Optimal Transport for Machine Learners
Course notes

Gabriel Peyré

CNRS and ENS, PSL Université

gabriel.peyre@ens.fr

May 13, 2025

Abstract

Optimal Transport is a foundational mathematical theory that connects optimization, partial differential equations, and probability. It offers a powerful framework for comparing probability distributions and has recently become an important tool in machine learning, especially for designing and evaluating generative models. These course notes cover the fundamental mathematical aspects of OT, including the Monge and Kantorovich formulations, Brenier's theorem, the dual and dynamic formulations, the Bures metric on Gaussian distributions, and gradient flows. It also introduces numerical methods such as linear programming, semi-discrete solvers, and entropic regularization. Applications in machine learning include topics like training neural networks via gradient flows, token dynamics in transformers, and the structure of GANs and diffusion models. These notes focus primarily on mathematical content rather than deep learning techniques.

Disclaimer: These notes are not intended to be a definitive resource on optimal transport. They should be viewed as an intermediate reference positioned between two extremal points: the book by Peyré and Cuturi [23], which focuses on computational aspects, and the book by Santambrogio [25], which emphasizes theoretical foundations.

1 Optimal Matching between Point Clouds

1.1 Monge Problem for Discrete Points

Matching Problem Given a cost matrix $(C_{i,j})_{i \in [\![n]\!], j \in [\![m]\!]}$ and assuming $n = m$, the optimal assignment problem aims to find a bijection σ within the set $\mathrm{Perm}(n)$ of permutations of n elements that solves

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

One could naively evaluate the cost function above using all permutations in the set $\mathrm{Perm}(n)$. However, this set has size $n!$, which becomes enormous even for small values of n. In general, the optimal σ is not unique.

1D Case If the cost is of the form $C_{i,j} = h(x_i - y_j)$, where $h : \mathbb{R} \to \mathbb{R}^+$ is convex (for example, $C_{i,j} = |x_i - y_j|^p$ for $p \geqslant 1$), it follows that an optimal σ necessarily defines an increasing map $x_i \mapsto y_{\sigma(i)}$, i.e.,

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

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

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

because
$$h(x_i - y_{\tilde{\sigma}(i')}) + h(x_{i'} - y_{\tilde{\sigma}(i)}) \leqslant h(x_i - y_{\sigma(i)}) + h(x_{i'} - y_{\sigma(i')}).$$
Therefore, the algorithm to compute an optimal transport is to sort the points, i.e., find some pair of permutations σ_X, σ_Y such that

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

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, the quicksort algorithm. Note that if $\varphi : \mathbb{R} \to \mathbb{R}$ is an increasing map, one can apply this technique to costs of the form $h(|\varphi(x) - \varphi(y)|)$ with a change of variable. A typical application is the grayscale histogram equalization of the luminance of images.

Note that if h is strictly convex, then all optimal assignement are increasing, so if the points are all distincts, there is a unique such increasing map. But if h is not strictly convex, for instance $c(x, y) = |x - y|$ then there exists non increasing optimal assignement, for instance in the book shifting problem, with overlapping uniform distribution (the mass at the intesection can stay at the same place).

This efficient strategy to compute the OT in 1-D does not extend to higher dimensions. In 2-D, if the cost is $c(x, y) = \|x - y\|$, then, as already noted by Monge, trajectories cannot cross, this is a consequence of the parallelogram inequality. However, this is not enough to uniquely determine an optimal matching.

Note that if h is concave instead of being convex, then the behavior is entirely different, and the optimal match tends to exchange the positions. In this case, there exists an $O(n^2)$ algorithm.

1.2 Matching Algorithms

Efficient algorithms exist to solve the optimal matching problem. The most well-known are the Hungarian and the auction algorithm, which run in $O(n^3)$ operations. Their derivation and analysis, however, are greatly simplified by introducing the Kantorovich relaxation and its associated dual problem. A typical application of these methods is equalizing the color palette between images, corresponding to a 3-D optimal transport.

2 Monge Problem between Measures

The presentation of the previous section could only handle two sets with the same number of points. To relax this to more general setting, one needs to consider probability distribution, so that the points are now weighted by some mass.

2.1 Measures

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

$$\Sigma_n := \left\{ a \in \mathbb{R}_+^n \; ; \; \sum_{i=1}^n a_i = 1 \right\}.$$

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

$$\alpha = \sum_{i=1}^n a_i \delta_{x_i} \tag{2}$$

where δ_x is the Dirac at position x, intuitively a unit of mass which is infinitely concentrated at location x. Such a measure describes a probability measure if, additionally, 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. a $= \frac{1}{n} \sum_i \delta_{x_i}$. In practice, in many applications, it 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 sets. A Dirac measure δ_x is then defined as $\delta_x(A) = 1$ is $x \in A$ and 0 otherwise, and this extends by linearity for discrete measures of the form (2) as

$$\alpha(A) = \sum_{x_i \in A} a_i$$

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 the integral of measurable functions (i.e. such that level sets $\{x \; ; \; f(x) < t\}$ are Borel sets), and we denote this pairing as

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

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

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

This can be applied in particular to the subspace of continuous functions that 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 with 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 functions and measures will be crucial to developing 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

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

(note that one can remove the $|\cdot|$ on the right-hand side, and such a quantity is often called a "dual norm"). One can 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 $\alpha = \sum_i a_i \delta_{x_i}$, $|\alpha| = \sum_i |a_i| \delta_{x_i}$ and if $\mathrm{d}\alpha(x) = \rho \mathrm{d}x$ for a positif reference measure $\mathrm{d}x$, then $\mathrm{d}|\alpha|(x) = |\rho(x)| \mathrm{d}x$.

3

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

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

For more general measures, for instance for those with a density, the notion of push-forward plays a fundamental role in describing 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 $y = T(x)$, so that $\mathrm{d}y = |\det(T'(x))|\mathrm{d}x$ 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 measures with the same number $n = m$ of points, one retrieves the optimal matching problem. Indeed, this corresponds to the setting of empirical measures $\alpha = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\beta = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$. In this case, $T_\sharp \alpha = \beta$ necessarily implies that σ is subjective, hence it is one-to-one, and it induces a permutation σ on the support, $T : x_i \mapsto x_{\sigma(i)}$, 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. The constraint set $T_\sharp \alpha = \beta$, which is the case for instance if $\alpha = \delta_x$ and β is not a single Dirac.

Note that 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.

Semi-discrete setting. It is also not a symmetric problem in α and β. For instance, this problem makes sense if α has a density with respect to Lebesgue and β is discrete. It is a semi-discrete problem, which can be understood as an optimal quantization problem. Indeed, for instance on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, one can consider a setting where α has a density with respect to Lebesgues and $\beta = \sum_j b_j \delta_{y_j}$ is discrete, supported on $\{y_1, \ldots, y_m\}$. In this case, a map T so that $T_\sharp \alpha = \beta$ defines a segemtation of the space into m cells $C_j := T^{-1}(y_j)$, so that $\alpha(C_j) = b_j$. This is often reffered to as the semi-discrete setting. If one exchanges the role of α and β (so that α is discrete) then there is not any valid T. Indeed, it is not possible to push-forward a discrete measure to a measure with density, so that the constraint set of the Monge problem is empty.

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

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

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 (pseudo-) distance (because it is not symmetric).

Proposition 1. $\tilde{\mathcal{W}}$ *is a non-symmetric 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

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

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

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

$$\leqslant \mathcal{W}_p(\alpha, \beta) + \mathcal{W}_p(\beta, \gamma) + 2\varepsilon.$$

Letting $\varepsilon \to 0$ gives the result. $\qquad\square$

This quantity $\tilde{\mathcal{W}}_p$ is problematic because it can have the value $+\infty$. It is the purpose of the Kantorovitch formulation exposed below to remedy this issue, to define a symmetric and well-behaved distance $\mathcal{W}_p$.

2.4 Existence and Uniqueness of the Monge Map

Brenier's theorem. The following celebrated theorem of [7] ensures that in $\mathbb{R}^d$ for $p = 2$, if at least one of the two input measures has a density, then Monge's problem has a unique solution, and it gives a precise description of the structure of this solution.

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 of 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

$$\forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \langle \nabla\varphi(x) - \nabla\varphi(x'), \, x - x' \rangle \geqslant 0.$$

Note however that in dimensions larger than 1, not all monotone fields are gradients of convex functions. For instance, a small enough 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 with a quadratic form $\varphi(x) = \langle Bx, x \rangle / 2$ and hence $A = \nabla\varphi = (B + B^\top)/2$. Optimal transport can thus play an important role in defining quantile functions in arbitrary dimensions, which in turn is useful for applications to quantile regression problems [9].

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 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 convex and Lipschitz so that $\nabla\varphi$ is well defined α-almost everywhere. Ensuring T to be smooth 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 exist 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 $T(x) \subset \partial\varphi$. 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-Ampère-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 $\det(\mathrm{Id} + \varepsilon A) = 1 + \varepsilon\operatorname{tr}(A) + o(\varepsilon)$.

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

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

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 $\gamma := (\mathcal{C}_\alpha)_\sharp^{-1}\mathcal{U}$ 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

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

$\qquad\square$

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 measures are uniform on the same number of Dirac masses, then this approach is 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 types of norm which behave similarly (i.e. metricize 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 derivative of the convex function

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

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 $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \boldsymbol{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \boldsymbol{\Sigma}_\beta)$ 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. An affine function maps a Gaussian to a Gaussian so that it remains to show that the mean and covariance of $T_\sharp \alpha$ are those of β. We consider X to be a random vector with law α, so that, denoting $Y := T(X)$ a random vector with law $T_\sharp \alpha$, its mean is

$$\mathbb{E}(Y) = \mathbb{E}(\mathbf{m}_\beta + A(X - \mathbf{m}_\alpha)) = \mathbf{m}_\beta + A\mathbb{E}(X - \mathbf{m}_\alpha) = \mathbf{m}_\beta$$

while it covariance is

$$\mathbb{E}((Y - \mathbf{m}_\beta)(Y - \mathbf{m}_\beta)^\top) = \mathbb{E}([A(X - \mathbf{m}_\alpha)][A(X - \mathbf{m}_\alpha)]^T) = A\mathbb{E}((X - \mathbf{m}_\alpha)(X - \mathbf{m}_\alpha)^T)A^T = A\mathbf{\Sigma}_\alpha A.$$

$\square$

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 [7], 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 [8] between positive definite matrices (see also [12]),

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

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 $\mathbf{\Sigma}_\alpha = \text{diag}(r_i)_i$ and $\mathbf{\Sigma}_\beta = \text{diag}(s_i)_i$ 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 (a_i, 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 [19] 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 $P \in \mathbb{R}_+^{n \times m}$, where $P_{i,j}$ 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 $\alpha = \sum_i a_i \delta_{x_i}$, $\beta = \sum_j b_j \delta_{y_j}$. Admissible couplings are only constrained to satisfy the conservation of mass

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

where we used the following matrix-vector notation

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

The set of matrices $U(a, b)$ 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 $U(a, b)$ if and only if $P^\top$ is in $U(b, a)$.

Kantorovitch, aiming for the planification of the economy, made a strong simplifying assumption: the cost of transportation should be linear in the amount of transport mass. Under this assumption, denoting $C_{i,j}$ the cost of moving a unit amount of mass from x_i to y_j, Kantorovich's optimal transport problem now reads

$$L_C(a, b) := \min_{P \in U(a,b)} \langle C, P \rangle := \sum_{i,j} C_{i,j} P_{i,j}. \tag{18}$$

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 plans, in the sense that there is always a solution P to (18) with $n + m - 1$ solutions. This is because there is $n + m$ linear constraints, but the rank of the constraint is $n + m - 1$ (there is a single redundancy among the constraint, which stems from the fact that both $P\mathbb{1}_m = \mathbb{1}_n$ and $P^\top \mathbb{1}_n = \mathbb{1}_m$, then $\sum_{i,j} P_{i,j} = 1$.

