Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

June 3, 2019

Chapter 1

Optimal Transport

The main reference for this chapter is the book "Computational Optimal Transport"[1]. We will only recap here some important points.

1.1 Radon Measures

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

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

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

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i} \tag{1.1}$$

where δ_x is the Dirac at position x, intuitively a unit of mass which is infinitely concentrated at location x. Such as measure describes a probability measure if, additionally, $\mathbf{a} \in \Sigma_n$, and more generally a positive measure if each of the "weights" described in vector $\mathbf{a}$ is positive itself.

Remark 1 (General measures). A convenient feature of OT is that it can deal with discrete and continuous "objects" within the same framework. Such objects only need to be modelled as measures. This corresponds to the notion of Radon measures $\mathcal{M}(\mathcal{X})$ on the space $\mathcal{X}$. The formal definition of that set requires that $\mathcal{X}$ is equipped with a distance, usually denoted d, because one can only access a measure by "testing" (integrating) it against continuous functions, denoted $f \in \mathcal{C}(\mathcal{X})$.

Integration of $f \in \mathcal{C}(\mathcal{X})$ against a discrete measure α computes a sum

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

More general measures, for instance on $\mathcal{X} = \mathbb{R}^d$ (where $d \in \mathbb{N}^*$ is the dimension), can have a density $\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x$ w.r.t. the Lebesgue measure, often denoted $\rho_\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}x}$, which means that

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

[1] optimaltransport.github.io

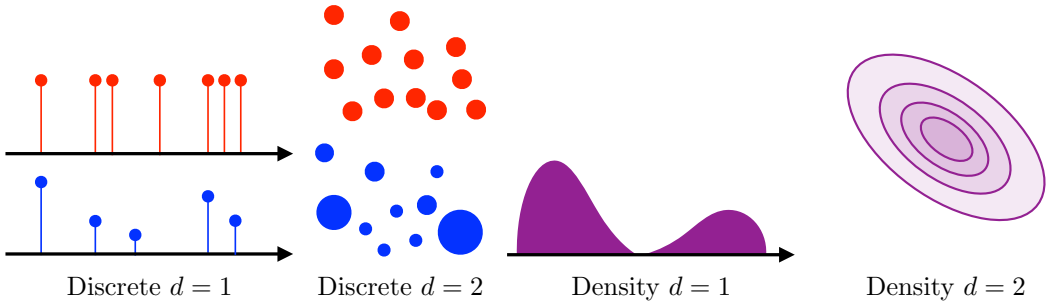


Figure 1.1: Schematic display of discrete distributions $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ (red corresponds to empirical uniform distribution $\mathbf{a}_i = 1/n$, and blue to arbitrary distributions) and densities $\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x$ (in violet), in both 1-D and 2-D. Discrete distributions in 1-D are displayed using vertical segments (with length equal to $\mathbf{a}_i$) and in 2-D using point clouds (radius equal to $\mathbf{a}_i$).

An arbitrary measure $\alpha \in \mathcal{M}(\mathcal{X})$ (which needs not to have a density nor be a sum of Diracs) is defined by the fact that it can be integrated agains any continuous function $f \in \mathcal{C}(\mathcal{X})$ and obtain $\int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) \in \mathbb{R}$. If $\mathcal{X}$ is not compact, one should also impose that f has compact support or at least as 0 limit at infinity. Measure as 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. We denote $\mathcal{M}_+(\mathcal{X})$ the set of all positive measures on $\mathcal{X}$. 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 $\alpha(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}\alpha = 1$. Figure **??** offers a visualization of the different classes of measures, beyond histograms, considered in this work.

Operators on measures. For some continuous map $T : \mathcal{X} \to \mathcal{Y}$, we define the pushforward operator $T_\sharp : \mathcal{M}(X) \to \mathcal{M}(Y)$. For discrete measures (**??**), the pushforward operation consists simply in moving the positions of all the points in the support of the measure

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

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

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

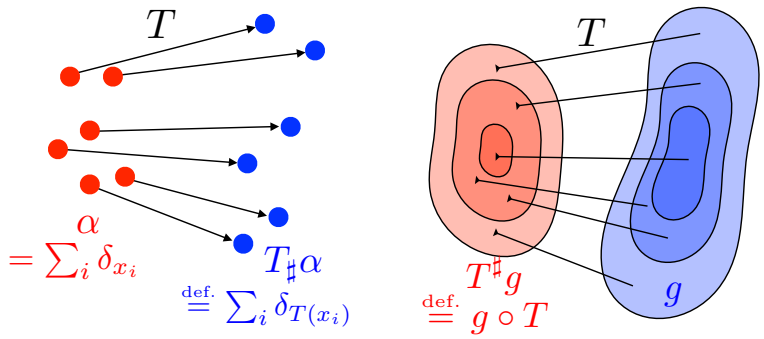
$$\forall\, h \in \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y)\mathrm{d}\beta(y) = \int_{\mathcal{X}} h(T(x))\mathrm{d}\alpha(x). \tag{1.2}$$

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

$$\beta(B) = \alpha(\{x \in \mathcal{X}\ ;\ T(x) \in B\}). \tag{1.3}$$

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

Intuitively, a measurable map $T : \mathcal{X} \to \mathcal{Y}$, can be interpreted as a function "moving" a single point from a measurable space to another. The more general extension $T_\sharp$ can now "move" an entire probability measure on $\mathcal{X}$ towards a new probability measure on $\mathcal{Y}$. The operator $T_\sharp$ "pushes forward" each elementary mass of a measure α on $\mathcal{X}$ by applying the map T to obtain then an elementary mass in $\mathcal{Y}$, to build on aggregate a new measure on $\mathcal{Y}$) written $T_\sharp \alpha$. Note that such a push-forward $T_\sharp : \mathcal{M}_+^1(\mathcal{X}) \to \mathcal{M}_+^1(\mathcal{Y})$ is a linear operator between measures in the sense that for two measures α_1, α_2 on $\mathcal{X}$, $T_\sharp(\alpha_1 + \alpha_2) = T_\sharp \alpha_1 + T_\sharp \alpha_2$.



Push-forward of measures Pull-back of functions

Figure 1.2: Comparison of push-forward $T_\sharp$ and pull-back $T^\sharp$.

Remark 2 (Push-forward for densities). Explicitly doing the change of variable in formula (**??**) for measures with densities $(\rho_\alpha, \rho_\beta)$ on $\mathbb{R}^d$ (assuming T is smooth and a bijection) shows that a push-forward acts on densities linearly as a change of variables in the integration formula, indeed

$$\rho_\alpha(x) = |\det(T'(x))| \rho_\beta(T(x)) \tag{1.4}$$

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 3 (Push-forward vs. pull-back). The push-forward $T_\sharp$ of measures should not be confounded with the pull-back of function $T^\sharp : \mathcal{C}(\mathcal{Y}) \to \mathcal{C}(\mathcal{X})$ which corresponds to the "warping" of functions. It is the linear map defined, for $g \in \mathcal{C}(\mathcal{Y})$ by $T^\sharp g = g \circ T$. Push-forward and pull-back are actually adjoint one from each others, in the sense that

$$\forall\, (\alpha, g) \in \mathcal{M}(\mathcal{X}) \times \mathcal{C}(Y), \quad \int_{\mathcal{Y}} g \mathrm{d}(T_\sharp \alpha) = \int_{\mathcal{X}} (T^\sharp g) \mathrm{d}\alpha.$$

It is important to realize that even if (α, β) have densities $(\rho_\alpha, \rho_\beta)$, $T_\sharp \alpha$ is not equal to $T^\sharp \rho_\beta$, because of the presence of the Jacobian in (**??**). This explains why OT should be used with caution to perform image registration, because it does not operate as an image warping method. Figure **??** illustrate the distinction between these push-forward and pull-back operators.

Remark 4 (Measures and random variables). Radon 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 α is the Radon measure $X \in \mathcal{M}_+^1(\mathcal{X})$ such that $\mathbb{P}(X \in A) = \alpha(A) = \int_A \mathrm{d}\alpha(x)$. Equivalently, it 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 (**??**), 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.

Convergence of random variable. Convergence of random variable (in probability, almost sure, in law), convergence of measures (strong, weak).

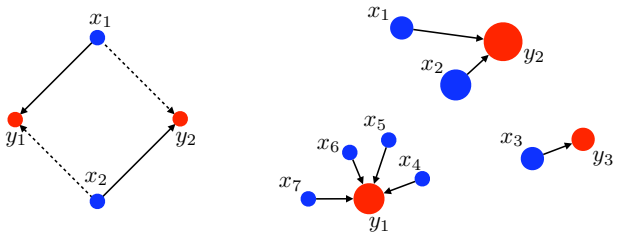


Figure 1.3: (left) blue dots from measure α and red dots from measure β are pairwise equidistant. Hence, either matching $\sigma = (1, 2)$ (full line) or $\sigma = (2, 1)$ (dotted line) is optimal. (right) a Monge map can associate the blue measure α to the red measure β. The weights α_i are displayed proportionally to the area of the disk marked at each location. The mapping here is such that $T(x_1) = T(x_2) = y_2$, $T(x_3) = y_3$, whereas for $4 \leqslant i \leqslant 7$ we have $T(x_i) = y_1$.

1.2 Monge Problem

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

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}. \tag{1.5}$$

One could naively evaluate the cost function above using all permutations in the set $\mathrm{Perm}(n)$. However, that set has size $n!$, which is gigantic even for small n. Consider for instance that such a set has more than 10^{100} elements [**?**] when n is as small as 70. That problem can therefore only be solved if there exist efficient algorithms to optimize that cost function over the set of permutations, which will be the subject of §**??**.

Remark 5 (Uniqueness). Note that the optimal assignment problem may have several optimal solutions. Suppose for instance that $n = m = 2$ and that the matrix $\mathbf{C}$ is the pairwise distance matrix between the 4 corners of a 2-dimensional square of side length 1, as represented in the left plot in Figure **??**. In that case only two assignments exist, and they share the same cost.

For discrete measures

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j} \tag{1.6}$$

the Monge problem [**?**] seeks for a map that associates to each point x_i a single point y_j, and which must push the mass of α toward the mass of β, which is to say that such a map $T : \{x_1, \ldots, x_n\} \to \{y_1, \ldots, y_m\}$ must verify that

$$\forall\, j \in [\![m]\!], \quad \mathbf{b}_j = \sum_{i : T(x_i) = y_j} \mathbf{a}_i \tag{1.7}$$

which we write in compact form as $T_\sharp \alpha = \beta$. This map should minimize some transportation cost, which is parameterized by a function $c(x, y)$ defined for points $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \;;\; T_\sharp \alpha = \beta \right\}. \tag{1.8}$$

Such a map between discrete points can be of course encoded, assuming all x's and y's are distinct, using indices $\sigma : [\![n]\!] \to [\![m]\!]$ so that $j = \sigma(i)$, and the mass conservation is written as

$$\sum_{i \in \sigma^{-1}(j)} \mathbf{a}_i = \mathbf{b}_j.$$

In the special case when $n = m$ and all weights are uniform, that is $\mathbf{a}_i = \mathbf{b}_j = 1/n$, then the mass conservation constraint implies that T is a bijection, such that $T(x_i) = y_{\sigma(i)}$, and the Monge problem is equivalent to the optimal matching problem (**??**) where the cost matrix is

$$\mathbf{C}_{i,j} \overset{\text{def.}}{=} c(x_i, y_j).$$

When $n \neq m$, note that, optimality aside, Monge maps may not even exist between an empirical measure to another. This happens when their weight vectors are not compatible, which is always the case when the target measure has more points than the source measure. For instance, the right plot in Figure **??** shows an (optimal) Monge map between α and β, but there is no Monge map from β to α.

Monge problem (**??**) 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

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\alpha(x) \;;\; T_\sharp \alpha = \beta \right\} \tag{1.9}$$

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 (**??**).

1.3 Kantorovitch Problem

The assignment problem has several limitations in practical settings, also encountered when using the Monge problem. Indeed, because the assignment problem is formulated as a permutation problem, it can only be used to compare two points clouds of the *same* size. A direct generalization to discrete measures with non-uniform weights can be carried out using Monge's formalism of pushforward maps, but that formulation may also be degenerate it there does not exist feasible solutions satisfying the mass conservation constraint (**??**) (see the end of Remark **??**). Additionally, the assignment Problem (**??**) is combinatorial, whereas the feasible set for the Monge Problem (**??**), consisting in all push-forward measures that satisfy the mass conservation constraint, is *non-convex*. Both are therefore difficult to solve in their original formulation.

