1

Introduction

The shortest path principle guides most decisions in life and sciences: When a commodity, a person or a single bit of information is available at a given point and needs to be sent at a target point, one should favor using the least possible effort. This is typically reached by moving an item along a straight line when in the plane or along geodesic curves in more involved metric spaces. The theory of optimal transport generalizes that intuition in the case where, instead of moving only one item at a time, one is concerned with the problem of moving simultaneously several items (or a continuous distribution thereof) from one configuration onto another. As schoolteachers might attest, planning the transportation of a group of individuals, with the constraint that they reach a given target configuration upon arrival, is substantially more involved than carrying it out for a single individual. Indeed, thinking in terms of groups or distributions requires a more advanced mathematical formalism which was first hinted at in the seminal work of Monge (1781). Yet, no matter how complicated that formalism might look at first sight, that problem has deep and concrete connections with our daily life. Transportation, be it of people, commodities or information, very rarely involves moving only one item. All major economic problems, in logistics, production

planning or network routing, involve moving distributions, and that thread appears in all of the seminal references on optimal transport. Indeed Tolstoı (1930), Hitchcock (1941) and Kantorovich (1942) were all guided by practical concerns. It was only a few years later, mostly after the 1980s, that mathematicians discovered, thanks to the works of Brenier (1991) and others, that this theory provided a fertile ground for research, with deep connections to convexity, partial differential equations and statistics. At the turn of the millenium, researchers in computer, imaging and more generally data sciences understood that optimal transport theory provided very powerful tools to study distributions in a different and more abstract context, that of comparing distributions readily available to them under the form of bags-of-features or descriptors.

Several reference books have been written on optimal transport, including the two recent monographs by Villani (2003; 2009), those by Rachev and Rüschendorf (1998; 1998) and more recently that by Santambrogio (2015). As exemplified by these books, the more formal and abstract concepts in that theory deserve in and by themselves several hundred pages. Now that optimal transport has gradually established itself as an applied tool (for instance, in economics, as put forward recently by Galichon (2016)), we have tried to balance that rich literature with a computational viewpoint, centered on applications to data science, notably imaging sciences and machine learning. We follow in that sense the motivation of the recent review by Kolouri *et al.* (2017) but try to cover more ground. Ultimately, our goal is to present an overview of the main theoretical insights that support the practical effectiveness of OT and spend more time explaining how to turn these insights into fast computational schemes. The main body of Chapters 2, 3, 4, 9, and 10 is devoted solely to the study of the geometry induced by optimal transport in the space of probability vectors or discrete histograms. Targeting more advanced readers, we also give in the same chapters, in light gray boxes, a more general mathematical exposition of optimal transport tailored for discrete measures. Discrete measures are defined by their probability weights, but also by the location at which these weights are defined. These locations are usually taken in a continuous metric space, giving a second important degree of freedom

to model random phenomena. Lastly, the third and most technical layer of exposition is indicated in dark gray boxes and deals with arbitrary measures that need not be discrete, and which can have in particular a density w.r.t. a base measure. This is traditionally the default setting for most classic textbooks on OT theory, but one that plays a less important role in general for practical applications. Chapters 5 to 8 deal with the interplay between continuous and discrete measures and are thus targeting a more mathematically inclined audience.

The field of computational optimal transport is at the time of this writing still an extremely active one. There are therefore a wide variety of topics that we have not touched upon in this survey. Let us cite in no particular order the subjects of distributionally robust optimization (Shafieezadeh Abadeh *et al.*, 2015; Esfahani and Kuhn, 2018; Lee and Raginsky, 2018; GAO *et al.*, 2018), in which parameter estimation is carried out by minimizing the worst posssible empirical risk of any data measure taken within a certain Wasserstein distance of the input data; convergence of the Langevin Monte Carlo sampling algorithm in the Wasserstein geometry (Dalalyan and Karagulyan, 2017; Dalalyan, 2017; Bernton, 2018); other numerical methods to solve OT with a squared Euclidian cost in low-dimensional settings using the Monge-Ampère equation (Froese and Oberman, 2011; Benamou *et al.*, 2014; Sulman *et al.*, 2011) which are only briefly mentioned in Remark 2.25.

Notation

- $[\![n]\!]$: set of integers $\{1, \dots, n\}$.
- $\mathbb{1}_{n,m}$: matrix of $\mathbb{R}^{n \times m}$ with all entries identically set to 1. $\mathbb{1}_n$: vector of ones.
- $\mathbb{I}_n$: identity matrix of size $n \times n$.
- For $u \in \mathbb{R}^n$, $\mathrm{diag}(u)$ is the $n \times n$ matrix with diagonal u and zero otherwise.
- Σ_n: probability simplex with n bins, namely the set of probability vectors in $\mathbb{R}^n_+$.

- $(\mathbf{a}, \mathbf{b})$: histograms in the simplices $\Sigma_n \times \Sigma_m$.
- (α, β): measures, defined on spaces $(\mathcal{X}, \mathcal{Y})$.
- $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}$: relative density of a measure α with respect to β.
- $\rho_\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}x}$: density of a measure α with respect to Lebesgue measure.
- $(\alpha = \sum_i \mathbf{a}_i \delta_{x_i}, \beta = \sum_j \mathbf{b}_j \delta_{y_j})$: discrete measures supported on $x_1, \ldots, x_n \in \mathcal{X}$ and $y_1, \ldots, y_m \in \mathcal{Y}$.
- $c(x, y)$: ground cost, with associated pairwise cost matrix $\mathbf{C}_{i,j} = (c(x_i, y_j))_{i,j}$ evaluated on the support of α, β.
- π: coupling measure between α and β, namely such that for any $A \subset \mathcal{X}, \pi(A \times \mathcal{Y}) = \alpha(A)$, and for any subset $B \subset \mathcal{Y}, \pi(\mathcal{X} \times B) = \beta(B)$. For discrete measures $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}$.
- $\mathcal{U}(\alpha, \beta)$: set of coupling measures, for discrete measures $\mathbf{U}(\mathbf{a}, \mathbf{b})$.
- $\mathcal{R}(c)$: set of admissible dual potentials; for discrete measures $\mathbf{R}(\mathbf{C})$.
- $T : \mathcal{X} \to \mathcal{Y}$: Monge map, typically such that $T_\sharp \alpha = \beta$.
- $(\alpha_t)_{t=0}^1$: dynamic measures, with $\alpha_{t=0} = \alpha_0$ and $\alpha_{t=1} = \alpha_1$.
- v: speed for Benamou–Brenier formulations; $J = \alpha v$: momentum.
- (f, g): dual potentials, for discrete measures $(\mathbf{f}, \mathbf{g})$ are dual variables.
- $(\mathbf{u}, \mathbf{v}) \stackrel{\text{def.}}{=} (e^{\mathbf{f}/\varepsilon}, e^{\mathbf{g}/\varepsilon})$: Sinkhorn scalings.
- $\mathbf{K} \stackrel{\text{def.}}{=} e^{-\mathbf{C}/\varepsilon}$: Gibbs kernel for Sinkhorn.
- s: flow for $\mathcal{W}_1$-like problem (optimization under divergence constraints).
- $\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ and $\mathcal{L}_c(\alpha, \beta)$: value of the optimization problem associated to the OT with cost $\mathbf{C}$ (histograms) and c (arbitrary measures).

- $\mathrm{W}_p(\mathbf{a}, \mathbf{b})$ and $\mathcal{W}_p(\alpha, \beta)$: p-Wasserstein distance associated to ground distance matrix $\mathbf{D}$ (histograms) and distance d (arbitrary measures).

- $\lambda \in \Sigma_S$: weight vector used to compute the barycenters of S measures.

- $\langle \cdot, \cdot \rangle$: for the usual Euclidean dot-product between vectors; for two matrices of the same size A and B, $\langle A, B \rangle \stackrel{\text{def.}}{=} \mathrm{tr}(A^\top B)$ is the Frobenius dot-product.

- $f \oplus g(x, y) \stackrel{\text{def.}}{=} f(x) + g(y)$, for two functions $f : \mathcal{X} \to \mathbb{R}, g : \mathcal{Y} \to \mathbb{R}$, defines $f \oplus g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

- $\mathbf{f} \oplus \mathbf{g} \stackrel{\text{def.}}{=} \mathbf{f}\mathbb{1}_m^\top + \mathbb{1}_n \mathbf{g}^\top \in \mathbb{R}^{n \times m}$ for two vectors $\mathbf{f} \in \mathbb{R}^n$, $\mathbf{g} \in \mathbb{R}^m$.

- $\alpha \otimes \beta$ is the product measure on $\mathcal{X} \times \mathcal{Y}$, *i.e.* $\int_{\mathcal{X} \times \mathcal{Y}} g(x, y) \mathrm{d}(\alpha \otimes \beta)(x, y) \stackrel{\text{def.}}{=} \int_{\mathcal{X} \times \mathcal{Y}} g(x, y) \mathrm{d}\alpha(x) \mathrm{d}\beta(y)$.

- $\mathbf{a} \otimes \mathbf{b} \stackrel{\text{def.}}{=} \mathbf{a}\mathbf{b}^\top \in \mathbb{R}^{n \times m}$.

- $\mathbf{u} \odot \mathbf{v} = (\mathbf{u}_i \mathbf{v}_i) \in \mathbb{R}^n$ for $(\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^n)^2$.

2

Theoretical Foundations

This chapter describes the basics of optimal transport, introducing first the related notions of optimal matchings and couplings between probability vectors $(\mathbf{a}, \mathbf{b})$, generalizing gradually this computation to transport between discrete measures (α, β), to cover lastly the general setting of arbitrary measures. At first reading, these last nuances may be omitted and the reader can only focus on computations between probability vectors, namely histograms, which is the only requisite to implement algorithms detailed in Chapters 3 and 4. More experienced readers will reach a better understanding of the problem by considering the formulation that applies to arbitrary measures, and will be able to apply it for more advanced problems (*e.g.* in order to move positions of clouds of points, or in a statistical setting where points are sampled from continuous densities).

2.1 Histograms and Measures

We will use interchangeably the terms histogram and probability vector for any element $\mathbf{a} \in \Sigma_n$ that belongs to the probability simplex

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

A large part of this review focuses exclusively on the study of the geometry induced by optimal transport on the simplex.

Remark 2.1 (Discrete measures)**.** A discrete measure with weights $\mathbf{a}$ and locations $x_1, \ldots, x_n \in \mathcal{X}$ reads

$$\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}, \tag{2.1}$$

where δ_x is the Dirac at position x, intuitively a unit of mass which is infinitely concentrated at location x. Such a measure describes a probability measure if, additionally, $\mathbf{a} \in \Sigma_n$ and more generally a positive measure if all the elements of vector $\mathbf{a}$ are nonnegative. To avoid degeneracy issues where locations with no mass are accounted for, we will assume when considering discrete measures that all the elements of $\mathbf{a}$ are positive.

Remark 2.2 (General measures)**.** A convenient feature of OT is that it can deal with measures that are either or both discrete and continuous within the same framework. To do so, one relies on the set 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 access a measure only 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

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

An arbitrary measure $\alpha \in \mathcal{M}(\mathcal{X})$ (which need not have a density nor be a sum of Diracs) is defined by the fact that it can be integrated against 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 has 0 limit at infinity. Measures are thus in some sense "less regular" than functions but more regular than distributions (which are dual to smooth functions). For instance, the derivative of a Dirac is not a measure. 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 2.1 offers a visualization of the different classes of measures, beyond histograms, considered in this work.

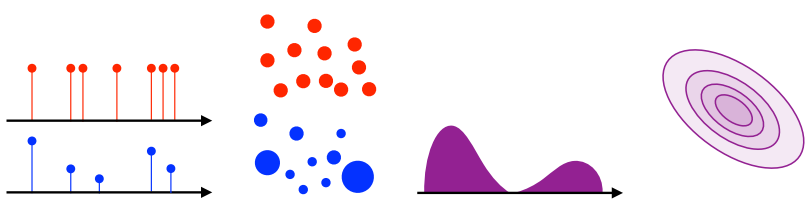


Figure 2.1: Schematic display of discrete distributions $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ (red corresponds to empirical uniform distribution $\mathbf{a}_i = 1/n$, and blue to arbitrary distributions) and densities $\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x$ (in purple), in both one and two dimensions. Discrete distributions in one-dimension are displayed as stem plots (with length equal to $\mathbf{a}_i$) and in two dimensions using point clouds (in which case their radius might be equal to $\mathbf{a}_i$ or, for a more visually accurate representation, their area).

2.2 Assignment and 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

$$\min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}. \tag{2.2}$$

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 (Dantzig, 1983) when n is as small as 70. That problem can therefore be solved only if there exist efficient algorithms to optimize that cost function over the set of permutations, which is the subject of §3.7.

Remark 2.3 (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 four corners of a 2-D square of side length 1, as represented in the left plot of Figure 2.2. In that case only two assignments exist, and they are both optimal.

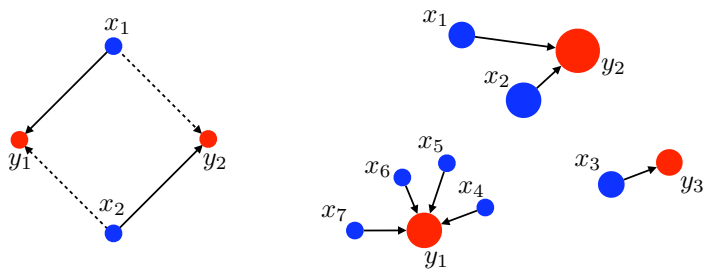


Figure 2.2: Left: blue dots from measure α and red dots from measure β are pairwise equidistant. Hence, either matching $\sigma = (1, 2)$ (full line) or $\sigma = (2, 1)$ (dotted line) is optimal. Right: a Monge map can associate the blue measure α to the red measure β. The weights α_i are displayed 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 $4 \leq i \leq 7$ we have $T(x_i) = y_1$.

Remark 2.4 (Monge problem between discrete measures)**.** For discrete measures

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}, \tag{2.3}$$

the Monge problem (1781) seeks 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, namely, such a map $T : \{x_1, \ldots, x_n\} \to \{y_1, \ldots, y_m\}$ must verify that

$$\forall\, j \in [\![m]\!], \quad \mathbf{b}_j = \sum_{i : T(x_i) = y_j} \mathbf{a}_i, \tag{2.4}$$

which we write in compact form as $T_\sharp \alpha = \beta$. Because all the elements of $\mathbf{b}$ are positive, that map is necessarily surjective. This map should minimize some transportation cost, which is parameterized by a function $c(x, y)$ defined for points $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \;:\; T_\sharp \alpha = \beta \right\}. \tag{2.5}$$

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

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

where the inverse $\sigma^{-1}(j)$ is to be understood as the preimage set of 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 (2.2), 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 a discrete 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, $n < m$. For instance, the right plot in Figure 2.2 shows an (optimal) Monge map between α and β, but there is no Monge map from β to α.

Remark 2.5 (Push-forward operator)**.** For a continuous map $T : \mathcal{X} \to \mathcal{Y}$, we define its corresponding push-forward operator $T_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$. For discrete measures (2.1), the push-forward operation consists simply in moving the positions of all the points in the support of the measure

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

For more general measures, for instance, for those with a density, the notion of push-forward plays a fundamental role to describe the spatial modification (or transport) of a probability measure. The formal definition reads as follows.