Linear programming algorithms. The reference algorithms to solve (18) are network simplexes. There exist instances of this method that scale like $O(n^3 \log n)$. Alternatives include interior points, which are usually inferior in 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 $a_i = b_j = \frac{1}{n}$ 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 $a_i = b_i = 1$ (up to scaling by $1/n$, these are thus probability measures). In this case, one can solve Monge's optimal matching problem (1), and it is convenient to rewrite it using permutation matrices. For a permutation $\sigma \in \text{Perm}(n)$, we write P_σ for the corresponding permutation matrix,

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

We denote the set of permutation matrices as

$$\mathcal{P}_n := \{ P_\sigma \; ; \; \sigma \in \text{Perm}(n) \},$$

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

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

so that the 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 the Monge and Kantorovitch problems

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

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

For this, we will make a detour through a 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 first introduce the notion of extremal points, which are intuitively the vertices of $\mathcal{C}$

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

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 solutions (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 $\mathcal{S} := \operatorname*{argmin}_{\mathrm{P} \in \mathcal{C}} \langle \mathrm{C}, \mathrm{P} \rangle$. 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 $\mathrm{Extr}(\mathcal{S}) \subset \mathrm{Extr}(\mathcal{C})$. $\qquad \square$

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 $P \in \mathcal{P}_n$, if $P = (Q+R)/2$ with $Q_{i,j}, R_{i,j} \in [0,1]$, since $P_{i,j} \in \{0,1\}$ 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 $P \in \mathcal{B}_n \backslash \mathcal{P}_n$, we need to split $P = (Q+R)/2$ where Q, R are distinct bistochastic matrices. P defines a bipartite graph linking two sets of n vertices. This graph is composed of isolated edges when $P_{i,j} = 1$ and connected edges corresponding to $0 < P_{i,j} < 1$. If i is such a connected vertex on the left (similarly for j on the right), because $\sum_j P_{i,j} = 1$, there are necessarily at least two edges (i, j_1) and (i, j_2) emanating from it (similarly 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 always 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 cycles. Along this cycle, the left-right and right-left edges satisfy

$$0 < P_{i_s, j_s}, P_{i_{s+1}, j_s} < 1.$$

The $(i_s)_s$ and $(j_s)_s$ are also all distinct because the cycle is the shortest. Let us pick

$$\varepsilon := \min_{0 \leqslant s \leqslant p} \left\{ P_{i_s, j_s}, P_{j_s, i_{s+1}}, 1 - P_{i_s, j_s}, 1 - P_{j_s, i_{s+1}} \right\}$$

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

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

We define the two matrices as

$$Q_{i,j} := \left\{ \begin{array}{ll} P_{i,j} & \text{if} \quad (i,j) \notin \mathcal{A} \cup \mathcal{B}, \\ P_{i,j} + \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{A}, \\ P_{i,j} - \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{B}, \end{array} \right. \quad \text{and} \quad R_{i,j} := \left\{ \begin{array}{ll} P_{i,j} & \text{if} \quad (i,j) \notin \mathcal{A} \cup \mathcal{B}, \\ P_{i,j} - \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{A}, \\ P_{i,j} + \varepsilon/2 & \text{if} \quad (i,j) \in \mathcal{B}, \end{array} \right. .$$

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 with a right-left edge in $\mathcal{B}$, (and the other way) the sum constraint on the row (and on the column) is maintained, so that $U, V \in \mathcal{B}_n$. Finally, note that $P = (P+Q)/2$. $\square$

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 $a = b = \mathbb{1}_n$, then there exists an optimal solution for Problem (18) $P_{\sigma^\star}$, 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 $a = b = \mathbb{1}_n/n$, 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 $P\mathbb{1}_m = \mathrm{a}, P\mathbb{1}_n = \mathrm{b}$ 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 $\pi_1 \coloneqq P_{1\sharp}\pi$ and $\pi_2 \coloneqq P_{2\sharp}\pi$ 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

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

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 functions, 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

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

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 $\pi = \sum_{i,j} P_{i,j} \delta_{(x_i, y_j)}$. 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.

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

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 through 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 $\pi_2 = \beta$.

A last important class of examples are semi-discrete problems (for instance to perform quantization or to fit statistical models), where α has a density with respect to Lebesgue and β is discrete. In this case, couplings π are singular, for instance, if $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, they are supported on d-dimensional subspace of $\mathbb{R}^d \times \mathbb{R}^d$.

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

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

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 measures is compact, and a linear function with a continuous $c(x, y)$ is weak-* continuous. And the set of constraints is non-empty, taking $\alpha \otimes \beta$. On non-compact domains, one needs to impose moment conditions 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 is unlikely that such a coupling is optimal. Indeed as stated by Brenier's theorem, optimal coupling for a square Euclidean loss on the contrary describes fully dependent variables, 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

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

Monge-Kantorovitch equivalence. The proof of Brenier theorem 1 (detailed in Section 5.3, Remark 8) to prove the existence of a Monge map relies on 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.

Corollary 2. *We assume α has a density with respect to Lebegues, and that $c(x, y) = \|x - y\|^2$. We denote T as the optimal transport solving Monge's formulation. Then $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is the unique optimal coupling solving Kantorovitch formulation. Monge and Kantorovitch's costs are the same.*

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 points 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 terms of random vectors, 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 imply uniqueness). Note however that the settings are radically different (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 histograms or measures. The proof of this result relies on a "gluing lemma", which we first prove in the discrete case.

14

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

$$\sum_k S_{i,j,k} = P_{i,j} \quad and \quad \sum_i S_{i,j,k} = Q_{j,k}.$$

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

Proof. One verifies that

$$S_{i,j,k} = \begin{cases} \frac{P_{i,j} Q_{j,k}}{b_j} & \text{if} \quad b_j \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{23}$$

is acceptable. Indeed, if $b_j \neq 0$

$$\sum_k S_{i,j,k} = \sum_k \frac{P_{i,j} Q_{j,k}}{b_j} = \frac{P_{i,j}}{b_j} (Q\mathbb{1}_m)_j = \frac{P_{i,j}}{b_j} b_j.$$

If $b_j = 0$, then necessarily $P_{i,j} = 0$ and $\sum_k S_{i,j,k} = 0 = P_{i,j}$. □

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

1. $D \in \mathbb{R}_+^{n \times n}$ *is symmetric;*

2. $D_{i,j} = 0$ *if and only if* $i = j$;

3. $\forall (i, j, k) \in [\![n]\!]^3, D_{i,k} \leqslant D_{i,j} + D_{j,k}$.
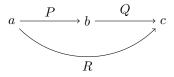
Then

$$W_p(a, b) := L_{D^p}(a, b)^{1/p} \tag{24}$$

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

$$\forall a, b, c \in \Sigma_n, \quad W_p(a, c) \leqslant W_p(a, b) + W_p(b, c).$$

Proof. For the symmetry, since D^p is symmetric, we use the fact that if $P \in U(a, b)$ is optimal for $W_p(a, b)$, then $P^\top \in U(b, a)$ is optimal for $W_p(b, a)$. For the definiteness, since $C = D^p$ has a null diagonal, $W_p(a, b) = 0$, with corresponding optimal transport matrix $P^\star = \text{diag}(a) = \text{diag}(b)$; by the positivity of all off-diagonal elements of D^p, $W_p(a, b) > 0$ whenever $a \neq b$ (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 $a, b, c \in \Sigma_n$, 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 $S \in \mathbb{R}_+^{n^3}$ with marginals $\sum_k S_{\cdot,\cdot,k} = P$ and $\sum_i S_{i,\cdot,\cdot} = Q$. We define $R = \sum_j S_{\cdot,j,\cdot}$, which is an element of $U(a, c)$.

$$a \xrightarrow{\quad P \quad} b \xrightarrow{\quad Q \quad} c$$
$$R$$

Note that if one assumes $b > 0$ then $R = P \, \text{diag}(1/b) Q$.

The triangle inequality follows from

$$W_p(a, c) = \Big(\min_{\tilde{R} \in U(a,c)} \langle \tilde{R}, D^p \rangle \Big)^{1/p} \leqslant \langle R, D^p \rangle^{1/p}$$

$$= \Big(\sum_{i,k} D_{ik}^p \sum_j S_{i,j,k} \Big)^{1/p} \leqslant \Big(\sum_{i,j,k} \big(D_{ij} + D_{j,k} \big)^p S_{i,j,k} \Big)^{1/p}$$

$$\leqslant \Big(\sum_{i,j,k} D_{ij}^p S_{i,j,k} \Big)^{1/p} + \Big(\sum_{i,j,k} D_{j,k}^p S_{i,j,k} \Big)^{1/p}$$

$$= \Big(\sum_{i,j} D_{i,j}^p \sum_k S_{i,j,k} \Big)^{1/p} + \Big(\sum_{j,k} D_{j,k}^p \sum_i S_{i,j,k} \Big)^{1/p}$$

$$= \Big(\sum_{i,j} D_{i,j}^p P_{i,j} \Big)^{1/p} + \Big(\sum_{j,k} D_{j,k}^p Q_{j,k} \Big)^{1/p} = W_p(a, b) + W_p(b, b).$$

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. $\qquad \square$

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*

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

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 $\beta = \sum_i b_j \delta_{y_j}$ and $\pi = \sum_{i,j} P_{i,j} \delta_{y_j}$, then this conditional distribution is defined on the support of β as $\pi_{y_j} = \sum_i \frac{P_{i,j}}{b_j} \delta_{x_i}$ (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 $\sigma = \sum_{i,j,k} S_{i,j,k} \delta_{x_i,y_j,z_k}$ where

$$S_{i,j,k} = \frac{P_{i,j}}{b_j} \frac{Q_{j,k}}{b_j} b_j.$$

$\qquad \square$

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

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

(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 $\Delta := \{(x,x)\}_x \subset \mathcal{X}^2$. 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 $\rho := (P_{\mathcal{X}, \mathcal{Z}})_\sharp \sigma$. Note that if π and ξ are couplings 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

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

$\square$

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. 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 behaves 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 $\mathrm{diam}(\mathcal{X}) \triangleq \sup\limits_{x,y} d(x,y)$.

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

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

$\square$

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 $f \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} f\mathrm{d}\alpha_k \to \int_{\mathcal{X}} f\mathrm{d}\alpha$.*

Remark 4 (Weak* convergence for discrete measures). In the special case of a single Dirac, $\delta_{x^{(n)}} \rightharpoonup \delta_x$ is equivalent to $\int f\mathrm{d}\delta_{x^{(n)}} = f(x^{(n)}) \to \int f\mathrm{d}\delta_x = f(x)$ for any continuous f. This in turn is equivalent to $x^{(n)} \to x$. More generally, a sequence of discrete measures of the form $\alpha_n := \sum_k a_i^{(n)}\delta_{x_i^{(n)}}$ converges toward some measure $\sum_j b_j\delta_{y_j}$ if and only if $a_i^{(n)} \to a_i$, and for those $a_i > 0$, $x_i^{(n)} \to y_{\sigma(j)}$ for some injection σ (in particular, the limit measure cannot have more point than the α_n). Furthermore, one must also have the balance of mass $\sum_{i:\sigma(i)=j} a_i = b_j$ for all j.

In terms 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 5 (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 $\mathbb{E}(X_i) = 0$ and $\mathbb{E}(X_iX_i^\top) = \mathrm{Id}$, 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*

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

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 $c_i = \min(a_i,b_i)$. By conservation of mass, for $\mathrm{P} \in \mathrm{U}(a,b)$, $\mathrm{P}_{i,i} \leqslant c_i$, thus

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

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

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

One has

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

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

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

satisfies

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

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

$$W_1(a, b) = \langle D, \hat{P} \rangle = \sum_{i,j} \frac{a_i b_j}{\langle \bar{a}, \mathbb{1} \rangle} = \sum_i \bar{a}_i = \sum_i \bar{b}_i == \frac{1}{2} \sum_i \bar{a}_i + \bar{b}_i = \frac{1}{2} \|a - b\|_{\text{TV}}.$$

$\square$

As explained in Remark 4, in the special case of Diracs, $\delta_{x_n} \rightharpoonup \delta_x$ is equivalent to $x_n \to x$. One can then contrast the strong topology with the Wasserstein distance if $x_n \neq x$,

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

This shows that for the strong topology, Diracs never converge, while they do converge for the Wasserstein distance. 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 3. On non-compact spaces, one needs also to impose the convergence of the moments up to order p. Note that there exist alternative distances which also metricize weak convergence. The simplest ones 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 the contrary, one has that for all n, since the two measures are mutually singular

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

so that there is no strong convergence.

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*

$$\frac{d_{\min}}{2}\|\alpha-\beta\|_{\mathrm{TV}} \leqslant \mathcal{W}_1(\alpha,\beta) \leqslant \frac{d_{\max}}{2}\|\alpha-\beta\|_{\mathrm{TV}} \quad where \quad \left\{ \begin{array}{l} d_{\min} := \inf\limits_{x\neq y} d(x,y) \\ d_{\max} := \sup\limits_{x,y} d(x,y) \end{array} \right.$$

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). $\square$

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 blow as $d_{\max}/d_{\min}$ increases, and on non-discrete space, if $d_{\min} = 0$, then the two distances are not equivalent, which is inlined with the fact that the strong and the weak topology do not coincide.

Remark 6 (Berry-Esseen theorem). The Wasserstein distance is a natural candidate to quantify the convergence in law in the central limit theorem (Remark 5). To obtain rates, one needs further assumption on the random vector, and the Berry-Esseen theorem ensures that if $\mathbb{E}(\|X_i\|^3) < +\infty$, then $\mathcal{W}_p(\alpha_k,\alpha) = O(\mathbb{E}(\|X_i\|^3)/\sqrt{n})$.

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)

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

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 later, this forces the solution P to be strictly positive on the support of $\mathrm{a}\otimes\mathrm{b}$. We will also show that as $\varepsilon \to +\infty$, then the solution $\mathrm{P} \to \mathrm{a}\otimes\mathrm{b}$

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 $U(a, b)$ has nm variables but $n + m$ constraints.

Proposition 12. P *is the unique solution to* (26) *if and only if there exists* $(u, v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ *such that*

$$\forall (i, j) \in [\![n]\!] \times [\![m]\!], \quad P_{i,j} = u_i K_{i,j} v_j \quad where \quad K_{i,j} := e^{-\frac{C_{i,j}}{\varepsilon}}, \tag{27}$$

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

Proof. Without loss of generality, we assume $a_i, b_j > 0$ (otherwise, we can set the corresponding u_i or v_j to 0).