Kantorovitch formulation for discrete measures. The key idea of [**?**] 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 $\mathbf{P} \in \mathbb{R}_+^{n \times m}$, where $\mathbf{P}_{i,j}$ 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 (**??**). Admissible couplings admit a far simpler characterization than Monge maps:

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \;;\; \mathbf{P}\mathbb{1}_m = \mathbf{a} \quad \text{and} \quad \mathbf{P}^{\mathrm{T}}\mathbb{1}_n = \mathbf{b} \right\}, \tag{1.10}$$

where we used the following matrix-vector notation

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

The set of matrices $\mathbf{U}(\mathbf{a}, \mathbf{b})$ 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 (as illustrated in the right plot of Figure **??**) was intrisically asymmetric, Kantorovich's relaxed formulation is always symmetric, in the sense that a coupling $\mathbf{P}$ is in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ if and only if $\mathbf{P}^{\mathrm{T}}$ is in $\mathbf{U}(\mathbf{b}, \mathbf{a})$.

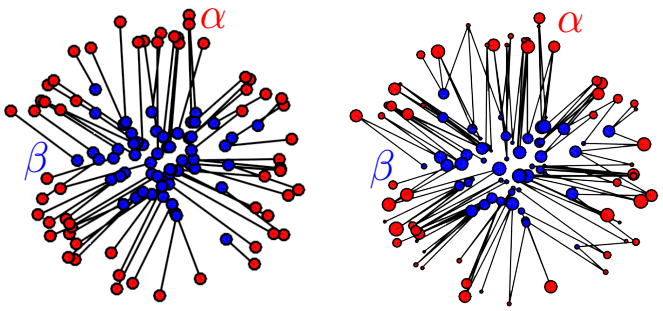


Figure 1.4: Comparison of optimal matching and generic couplings. A black segment between x_i and y_j indicates a non-zero element in the displayed optimal coupling $\mathbf{P}_{i,j}$ solving (**??**). Left: optimal matching, corresponding to the setting of Proposition (**??**) (empirical measures with the same number $n = m$ of points). Right: these two weighted point clouds cannot be matched; instead a Kantorovich coupling can be used to associate two arbitrary discrete measures.

Kantorovich's optimal transport problem now reads

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\, \mathbf{P}\rangle \overset{\text{def.}}{=} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}. \tag{1.11}$$

This is a linear program (see Chapter **??**), and as is usually the case with such programs, its solutions are not necessarily unique.

Permutation Matrices as Couplings For a permutation $\sigma \in \mathrm{Perm}(n)$, we write $\mathbf{P}_\sigma$ for the corresponding permutation matrix,

$$\forall\,(i,j)\in[\![n]\!]^2, \quad (\mathbf{P}_\sigma)_{i,j} = \begin{cases} 1/n & \text{if} \quad j=\sigma_i, \\ 0 & \text{otherwise.} \end{cases} \tag{1.12}$$

One can check that in that case

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

which shows that the assignment problem (**??**) can be recast as a Kantorovich problem (**??**) where the couplings $\mathbf{P}$ are restricted to be exactly permutation matrices:

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

Next, one can easily check that the set of permutation matrices is strictly included in the so-called Birkhoff polytope $\mathbf{U}(\mathbb{1}_n/n, \mathbb{1}_n, n)$. Indeed, for any permutation σ we have $\mathbf{P}_\sigma\mathbb{1} = \mathbb{1}_n$ and ${\mathbf{P}_\sigma}^{\mathrm{T}}\mathbb{1} = \mathbb{1}_n$, whereas $\mathbb{1}_n{\mathbb{1}_n}^{\mathrm{T}}/n^2$ is a valid coupling but not a permutation matrix. Therefore, one has naturally that

$$\min_{\sigma\in\mathrm{Perm}(n)} \langle \mathbf{C},\, \mathbf{P}_\sigma\rangle \leqslant \mathrm{L}_{\mathbf{C}}(\mathbb{1}_n/n, \mathbb{1}_n/n).$$

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 (**??**) between two uniform measures $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, which shows that the Kantorovich relaxation is *tight* when considered on assignment problems. Figure **??** shows on the left a 2-D example of optimal matching corresponding to this special case.

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

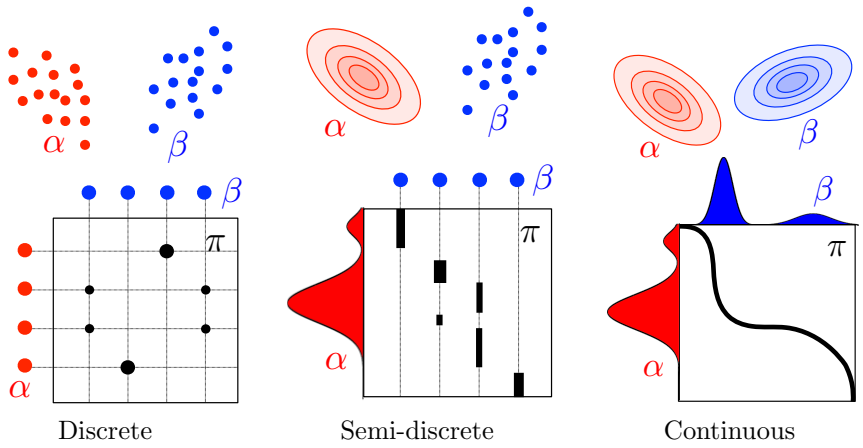


Figure 1.5: Schematic viewed of input measures (α, β) and couplings $\mathcal{U}(\alpha, \beta)$ encountered in the three main scenario for Kantorovich OT. Chapter **??** is dedicated to the semi-discrete setup.

Proof. Birkhoff's theorem states that the set of extremal points of $\mathbf{U}(\mathbb{1}_n/n, \mathbb{1}_n/n)$ is equal to the set of permutation matrices. A fundamental theorem of linear programming [**?**, Theorem 2.7] states that the minimum of a linear objective in a non-empty polyhedron, if finite, is reached at an extremal point of the polyhedron. □

Kantorovitch formulation for arbitrary measures. The definition of $\mathcal{L}_c$ in (**??**) can be 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 discrete case is a special situation where one imposes this product measure to be of the form $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}$. In the general case, the mass conservation constraint (**??**) should be rewritten as a marginal constraint on joint probability distributions

$$\mathcal{U}(\alpha, \beta) \overset{\text{def.}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \;;\; P_{\mathcal{X}\sharp}\pi = \alpha \quad \text{and} \quad P_{\mathcal{Y}\sharp}\pi = \beta \right\}. \tag{1.13}$$

Here $P_{\mathcal{X}\sharp}$ and $P_{\mathcal{Y}\sharp}$ are the push-forward (see Definition **??**) by the projections $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$. Figure **??** shows a schematic visualization of the coupling constraints for different class of problem (discrete measures and densities). Using (**??**), these marginal constraints are 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 $B \subset \mathcal{Y}$.

The Kantorovich problem (**??**) is then generalized as

$$\mathcal{L}_c(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y). \tag{1.14}$$

This is an infinite-dimensional linear program over a space of measures. Figure **??** shows examples of discrete and continuous optimal coupling solving (**??**). Figure **??** shows other examples of optimal 1-D couplings, involving discrete and continuous marginals.

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

Wasserstein distances. An important feature of OT is that it defines a distance between histograms and probability measures as soon as the cost matrix satisfies certain suitable properties. Indeed, OT can be understood as a canonical way to lift a ground distance between points to a distance between histogram or measures.

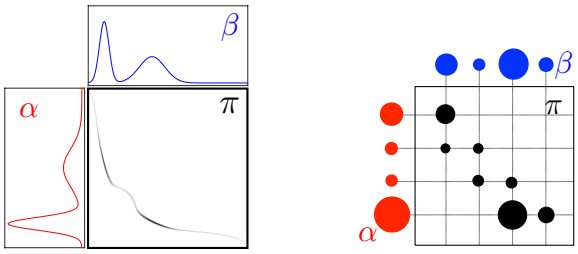


Figure 1.6: Left: "continuous" coupling π solving (**??**) between two 1-D measure with density. The coupling is localized along the graph of the Monge map $(x, T(x))$ (displayed in black). Right: "discrete" coupling T solving (**??**) between two discrete measures of the form (**??**). The non-zero entries $T_{i,j}$ are display with a black disk at position (i, j) with radius proportional to $T_{i,j}$.

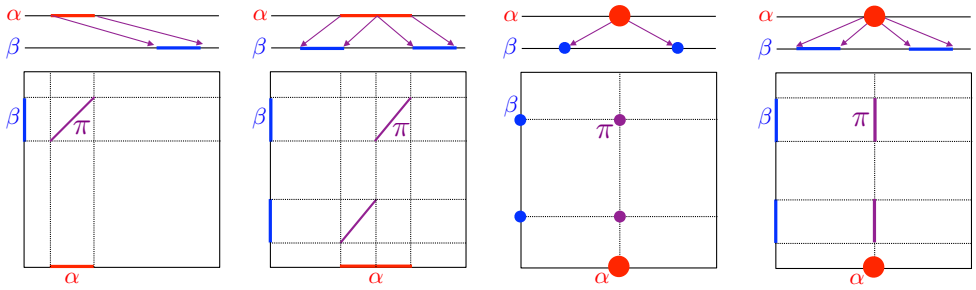


Figure 1.7: Four simple examples of optimal couplings between 1-D distributions, represented as maps above (arrows) and couplings below. Inspired by [**?**].

We first consider the case where, using a term first introduce by [**?**], the "ground metric" matrix $\mathbf{C}$ is fixed, representing substitution costs between bins, and shared across several histograms we would like to compare. The following proposition states that OT provides a meaningful distance between histograms supported on these bins.

Proposition 2. *We suppose $n = m$, and that for some $p \geqslant 1$, $\mathbf{C} = \mathbf{D}^p = (\mathbf{D}_{i,j}^p)_{i,j} \in \mathbb{R}^{n \times n}$ where $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is a distance on $[\![n]\!]$,* i.e.

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

Then

$$\mathrm{W}_p(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \mathrm{L}_{\mathbf{D}^p}(\mathbf{a}, \mathbf{b})^{1/p} \tag{1.15}$$

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

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

Proof. Symmetry and definiteness of the distance are easy to prove: since $\mathbf{C} = \mathbf{D}^p$ has a null diagonal, $\mathrm{W}_p(\mathbf{a}, \mathbf{a}) = 0$, with corresponding optimal transport matrix $\mathbf{P}^\star = \mathrm{diag}(\mathbf{a})$; by the positivity of all off-diagonal elements of $\mathbf{D}^p$, $\mathrm{W}_p(\mathbf{a}, \mathbf{b}) > 0$ whenever $\mathbf{a} \neq \mathbf{b}$ (because in this case, an admissible coupling necessarily has a non-zero element outside the diagonal); by symmetry of $\mathbf{D}^p$, $\mathrm{W}_p(\mathbf{a}, \mathbf{b}) = 0$ is itself a symmetric function.

To prove the triangle inequality of Wasserstein distances for arbitrary measures, [**?**, Theorem 7.3] uses the gluing lemma, which stresses the existence of couplings with a prescribed structure. In the discrete setting,

the explicit constuction of this glued coupling is simple. Let $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \Sigma_n$. Let $\mathbf{P}$ and $\mathbf{Q}$ be two optimal solutions of the transport problems between $\mathbf{a}$ and $\mathbf{b}$, and $\mathbf{b}$ and $\mathbf{c}$ respectively. We define $\bar{\mathbf{b}}_j \overset{\text{def.}}{=} \mathbf{b}_j$ if $\mathbf{b}_j > 0$ and set otherwise $\bar{\mathbf{b}}_j = 1$ (or actually any other value). We then define

$$\mathbf{S} \overset{\text{def.}}{=} \mathbf{P} \operatorname{diag}(1/\bar{\mathbf{b}}) \mathbf{Q} \in \mathbb{R}_+^{n \times n}.$$

We remark that $\mathbf{S} \in U(\mathbf{a}, \mathbf{c})$ because

$$\mathbf{S}\mathbb{1}_n = \mathbf{P} \operatorname{diag}(1/\bar{\mathbf{b}}) \mathbf{Q}\mathbb{1}_n = \mathbf{P}(\mathbf{b}/\bar{\mathbf{b}}) = \mathbf{P}\mathbb{1}_{\operatorname{Supp}(\mathbf{b})} = \mathbf{a}$$

where we denoted $\mathbb{1}_{\operatorname{Supp}(\mathbf{b})}$ the indicator of the support of $\mathbf{b}$, and we use the fact that $\mathbf{P}\mathbb{1}_{\operatorname{Supp}(\mathbf{b})} = \mathbf{P}\mathbb{1} = \mathbf{b}$ because necessarily $\mathbf{P}_{i,j} = 0$ for $j \notin \operatorname{Supp}(\mathbf{b})$. Similarly one verifies that $\mathbf{S}^\top \mathbb{1}_n = \mathbf{c}$.