Definition 2.1 (Push-forward)**.** For $T : \mathcal{X} \to \mathcal{Y}$, the push-forward measure $\beta = T_\sharp \alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ satisfies

$$\forall\, h \in \mathcal{C}(\mathcal{Y}), \quad \int_{\mathcal{Y}} h(y) \mathrm{d}\beta(y) = \int_{\mathcal{X}} h(T(x)) \mathrm{d}\alpha(x). \tag{2.6}$$

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

$$\beta(B) = \alpha(\{x \in \mathcal{X} \;:\; T(x) \in B\}) = \alpha(T^{-1}(B)). \tag{2.7}$$

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. $T_\sharp$ is an extension of T that can move an entire probability measure on $\mathcal{X}$ toward 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 $\mathcal{Y}$. Note

that a push-forward operator $T_\sharp : \mathcal{M}_+^1(\mathcal{X}) \to \mathcal{M}_+^1(\mathcal{Y})$ is *linear* in the sense that for two measures α_1, α_2 on $\mathcal{X}$, $T_\sharp(\alpha_1 + \alpha_2) = T_\sharp\alpha_1 + T_\sharp\alpha_2$.

Remark 2.6 (Push-forward for multivariate densities)**.** Explicitly doing the change of variables in formula (2.6) for measures with densities $(\rho_\alpha, \rho_\beta)$ on $\mathbb{R}^d$ (assuming T is smooth and bijective) shows that a push-forward acts on densities linearly as a change of variables in the integration formula. Indeed, one has

$$\rho_\alpha(x) = |\det(T'(x))|\rho_\beta(T(x)), \tag{2.8}$$

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

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

Remark 2.7 (Monge problem between arbitrary measures)**.** The Monge problem (2.5) can be extended to the case where two arbitrary probability measures (α, β), supported on two spaces $(\mathcal{X}, \mathcal{Y})$ can be linked through a map $T : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\alpha(x) \;:\; T_\sharp\alpha = \beta \right\}. \tag{2.9}$$

The constraint $T_\sharp\alpha = \beta$ means that T pushes forward the mass of α to β, using the push-forward operator defined in Remark 2.5

Remark 2.8 (Push-forward vs. pull-back)**.** The push-forward $T_\sharp$ of measures should not be confused with the pull-back of functions $T^\sharp : \mathcal{C}(\mathcal{Y}) \to \mathcal{C}(\mathcal{X})$ which corresponds to "warping" between functions, defined as the linear map which to $g \in \mathcal{C}(\mathcal{Y})$ associates $T^\sharp g = g \circ T$. Push-forward and pull-back are actually adjoint to one another, in the sense that

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

Note that even if (α, β) have densities $(\rho_\alpha, \rho_\beta)$ with respect to a fixed measure (*e.g.* Lebesgue on $\mathbb{R}^d$), $T_\sharp \alpha$ does not have $T^\sharp \rho_\beta$ as density, because of the presence of the Jacobian in (2.8). This explains why OT should be used with caution to perform image registration, because it does not operate as an image warping method. Figure 2.3 illustrates the distinction between these push-forward and pull-back operators.

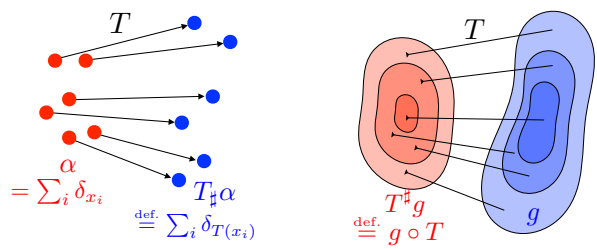


Figure 2.3: Comparison of the push-forward operator $T_\sharp$, which can take as an input any measure, and the pull-back operator $T^\sharp$, which operates on functions, notaly densities.

Remark 2.9 (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 unspecified) probability space $(\Omega, \mathbb{P})$, and its distribution α is the Radon measure $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ such that $\mathbb{P}(X \in A) = \alpha(A) = \int_A \mathrm{d}\alpha(x)$. 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 (2.6), is equivalent to defining another random variable $Y = T(X) : \omega \in \Omega \to T(X(\omega)) \in Y$, so that β is the distribution of Y. Drawing a random sample y from Y is thus simply achieved by computing $y = T(x)$, where x is drawn from X.

2.3 Kantorovich Relaxation

The assignment problem, and its generalization found in the Monge problem laid out in Remark 2.4, is not always relevant to studying discrete measures, such as those found in practical problems. Indeed, because the assignment problem is formulated as a permutation problem, it can only be used to compare *uniform* histograms of the *same* size. A direct generalization to discrete measures with nonuniform weights can be carried out using Monge's formalism of push-forward maps, but that formulation may also be degenerate in the absence of feasible solutions satisfying the mass conservation constraint (2.4) (see the end of Remark 2.4). Additionally, the assignment problem (2.5) is combinatorial, and the feasible set for the Monge problem (2.9), despite being continuously parameterized as the set consisting in all push-forward measures that satisfy the mass conservation constraint, is *nonconvex*. Both are therefore difficult to solve when approached in their original formulation.

Kantorovich's relaxation The key idea of Kantorovich (1942) is to relax the deterministic nature of transportation, namely the fact that a source point x_i can only be assigned to another point or location y_{σ_i} or $T(x_i)$ 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 toward 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 toward bin j, or from the mass found at x_i toward y_j in the formalism of discrete measures (2.3). Admissible couplings admit a far simpler characterization than Monge maps,

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \;:\; \mathbf{P}\mathbb{1}_m = \mathbf{a} \quad \text{and} \quad \mathbf{P}^{\mathrm{T}}\mathbb{1}_n = \mathbf{b} \right\}, \tag{2.10}$$

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 and defined by $n + m$ equality constraints, and therefore is a convex polytope (the convex hull of a finite set of matrices) (Brualdi, 2006, §8.1).

Additionally, whereas the Monge formulation (as illustrated in the right plot of Figure 2.2) 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 $\mathbf{U}(\mathbf{b}, \mathbf{a})$. Kantorovich's optimal transport problem now reads

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \overset{\text{def.}}{=} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}. \tag{2.11}$$

This is a linear program (see Chapter 3), and as is usually the case with such programs, its optimal solutions are not necessarily unique.

Remark 2.10 (Mines and factories)**.** The Kantorovich problem finds a very natural illustration in the following resource allocation problem (see also Hitchcock (1941)). Suppose that an operator runs n warehouses and m factories. Each warehouse contains a valuable raw material that is needed by the factories to run properly. More precisely, each warehouse is indexed with an integer i and contains $\mathbf{a}_i$ units of the raw material. These raw materials must all be moved to the factories, with a prescribed quantity $\mathbf{b}_j$ needed at factory j to function properly. To transfer resources from a warehouse i to a factory j, the operator can use a transportation company that will charge $\mathbf{C}_{i,j}$ to move a single unit of the resource from location i to location j. We assume that the transportation company has the monopoly to transport goods and applies the same linear pricing scheme to all actors of the economy: the cost of shipping a units of the resource from i to j is equal to $a \times \mathbf{C}_{i,j}$.

Faced with the problem described above, the operator chooses to solve the linear program described in Equation (2.11) to obtain a transportation plan $\mathbf{P}^\star$ that quantifies for each pair i, j the amount of goods $\mathbf{P}_{i,j}$ that must transported from warehouse i to factory j. The

operator pays on aggregate a total of $\langle \mathbf{P}^\star, \mathbf{C} \rangle$ to the transportation company to execute that plan.

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

$$\forall\, (i,j) \in [\![n]\!]^2, \quad (\mathbf{P}_\sigma)_{i,j} = \begin{cases} 1/n & \text{if} \quad j = \sigma_i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

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 (2.2) can be recast as a Kantorovich problem (2.11) 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 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, the minimum of $\langle \mathbf{C}, \mathbf{P} \rangle$ is necessarily smaller when considering all transportation than when considering only permutation matrices:

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

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 (2.11) between two uniform measures $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$. The Kantorovich relaxation is therefore *tight* when considered on assignment problems. Figure 2.4 shows on the left a 2-D example of optimal matching corresponding to this special case.

Proposition 2.1 (Kantorovich for matching)**.** If $m = n$ and $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, then there exists an optimal solution for Problem (2.11) $\mathbf{P}_{\sigma^\star}$,

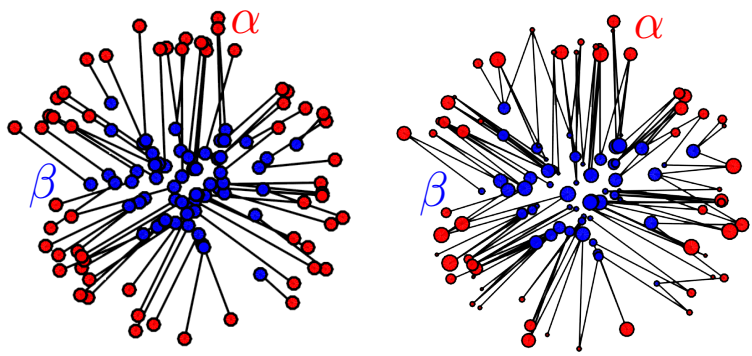


Figure 2.4: Comparison of optimal matching and generic couplings. A black segment between x_i and y_j indicates a nonzero element in the displayed optimal coupling $\mathbf{P}_{i,j}$ solving (2.11). Left: optimal matching, corresponding to the setting of Proposition 2.1 (empirical measures with the same 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.

which is a permutation matrix associated to an optimal permutation $\sigma^\star \in \mathrm{Perm}(n)$ for Problem (2.2).

Proof. Birkhoff's theorem (1946) 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 (Bertsimas and Tsitsiklis, 1997, Theorem 2.7) states that the minimum of a linear objective in a nonempty polyhedron, if finite, is reached at an extremal point of the polyhedron. □

Remark 2.11 (Kantorovich problem between discrete measures)**.** For discrete measures α, β of the form (2.3), we store in the matrix $\mathbf{C}$ all pairwise costs between points in the supports of α, β, namely $\mathbf{C}_{i,j} \overset{\text{def.}}{=} c(x_i, y_j)$, to define

$$\mathcal{L}_c(\alpha, \beta) \overset{\text{def.}}{=} \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}). \tag{2.13}$$

Therefore, the Kantorovich formulation of optimal transport between discrete measures is the same as the problem between their associated probability weight vectors $\mathbf{a}, \mathbf{b}$ except that the cost ma-

trix $\mathbf{C}$ depends on the support of α and β. The notation $\mathcal{L}_c(\alpha, \beta)$, however, is useful in some situations, because it makes explicit the dependency with respect to *both* probability weights and supporting points, the latter being exclusively considered through the cost function c.

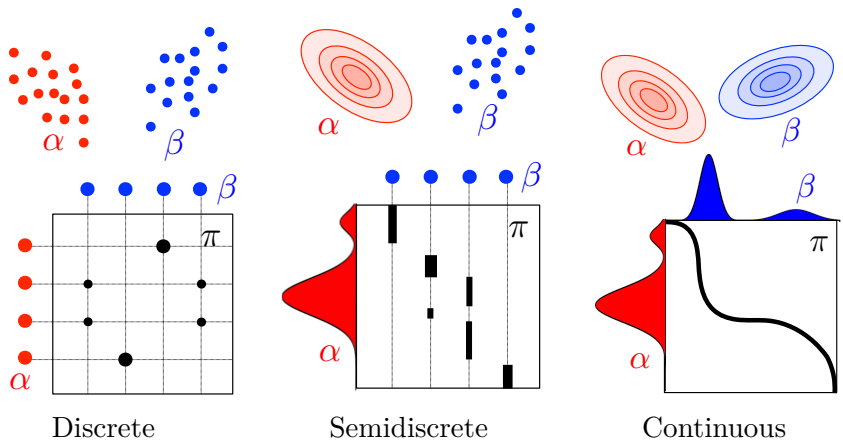


Figure 2.5: Schematic viewed of input measures (α, β) and couplings $\mathcal{U}(\alpha, \beta)$ encountered in the three main scenarios for Kantorovich OT. Chapter 5 is dedicated to the semidiscrete setup.

Remark 2.12 (Using optimal assignments and couplings)**.** The optimal transport plan itself (either as a coupling $\mathbf{P}$ or a Monge map T when it exists) has found many applications in data sciences, and in particular image processing. It has, for instance, been used for contrast equalization (Delon, 2004) and texture synthesis Gutierrez *et al.* (2017). A significant part of applications of OT to imaging sciences is for image matching (Zhu *et al.*, 2007; Wang *et al.*, 2013; Museyko *et al.*, 2009; Li *et al.*, 2013), image fusion (Courty *et al.*, 2016), medical imaging (Wang *et al.*, 2011) and shape registration (Makihara and Yagi, 2010; Lai and Zhao, 2017; Su *et al.*, 2015), and image watermarking (Mathon *et al.*, 2014). In astrophysics, OT has been used for reconstructing the early universe (Frisch *et al.*, 2002). OT has also been used for music transcription (Fla-

mary *et al.*, 2016), and finds numerous applications in economics to interpret matching data (Galichon, 2016). Lastly, let us note that the computation of transportation maps computed using OT techniques (or inspired from them) is also useful to perform sampling (Reich, 2013; Oliver, 2014) and Bayesian inference (Kim *et al.*, 2013; El Moselhy and Marzouk, 2012).

Remark 2.13 (Kantorovich problem between arbitrary measures)**.** Definition (2.13) of $\mathcal{L}_c$ is extended to arbitrary measures by considering couplings $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ which are joint distributions over the product space. The 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 (2.10) should be rewritten as a marginal constraint on joint probability distributions

$$\mathcal{U}(\alpha,\beta) \overset{\text{def.}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \;:\; P_{\mathcal{X}\sharp}\pi = \alpha \quad \text{and} \quad P_{\mathcal{Y}\sharp}\pi = \beta \right\}. \tag{2.14}$$

Here $P_{\mathcal{X}\sharp}$ and $P_{\mathcal{Y}\sharp}$ are the push-forwards (see Definition 2.1) of the projections $P_{\mathcal{X}}(x,y) = x$ and $P_{\mathcal{Y}}(x,y) = y$. Figure 2.5 shows how these coupling constraints translate for different classes of problems (discrete measures and densities). Using (2.7), 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 (2.11) is then generalized as

$$\mathcal{L}_c(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y). \tag{2.15}$$

This is an infinite-dimensional linear program over a space of measures. If $(\mathcal{X},\mathcal{Y})$ are compact spaces and c is continuous, then it is easy to show that it always has solutions. Indeed $\mathcal{U}(\alpha,\beta)$ is compact for the weak topology of measures (see Remark 2.2), $\pi \mapsto \int c\mathrm{d}\pi$ is a continuous function for this topology and the constraint set is nonempty (for instance, $\alpha \otimes \beta \in \mathcal{U}(\alpha,\beta)$). Figure 2.6 shows examples of discrete and continuous optimal coupling solving (2.15).

Figure 2.7 shows other examples of optimal 1-D couplings, involving discrete and continuous marginals.