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

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

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

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

$$\frac{\partial \mathcal{E}(P, f, g)}{\partial P_{i,j}} = C_{i,j} + \varepsilon(\log (P_{i,j}) + 1) - f_i - g_j = 0.$$

which results, in an optimal P coupling of the regularized problem, in the expression $P_{i,j} = e^{\frac{f_i + g_j - C_{i,j}}{\varepsilon} - 1}$ which can be rewritten in the form provided in the proposition using non-negative vectors $u := (e^{f_i/\varepsilon} - 1)_i$ and $v := (e^{g_j/\varepsilon})_j$. $\qquad \square$

The factorization of the optimal solution exhibited in Equation (27) can be conveniently rewritten in matrix form as $P = \text{diag}(u)K \text{diag}(v)$. u, v must therefore satisfy the following non-linear equations which correspond to the mass conservation constraints inherent to $U(a, b)$,

$$\text{diag}(u)K \text{diag}(v)\mathbb{1}_m = a, \quad and \quad \text{diag}(v)K^\top \text{diag}(u)\mathbb{1}_n = b, \tag{28}$$

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

$$u \odot (Kv) = a \quad and \quad v \odot (K^\top u) = b \tag{29}$$

where $\odot$ corresponds to the entry-wise multiplication of vectors. This problem is known in the numerical analysis community as the matrix scaling problem (see [22] and references therein).

The problem of normalizing a positive matrix K by diagonal scaling is well known, in particular when $n = m$ and a and b are uniform. This corresponds to the question of diagonal scaling toward bi-stochasticity, which is a very old problem. The previous result shows that there is a unique such scaled matrix P, thanks to the strong convexity of the regularized problem. However, the question remains to find in practice this scaled matrix. Note also that if some entry of K vanish (or equivalently if the cost matrix C can have infinite values)

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 (29) and then v to satisfy its right-hand side. These two updates define Sinkhorn's algorithm

$$u^{(\ell+1)} := \frac{a}{Kv^{(\ell)}} \quad and \quad v^{(\ell+1)} := \frac{b}{K^\top u^{(\ell+1)}}, \tag{30}$$

initialized with an arbitrary positive vector, for instance $v^{(0)} = \mathbb{1}_m$. 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 (28) then so do $\lambda u, v/\lambda$ for any $\lambda > 0$). It turns out however that these iterations converge, as we detail next.

A chief advantage, besides its simplicity, of Sinkhorn's algorithm is that the only computationally expensive step is matrix-vector multiplication by the Gibbs kernel so that its complexity scales like Cnm where C is the number of Sinkhorn iteration, which can be shown to be of the order $1/\varepsilon^2$ f one is interested in reaching an accuracy ε on the (unregularized) transportation cost. Note however that in many situations, 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 $-\sum_i \log(P_{i,j})$. These algorithms 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 problems with a fixed cost matrix C, one can replace many matrix-vector multiplications with matrix-matrix multiplications, 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 below) is to consider the relative entropy, also called Kullback-Leibler divergence, which is defined as

$$\mathrm{KL}(P|Q) := \sum_{i,j} P_{i,j} \log\left(\frac{P_{i,j}}{Q_{i,j}}\right) - P_{i,j} + Q_{i,j}. \tag{31}$$

with the convention $0 \log(0) = 0$ and $\mathrm{KL}(P|Q) = +\infty$ if there exists some (i, j) such that $Q_{i,j} = 0$ but $P_{i,j} \neq 0$. For the specific case of comparing probability distribution, this further simplifies to

$$\mathrm{KL}(P|Q) = \sum_{i,j} P_{i,j} \log\left(\frac{P_{i,j}}{Q_{i,j}}\right).$$

The Shannon-Boltzmann neg-entropy is obtained when considering $Q = \mathbb{1}_{n \times m}$, i.e.

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

KL is a particular instance (and the unique case) of both a φ-divergence (as defined in Section 7.1) 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 $\mathrm{KL}(P|Q) \geqslant 0$ and $\mathrm{KL}(P|Q) = 0$ if and only if $P = Q$ (see Proposition 2). Indeed, it reads

$$\mathrm{KL}(P|Q) = \sum_{i,j} \varphi(P_{i,j}/Q_{i,j})Q_{i,j}.$$

where $\varphi(s) = s \log(s)$. For any convex φ such that $\varphi(1) = 0$, one has indeed by Jensen

$$\sum_{i,j} \varphi(P_{i,j}/Q_{i,j})Q_{i,j} \geqslant \varphi(\sum_{i,j} P_{i,j}/Q_{i,j}Q_{i,j}) = \varphi(\sum_{i,j} P_{i,j}) = \varphi(1) = 0.$$

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

$$\min_{P \in U(a,b)} \langle P, C \rangle - \varepsilon \, \mathrm{KL}(P|a \otimes b). \tag{32}$$

Note also that for later, it will be important to have this normalization when considering unbalanced OT .

But, 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, (32) and (26) have the same unique solution.

Proposition 13. *For* $P \in U(a, b)$*, one has*

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

Proof. This follows from

$$\sum_{i,j} P_{i,j} \log \frac{P_{i,j}}{a_i b_j} = \sum_{i,j} P_{i,j} \log \frac{P_{i,j}}{a'_i b'_j} + \sum_{i,j} P_{i,j} \left(\log \frac{a'_i}{a_i} + \log \frac{b'_j}{b_j} \right).$$

$\square$

The choice of using the reference measure $a \otimes b$ is however important to deal with situations 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.

One has the following convergence property.

Proposition 14 (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*

$$P_\varepsilon \xrightarrow{\varepsilon \to 0} \underset{P}{\mathrm{argmin}} \; \{-H(P) \; ; \; P \in U(a, b), \langle P, C \rangle = L_C(a, b)\} \tag{33}$$

so that in particular

$$L_C^\varepsilon(a, b) \xrightarrow{\varepsilon \to 0} L_C(a, b).$$

One has

$$P_\varepsilon \xrightarrow{\varepsilon \to \infty} a \otimes b. \tag{34}$$

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 $U(a, b)$ is bounded, we can extract a sequence (that we do not relabel for the sake of simplicity) such that $P_\ell \to P^\star$. Since $U(a, b)$ is closed, $P^\star \in U(a, b)$. We consider any P such that $\langle C, P \rangle = L_C(a, b)$. By optimality of P and P_ℓ for their respective optimization problems (for $\varepsilon = 0$ and $\varepsilon = \varepsilon_\ell$), one has

$$0 \leqslant \langle C, P_\ell \rangle - \langle C, P \rangle \leqslant \varepsilon_\ell (KL(P_\ell|a \otimes b) - KL(P|a \otimes b)). \tag{35}$$

Since KL is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle C, P^\star \rangle = \langle C, P \rangle$ so that $P^\star$ is a feasible point of (33). Furthermore, dividing by ε_ℓ in (35) and taking the limit shows that $KL(P|a \otimes b) \leqslant KL(P^\star|a \otimes b)$, which shows that $P^\star$ is a solution of (33). Since the solution $P_0^\star$ to this program is unique by strict convexity of $KL(\cdot|a \otimes b)$, one has $P^\star = P_0^\star$, and the whole sequence is converging.

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

$$\langle C, P_\varepsilon \rangle + \varepsilon KL(P_\varepsilon|a \otimes b) \leqslant \langle C, a \otimes b \rangle + \varepsilon \times 0$$

and since $\langle C, P_\varepsilon \rangle \geqslant 0$, this leads to

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

so that $KL(P_\varepsilon|a \otimes b) \to 0$ and thus $P_\varepsilon \to a \otimes b$ since KL is a valid divergence. $\square$

4.4 General Formulation

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

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

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence (31)

$$\operatorname{KL}(\pi | \xi) := \int_{\mathcal{X} \times \mathcal{Y}} \log\left(\frac{d\pi}{d\xi}(x, y)\right)d\pi(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} (d\xi(x, y) - d\pi(x, y)), \tag{37}$$

and by convention $\operatorname{KL}(\pi | \xi) = +\infty$ if π does not have a density $\frac{d\pi}{d\xi}$ 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 $\operatorname{KL}(\cdot | \alpha \otimes \beta)$ plays no specific role (because Proposition 13 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 particles of gas 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 gas, and particles do not move in a deterministic straight line as in optimal transport for the Euclidean cost, but rather according to a stochastic Brownian bridge.

Remark 7 (Probabilistic interpretation). If $(X, Y) \sim \pi$ have marginals $X \sim \alpha$ and $Y \sim \beta$, then $\operatorname{KL}(\pi | \alpha \otimes \beta) = \mathcal{I}(X, Y)$ 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 carries over to general measures. This analysis is revisited in Section 8.2Â using convex duality.

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*

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

Proof. The objective is indeed equal to

$$\varepsilon \sum_{i,j} P_{i,j} \frac{C_{i,j}}{\varepsilon} + \log\left(\frac{P_{i,j}}{a_i b_j}\right) P_{i,j} = \varepsilon \sum_{i,j} P_{i,j} \log\left(\frac{P_{i,j}}{a_i b_j e^{-C_{i,j}/\varepsilon}}\right) = \varepsilon \sum_{i,j} P_{i,j} \log\left(\frac{P_{i,j}}{K_{i,j}}\right) + cst.$$

$\square$

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

$$P_\varepsilon = \operatorname{Proj}_{U(a,b)}^{\operatorname{KL}}(K) := \underset{P \in U(a,b)}{\operatorname{argmin}} \operatorname{KL}(P | K). \tag{38}$$

Denoting

$$\mathcal{C}_a^1 := \{P \; ; \; P\mathbb{1}_m = a\} \quad \text{and} \quad \mathcal{C}_b^2 := \{P \; ; \; P^\top \mathbb{1}_m = b\}$$

the rows and columns constraints, one has $U(a, b) = \mathcal{C}_a^1 \cap \mathcal{C}_b^2$. One can use Bregman iterative projections [6]

$$P^{(\ell+1)} := \text{Proj}_{\mathcal{C}_a^1}^{\text{KL}}(P^{(\ell)}) \quad \text{and} \quad P^{(\ell+2)} := \text{Proj}_{\mathcal{C}_b^2}^{\text{KL}}(P^{(\ell+1)}). \tag{39}$$

Since the sets $\mathcal{C}_a^1$ and $\mathcal{C}_b^2$ are affine, these iterations are known to converge to the solution of (38), see [6].

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

Proposition 16. *One has*

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

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

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

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 $p = sp/\sum_i q_i$. $\qquad\square$

These iterations are equivalent to Sinkhorn iterations (30) since defining

$$P^{(2\ell)} := \text{diag}(u^{(\ell)}) K \, \text{diag}(v^{(\ell)}),$$

one has

$$P^{(2\ell+1)} := \text{diag}(u^{(\ell+1)}) K \, \text{diag}(v^{(\ell)})$$
$$\text{and} \quad P^{(2\ell+2)} := \text{diag}(u^{(\ell+1)}) K \, \text{diag}(v^{(\ell+1)})$$

In practice however one should prefer using (30) 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 interest 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 also with ε). It is also possible to decay ε during the iterates to improve the speed and rely on multiscale strategies in low dimensions.

Convergence for the Hilbert metric As initially explained by [13], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert projective metric on $\mathbb{R}_{+,*}^n$ (positive vectors), defined as

$$\forall (u, u') \in (\mathbb{R}_{+,*}^n)^2, \quad d_\mathcal{H}(u, u') := \|\log(u) - \log(v)\|_V$$

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 $u \sim u'$ means that $\exists s > 0, u = su'$ (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 [5] and [24] to provide quantitative proof of the Perron-Frobenius theorem (convergence of iterations of positive matrices). Sinkhorn should be thought as a non-linear generalization of Perron-Frobenius.

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

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

Note that this results extends to arbitrary convex cones and affine mapping from the cone to its interior, with the

The following theorem proved by [13], makes use of this Theorem 3 to show the linear convergence of Sinkhorn's iterations.

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

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

One also has

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

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

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

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

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

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

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

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

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

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

$$= d_{\mathcal{H}}\left(a, u^{(\ell)} \odot (Kv^{(\ell)})\right) + \lambda(K)d_{\mathcal{H}}(u^{(\ell)}, u^\star),$$

which gives the first part of (41) since $u^{(\ell)} \odot (Kv^{(\ell)}) = P^{(\ell)}\mathbb{1}_m$ (the second one being similar). The proof of (42) follows from [13, Lemma 3] $\qquad\square$

The bound (41) shows that some error measures on the marginal constraints violation, for instance, $\|P^{(\ell)}\mathbb{1}_m - a\|_1$ and $\|{P^{(\ell)}}^\top \mathbb{1}_n - b\|_1$, are useful stopping criteria to monitor the convergence. This theorem shows that the Sinkhorn algorithm converges linearly, but the rates become exponentially bad as $\varepsilon \to 0$, since it scales like $e^{-\|c\|_\infty/\varepsilon}$. In practice, one eventually observes a linear rate after enough iteration, because the local linear rate is much better, usually of the order $1 - \varepsilon$. Note also that while we wrote the proof for discrete measures, they carry over without modification to arbitrary measures. But an important limitation of this analysis is that it is restricted to compactly supported measures, since the cost needs to be uniformly bounded (for instance, analyzing Gaussian distribution requires a different approach).