The triangle inequality follows from

$$\begin{aligned}
\mathrm{W}_p(\mathbf{a}, \mathbf{c}) &= \left(\min_{\mathbf{P} \in U(\mathbf{a},\mathbf{c})} \langle \mathbf{P}, \mathbf{D}^p \rangle \right)^{1/p} \leqslant \langle \mathbf{S}, \mathbf{D}^p \rangle^{1/p} \\
&= \left(\sum_{ik} \mathbf{D}_{ik}^p \sum_j \frac{\mathbf{P}_{ij}\mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} \leqslant \left(\sum_{ijk} (\mathbf{D}_{ij} + \mathbf{D}_{jk})^p \frac{\mathbf{P}_{ij}\mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} \\
&\leqslant \left(\sum_{ijk} \mathbf{D}_{ij}^p \frac{\mathbf{P}_{ij}\mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} + \left(\sum_{ijk} \mathbf{D}_{jk}^p \frac{\mathbf{P}_{ij}\mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} \\
&= \left(\sum_{ij} \mathbf{D}_{ij}^p \mathbf{P}_{ij} \sum_k \frac{\mathbf{Q}_{jk}}{\bar{\mathbf{b}}_j} \right)^{1/p} + \left(\sum_{jk} \mathbf{D}_{jk}^p \mathbf{Q}_{jk} \sum_i \frac{\mathbf{P}_{ij}}{\bar{\mathbf{b}}_j} \right)^{1/p} \\
&= \left(\sum_{ij} \mathbf{D}_{ij}^p \mathbf{P}_{ij} \right)^{1/p} + \left(\sum_{jk} \mathbf{D}_{jk}^p \mathbf{Q}_{jk} \right)^{1/p} \\
&= \mathrm{W}_p(\mathbf{a}, \mathbf{b}) + \mathrm{W}_p(\mathbf{b}, \mathbf{b}).
\end{aligned}$$

The first inequality is due to the suboptimality of $\mathbf{S}$, the second is the usual triangle inequality for elements in $\mathbf{D}$, and the third comes from Minkowski's inequality. □

Proposition **??** generalizes from histogram to arbitrary measures that need not be discrete.

Proposition 3. *We assume $\mathcal{X} = \mathcal{Y}$, and that for some $p \geqslant 1$, $c(x, y) = d(x, y)^p$ where d is a distance on $\mathcal{X}$,* i.e.

(i) $d(x, y) = d(y, x) \geqslant 0$;
(ii) $d(x, y) = 0$ if and only if $x = y$;
(ii) $\forall\, (x, y, z) \in \mathcal{X}^3, d(x, z) \leqslant d(x, y) + d(y, z)$.

Then

$$\mathcal{W}_p(\alpha, \beta) \overset{\text{def.}}{=} \mathcal{L}_{d^p}(\alpha, \beta)^{1/p} \tag{1.16}$$

(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 proof follows the same approach as that for Proposition **??** and relies on the existence of a coupling between (α, γ) obtained by "guying" optimal couplings between (α, β) and (β, γ). □

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. In particular, "classical" distances (or divergences) are not even defined between discrete distributions (the L^2 norm can only be applied to continuous measures with a density with respect to a base measure, and the discrete ℓ^2 norm requires the positions (x_i, y_j) to be fixed to work). In sharp contrast, one has that for any $p > 0$, $\mathcal{W}_p^p(\delta_x, \delta_y) = d(x, y)$. Indeed, it suffices to notice that $\mathcal{U}(\delta_x, \delta_y) = \{\delta_{x,y}\}$ and therefore the Kantorovich problem having only one feasible solution, $\mathcal{W}_p^p(\delta_x, \delta_y)$ is necessarily $(d(x, y)^p)^{1/p} = d(x, y)$. This shows that $\mathcal{W}_p(\delta_x, \delta_y) \to 0$ if $x \to y$. This property corresponds to the fact that $\mathcal{W}_p$ is a way to quantify the weak convergence as we now define.

Definition 2 (Weak convergence)**.** *$(\alpha_k)_k$ converges weakly to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any continuous function $g \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} g \mathrm{d}\alpha_k \to \int_{\mathcal{X}} g \mathrm{d}\alpha$. This notion of weak convergence corresponds to the convergence in law of random vectors.*

This convergence can be shown to be equivalent to $\mathcal{W}_p(\alpha_k, \alpha) \to 0$ [**?**, Theorem 6.8] (together with a convergence of the moments up to order p for unbounded metric spaces).

Note that there exists alternative distances which also metrize weak convergence. The simplest one are Hilbertian norms, defined as

$$\|\alpha\|_k^2 \overset{\text{def.}}{=} \mathbb{E}_{\alpha\otimes\alpha}(k) = \int_{\mathcal{X}\times\mathcal{X}} k(x, y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y)$$

for a suitable choice of kernel $k : \mathcal{X}^2 \to \mathbb{R}$. The most famous of such kernel is the Gaussian one $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ for some choice of bandwidth $\sigma > 0$.

This convergence should not be confounded with the strong convergence of measures, which is metrized by the TV norm $\|\alpha\|_{TV} \overset{\text{def.}}{=} |\alpha|(\mathcal{X})$, which is the total mass of the absolute value of the measure.

Algorithms Since (**??**)Â is a linear program, it is possible to use any classical linear program solver, such as interior point methods or simplex. In practice, the network simplex is an efficient option, and it used pivoting rule adapted to the OT constraint set. In the case of the assignment problem, $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, there exists faster combinatorial optimization scheme, the most famous ones being the Hungarian algorithm and the auction algorithm, which have roughly $O(n^3)$ complexity. Section **??** details an approximate algorithm, which is typically faster, and amenable to parallelisation, but do not compute exactly the solution to the OT problem.

1.4 Duality

The Kantorovich problem (**??**) is a constrained convex minimization problem, and as such, it can be naturally paired with a so-called dual problem, which is a constrained concave maximization problem. The following fundamental proposition, which is a special case of Fenchel-Rockafellar duality theory, explains the relationship between the primal and dual problems.

Proposition 4. *One has*

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \max_{(\mathbf{f},\mathbf{g})\in\mathbf{R}(\mathbf{a},\mathbf{b})} \langle \mathbf{f}, \mathbf{a}\rangle + \langle \mathbf{g}, \mathbf{b}\rangle \tag{1.17}$$

where the set of admissible potentials is

$$\mathbf{R}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \{(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m \;;\; \forall\, (i, j) \in [\![n]\!] \times [\![m]\!], \mathbf{f} \oplus \mathbf{g} \leqslant \mathbf{C}\} \tag{1.18}$$

Proof. This result is a direct consequence of the more general result on the strong duality for linear programs [**?**, p.148,Theo.4.4]. The easier part of that result, namely that the right-hand side of Equation (**??**)

is a lower bound on $\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ is discussed in **??**. For the sake of completeness, let us derive this dual problem with the use of Lagrangian duality. The Lagangian associate to (**??**) reads

$$\min_{\mathbf{P}\geqslant 0} \max_{(\mathbf{f},\mathbf{g})\in\mathbb{R}^n\times\mathbb{R}^m} \langle \mathbf{C},\, \mathbf{P}\rangle + \langle \mathbf{a}-\mathbf{P}\mathbb{1}_m,\, \mathbf{f}\rangle + \langle \mathbf{b}-\mathbf{P}^\top\mathbb{1}_n,\, \mathbf{g}\rangle. \tag{1.19}$$

For linear program, one can always exchange the min and the max and get the same value of the linear program, and one thus consider

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

We conclude by remarking that

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

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

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

$$\mathrm{Supp}(\mathbf{P}) \subset \left\{(i,j)\in[\![n]\!]\times[\![m]\!]\;;\; \mathbf{f}_i+\mathbf{g}_j = \mathbf{C}_{i,j}\right\}. \tag{1.20}$$

To extend this primal-dual construction to arbitrary measures, it is important to realize that measures are naturally paired in duality with continuous functions (a measure can only be accessed through integration against continuous functions). The duality is formalized in the following proposition, which boils down to Proposition **??** when dealing with discrete measures.

Proposition 5. *One has*

$$\mathcal{L}_c(\alpha,\beta) = \max_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y), \tag{1.21}$$

where the set of admissible dual potentials is

$$\mathcal{R}(c) \stackrel{\text{def.}}{=} \{(f,g)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})\;;\; \forall (x,y), f(x)+g(y)\leqslant c(x,y)\}. \tag{1.22}$$

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

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

$$\mathrm{Supp}(\pi) \subset \{(x,y)\in\mathcal{X}\times\mathcal{Y}\;;\; f(x)+g(y)=c(x,y)\}. \tag{1.23}$$

Note that in contrast to the primal problem (**??**), showing the existence of solutions to (**??**) 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 **??**, one can however show that optimal (f,g) are necessarily Lipschitz regular, which enable to replace the constraint by a compact one.

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

Theorem 1 (Brenier)**.** *In the case $\mathcal{X}=\mathcal{Y}=\mathbb{R}^d$ and $c(x,y)=\|x-y\|^2$, if at least one of the two inputs measures (denoted α) has a density ρ_α with respect to the Lebesgue measure, then the optimal π in the Kantorovich formulation (**??**) is unique, and is supported on the graph $(x,T(x))$ of a "Monge map" $T:\mathbb{R}^d\to\mathbb{R}^d$. This means that $\pi=(\mathrm{Id},T)_\sharp\mu$,* i.e.

$$\forall\, h\in\mathcal{C}(\mathcal{X}\times\mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\mathrm{d}\pi(x,y) = \int_{\mathcal{X}} h(x,T(x))\mathrm{d}\mu(x). \tag{1.24}$$

*Furthermore, this map T is uniquely defined as the gradient of a convex function φ, $T(x) = \nabla\varphi(x)$, where φ is the unique (up to an additive constant) convex function such that $(\nabla\varphi)_\sharp\mu = \nu$. This convex function is related to the dual potential f solving (**??**) as $\varphi(x) = \frac{\|x\|^2}{2} - f(x)$.*

Proof. We sketch the main ingredients of the proof, more details can be found for instance in [**?**]. We remark that $\int c\mathrm{d}\pi = C_{\alpha,\beta} - 2\int\langle x,\, y\rangle\mathrm{d}\pi(x,y)$ where the constant is $C_{\alpha,\beta} = \int\|x\|^2\mathrm{d}\alpha(x) + \int\|y\|^2\mathrm{d}\beta(y)$. Instead of solving (**??**), one can thus consider the following problem

$$\max_{\pi\in\mathcal{U}(\alpha,\beta)}\int_{\mathcal{X}\times\mathcal{Y}}\langle x,\, y\rangle\mathrm{d}\pi(x,y),$$

whose dual reads

$$\min_{(\varphi,\psi)}\left\{\int_{\mathcal{X}}\varphi\mathrm{d}\alpha + \int_{\mathcal{Y}}\psi\mathrm{d}\beta\ ;\ \forall(x,y),\quad \varphi(x)+\psi(y)\geqslant\langle x,\, y\rangle\right\}. \tag{1.25}$$

The relation between these variables and those of (**??**) is $(\varphi,\psi) = (\frac{\|\cdot\|^2}{2} - f, \frac{\|\cdot\|^2}{2} - g)$. One can replace the constraint by

$$\forall\, y,\quad \psi(y)\geqslant\varphi^*(y)\stackrel{\text{def.}}{=}\sup_x\ \langle x,\, y\rangle - \varphi(x). \tag{1.26}$$

Here φ^* is the Legendre transform of φ and is a convex function as a supremum of linear forms (see also (**??**)). Since the objective appearing in (**??**) is linear and the integrating measures positive, one can minimize explicitly with respect to φ and set $\psi = \varphi^*$ in order to consider the unconstraint problem

$$\min_{\varphi}\int_{\mathcal{X}}\varphi\mathrm{d}\alpha + \int_{\mathcal{Y}}\varphi^*\mathrm{d}\beta, \tag{1.27}$$

see also Section **??** for a generalization of this idea to generic costs $c(x,y)$. By iterating this argument twice, one can replace φ by φ^{**}, which is a convex function, and thus impose in (**??**) that φ is convex. Condition (**??**) shows that an optimal π is supported on $\{(x,y)\ ;\ \varphi(x)+\varphi^*(y) = \langle x,\, y\rangle\}$ which shows that such an y is optimal for the minimization (**??**) of the Legendre transform, whose optimality condition reads $y\in\partial\varphi(x)$. Since φ is convex, it is differentiable almost everywhere, and since α has a density, it is also differentiable α-almost everywhere. This shows that for each x, the associated y is uniquely defined α-almost everywhere as $y = \nabla\varphi(x)$, and shows that necessarily $\pi = (\mathrm{Id},\nabla\varphi)_\sharp\alpha$. □

This results shows that in the setting of $\mathcal{W}_2$ with non-singular densities, the Monge problem (**??**) and its Kantorovich relaxation (**??**) are equal (the relaxation is tight). This is the continuous analog of Proposition **??** for the assignment case (**??**), which states that the minimum of the optimal transport problem is achieved, when the marginals are equal and uniform, at a permutation matrix (a discrete map). 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 the natural generalization of the notion of increasing functions in dimension more than one. Optimal transport can thus plays an important role to define quantile functions in arbitrary dimensions, which in turn is useful for applications to quantile regression problems [**?**].

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 function.