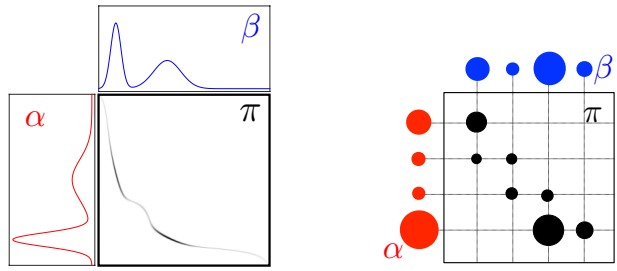


Figure 2.6: Left: "continuous" coupling π solving (2.14) between two 1-D measures with density. The coupling is localized along the graph of the Monge map $(x, T(x))$ (displayed in black). Right: "discrete" coupling T solving (2.11) between two discrete measures of the form (2.3). The positive entries $T_{i,j}$ are displayed as black disks at position (i, j) with radius proportional to $T_{i,j}$.

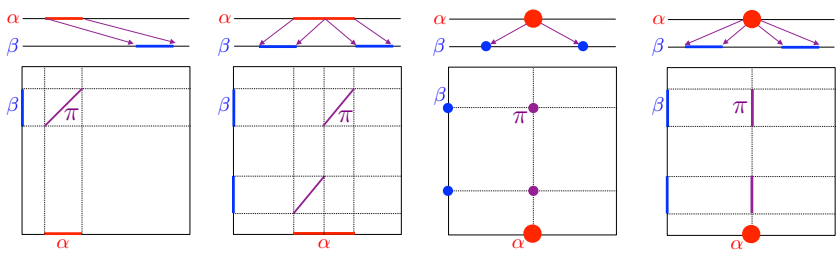


Figure 2.7: Four simple examples of optimal couplings between 1-D distributions, represented as maps above (arrows) and couplings below. Inspired by Lévy and Schwindt (2018).

Remark 2.14 (Probabilistic interpretation)**.** Kantorovich's problem can be reinterpreted through the prism of random variables, following Remark 2.9. Indeed, Problem (2.15) is equivalent to

$$\mathcal{L}_c(\alpha, \beta) = \min_{(X,Y)} \left\{ \mathbb{E}_{(X,Y)}(c(X,Y)) \ : \ X \sim \alpha, Y \sim \beta \right\}, \qquad (2.16)$$

where (X, Y) is a couple of random variables over $\mathcal{X} \times \mathcal{Y}$ and $X \sim \alpha$

(resp., $Y \sim \beta$) means that the law of X (resp., Y), represented as a measure, must be α (resp., β). The law of the couple (X, Y) is then $\pi \in \mathcal{U}(\alpha, \beta)$ over the product space $\mathcal{X} \times \mathcal{Y}$.

2.4 Metric Properties of Optimal Transport

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.

We first consider the case where, using a term first introduced by Rubner *et al.* (2000), 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 valid distance between histograms supported on these bins.

Proposition 2.2. We suppose $n = m$ and that for some $p \geq 1$, $\mathbf{C} = \mathbf{D}^p = (\mathbf{D}_{i,j}^p)_{i,j} \in \mathbb{R}^{n \times n}$, where $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is a distance on $[\![n]\!]$, *i.e.*

(i) $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is symmetric;

(ii) $\mathbf{D}_{i,j} = 0$ if and only if $i = j$;

(iii) $\forall\, (i, j, k) \in [\![n]\!]^3, \mathbf{D}_{i,k} \leq \mathbf{D}_{i,j} + \mathbf{D}_{j,k}$.

Then

$$\mathrm{W}_p(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \mathrm{L}_{\mathbf{D}^p}(\mathbf{a}, \mathbf{b})^{1/p} \tag{2.17}$$

(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

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

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 nonzero element outside the diagonal); by symmetry of $\mathbf{D}^p$, $\mathrm{W}_p(\mathbf{a}, \mathbf{b})$ is itself a symmetric function.

To prove the triangle inequality of Wasserstein distances for arbitrary measures, Villani (2003, 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. To avoid issues that may arise from null coordinates in $\mathbf{b}$, we define a vector $\tilde{\mathbf{b}}$ such that $\tilde{\mathbf{b}}_j \stackrel{\text{def.}}{=} \mathbf{b}_j$ if $\mathbf{b}_j > 0$, and $\tilde{\mathbf{b}}_j \stackrel{\text{def.}}{=} 1$ otherwise, to write

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

and notice that $\mathbf{S} \in \mathbf{U}(\mathbf{a}, \mathbf{c})$ because

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

where we denoted $\mathbb{1}_{\mathrm{Supp}(\mathbf{b})}$ the vector of size n with ones located at those indices j where $\mathbf{b}_j > 0$ and zero otherwise, and we use the fact that $\mathbf{P}\mathbb{1}_{\mathrm{Supp}(\mathbf{b})} = \mathbf{P}\mathbb{1} = \mathbf{a}$ because necessarily $\mathbf{P}_{i,j} = 0$ for those j where $\mathbf{b}_j = 0$. Similarly one verifies that $\mathbf{S}^{\mathrm{T}}\mathbb{1}_n = \mathbf{c}$. The triangle inequality follows then from

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

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

Minkowski's inequality. One thus has

$$\begin{aligned}\mathrm{W}_p(\mathbf{a},\mathbf{c}) &\leq \left(\sum_{ij} \mathbf{D}_{ij}^p \mathbf{P}_{ij} \sum_k \frac{\mathbf{Q}_{jk}}{\tilde{\mathbf{b}}_j}\right)^{1/p} + \left(\sum_{jk} \mathbf{D}_{jk}^p \mathbf{Q}_{jk} \sum_i \frac{\mathbf{P}_{ij}}{\tilde{\mathbf{b}}_j}\right)^{1/p} \\ &= \left(\sum_{ij} \mathbf{D}_{ij}^p \mathbf{P}_{ij}\right)^{1/p} + \left(\sum_{jk} \mathbf{D}_{jk}^p \mathbf{Q}_{jk}\right)^{1/p} \\ &= \mathrm{W}_p(\mathbf{a},\mathbf{b}) + \mathrm{W}_p(\mathbf{b},c),\end{aligned}$$

which concludes the proof. □

Remark 2.15 (The cases $0 < p \leq 1$)**.** Note that if $0 < p \leq 1$, then $\mathbf{D}^p$ is itself distance. This implies that while for $p \geq 1$, $\mathrm{W}_p(\mathbf{a},\mathbf{b})$ is a distance, in the case $p \leq 1$, it is actually $\mathrm{W}_p(\mathbf{a},\mathbf{b})^p$ which defines a distance on the simplex.

Remark 2.16 (Applications of Wasserstein distances)**.** The fact that the OT distance automatically "lifts" a ground metric between bins to a metric between histograms on such bins makes it a method of choice for applications in computer vision and machine learning to compare histograms. In these fields, a classical approach is to "pool" local features (for instance, image descriptors) and compute a histogram of the empirical distribution of features (a so-called bag of features) to perform retrieval, clustering or classification; see, for instance, (Oliva and Torralba, 2001). Along a similar line of ideas, OT distances can be used over some lifted feature spaces to perform signal and image analysis (Thorpe *et al.*, 2017). Applications to retrieval and clustering were initiated by the landmark paper (Rubner *et al.*, 2000), with renewed applications following faster algorithms for threshold matrices $\mathbf{C}$ that fit for some applications, for example, in computer vision (Pele and Werman, 2008; Pele and Werman, 2009). More recent applications stress the use of the earth mover's distance for bags-of-words, either to carry out dimensionality reduction (Rolet *et al.*, 2016) and classify texts (Kusner *et al.*, 2015; Huang *et al.*, 2016), or to define an alternative loss to train multiclass classifiers that output bags-of-words (Frogner

et al., 2015). Kolouri *et al.* (2017) provides a recent overview of such applications to signal processing and machine learning.

Remark 2.17 (Wasserstein distance between measures)**.** Proposition 2.2 can be generalized to deal with arbitrary measures that need not be discrete.

Proposition 2.3. We assume $\mathcal{X} = \mathcal{Y}$ and that for some $p \geq 1$, $c(x, y) = d(x, y)^p$, where d is a distance on $\mathcal{X}$, *i.e.*

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

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

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

Then the p-Wasserstein distance on $\mathcal{X}$,

$$\mathcal{W}_p(\alpha, \beta) \stackrel{\text{def.}}{=} \mathcal{L}_{d^p}(\alpha, \beta)^{1/p} \tag{2.18}$$

(note that $\mathcal{W}_p$ depends on d), is indeed a distance, namely $\mathcal{W}_p$ is symmetric, nonnegative, $\mathcal{W}_p(\alpha, \beta) = 0$ if and only if $\alpha = \beta$, and it satisfies the triangle inequality

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

Proof. The proof follows the same approach as that for Proposition 2.2 and relies on the existence of a coupling between (α, γ) obtained by "gluing" optimal couplings between (α, β) and (β, γ). □

Remark 2.18 (Geometric intuition and weak convergence)**.** The Wasserstein distance $\mathcal{W}_p$ has many important properties, the most important being that it is a weak distance, *i.e.* it allows one to compare singular distributions (for instance, discrete ones) whose supports do not overlap and to quantify the spatial shift between the supports of two 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 that positions (x_i, y_j) take values in a predetermined discrete set to work properly). 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 2.2 (Weak convergence)**.** On a compact domain $\mathcal{X}$, $(\alpha_k)_k$ converges weakly to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any continuous function $g \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} g \mathrm{d}\alpha_k \to \int_{\mathcal{X}} g \mathrm{d}\alpha$. One needs to add additional decay conditions on g on noncompact domains. This notion of weak convergence corresponds to the convergence in the law of random vectors.

This convergence can be shown to be equivalent to $\mathcal{W}_p(\alpha_k, \alpha) \to 0$ (Villani, 2009, Theorem 6.8) (together with a convergence of the moments up to order p for unbounded metric spaces).

Remark 2.19 (Translations)**.** A nice feature of the Wasserstein distance over a Euclidean space $\mathcal{X} = \mathbb{R}^d$ for the ground cost $c(x, y) = \|x - y\|^2$ is that one can factor out translations; indeed, denoting $T_\tau : x \mapsto x - \tau$ the translation operator, one has

$$\mathcal{W}_2(T_{\tau\sharp}\alpha, T_{\tau'\sharp}\beta)^2 = \mathcal{W}_2(\alpha, \beta)^2 - 2\langle \tau - \tau', \mathbf{m}_\alpha - \mathbf{m}_\beta \rangle + \|\tau - \tau'\|^2,$$

where $\mathbf{m}_\alpha \overset{\text{def.}}{=} \int_{\mathcal{X}} x \mathrm{d}\alpha(x) \in \mathbb{R}^d$ is the mean of α. In particular, this implies the nice decomposition of the distance as

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

where $(\tilde{\alpha}, \tilde{\beta})$ are the "centered" zero mean measures $\tilde{\alpha} = T_{\mathbf{m}_\alpha\sharp}\alpha$.

Remark 2.20 (The case $p = +\infty$)**.** Informally, the limit of $\mathcal{W}_p^p$ as $p \to +\infty$ is

$$\mathcal{W}_\infty(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \sup_{(x,y)\in \mathrm{Supp}(\pi)} d(x, y), \tag{2.19}$$

where the sup should be understood as the essential supremum according to the measure π on $\mathcal{X}^2$. In contrast to the cases $p < +\infty$, this is a nonconvex optimization problem, which is difficult to solve numerically and to study theoretically. The $\mathcal{W}_\infty$ distance is related to the Hausdorff distance between the supports of (α, β); see § 10.6.1. We refer to (Champion *et al.*, 2008) for details.

2.5 Dual Problem

The Kantorovich problem (2.11) 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 explains the relationship between the primal and dual problems.

Proposition 2.4. The Kantorovich problem (2.11) admits the dual

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \max_{(\mathbf{f},\mathbf{g})\in \mathbf{R}(\mathbf{C})} \langle \mathbf{f}, \mathbf{a}\rangle + \langle \mathbf{g}, \mathbf{b}\rangle, \tag{2.20}$$

where the set of admissible dual variables is

$$\mathbf{R}(\mathbf{C}) \stackrel{\text{def.}}{=} \{(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m \;:\; \forall\, (i, j) \in [\![n]\!] \times [\![m]\!], \mathbf{f} \oplus \mathbf{g} \leq \mathbf{C}\}\,. \tag{2.21}$$

Such dual variables are often referred to as "Kantorovich potentials."

Proof. This result is a direct consequence of the more general result on the strong duality for linear programs (Bertsimas and Tsitsiklis, 1997, p. 148, Theo. 4.4). The easier part of the proof, namely, establishing that the right-hand side of Equation (2.20) is a lower bound of $\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$, is discussed in Remark 3.2 in the next section. For the sake of completeness, let us derive our result using Lagrangian duality. The Lagangian associated to (2.11) reads

$$\min_{\mathbf{P}\geq 0} \max_{(\mathbf{f},\mathbf{g})\in\mathbb{R}^n\times\mathbb{R}^m} \langle \mathbf{C}, \mathbf{P}\rangle + \langle \mathbf{a} - \mathbf{P}\mathbb{1}_m, \mathbf{f}\rangle + \langle \mathbf{b} - \mathbf{P}^{\mathrm{T}}\mathbb{1}_n, \mathbf{g}\rangle. \tag{2.22}$$

We exchange the min and the max above, which is always possible when considering linear programs (in finite dimension), to obtain

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

We conclude by remarking that

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

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

The primal-dual optimality relation for the Lagrangian (2.22) allows us to locate the support of the optimal transport plan (see also §3.3)

$$\{(i,j)\in[\![n]\!]\times[\![m]\!] : \mathbf{P}_{i,j} > 0\} \subset \Big\{(i,j)\in[\![n]\!]\times[\![m]\!] \;:\; \mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}\Big\}. \tag{2.23}$$

Remark 2.21. Following the interpretation given to the Kantorovich problem in Remark 2.10, we follow with an intuitive presentation of the dual. Recall that in that setup, an operator wishes to move at the least possible cost an overall amount of resources from warehouses to factories. The operator can do so by solving (2.11), follow the instructions set out in $\mathbf{P}^\star$, and pay $\langle \mathbf{P}^\star, \mathbf{C}\rangle$ to the transportation company.

Outsourcing logistics. Suppose that the operator does not have the computational means to solve the linear program (2.11). He decides instead to outsource that task to a vendor. The vendor chooses a pricing scheme with the following structure: the vendor splits the logistic task into that of collecting and then delivering the goods and will apply a collection price $\mathbf{f}_i$ to collect a unit of resource at each warehouse i (no matter where that unit is sent to) and a price $\mathbf{g}_j$ to deliver a unit of resource to factory j (no matter from which warehouse that unit comes from). On aggregate, since there are exactly $\mathbf{a}_i$ units at warehouse i and $\mathbf{b}_j$ needed at factory j, the vendor asks as a consequence of that pricing scheme a price of $\langle \mathbf{f}, \mathbf{a}\rangle + \langle \mathbf{g}, \mathbf{b}\rangle$ to solve the operator's logistic problem.

Setting prices. Note that the pricing system used by the vendor allows quite naturally for arbitrarily negative prices. Indeed, if the

vendor applies a price vector $\mathbf{f}$ for warehouses and a price vector $\mathbf{g}$ for factories, then the total bill will not be changed by simultaneously decreasing all entries in $\mathbf{f}$ by an arbitrary number and increasing all entries of $\mathbf{g}$ by that same number, since the total amount of resources in all warehouses is equal to those that have to be delivered to the factories. In other words, the vendor can give the illusion of giving an extremely good deal to the operator by paying him to collect some of his goods, but compensate that loss by simply charging him more for delivering them. Knowing this, the vendor, wishing to charge as much as she can for that service, sets vectors $\mathbf{f}$ and $\mathbf{g}$ to be as high as possible.