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*

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

where the set of admissible potentials is

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

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

$$\min_{\mathrm{P}\geqslant 0} \ \max_{(\mathrm{f,g})\in \mathbb{R}^n \times \mathbb{R}^m} \ \langle \mathrm{C, P}\rangle + \langle \mathrm{a - P\mathbb{1}}_m, \mathrm{f}\rangle + \langle \mathrm{b - P^\top \mathbb{1}}_n, \mathrm{g}\rangle. \tag{45}$$

For a linear program, if the primal set of constraints 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

$$\max_{(\mathrm{f,g})\in \mathbb{R}^n \times \mathbb{R}^m} \ \langle \mathrm{a, f}\rangle + \langle \mathrm{b, g}\rangle + \min_{\mathrm{P}\geqslant 0} \ \langle \mathrm{C - f\mathbb{1}}_m^\top - \mathbb{1}_n g^\top, \mathrm{P}\rangle.$$

We conclude by remarking that

$$\min_{\mathrm{P}\geqslant 0} \langle \mathrm{Q, P}\rangle = \left\{ \begin{array}{ll} 0 & \text{if} \quad \mathrm{Q}\geqslant 0 \\ -\infty & \text{otherwise} \end{array} \right.$$

so that the constraint reads $\mathrm{C - f\mathbb{1}}_m^\top - \mathbb{1}_n g^\top = \mathrm{C - f}\oplus \mathrm{g} \geqslant 0$. $\qquad\square$

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

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

The formulation (43) shows that $(\mathrm{a,b})\mapsto \mathrm{L_C(a,b)}$ is a convex function (as a supremum of linear functions). From the primal problem (18), one also sees that $\mathrm{C}\mapsto \mathrm{L_C(a,b)}$ 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 $\langle f, \alpha\rangle := \int f\mathrm{d}\alpha$.

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

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

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.* $(f_i, g_j) = (f(x_i), g(y_j))$. The primal-dual optimality conditions allow for tracking the support of the optimal plan, and (46) is generalized as

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

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 enables to replacement of 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

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

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.

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

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) 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 space.

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 the existence of solutions to the dual problem. Indeed, since one can impose this Lipschitz on the dual problems, the constraint set is compact via the 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

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

$\square$

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

$$h^*(y) := \sup_x \langle x, y \rangle - h(y).$$

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 8 (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 relationship, 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 has $\varphi(x) + \varphi^*(y) \leqslant \langle x, y \rangle$ 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 envelope of f (the largest convex function below f). These properties carry 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 $f^{c\bar{c}} := (f^c)^{\bar{c}}$:*

(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 $f^c \leqslant f^{c\bar{c}c}$. $\qquad\square$

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

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

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 smoothing, which is the classical method of augmented Lagrangian, and that we will develop here using entropic regularization, which 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

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

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,

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

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,

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

The Laguerre cells associated to the dual weights g

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

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 on the right-hand side vanish (because most of the Laguerre cells associated with $\mathrm{g}+\varepsilon\delta_j$ 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 to solve the dual semi-discrete problem, one needs to select the weights g 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 [2]. 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 [10] 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,

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

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 [14]. 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

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

The step size τ_ℓ should decay fast enough to zero 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

$$\tau_\ell := \frac{\tau_0}{1 + \ell/\ell_0}, \tag{58}$$

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.

31

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

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

This problem is at the heart of the computation of efficient vector quantizers 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

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

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 the Lloyd algorithm, iteratively applies this fixed point. It is not guaranteed to converge (it could in theory cycle) but in practice, it always converges 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

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

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 $x, y \in \mathcal{X}$,

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

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

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

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 c-concave (i.e. it is the c-transform of a function). $\qquad \square$

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 [20].

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 $z_1 \leqslant z_2 \leqslant \dots$. 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 $\xi := \alpha - \beta$, 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 $s : \mathbb{R}^d \to \mathbb{R}^d$

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

$$= \inf_{s(x) \in \mathbb{R}^d} \int \|s(x)\| \mathrm{d}x + \sup_f \int f(x)(\mathrm{d}\xi - \mathrm{div}(s)\mathrm{d}x)$$

one obtains an optimization problem under a 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 [3]. 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 problem.

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). Similarly, 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 graphs (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

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

These dual norms are often called "integral probability metrics"; see [28].

Example 1 (Total variation). The total variation norm (Example 5) is a dual norm associated with 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 [27].

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,

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

34

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 $|\mathbf{f}_k| \leqslant 1$. The flat norm is sometimes called the "Kantorovich–Rubinstein" norm [17] and has been used as a fidelity term for inverse problems in imaging [21]. 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 obtains 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 $|\langle \alpha_{n_k}, f_{n_k} - f \rangle| \leqslant \|f_{n_k} - f\|_\infty$). $\qquad \square$

Corollary 3. *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 place 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 it is also the topology of convergence in law. $\qquad \square$

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 [26] 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 $\mathcal{X} = \mathbb{R}^d$. Another type of kernels are energy distances, which are more global (scale free) and are studied in Section 8.3.

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

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

These norms are often referred to as "maximum mean discrepancy" (MMD) (see [16]) and have also been called "kernel norms" in shape analysis [15]. 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 [4, 18, 26] for more details on RKHS functional spaces.

Remark 9 (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

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

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

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

7 Divergences and Dual Norms

7.1 φ-divergences

We now consider a radically different class of methods to compare distributions, which are simpler to compute ($O(n)$ for discrete distributions) but never metrize the weak* convergence. Note that yet another way is possible, using Bregman divergence, which might metrize the weak* convergence in the case where the associated entropy function is weak* regular.

Definition 5 (Entropy function). *A function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is an entropy function if it is lower semicontinuous, convex, $\mathrm{dom}\, \varphi \subset [0, \infty[$, and satisfies the following feasibility condition: $\mathrm{dom}\, \varphi \cap]0, \infty[\neq \emptyset$. The speed of growth of φ at ∞ is described by*

$$\varphi'_\infty = \lim_{x \to +\infty} \varphi(x)/x \in \mathbb{R} \cup \{\infty\}.$$

If $\varphi'_\infty = \infty$, then φ grows faster than any linear function and φ is said *superlinear*. Any entropy function φ induces a φ-divergence (also known as Ciszár divergence [11, 1] or f-divergence) as follows.

Definition 6 (φ-Divergences)**.** *Let φ be an entropy function. For $\alpha, \beta \in \mathcal{M}(\mathcal{X})$, let $\frac{d\alpha}{d\beta}\beta + \alpha^\perp$ be the Lebesgue decomposition*[1] *of α with respect to β. The divergence $\mathcal{D}_\varphi$ is defined by*

$$\mathcal{D}_\varphi(\alpha|\beta) := \int_{\mathcal{X}} \varphi\left(\frac{d\alpha}{d\beta}\right) d\beta + \varphi'_\infty \alpha^\perp(\mathcal{X}) \tag{68}$$

if α, β are nonnegative and ∞ otherwise.

The additional term $\varphi'_\infty \alpha^\perp(\mathcal{X})$ in (68) is important to ensure that $\mathcal{D}_\varphi$ defines a continuous functional (for the weak topology of measures) even if φ has a linear growth at infinity, as this is, for instance, the case for the absolute value (72) defining the TV norm. If φ as a superlinear growth, e.g.the usual entropy (71), then $\varphi'_\infty = +\infty$ so that $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if α does not have a density with respect to β.

In the discrete setting, assuming

$$\alpha = \sum_i a_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_i b_i \delta_{x_i} \tag{69}$$

are supported on the same set of n points $(x_i)_{i=1}^n \subset \mathcal{X}$, (68) defines a divergence on Σ_n

$$D_\varphi(a|b) = \sum_{i \in \text{Supp}(b)} \varphi\left(\frac{a_i}{b_i}\right) b_i + \varphi'_\infty \sum_{i \notin \text{Supp}(b)} a_i, \tag{70}$$

where $\text{Supp}(b) := \{i \in [\![n]\!] \; ; \; b_i \neq 0\}$.

Proposition 1. *If φ is an entropy function, then $\mathcal{D}_\varphi$ is jointly 1-homogeneous, convex and weakly* lower semicontinuous in (α, β).*

Proof. One defines the associated perspective function

$$\forall (u, v) \in (\mathbb{R}_+)^2, \quad \psi(u, v) = \begin{cases} \varphi(u/v)v & \text{if} \quad v \neq 0, \\ u\varphi_\infty & \text{if} \quad v = 0 \end{cases}$$

so that (we only do the proof in discrete for simplicity)

$$D_\varphi(a|b) = \sum_i \psi(a_i, b_j),$$

and we will show that it is convex on $(\mathbb{R}_+)^2$. We will show this convexity on $\mathbb{R}_+ \times \mathbb{R}_+^*$ and this convexity extends to $(\mathbb{R}_+)^2$ by taking the limit $v \to 0$. Indeed, for any $\lambda \in [0, 1]$, $\tau = 1 - \lambda$

$$\varphi\left(\frac{\tau u_1 + \lambda u_2}{\tau v_1 + \lambda v_2}\right)(\tau v_1 + \lambda v_2) = \varphi\left(\frac{\tau u_1}{\tau v_1 + \lambda v_2}\frac{u_1}{v_1} + \frac{\lambda u_2}{\tau v_1 + \lambda v_2}\frac{u_2}{v_2}\right)(\tau v_1 + \lambda v_2) \leqslant \tau\varphi\left(\frac{u_1}{v_1}\right)v_1 + \lambda\varphi\left(\frac{u_2}{v_2}\right)v_2.$$

$$\square$$

The following proposition shows that $\mathcal{D}_\varphi$ is a "distance-like" function.

Proposition 2. *For probability distribution $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X})$, then $\mathcal{D}_\varphi(\alpha, \beta) \geqslant 0$. Assuming φ to be strictly convex and $\varphi(1) = 0$, then one has $\mathcal{D}_\varphi(\alpha, \beta) = 0$ if and only if $\alpha = \beta$. This property extends to arbitrary distribution $(\alpha, \beta) \in \mathcal{M}_+(\mathcal{X})$ if one furthermore imposes that $\varphi \geqslant 0$.*

Proof. For probability measure, this follows from Jensen inequality, since

$$\int \varphi\left(\frac{d\alpha}{d\beta}\right) d\beta \geqslant \varphi\left(\int \frac{d\alpha}{d\beta} d\beta\right) = \varphi(1) = 0$$

and the case of equality for a strictly convex function is only when $\frac{d\alpha}{d\beta}$ is constant (and thus equal to 1). In the general case, if $\varphi \geqslant 0$ then the divergence is positive by construction. $\square$

---

[1]The Lebesgue decomposition theorem asserts that, given β, α admits a unique decomposition as the sum of two measures $\alpha^s + \alpha^\perp$ such that α^s is absolutely continuous with respect to β and $\alpha^\perp$ and β are singular.

Example 4 (Kullback–Leibler divergence). The Kullback–Leibler divergence $\mathrm{KL} := \mathcal{D}_{\varphi_{\mathrm{KL}}}$, also known as the relative entropy, was already introduced in (37) and (31). It is the divergence associated to the Shannon–Boltzman entropy function φ_{KL}, given by

$$\varphi_{\mathrm{KL}}(s) = \begin{cases} s\log(s) - s + 1 & \text{for } s > 0, \\ 1 & \text{for } s = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{71}$$

Example 5 (Total variation). The total variation distance $\mathrm{TV} := \mathcal{D}_{\varphi_{\mathrm{TV}}}$ is the divergence associated to

$$\varphi_{\mathrm{TV}}(s) = \begin{cases} |s - 1| & \text{for } s \geqslant 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{72}$$

It actually defines a norm on the full space of measure $\mathcal{M}(\mathcal{X})$ where

$$\mathrm{TV}(\alpha|\beta) = \|\alpha - \beta\|_{\mathrm{TV}}, \quad \text{where} \quad \|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}|\alpha|(x). \tag{73}$$

If α has a density ρ_α on $\mathcal{X} = \mathbb{R}^d$, then the TV norm is the L^1 norm on functions, $\|\alpha\|_{\mathrm{TV}} = \int_{\mathcal{X}} |\rho_\alpha(x)|\mathrm{d}x = \|\rho_\alpha\|_{L^1}$. If α is discrete as in (69), then the TV norm is the ℓ^1 norm of vectors in $\mathbb{R}^n$, $\|\alpha\|_{\mathrm{TV}} = \sum_i |\mathrm{a}_i| = \|\mathrm{a}\|_{\ell^1}$.

Remark 10 (Strong vs. weak topology). The total variation norm (73) defines the so-called "strong" topology on the space of measure. On a compact domain $\mathcal{X}$ of radius R, one has

$$\mathcal{W}_1(\alpha, \beta) \leqslant R\|\alpha - \beta\|_{\mathrm{TV}}$$

so that this strong notion of convergence implies the weak convergence metrized by Wasserstein distances. The converse is, however, not true, since δ_x does not converge strongly to δ_y if $x \to y$ (note that $\|\delta_x - \delta_y\|_{\mathrm{TV}} = 2$ if $x \neq y$). A chief advantage is that $\mathcal{M}_+^1(\mathcal{X})$ (once again on a compact ground space $\mathcal{X}$) is compact for the weak topology so that from any sequence of probability measures $(\alpha_k)_k$, one can always extract a converging subsequence, which makes it a suitable space for several optimization problems.

