\mathcal{H}}(\mathrm{v},\mathrm{v}')$. This shows that

$$d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)},\mathrm{u}^\star)=d_{\mathcal{H}}\left(\frac{\mathrm{a}}{\mathrm{K}\mathrm{v}^{(\ell)}},\frac{\mathrm{a}}{\mathrm{K}\mathrm{v}^\star}\right)=d_{\mathcal{H}}(\mathrm{K}\mathrm{v}^{(\ell)},\mathrm{K}\mathrm{v}^\star)\leqslant\lambda(\mathrm{K})d_{\mathcal{H}}(\mathrm{v}^{(\ell)},\mathrm{v}^\star).$$

where we used Theorem 3. This shows (40). One also has, using the triangular inequality

$$\begin{aligned}d_{\mathcal{H}}(\mathrm{u}^{(\ell)},\mathrm{u}^\star)&\leqslant d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)},\mathrm{u}^{(\ell)})+d_{\mathcal{H}}(\mathrm{u}^{(\ell+1)},\mathrm{u}^\star)\leqslant d_{\mathcal{H}}\left(\frac{\mathrm{a}}{\mathrm{K}\mathrm{v}^{(\ell)}},\mathrm{u}^{(\ell)}\right)+\lambda(\mathrm{K})d_{\mathcal{H}}(\mathrm{u}^{(\ell)},\mathrm{u}^\star)\\&=d_{\mathcal{H}}\left(\mathrm{a},\mathrm{u}^{(\ell)}\odot(\mathrm{K}\mathrm{v}^{(\ell)})\right)+\lambda(\mathrm{K})d_{\mathcal{H}}(\mathrm{u}^{(\ell)},\mathrm{u}^\star),\end{aligned}$$

which gives the first part of (41) since $\mathrm{u}^{(\ell)}\odot(\mathrm{K}\mathrm{v}^{(\ell)})=\mathrm{P}^{(\ell)}\mathbb{1}_m$ (the second one being similar). The proof of (42) follows from [11, Lemma 3] □

The bound (41) shows that some error measures on the marginal constraints violation, for instance $\|\mathrm{P}^{(\ell)}\mathbb{1}_m-\mathrm{a}\|_1$ and $\|{\mathrm{P}^{(\ell)}}^\top\mathbb{1}_n-\mathrm{b}\|_1$, are useful stopping criteria to monitor the convergence. This theorem shows that Sinkhorn algorithm converges linearly, but the rates becomes exponentially bad as $\varepsilon\to0$, since it scales like $e^{-1/\varepsilon}$. In practice, one eventually observes a linear rate after enough iteration, because the local linear rate is much better, usually of the order $1-\varepsilon$.

5 Dual Problem

5.1 Discrete dual

The Kantorovich problem (18) is a linear program, so that one can equivalently compute its value by solving a dual linear program.

Proposition 17. *One has*

$$\mathrm{L}_{\mathrm{C}}(\mathrm{a},\mathrm{b})=\max_{(\mathrm{f},\mathrm{g})\in\mathrm{R}(\mathrm{a},\mathrm{b})}\langle\mathrm{f},\,\mathrm{a}\rangle+\langle\mathrm{g},\,\mathrm{b}\rangle\tag{43}$$

where the set of admissible potentials is

$$\mathrm{R}(\mathrm{a},\mathrm{b})\overset{\text{def.}}{=}\{(\mathrm{f},\mathrm{g})\in\mathbb{R}^n\times\mathbb{R}^m\;;\;\forall\,(i,j)\in[\![n]\!]\times[\![m]\!],\mathrm{f}\oplus\mathrm{g}\leqslant\mathrm{C}\}\tag{44}$$

Proof. For the sake of completeness, let us derive this dual problem with the use of Lagrangian duality. The Lagangian associate to (18) reads

$$\min_{\mathrm{P}\geqslant0}\max_{(\mathrm{f},\mathrm{g})\in\mathbb{R}^n\times\mathbb{R}^m}\langle\mathrm{C},\,\mathrm{P}\rangle+\langle\mathrm{a}-\mathrm{P}\mathbb{1}_m,\,\mathrm{f}\rangle+\langle\mathrm{b}-\mathrm{P}^\top\mathbb{1}_n,\,\mathrm{g}\rangle.\tag{45}$$

For linear program, if the primal set of constraint is non-empty, one can always exchange the min and the max and get the same value of the linear program, and one thus consider

$$\max_{(\mathrm{f},\mathrm{g})\in\mathbb{R}^n\times\mathbb{R}^m}\langle\mathrm{a},\,\mathrm{f}\rangle+\langle\mathrm{b},\,\mathrm{g}\rangle+\min_{\mathrm{P}\geqslant0}\langle\mathrm{C}-\mathrm{f}\mathbb{1}_m^\top-\mathbb{1}_n g^\top,\,\mathrm{P}\rangle.$$

We conclude by remarking that

$$\min_{\mathrm{P}\geqslant0}\langle\mathrm{Q},\,\mathrm{P}\rangle=\begin{cases}0&\text{if}\quad\mathrm{Q}\geqslant0\\-\infty&\text{otherwise}\end{cases}$$

so that the constraint reads $\mathrm{C}-\mathrm{f}\mathbb{1}_m^\top-\mathbb{1}_n\mathrm{g}^\top=\mathrm{C}-\mathrm{f}\oplus\mathrm{g}\geqslant0$. □

The primal-dual optimality relation for the Lagrangian (45) allows to locate the support of the optimal transport plan

$$\mathrm{Supp}(\mathrm{P})\subset\{(i,j)\in[\![n]\!]\times[\![m]\!]\;;\;\mathrm{f}_i+\mathrm{g}_j=\mathrm{C}_{i,j}\}\,.\tag{46}$$

The formulation (43) shows that $(\mathrm{a},\mathrm{b})\mapsto\mathrm{L}_{\mathrm{C}}(\mathrm{a},\mathrm{b})$ is a convex function (as a supremum of linear functions). From the primal problem (18), one also sees that $\mathrm{C}\mapsto\mathrm{L}_{\mathrm{C}}(\mathrm{a},\mathrm{b})$ is concave.