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

$$\det(\partial^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x) \tag{1.28}$$

where $\partial^2\varphi(x)\in\mathbb{R}^{d\times d}$ is the hessian of φ. The Monge-Ampère operator $\det(\partial^2\varphi(x))$ can be understood as a non-linear degenerate Laplacian. In the limit of small displacements, $\varphi = \mathrm{Id} + \varepsilon\varphi$, one indeed recovers the Laplacian Δ as a linearization since for smooth maps

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

The convexity constraint forces $\det(\partial^2\varphi(x))\geqslant 0$ and is necessary for this equation to have a solution.

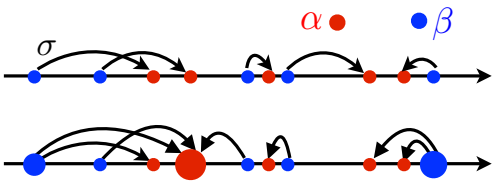


Figure 1.8: 1-D optimal couplings: each arrow $x_i \to y_j$ indicate a non-zero $\mathbf{P}_{i,j}$ in the optimal coupling. Top: empirical measures with same number of points (optimal matching). Bottom: generic case. This corresponds to monotone rearrangements, if $x_i \leqslant x_{i'}$ are such that $\mathbf{P}_{i,j} \neq 0, \mathbf{P}_{i',j'} \neq 0$, then necessarily $y_j \leqslant y_{j'}$.

Special cases In general, computing OT distances is numerically involved. We review special favorable cases where the resolution of the OT problem is easy.

Remark 6 (Binary Cost Matrix and 1-Norm). One can easily check that when the cost matrix $\mathbf{C}$ is zero on the diagonal and 1 elsewhere, namely when $\mathbf{C} = \mathbb{1}_{n\times n} - I_n$, the OT distance between $\mathbf{a}$ and $\mathbf{b}$ is equal to the 1-norm of their difference, $\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\|_1$. One can also easily check that this result extends to discrete and discrete measures in the case where $c(x,y)$ is 0 if $x = y$ and 1 when $x \neq y$. The OT distance between two discrete measures α and β is equal to their total variation distance.

Remark 7 (1-D case – Empirical measures). Here $\mathcal{X} = \mathbb{R}$. Assuming $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, and assuming (without loss of generality) that the points are ordered, *i.e.* $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ and $y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n$, then one has the simple formula

$$\mathcal{W}_p(\alpha,\beta)^p = \sum_{i=1}^p |x_i - y_i|^p, \tag{1.29}$$

i.e. locally (if one assumes distinct points), $\mathcal{W}_p(\alpha,\beta)$ is the ℓ^p norm between two vectors of ordered values of α and β. That statement is only valid locally, in the sense that the order (and those vector representations) might change whenever some of the values change. That formula is a simple consequence of the more general remark given below. Figure **??**, top row, illustrates the 1-D transportation map between empirical measures with the same number of points. The bottom row shows how this monotone map generalizes to arbitrary discrete measures. It is possible to leverage this 1-D computation to also compute efficiently OT on the circle, see [**?**]. Note that in the case of concave cost of the distance, for instance when $p < 1$, the behaviour of the optimal transport plan is very different, see [**?**], which describes an efficient solver in this case.

Remark 8 (1-D case – Generic case). For a measure α on $\mathbb{R}$, we introduce the cumulative function

$$\forall\, x \in \mathbb{R}, \quad \mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^x \mathrm{d}\alpha, \tag{1.30}$$

which is a function $\mathcal{C}_\alpha : \mathbb{R} \to [0,1]$, and 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 generalized quantile function of α. For any $p \geqslant 1$, one has

$$\mathcal{W}_p(\alpha,\beta)^p = \|\mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1}\|_{L^p([0,1])}^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(r) - \mathcal{C}_\beta^{-1}(r)|^p \mathrm{d}r. \tag{1.31}$$

This 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, or, equivalently, that the Wasserstein distance for measures on the real line is a Hilbertian metric. This makes the geometry of 1-D optimal transport very simple, but also very different from its

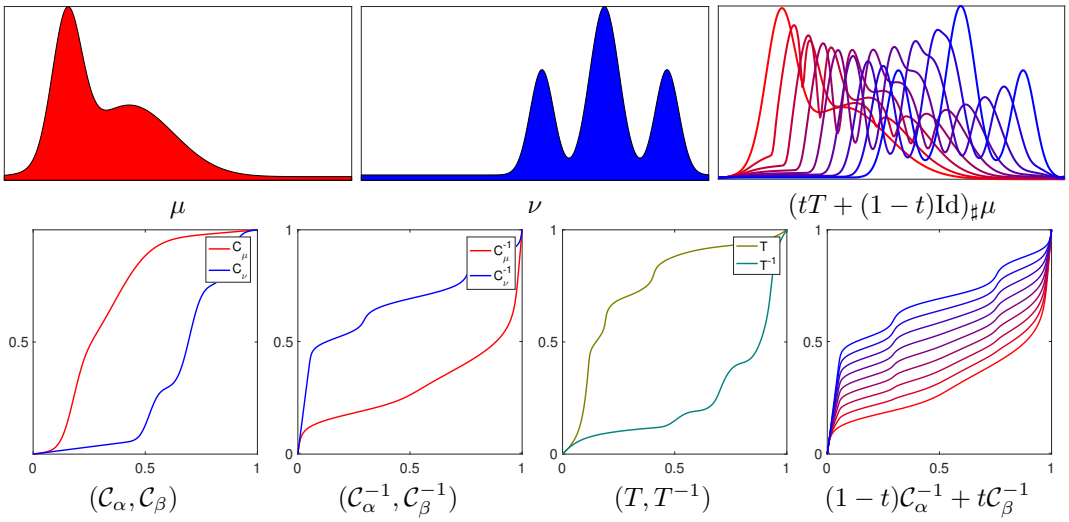


Figure 1.9: Computation of OT and displacement interpolation between two 1-D measures, using cumulant function as detailed in (**??**).

geometry in higher dimensions, which is not Hilbertian as discussed in Proposition **??** and more generally in §**??**. For $p = 1$, one even has the simpler formula

$$\mathcal{W}_1(\alpha, \beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x \tag{1.32}$$

$$= \int_{\mathbb{R}} \left| \int_{-\infty}^{x} \mathrm{d}(\alpha - \beta) \right| \mathrm{d}x. \tag{1.33}$$

which shows that $\mathcal{W}_1$ is a norm (see §**??** for the generalization to arbitrary dimensions). An optimal Monge map T such that $T_\sharp \alpha = \beta$ is then defined by

$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha. \tag{1.34}$$

Figure **??** illustrates the computation of 1-D OT through cumulative functions. It also displays displacement interpolations, computed as detailed in (**??**), see also Remark **??**. For a detailed survey of the properties of optimal transport in 1-D, we refer the reader to [**?**, Chapter 2].

Remark 9 (Distance between Gaussians). If $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$ are two Gaussians in $\mathbb{R}^d$, then one can show that the following map

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha), \tag{1.35}$$

where

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

is such that $T_\sharp \rho_\alpha = \rho_\beta$. Indeed, one simply has to notice that the change of variables formula (**??**) is satisfied since

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

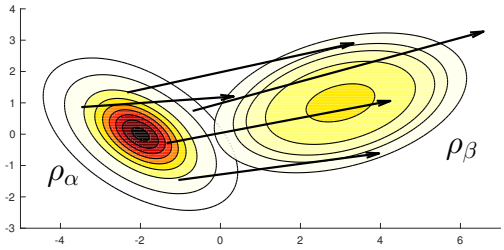


Figure 1.10: Two Gaussians ρ_α and ρ_β, represented using the contour plots of their densities, with respective mean and variance matrices $\mathbf{m}_\alpha = (-2, 0), \mathbf{\Sigma}_\alpha = \frac{1}{2}\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$ and $\mathbf{m}_\beta = (3, 1), \mathbf{\Sigma}_\beta = \begin{pmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$. The arrows originate at random points x taken on the plane and end at the corresponding mappings of those points $T(x) = \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$.

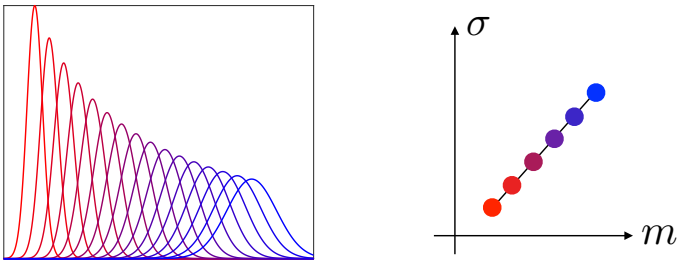


Figure 1.11: Computation of displacement interpolation between two 1-D Gaussians. Denoting $\mathcal{G}_{m,\sigma}(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-m)^2}{2s^2}}$ the Gaussian density, it thus shows the interpolation $\mathcal{G}_{(1-t)m_0+tm_1,(1-t)\sigma_0+t\sigma_1}$.

and since T is a linear map we have that

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

and we therefore recover $\rho_\alpha = |\det T'| \rho_\beta$ meaning $T_\sharp \alpha = \beta$. Notice now that T is the gradient of the convex function $\psi : x \mapsto \frac{1}{2}\langle x - \mathbf{m}_\alpha,\, A(x - \mathbf{m}_\alpha)\rangle + \langle \mathbf{m}_\beta,\, x\rangle$ to conclude, using Brenier's theorem [**?**] (see Remark **??**) that T is optimal. Both that map T and the corresponding potential ψ are illustrated in Figures **??** and **??**

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

$$\mathcal{W}_2^2(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \tag{1.36}$$

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

$$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \overset{\text{def.}}{=} \operatorname{tr}\left(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - 2(\mathbf{\Sigma}_\alpha^{1/2} \mathbf{\Sigma}_\beta \mathbf{\Sigma}_\alpha^{1/2})^{1/2}\right), \tag{1.37}$$

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

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

For 1-D Gaussians, $\mathcal{W}_2$ is thus the Euclidean distance on the 2-D plane $(\mathbf{m}, \sqrt{\mathbf{\Sigma}})$, as illustrated in Figure **??**. For a detailed treatment of the Wasserstein geometry of Gaussian distributions, we refer to [**?**].

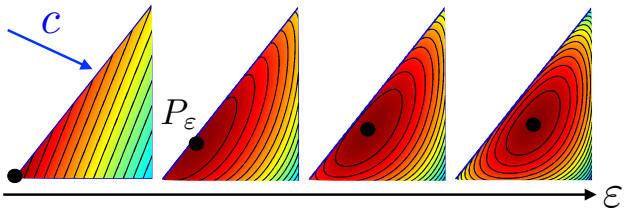


Figure 1.12: Impact of ε on the optimization of a linear function on the simplex, solving $\mathbf{P}_\varepsilon = \operatorname{argmin}_{\mathbf{P}\in\Sigma_3} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon \mathbf{H}(\mathbf{P})$ for a varying ε.

1.5 Sinkhorn

This section introduces a family of numerical scheme to approximate solutions to Kantorovich formulation of optimal transport and its many generalizations. It operates by adding an entropic regularization penalty to the original problem. This regularization has several important advantages, but a few stand out particularly: The minimization of the regularized problen can be solved using a simple alternate minimization scheme; that scheme translates into iterations that are simple matrix products, making them particularly suited to execution of GPU; the resulting approximate distance is smooth with respect to input histogram weights and positions of the Diracs.

Entropic Regularization. The discrete entropy of a coupling matrix is defined as

$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def.}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1), \tag{1.38}$$

with an analogous definition for vectors, with the convention that $\mathbf{H}(\mathbf{a}) = -\infty$ if one of the entries $\mathbf{a}_j$ is 0 or negative. The function $\mathbf{H}$ is 1-strongly concave, because its hessian is $\partial^2\mathbf{H}(P) = -\operatorname{diag}(1/\mathbf{P}_{i,j})$ and $\mathbf{P}_{i,j} \leqslant 1$. The idea of the entropic regularization of optimal transport is to use $-\mathbf{H}$ as a regularizing function to obtain approximate solutions to the original transport problem (**??**):

$$\mathrm{L}^\varepsilon_{\mathbf{C}}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon\mathbf{H}(\mathbf{P}). \tag{1.39}$$

Since the objective is a ε-strongly convex function, problem **??** has a unique optimal solution. The idea to regularize the optimal transport problem by an entropic term can be traced back to modeling ideas in transportation theory [**?**]: Actual traffic patterns in a network do not agree with those predicted by the solution of the optimal transport problem. Indeed, the former are more diffuse than the latter, which tend to rely on a few routes as a result of the sparsity of optimal couplings to the solution of **??**. To balance for that, researchers in transportation proposed a model, called the "gravity" model [**?**], that is able to form a more "blurred" traffic prediction.

Figure **??** illustrates the effect of the entropy to regularize a linear program over the simples Σ_3 (which can thus be visualized as a triangle in 2-D). Note how the entropy pushes the original LP solution away from the boundary of the triangle. The optimal $\mathbf{P}_\varepsilon$ progressively moves toward an "entropic center" of the triangle. This is further detailed in the proposition below. The convergence of the solution of that regularized problem towards an optimal solution of the original linear program has been studied by [**?**].