Checking prices. In the absence of another competing vendor, the operator must therefore think of a quick way to check that the vendor's prices are reasonable. A possible way to do so would be for the operator to compute the price $\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ of the most efficient plan by solving problem (2.11) and check if the vendor's offer is at the very least no larger than that amount. However, recall that the operator cannot afford such a lengthy computation in the first place. Luckily, there is a far more efficient way for the operator to check whether the vendor has a competitive offer. Recall that $\mathbf{f}_i$ is the price charged by the vendor for picking a unit at i and $\mathbf{g}_j$ to deliver one at j. Therefore, the vendor's pricing scheme implies that transferring one unit of the resource from i to j costs exactly $\mathbf{f}_i + \mathbf{g}_j$. Yet, the operator also knows that the cost of shipping one unit from i to j as priced by the transporting company is $\mathbf{C}_{i,j}$. Therefore, if for any pair i, j the aggregate price $\mathbf{f}_i + \mathbf{g}_j$ is strictly larger that $\mathbf{C}_{i,j}$, the vendor is charging more than the fair price charged by the transportation company for that task, and the operator should refuse the vendor's offer.

Optimal prices as a dual problem. It is therefore in the interest of the operator to check that for all pairs i, j the prices offered by the vendor verify $\mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{i,j}$. Suppose that the operator does check that the vendor has provided price vectors that do comply with these $n \times m$ inequalities. Can he conclude that the vendor's proposal is attractive? Doing a quick back of the hand calculation, the operator does indeed conclude that it is in his interest to accept that offer. Indeed, since any of his transportation plans $\mathbf{P}$ would have a cost $\langle \mathbf{P}, \mathbf{C} \rangle = \sum_{i,j} \mathbf{P}_{i,j} \mathbf{C}_{i,j}$, the operator can conclude by applying these $n \times m$ inequalities that

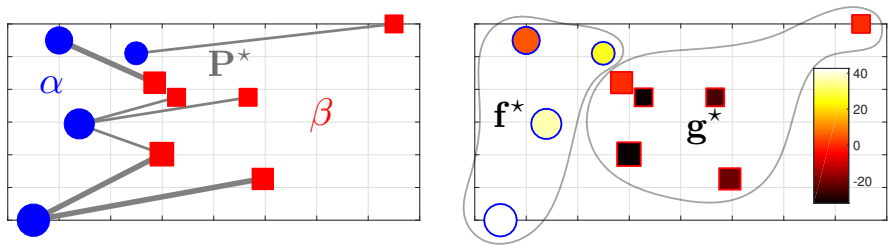


Figure 2.8: Consider in the left plot the optimal transport problem between two discrete measures α and β, represented respectively by blue dots and red squares. The area of these markers is proportional to the weight at each location. That plot also displays the optimal transport $\mathbf{P}^\star$ using a quadratic Euclidean cost. The corresponding dual (Kantorovich) potentials $\mathbf{f}^\star$ and $\mathbf{g}^\star$ that correspond to that configuration are also displayed on the right plot. Since there is a "price" $\mathbf{f}_i^\star$ for each point in α (and conversely for $\mathbf{g}$ and β), the color at that point represents the obtained value using the color map on the right. These potentials can be interpreted as relative prices in the sense that they indicate the individual cost, under the best possible transport scheme, to move a mass away at each location in α, or on the contrary to send a mass toward any point in β. The optimal transport cost is therefore equal to the sum of the squared lengths of all the arcs on the left weighted by their thickness or, alternatively, using the dual formulation, to the sum of the values (encoded with colors) multiplied by the area of each marker on the right plot.

for any transport plan $\mathbf{P}$ (including the optimal one $\mathbf{P}^\star$), the marginal constraints imply

$$\sum_{i,j} \mathbf{P}_{i,j}\mathbf{C}_{i,j} \geq \sum_{i,j} \mathbf{P}_{i,j}\left(\mathbf{f}_i + \mathbf{g}_j\right) = \left(\sum_i \mathbf{f}_i \sum_j \mathbf{P}_{i,j}\right) + \left(\sum_j \mathbf{g}_j \sum_i \mathbf{P}_{i,j}\right)$$
$$= \langle \mathbf{f}, \mathbf{a}\rangle + \langle \mathbf{g}, \mathbf{b}\rangle,$$

and therefore observe that *any* attempt at doing the job by himself would necessarily be more expensive than the vendor's price.

Knowing this, the vendor must therefore find a set of prices $\mathbf{f}, \mathbf{g}$ that maximize $\langle \mathbf{f}, \mathbf{a}\rangle + \langle \mathbf{g}, \mathbf{b}\rangle$ but that must satisfy at the very least for all i, j the basic inequality that $\mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{i,j}$ for his offer to be accepted, which results in Problem (2.20). One can show, as we do later in §3.1, that the best price obtained by the vendor is in fact exactly equal to the best possible cost the operator would obtain by computing $\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$.

Figure 2.8 illustrates the primal and dual solutions resulting from the same transport problem. On the left, blue dots represent warehouses

and red dots stand for factories; the areas of these dots stand for the probability weights $\mathbf{a}, \mathbf{b}$, links between them represent an optimal transport, and their width is proportional to transfered amounts. Optimal prices obtained by the vendor as a result of optimizing Problem (2.20) are shown on the right. Prices have been chosen so that their mean is equal to 0. The highest relative prices come from collecting goods at an isolated warehouse on the lower left of the figure, and delivering goods at the factory located in the upper right area.

Remark 2.22 (Dual problem between arbitrary measures)**.** 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 be accessed only through integration against continuous functions). The duality is formalized in the following proposition, which boils down to Proposition 2.4 when dealing with discrete measures.

Proposition 2.5. One has

$$\mathcal{L}_c(\alpha,\beta) = \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y), \tag{2.24}$$

where the set of admissible dual potentials is

$$\mathcal{R}(c) \stackrel{\text{def.}}{=} \{(f,g)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y}) \;:\; \forall (x,y), f(x)+g(y)\leq c(x,y)\}\,. \tag{2.25}$$

Here, (f,g) is a pair of continuous functions and are also called, as in the discrete case, "Kantorovich potentials."

The discrete case (2.20) 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 us to track the support of the optimal plan, and (2.23) is generalized as

$$\mathrm{Supp}(\pi) \subset \{(x,y)\in\mathcal{X}\times\mathcal{Y} \;:\; f(x)+g(y)=c(x,y)\}\,. \tag{2.26}$$

Note that in contrast to the primal problem (2.15), showing the existence of solutions to (2.24) is nontrivial, because the constraint

set $\mathcal{R}(c)$ is not compact and the function to minimize noncoercive. Using the machinery of c-transform detailed in § 5.1, in the case $c(x,y) = d(x,y)^p$ with $p \geq 1$, one can, however, show that optimal (f,g) are necessarily Lipschitz regular, which enables us to replace the constraint by a compact one.

Remark 2.23 (Unconstrained dual). In the case $\int_{\mathcal{X}} \mathrm{d}\alpha = \int_{\mathcal{Y}} \mathrm{d}\beta = 1$, the constrained dual problem (2.24) can be replaced by an unconstrained one,

$$\mathcal{L}_c(\alpha,\beta) = \sup_{(f,g)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f\mathrm{d}\alpha + \int_{\mathcal{Y}} g\mathrm{d}\beta + \min_{\mathcal{X}\otimes\mathcal{Y}} (c - f \oplus g), \tag{2.27}$$

where we denoted $(f \oplus g)(x,y) = f(x) + g(y)$. Here the minimum should be considered as the essential supremum associated to the measure $\alpha \otimes \beta$, i.e., it does not change if f or g is modified on sets of zero measure for α and β. This alternative dual formulation was pointed out to us by Francis Bach. It is obtained from the primal problem (2.15) by adding the redundant constraint $\int \mathrm{d}\pi = 1$.

Remark 2.24 (Monge–Kantorovich equivalence—Brenier theorem). The following theorem is often attributed to Brenier (1991) and ensures that in $\mathbb{R}^d$ for $p = 2$, if at least one of the two input measures has a density, and for measures with second order moments, then the Kantorovich and Monge problems are equivalent. The interested reader should also consult variants of the same result published more or less at the same time by Cuesta and Matran (1989) and Rüschendorf and Rachev (1990), including notably the original result in (Brenier, 1987) and a precursor by Knott and Smith (1984).

Theorem 2.1 (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 input measures (denoted α) has a density ρ_α with respect to the Lebesgue measure, then the optimal π in the Kantorovich formulation (2.15) 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 $\pi = (\mathrm{Id}, T)_\sharp \alpha$, *i.e.*

$$\forall\, h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\mathrm{d}\pi(x,y) = \int_{\mathcal{X}} h(x,T(x))\mathrm{d}\alpha(x). \tag{2.28}$$

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 $(\nabla\varphi)_\sharp \alpha = \beta$. This convex function is related to the dual potential f solving (2.24) 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 (Santambrogio, 2015). 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 (2.15), one can thus consider the 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

$$\min_{(\varphi,\psi)} \left\{ \int_{\mathcal{X}} \varphi \mathrm{d}\alpha + \int_{\mathcal{Y}} \psi \mathrm{d}\beta \;:\; \forall (x,y), \quad \varphi(x) + \psi(y) \geq \langle x,\, y\rangle \right\}. \tag{2.29}$$

The relation between these variables and those of (2.25) is $(\varphi, \psi) = (\frac{\|\cdot\|^2}{2} - f, \frac{\|\cdot\|^2}{2} - g)$. One can replace the constraint by

$$\forall\, y, \quad \psi(y) \geq \varphi^*(y) \stackrel{\text{def.}}{=} \sup_x \langle x,\, y\rangle - \varphi(x). \tag{2.30}$$

Here φ^* is the Legendre transform of φ and is a convex function as a supremum of linear forms (see also (4.54)). Since the objective appearing in (2.31) is linear and the integrating measures positive, one can minimize explicitly with respect to ψ and set $\psi = \varphi^*$ in order to consider the unconstrained problem

$$\min_{\varphi} \int_{\mathcal{X}} \varphi \mathrm{d}\alpha + \int_{\mathcal{Y}} \varphi^* \mathrm{d}\beta; \tag{2.31}$$

see also §3.2 and §5.1, where that idea is applied respectively in the discrete setting and for 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 (2.31) that φ is convex. Condition (2.26) shows that an optimal π is supported on $\{(x, y) \ : \ \varphi(x) + \varphi^*(y) = \langle x, y\rangle\}$, which shows that such a y is optimal for the minimization (2.30) 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 it shows that necessarily $\pi = (\mathrm{Id}, \nabla\varphi)_\sharp\alpha$. □

This result shows that in the setting of $\mathcal{W}_2$ with no-singular densities, the Monge problem (2.9) and its Kantorovich relaxation (2.15) are equal (the relaxation is tight). This is the continuous counterpart of Proposition 2.1 for the assignment case (2.1), which states that the minimum of the optimal transport problem is achieved at a permutation matrix (a discrete map) when the marginals are equal and uniform. Brenier's theorem, stating that an optimal transport map must be the gradient of a convex function, provides a useful generalization of the notion of increasing functions in dimension more than one. This is the main reason why optimal transport can be used to define quantile functions in arbitrary dimensions, which is in turn useful for applications to quantile regression problems (Carlier *et al.*, 2016).

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.

Remark 2.25 (Monge–Ampère equation)**.** For measures with densities, using (2.8), one obtains that φ is the unique (up to the

addition of a constant) convex function which solves the following Monge–Ampère-type equation:

$$\det(\partial^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x) \tag{2.32}$$

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 nonlinear degenerate Laplacian. In the limit of small displacements, $\varphi = \mathrm{Id} + \varepsilon\psi$, one indeed recovers the Laplacian Δ as a linearization since for smooth maps

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

The convexity constraint forces $\det(\partial^2\varphi(x)) \geq 0$ and is necessary for this equation to have a solution. There is a large body of literature on the theoretical analysis of the Monge–Ampère equation, and in particular the regularity of its solution—see, for instance, (Gutiérrez, 2016); we refer the interested read to the review paper by Caffarelli (2003). A major difficulty is that in full generality, solutions need not be smooth, and one has to resort to the machinery of Alexandrov solutions when the input measures are arbitrary (*e.g.* Dirac masses). Many solvers have been proposed in the simpler case of the Monge–Ampère equation $\det(\partial^2\varphi(x)) = f(x)$ for a fixed right-hand-side f; see, for instance, (Benamou *et al.*, 2016b) and the references therein. In particular, capturing anisotropic convex functions requires special care, and usual finite differences can be inaccurate. For optimal transport, where f actually depends on $\nabla\varphi$, the discretization of Equation (2.32), and the boundary condition result in technical challenges outlined in (Benamou *et al.*, 2014) and the references therein. Note also that related solvers based on fixed-point iterations have been applied to image registration (Haker *et al.*, 2004).

2.6 Special Cases

In general, computing OT distances is numerically involved. Before detailing in §§3,4, and 7 different numerical solvers, we first review special favorable cases where the resolution of the OT problem is relatively easy.

Remark 2.26 (Binary cost matrix and 1-norm)**.** One can easily check that when the cost matrix $\mathbf{C}$ is 0 on the diagonal and 1 elsewhere, namely, when $\mathbf{C} = \mathbb{1}_{n\times n} - \mathbb{I}_n$, the 1-Wasserstein distance between $\mathbf{a}$ and $\mathbf{b}$ is equal to the 1-norm of their difference, $\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_1$.

Remark 2.27 (Kronecker cost function and total variation)**.** In addition to Remark 2.26 above, one can also easily check that this result extends to arbitrary 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 (see also Example 8.2).

Remark 2.28 (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.* $x_1 \leq x_2 \leq \cdots \leq x_n$ and $y_1 \leq y_2 \leq \cdots \leq y_n$, then one has the simple formula

$$\mathcal{W}_p(\alpha, \beta)^p = \frac{1}{n}\sum_{i=1}^n |x_i - y_i|^p, \tag{2.33}$$

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 valid only 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 setting detailed in Remark 2.30. Figure 2.9, 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 also possible to leverage this 1-D computation to also compute efficiently OT on the circle as shown by Delon *et al.* (2010). Note that if the cost is a concave function of the distance, notably when $p < 1$, the behavior of the optimal transport plan is very different, yet efficient solvers also exist (Delon *et al.*, 2012).

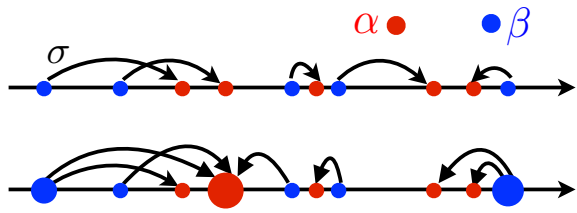


Figure 2.9: 1-D optimal couplings: each arrow $x_i \to y_j$ indicates a nonzero $\mathbf{P}_{i,j}$ in the optimal coupling. Top: empirical measures with same number of points (optimal matching). Bottom: generic case. This corresponds to monotone rearrangements, if $x_i \leq x_{i'}$ are such that $\mathbf{P}_{i,j} \neq 0, \mathbf{P}_{i',j'} \neq 0$, then necessarily $y_j \leq y_{j'}$.