Proposition 3 (Dual expression). *A φ-divergence can be expressed using the Legendre transform*

$$\varphi^{*,\geqslant 0}(s) := \sup_{t \in \mathbb{R}^+} st - \varphi(t)$$

(notice that we restrict the function to the positive real) of φ as

$$\mathcal{D}_\varphi(\alpha|\beta) = \sup_{f:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) - \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x). \tag{74}$$

which equivalently reads that the Legendre transform of $\mathcal{D}_\varphi(\cdot|\beta)$ reads

$$\forall f \in \mathcal{C}(\mathcal{X}), \quad \mathcal{D}_\varphi^*(f|\beta) = \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x). \tag{75}$$

Proof. For simplicity, we consider super-linear entropy so that $\varphi_\infty' = +\infty$, and thus this impose $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if α does not has a density $\rho \geqslant 0$ with respect to β, $\mathrm{d}\alpha = \rho\mathrm{d}\beta$. Thus the Legendre-Fenchel transform of $\mathcal{D}_\varphi(\cdot|\beta)$ reads

$$\mathcal{D}_\varphi^*(f|\beta) = \sup_{\rho \geqslant 0} \int_{\mathcal{X}} f(x)\rho(x)\mathrm{d}\beta(x) - \int_{\mathcal{X}} \varphi(\rho(x))\mathrm{d}\beta(x)$$

$$= \int_{\mathcal{X}} \sup_{\rho(x) \geqslant 0} (f(x)\rho(x) - \varphi(\rho(x)))\, \mathrm{d}\beta(x) = \int_{\mathcal{X}} \varphi^{*,\geqslant 0}(f(x))\mathrm{d}\beta(x).$$

The proposed formula intuitively corresponds to using the idempotence property $\mathcal{D}_\varphi^{**}(\cdot|\beta) = \mathcal{D}_\varphi(\cdot|\beta)$. $\qquad \square$

7.2 GANs via Duality

The goal is to fit a generative parametric model $\alpha_\theta = g_{\theta,\sharp}\zeta$ to empirical data $\beta = \frac{1}{m}\sum_j \delta_{y_j}$, where $\zeta \in \mathcal{M}_+^1(\mathcal{Z})$ is a fixed density over the latent space and $g_\theta : \mathcal{Z} \to \mathcal{X}$ is the generator, often a neural network. We consider first a dual norm (63) minimization, in which case one aims at solving a min-max saddle point problem

$$\min_\theta \|\alpha_\theta - \beta\|_B = \min_\theta \sup_{f\in B} \int_{\mathcal{X}} f(x)\mathrm{d}(\alpha_\theta - \beta)(x) = \min_\theta \sup_{f\in B} \int_{\mathcal{Z}} f(g_\theta(z))\mathrm{d}\zeta - \frac{1}{m}\sum_j f(y_j).$$

Instead of a dual norm, one can consider any convex function and represent is as a maximization, for instance, a φ-divergence, which, thanks to the dual formula (74), leads to

$$\min_\theta \mathcal{D}_\varphi(\alpha_\theta|\beta) = \min_\theta \sup_f \int_{\mathcal{X}} f(x)\mathrm{d}\alpha_\theta(x) - \mathcal{D}_\varphi^*(f|\beta) = \min_\theta \sup_f \int_{\mathcal{X}} f(g_\theta(z))\mathrm{d}\zeta(z) - \frac{1}{m}\sum_j \varphi^*(f(y_j)).$$

The GAN's idea corresponds to replacing f by a parameterized network $f = f_\xi$ and doing the maximization over the parameter ξ. For instance, Wasserstein GAN consider weight clipping by constraining $\|\xi\|_\infty \leqslant 1$ in order to ensure $f_\xi \in B = \{f \ ; \ \mathrm{Lip}(f) \leqslant 1\}$. This set of networks is both in practice smaller and non-convex so no theoretical analysis of this method currently exists.

8 Sinkhorn Divergences

8.1 Dual of Sinkhorn

Discrete dual. The following proposition details the dual problem associated to (26).

Proposition 24. *One has*

$$\mathrm{L}_{\mathrm{C}}^\varepsilon(\mathrm{a},\mathrm{b}) = \max_{\mathrm{f}\in\mathbb{R}^n,\mathrm{g}\in\mathbb{R}^m} \langle \mathrm{f}, \mathrm{a}\rangle + \langle \mathrm{g}, \mathrm{b}\rangle - \varepsilon\sum_{i,j}\exp\left(\frac{\mathrm{f}_i + \mathrm{g}_j - \mathrm{C}_{i,j}}{\varepsilon}\right)\mathrm{a}_i\mathrm{b}_j + \varepsilon. \tag{76}$$

The optimal (f,g) *are linked to scalings* (u,v) *appearing in* (27) *through*

$$(\mathrm{u},\mathrm{v}) = (\mathrm{a}_i e^{\mathrm{f}/\varepsilon}, \mathrm{b}_j e^{\mathrm{g}/\varepsilon}). \tag{77}$$

Proof. We introduce Lagrange multipliers and consider

$$\min_{\mathrm{P}\geqslant 0} \max_{\mathrm{f},\mathrm{g}} \langle \mathrm{C}, \mathrm{P}\rangle + \varepsilon\mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) + \langle \mathrm{a} - \mathrm{P}\mathbb{1}, \mathrm{f}\rangle + \langle \mathrm{b} - \mathrm{P}^\top\mathbb{1}, \mathrm{g}\rangle.$$

One can check that strong duality holds since the function is continuous and that one can exchange the min with the max to get

$$\max_{\mathrm{f},\mathrm{g}} \langle \mathrm{f}, \mathrm{a}\rangle + \langle \mathrm{g}, \mathrm{b}\rangle - \varepsilon\min_{\mathrm{P}\geqslant 0} \langle \frac{\mathrm{f}\oplus\mathrm{g}-\mathrm{C}}{\varepsilon}, \mathrm{P}\rangle - \mathrm{KL}(\mathrm{P}|\mathrm{a}\otimes\mathrm{b}) = \langle \mathrm{f}, \mathrm{a}\rangle + \langle \mathrm{g}, \mathrm{b}\rangle - \varepsilon\mathrm{KL}^*\left(\frac{\mathrm{f}\oplus\mathrm{g}-\mathrm{C}}{\varepsilon}|\mathrm{a}\otimes\mathrm{b}\right).$$

One concludes by using (75) for $\varphi(r) = r\log(r) - r + 1$

$$\mathrm{KL}^*(H|\mathrm{a}\otimes\mathrm{b}) = \sum_{i,j}\varphi^*(e^{H_{i,j}})\mathrm{a}_i\mathrm{b}_j.$$

and we have after simple computation that $\varphi^*(s) = e^s - 1$. $\qquad\square$

Discrete soft c-transforms. Since the dual problem (76) is smooth, one can consider an alternating minimization. For a fixed g, one can minimize with respect to f, which leads to the following equation to be solved when zeroing the derivative with respect to f

$$\mathrm{a}_i - e^{\frac{\mathrm{f}_i}{\varepsilon}} \mathrm{a}_i \sum_j \exp\left(\frac{\mathrm{g}_j - \mathrm{C}_{i,j}}{\varepsilon}\right) \mathrm{b}_j = 0$$

which leads to the explicit solution

$$\mathrm{f}_i = -\varepsilon \log \sum_j \exp\left(\frac{\mathrm{g}_j - \mathrm{C}_{i,j}}{\varepsilon}\right) \mathrm{b}_j.$$

We conveniently introduce the soft-min operator of some vector $h \in \mathbb{R}^m$

$$\min_{\mathrm{b}}^\varepsilon(h) := -\varepsilon \log \sum_j e^{-h_j/\varepsilon} \mathrm{b}_j$$

which is a smooth approximation of the minimum operator, and the optimal f for a fixed g is computed by a soft version of the c-transform

$$\mathrm{f}_i = \min_{\mathrm{b}}^\varepsilon(\mathrm{C}_{i,\cdot} - \mathrm{g}). \tag{78}$$

In a similar way, the optimal g for a fixed f is

$$\mathrm{g}_j = \min_{\mathrm{a}}^\varepsilon(\mathrm{C}_{\cdot,j} - \mathrm{f}). \tag{79}$$

Exponentiating these iterations, one retrieves exactly the Sinkhorn algorithm. These iterations are however unstable for small ε. To be able to apply the algorithm in this regime, one needs to stabilize it using the celebrated log-sum-exp trick. It follows from noticing that similarly to the minimum operator, one has

$$\min_{\mathrm{b}}^\varepsilon(h - \mathrm{cst}) = \min_{\mathrm{b}}^\varepsilon(h) - \mathrm{cst}$$

and to replace the computation of $\min_{\mathrm{b}}^\varepsilon(h)$ by its stabilized version (equal when using infinite precision computation) $\min_{\mathrm{b}}^\varepsilon(h - \min(h)) + \min(h)$.

Continuous dual and soft-transforms. For generic (non-necessarily discrete) input measures (α, β), the dual problem (76) reads

$$\sup_{f,g \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(x) \mathrm{d}\beta(x) - \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} \left(e^{\frac{f \oplus g - c}{\varepsilon}} - 1\right) \mathrm{d}\alpha \otimes \mathrm{d}\beta \tag{80}$$

This corresponds to a smoothing of the constraint $\mathcal{R}(c)$ appearing in the original problem (47), which is retrieved in the limit $\varepsilon \to 0$.

The corresponding soft c-transform, which minimize this dual problem with respect to either f or g reads

$$\forall\, y \in \mathcal{Y}, \quad f^{c,\varepsilon}(y) := -\varepsilon \log\left(\int_{\mathcal{X}} e^{\frac{-c(x,y) + f(x)}{\varepsilon}} \mathrm{d}\alpha(x)\right),$$

$$\forall\, x \in \mathcal{X}, \quad g^{\bar{c},\varepsilon}(x) := -\varepsilon \log\left(\int_{\mathcal{Y}} e^{\frac{-c(x,y) + g(y)}{\varepsilon}} \mathrm{d}\beta(y)\right).$$

In the case of discrete measures, on retrieves the formula (78) and (79).

We omit the details, but similarly to the unregularized case, one can define an entropic semi-discrete problem and develop a stochastic optimization method to solve it.

The soft c-transform, besides being useful to define Sinkhorn's algorithm, is also important to show the existence of dual solutions.

Proposition 25. *The dual problem* (80) *has solutions, and the set of solutions is of the form* $(f^\star + \lambda, g^\star - \lambda)$ *for* $\lambda \in \mathbb{R}$.

Proof. Several proof strategies show the existence of a solution, including (i) using the contractance of Sinkhorn for the Hilbert metric (see below), (ii) using the Lipschitz smoothness of the c-transforms which implies one can consider a compact set of continuous functions, (iii) using the fact that $f \mapsto f^{c,\varepsilon}$ is a monotone map. Uniqueness follows from strict convexity of $H \mapsto \int e^{\frac{H}{\varepsilon}} \mathrm{d}a \otimes \beta$ on the space of functions defined up to an additive constant, and that the kernel of the linear map $(f,g) \mapsto H = f \oplus g - c$ are constant functions of the form $f(x) = -g(x) = \lambda$. $\qquad\square$

Remark 11 (Convexity properties). The $\log \int \exp$ operator behaves similarly to the max operator, and in particular it preserves convexity. Similarly to the unregularized case studied in Remark 8, in the particular case $c(x,y) = -\langle x, y \rangle$, one thus has that $f^{c,\varepsilon}$ is always a concave function, so that when $c(x,y) = \|x-y\|^2/2$, one has that the optimal potentials are of the form $f^\star(x) = \|x\|^2 - \varphi^\star(x)$ where $\varphi^\star$ is a convex function.

Remark 12 (Gaussian marginals). Another remarkable property of these soft c-transforms in the case $c(x,y) = \|x-y\|^2$ and (α, β) are Gaussians is that it preserves quadratic functions since the product and convolution of Gaussian functions are Gaussians. This implies that for Gaussian marginal, the optimal potentials $(f^\star, g^\star)$ are quadratic, and that the optimal entropic-regularized coupling $\pi^\star$ is a Gaussian.

8.2 Convergence Analysis

This section revisits the convergence analysis of Section 4.5

Bregman sub-linear convergence. One can show the following KL conservation during the iterations.