Proposition 6 (Convergence with ε)**.** *The unique solution $\mathbf{P}_\varepsilon$ of (**??**) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$\mathbf{P}_\varepsilon \stackrel{\varepsilon\to 0}{\longrightarrow} \operatorname*{argmin}_{\mathbf{P}} \{-\mathbf{H}(\mathbf{P}) \;;\; \mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b}), \langle \mathbf{P}, \mathbf{C}\rangle = \mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b})\} \tag{1.40}$$

so that in particular

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

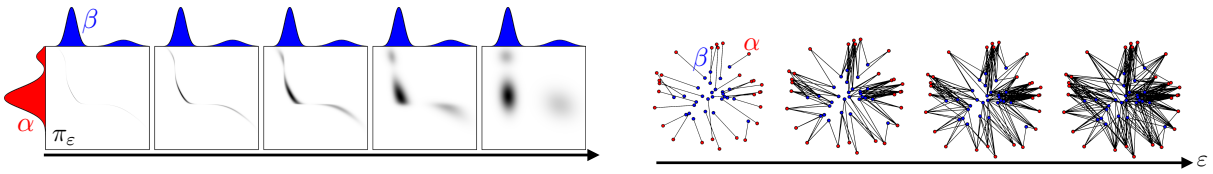


Figure 1.13: Impact of ε on coupling between densities and discrete distributions, illustrating Proposition **??**. Left: between two 1-D densities. Right: between two 2-D discrete empirical densities with same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

One has

$$\mathbf{P}_\varepsilon \overset{\varepsilon\to\infty}{\longrightarrow} \mathbf{a}\mathbf{b}^{\mathrm{T}} = (\mathbf{a}_i\mathbf{b}_j)_{i,j}. \tag{1.41}$$

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

$$0 \leqslant \langle \mathbf{C}, \mathbf{P}_\ell\rangle - \langle \mathbf{C}, \mathbf{P}\rangle \leqslant \varepsilon_\ell(\mathbf{H}(\mathbf{P}_\ell) - \mathbf{H}(\mathbf{P})). \tag{1.42}$$

Since $\mathbf{H}$ is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle \mathbf{C}, \mathbf{P}^\star\rangle = \langle \mathbf{C}, \mathbf{P}\rangle$ so that $\mathbf{P}^\star$ is a feasible point of (**??**). Furthermore, dividing by ε_ℓ in (**??**) and taking the limit shows that $\mathbf{H}(\mathbf{P}) \leqslant \mathbf{H}(\mathbf{P}^\star)$, which shows that $\mathbf{P}^\star$ is a solution of (**??**). Since the solution $\mathbf{P}_0^\star$ to this program is unique by strict convexity of $-\mathbf{H}$, one has $\mathbf{P}^\star = \mathbf{P}_0^\star$, and the whole sequence is converging. □

Formula (**??**) states that for low regularization, the solution converges to the maximum entropy optimal transport coupling. In sharp contrast, (**??**) shows that for large regularization, the solution converges to the coupling with maximal entropy between two prescribed marginals $\mathbf{a},\mathbf{b}$, namely the joint probability between two independent random variables with prescribed distributions. A refined analysis of this convergence is performed in [**?**], including a first order expansion in ε (resp. $1/\varepsilon$) near $\varepsilon = 0$ (resp $\varepsilon = +\infty$). Figure **??** shows visually the effect of these two convergence. A key insight is that, as ε increases, the optimal coupling becomes less and less sparse (in the sense of having entries larger than a prescribed thresholds), which in turn as the effect of both accelerating computational algorithms (as we study in §**??**) but also leading to faster statistical convergence (as exposed in §**??**).

Defining the Kullback-Leibler divergence between couplings as

$$\mathbf{KL}(\mathbf{P}|\mathbf{K}) \overset{\text{def.}}{=} \sum_{i,j} \mathbf{P}_{i,j}\log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}}\right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}, \tag{1.43}$$

the unique solution $\mathbf{P}_\varepsilon$ of (**??**) is a projection onto $\mathbf{U}(\mathbf{a},\mathbf{b})$ of the Gibbs kernel associated to the cost matrix $\mathbf{C}$ as

$$\mathbf{K}_{i,j} \overset{\text{def.}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$$

Indeed one has that using the definition above

$$\mathbf{P}_\varepsilon = \mathrm{Proj}^{\mathbf{KL}}_{\mathbf{U}(\mathbf{a},\mathbf{b})}(\mathbf{K}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathbf{KL}(\mathbf{P}|\mathbf{K}). \tag{1.44}$$

Remark 10 (General formulation). One can consider arbitrary measures by replacing the discrete entropy by the relative entropy with respect to the product measure $\mathrm{d}\alpha\otimes\mathrm{d}\beta(x,y) \overset{\text{def.}}{=} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$, and propose a regularized counterpart to (**??**) using

$$\mathcal{L}_c^\varepsilon(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{X\times Y} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) \tag{1.45}$$

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

$$\begin{aligned}\mathrm{KL}(\pi|\xi) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\Big)\mathrm{d}\pi(x,y) + \\ \int_{\mathcal{X}\times\mathcal{Y}} (\mathrm{d}\xi(x,y) - \mathrm{d}\pi(x,y)),\end{aligned} \tag{1.46}$$

and by convention $\mathrm{KL}(\pi|\xi) = +\infty$ if π does not have a density $\frac{\mathrm{d}\pi}{\mathrm{d}\xi}$ with respect to ξ. It is important to realize that the reference measure $\alpha \otimes \beta$ chosen in (**??**) to define the entropic regularizing term $\mathrm{KL}(\cdot|\alpha\otimes\beta)$ plays no specific role, only its support matters.

Formula (**??**) can be re-factored as a projection problem

$$\min_{\pi\in\mathcal{U}(\alpha,\beta)} \mathrm{KL}(\pi|\mathcal{K}) \tag{1.47}$$

where $\mathcal{K}$ is the Gibbs distributions $\mathrm{d}\mathcal{K}(x,y) \overset{\text{def.}}{=} e^{-\frac{c(x,y)}{\varepsilon}}\mathrm{d}\mu(x)\mathrm{d}\nu(y)$. This problem is often referred to as the "static Schrödinger problem" [**?**, **?**], since it was initially considered by Schrödinger in statistical physics [**?**]. As $\varepsilon \to 0$, the unique solution to (**??**) converges to the maximum entropy solution to (**??**), see [**?**, **?**]. §**??** details an alternate "dynamic" formulation of the Schrödinger problem over the space of paths connecting the points of two measures.

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

Proposition 7. *The solution to* (**??**) *is unique and has the form*

$$\forall\,(i,j) \in [\![n]\!]\times[\![m]\!], \quad \mathbf{P}_{i,j} = \mathbf{u}_i\mathbf{K}_{i,j}\mathbf{v}_j \tag{1.48}$$

for two (unknown) scaling variable $(\mathbf{u},\mathbf{v}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$.

Proof. Introducing two dual variables $\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m$ for each marginal constraint, the Lagrangian of (**??**) reads

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

Considering first order conditions, we have

$$\frac{\partial\mathcal{E}(\mathbf{P},\mathbf{f},\mathbf{g})}{\partial\mathbf{P}_{i,j}} = \mathbf{C}_{i,j} - \varepsilon\log(\mathbf{P}_{i,j}) - \mathbf{f}_i - \mathbf{g}_j.$$

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

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

$$\mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})\mathbb{1}_m = \mathbf{a}, \quad \text{and} \quad \mathrm{diag}(\mathbf{v})\mathbf{K}^\top\,\mathrm{diag}(\mathbf{u})\mathbb{1}_n = \mathbf{b}, \tag{1.49}$$

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

$$\mathbf{u}\odot(\mathbf{K}\mathbf{v}) = \mathbf{a} \quad \text{and} \quad \mathbf{v}\odot(\mathbf{K}^{\mathrm{T}}\mathbf{u}) = \mathbf{b} \tag{1.50}$$

where $\odot$ corresponds to entry-wise multiplication of vectors. That problem is known in the numerical analysis community as the matrix scaling problem (see [**?**] and references therein). An intuitive way to try to solve

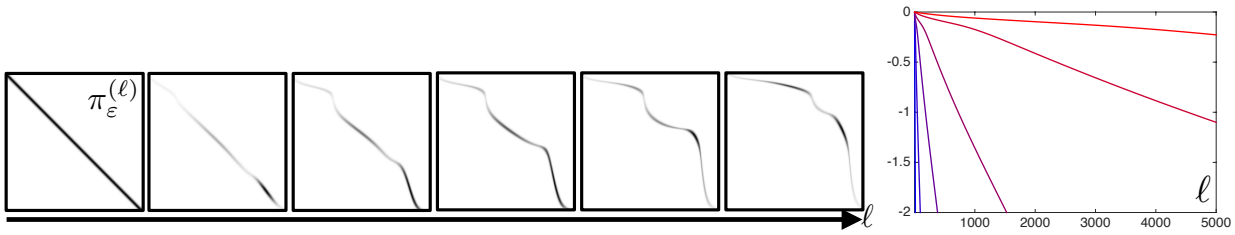


Figure 1.14: Left: evolution of the coupling $\pi_\varepsilon^\ell = \mathrm{diag}(\mathbf{U}^{(\ell)})\mathbf{K}\,\mathrm{diag}(\mathbf{V}^{(\ell)})$ computed at iteration ℓ of Sinkhorn's iterations, for 1-D densities. Right: impact of ε the convergence rate of Sinkhorn, as measured in term of marginal constraint violation $\log(\|\pi_\varepsilon^\ell \mathbb{1}_m - \mathbf{b}\|_1)$.

these equations is to solve them iteratively, by modifying first $\mathbf{u}$ so that it satisfies the left-hand side of Equation (**??**) and then $\mathbf{v}$ to satisfy its right-hand side. These two updates define Sinkhorn's algorithm:

$$\mathbf{u}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \text{and} \quad \mathbf{v}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(\ell+1)}}, \tag{1.51}$$

initialized with an arbitrary positive vector $\mathbf{v}^{(0)} = \mathbb{1}_m$. 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 $\mathbf{u}, \mathbf{v}$, since $\mathbf{u}, \mathbf{v}$ are only defined up to a multiplicative constant (if $\mathbf{u}, \mathbf{v}$ satisfy (**??**) then so do $\lambda\mathbf{u}, \mathbf{v}/\lambda$ for any $\lambda > 0$). It turns out however that these iterations converge (see Remark **??** for a justification using iterative projections, and Remark **??** for a strict contraction result) and all result in the same optimal coupling $\mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$. Figure **??**, top row, shows the evolution of the coupling $\mathrm{diag}(\mathbf{U}^{(\ell)})\mathbf{K}\,\mathrm{diag}(\mathbf{V}^{(\ell)})$ computed by Sinkhorn iterations. It evolves from the Gibbs kernel $\mathbf{K}$ towards the optimal coupling solving (**??**) by progressively shifting the mass away from the diagonal.

Remark 11 (Relation with iterative projections). Denoting

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

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

$$\mathbf{P}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Proj}_{\mathcal{C}_{\mathbf{a}}^1}^{\mathbf{KL}}(\mathbf{P}^{(\ell)}) \quad \text{and} \quad \mathbf{P}^{(\ell+2)} \overset{\text{def.}}{=} \mathrm{Proj}_{\mathcal{C}_{\mathbf{b}}^2}^{\mathbf{KL}}(\mathbf{P}^{(\ell+1)}). \tag{1.52}$$

Since the sets $\mathcal{C}_{\mathbf{a}}^1$ and $\mathcal{C}_{\mathbf{b}}^2$ are affine, these iterations are known to converge to the solution of (**??**), see [**?**]. These iterate are equivalent to Sinkhorn iterations (**??**) since defining

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

one has

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

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

Remark 12 (Hilbert metric). As initially explained by [**?**], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert projective metric on $\mathbb{R}_{+,*}^n$ (positive vectors), defined as

$$\forall\,(\mathbf{u}, \mathbf{u}') \in (\mathbb{R}_{+,*}^n)^2, \quad d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') \overset{\text{def.}}{=} \log \max_{i,i'} \frac{\mathbf{u}_i \mathbf{u}'_{i'}}{\mathbf{u}_{i'} \mathbf{u}'_i}.$$

This can be shows to be a distance on the projective cone $\mathbb{R}^n_{+,*}/\sim$, where $\mathbf{u}\sim\mathbf{u}'$ means that $\exists s>0, \mathbf{u}=s\mathbf{u}'$ (the vector are equal up to rescaling, hence the naming "projective"). This means that $d_{\mathcal{H}}$ satisfies the triangular inequality and $d_{\mathcal{H}}(\mathbf{u},\mathbf{u}')=0$ if and only if $\mathbf{u}\sim\mathbf{u}'$. This is a projective version of Hilbert's original distance on bounded open convex sets [**?**]. The projective cone $\mathbb{R}^n_{+,*}/\sim$ is a complete metric space for this distance. It was introduced independently by [**?**] and [**?**] to provide a quantitative proof of Perron-Frobenius theorem, which, as explained in Remark **??** is linked to a local linearization of Sinkhorn's iterates. They proved the following fundamental theorem, which shows that a positive matrix is a strict contraction on the cone of positive vectors.