Remark 2.29 (Histogram equalization)**.** One-dimensional optimal transport can be used to perform histogram equalization, with applications to the normalization of the palette of grayscale images, see Figure 2.10. In this case, one denotes $(\bar{x}_i)_i$ and $(\bar{y}_j)_j$ the gray color levels (0 for black, 1 for white, and all values in between) of all pixels of the two input images enumerated in a predefined order (*i.e.* columnwise). Assuming the number of pixels in each image is the same and equal to $n \times m$, sorting these color levels defines $x_i = \bar{x}_{\sigma_1(i)}$ and $y_j = \bar{y}_{\sigma_2(j)}$ as in Remark 2.28, where $\sigma_1, \sigma_2 : \{1, \ldots, nm\} \to \{1, \ldots, nm\}$ are permutations, so that $\sigma \overset{\text{def.}}{=} \sigma_2 \circ \sigma_1^{-1}$ is the optimal assignment between the two discrete distributions. For image processing applications, $(\bar{y}_{\sigma(i)})_i$ defines the color values of an equalized version of $\bar{x}$, whose empirical distribution matches exactly the one of $\bar{y}$. The equalized version of that image can be recovered by folding back that nm-dimensional vector as an image of size $n \times m$. Also, $t \in [0,1] \mapsto (1-t)\bar{x}_i + t\bar{y}_{\sigma(i)}$ defines an interpolation between the original image and the equalized one, whose empirical distribution of pixels is the displacement interpolation (as defined in (7.7)) between those of the inputs.

Remark 2.30 (1-D case—Generic case)**.** For a measure α on $\mathbb{R}$, we introduce the cumulative distribution function from $\mathbb{R}$ to $\to [0,1]$

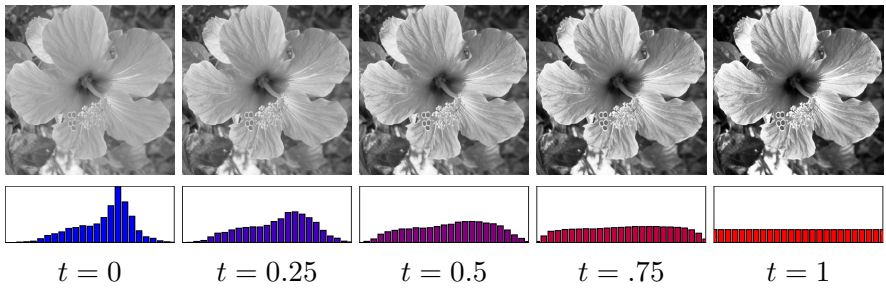


Figure 2.10: Histogram equalization for image processing, where t parameterizes the displacement interpolation between the histograms.

defined as

$$\forall\, x \in \mathbb{R}, \quad \mathcal{C}_\alpha(x) \overset{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha, \tag{2.34}$$

and its pseudoinverse $\mathcal{C}_\alpha^{-1} : [0,1] \to \mathbb{R} \cup \{-\infty\}$

$$\forall\, r \in [0,1], \quad \mathcal{C}_\alpha^{-1}(r) = \min_x \ \{x \in \mathbb{R} \cup \{-\infty\} \ : \ \mathcal{C}_\alpha(x) \geq r\}\,. \tag{2.35}$$

That function is also called the generalized quantile function of α. For any $p \geq 1$, one has

$$\mathcal{W}_p(\alpha,\beta)^p = \left\| \mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1} \right\|_{L^p([0,1])}^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(r) - \mathcal{C}_\beta^{-1}(r)|^p \mathrm{d}r. \tag{2.36}$$

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 geometry in higher dimensions, which is not Hilbertian as discussed in Proposition 8.1 and more generally in §8.3. For $p = 1$, one even has the simpler formula

$$\mathcal{W}_1(\alpha,\beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x \tag{2.37}$$

$$= \int_{\mathbb{R}} \left| \int_{-\infty}^{x} \mathrm{d}(\alpha - \beta) \right| \mathrm{d}x, \tag{2.38}$$

which shows that $\mathcal{W}_1$ is a norm (see §6.2 for the generalization to arbitrary dimensions). An optimal Monge map T such that $T_\sharp \alpha = \beta$ is then defined by

$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha. \tag{2.39}$$

Figure 2.11 illustrates the computation of 1-D OT through cumulative functions. It also displays displacement interpolations, computed as detailed in (7.7); see also Remark 9.6. For a detailed survey of the properties of optimal transport in one dimension, we refer the reader to Santambrogio, 2015, Chapter 2.

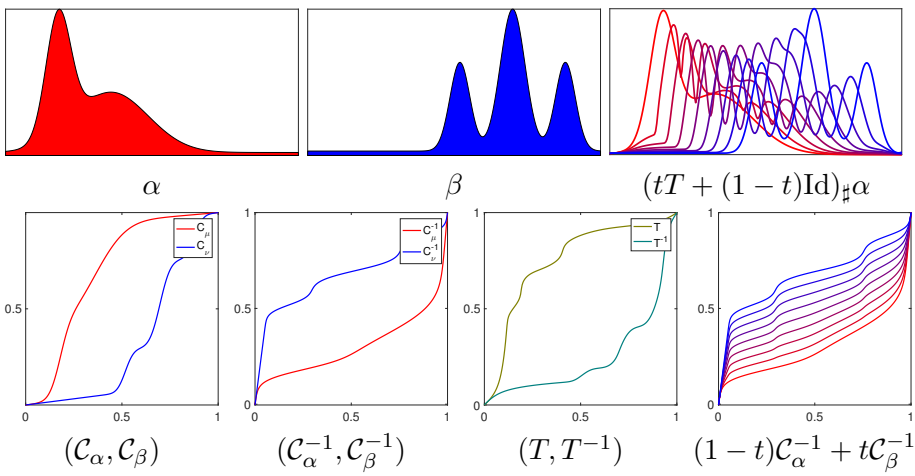


Figure 2.11: Computation of OT and displacement interpolation between two 1-D measures, using cumulant function as detailed in (2.39).

Remark 2.31 (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

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha), \tag{2.40}$$

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 (2.8) is satisfied since

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

and since T is a linear map we have that

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

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 (1991) (see Remark 2.24), that T is optimal. Both that map T and the corresponding potential ψ are illustrated in Figures 2.12 and 2.13

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

$$\mathcal{W}_2^2(\alpha,\beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2, \tag{2.41}$$

where $\mathcal{B}$ is the so-called Bures metric (1969) between positive definite matrices (see also Forrester and Kieburg (2016)),

$$\mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2 \overset{\text{def.}}{=} \operatorname{tr}\left(\boldsymbol{\Sigma}_\alpha + \boldsymbol{\Sigma}_\beta - 2(\boldsymbol{\Sigma}_\alpha^{1/2}\boldsymbol{\Sigma}_\beta\boldsymbol{\Sigma}_\alpha^{1/2})^{1/2}\right), \tag{2.42}$$

where $\boldsymbol{\Sigma}^{1/2}$ 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 $\boldsymbol{\Sigma}_\alpha = \operatorname{diag}(r_i)_i$ and $\boldsymbol{\Sigma}_\beta = \operatorname{diag}(s_i)_i$ are diagonals, the Bures metric is the Hellinger distance

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

For 1-D Gaussians, $\mathcal{W}_2$ is thus the Euclidean distance on the 2-D plane plotting the mean and the standard deviation of a Gaussian

$(\mathbf{m}, \sqrt{\boldsymbol{\Sigma}})$, as illustrated in Figure 2.14. For a detailed treatment of the Wasserstein geometry of Gaussian distributions, we refer to Takatsu (2011), and for additional considerations on the Bures metric the reader can consult the very recent references (Malagò *et al.*, 2018; Bhatia *et al.*, 2018). One can also consult (Muzellec and Cuturi, 2018) for a a recent application of this metric to compute probabilistic embeddings for words, (Shafieezadeh Abadeh *et al.*, 2018) to see how it is used to compute a robust extension to Kalman filtering, or (Mallasto and Feragen, 2017) in which it is applied to covariance functions in reproducing kernel Hilbert spaces.

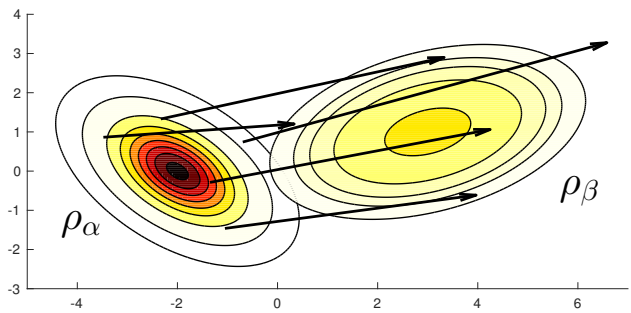


Figure 2.12: Two Gaussians ρ_α and ρ_β, represented using the contour plots of their densities, with respective mean and variance matrices $\mathbf{m}_\alpha = (-2, 0), \boldsymbol{\Sigma}_\alpha = \frac{1}{2}\begin{pmatrix} 1 - \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$ and $\mathbf{m}_\beta = (3, 1), \boldsymbol{\Sigma}_\beta = \begin{pmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$. The arrows originate at random points x taken on the plane and end at the corresponding mappings of those points $T(x) = \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$.

Remark 2.32 (Distance between elliptically contoured distributions)**.** Gelbrich provides a more general result than that provided in Remark 2.31: the Bures metric between Gaussians extends more generally to *elliptically contoured distributions* (1990). In a nutshell, one can first show that for two measures with given mean and covariance matrices, the distance between the two Gaussians with these respective parameters is a lower bound of the Wasserstein distance between the two measures (Gelbrich, 1990, Theorem 2.1). Additionally, the closed form (2.41) extends to families of elliptically contoured densities: If two densities ρ_α and ρ_β belong to such

a family, namely when ρ_α and ρ_β can be written for any point x using a mean and positive definite parameter,

$$\rho_\alpha(x) = \frac{1}{\sqrt{\det(\mathbf{A})}} h(\langle x - \mathbf{m}_\alpha, \mathbf{A}^{-1}(x - \mathbf{m}_\alpha)\rangle)$$
$$\rho_\beta(x) = \frac{1}{\sqrt{\det(\mathbf{B})}} h(\langle x - \mathbf{m}_\beta, \mathbf{B}^{-1}(x - \mathbf{m}_\beta)\rangle),$$

for the same nonnegative valued function h such that the integral

$$\int_{\mathbb{R}^d} h(\langle x, x\rangle) dx = 1,$$

then their optimal transport map is also the linear map (2.40) and their Wasserstein distance is also given by the expression (2.41), with a slightly different scaling of the Bures metric that depends only the generator function h. For instance, that scaling is 1 for Gaussians ($h(t) = e^{-t/2}$) and $1/(d+2)$ for uniform distributions on ellipsoids (h the indicator function for $[0, 1]$). This result follows from the fact that the covariance matrix of an elliptic distribution is a constant times its positive definite parameter (Gómez *et al.*, 2003, Theo. 4(ii)) and that the Wasserstein distance between elliptic distributions is a function of the Bures distance between their covariance matrices (Gelbrich, 1990, Cor. 2.5).

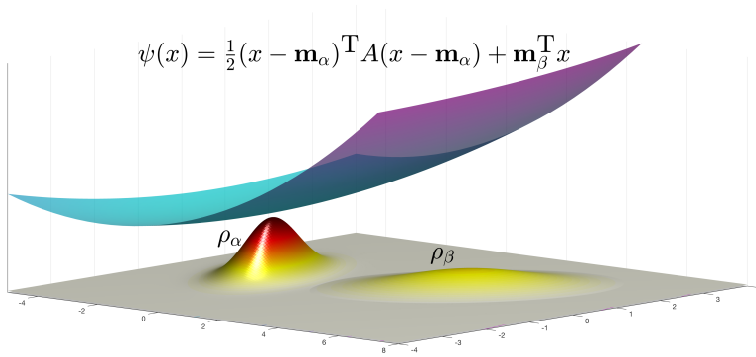


Figure 2.13: Same Gaussians ρ_α and ρ_β as defined in Figure 2.12, represented this time as surfaces. The surface above is the Brenier potential ψ defined up to an additive constant (here $+50$) such that $T = \nabla\psi$. For visual purposes, both Gaussian densities have been multiplied by a factor of 100.

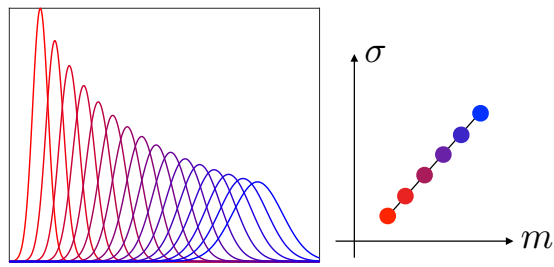


Figure 2.14: Computation of displacement interpolation between two 1-D Gaussians. Denoting $\mathcal{G}_{m,\sigma}(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-m)^2}{2s^2}}$ the Gaussian density, it thus shows the interpolation $\mathcal{G}_{(1-t)m_0+tm_1,(1-t)\sigma_0+t\sigma_1}$.

3

Algorithmic Foundations

This chapter describes the most common algorithmic tools from combinatorial optimization and linear programming that can be used to solve the discrete formulation of optimal transport, as described in the primal problem (2.11) or alternatively its dual (2.20).

The origins of these algorithms can be traced back to World War II, either right before with Tolstoı's seminal work (1930) or during the war itself, when Hitchcock (1941) and Kantorovich (1942) formalized the generic problem of dispatching available resources toward consumption sites in an optimal way. Both of these formulations, as well as the later contribution by Koopmans (1949), fell short of providing a *provably* correct algorithm to solve that problem (the cycle violation method was already proposed as a heuristic by Tolstoı (1939)). One had to wait until the field of linear programming fully blossomed, with the proposal of the simplex method, to be at last able to solve rigorously these problems.

The goal of linear programming is to solve optimization problems whose objective function is linear and whose constraints are linear (in)equalities in the variables of interest. The optimal transport problem fits that description and is therefore a particular case of that wider

class of problems. One can argue, however, that optimal transport is truly special among all linear program. First, Dantzig's early motivation to solve linear programs was greatly related to that of solving transportation problems (Dantzig, 1949, p. 210). Second, despite being only a particular case, the optimal transport problem remained in the spotlight of optimization, because it was understood shortly after that optimal transport problems were related, and in fact equivalent, to an important class of linear programs known as minimum cost network flows (Korte and Vygen, 2012, p. 213, Lem. 9.3) thanks to a result by Ford and Fulkerson (1962). As such, the OT problem has been the subject of particular attention, ever since the birth of mathematical programming (Dantzig, 1951), and is still widely used to introduce optimization to a new audience (Nocedal and Wright, 1999, §1, p. 4).