Proposition 4. *Denoting* (f_k, g_k) *the dual variable generated by Sinkhorn's algorithm and* $\pi_k = \exp(\frac{c - f \oplus g}{\varepsilon})\alpha \otimes \beta$ *the associated primal variable. The dual cost is* $D(f,g)$ *while the primal one is* $\mathrm{KL}(\pi_{k+1}|\xi)$ *where* $\xi := e^{-c/\varepsilon}\alpha \otimes \beta$ *is the Gibbs kernel. One has*

$$D(f_{k+1}, g_{k+1}) - D(f_k, g_k) = \varepsilon\,\mathrm{KL}(\pi^\star|\pi_k) - \varepsilon\,\mathrm{KL}(\pi^\star|\pi_{k+1}) = \varepsilon\,\mathrm{KL}(\alpha|\pi_{k,1}) + \varepsilon\,\mathrm{KL}(\beta|\pi_{k,2}) > 0.$$

Proof. TODO. $\qquad\square$

One sees that for even index k, $\mathrm{KL}(\beta|\pi_{k,2}) = 0$ while for even k, $\mathrm{KL}(\alpha|\pi_{k,1}) = 0$ (Sinkhorn corresponds to alternating projection on the constraints).

This lemma shows the strict decay of the dual and primal energies. It implies that (assuming $D(f_0, g_0) = 0$, using $f_0 = g_0 = 0$)

$$\varepsilon \sum_k \mathrm{KL}(\alpha|\pi_{k,1}) + \mathrm{KL}(\beta|\pi_{k,2}) = D(f^\star, g^\star)$$

so that necessarily $\mathrm{KL}(\alpha|\pi_{k,1})$ and $\mathrm{KL}(\beta|\pi_{k,2})$ are converging to zero. One cannot deduce from this a rate on the last iterate (since it is not clear $\mathrm{KL}(\alpha|\pi_1^k)$ is decaying), but at least

$$\min_{\ell \leqslant k} \mathrm{KL}(\alpha|\pi_{k,1}) \leqslant \frac{D(f^\star, g^\star)}{\varepsilon k} = \frac{\mathrm{KL}(\pi^\star|\xi)}{k}.$$

where $\xi := \alpha \otimes \beta e^{-\frac{c}{\varepsilon}}$ is the Gibbs kernel.

Hilbert-metric linear convergence. The Hilbert metric convergence result is equivalent to convergence on the dual potentials (f,g) according to the variational norms. For bounded cost c (e.g. on compact spaces),

$$\|f_k - f^\star\|_V = O(\lambda^k) \quad \text{and} \quad \|g_k - f^\star\|_V = O(\lambda^k)$$

$$\|\log \frac{\mathrm{d}\pi_k}{\mathrm{d}\pi^\star}\|_\infty = \|(f_k - f^\star) \oplus (g_k - g^\star)\|_\infty \leqslant \|f_k - f^\star\|_V + \|g_k - g^\star\|_V$$

where the contraction ratio only depends on the radius $R := \sup_{x,y} c(x,y)$ of the cost, $\lambda = \frac{\sqrt{\bar{\eta}}-1}{\sqrt{\bar{\eta}}+1} \leqslant \tanh(R/(2\varepsilon)) < 1$. One also has the following bounds

$$\|f_k - f^\star\|_V \leqslant \frac{\|\log \frac{\mathrm{d}\pi_{k,1}}{\mathrm{d}\alpha}\|_\infty}{1-\lambda}$$

which can be used to provide a posterior estimate of the rate of convergence and serves as a stopping criterion.

8.3 Sinkhorn Divergences

Entropic bias. A major issue of the value of Sinkhorn problem (36) is that $\mathcal{L}_c^\varepsilon(\alpha, \beta) > 0$. So in particular,

$$\alpha_\varepsilon = \underset{\beta}{\operatorname{argmin}} \; \mathcal{L}_c^\varepsilon(\alpha, \beta)$$

does not satisfy $\alpha_\varepsilon = \alpha$ unless $\varepsilon = 0$. The following proposition shows that the bias induced by this entropic regularization has a catastrophic influence in the large ε limit.

Proposition 26. *One has $\mathcal{L}_c^\varepsilon(\alpha, \beta) \to \int c\mathrm{d}\alpha \otimes \beta$ as $\varepsilon \to +\infty$.*

Proof. The intuition of the proof follows from the fact that the optimal coupling converges to $\alpha \otimes \beta$. □

So in the large ε limit, $\mathcal{L}_c^\varepsilon$ behaves like an inner product and not like a norm. For instance, in the case

$$\alpha_\varepsilon \to \underset{\beta}{\min} \; \left\langle \int c(x, \cdot)\mathrm{d}\alpha(x), \; \beta \right\rangle = \delta_{y^\star(\alpha)} \quad \text{where} \quad y^\star(\alpha) = \underset{y}{\operatorname{argmin}} \int c(x, y)\mathrm{d}\alpha(x).$$

For instance, when $c(x,y) = \|x-y\|^2$ then α_ε collapses towards a Dirac located at the mean $\int x\mathrm{d}\alpha(x)$ of α.

Sinkhorn divergences. The usual way to go from an inner product to a norm is to use the polarization formula, we thus also consider the Sinkhorn cost, to define the debiased Sinkhorn divergence

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) := \mathcal{L}_c^\varepsilon(\alpha, \beta) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha, \beta).$$

It is not (yet) at all clear why this quantity should be positive.

Before going on, we prove a fundamental lemma which states that the dual cost has a simple form where the regularization vanishes at a solution (and it vanishes also during Sinkhorn's iteration by the same proof).

Lemma 3. *Denoting $(f_{\alpha,\beta}, g_{\alpha,\beta})$ optimal dual potentials (which can be shown to be unique up to an additive constant), one has*

$$\mathcal{L}_c^\varepsilon(\alpha, \beta) = \langle f_{\alpha,\beta}, \; \alpha \rangle + \langle g_{\alpha,\beta}, \; \beta \rangle. \tag{81}$$

Proof. We first notice that at optimality, the relation

$$f_{\alpha,\beta} = -\varepsilon \log \int_{\mathcal{Y}} e^{\frac{g_{\alpha,\beta}(y) - c(x,y)}{\varepsilon}} \mathrm{d}\beta(y)$$

after taking the exponential, equivalently reads

$$1 = \int_{\mathcal{Y}} e^{\frac{f_{\alpha,\beta}(x) + g_{\alpha,\beta}(y) - c(x,y)}{\varepsilon}} \mathrm{d}\beta(y) \quad \Longrightarrow \quad \int_{\mathcal{X}\times\mathcal{Y}} \left(e^{\frac{f_{\alpha,\beta} \oplus g_{\alpha,\beta} - c}{\varepsilon}} - 1 \right) \mathrm{d}\alpha \otimes \beta = 0.$$

Plugging this in formula (80), one obtains the result. □

Let us first show that its asymptotic makes sense.

Proposition 27. *One has $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \to \mathcal{L}_c(\alpha, \beta)$ when $\varepsilon \to 0$ and*

$$\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \to \frac{1}{2}\int -c\,\mathrm{d}(\alpha - \beta) \otimes \mathrm{d}(\alpha - \beta) \quad when \quad \varepsilon \to +\infty.$$

Proof. For discrete measures, the convergence is already proved in Proposition (34), we now give a general treatment. **Case $\varepsilon \to 0$.** **Case $\varepsilon \to +\infty$.** We denote $(f_\varepsilon, g_\varepsilon)$ optimal dual potential. Optimality condition on f_ε (equivalently Sinkhorn fixed point on f_ε) reads

$$f_\varepsilon = -\varepsilon \log \int \exp\left(\frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon}\right)\mathrm{d}\beta(y) = -\varepsilon \log \int \left(1 + \frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon} + o(1/\varepsilon)\right)\mathrm{d}\beta(y)$$

$$= -\varepsilon \int \left(\frac{g_\varepsilon(y) - c(\cdot, y)}{\varepsilon} + o(1/\varepsilon)\right)\mathrm{d}\beta(y) = -\int g_\varepsilon\,\mathrm{d}\beta + \int c(\cdot, y)\mathrm{d}\beta(y) + o(1).$$

Plugging this relation in the dual expression (81)

$$\mathcal{L}_c^\varepsilon(\alpha, \beta) = \int f_\varepsilon\,\mathrm{d}\alpha + \int g_\varepsilon\,\mathrm{d}\beta = -\iint c(x, y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y) + o(1).$$

$\square$

In the case where $-c$ defines a conditionally positive definite kernel, then $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta)$ converges to the square of a Hilbertian kernel norm. A typical example is when $c(x, y) = \|x - y\|^p$ for $0 < p < 2$, which corresponds to the so-called Energy distance kernel. This kernel norm is the dual of a homogeneous Sobolev norm.

We now show that this debiased Sinkhorn divergence is positive.

Proposition 28. *If $k(x, y) = e^{-c(x,y)/\varepsilon}$ is positive definite, then $\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 0$ and is zero if and only if $\alpha = \beta$.*

Proof. In the following, we denote $(f_{\alpha,\beta}, g_{\alpha,\beta})$ optimal dual potential for the dual Schrodinger problem between α and β. We denote $f_{\alpha,\alpha} = g_{\alpha,\alpha}$ (one can assume they are equal by symmetry) the solution for the problem between α and itself. Using the suboptimal function $(f_{\alpha,\alpha}, g_{\beta,\beta})$ in the dual maximization problem, and using relation (81) for the simplified expression of the dual cost, one obtains

$$\mathcal{L}_c^\varepsilon(\alpha, \beta) \geqslant \langle f_{\alpha,\alpha}, \alpha\rangle + \langle g_{\beta,\beta}, \beta\rangle - \varepsilon \langle e^{\frac{f_{\alpha,\beta} \oplus g_{\alpha,\beta} - c}{\varepsilon}} - 1, \alpha \otimes \beta\rangle$$

But one has $\langle f_{\alpha,\alpha}, \alpha\rangle = \frac{1}{2}\mathcal{L}_c^\varepsilon(\alpha, \alpha)$ and same for β, so that the previous inequality equivalently reads

$$\frac{1}{\varepsilon}\bar{\mathcal{L}}_c^\varepsilon(\alpha, \beta) \geqslant 1 - \langle e^{\frac{f_{\alpha,\beta} \oplus g_{\alpha,\beta} - c}{\varepsilon}}, \alpha \otimes \beta\rangle = 1 - \langle \tilde\alpha, \tilde\beta\rangle_k$$

where $\tilde\alpha = e^{f_{\alpha,\alpha}}\alpha$, $\tilde\beta = e^{f_{\beta,\beta}}\beta$ and we introduced the inner product (which is a valid one because k is positive) $\langle \tilde\alpha, \tilde\beta\rangle_k := \int k(x, y)\mathrm{d}\tilde\alpha(x)\mathrm{d}\tilde\beta(y)$. One notes that the Sinkhorn fixed point equation, once exponentiated, reads $e^{f_{\alpha,\alpha}} \odot [k(\tilde\alpha)] = 1$ and hence

$$\|\tilde\alpha\|_k^2 = \langle k(\tilde\alpha), \tilde\alpha\rangle = \langle e^{f_{\alpha,\alpha}} \odot k(\tilde\alpha), \alpha\rangle = \langle 1, \alpha\rangle = 1$$

and similarly $\|\tilde\beta\|_k^2 = 1$. So by Cauchy-Schwartz, one has $1 - \langle \tilde\alpha, \tilde\beta\rangle_k \geqslant 0$. Showing strict positivity is more involved, and is not proved here. $\square$

One can furthermore show that this debiased divergence metrizes the convergence in law.

9 Wasserstein (gradient) Flows

The goal of this section is to expose the connection between optimal transport and certain evolutions over the space of probability distributions, particularly solutions to some PDEs and generative models using diffusion. The exposition is informal, focusing on intuition rather than rigorous proof. We work over the space $\mathcal{X} = \mathbb{R}^d$. It is also the opportunity to draw some connexions with recent applications in ML, most notably analyzing the training dynamic of MLP (where neurons are transported), modeling deep transformers (where tokens are transported), and flow matching for generative models (where features, such as image's pixels, are transported).

9.1 Evolutions over the Space of Measures

We consider the evolution $t \mapsto \alpha_t \in \mathcal{P}(\mathbb{R}^d)$. Such evolution can be described in a "Lagrangian" way as the advection of particles along a (time-dependent) vector field $v_t(x)$ in $\mathbb{R}^d$. At the particle level, this advection is governed by

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t)), \tag{82}$$

such that $x(0)$ is mapped to $x(t)$ by a "transport" mapping $T_t : x(0) \mapsto x(t)$. The fact that α_t is the density of advected particles implies $\alpha_t = (T_t)_\sharp \alpha_0$. For discrete measures, $\alpha_t = \frac{1}{n} \sum_{i=1}^n \delta_{x_i(t)}$, meaning each $x_i(t)$ solves (82).

In the Eulerian interpretation, over the measure itself, the ODE for particles becomes the PDE

$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(v_t \alpha_t) = 0. \tag{83}$$

Here, writing $\mathrm{div}(v_t \alpha_t)$ is an abuse of notation since divergence is strictly valid for densities. Instead, it refers to the measure defined by $\mathrm{div}\left(v_t \frac{\mathrm{d}\alpha_t}{\mathrm{d}x}\right) \mathrm{d}x$.