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

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

Remark 13 (Global convergence). The following theorem, proved by [**?**], makes use of this Theorem **??** to show the linear convergence of Sinkhorn's iterations.

Theorem 3. *One has* $(\mathbf{u}^{(\ell)},\mathbf{v}^{(\ell)})\to(\mathbf{u}^\star,\mathbf{v}^\star)$ *and*

$$d_{\mathcal{H}}(\mathbf{u}^{(\ell)},\mathbf{u}^\star)=O(\lambda(\mathbf{K})^{2\ell}),\quad d_{\mathcal{H}}(\mathbf{v}^{(\ell)},\mathbf{v}^\star)=O(\lambda(\mathbf{K})^{2\ell}). \tag{1.53}$$

One also has

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell)},\mathbf{u}^\star)&\leqslant\frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell)}\mathbb{1}_m,\mathbf{a})}{1-\lambda(\mathbf{K})} \\ d_{\mathcal{H}}(\mathbf{v}^{(\ell)},\mathbf{v}^\star)&\leqslant\frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell),\top}\mathbb{1}_n,\mathbf{b})}{1-\lambda(\mathbf{K})} \end{aligned} \tag{1.54}$$

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

$$\|\log(\mathbf{P}^{(\ell)})-\log(\mathbf{P}^\star)\|_\infty\leqslant d_{\mathcal{H}}(\mathbf{u}^{(\ell)},\mathbf{u}^\star)+d_{\mathcal{H}}(\mathbf{v}^{(\ell)},\mathbf{v}^\star) \tag{1.55}$$

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

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

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

This shows that

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

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

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

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

The bound (**??**) shows that some error measures on the marginal constraints violation, for instance $\|\mathbf{P}^{(\ell)}\mathbb{1}_m - \mathbf{a}\|_1$ and $\|{\mathbf{P}^{(\ell)}}^{\mathrm{T}}\mathbb{1}_n - \mathbf{b}\|_1$, are useful stopping criteria to monitor the convergence.

Figure **??**, bottom row, highlights this linear rate on the constraint violation, and shows how this rate degrades as $\varepsilon \to 0$. These results are proved in [**?**] and are tightly connected to nonlinear Perron-Frobenius Theory [**?**]. Perron-Frobenius theory corresponds to the linearization of the iterations, see (**??**). This convergence analysis is extended in [**?**], who shows that each iteration of Sinkhorn increases the permanent of the scaled coupling matrix.

Regularized Dual and Log-domain Computations The following proposition details the dual problem associated to (**??**).

Proposition 8. *One has*

$$\mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \max_{\mathbf{f}\in\mathbb{R}^n,\mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f},\, \mathbf{a}\rangle + \langle \mathbf{g},\, \mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon},\, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle. \tag{1.56}$$

The optimal $(\mathbf{f},\mathbf{g})$ *are linked to scalings* $(\mathbf{u},\mathbf{v})$ *appearing in* (**??**) *through*

$$(\mathbf{u},\mathbf{v}) = (e^{\mathbf{f}/\varepsilon}, e^{\mathbf{g}/\varepsilon}). \tag{1.57}$$

Proof. We start from the end of the proof of Proposition **??**, which links the optimal primal solution $\mathbf{P}$ and dual multipliers $\mathbf{f}$ and $\mathbf{g}$ for the marginal constraints as $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon}e^{-\mathbf{C}_{i,j}/\varepsilon}e^{\mathbf{g}_j/\varepsilon}$. Substituting in the Lagrangian $\mathcal{E}(\mathbf{P},\mathbf{f},\mathbf{g})$ of Equation (**??**) the optimal $\mathbf{P}$ as a function of $\mathbf{f}$ and $\mathbf{g}$, we obtain that the Lagrange dual function equals

$$\mathbf{f},\mathbf{g} \mapsto \langle e^{\mathbf{f}/\varepsilon},\, (\mathbf{K}\odot\mathbf{C})\, e^{\mathbf{g}/\varepsilon}\rangle - \varepsilon\mathbf{H}(\mathrm{diag}(e^{\mathbf{f}/\varepsilon})\mathbf{K}\,\mathrm{diag}(e^{\mathbf{g}/\varepsilon})). \tag{1.58}$$

The entropy of $\mathbf{P}$ scaled by ε, namely $\varepsilon\langle \mathbf{P},\, \log\mathbf{P} - \mathbb{1}_{n\times m}\rangle$ can be stated explicitly as a function of $\mathbf{f},\mathbf{g},\mathbf{C}$

$$\begin{aligned}&\langle \mathrm{diag}(e^{\mathbf{f}/\varepsilon})\mathbf{K}\,\mathrm{diag}(e^{\mathbf{g}/\varepsilon}),\, \mathbf{f}\mathbb{1}_m{}^{\mathrm{T}} + \mathbb{1}_n\mathbf{g}^{\mathrm{T}} - \mathbf{C} - \varepsilon\mathbb{1}_{n\times m}\rangle\\ &= -\langle e^{\mathbf{f}/\varepsilon},\, (\mathbf{K}\odot\mathbf{C})\, e^{\mathbf{g}/\varepsilon}\rangle + \langle \mathbf{f},\, \mathbf{a}\rangle + \langle \mathbf{g},\, \mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon},\, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle\end{aligned}$$

therefore, the first term in (**??**) cancels out with the first term in the entropy above. The remaining times are those displayed in (**??**). □

Remark 14. Dual for generic measures For generic (non-necessarily discrete) input measures (α,β), the dual problem (**??**) reads

$$\sup_{f,g\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(x)\mathrm{d}\beta(x) - \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{-c(x,y)+f(x)+g(y)}{\varepsilon}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$$

This corresponds to a smoothing of the constraint $\mathcal{R}(c)$ appearing in the original problem (**??**), which is retrieved in the limit $\varepsilon\to 0$. Proving existence (*i.e.* the sup is actually a max) of these Kantorovich potentials (f,g) in the case of entropic transport is less easy than for classical OT (because one cannot use c-transform and potentials are not automatically Lipschitz). Proof of existence can be done using the convergence of Sinkhorn iterations, see [**?**] for more details.

Remark 15 (Sinkhorn as a Block Coordinate Ascent on the Dual Problem). A simple approach to solve the unconstrained maximization problem (**??**) is to use an exact *block coordinate ascent* strategy, namely to update alternatively $\mathbf{f}$ and $\mathbf{g}$ to cancel their gradients with respect to the objective of (**??**). Indeed, one can easily notice that, writing $Q(\mathbf{f},\mathbf{g})$ for the objective of (**??**) that

$$\nabla|_{\mathbf{f}}\, Q(\mathbf{f},\mathbf{g}) = \mathbf{a} - e^{\mathbf{f}/\varepsilon}\odot\left(\mathbf{K}e^{\mathbf{g}/\varepsilon}\right), \tag{1.59}$$

$$\nabla|_{\mathbf{g}}\, Q(\mathbf{f},\mathbf{g}) = \mathbf{b} - e^{\mathbf{g}/\varepsilon}\odot\left(\mathbf{K}^T e^{\mathbf{f}/\varepsilon}\right). \tag{1.60}$$

Block coordinate ascent can therefore be implemented in a closed form by applying successively the following updates, starting from any arbitrary $\mathbf{g}^{(0)}$, for $l \geqslant 0$,

$$\mathbf{f}^{(\ell+1)} = \varepsilon \log \mathbf{a} - \varepsilon \log \left(\mathbf{K} e^{\mathbf{g}^{(\ell)}/\varepsilon} \right), \tag{1.61}$$

$$\mathbf{g}^{(\ell+1)} = \varepsilon \log \mathbf{b} - \varepsilon \log \left(\mathbf{K}^{\mathrm{T}} e^{\mathbf{f}^{(\ell+1)}/\varepsilon} \right). \tag{1.62}$$

Such iterations are mathematically equivalent to the Sinkhorn iterations (**??**) when considering the primal-dual relations highlighted in (**??**). Indeed, we recover that at any iteration

$$(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)}) = \varepsilon(\log(\mathbf{u}^{(\ell)}), \log(\mathbf{v}^{(\ell)})).$$

Remark 16 (Soft-min rewriting). Iterations (**??**) and (**??**) can be given an alternative interpretation, using the following notation. Given a vector $\mathbf{z}$ of real numbers we write $\min_\varepsilon \mathbf{z}$ for the *soft-minimum* of its coordinates, namely

$$\min_\varepsilon \mathbf{z} = -\varepsilon \log \sum_i e^{-\mathbf{z}_i/\varepsilon}.$$

Note that $\min_\varepsilon (\mathbf{z})$ converges to $\min \mathbf{z}$ for any vector $\mathbf{z}$ as $\varepsilon \to 0$. Indeed, $\min_\varepsilon$ can be interpreted as a differentiable approximation of the min function. Using these notations, Equations (**??**) and (**??**) can be rewritten

$$(\mathbf{f}^{(\ell+1)})_i = \min_\varepsilon \, (\mathbf{C}_{ij} - \mathbf{g}_j^{(\ell)})_j + \varepsilon \log \mathbf{a}_i, \tag{1.63}$$

$$(\mathbf{g}^{(\ell+1)})_j = \min_\varepsilon \, (\mathbf{C}_{ij} - \mathbf{f}_i^{(\ell)})_i + \varepsilon \log \mathbf{b}_j. \tag{1.64}$$

Here the term $\min_\varepsilon \, (\mathbf{C}_{ij} - \mathbf{g}_j^{(\ell)})_j$ denotes the soft-minimum of all values of the j-th column of matrix $(\mathbf{C} - \mathbb{1}_n (\mathbf{g}^{(\ell)})^\top)$. To simplify notations, we introduce an operator that takes a matrix as input and outputs now a column vector of the soft-minimum values of its columns or rows. Namely, for any matrix $A \in \mathbb{R}^{n \times m}$, we define

$$\mathrm{Min}_\varepsilon^{\mathrm{row}} (\mathbf{A}) \stackrel{\text{def.}}{=} \left(\min_\varepsilon \, (\mathbf{A}_{i,j})_j \right)_i \in \mathbb{R}^n,$$

$$\mathrm{Min}_\varepsilon^{\mathrm{col}} (\mathbf{A}) \stackrel{\text{def.}}{=} \left(\min_\varepsilon \, (\mathbf{A}_{i,j})_i \right)_j \in \mathbb{R}^m.$$

Note that these operations are equivalent to the entropic c-transform introduced in §**??** (see in particular (**??**)). Using these notations, Sinkhorn's iterates read

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{row}} \, (\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)^{\mathrm{T}}}) + \varepsilon \log \mathbf{a}, \tag{1.65}$$

$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{col}} \, (\mathbf{C} - \mathbf{f}^{(\ell)} \mathbb{1}_m{}^{\mathrm{T}}) + \varepsilon \log \mathbf{b}. \tag{1.66}$$

Note that as $\varepsilon \to 0$, $\min_\varepsilon$ converges to min, but the iterations do not converge anymore in the limit $\varepsilon = 0$, because alternate minimization does not converge for constrained problems (which is the case for the unregularized dual (**??**)).

Remark 17 (Log-domain Sinkhorn). While mathematically equivalent to the Sinkhorn updates (**??**), iterations (**??**) and (**??**) suggest to use the *log-sum-exp* stabilization trick to avoid underflow for small values of ε. Writing $\underline{\mathrm{z}} = \min \mathbf{z}$, that trick suggests to evaluate $\min_\varepsilon \mathbf{z}$ as

$$\min_\varepsilon \mathbf{z} = \underline{\mathrm{z}} - \varepsilon \log \sum_i e^{-(\mathbf{z}_i - \underline{\mathrm{z}})/\varepsilon}. \tag{1.67}$$

Instead of substracting $\underline{\mathrm{z}}$ to stabilize the log domain iterations as in (**??**), one can actually substract the previously computed scalings. This leads to the following stabilized iteration

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{row}} \, (\mathbf{S}(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)})) - \mathbf{f}^{(\ell)} + \varepsilon \log(\mathbf{a}) \tag{1.68}$$

$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{col}} \, (\mathbf{S}(\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell)})) - \mathbf{g}^{(\ell)} + \varepsilon \log(\mathbf{b}), \tag{1.69}$$

where we defined

$$\mathbf{S}(\mathbf{f},\mathbf{g}) = \big(\mathbf{C}_{i,j} - \mathbf{f}_i - \mathbf{g}_j\big)_{i,j}.$$

In contrast to the original iterations (**??**), these log-domain iterations (**??**) and (**??**) are stable for arbitrary $\varepsilon > 0$, because the quantity $\mathbf{S}(\mathbf{f},\mathbf{g})$ stays bounded during the iterations. The downside is that it requires nm computations of exp at each step. Computing a $\mathrm{Min}_\varepsilon^{\mathrm{row}}$ or $\mathrm{Min}_\varepsilon^{\mathrm{col}}$ is typically substantially slower than matrix multiplications, and requires computing line by line soft-minima of matrices $\mathbf{S}$. There is therefore no efficient way to parallelize the application of Sinkhorn maps for several marginals simultaneously. In Euclidean domain of small dimension, it is possible to develop efficient multiscale solvers with a decaying ε strategy to significantly speed up the computation using sparse grids [**?**].