3.1 The Kantorovich Linear Programs

We have already introduced in Equation (2.11) the primal OT problem:

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \sum_{i \in [\![n]\!], j \in [\![m]\!]} \mathbf{C}_{i,j} \mathbf{P}_{i,j}. \tag{3.1}$$

To make the link with the linear programming literature, one can cast the equation above as a linear program in *standard* form, that is, a linear program with a linear objective; equality constraints defined with a matrix and a constant vector; and nonnegative constraints on variables. Let $\mathbb{I}_n$ stand for the identity matrix of size n and let $\otimes$ be Kronecker's product. The $(n+m) \times nm$ matrix

$$\mathbf{A} = \begin{bmatrix} \mathbb{1}_n{}^{\mathrm{T}} \otimes \mathbb{I}_m \\ \mathbb{I}_n \otimes \mathbb{1}_m{}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{(n+m) \times nm}$$

can be used to encode the row-sum and column-sum constraints that need to be satisfied for any $\mathbf{P}$ to be in $\mathbf{U}(\mathbf{a}, \mathbf{b})$. To do so, simply cast a matrix $\mathbf{P} \in \mathbb{R}^{n \times m}$ as a vector $\mathbf{p} \in \mathbb{R}^{nm}$ such that the $i + n(j-1)$'s element of $\mathbf{p}$ is equal to $\mathbf{P}_{ij}$ ($\mathbf{P}$ is enumerated columnwise) to obtain the following equivalence:

$$\mathbf{P} \in \mathbb{R}^{n \times m} \in \mathbf{U}(\mathbf{a}, \mathbf{b}) \Leftrightarrow \mathbf{p} \in \mathbb{R}_+^{nm}, \mathbf{A}\mathbf{p} = \left[\begin{smallmatrix} \mathbf{a} \\ \mathbf{b} \end{smallmatrix}\right].$$

Therefore we can write the original optimal transport problem as

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \min_{\substack{\mathbf{p}\in\mathbb{R}_+^{nm} \\ \mathbf{A}\mathbf{p}=\left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{b}\end{smallmatrix}\right]}} \mathbf{c}^{\mathrm{T}}\mathbf{p}, \tag{3.2}$$

where the nm-dimensional vector $\mathbf{c}$ is equal to the stacked columns contained in the cost matrix $\mathbf{C}$.

Remark 3.1. Note that one of the $n+m$ constraints described above is redundant or that, in other words, the line vectors of matrix A are not linearly independent. Indeed, summing all n first lines and the subsequent m lines results in the same vector (namely $A\left[\begin{smallmatrix}\mathbb{1}_n\\ \mathbb{0}_m\end{smallmatrix}\right] = A\left[\begin{smallmatrix}\mathbb{0}_n\\ \mathbb{1}_m\end{smallmatrix}\right] = \mathbb{1}_{nm}{}^{\mathrm{T}}$). One can show that removing a line in A and the corresponding entry in $\left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{b}\end{smallmatrix}\right]$ yields a properly defined linear system. For simplicity, and to avoid treating asymmetrically $\mathbf{a}$ and $\mathbf{b}$, we retain in what follows a redundant formulation, keeping in mind that degeneracy will pop up in some of our computations.

The dual problem corresponding to Equation (3.2) is, following duality in linear programming (Bertsimas and Tsitsiklis, 1997, p. 143) defined as

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \max_{\substack{\mathbf{h}\in\mathbb{R}^{n+m} \\ A^{\mathrm{T}}\mathbf{h}\leq\mathbf{c}}} \left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{b}\end{smallmatrix}\right]^{\mathrm{T}}\mathbf{h}. \tag{3.3}$$

Note that this program is exactly equivalent to that presented in Equation (2.4).

Remark 3.2. We provide a simple derivation of the duality result above, which can be seen as a rigorous formulation of the arguments developed in Remark 2.21 to introduce duality. Strong duality, namely the fact that the optima of both primal (3.2) and dual (3.3) problems do indeed coincide, requires a longer proof (Bertsimas and Tsitsiklis, 1997, §4.10). To simplify notation, we write $\mathbf{q} = \left[\begin{smallmatrix}\mathbf{a}\\ \mathbf{b}\end{smallmatrix}\right]$. Consider now a relaxed primal problem of the optimal transport problem, where the constraint $\mathbf{A}\mathbf{p} = \mathbf{q}$ is no longer necessarily enforced but bears a cost $\mathbf{h}^{\mathrm{T}}(\mathbf{A}\mathbf{p}-\mathbf{q})$ parameterized by an arbitrary cost vector $\mathbf{h}\in\mathbb{R}^{n+m}$. This relaxation, whose optimum depends directly on the cost vector $\mathbf{h}$, can be written as

$$\mathrm{H}(\mathbf{h}) \stackrel{\text{def.}}{=} \min_{\mathbf{p}\in\mathbb{R}_+^{nm}} \mathbf{c}^{\mathrm{T}}\mathbf{p} - \mathbf{h}^{\mathrm{T}}(\mathbf{A}\mathbf{p}-\mathbf{q}).$$

Note first that this relaxed problem has no marginal constraints on $\mathbf{p}$. Because that minimization allows for many more $\mathbf{p}$ solutions, we expect $\mathrm{H}(\mathbf{h})$ to be smaller than $\bar{z} = \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$. Indeed, writing $\mathbf{p}^\star$ for any optimal solution of the primal problem (3.1), we obtain

$$\min_{\mathbf{p} \in \mathbb{R}_+^{nm}} \mathbf{c}^{\mathrm{T}}\mathbf{p} - \mathbf{h}^{\mathrm{T}}(\mathbf{A}\mathbf{p} - \mathbf{q}) \leq \mathbf{c}^{\mathrm{T}}\mathbf{p}^\star - \mathbf{h}^{\mathrm{T}}(\mathbf{A}\mathbf{p}^\star - \mathbf{q}) = \mathbf{c}^{\mathrm{T}}\mathbf{p}^\star = \bar{z}.$$

The approach above defines therefore a problem which can be used to compute an optimal upper bound for the original problem (3.1), for any cost vector $\mathbf{h}$; that function is called the Lagrange dual function of L. The goal of duality theory is now to compute the best lower bound $\underline{z}$ by *maximizing* H over *any* cost vector $\mathbf{h}$, namely

$$\underline{z} = \max_{\mathbf{h}} \left(\mathrm{H}(\mathbf{h}) = \max_{\mathbf{h}} \mathbf{h}^{\mathrm{T}}\mathbf{q} + \min_{\mathbf{p} \in \mathbb{R}_+^{nm}} \left(\mathbf{c} - A^{\mathrm{T}}\mathbf{h}\right)^{\mathrm{T}} \mathbf{p} \right).$$

The second term involving a minimization on $\mathbf{p}$ can be easily shown to be $-\infty$ if any coordinate of $\mathbf{c}^{\mathrm{T}} - A^{\mathrm{T}}\mathbf{h}$ is negative. Indeed, if for instance for a given index $i \leq n + m$ we have $\mathbf{c}_i - (A^{\mathrm{T}}\mathbf{h})_i < 0$, then it suffices to take for $\mathbf{p}$ the canonical vector $\mathbf{e}_i$ multiplied by any arbitrary large positive value to obtain an unbounded value. When trying to maximize the lower bound $\mathrm{H}(\mathbf{h})$ it therefore makes sense to restrict vectors $\mathbf{h}$ to be such that $A^{\mathrm{T}}\mathbf{h} \leq \mathbf{c}$, in which case the best possible lower bound becomes

$$\underline{z} = \max_{\substack{\mathbf{h} \in \mathbb{R}^{n+m} \\ A^{\mathrm{T}}\mathbf{h} \leq \mathbf{c}}} \mathbf{h}^{\mathrm{T}}\mathbf{q}.$$

We have therefore proved a weak duality result, namely that $\underline{z} \leq \bar{z}$.

3.2 C-Transforms

We present in this section an important property of the dual optimal transport problem (3.3) which takes a more important meaning when used for the semidiscrete optimal transport problem in §5.1. This section builds upon the original formulation (2.20) that splits dual variables according to row and column sum constraints:

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \max_{(\mathbf{f}, \mathbf{g}) \in \mathbf{R}(\mathbf{C})} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle. \tag{3.4}$$

Consider any dual feasible pair $(\mathbf{f}, \mathbf{g})$. If we "freeze" the value of $\mathbf{f}$, we can notice that there is no better vector solution for $\mathbf{g}$ than the $\mathbf{C}$-transform vector of $\mathbf{f}$, denoted $\mathbf{f}^{\mathbf{C}} \in \mathbb{R}^m$ and defined as

$$(\mathbf{f}^{\mathbf{C}})_j = \min_{i \in [\![n]\!]} \mathbf{C}_{ij} - \mathbf{f}_i,$$

since it is indeed easy to prove that $(\mathbf{f}, \mathbf{f}^{\mathbf{C}}) \in \mathbf{R}(\mathbf{C})$ and that $\mathbf{f}^{\mathbf{C}}$ is the largest possible vector such that this constraint is satisfied. We therefore have that

$$\langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle \leq \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{f}^{\mathbf{C}}, \mathbf{b} \rangle.$$

This result allows us first to reformulate the dual problem as a piecewise affine concave maximization problem expressed in a single variable $\mathbf{f}$ as

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \max_{\mathbf{f} \in \mathbb{R}^n} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{f}^{\mathbf{C}}, \mathbf{b} \rangle. \tag{3.5}$$

Putting that result aside, the same reasoning applies of course if we now "freeze" the values of $\mathbf{g}$ and consider instead the $\bar{\mathbf{C}}$-transform of $\mathbf{g}$, namely vector $\mathbf{g}^{\bar{\mathbf{C}}} \in \mathbb{R}^n$ defined as

$$(\mathbf{g}^{\bar{\mathbf{C}}})_i = \min_{j \in [\![m]\!]} \mathbf{C}_{ij} - \mathbf{g}_j,$$

with a different increase in objective

$$\langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle \leq \langle \mathbf{g}^{\bar{\mathbf{C}}}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle.$$

Starting from a given $\mathbf{f}$, it is therefore tempting to alternate $\mathbf{C}$ and $\bar{\mathbf{C}}$ transforms several times to improve $\mathbf{f}$. Indeed, we have the sequence of inequalities

$$\langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{f}^{\mathbf{C}}, \mathbf{b} \rangle \leq \langle \mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}}, \mathbf{a} \rangle + \langle \mathbf{f}^{\mathbf{C}}, \mathbf{b} \rangle \leq \langle \mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}}, \mathbf{a} \rangle + \langle \mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}\mathbf{C}}, \mathbf{b} \rangle \leq \dots$$

One may hope for a strict increase in the objective at each of these iterations. However, this does not work because alternating $\mathbf{C}$ and $\bar{\mathbf{C}}$ transforms quickly hits a plateau.

Proposition 3.1. The following identities, in which the inequality sign between vectors should be understood elementwise, hold:

(i) $\mathbf{f} \leq \mathbf{f}' \Rightarrow \mathbf{f}^{\mathbf{C}} \geq \mathbf{f}'^{\mathbf{C}}$,

(ii) $\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}} \geq \mathbf{f}$, $\mathbf{g}^{\bar{\mathbf{C}}\mathbf{C}} \geq \mathbf{g}$,

(iii) $\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}\mathbf{C}} = \mathbf{f}^{\mathbf{C}}$.

Proof. The first inequality follows from the definition of $\mathbf{C}$-transforms. Expanding the definition of $\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}}$ we have

$$\left(\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}}\right)_i = \min_{j\in[\![m]\!]} \mathbf{C}_{ij} - \mathbf{f}_j^{\mathbf{C}} = \min_{j\in[\![m]\!]} \mathbf{C}_{ij} - \min_{i'\in[\![n]\!]} \mathbf{C}_{i'j} - \mathbf{f}_{i'}.$$

Now, since $-\min_{i'\in[\![n]\!]} \mathbf{C}_{i'j} - \mathbf{f}_{i'} \geq -(\mathbf{C}_{ij} - \mathbf{f}_i)$, we recover

$$\left(\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}}\right)_i \geq \min_{j\in[\![m]\!]} \mathbf{C}_{ij} - \mathbf{C}_{ij} + \mathbf{f}_i = \mathbf{f}_i.$$

The relation $\mathbf{g}^{\bar{\mathbf{C}}\mathbf{C}} \geq \mathbf{g}$ is obtained in the same way. Now, set $\mathbf{g} = \mathbf{f}^{\mathbf{C}}$. Then, $\mathbf{g}^{\bar{\mathbf{C}}} = \mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}} \geq \mathbf{f}$. Therefore, using result (i) we have $\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}\mathbf{C}} \leq \mathbf{f}^{\mathbf{C}}$. Result (ii) yields $\mathbf{f}^{\mathbf{C}\bar{\mathbf{C}}\mathbf{C}} \geq \mathbf{f}^{\mathbf{C}}$, proving the equality. □

3.3 Complementary Slackness

Primal (3.2) and dual (3.3), (2.20) problems can be solved independently to obtain optimal primal $\mathbf{P}^\star$ and dual $(\mathbf{f}^\star, \mathbf{g}^\star)$ solutions. The following proposition characterizes their relationship.

Proposition 3.2. Let $\mathbf{P}^\star$ and $\mathbf{f}^\star, \mathbf{g}^\star$ be optimal solutions for the primal (2.24) and dual (2.11) problems, respectively. Then, for any pair $(i, j) \in [\![n]\!] \times [\![m]\!]$, $\mathbf{P}^\star_{i,j}(\mathbf{C}_{i,j} - \mathbf{f}^\star_i + \mathbf{g}^\star_j) = 0$ holds. In other words, if $\mathbf{P}^\star_{i,j} > 0$, then necessarily $\mathbf{f}^\star_i + \mathbf{g}^\star_j = \mathbf{C}_{i,j}$; if $\mathbf{f}^\star_i + \mathbf{g}^\star_j < \mathbf{C}_{i,j}$ then necessarily $\mathbf{P}^\star_{i,j} = 0$.

Proof. We have by strong duality that $\langle \mathbf{P}^\star, \mathbf{C}\rangle = \langle \mathbf{f}^\star, \mathbf{a}\rangle + \langle \mathbf{g}^\star, \mathbf{b}\rangle$. Recall that $\mathbf{P}^\star \mathbb{1}_m = \mathbf{a}$ and $\mathbf{P}^{\star\mathrm{T}} \mathbb{1}_n = \mathbf{b}$; therefore

$$\begin{aligned}\langle \mathbf{f}^\star, \mathbf{a}\rangle + \langle \mathbf{g}^\star, \mathbf{b}\rangle &= \langle \mathbf{f}^\star, \mathbf{P}^\star \mathbb{1}_m\rangle + \langle \mathbf{g}^\star, \mathbf{P}^{\star\mathrm{T}} \mathbb{1}_n\rangle \\ &= \langle \mathbf{f}^\star {\mathbb{1}_m}^{\mathrm{T}}, \mathbf{P}^\star\rangle + \langle \mathbb{1}_n {\mathbf{g}^\star}^{\mathrm{T}}, \mathbf{P}^\star\rangle,\end{aligned}$$

which results in

$$\langle \mathbf{P}^\star, \mathbf{C} - \mathbf{f}^\star \oplus \mathbf{g}^\star\rangle = 0.$$

Because $(\mathbf{f}^\star, \mathbf{g}^\star)$ belongs to the polyhedron of dual constraints (2.21), each entry of the matrix $\mathbf{C} - \mathbf{f}^\star \oplus \mathbf{g}^\star$ is necessarily nonnegative. Therefore,

since all the entries of $\mathbf{P}$ are nonnegative, the constraint that the dot-product above is equal to 0 enforces that, for any pair of indices (i, j) such that $\mathbf{P}_{i,j} > 0$, $\mathbf{C}_{i,j} - (\mathbf{f}_i + \mathbf{g}_j)$ must be zero, and for any pair of indices (i, j) such that $\mathbf{C}_{i,j} > \mathbf{f}_i + \mathbf{g}_j$ that $\mathbf{P}_{i,j} = 0$. □

The converse result is also true. We define first the idea that two variables for the primal and dual problems are complementary.