5.2 General formulation

To extend this primal-dual construction to arbitrary measures, it is important to realize that measures are naturally paired in duality with continuous functions, using the pairing $\langle f, \alpha \rangle \overset{\text{def.}}{=} \int f \mathrm{d}\alpha$.

Proposition 18. *One has*

$$\mathcal{L}_c(\alpha, \beta) = \max_{(f,g) \in \mathcal{R}(c)} \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y) \mathrm{d}\beta(y), \tag{47}$$

where the set of admissible dual potentials is

$$\mathcal{R}(c) \overset{\text{def.}}{=} \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \; ; \; \forall (x,y), f(x) + g(y) \leqslant c(x,y)\}. \tag{48}$$

Here, (f,g) is a pair of continuous functions, and are often called "Kantorovich potentials".

The discrete case (43) corresponds to the dual vectors being samples of the continuous potentials, *i.e.* $(\mathrm{f}_i, \mathrm{g}_j) = (f(x_i), g(y_j))$. The primal-dual optimality conditions allow to track the support of optimal plan, and (46) is generalized as

$$\mathrm{Supp}(\pi) \subset \{(x,y) \in \mathcal{X} \times \mathcal{Y} \; ; \; f(x) + g(y) = c(x,y)\}. \tag{49}$$

Note that in contrast to the primal problem (21), showing the existence of solutions to (47) is non-trivial, because the constraint set $\mathcal{R}(c)$ is not compact and the function to minimize non-coercive. Using the machinery of c-transform detailed in Section 5.3, one can however show that optimal (f,g) are necessarily Lipschitz regular, which enable to replace the constraint by a compact one.

5.3 c-transforms

Definition. Keeping a dual potential g fixed, one can try to minimize in closed form the dual problem (47), which leads to consider

$$\sup_{g \in \mathcal{C}(\mathcal{Y})} \left\{ \int g \mathrm{d}\beta \; ; \; \forall\, (x,y), g(y) \leqslant c(x,y) - f(x) \right\}.$$

The constraint can be replaced by

$$\forall\, y \in \mathcal{Y}, \quad g(y) \leqslant f^c(y)$$

where we define the c-transform as

$$\forall\, y \in \mathcal{Y}, \quad f^c(y) \overset{\text{def.}}{=} \inf_{x \in \mathcal{X}} c(x,y) - f(x). \tag{50}$$

Since β is positive, the maximization of $\int g \mathrm{d}\beta$ is thus achieved at those functions such that $g = f^c$ on the support of β, which means β-almost everywhere.

Similarly, we defined the $\bar{c}$-transform, which a transform for the symetrized cost $\bar{c}(y,x) = c(x,y)$, i.e.

$$\forall\, x \in \mathcal{X}, \quad g^{\bar{c}}(x) \overset{\text{def.}}{=} \inf_{y \in \mathcal{Y}} c(x,y) - g(y),$$

and one checks that any function f such that $f = g^{\bar{c}}$ α-almost everywhere is solution to the dual problem for a fixed g.

The map $(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mapsto (g^{\bar{c}}, f^c) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})$ replaces dual potentials by "better" ones (improving the dual objective $\mathcal{E}$). Functions that can be written in the form f^c and $g^{\bar{c}}$ are called c-concave and $\bar{c}$-concave functions.

Note that these partial minimizations define maximizers on the support of respectively α and β, while the definitions (50) actually define functions on the whole spaces $\mathcal{X}$ and $\mathcal{Y}$. This is thus a way to extend in a canonical way solutions of (47) on the whole spaces.

Furthermore, if c is Lipschitz, then f^c and g^c are also Lipschitz functions, as we now show. This property is crucial to show existence of solution to the dual problem. Indeed, since one can impose this Lipschitz on the dual problems, the constraint set is compact via Ascoli theorem.

Proposition 19. *If c is L-Lipschitz with respect to the second variable, then f^c is L-Lipschitz.*

Proof. We apply to $F_x = c(x,\cdot) - f(x)$ the fact that if all the F_x are L-Lipschitz, then the Lipschitz constant of $F = \min_x F_x$ is L. Indeed, using the fact that $|\inf(A) - \inf(B)| \leqslant \sup |A - B|$ for two function A and B, then

$$|F(y) - F(y')| = |\inf_x(F_x(y)) - \inf_x(F_x(y'))| \leqslant \sup_x |F_x(y) - F_x(y')| \leqslant \sup_x Ld(y,y') = Ld(y,y').$$

□

Euclidean case. The special case $c(x,y) = -\langle x,\, y\rangle$ in $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ is of utmost importance because it allows one to study the W_2 problem, since for any $\pi \in \mathcal{U}(\alpha,\beta)$

$$\int \|x-y\|^2 \mathrm{d}\pi(x,y) = \mathrm{cst} - 2\int \langle x,\, y\rangle \mathrm{d}\pi(x,y) \quad \text{where} \quad \mathrm{cst} = \int \|x\|^2 \mathrm{d}\alpha(x) + \int \|y\|^2 \mathrm{d}\beta(y).$$

For this special choice of cost, one has $f^c = -(-f)^*$ where h^* is the Fenchel-Legendre transform

$$h^*(y) \stackrel{\text{def.}}{=} \sup_x \langle x,\, y\rangle - h(y).$$

One has that h^* is always convex, so that f^c is always concave. For a general cost, one thus denotes functions of the form f^c as being c-concave.