1.6 Extensions

Wasserstein Barycenters. Given input histogram $\{\mathbf{b}_s\}_{s=1}^S$, where $b_s \in \Sigma_{n_s}$, and weights $\lambda \in \Sigma_S$, a Wasserstein barycenter is computed by minimizing

$$\min_{\mathbf{a}\in\Sigma_n} \sum_{s=1}^S \lambda_s \mathrm{L}_{\mathbf{C}_s}(\mathbf{a},\mathbf{b}_s) \tag{1.70}$$

where the cost matrices $\mathbf{C}_s \in \mathbb{R}^{n\times n_s}$ need to be specified. A typical setup is "Eulerian", so that all the barycenters are defined on the same grid, $n_s = n$, $\mathbf{C}_s = \mathbf{C} = \mathbf{D}^p$ is set to be a distance matrix, so that one solves

$$\min_{\mathbf{a}\in\Sigma_n} \sum_{s=1}^S \lambda_s \,\mathrm{W}_p^p(\mathbf{a},\mathbf{b}_s).$$

This barycenter problem (**??**) was originally introduced by [**?**] following earlier ideas of [**?**]. They proved in particular uniqueness of the barycenter for $c(x,y) = \|x-y\|^2$ over $\mathcal{X} = \mathbb{R}^d$, if one of the input measure has a density with respect to the Lebesgue measure (and more generally under the same hypothesis as the one guaranteeing the existence of a Monge map, see Remark **??**).

The barycenter problem for histograms (**??**) is in fact a linear program, since one can look for the S couplings $(\mathbf{P}_s)_s$ between each input and the barycenter itself

$$\min_{\mathbf{a}\in\Sigma_n,(\mathbf{P}_s\in\mathbb{R}^{n\times n_s})_s} \left\{ \sum_{s=1}^S \lambda_s \langle \mathbf{P}_s,\, \mathbf{C}_s\rangle \;;\; \forall\, s, \mathbf{P}_s^\top \mathbb{1}_{n_s} = \mathbf{a}, \mathbf{P}_s^\top \mathbb{1}_n = \mathbf{b}_s \right\}.$$

Although this problem is an LP, its scale forbids the use generic solvers for medium scale problems. One can therefore resort to using first order methods such as subgradient descent on the dual [**?**].

Remark 18. Barycenter of arbitrary measures Given a set of input measure $(\beta_s)_s$ defined on some space $\mathcal{X}$, the barycenter problem becomes

$$\min_{\alpha\in\mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^S \lambda_s \mathcal{L}_c(\alpha,\beta_s). \tag{1.71}$$

In the case where $\mathcal{X} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^2$, [**?**] shows that if one of the input measures has a density, then this barycenter is unique. Problem (**??**) can be viewed as a generalization of the problem of computing barycenters of points $(x_s)_{s=1}^S \in \mathcal{X}^S$ to arbitrary measures. Indeed, if $\beta_s = \delta_{x_s}$ is a single Dirac mass, then a solution to (**??**) is $\delta_{x^\star}$ where $x^\star$ is a Fréchet mean solving (**??**). Note that for $c(x,y) = \|x-y\|^2$, the mean of the barycenter $\alpha^\star$ is necessarily the barycenter of the mean, *i.e.*

$$\int_{\mathcal{X}} x \mathrm{d}\alpha^\star(x) = \sum_s \lambda_s \int_{\mathcal{X}} x \mathrm{d}\alpha_s(x),$$

and the support of $\alpha^\star$ is located in the convex hull of the supports of the $(\alpha_s)_s$. The consistency of the approximation of the infinite dimensional optimization (**??**) when approximating the input distribution using discrete ones (and thus solving (**??**) in place) is studied in [**?**]. Let us also note that it is possible to re-cast (**??**) as a multi-marginal OT problem, see Remark **??**.

One can use entropic smoothing and approximate the solution of (**??**) using

$$\min_{\mathbf{a}\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}^\varepsilon_{\mathbf{C}_s}(\mathbf{a}, \mathbf{b}_s) \tag{1.72}$$

for some $\varepsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent [**?**]. An alternative is to use descent method (typically quasi-Newton) on the semi-dual [**?**], which is useful to integrate additional regularizations on the barycenter (e.g. to impose some smoothness). A simple but effective approach, as remarked in [**?**] is to rewrite (**??**) as a (weighted) KL projection problem

$$\min_{(\mathbf{P}_s)_s} \left\{ \sum_s \lambda_s \mathbf{KL}(\mathbf{P}_s|\mathbf{K}_s) \;;\; \forall\, s, {\mathbf{P}_s}^{\mathrm{T}}\mathbb{1}_m = \mathbf{b}_s,\; \mathbf{P}_1\mathbb{1}_1 = \ldots = \mathbf{P}_S\mathbb{1}_S \right\} \tag{1.73}$$

where we denoted $\mathbf{K}_s \overset{\text{def.}}{=} e^{-\mathbf{C}_s/\varepsilon}$. Here, the barycenter $\mathbf{a}$ is implicitly encoded in the row marginals of all the couplings $\mathbf{P}_s \in \mathbb{R}^{n\times n_s}$ as $\mathbf{a} = \mathbf{P}_1\mathbb{1}_1 = \ldots = \mathbf{P}_S\mathbb{1}_S$. As detailed in [**?**], one can generalize Sinkhorn to this problem, which also corresponds to iterative projection. This can also be seen as a special case of the generalized Sinkhorn detailed in §**??**. The optimal couplings $(\mathbf{P}_s)_s$ solving (**??**) are computed in scaling form as

$$\mathbf{P}_s = \mathrm{diag}(\mathbf{u}_s)\mathbf{K}\,\mathrm{diag}(\mathbf{v}_s), \tag{1.74}$$

and the scalings are sequentially updated as

$$\forall\, s \in [\![1, S]\!], \quad \mathbf{v}_s^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}_s}{\mathbf{K}_s^{\mathrm{T}}\mathbf{u}_s^{(\ell)}}, \tag{1.75}$$

$$\forall\, s \in [\![1, S]\!], \quad \mathbf{u}_s^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{a}^{(\ell+1)}}{\mathbf{K}_s\mathbf{v}_s^{(\ell+1)}}, \tag{1.76}$$

$$\text{where} \quad \mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \prod_s (\mathbf{K}_s\mathbf{v}_s^{(\ell+1)})^{\lambda_s}. \tag{1.77}$$

An alternative way to derive these iterations is to perform alternate minimization on the variables of a dual problem, which detailed in the following proposition.

Proposition 9. *The optimal $(\mathbf{u}_s, \mathbf{v}_s)$ appearing in (**??**) can be written as $(\mathbf{u}_s, \mathbf{v}_s) = (e^{\mathbf{f}_s/\varepsilon}, e^{\mathbf{g}_s/\varepsilon})$ where $(\mathbf{f}_s, \mathbf{g}_s)_s$ are the solutions of the following program (whose value matches the one of (**??**))*

$$\max_{(\mathbf{f}_s,\mathbf{g}_s)_s} \left\{ \sum_s \lambda_s \left(\langle \mathbf{g}_s,\, \mathbf{b}_s\rangle - \varepsilon \langle \mathbf{K}_s e^{\mathbf{g}_s/\varepsilon},\, e^{\mathbf{f}_s/\varepsilon}\rangle \right) \;;\; \sum_s \lambda_s \mathbf{f}_s = 0 \right\}. \tag{1.78}$$

Proof. Introducing Lagrange multipliers in (**??**) leads to

$$\min_{(\mathbf{P}_s)_s, \mathbf{a}} \max_{(\mathbf{f}_s,\mathbf{g}_s)_s} \sum_s \lambda_s \Big(\varepsilon \mathbf{KL}(\mathbf{P}_s|\mathbf{K}_s) + \langle \mathbf{a} - \mathbf{P}_s\mathbb{1}_m,\, \mathbf{f}_s\rangle + \langle \mathbf{b}_s - {\mathbf{P}_s}^{\mathrm{T}}\mathbb{1}_m,\, \mathbf{g}_s\rangle \Big).$$

Strong duality holds, so that one can exchange the min and the max, and gets

$$\max_{(\mathbf{f}_s,\mathbf{g}_s)_s} \sum_s \lambda_s \left(\langle \mathbf{g}_s,\, \mathbf{b}_s\rangle + \min_{\mathbf{P}_s} \varepsilon \mathbf{KL}(\mathbf{P}_s|\mathbf{K}_s) - \langle \mathbf{P}_s,\, \mathbf{f}_s \oplus \mathbf{g}_s\rangle \right) + \min_{\mathbf{a}} \langle \sum_s \lambda_s \mathbf{f}_s,\, \mathbf{a}\rangle.$$

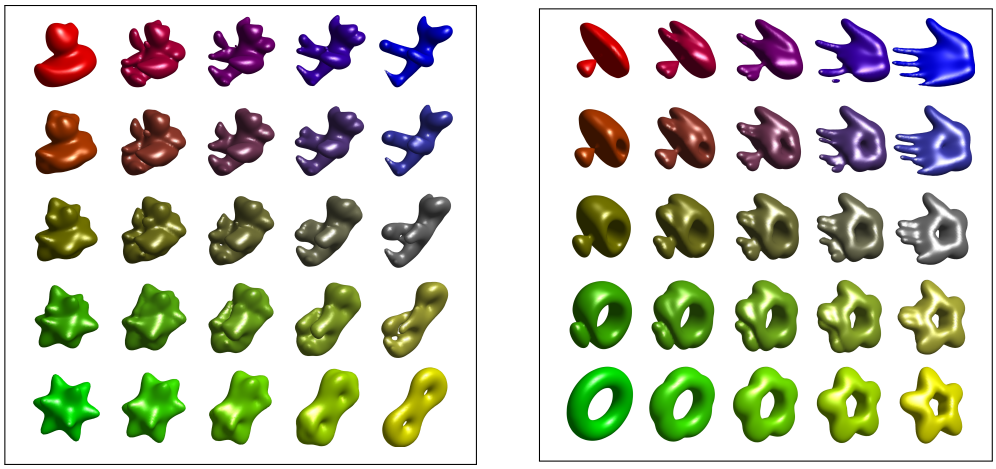


Figure 1.15: Barycenters between 4 input 3-D shapes using entropic regularization (**??**). The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square. Shapes are represented as measures that are uniform within the boundaries of the shape and null outside.

The explicit minimization on $\mathbf{a}$ gives the constraint $\sum_s \lambda_s \mathbf{f}_s = 0$ together with

$$\max_{(\mathbf{f}_s,\mathbf{g}_s)_s} \sum_s \lambda_s \langle \mathbf{g}_s, \mathbf{b}_s \rangle - \varepsilon \mathbf{KL}^* \left(\frac{\mathbf{f}_s \oplus \mathbf{g}_s}{\varepsilon} | \mathbf{K}_s \right)$$

where $\mathbf{KL}^*(\cdot|\mathbf{K}_s)$ is the Legendre transform (**??**) of the function $\mathbf{KL}^*(\cdot|\mathbf{K}_s)$. This Legendre transform reads

$$\mathbf{KL}^*(\mathbf{U}|\mathbf{K}) = \sum_{i,j} \mathbf{K}_{i,j}(e^{\mathbf{U}_{i,j}} - 1), \tag{1.79}$$

which shows the desired formula. To show (**??**), since this function is separable, one needs to compute

$$\forall\, (u,k) \in \mathbb{R}_+^2, \quad \mathbf{KL}^*(u|k) \overset{\text{def.}}{=} \max_r ur - (r \log(r/k) - r + k)$$

whose optimality condition reads $u = \log(r/k)$, i.e. $r = ke^u$, hence the result. □

Minimizing (**??**) with respect to each $\mathbf{g}_s$, while keeping all the other variable fixed, is obtained in closed form by (**??**). Minimizing (**??**) with respect to all the $(\mathbf{f}_s)_s$ requires to solve for $\mathbf{a}$ using (**??**) and leads to the expression (**??**).

Figures **??** and **??** show applications to 2-D and 3-D shapes interpolation. Figure **??** shows a computation of barycenters on a surface, where the ground cost is the square of the geodesic distance. For this figure, the computations are performed using the geodesic in heat approximation detailed in Remark **??**. We refer to [**?**] for more details and other applications to computer graphics and imaging sciences.