Definition 3.1. A matrix $\mathbf{P} \in \mathbb{R}^{n \times m}$ and a pair of vectors $(\mathbf{f}, \mathbf{g})$ are complementary w.r.t. $\mathbf{C}$ if for all pairs of indices (i, j) such that $\mathbf{P}_{i,j} > 0$ one also has $\mathbf{C}_{i,j} = \mathbf{f}_i + \mathbf{g}_j$.

If a pair of feasible primal and dual variables is complementary, then we can conclude they are optimal.

Proposition 3.3. If $\mathbf{P}$ and $(\mathbf{f}, \mathbf{g})$ are complementary and feasible solutions for the primal (2.24) and dual (2.11) problems, respectively, then $\mathbf{P}$ and $(\mathbf{f}, \mathbf{g})$ are both primal and dual optimal.

Proof. By weak duality, we have that

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \leq \langle \mathbf{P}, \mathbf{C} \rangle = \langle \mathbf{P}, \mathbf{f} \oplus \mathbf{g} \rangle = \langle \mathbf{a}, \mathbf{f} \rangle + \langle \mathbf{b}, \mathbf{g} \rangle \leq \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$$

and therefore $\mathbf{P}$ and $(\mathbf{f}, \mathbf{g})$ are respectively primal and dual optimal. □

3.4 Vertices of the Transportation Polytope

Recall that a vertex or an extremal point of a convex set is formally a point $\mathbf{x}$ in that set such that, if there exiss $\mathbf{y}$ and $\mathbf{z}$ in that set with $\mathbf{x} = (\mathbf{y} + \mathbf{z})/2$, then necessarily $\mathbf{x} = \mathbf{y} = \mathbf{z}$. A linear program with a nonempty and bounded feasible set attains its minimum at a vertex (or extremal point) of the feasible set (Bertsimas and Tsitsiklis, 1997, p. 65, Theo. 2.7). Since the feasible set $\mathbf{U}(\mathbf{a}, \mathbf{b})$ of the primal optimal transport problem (3.2) is bounded, one can restrict the search for an optimal $\mathbf{P}$ to the set of extreme points of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$. Matrices $\mathbf{P}$ that are extremal in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ have an interesting structure that has been the subject of extensive research (Brualdi, 2006, §8). That structure requires describing the transport problem using the formalism of bipartite graphs.

3.4.1 Tree Structure of the Support of All Vertices of U(a, b)

Let $V = (1, 2, \ldots, n)$ and $V' = (1', 2', \ldots, m')$ be two sets of nodes. Note that we add a prime to the labels of set V' to disambiguate them from those of V. Consider their union $V \cup V'$, with $n + m$ nodes, and the set $\mathcal{E}$ of all nm directed edges $\{(i, j'), i \in [\![n]\!], j \in [\![m]\!]\}$ between them (here we just add a prime to an integer $j \leq m$ to form j' in V'). To each edge (i, j') we associate the corresponding cost value $\mathbf{C}_{ij}$. The complete bipartite graph $\mathcal{G}$ between V and V' is $(V \cup V', E)$. A transport plan is a flow on that graph satisfying source ($\mathbf{a}_i$ flowing out of each node i) and sink ($\mathbf{b}_j$ flowing into each node j') constraints, as described informally in Figure 3.1. An extremal point in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ has the following property (Brualdi, 2006, p. 338, Theo. 8.1.2).

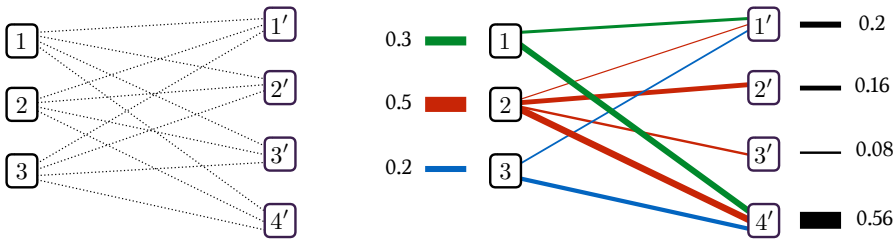


Figure 3.1: The optimal transport problem as a bipartite network flow problem. Here $n = 3, m = 4$. All coordinates of the source histogram, $\mathbf{a}$, are depicted as source nodes on the left labeled $1, 2, 3$, whereas all coordinates of the target histogram $\mathbf{b}$ are labeled as nodes $1', 2', 3', 4'$. The graph is bipartite in the sense that all source nodes are connected to all target nodes, with no additional edges. To each edge (i, j') is associated a cost $\mathbf{C}_{ij}$. A feasible flow is represented on the right. Proposition 3.4 shows that this flow is not extremal since it has at least one cycle given by $((1, 1'), (2, 1'), (2, 4'), (1, 4'))$.

Proposition 3.4 (Extremal solutions)**.** Let $\mathbf{P}$ be an extremal point of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$. Let $S(\mathbf{P}) \subset \mathcal{E}$ be the subset of edges $\{(i, j'), i \in [\![n]\!], j \in [\![m]\!]$ such that $\mathbf{P}_{ij} > 0\}$. Then the graph $G(\mathbf{P}) \stackrel{\text{def.}}{=} (V \cup V', S(\mathbf{P}))$ has no cycles. In particular, $\mathbf{P}$ cannot have more than $n+m-1$ nonzero entries.

Proof. We proceed by contradiction. Suppose that $\mathbf{P}$ is an extremal point of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$ and that its corresponding set $S(\mathbf{P})$ of

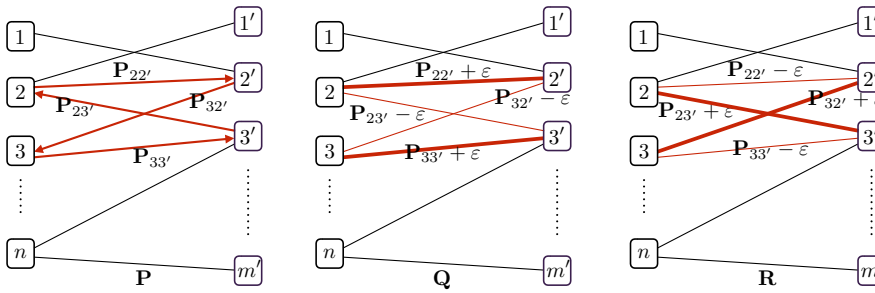


Figure 3.2: A solution $\mathbf{P}$ with a cycle in the graph of its support can be perturbed to obtain two feasible solutions $\mathbf{Q}$ and $\mathbf{R}$ such that $\mathbf{P}$ is their average, therefore disproving that $\mathbf{P}$ is extremal.

edges, denoted F for short, is such that the graph $G = (V \cup V', F)$ contains a cycle, namely there exists $k > 1$ and a sequence of distinct indices $i_1, \dots, i_{k-1} \in [\![n]\!]$ and $j_1, \dots, j_{k-1} \in [\![m]\!]$ such that the set of edges

$$H = \{(i_1, j'_1), (i_2, j'_1), (i_2, j'_2), \dots, (i_k, j'_k), (i_1, j'_k)\}$$

forms a subset of F. We now construct two feasible matrices $\mathbf{Q}$ and $\mathbf{R}$ such that $\mathbf{P} = (\mathbf{Q} + \mathbf{R})/2$. To do so, consider a *directed* cycle $\bar{H}$ corresponding to H, namely the sequence of pairs $i_1 \to j'_1, j'_1 \to i_2, i_2 \to j'_2, \dots, i_k \to j'_k, j'_k \to i_1$, as well as the elementary amount of flow $\varepsilon < \min_{(i,j') \in F} \mathbf{P}_{ij}$. Consider a perturbation matrix $\mathbf{E}$ whose (i, j) entry is equal to ε if $i \to j' \in \bar{H}$, $-\varepsilon$ if $j \to i' \in \bar{H}$, and zero otherwise. Define matrices $\mathbf{Q} = \mathbf{P} + \mathbf{E}$ and $\mathbf{R} = \mathbf{P} - \mathbf{E}$ as illustrated in Figure 3.2. Because ε is small enough, all elements in $\mathbf{Q}$ and $\mathbf{R}$ are nonnegative. By construction, $\mathbf{E}$ has either lines (resp., columns) with all entries equal to 0 or exactly one entry equal to ε and another equal to $-\varepsilon$ for those indexed by $i_1, \dots, i_k$ (resp., $j_1, \dots, j_k$). Therefore, $\mathbf{E}$ is such that $\mathbf{E}\mathbb{1}_m = \mathbb{0}_n$ and $\mathbf{E}^{\mathrm{T}}\mathbb{1}_n = \mathbb{0}_m$, and we have that $\mathbf{Q}$ and $\mathbf{R}$ have the same marginals as $\mathbf{P}$, and are therefore feasible. Finally $\mathbf{P} = (\mathbf{Q} + \mathbf{R})/2$ which, since $\mathbf{Q}, \mathbf{R} \neq \mathbf{P}$, contradicts the fact that $\mathbf{P}$ is an extremal point. Since a graph with k nodes and no cycles cannot have more than $k - 1$ edges, we conclude that $S(\mathbf{P})$ cannot have more than $n + m - 1$ edges, and therefore $\mathbf{P}$ cannot have more than $n + m - 1$ nonzero entries. □

3.4.2 The North-West Corner Rule

The north-west (NW) corner rule is a heuristic that produces a vertex of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$ in up to $n + m$ operations. This heuristic can play a role in initializing any algorithm working on the primal, such as the network simplex outlined in the next section.

The rule starts by giving the highest possible value to $\mathbf{P}_{1,1}$ by setting it to $\min(\mathbf{a}_1, \mathbf{b}_1)$. At each step, the entry $\mathbf{P}_{i,j}$ is chosen to saturate either the row constraint at i, the column constraint at j, or both if possible. The indices i, j are then updated as follows: i is incremented in the first case, j is in the second, and both i and j are in the third case. The rule proceeds until $\mathbf{P}_{n,m}$ has received a value.

Formally, the algorithm works as follows: i and j are initialized to 1, $r \leftarrow \mathbf{a}_1, c \leftarrow \mathbf{b}_1$. While $i \leq n$ and $j \leq m$, set $t \leftarrow \min(r, c)$, $\mathbf{P}_{i,j} \leftarrow t$, $r \leftarrow r - t$, $c \leftarrow s - t$; if $r = 0$ then increment i, and update $r \leftarrow \mathbf{a}_i$ if $i \leq n$; if $c = 0$ then increment j, and update $c \leftarrow \mathbf{b}_j$ if $j \leq n$; repeat. Here is an example of this sequence assuming $\mathbf{a} = [0.2, 0.5, 0.3]$ and $\mathbf{b} = [0.5, 0.1, 0.4]$:

$$\begin{bmatrix} \bullet & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2 & 0 & 0 \\ \bullet & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2 & 0 & 0 \\ 0.3 & \bullet & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 0.2 & 0 & 0 \\ 0.3 & 0.1 & \bullet \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2 & 0 & 0 \\ 0.3 & 0.1 & 0.1 \\ 0 & 0 & \bullet \end{bmatrix} \rightarrow \begin{bmatrix} 0.2 & 0 & 0 \\ 0.3 & 0.1 & 0.1 \\ 0 & 0 & 0.3 \end{bmatrix}$$

We write $\mathbf{NW}(\mathbf{a}, \mathbf{b})$ for the unique plan that can be obtained through this heuristic.

Note that there is, however, a much larger number of NW corner solutions that can be obtained by permuting arbitrarily the order of $\mathbf{a}$ and $\mathbf{b}$ first, computing the corresponding NW corner table, and recovering a table of $\mathbf{U}(\mathbf{a}, \mathbf{b})$ by inverting again the order of columns and rows: setting $\sigma = (3, 1, 2), \sigma' = (3, 2, 1)$ gives $\mathbf{a}_\sigma = [0.3, 0.2, 0.5], \mathbf{b}_{\sigma'} =$

$[0.4, 0.1, 0.5]$, and $\sigma^{-1} = (2, 3, 1), \sigma' = (3, 2, 1)$. Observe that

$$\mathbf{NW}(\mathbf{a}_\sigma, \mathbf{b}_{\sigma'}) = \begin{bmatrix} 0.3 & 0 & 0 \\ 0.1 & 0.1 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} \in \mathbf{U}(\mathbf{a}_\sigma, \mathbf{b}_{\sigma'}),$$

$$\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(\mathbf{a}_\sigma, \mathbf{b}_{\sigma'}) = \begin{bmatrix} 0 & 0.1 & 0.1 \\ 0.5 & 0 & 0 \\ 0 & 0 & 0.3 \end{bmatrix} \in \mathbf{U}(\mathbf{a}, \mathbf{b}).$$

Let $\mathcal{N}(\mathbf{a}, \mathbf{b})$ be the set of all NW corner solutions that can be produced this way:

$$\mathcal{N}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \{\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}), \sigma, \sigma' \in S_d\}.$$

All NW corner solutions have by construction up to $n + m - 1$ nonzero elements. The NW corner rule produces a table which is by construction unique for $\mathbf{a}_\sigma$ and $\mathbf{b}'_\sigma$, but there is an exponential number of pairs or row/column permutations (σ, σ') that may yield the same table (Stougie, 2002, p. 2). $\mathcal{N}(\mathbf{a}, \mathbf{b})$ forms a subset of (usually strictly included in) the set of extreme points of $\mathbf{U}(\mathbf{a}, \mathbf{b})$ (Brualdi, 2006, Cor. 8.1.4).

3.5 A Heuristic Description of the Network Simplex

Consider a feasible matrix $\mathbf{P}$ whose graph $G(\mathbf{P}) = (V \cup V', S(\mathbf{P}))$ has no cycles. $\mathbf{P}$ has therefore no more than $n + m - 1$ nonzero entries and is a vertex of $\mathbf{U}(\mathbf{a}, \mathbf{b})$ by Proposition 3.4. Following Proposition 3.3, it is therefore sufficient to obtain a dual solution $(\mathbf{f}, \mathbf{g})$ which is feasible (*i.e.* $\mathbf{C} - \mathbf{f} \oplus \mathbf{g}$ has nonnegative entries) and complementary to $\mathbf{P}$ (pairs of indices (i, j') in $S(\mathbf{P})$ are such that $\mathbf{C}_{i,j} = \mathbf{f}_i + \mathbf{g}_j$), to prove that $\mathbf{P}$ is optimal. The network simplex relies on two simple principles: to each feasible primal solution $\mathbf{P}$ one can associate a complementary pair $(\mathbf{f}, \mathbf{g})$. If that pair is feasible, then we have reached optimality. If not, one can consider a modification of $\mathbf{P}$ that remains feasible and whose complementary pair $(\mathbf{f}, \mathbf{g})$ is modified so that it becomes closer to feasibility.

3.5.1 Obtaining a Dual Pair Complementary to P

The simplex proceeds by associating first to any extremal solution $\mathbf{P}$ a pair of $(\mathbf{f}, \mathbf{g})$ complementary dual variables. This is simply carried out by finding two vectors $\mathbf{f}$ and $\mathbf{g}$ such that for any (i, j') in $S(\mathbf{P})$, $\mathbf{f}_i + \mathbf{g}_j$ is equal to $\mathbf{C}_{i,j}$. Note that this, in itself, does not guarantee that $(\mathbf{f}, \mathbf{g})$ is feasible.