Remark 7 (Proof of Brenier's theorem). In the case $c(x,y) = \|x-y\|^2$, using instead $c(x,y) = -\langle x,\, y\rangle$, the primal-dual relation ship, together with the fact that one can replace (f,g) by $(f^{cc}, f^{ccc} = f^c)$ one sees that

$$\mathrm{supp}(\pi) \subset \{(x,y)\ ;\ \varphi(x) + \varphi^*(y) = \langle x,\, y\rangle\}$$

where we have denoted $\varphi = -f^{cc}$ which is a convex function and $-g = \varphi^*$. One always have $\varphi(x) + \varphi^*(y) \leqslant \langle x,\, y\rangle$ from the definition of the Legendre transform, and the set of y such that this equality holds is precisely the sub-differential $\partial\varphi(x)$. In the special case where α has a density, since a convex function is differentiable Lebesgue-almost everywhere, it is also α-everywhere differentiable, so it is legit to use $T = \nabla\varphi$ as an optimal transport plan.

The failure of alternate optimization. A crucial property of the Legendre transform is that $f^{***} = f^*$, and that f^{**} is the convex enveloppe of f (the largest convex function below f). These properties carries over for the more general setting of c-transforms.

Proposition 20. *The following identities, in which the inequality sign between vectors should be understood elementwise, hold, denoting $f^{c\bar{c}} \stackrel{\text{def.}}{=} (f^c)^{\bar{c}}$:*

(i) $f \leqslant f' \Rightarrow f^c \geqslant f'^c$,

(ii) $f^{c\bar{c}} \geqslant f$,

(iii) $g^{\bar{c}c} \geqslant g$,

(iv) $f^{c\bar{c}c} = f^c$.

Proof. The first inequality (i) follows from the definition of c-transforms (because of the $-$ sign). To prove (ii), expanding the definition of $f^{c\bar{c}}$ we have

$$\left(f^{c\bar{c}}\right)(x) = \min_y c(x,y) - f^c(y) = \min_y c(x,y) - \min_{x'}(c(x',y) - f(x')).$$

Now, since $-\min_{x'} c(x',y) - f(x') \geqslant -(c(x,y) - f(x))$, we recover

$$(f^{c\bar{c}})(x) \geqslant \min_y c(x,y) - c(x,y) + f(x) = f(x).$$

The relation $g^{\bar{c}c} \geqslant g$ is obtained in the same way. Now, to prove (iv), we first apply (ii) and then (i) with $f' = f^{c\bar{c}}$ to have $f^c \geqslant f^{c\bar{c}c}$. Then we apply (iii) to $g = f^c$ to obtain $f^c \leqslant f^{c\bar{c}c}$. □

This invariance property shows that one can "improve" only once the dual potential this way. Indeed, starting from any pair (f, g), one obtains the following iterates by alternating maximization

$$(f, g) \mapsto (f, f^c) \mapsto (f^{cc}, f^c) \mapsto (f^{cc}, f^{ccc}) = (f^{cc}, f^c) \dots \tag{51}$$

so that one reaches a stationary point. This failure is the classical behavior of alternating maximization on a non-smooth problem, where the non-smooth part of the functional (here the constraint) mixes the two variables. The workaround is to introduce a smoothing, which is the classical method of augmented Lagrangian, and that we will develop here using entropic regularization, and corresponds to Sinkhorn's algorithm.

6 Semi-discrete and W_1

6.1 Semi-dual

From the dual problem (47), that we write as

$$\sup_{f,g \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \mathcal{E}(f, g)$$

where $\mathcal{E}$ takes into account the constraints one can "marginalize" it with respect to g by minimizing over it to obtain the following "semi-dual" problem

$$\sup_{f \in \mathcal{C}(\mathcal{X})} \tilde{\mathcal{E}}(f) \overset{\text{def.}}{=} \mathcal{E}(f, f^c) = \sup_g \mathcal{E}(f, g) = \int_{\mathcal{X}} f \mathrm{d}\alpha + \int_{\mathcal{Y}} f^c \mathrm{d}\beta. \tag{52}$$

Note that marginalizing a concave problem retains concavity so that $\tilde{\mathcal{E}}$ is still concave. The major advantage of this new "semi-dual" problem is that it is an unconstraint problem, which allows the use of simpler optimization algorithms, as we will now see.

6.2 Semi-discrete

A case of particular interest is when $\beta = \sum_j \mathrm{b}_j \delta_{y_j}$ is discrete (of course the same construction applies if α is discrete by exchanging the role of α, β). One can adapt the definition of the $\bar{c}$ transform (50) to this setting by restricting the minimization to the support $(y_j)_j$ of β,

$$\forall\, \mathrm{g} \in \mathbb{R}^m,\ \forall\, x \in \mathcal{X}, \quad \mathrm{g}^{\bar{c}}(x) \overset{\text{def.}}{=} \min_{j \in [\![m]\!]} c(x, y_j) - \mathrm{g}_j. \tag{53}$$

This transform maps a vector g to a continuous function $\mathrm{g}^{\bar{c}} \in \mathcal{C}(\mathcal{X})$. Note that this definition coincides with (50) when imposing that the space $\mathcal{X}$ is equal to the support of β.

Crucially, using the discrete $\bar{c}$-transform, when β is a discrete measure, yields a finite-dimensional optimization,

$$\mathcal{L}_c(\alpha, \beta) = \max_{\mathrm{g} \in \mathbb{R}^m} \mathcal{E}(\mathrm{g}) \overset{\text{def.}}{=} \int_{\mathcal{X}} \mathrm{g}^{\bar{c}}(x) \mathrm{d}\alpha(x) + \sum_j \mathrm{g}_j \mathrm{b}_j. \tag{54}$$

The Laguerre cells associated to the dual weights g

$$\mathbb{L}_j(\mathrm{g}) \overset{\text{def.}}{=} \{x \in \mathcal{X}\ ;\ \forall\, j' \neq j, c(x, y_j) - \mathrm{g}_j \leqslant c(x, y_{j'}) - \mathrm{g}_{j'}\}$$

induce a disjoint decomposition of $\mathcal{X} = \bigcup_j \mathbb{L}_j(\mathrm{g})$. When g is constant, the Laguerre cells decomposition corresponds to the Voronoi diagram partition of the space.