Wasserstein Loss. In statistics, text processing or imaging, one must usually compare a probability distribution β arising from measurements to a model, namely a parameterized family of distributions $\{\alpha_\theta, \theta \in \Theta\}$ where Θ is a subset of an Euclidean space. Such a comparison is done through a "loss" or a "fidelity" term, which, in this section, is the Wasserstein distance. In the simplest scenario, the computation of a suitable parameter θ is obtained by minimizing directly

$$\min_{\theta \in \Theta} \mathcal{E}(\theta) \overset{\text{def.}}{=} \mathcal{L}_c(\alpha_\theta, \beta). \tag{1.80}$$

Of course, one can consider more complicated problems: for instance, the barycenter problem described in §**??** consists in a sum of such terms. However, most of these more advanced problems can be usually

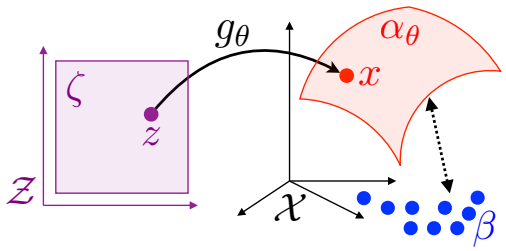


Figure 1.16: Schematic display of the density fitting problem **??**.

solved by adapting tools defined for basic case: either using the chain rule to compute explicitly derivatives, or using automatic differentiation.

The Wasserstein distance between two histograms or two densities is convex with respect to these inputs, as shown by (**??**) and (**??**) respectively. Therefore, when the parameter θ is itself a histogram, namely $\Theta = \Sigma_n$ and $\alpha_\theta = \theta$, or more generally when θ describes K weights in the simplex, $\Theta = \Sigma_K$, and $\alpha_\theta = \sum_{i=1}^K \theta_i \alpha_i$ is a convex combination of known atoms $\alpha_1, \ldots, \alpha_K$ in Σ_N, Problem (**??**) remains convex (the first case corresponds to the barycenter problem, the second to one iteration of the dictionary learning problem with a Wasserstein loss [**?**]). However, for more general parameterizations $\theta \mapsto \alpha_\theta$, Problem (**??**) is in general not convex.

A practical problem of paramount importance in statistic and machine learning is density fitting. Given some discrete samples $(x_i)_{i=1}^n \subset \mathcal{X}$ from some unknown distribution, the goal is to fit a parametric model $\theta \mapsto \alpha_\theta \in \mathcal{M}(\mathcal{X})$ to the observed empirical input measure β

$$\min_{\theta \in \Theta} \mathcal{L}(\alpha_\theta, \beta) \quad \text{where} \quad \beta = \frac{1}{n} \sum_i \delta_{x_i}, \tag{1.81}$$

where $\mathcal{L}$ is some "loss" function between a discrete and a "continuous" (arbitrary) distribution (see Figure **??**).

In the case where α_θ as a densify $\rho_\theta \overset{\text{def.}}{=} \rho_{\alpha_\theta}$ with respect to the Lebesgue measure (or any other fixed reference measure), the maximum likelihood estimator (MLE) is obtained by solving

$$\min_\theta \mathcal{L}_{\text{MLE}}(\alpha_\theta, \beta) \overset{\text{def.}}{=} -\sum_i \log(\rho_\theta(x_i)).$$

This corresponds to using an empirical counterpart of a Kullback-Leibler loss since, assuming the x_i are i.i.d. samples of some $\bar{\beta}$, then

$$\mathcal{L}_{\text{MLE}}(\alpha, \beta) \overset{n \to +\infty}{\longrightarrow} \text{KL}(\alpha | \bar{\beta})$$

This MLE approach is known to lead to optimal estimation procedures in many cases (see for instance [**?**]). However, it fails to work when estimating singular distributions, typically when the α_θ does not has a density (so that $\mathcal{L}_{\text{MLE}}(\alpha_\theta, \beta) = +\infty$) or when $(x_i)_i$ are samples from some singular $\bar{\beta}$ (so that the α_θ should share the same support as β for $\text{KL}(\alpha|\bar{\beta})$ to be finite, but this support is usually unknown). Another issue is that in several cases of practical interest, the density ρ_θ is inaccessible (or too hard to compute).

A typical setup where both problems (singular and unknown densities) occur is for so-called generative models, where the parametric measure is written as a push-forward of a fixed reference measure $\zeta \in \mathcal{M}(\mathcal{Z})$

$$\alpha_\theta = h_{\theta,\sharp}\zeta \quad \text{where} \quad h_\theta : \mathcal{Z} \to \mathcal{X}$$

where the push-forward operator is introduced in Definition **??**. The space $\mathcal{Z}$ is usually low-dimensional, so that the support of α_θ is localized along a low-dimensional "manifold" and the resulting density is highly singular (it does not have a density with respect to Lebesgue measure). Furthermore, computing this density is usually intractable, while generating i.i.d. samples from α_θ is achieved by computing $x_i = h_\theta(z_i)$ where $(z_i)_i$ are i.i.d. samples from ζ.

In order to cope with such difficult scenario, one has to use weak metrics in place of the MLE functional $\mathcal{L}_{\text{MLE}}$, which needs to be written in dual form as

$$\mathcal{L}(\alpha,\beta) \stackrel{\text{def.}}{=} \max_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \left\{ \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{X}} g(x)\mathrm{d}\beta(x) \;;\; (f,g)\in\mathcal{R} \right\}. \tag{1.82}$$

Dual norms exposed in §?? correspond to imposing $\mathcal{R} = \{(f,-f) \;;\; f\in B\}$, while optimal transport (??) sets $\mathcal{R} = \mathcal{R}(c)$ as defined in (??).

For a fixed θ, evaluating the energy to be minimized in (??) using such a loss function corresponds to solving a semi-discrete optimal transport, which is the focus of Chapter ??. Minimizing the energy with respect to θ is much more involved, and is typically highly non-convex.

The class of estimators obtained using $\mathcal{L} = \mathcal{L}_c$, often called "Minimum Kantorovitch Estimators" (MKE), was initially introduced in [?], see also [?].

Gromov-Wasserstein. Optimal transport needs a ground cost $\mathbf{C}$ to compare histograms $(\mathbf{a},\mathbf{b})$, it can thus not be used if the histograms are not defined on the same underlying space, or if one cannot pre-register these spaces to define a ground cost. To address this issue, one can instead only assume a weaker assumption, namely that one has at its disposal two matrices $\mathbf{D}\in\mathbb{R}^{n\times n}$ and $\mathbf{D}'\in\mathbb{R}^{m\times m}$ that represent some relationship between the points on which the histograms are defined. A typical scenario is when these matrices are (power of) distance matrices. The Gromov-Wasserstein problem reads

$$\mathrm{GW}((\mathbf{a},\mathbf{D}),(\mathbf{b},\mathbf{D}'))^2 \stackrel{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}) \stackrel{\text{def.}}{=} \sum_{i,j,i',j'} |\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|^2 \mathbf{P}_{i,j}\mathbf{P}_{i',j'}. \tag{1.83}$$

This is a non-convex problem, which can be recast as a Quadratic Assignment Problem (QAP) [?] and is in full generality NP-hard to solve for arbitrary inputs. It is in fact equivalent to a graph matching problem [?] for a particular cost.

One can show that GW satisfies the triangular inequality, and in fact it defines a distance between metric spaces equipped with a probability distribution (here assumed to be discrete in definition (??)) up to isometries preserving the measures. This distance was introduced and studied in details by Memoli in [?]. An in-depth mathematical exposition (in particular, its geodesic structure and gradient flows) is given in [?]. See also [?] for applications in computer vision. This distance is also tightly connected with the Gromov-Hausdorff distance [?] between metric spaces, which have been used for shape matching [?, ?].

Remark 19. Gromov-Wasserstein distance The general setting corresponds to computing couplings between metric measure spaces $(\mathcal{X},d_{\mathcal{X}},\alpha_{\mathcal{X}})$ and $(\mathcal{Y},d_{\mathcal{Y}},\alpha_{\mathcal{Y}})$ where $(d_{\mathcal{X}},d_{\mathcal{Y}})$ are distances and $(\alpha_{\mathcal{X}},\alpha_{\mathcal{Y}})$ are measures on their respective spaces. One defines

$$\mathcal{GW}((\alpha_{\mathcal{X}},d_{\mathcal{X}}),(\alpha_{\mathcal{Y}},d_{\mathcal{Y}}))^2 \stackrel{\text{def.}}{=} \min_{\pi\in\mathbf{U}(\alpha_{\mathcal{X}},\alpha_{\mathcal{Y}})} \int_{\mathcal{X}^2\times\mathcal{Y}^2} |d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')|^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y'). \tag{1.84}$$

$\mathcal{GW}$ defines a distance between metric measure spaces up to isometries, where one says that $(\alpha_{\mathcal{X}},d_{\mathcal{X}})$ and $(\alpha_{\mathcal{Y}},d_{\mathcal{Y}})$ are isometric if there exists $\varphi:\mathcal{X}\to\mathcal{Y}$ such that $\varphi_\sharp\alpha_{\mathcal{X}} = \alpha_{\mathcal{Y}}$ and $d_{\mathcal{Y}}(\varphi(x),\varphi(x')) = d_{\mathcal{X}}(x,x')$.

Remark 20. Gromov-Wasserstein geodesics The space of metric spaces (up to isometries) endowed with this $\mathcal{GW}$ distance (??) has a geodesic structure. [?] shows that the geodesic between $(\mathcal{X}_0,d_{\mathcal{X}_0},\alpha_0)$ and $(\mathcal{X}_1,d_{\mathcal{X}_1},\alpha_1)$ can be chosen to be $t\in[0,1]\mapsto(\mathcal{X}_0\times\mathcal{X}_1,d_t,\pi^\star)$ where $\pi^\star$ is a solution of (??) and for all $((x_0,x_1),(x_0',x_1'))\in(\mathcal{X}_0\times\mathcal{X}_1)^2$,

$$d_t((x_0,x_1),(x_0',x_1')) \stackrel{\text{def.}}{=} (1-t)d_{\mathcal{X}_0}(x_0,x_0') + t d_{\mathcal{X}_1}(x_1,x_1').$$

This formula allows one to define and analyze gradient flows which minimize functionals involving metric spaces, see [?]. It is however difficult to handle numerically, because it involves computations over the product space $\mathcal{X}_0\times\mathcal{X}_1$. A heuristic approach is used in [?] to define geodesics and barycenters of metric measure spaces while imposing the cardinality of the involved spaces and making use of the entropic smoothing (??) detailed below.

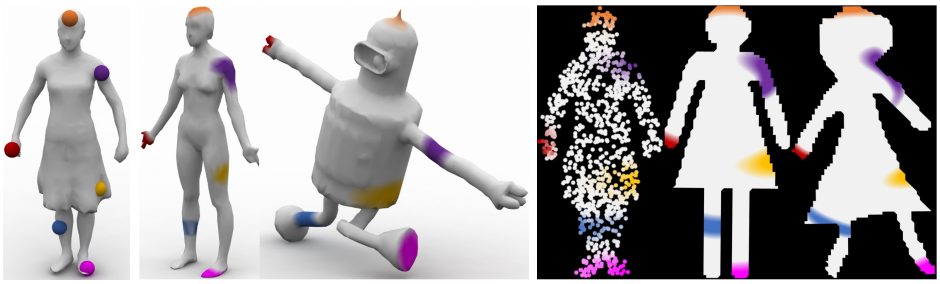


Figure 1.17: Example of fuzzy correspondences computed by solving GW problem (**??**) with Sinkhorn iterations (**??**). Extracted from [**?**].

To approximate the computation of GW, and to help convergence of minimization schemes to better minima, one can consider the entropic regularized variant

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}) - \varepsilon\mathbf{H}(\mathbf{P}). \tag{1.85}$$

As proposed initially in [**?**, **?**], and later revisited in [**?**] for applications in graphics, one can use iteratively Sinkhorn's algorithm to progressively compute a stationary point of (**??**). Indeed, successive linearizations of the objective function lead to consider the succession of updates

$$\mathbf{P}^{(\ell+1)} \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}^{(\ell)} \rangle - \varepsilon H(\mathbf{P}) \quad \text{where} \tag{1.86}$$

$$\mathbf{C}^{(\ell)} \overset{\text{def.}}{=} \nabla\mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}^{(\ell)}) = -\mathbf{D}'^{\mathrm{T}}\mathbf{P}^{(\ell)}\mathbf{D},$$

which can be interpreted as a mirror-descent scheme [**?**]. Each update can thus be solved using Sinkhorn iterations (**??**) with cost $\mathbf{C}^{(\ell)}$. Figure (**??**) illustrates the use of this entropic Gromov-Wasserstein to compute soft maps between domains.

Bibliography

[1] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

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

[4] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[5] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[6] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[7] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[8] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[9] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[10] Philippe G Ciarlet. Introduction à l'analyse numérique matricielle et à l'optimisation. 1982.

[11] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[12] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

[13] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[14] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[15] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[16] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[17] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[18] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[19] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[20] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier*. Ellipses, 2004.

[21] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.

[22] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[23] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging*. Springer, 2009.

[24] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[25] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis*. Cambridge university press, 2015.