More rigorously, this PDE (83) should be understood in the weak sense, allowing it to be defined even for discrete measures with particles evolving according to (82). Specifically, for any smooth function $x \to \varphi(x)$ and for almost every t,

$$\partial_t \left[\int_{\mathbb{R}^d} \varphi(x) \mathrm{d}\alpha_t(x) \right] - \int_{\mathbb{R}^d} \langle v_t, \nabla_x \varphi(x) \rangle \mathrm{d}\alpha_t(x) = 0.$$

From measure evolutions to vector fields. It is important to note that for a given evolution α_t, there are infinitely many possible choices of vector fields v_t satisfying

$$\partial_t \alpha_t + \mathrm{div}(v_t \alpha_t) = 0. \tag{84}$$

This is because modifying v_t by a divergence-free field does not affect the density evolution. The linear space of vector fields that leave a measure α invariant is

$$\mathcal{H}_\alpha := \{ v : \mathrm{div}(\alpha v) = 0 \}.$$

It is non-trivial because it corresponds to the kernel of a "weighted" divergence. For instance, if α is an isotropic Gaussian, $\mathcal{H}_\alpha$ contains vector fields inducing rotations (i.e., associated with anti-symmetric matrices).

Dacorogna and Moser inversion. Consequently, infinitely many particle evolutions result in the same density. Reconstructing particle evolution from observed density evolution (sometimes called "trajectory inference") is thus an ill-posed inverse problem. A simple choice is to impose that $\alpha_t v_t$ is a gradient field, thus leading to the inversion of a Laplacian (which is possible assuming boundary conditions, for instance, vanishing at infinity)

$$v_t = -\frac{1}{\alpha_t} \nabla \Delta^{-1} (\partial_t \alpha_t). \tag{85}$$

which was initially proposed by Dacorogna and Moser. A difficulty with this choice is that it is not well defined when α_t vanishes, and also it is not a gradient field, which might be desirable in some cases. In the following, we will consider techniques where v_t is a gradient by design, it is well-defined and can be computed more efficiently without explicitly inverting a Laplacian.

Least square inversion. A more fruitful method (corresponding to both flow matching, optimal transport, and Wasserstein flow construction) is to solve the inversion from α_t to v_t in a least square manner, solving

$$\min_v \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, \mathrm{d}\alpha_t(x) \, \mathrm{d}t \quad \text{subject to} \quad \mathrm{div}(\alpha v) + \frac{\partial \alpha}{\partial t} = 0. \tag{86}$$

The optimality condition of this constraint problem imposes that the optimal v_t is a gradient field, i.e. there exists ψ_t such that $v_t = \nabla \psi_t$. This means that the vector field is obtained by formally inverting a weighted Laplacian $\Delta_\alpha(\varphi) := \mathrm{div}(\nabla \varphi \alpha)$

$$v_t = -\nabla \Delta_\alpha^{-1}(\partial_t \alpha_t). \tag{87}$$

In general is hard to perform this inversion, but we will see that under specific choices for α_t, simpler formula are available.

9.2 Generative Models via Flow Matching

Generative models aim to build a transportation map T between a reference distribution α (typically an isotropic Gaussian) and the target data distribution β. It is easy to see that such a map always exists for any β, but finding an explicit constructive method for T is surprisingly non-trivial. Optimal transport is one approach to achieving this, but it is computationally expensive and raises questions about how to estimate it from samples. A recent idea, first introduced in diffusion models and later systematically developed in flow matching by Yaron Lipman and his collaborators, is to obtain T by integrating a time-dependent vector field v_t. This vector field v_t is obtained by constructing an interpolation α_t and then finding v_t using the least square formula (87). As we will explain, for a specific class of interpolation (obtained by a parametric push-forward), this v_t can be obtained by avoiding explicitly inverting a Laplacian and instead computing a simple conditional expectation. This conditional expectation can itself be estimated by solving another least square, but this time unconstraint, making the estimation feasible from finite samples of α and β.

We assume that α_t is defined via a "projection" (in a loose sense) of a latent distribution $\pi \in \mathcal{P}(\mathbb{R}^{d'})$, using an operator $P_t : \mathbb{R}^{d'} \to \mathbb{R}^d$ where $d' \gg d$, i.e.

$$\forall t \in [0, 1], \quad \alpha_t := (P_t)_\sharp \pi. \tag{88}$$

The most usual case is when $d' = 2d$, we denote $(x, y) \in \mathbb{R}^{d'} = \mathbb{R}^d \times \mathbb{R}^d$ and asume $P_0(x, y) = x$ and $P_1(x, y) = y$ so that π is a probabilistic coupling between α_0 and α_1, i.e. π has marginals (α_0, α_1). For instance, one can use $\pi = \alpha_0 \otimes \alpha_1$, the trivial coupling. An even more special case is to assume $P_t(x, y) = (1 - t)x + ty$ is a linear interpolation. But one can use more complex constructions, as long as it is simple to sample from π.

If $\pi = \alpha \otimes \beta$ and $\alpha = \frac{1}{n} \sum_i \delta_{x_i}$, $\beta = \frac{1}{m} \sum_j \delta_{y_j}$, then α_t consists of $n \times m$ Dirac masses

$$\alpha_t = \frac{1}{nm} \sum_{i,j} \delta_{P_t(x_i, y_j)}.$$

If $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is a Brenier-type coupling, then $\alpha_t = ((1 - t)\mathrm{Id} + tT)_\sharp \alpha$ is the so-called McCann OT interpolation.

This interpolation is not directly useful for sampling from β, but it can be used to define a flow field v_t so that the Eulerian advection equation (83) holds. This flow field is computed by solving an unconstrained least square problem, or equivalently it is a conditional expectation.

45

Proposition 29. *The solution of the following flow-matching problem*

$$\min_{(v_t)_t} \int_{\mathbb{R}^{d'}} \|v_t(P_t(u)) - [\partial_t P_t](u)\|^2 \, d\pi(u). \tag{89}$$

or equivalently the conditional expectation

$$v_t(z) = \mathbb{E}_{u \sim \pi}\big([\partial_t P_t](u) \,\big|\, z = P_t(u)\big). \tag{90}$$

satisfies the continuity equation (84)

The solution of (89) is the conditional expectation of the velocities ∂P_t, intuitively, this means that $v_t(z)$ at some point z should be the average velocity of all trajectories passing through z. Numerically, $(x, t) \to v_t(x)$ can be parameterized by a neural network (e.g., a U-Net for vision tasks) and estimated using stochastic gradient descent on the objective in (89). Once v_t is estimated, integrating the ODE $\dot{x} = v_t(x)$ defines the transport map T_t, ensuring that $\alpha_t = (T_t)_\sharp \alpha_0$ produces the same interpolation as (88), though with a different particle system. Instead of a coupling, this approach uses a deterministic map. The sampling procedure consists in first drawing $X_0 \sim \alpha$, and then integrating the ODE $\dot{X}_t = v_t(X_t)$ starting with $X_{t=0} = X_0$. The resulting $X_{t=1}$ is distributed according to $\alpha_1 = \beta$.

In the special case where $P_t(x, y) = (1 - t)x + ty$ is a linear interpolation and $\pi = \alpha \otimes \beta$, then α_t is a convolution of rescaled versions of α_0 and α_1. Then the flow matching (89) boils down to

$$\min_{(v_t)_t} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|v_t((1 - t)x + ty) - (y - x)\|^2 \, d\alpha_0(x) d\alpha_1(y).$$

If furthermore, α_0 is an isotropic Gaussian, this is exactly (up to an exponential change of variable in the time variable) the diffusion model method. In this case, one can show that the obtained v_t has a closed form

$$v_t(x) = ?? + \nabla \log(\alpha_t).$$

So in particular, this v_t is a gradient field, so it is also the solution of the (constrained) least square problem (87). But note that the least square (89) is much simpler because it is un-constrained.

Proof. We now prove the flow matching formula (90), i.e. that it defines a valid flow between α_0 and α_1. First, let us recall the definition of the interpolated density α_t and the velocity field v_t. According to (88), the density α_t is defined heuristically as:

$$\alpha_t(z) = \int \delta(z - P_t(u)) \, d\pi(u)$$

and rigorously for any test function $\varphi(z)$ as:

$$\int \varphi(z) \, d\alpha_t(z) = \int \varphi(P_t(u)) \, d\pi(u). \tag{91}$$

Following (90), the velocity field v_t is defined heuristically as:

$$v_t(z) = \frac{1}{\alpha_t(z)} \int \delta(z - P_t(u))[\partial_t P_t](u) \, d\pi(u),$$

and rigorously for any vector field $m(z)$ as:

$$\int \langle m(z), v_t(z) \rangle \, d\alpha_t(z) = \int \langle m(P_t(u)), [\partial_t P_t](u) \rangle \, d\pi(u). \tag{92}$$

We aim to prove that the density α_t satisfies the continuity equation:

$$\frac{\partial \alpha_t}{\partial t} + \text{div}(\alpha_t v_t) = 0.$$

In a rigorous sense, this means showing that for any smooth test function $\varphi(z)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z) \, \mathrm{d}\alpha_t(z) - \int \langle v_t(z), \nabla\varphi(z)\rangle \, \mathrm{d}\alpha_t(z) = 0. \tag{93}$$

To prove (93), we compute both terms separately and show that they cancel out. First, consider the time derivative of $\int \varphi(z) \, \mathrm{d}\alpha_t(z)$. Using (91), we differentiate under the integral sign:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z) \, \mathrm{d}\alpha_t(z) = \int \frac{\mathrm{d}}{\mathrm{d}t} \varphi(P_t(u)) \, \mathrm{d}\pi(u).$$

Applying the chain rule to $t \to \varphi \circ P_t$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \varphi(P_t(u)) = \langle \nabla\varphi(P_t(u)), [\partial_t \nabla P_t](u)\rangle.$$

Thus:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z) \, \mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(P_t(u)), [\partial_t P_t](u)\rangle \, \mathrm{d}\pi(u). \tag{94}$$

Next, for the term involving $\mathrm{div}(\alpha_t v_t)$, we use the definition of v_t in (92) with $m(z) = \nabla\varphi(z)$. Substituting this into the expression for v_t, we get:

$$\int \langle v_t(z), \nabla\varphi(z)\rangle \, \mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(P_t(u)), [\partial_t P_t](u)\rangle \, \mathrm{d}\pi(u).$$

Comparing this result with (94), we see that:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z) \, \mathrm{d}\alpha_t(z) - \int \langle v_t(z), \nabla\varphi(z)\rangle \, \mathrm{d}\alpha_t(z) = 0.$$

$\square$

9.3 Benamou-Brenier dynamic formulation of OT

Instead of imposing access to the full dynamics $(\alpha_t)_{t=0}^1$, we assume only the knowledge of α_0 and α_1 and seek for an interpolation minimizing the least square energy (86). A fundamental result, proved by Benamou and Brenier is that the value of this "geodesic" energy is equal to the Wasserstein-2 distance,

$$\mathrm{W}_2(\alpha_0, \alpha_1)^2 = \min_{\alpha, v} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, \mathrm{d}\alpha_t(x) \, \mathrm{d}t \quad \text{subject to} \quad \mathrm{div}(\alpha v) + \frac{\partial \alpha}{\partial t} = 0. \tag{95}$$

where here implicitly, we impose $\alpha_{t=0} = \alpha_0, \alpha_{t=1} = \alpha_1$ in the above minimization. While this is not the focus of this chapter, let us note that, while the initial problem (95) is non-convex, after the change of variable $m_t := \alpha_t v_t$, it becomes convex in (m_t, α_t). This beautiful property enables solving the geodesic interpolation using convex optimization technics, once the domain is discretized.

The solution α_t of (95) can be obtained by evolving particules in straight lines along the Monge map T solving Monge's problem

$$\min_T \left\{ \int_{\mathbb{R}^d} \|T(x) - x\|^2 \mathrm{d}\alpha_0(x) \; : \; T_\sharp \alpha_0 = \alpha_1 \right\}$$

i.e. defining $T_t := (1-t)\mathrm{Id} + tT$, one has $\alpha_t = (T_t)_\sharp \alpha_0$.

9.4 Wasserstein Gradient Flows

We now consider a function $f(\alpha)$ and seek a minimizing evolution $(\alpha_t)_t$. The general strategy of minimizing movement over a metric space is to construct a discrete-time evolution using an implicit Euler scheme:

$$\alpha_{t+\tau} := \arg\min_\alpha \frac{1}{2\tau} \mathrm{W}_2(\alpha_t, \alpha)^2 + f(\alpha). \tag{96}$$

47

Euclidean gradient flows. If we restrict (96) to finite dimensions and assume $\alpha_t = \delta_{x(t)}$ and $\alpha = \delta_x$ (single Dirac measures), this matches the implicit Euler scheme:

$$x(t + \tau) := \arg\min_x \frac{1}{2\tau}\|x - x(t)\|^2 + h(x),$$

where $h(x) = f(\delta_x)$. Its solution is formally given by the implicit Euler formula:

$$x(t + \tau) = (\mathrm{Id} + \tau\nabla h)^{-1}(x(t)).$$

In contrast, the explicit Euler scheme is:

$$x(t + \tau) = (\mathrm{Id} - \tau\nabla h)(x(t)) = x(t) - \tau\nabla h(x(t)).$$

Both schemes converge as $\tau \to 0$ to:
$$\dot{x}(t) = -\nabla h(x(t)). \tag{97}$$

Wasserstein gradient formula. The implicit Euler scheme has the advantage that it does not require h or f to be smooth. For f, this is crucial to handle evolution over arbitrary measures (with or without densities) seamlessly.