This allows one to conveniently rewrite the minimized energy as

$$\mathcal{E}(\mathrm{g}) = \sum_{j=1}^m \int_{\mathbb{L}_j(\mathrm{g})} (c(x, y_j) - \mathrm{g}_j)\, \mathrm{d}\alpha(x) + \langle \mathrm{g}, \mathrm{b} \rangle. \tag{55}$$

The following proposition provides a formula for the gradient of this convex function.

Proposition 21. *If α has a density with respect to Lebesgue measure and if c is smooth away from the diagonal, then $\mathcal{E}$ is differentiable and*

$$\forall\, j \in [\![m]\!], \quad \nabla\mathcal{E}(\mathrm{g})_j = \mathrm{b}_j - \int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha.$$

Proof. One has

$$\mathcal{E}(\mathrm{g}+\varepsilon\delta_j) - \mathcal{E}(\mathrm{g}) - \varepsilon\left(\mathrm{b}_j - \int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha\right) = \sum_k \int_{\mathbb{L}_k(\mathrm{g}+\varepsilon\delta_j)} c(x,x_k)\mathrm{d}\alpha(x) - \int_{\mathbb{L}_k(\mathrm{g})} c(x,x_k)\mathrm{d}\alpha(x).$$

Most of the terms in the right hand side vanish (because most the Laguerre cells associated to $\mathrm{g}+\varepsilon\delta_j$ are equal to those of g) and the only terms remaining correspond to neighboring cells (j,k) such that $\mathbb{L}_j(\mathrm{g})\cap\mathbb{L}_k(\mathrm{g}) \neq \emptyset$ (for the cost $\|x-y\|^2$ and $\mathrm{g}=0$ this forms the Delaunay triangulation). On these pairs, the right integral differs on a volume of the order of ε (since α has a density) and the function being integrated only varies on the order of ε (since the cost is smooth). So the right hand side is of the order of ε^2. □

The first order optimality condition shows that in order to solve the dual semi discrete problem, one needs to select the weights g in order to drive the Laguerre cell in a configuration such that $\int_{\mathbb{L}_j(\mathrm{g})} \mathrm{d}\alpha = \mathrm{b}_j$, i.e. each cell should capture the correct amount of mass. In this case, the optimal transport T such that $T_\sharp\alpha = \beta$ (which exists and is unique according to Brenier's theorem if α has a density) is piecewise constant and map $x \in \mathbb{L}_j(\mathrm{g})$ to y_j.

In the special case $c(x,y) = \|x-y\|^2$, the decomposition in Laguerre cells is also known as a "power diagram". In this case, the cells are polyhedral and can be computed efficiently using computational geometry algorithms; see [1]. The most widely used algorithm relies on the fact that the power diagram of points in $\mathbb{R}^d$ is equal to the projection on $\mathbb{R}^d$ of the convex hull of the set of points $((y_j, \|y_j\|^2 - \mathrm{g}_j))_{j=1}^m \subset \mathbb{R}^{d+1}$. There are numerous algorithms to compute convex hulls; for instance, that of [9] in two and three dimensions has complexity $O(m\log(Q))$, where Q is the number of vertices of the convex hull.

Stochastic optimization. The semidiscrete formulation (55) is also appealing because the energies to be minimized are written as an expectation with respect to the probability distribution α,

$$\mathcal{E}(\mathrm{g}) = \int_{\mathcal{X}} E(\mathrm{g},x)\mathrm{d}\alpha(x) = \mathbb{E}_X(E(\mathrm{g},X)) \quad \text{where} \quad E(\mathrm{g},x) \overset{\text{def.}}{=} \mathrm{g}^{\bar{c}}(x) - \langle \mathrm{g},\, \mathrm{b}\rangle, \tag{56}$$

and X denotes a random vector distributed on $\mathcal{X}$ according to α. Note that the gradient of each of the involved functional reads

$$\nabla_{\mathrm{g}} E(x,\mathrm{g}) = (\mathbb{1}_{\mathbb{L}_j(\mathrm{g})}(x) - \mathrm{b}_j)_{j=1}^m \in \mathbb{R}^m$$

where $\mathbb{1}_{\mathbb{L}_j(\mathrm{g})}$ is the indicator function of the Laguerre cell. One can thus use stochastic optimization methods to perform the maximization, as proposed in [12]. This allows us to obtain provably convergent algorithms without the need to resort to an arbitrary discretization of α (either approximating α using sums of Diracs or using quadrature formula for the integrals). The measure α is used as a black box from which one can draw independent samples, which is a natural computational setup for many high-dimensional applications in statistics and machine learning.

Initializing $\mathrm{g}^{(0)} = \mathbf{0}_m$, the stochastic gradient descent algorithm (SGD; used here as a maximization method) draws at step ℓ a point $x_\ell \in \mathcal{X}$ according to distribution α (independently from all past and future samples $(x_\ell)_\ell$) to form the update

$$\mathrm{g}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{g}^{(\ell)} + \tau_\ell \nabla_{\mathrm{g}} E(\mathrm{g}^{(\ell)}, x_\ell). \tag{57}$$

The step size τ_ℓ should decay fast enough to zero in order to ensure that the "noise" created by using $\nabla_{\mathrm{g}} E(x_\ell,\mathrm{g})$ as a proxy for the true gradient $\nabla\mathcal{E}(\mathrm{g})$ is canceled in the limit. A typical choice of schedule is

$$\tau_\ell \overset{\text{def.}}{=} \frac{\tau_0}{1+\ell/\ell_0}, \tag{58}$$

where ℓ_0 indicates roughly the number of iterations serving as a warmup phase. One can prove the convergence result

$$\mathcal{E}(\mathrm{g}^\star) - \mathbb{E}(\mathcal{E}(\mathrm{g}^{(\ell)})) = O\left(\frac{1}{\sqrt{\ell}}\right),$$

where $\mathrm{g}^\star$ is a solution of (56) and where $\mathbb{E}$ indicates an expectation with respect to the i.i.d. sampling of $(x_\ell)_\ell$ performed at each iteration.

Optimal quantization. The optimal quantization problem of some measure α corresponds to the resolution of

$$\mathcal{Q}_m(\alpha) = \min_{Y=(y_j)_{j=1}^m, (\mathrm{b}_j)_{j=1}^m} W_p(\alpha, \sum_j \mathrm{b}_j \delta_{y_j}).$$

This problem is at the heart of the computation of efficient vector quantizer in information theory and compression, and is also the basic problem to solve for clustering in unsupervised learning. The asymptotic behavior of $\mathcal{Q}_m$ is of fundamental importance, and its precise behavior is in general unknown. For a measure with a density in Euclidean space, it scales like $O(1/n^{1/d})$, so that quantization generally suffers from the curse of dimensionality.

This optimal quantization problem is convex with respect to b, but is unfortunately non-convex with respect to $Y = (y_j)_j$. Its resolution is in general NP-hard. The only setting where this problem is simple is the 1-D case, in which case the optimal sampling is simply $y_j = \mathcal{C}_\alpha^{-1}(j/m)$.

Solving explicitly for the minimization over b in the formula (54) (exchanging the role of the min and the max) shows that necessarily, at optimality, one has $\mathrm{g} = 0$, so that the optimal transport maps the Voronoi cells $\mathbb{L}_j(\mathrm{g} = 0)$, which we denote $\mathbb{V}_j(Y)$ to highlight the dependency on the quantization points $Y = (y_j)_j$

$$\mathbb{V}_j(Y) = \{x \;;\; \forall\, j', c(x, y_{j'}) \leqslant c(x, y_j)\}.$$

This also shows that the quantization energy can be rewritten in a more intuitive way, which accounts for the average quantization error induced by replacing a point x by its nearest centroid

$$\mathcal{Q}_m(\alpha) = \min_Y \mathcal{F}(Y) \overset{\text{def.}}{=} \int_{\mathcal{X}} \min_{1 \leqslant j \leqslant m} c(x, y_j) \mathrm{d}\alpha(x).$$

At any local minimizer (at least if α has a density so that this function is differentiable) of this energy over Y, one sees that each y_j should be a centroid of its associated Voronoi region,

$$y_j \in \operatorname*{argmin}_y \int_{\mathbb{V}_j(Y)} c(x, y) \mathrm{d}\alpha(x).$$

For instance, when $c(x, y) = \|x - y\|^2$, one sees that any local minimizer should satisfy the fixed point equation

$$y_j = \frac{\int_{\mathbb{V}_j(Y)} x \mathrm{d}\alpha(x)}{\int_{\mathbb{V}_j(Y)} \mathrm{d}\alpha}.$$

The celebrated k-means algorithm, also known as Lloyd algorithm, iteratively apply this fixed point. It is not guaranteed to converge (it could in theory cycle) but in practice it always converge to a local minimum. A practical issue to obtain a good local minimizer is to seed a good initial configuration. The intuitive way to achieve this is to spread them as much as possible, and a well known algorithm to do so is the k-means++ methods, which achieve without even any iteration a quantization cost which is of the order of $\log(m)\mathcal{Q}_m(\alpha)$.

6.3 W_1

c-transform for W_1. Here we assume that d is a distance on $\mathcal{X} = \mathcal{Y}$, and we solve the OT problem with the ground cost $c(x, y) = d(x, y)$. The following proposition highlights key properties of the c-transform (50) in this setup. In the following, we denote the Lipschitz constant of a function $f \in \mathcal{C}(\mathcal{X})$ as

$$\mathrm{Lip}(f) \overset{\text{def.}}{=} \sup\left\{\frac{|f(x) - f(y)|}{d(x, y)} \;;\; (x, y) \in \mathcal{X}^2, x \neq y\right\}.$$

Proposition 22. *Suppose $\mathcal{X}=\mathcal{Y}$ and $c(x,y)=d(x,y)$. Then, there exists g such that $f=g^c$ if and only* $\mathrm{Lip}(f)\leqslant 1$. *Furthermore, if* $\mathrm{Lip}(f)\leqslant 1$, *then* $f^c=-f$.

Proof. First, suppose $f=g^c$ for some g. Then, for $x,y\in\mathcal{X}$,

$$\begin{aligned}|f(x)-f(y)| &= \left|\inf_{z\in\mathcal{X}}[d(x,z)-g(z)] \;-\; \inf_{z\in\mathcal{X}}[d(y,z)-g(z)]\right| \\ &\leqslant \sup_{z\in\mathcal{X}}|d(x,z)-d(y,z)|\leqslant d(x,y).\end{aligned}$$

The first equality follows from the definition of g^c, the next inequality from the identity $|\inf A-\inf B|\leqslant \sup|A-B|$, and the last from the reversed triangle inequality. This shows that $\mathrm{Lip}(f)\leqslant 1$.

If f is 1-Lipschitz, for all $x,y\in\mathcal{X}$, $f(y)-d(x,y)\leqslant f(x)\leqslant f(y)+d(x,y)$, which shows that

$$\begin{aligned}f^c(y)&=\inf_{x\in\mathcal{X}}[d(x,y)-f(x)]\geqslant \inf_{x\in\mathcal{X}}[d(x,y)-f(y)-d(x,y)]=-f(y),\\ f^c(y)&=\inf_{x\in\mathcal{X}}[d(x,y)-f(x)]\leqslant \inf_{x\in\mathcal{X}}[d(x,y)-f(y)+d(x,y)]=-f(y),\end{aligned}$$

because $\inf_x d(x,y)=0$ (for $x=y$) and thus $f^c=-f$.