As $\tau \to 0$, under certain conditions on f, (96) defines a continuous evolution $t \mapsto \alpha_t$. As discussed earlier, this evolution can be described as a Lagrangian evolution (82). A key point is that it provides an explicit vector field v_t (depending on α_t), denoted as $\nabla_{\mathrm{W}} f(\alpha)$, called the Wasserstein gradient. In the weak sense, α_t satisfies:
$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(-\nabla_{\mathrm{W}} f(\alpha_t)\alpha_t) = 0. \tag{98}$$

Here, $\nabla_{\mathrm{W}} f(\alpha_t)(x) \in \mathbb{R}^d$ is a vector field and is a gradient (similar to the computation in (86)), computable as:
$$\nabla_{\mathrm{W}} f(\alpha) = \nabla_{\mathbb{R}^d}\varphi, \quad \text{where } \varphi := \delta f(\alpha).$$

The function $\delta f(\alpha) \in \mathcal{C}(\mathbb{R}^d)$ is known as the first variation or Fréchet (directional) derivative, satisfying for any $\beta \in \mathcal{P}(\mathbb{R}^d)$:
$$f((1 - \tau)\alpha + \tau\beta) = f(\alpha + \tau\rho) = f(\alpha) + \tau\int[\delta f(\alpha)](x)\mathrm{d}\rho(x) + o(\tau),$$

where $\rho = \beta - \alpha$ is a zero-mean measure. The idea of doing gradient flow for the Wasserstein metric was first introduced by John D. Lafferty in his PhD, and published in "The Density Manifold and Configuration Space Quantization", under the name "density manifold". It was systematically studied by Felix Otto, who studied the properties of this space.

Heuristic derivation of the Wasserstein gradient formula. An heuristic way to see why (98) holds with this specific choice of vector field $\nabla_{\mathrm{W}} f(\alpha)$ is first re-write (96) as a minimization over displacement fields v so that $\alpha = (\mathrm{Id} + \tau v)_\sharp \alpha_t$, considering

$$\min_v \frac{1}{2\tau}\tau^2\|v\|_{L^2(\alpha_t)}^2 + f((\mathrm{Id} + \tau v)_\sharp \alpha_t).$$

Then we perform a first-order Taylor expansion of this formulation using

$$(\mathrm{Id} + \tau v)_\sharp \alpha_t = \alpha_t + \tau\mathrm{div}(v\alpha_t) + o(\tau)$$

$$f((\mathrm{Id} + \tau v)_\sharp \alpha_t) = f(\alpha_t) - \tau\int \delta f(\alpha_t)\mathrm{div}(v\alpha_t)\mathrm{d}dx + o(\tau)$$

$$= f(\alpha_t) + \tau\int \langle\nabla_{\mathbb{R}^d}\delta f(\alpha_t)(x), v(x)\rangle\mathrm{d}\alpha_t(x) + o(\tau)$$

to obtain the following first order expansion in τ of the problem minimized in (96)

$$\min_v f(\alpha_t) + \tau \int \left[\frac{1}{2} \|v(x)\|^2 + \langle \nabla_W f(\alpha_t)(x), v(x) \rangle \right] \mathrm{d}\alpha_t(x) + o(\tau)$$

We now detail examples of such Wasserstein gradient flows.

Discrete evolutions. If $f(\alpha)$ can be evaluated on discrete distributions and ∇_W is continuous in this case, the flow (98) maintains the number of Dirac masses, $\alpha_t = \frac{1}{n} \sum_i \delta_{x_i(t)}$. The particles $X(t) := (x_i(t))_i$ evolve according to a system of coupled ODEs:

$$\dot{X}(t) = -\nabla F(X), \tag{99}$$

where $F(X) := f\left(\frac{1}{n} \sum_i \delta_{x_i} \right)$.

Linear Functionals. The simplest example of flows is for linear functions

$$f(\alpha) = \int h(x) \mathrm{d}\alpha(x). \tag{100}$$

Here, $\delta f(\alpha) = h$ is a fixed function (independent of α). The flow (98) becomes:

$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(-\nabla h \alpha_t) = 0.$$

This implies particles move independently according to the usual gradient flow (97).

Shannon Neg-Entropy. A very different behavior is obtained by considering functions which require α_t to have a density, the canonical example being Shannon neg-entropy

$$f(\alpha) = \int \log \left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x) \right) \mathrm{d}\alpha(x). \tag{101}$$

Here, $\delta f(\alpha) = \log \left(\frac{\mathrm{d}\alpha}{\mathrm{d}x} \right)$, so $\nabla_W f(\alpha) = \frac{\nabla \alpha}{\alpha}$ (often called the score). The flow (98) becomes the heat equation:

$$\partial_t \alpha_t = \Delta(\alpha).$$

Other entropy functionals lead to nonlinear diffusion equations. For example, generalized entropy of the form (a.k.a φ-divergences with respect to Lebesgue)

$$f(\alpha) = \int g \left(\frac{\mathrm{d}\alpha}{\mathrm{d}x} \right) \mathrm{d}x, \tag{102}$$

for a 1-D function $g(s)$, leads to nonlinear diffusions:

$$\frac{\partial \alpha_t}{\partial t} = \Delta(\tilde{g}(\alpha)),$$

where $sg'(s) = \tilde{g}'(s)$. For example, $g(s) = s \log(s)$ corresponds to (101), while $g(s) = s^{p-1}/(p-1)$, $p > 1$, yields slow diffusion. A celebrated, and non-trivial, theorem by McCann is that a function of the form (102), for $g : \mathbb{R}^+ \to \mathbb{R}$, is geodesically convex on $\mathcal{P}(\mathbb{R}^d)$ if $g(0) = 0$, g is convex increasing super-linearly and $s \mapsto g(s^{-d})s^d$ is convex decreasing. Examples of such functions are $g(s) = s^q$ for $q > 1$ and Shannon entropy $g(s) = s \log(s)$. Note, however, that $-\log(t)$ (associated with the reverse KL divergence) is not super-linear, so one cannot conclude geodesic convexity.

Interaction Energies. In a similar spirit, to obtain non-linear evolutions, but without requiring the measure to have density, one can consider

$$f(\alpha) := \iint k(x, y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y). \tag{103}$$

For a symmetric kernel k:

$$\delta f(\alpha)(x) = 2 \int k(x, y) \mathrm{d}\alpha(y), \quad \nabla_W f(\alpha)(x) = 2 \int \nabla_x k(x, y) \mathrm{d}\alpha(y).$$

For $\alpha_0 = \frac{1}{n} \sum_i \delta_{x_i}$, the flow (98) implies particles $(x_i(t))_i$ obey:

$$\dot{x}_i(t) = -2 \sum_j \nabla k(x_i(t), x_j(t)).$$

Convergence of the flow. In general, analyzing (98) is challenging. A simple case is when f is geodesically convex. Denoting T as the optimal transport map from α_0 to α_1, the function $t \mapsto f(((1-t)\mathrm{Id} + tT)_\sharp \alpha_0)$ is convex. In this case, α_t is well-defined and converges to a global minimizer of f. This applies to linear (100), quadratic (103) with convex $h(x)$ and $k(x, y)$, and Shannon entropy (101).

9.5 Application: Training Two-Layer MLPs as Wasserstein Flows

In this section, we replace the variable x with θ to align with customary notation in machine learning. We consider a two-layer MLP $g_\theta : u \in \mathbb{R}^d \to \mathbb{R}$ with n neurons:

$$g_\theta(u) := \frac{1}{n} \sum_i \psi(\theta_i, u), \quad \text{where } \psi(\theta_i, u) := a_i \langle u, w_i \rangle,$$

and $\theta_i := (w_i, a_i) \in \mathbb{R}^{d+1}$ represents the parameters of a neuron. Importantly, these functions are invariant under permutations of the neurons. Thus, we can rewrite it using a probability distribution $\alpha = \frac{1}{n} \sum_i \delta_{\theta_i}$ as:

$$G_\alpha(u) := \int_{\mathbb{R}^{d+1}} \psi(\theta, u) \, \mathrm{d}\alpha(\theta).$$

This formulation has the advantage that α can represent a continuous density with an infinite number of neurons.

Now, we consider training the MLP by minimizing an empirical risk to predict labels $y_k \in \mathbb{R}$ from features $u_k \in \mathbb{R}^d$:

$$\min_\alpha f(\alpha) := \frac{1}{N} \sum_{k=1}^N \ell(G_\alpha(u_k), y_k).$$

Since $\alpha \mapsto G_\alpha$ is linear, if ℓ is convex, then f is a convex function of α. However, this observation is not particularly useful because α is infinite-dimensional, making standard minimization infeasible. The typical approach is to perform gradient descent on the neuron parameters, as described in (99):

$$\dot{\theta} = -\nabla F(\theta), \quad \text{where } F(\theta) := f\left(\frac{1}{n} \sum_i \delta_{\theta_i}\right).$$

This is equivalent to the PDE (98) on the neuron density, where the Wasserstein gradient is used. Let us write in more detail this PDE in this specific case. For $\ell(s, s') = \frac{1}{2}(s - s')^2$, the first variation of f is:

$$\delta f(\alpha)(\theta) = \frac{1}{N} \sum_{k=1}^N \left(\int \psi(\theta', u_k) \, \mathrm{d}\alpha(\theta') - y_k \right) \psi(\theta, u_k),$$

50

which can be rewritten as:

$$\delta f(\alpha)(\theta) = \int k(\theta, \theta') \, d\alpha(\theta') + g(\theta),$$

where the kernel and potential functions are:

$$k(\theta, \theta') := \frac{1}{N} \sum_k \psi(\theta, u_k) \psi(\theta', u_k), \tag{104}$$

$$g(\theta) := -\frac{1}{N} \sum_k y_k \psi(\theta, u_k). \tag{105}$$

Thus, the Wasserstein gradient becomes:

$$\nabla_W f(\alpha)(\theta) = \int \nabla_\theta k(\theta, \theta') \, d\alpha(\theta') + \nabla_\theta g(\theta).$$

This corresponds to a Wasserstein flow of the sum of a quadratic and a linear interaction potential. These functions are not geodesically convex because neither k nor g are convex, making convergence analysis challenging.

A breakthrough was achieved by Chizat and Bach, who proved that if the initialization has enough Dirac masses, the flow cannot become stuck in local minima or saddle points. This result leverages the classical convexity of the function f and the 1-homogeneity of $\psi((a, w), u)$ with respect to the external weights a.

9.6 Application: Evolution in Depth of Transformers

We consider very deep transformers, focusing on a single-head attention mechanism for simplicity while ignoring MLP layers, layer normalization, causality, and masking. This framework is best suited to modeling encoders and vision transformers.

After tokenization, embedding, and positional encoding, each input (from a set of tokens) is represented as a point cloud $(x_i)_{i=1}^n$ of n points in the space of vectorized tokens. An attention layer with skip connection and rescaling by $1/T$ (where T is the depth) defines a transformation of the tokens:

$$x_i \mapsto x_i + \frac{1}{T} \sum_j \frac{e^{\langle Qx_i, Kx_j \rangle} V x_j}{\sum_\ell e^{\langle Qx_i, Kx_\ell \rangle}},$$

where $\theta = (K, Q, V)$ are the parameters of the attention layer, represented by three matrices.

To handle an arbitrary number of tokens, we define $\alpha = \frac{1}{n} \sum_i \delta_{x_i}$ as the empirical measure of tokens and rewrite the transformer mapping as:

$$x_i \mapsto x_i + \frac{1}{T} \Gamma_\theta[\alpha](x_i),$$

where

$$\Gamma_\theta[\alpha](x) := \int \frac{e^{\langle Qx, Ky \rangle} V y \, d\alpha(y)}{\int e^{\langle Qx, Kz \rangle} \, d\alpha(z)}.$$

In terms of the evolution of the token distribution α, this means α is pushed forward by the "in-context" mapping $\Gamma_{\theta_t}[\alpha]$, which depends on the context α, the tokens, and the depth-dependent parameters θ_t. Denoting $t \in [0, 1]$ as the depth and $\tau = 1/T$ as the step size, this gives:

$$\alpha_{t+\tau} = (\mathrm{Id} + \tau \Gamma_{\theta_t}[\alpha_t])_\sharp \alpha_t.$$

As $\tau \to 0$, this converges to the following conservation equation:

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t \Gamma[\alpha_t]) = 0.$$

An interesting remark is that, when $V = KQ^T$, then $\Gamma[\alpha]$ is a gradient vector field, but it is not a gradient of a first variation, so that this PDE is not a Wasserstein gradient flow. This formulation was first introduced by Michael Sander in the Sinkformer's paper, modeling deep transformers as PDEs. The key challenge lies in understanding the training of the network, which corresponds to optimizing the parameters $(\theta_t)_t$. This remains an open problem.

References

[1] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

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

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

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

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

[6] 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.

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

[8] 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.

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

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

[11] Imre Ciszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

[21] 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.

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

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

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

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

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

[27] 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.

[28] 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.