Applying this property to $-f$ which is also 1-Lipschitz shows that $(-f)^c=f$ so that f is indeed c-concave (i.e. it is the c-transform of a function). □

Using the iterative c-transform scheme (51), one can replace the dual variable (f,g) by $(f^{cc},f^c)=(-f^c,f^c)$, or equivalently by any pair $(f,-f)$ where f is 1-Lipschitz. This leads to the following alternative expression for the $\mathcal{W}_1$ distance

$$\mathcal{W}_1(\alpha,\beta)=\max_f\left\{\int_{\mathcal{X}} f\mathrm{d}(\alpha-\beta)\;;\;\mathrm{Lip}(f)\leqslant 1\right\}. \tag{59}$$

This expression shows that $\mathcal{W}_1$ is actually a norm, i.e.$\mathcal{W}_1(\alpha,\beta)=\|\alpha-\beta\|_{\mathcal{W}_1}$, and that it is still valid for any measures (not necessary positive) as long as $\int_{\mathcal{X}}\alpha=\int_{\mathcal{X}}\beta$. This norm is often called the Kantorovich-Rubinstein norm [18].

For discrete measures of the form (2), writing $\alpha-\beta=\sum_k \mathrm{m}_k\delta_{z_k}$ with $z_k\in\mathcal{X}$ and $\sum_k \mathrm{m}_k=0$, the optimization (59) can be rewritten as

$$\mathcal{W}_1(\alpha,\beta)=\max_{(\mathrm{f}_k)_k}\left\{\sum_k \mathrm{f}_k\mathrm{m}_k\;;\;\forall\,(k,\ell),|\mathrm{f}_k-\mathrm{f}_\ell|\leqslant d(z_k,z_\ell),\right\} \tag{60}$$

which is a finite-dimensional convex program with quadratic-cone constraints. It can be solved using interior point methods or, as we detail next for a similar problem, using proximal methods.

When using $d(x,y)=|x-y|$ with $\mathcal{X}=\mathbb{R}$, we can reduce the number of constraints by ordering the z_k's via $z_1\leqslant z_2\leqslant\ldots$. In this case, we only have to solve

$$\mathcal{W}_1(\alpha,\beta)=\max_{(\mathrm{f}_k)_k}\left\{\sum_k \mathrm{f}_k\mathrm{m}_k\;;\;\forall\,k,|\mathrm{f}_{k+1}-\mathrm{f}_k|\leqslant z_{k+1}-z_k\right\},$$

which is a linear program. Note that furthermore, in this 1-D case, a closed form expression for $\mathcal{W}_1$ using cumulative functions is given in (12).

W_1 on Euclidean spaces In the special case of Euclidean spaces $\mathcal{X}=\mathcal{Y}=\mathbb{R}^d$, using $c(x,y)=\|x-y\|$, the global Lipschitz constraint appearing in (59) can be made local as a uniform bound on the gradient of f,

$$\mathcal{W}_1(\alpha,\beta)=\sup_f\left\{\int_{\mathbb{R}^d} f(\mathrm{d}\alpha-\mathrm{d}\beta)\;;\;\|\nabla f\|_\infty\leqslant 1\right\}. \tag{61}$$

Here the constraint $\|\nabla f\|_\infty \leqslant 1$ signifies that the norm of the gradient of f at any point x is upper bounded by 1, $\|\nabla f(x)\|_2 \leqslant 1$ for any x.

Considering the dual problem to (61), denoting $\xi \overset{\text{def.}}{=} \alpha - \beta$, and using that

$$\iota_{\|\cdot\|_{\mathbb{R}^d} \leqslant 1}(u) = \max_v \langle u, v \rangle - \|v\|_{\mathbb{R}^d}$$

one has a maximization on flow vector fields $s : \mathbb{R}^d \to \mathbb{R}^d$

$$\begin{aligned} \mathcal{W}_1(\alpha, \beta) &= \sup_f \inf_{s(x) \in \mathbb{R}^d} \int_{\mathbb{R}^d} f \mathrm{d}\xi - \int \langle \nabla f(x), s(x) \rangle \mathrm{d}x + \int \|s(x)\|_{\mathbb{R}^d} \mathrm{d}x \\ &= \inf_{s(x) \in \mathbb{R}^d} \int \|s(x)\| \mathrm{d}x + \sup_f \int f(x)(\mathrm{d}\xi - \mathrm{div}(s)\mathrm{d}x) \end{aligned}$$

one obtains an optimization problem under fixed divergence constraint

$$\mathcal{W}_1(\alpha, \beta) = \inf_s \left\{ \int_{\mathbb{R}^d} \|s(x)\|_{\mathbb{R}^d} \mathrm{d}x \;;\; \mathrm{div}(s) = \alpha - \beta \right\}, \tag{62}$$

which is often called the Beckmann formulation [2]. Here the vectorial function $s(x) \in \mathbb{R}^2$ can be interpreted as a flow field, describing locally the movement of mass. Outside the support of the two input measures, $\mathrm{div}(s) = 0$, which is the conservation of mass constraint. Once properly discretized using finite elements, Problems (61) and (62) become a nonsmooth convex optimization problems.

The previous formulations (61) and (62) of $\mathcal{W}_1$ can be generalized to the setting where $\mathcal{X}$ is a Riemannian manifold, i.e. $c(x, y) = d(x, y)$ where d is the associated geodesic distance (and then for smooth manifolds, the gradient and divergence should be understood as the differential operators on manifold). In a similar way it can be extended on a graph (where the geodesic distance is the length of the shortest path), in this case, the gradient and divergence are the corresponding finite difference operations operating along the edges of the graph. In this setting, the corresponding linear program can be solved using a min-cost flow simplex in complexity $O(n^2 \log(n))$ for sparse graph (e.g. grids).

6.4 Dual norms (Integral Probability Metrics)

Formulation (61) is a special case of a dual norm. A dual norm is a convenient way to design "weak" norms that can deal with arbitrary measures. For a symmetric convex set B of measurable functions, one defines

$$\|\alpha\|_B \overset{\text{def.}}{=} \sup_f \left\{ \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) \;;\; f \in B \right\}. \tag{63}$$

These dual norms are often called "integral probability metrics"; see [26].

Example 1 (Total variation). The total variation norm is a dual norm associated to the whole space of continuous functions

$$B = \{ f \in \mathcal{C}(\mathcal{X}) \;;\; \|f\|_\infty \leqslant 1 \}.$$

The total variation distance is the only nontrivial divergence that is also a dual norm; see [25].

Example 2 ($\mathcal{W}_1$ norm). $\mathcal{W}_1$ as defined in (61), is a special case of dual norm (63), using

$$B = \{ f \;;\; \mathrm{Lip}(f) \leqslant 1 \}$$

the set of 1-Lipschitz functions.

Example 3 (Flat norm and Dudley metric). If the set B is bounded, then $\|\cdot\|_B$ is a norm on the whole space $\mathcal{M}(\mathcal{X})$ of measures. This is not the case of $\mathcal{W}_1$, which is only defined for α such that $\int_{\mathcal{X}} \mathrm{d}\alpha = 0$ (otherwise

$\|\alpha\|_B = +\infty$). This can be alleviated by imposing a bound on the value of the potential f, in order to define for instance the flat norm,

$$B = \{f \;;\; \mathrm{Lip}(f) \leqslant 1 \quad \text{and} \quad \|f\|_\infty \leqslant 1\}\,. \tag{64}$$

It metrizes the weak convergence on the whole space $\mathcal{M}(\mathcal{X})$. Formula (60) is extended to compute the flat norm by adding the constraint $|\mathrm{f}_k| \leqslant 1$. The flat norm is sometimes called the "Kantorovich–Rubinstein" norm [15] and has been used as a fidelity term for inverse problems in imaging [19]. The flat norm is similar to the Dudley metric, which uses

$$B = \{f \;;\; \|\nabla f\|_\infty + \|f\|_\infty \leqslant 1\}\,.$$

The following proposition shows that to metrize the weak convergence, the dual ball B should not be too large (because otherwise one obtain a strong norm), namely one needs $\mathcal{C}(\mathcal{X}) \subset \overline{\mathrm{Span}(B)}$.

Proposition 23. *(i) If $\mathcal{C}(\mathcal{X}) \subset \overline{\mathrm{Span}(B)}$ (i.e. if the span of B is dense in continuous functions for the sup-norm $\|\cdot\|_\infty$), then $\|\alpha_k - \alpha\|_B \to 0$ implies $\alpha_k \rightharpoonup \alpha$.*
(ii) If $B \subset \mathcal{C}(\mathcal{X})$ is compact for $\|\cdot\|_\infty$ then $\alpha_k \rightharpoonup \alpha$ implies $\|\alpha_k - \alpha\|_B \to 0$.

Proof. (i) If $\|\alpha_k - \alpha\|_B \to 0$, then by duality, for any $f \in B$, since $\langle f,\, \alpha_k - \alpha\rangle \leqslant \|\alpha_k - \alpha\|_B$ then $\langle f,\, \alpha_k\rangle \to \langle f,\, \alpha\rangle$. By linearity, this property extends to $\mathrm{Span}(B)$. By density, this extends to $\overline{\mathrm{Span}(B)}$, indeed $|\langle f,\, \alpha_k\rangle - \langle f',\, \alpha_k\rangle| \leqslant \|f - f'\|_\infty$.
(ii) We assume $\alpha_k \rightharpoonup \alpha$ and we consider a sub-sequence α_{n_k} such that

$$\|\alpha_{n_k} - \alpha\|_B \longrightarrow \limsup_k \|\alpha_k - \alpha\|_B$$

Since B is compact, the maximum appearing in the definition of $\|\alpha_{n_k} - \alpha\|_B$ is reached, so that there exists some 1-Lipschitz function f_{n_k} so that $\langle \alpha_{n_k} - \alpha,\, f_{n_k}\rangle = \|\alpha_{n_k} - \alpha\|_B$. Once again, by compacity, we can extract from $(f_{n_k})_k$ a (not relabelled for simplicity) subsequence converging to some $f \in B$. One has $\|\alpha_{n_k} - \alpha\|_B = \langle \alpha_{n_k} - \alpha,\, f_{n_k}\rangle$, and this quantity converges to 0 because one can decompose it as

$$\langle \alpha_{n_k} - \alpha,\, f_{n_k}\rangle = \langle \alpha_{n_k} - \alpha,\, f\rangle + \langle \alpha_{n_k},\, f_{n_k} - f\rangle - \langle \alpha,\, f_{n_k} - f\rangle$$

and these three terms goes to zero because $\alpha_{n_k} - \alpha \rightharpoonup 0$ (first term) and $\|f_{n_k} - f\|_\infty \to 0$ (two others, recall that $|\langle \alpha_{n_k},\, f_{n_k} - f\rangle| \leqslant \|f_{n_k} - f\|_\infty$). □

Corollary 2. *On a compact space, the Wasserstein-p distance metrizes the weak convergence.*

Proof. Denoting $B = \{f \;;\; \mathrm{Lip}(f) \leqslant 1\}$.

For (i), one has that then $\mathrm{Span}(B)$ is the space of Lipschitz functions. The adherence of Lipschitz functions for $\|\cdot\|_\infty$ is the space of continuous functions. For (ii), for probability distributions, without loss of generality, functions f in B can be taken up to an additive constant, so that we can impose $f(x_0) = 0$ for some fixed $x_0 \in \mathcal{X}$, and since $\mathcal{X}$ is compact, $\|f\|_\infty \leqslant \mathrm{diam}(\mathcal{X})$ so that we can consider in placed of B another ball of equicontinuous bounded functions. By Ascoli-Arzelà theorem, it is hence compact. Proposition 8 shows that W_p has the same topology as W_1 so that it is also the topology of convergence in law. □

Dual RKHS Norms and Maximum Mean Discrepancies. It is also possible to define "Euclidean" norms (built using quadratic functionals) on measures using the machinery of kernel methods and more specifically reproducing kernel Hilbert spaces (RKHS; see [24] for a survey of their applications in data sciences), of which we recall first some basic definitions.

Definition 4. *A symmetric function k defined on $\mathcal{X} \times \mathcal{X}$ is said to be positive definite if for any $n \geqslant 0$, for any family $x_1, \ldots, x_n \in Z$ the matrix $(k(x_i, x_j))_{i,j}$ is positive (i.e. has positive eigenvalues), i.e. for all $r \in \mathbb{R}^n$*

$$\sum_{i,j=1}^n r_i r_j k(x_i, x_j) \geqslant 0, \tag{65}$$

The kernel is said to be conditionally positive if positivity only holds in (65) for zero mean vectors r (i.e. such that $\langle r, \mathbb{1}_n \rangle = 0$).

One of the most popular kernels is the Gaussian one $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$, which is a positive universal kernel on $\mathcal{X} = \mathbb{R}^d$.

If k is conditionally positive, one defines the following norm for $\xi = \alpha - \beta$ being a signed measure

$$\|\xi\|_k^2 \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\mathrm{d}\xi(x)\mathrm{d}\xi(y). \tag{66}$$

These norms are often referred to as "maximum mean discrepancy" (MMD) (see [14]) and have also been called "kernel norms" in shape analysis [13]. This expression (66) can be rephrased, introducing two independent random vectors (X, X') on $\mathcal{X}$ distributed with law α, as

$$\|\alpha\|_k^2 = \mathbb{E}_{X,X'}(k(X,X')).$$

One can show that $\|\cdot\|_k^2$ is the dual norm in the sense of (63) associated to the unit ball B of the RKHS associated to k. We refer to [3, 16, 24] for more details on RKHS functional spaces.

Remark 8 (Universal kernels). According to Proposition 23, the MMD norm $\|\cdot\|_k$ metrizes the weak convergence if the span of the dual ball B is dense in the space of continuous functions $\mathcal{C}(\mathcal{X})$. This means that finite sums of the form $\sum_{i=1}^n a_i k(x_i, \cdot)$ (for arbitrary choice of n and points $(x_i)_i$) are dense in $\mathcal{C}(\mathcal{X})$ for the uniform norm $\|\cdot\|_\infty$. For translation-invariant kernels over $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = k_0(x-y)$, this is equivalent to having a nonvanishing Fourier transform, $\hat{k}_0(\omega) > 0$.

In the special case where α is a discrete measure, one thus has the simple expression

$$\|\alpha\|_k^2 = \sum_{i=1}^n \sum_{i'=1}^n \mathrm{a}_i \mathrm{a}_{i'} \mathrm{k}_{i,i'} = \langle \mathrm{ka}, \mathrm{a} \rangle \quad \text{where} \quad \mathrm{k}_{i,i'} \overset{\text{def.}}{=} k(x_i, x_{i'}).$$

In particular, when $\alpha = \sum_{i=1}^n \mathrm{a}_i \delta_{x_i}$ and $\beta = \sum_{i=1}^n \mathrm{b}_i \delta_{x_i}$ are supported on the same set of points, $\|\alpha - \beta\|_k^2 = \langle \mathrm{k}(\mathrm{a}-\mathrm{b}), \mathrm{a}-\mathrm{b} \rangle$, so that $\|\cdot\|_k$ is a Euclidean norm (proper if k is positive definite, degenerate otherwise if k is semidefinite) on the simplex Σ_n. To compute the discrepancy between two discrete measures, one can use

$$\|\alpha - \beta\|_k^2 = \sum_{i,i'} \mathrm{a}_i \mathrm{a}_{i'} k(x_i, x_{i'}) + \sum_{j,j'} \mathrm{b}_j \mathrm{b}_{j'} k(y_j, y_{j'}) - 2\sum_{i,j} \mathrm{a}_i \mathrm{b}_j k(x_i, y_j). \tag{67}$$

References

[1] Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[2] Martin Beckmann. A continuous model of transportation. *Econometrica*, 20:643–660, 1952.

[3] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2003.

[4] Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

[5] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[6] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[7] Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

[8] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression beyond correct specification. *arXiv preprint arXiv:1610.06833*, 2016.

[9] Timothy M Chan. Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete & Computational Geometry*, 16(4):361–368, 1996.

[10] Peter J Forrester and Mario Kieburg. Relating the Bures measure to the Cauchy two-matrix model. *Communications in Mathematical Physics*, 342(1):151–187, 2016.

[11] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[12] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[13] Joan Glaunes, Alain Trouvé, and Laurent Younes. Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2004.

[14] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007.

[15] Leonid G Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

[16] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[17] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[18] LV Kantorovich and G.S. Rubinstein. On a space of totally additive functions. *Vestn Leningrad Universitet*, 13:52–59, 1958.

[19] Jan Lellmann, Dirk A Lorenz, Carola Schönlieb, and Tuomo Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[20] Arkadi Nemirovski and Uriel Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302:435–460, 1999.

[21] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[22] Hans Samelson et al. On the perron-frobenius theorem. *Michigan Mathematical Journal*, 4(1):57–59, 1957.

[23] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

[24] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[25] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,φ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[26] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.