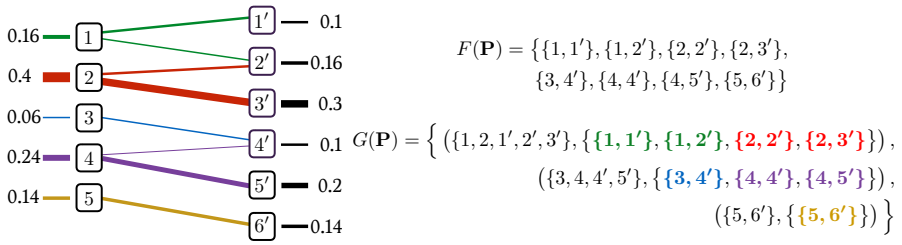


Figure 3.3: A feasible transport $\mathbf{P}$ and its corresponding set of edges $S(\mathbf{P})$ and graph $G(\mathbf{P})$. As can be seen, the graph $G(\mathbf{P}) = (\{1, \ldots, 5, 1', \ldots, 6'\}, S(\mathbf{P}))$ is a forest, meaning that it can be expressed as the union of tree graphs, three in this case.

Let s be the cardinality of $S(\mathbf{P})$. Because $\mathbf{P}$ is extremal, $s \leq n+m-1$. Because $G(\mathbf{P})$ has no cycles, $G(\mathbf{P})$ is either a tree or a forest (a union of trees), as illustrated in Figure 3.3. Aiming for a pair $(\mathbf{f}, \mathbf{g})$ that is complementary to $\mathbf{P}$, we consider the following set of s linear equality constraints on $n + m$ variables:

$$\begin{array}{rcl} \mathbf{f}_{i_1} + \mathbf{g}_{j_1} & = & \mathbf{C}_{i_1,j_1} \\ \mathbf{f}_{i_2} + \mathbf{g}_{j_1} & = & \mathbf{C}_{i_2,j_1} \\ \vdots & = & \vdots \\ \mathbf{f}_{i_s} + \mathbf{g}_{j_s} & = & \mathbf{C}_{i_s,j_s}, \end{array} \tag{3.6}$$

where the elements of $S(\mathbf{P})$ are enumerated as $(i_1, j'_1), \ldots, (i_s, j'_s)$.

Since $s \leq n+m-1 < n+m$, the linear system (3.6) above is always undetermined. This degeneracy can be interpreted in part because the parameterization of $\mathbf{U}(\mathbf{a}, \mathbf{b})$ with $n + m$ constraints results in $n + m$ dual variables. A more careful formulation, outlined in Remark 3.1, would have resulted in an equivalent formulation with only $n + m - 1$ constraints and therefore $n + m - 1$ dual variables. However, s can also be strictly smaller than $n + m - 1$: This happens when $G(\mathbf{P})$ is the

disjoint union of two or more trees. For instance, there are $5 + 6 = 11$ dual variables (one for each node) in Figure 3.3, but only 8 edges among these 11 nodes, namely 8 linear equations to define $(\mathbf{f}, \mathbf{g})$. Therefore, there will be as many undetermined dual variables under that setting as there will be connected components in $G(\mathbf{P})$.

Consider a tree among those listed in $G(\mathbf{P})$. Suppose that tree has k nodes $i_1, \ldots, i_k$ among source nodes and l nodes $j'_1, \ldots, j'_l$ among target nodes, resulting in $r \stackrel{\text{def.}}{=} k + l$, and $r - 1$ edges, corresponding to k variables in $\mathbf{f}$ and l variables in $\mathbf{g}$, linked with $r - 1$ linear equations. To lift an indetermination, we can choose arbitrarily a root node in that tree and assign the value 0 to its corresponding dual variable. From there, we can traverse the tree using a breadth-first or depth-first search to obtain a sequence of simple variable assignments that determines the values of all other dual variables in that tree, as illustrated in Figure 3.4. That procedure can then be repeated for all trees in the graph of $\mathbf{P}$ to obtain a pair of dual variables $(\mathbf{f}, \mathbf{g})$ that is complementary to $\mathbf{P}$.

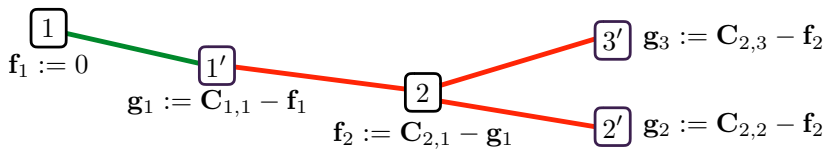


Figure 3.4: The five dual variables $\mathbf{f}_1, \mathbf{f}_2, \mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ corresponding to the five nodes appearing in the first tree of the graph $G(\mathbf{P})$ illustrated in Figure 3.3 are linked through four linear equations that involve corresponding entries in the cost matrix $\mathbf{C}$. Because that system is degenerate, we choose a root in that tree (node 1 in this example) and set its corresponding variable to 0 and proceed then by traversing the tree (either breadth-first or depth-first) from the root to obtain iteratively the values of the four remaining dual variables.

3.5.2 Network Simplex Update

The dual pair $(\mathbf{f}, \mathbf{g})$ obtained previously might be feasible, in the sense that for all i, j we have $\mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{i,j}$, in which case we have reached the optimum by Proposition 3.3. When that is not the case, namely when there exists i, j such that $\mathbf{f}_i + \mathbf{g}_j > \mathbf{C}_{i,j}$, the network simplex algorithm kicks in. We first initialize a graph G to be equal to the graph $G(\mathbf{P})$ corresponding to the feasible solution $\mathbf{P}$ and add the violating edge

(i, j') to G. Two cases can then arise:

(a) G is (still) a forest, which can happen if (i, j') links two existing subtrees. The approach outlined in §3.5.1 can be used on graph G to recover a new complementary dual vector $(\mathbf{f}, \mathbf{g})$. Note that this addition simply removes an indetermination among the $n + m$ dual variables and does not result in any change in the primal variable $\mathbf{P}$. That update is usually called degenerate in the sense that (i, j') has now entered graph G although $\mathbf{P}_{i,j}$ remains 0. $G(\mathbf{P})$ is, however, contained in G.

(b) G now has a cycle. In that case, we need to remove an edge in G to ensure that G is still a forest, yet also modify $\mathbf{P}$ so that $\mathbf{P}$ is feasible and $G(\mathbf{P})$ remains included in G. These operations can all be carried out by increasing the value of $\mathbf{P}_{i,j}$ and modifying the other entries of $\mathbf{P}$ appearing in the detected cycle, in a manner very similar to the one we used to prove Proposition 3.4. To be more precise, let us write that cycle $(i_1, j'_1), (j'_1, i_2), (i_2, j'_2), \ldots, (i_l, j'_l), (j'_l, i_{l+1})$ with the convention that $i_1 = i_{l+1} = i$ to ensure that the path is a cycle that starts and ends at i, whereas $j_1 = j$, to highlight the fact that the cycle starts with the added edge $\{i, j\}$, going in the right direction. Increase now the flow of all "positive" edges (i_k, j'_k) (for $k \leq l$), and decrease that of "negative" edges (j'_k, i_{k+1}) (for $k \leq l$), to obtain an updated primal solution $\tilde{\mathbf{P}}$, equal to $\mathbf{P}$ for all but the following entries:

$$\forall k \leq l, \quad \tilde{\mathbf{P}}_{i_k,j_k} := \mathbf{P}_{i_k,j_k} + \theta; \quad \tilde{\mathbf{P}}_{i_{k+1},j_k} := \mathbf{P}_{i_{k+1},j_k} - \theta.$$

Here, θ is the largest possible increase at index i, j using that cycle. The value of θ is controlled by the smallest flow negatively impacted by the cycle, namely $\min_k \mathbf{P}_{i_{k+1},j_k}$. That update is illustrated in Figure 3.5. Let $k^\star$ be an index that achieves that minimum. We then close the update by removing $(i_{k^\star+1}, j_{k^\star})$ from G, to compute new dual variables $(\mathbf{f}, \mathbf{g})$ using the approach outlined in §3.5.1.

3.5.3 Improvement of the Primal Solution

Although this was not necessarily our initial motivation, one can show that the manipulation above can only improve the cost of $\mathbf{P}$. If the added edge has not created a cycle, case (a) above, the primal solution remains unchanged. When a cycle is created, case (b), $\mathbf{P}$ is updated to $\tilde{\mathbf{P}}$, and the following equality holds:

$$\langle \tilde{\mathbf{P}}, \mathbf{C} \rangle - \langle \mathbf{P}, \mathbf{C} \rangle = \theta \left(\sum_{k=1}^{l} \mathbf{C}_{i_k, j_k} - \sum_{k=1}^{l} \mathbf{C}_{i_{k+1}, j_k} \right).$$

We now use the dual vectors $(\mathbf{f}, \mathbf{g})$ computed at the end of the previous iteration. They are such that $f_{i_k} + g_{i_k} = \mathbf{C}_{i_k, j_k}$ and $f_{i_{k+1}} + g_{i_k} = \mathbf{C}_{i_{k+1}, j_k}$ for *all* edges initially in G, resulting in the identity

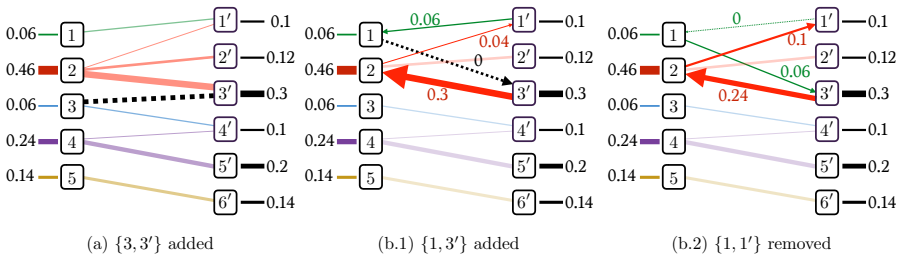
$$\begin{aligned} \sum_{k=1}^{l} \mathbf{C}_{i_k, j_k} - \sum_{k=1}^{l} \mathbf{C}_{i_{k+1}, j_k} &= \mathbf{C}_{i,j} + \sum_{k=2}^{l} \mathbf{f}_{i_k} + \mathbf{g}_{j_k} - \sum_{k=1}^{l} \mathbf{f}_{i_{k+1}} + \mathbf{g}_{j_k} \\ &= \mathbf{C}_{i,j} - (\mathbf{f}_i + \mathbf{g}_j). \end{aligned}$$

That term is, by definition, negative, since i, j were chosen because $C_{i,j} < \mathbf{f}_i - \mathbf{g}_j$. Therefore, if $\theta > 0$, we have that

$$\langle \tilde{\mathbf{P}}, \mathbf{C} \rangle = \langle \mathbf{P}, \mathbf{C} \rangle + \theta \left(\mathbf{C}_{i,j} - (\mathbf{f}_i - \mathbf{f}_g) \right) < \langle \mathbf{P}, \mathbf{C} \rangle.$$

If $\theta = 0$, which can happen if G and $G(\mathbf{P})$ differ, the graph G is simply changed, but $\mathbf{P}$ is not.

The network simplex algorithm can therefore be summarized as follows: Initialize the algorithm with an extremal solution $\mathbf{P}$, given for instance by the NW corner rule as covered in §3.4.2. Initialize the graph G with $G(\mathbf{P})$. Compute a pair of dual variables $(\mathbf{f}, \mathbf{g})$ that are complementary to $\mathbf{P}$ using the linear system solve using the tree structure(s) in G as described in §3.5.1. (i) Look for a violating pair of indices to the constraint $\mathbf{C} - \mathbf{f} \oplus \mathbf{g} \geq 0$; if none, $\mathbf{P}$ is optimal and stop. If there is a violating pair (i, j'), (ii) add the edge (i, j') to G. If G still has no cycles, update $(\mathbf{f}, \mathbf{g})$ accordingly; if there is a cycle, direct it making sure (i, j') is labeled as positive, and remove a negative edge in that cycle with the smallest flow value, updating $\mathbf{P}, G$ as illustrated in Figure 3.5, then build a complementary pair $\mathbf{f}, \mathbf{g}$ accordingly; return



(a) $\{3, 3'\}$ added (b.1) $\{1, 3'\}$ added (b.2) $\{1, 1'\}$ removed

Figure 3.5: Adding an edge $\{i, j\}$ to the graph $G(\mathbf{P})$ can result in either (a) the graph remains a forest after this addition, in which case $\mathbf{f}, \mathbf{g}$ can be recomputed following the approach outlined in §3.5.1; (b.1) the addition of that edge creates a cycle, from which we can define a directed path; (b.2) the path can be used to increase the value of $\mathbf{P}_{i,j}$ and propagate that change along the cycle to maintain the flow feasibility constraints, until the flow of one of the edges that is negatively impacted by the cycle is decreased to 0. This removes the cycle and updates $\mathbf{P}$.

to (i). Some of the operations above require graph operations (cycle detection, tree traversals) which can be implemented efficiently in this context, as described in (Bertsekas, 1998, §5).

Orlin (1997) was the first to prove the polynomial time complexity of the network simplex. Tarjan (1997) provided shortly after an improved bound in $O\left((n+m)nm\log(n+m)\log\left((n+m)\|\mathbf{C}\|_\infty\right)\right)$ which relies on more efficient data structures to help select pivoting edges.

3.6 Dual Ascent Methods

Dual ascent methods precede the network simplex by a few decades, since they can be traced back to work by Borchardt and Jocobi (1865) and later König and Egerváry, as recounted by Kuhn (1955). The Hungarian algorithm is the best known algorithm in that family, and it can work only in the particular case when $\mathbf{a}$ and $\mathbf{b}$ are equal and are both uniform, namely $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$. We provide in what follows a concise description of the more general family of dual ascent methods. This requires the knowledge of the maximum flow problem (Bertsimas and Tsitsiklis, 1997, §7.5). By contrast to the network simplex, presented above in the primal, dual ascent methods maintain at each iteration dual feasible solutions whose objective is progressively improved by adding a sparse vector to $\mathbf{f}$ and $\mathbf{g}$. Our presentation is mostly derived

from that of (Bertsimas and Tsitsiklis, 1997, §7.7) and starts with the following definition.

Definition 3.2. For $S \subset [\![n]\!], S' \subset [\![m]\!]' \stackrel{\text{def.}}{=} \{1', \dots, m'\}$ we write $\mathbb{1}_S$ for the vector in $\mathbb{R}^n$ of zeros except for ones at the indices enumerated in S, and likewise for the vector $\mathbb{1}_{S'}$ in $\mathbb{R}^m$ with indices in S'.

In what follows, $(\mathbf{f}, \mathbf{g})$ is a feasible dual pair in $\mathbf{R}(\mathbf{C})$. Recall that this simply means that for all pairs $(i, j') \in [\![n]\!] \times [\![m]\!]'$, $\mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{ij}$. We say that (i, j') is a *balanced* pair (or edge) if $\mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{ij}$ and *inactive* otherwise, namely if $\mathbf{f}_i + \mathbf{g}_j < \mathbf{C}_{ij}$. With this convention, we start with a simple result describing how a feasible dual pair $(\mathbf{f}, \mathbf{g})$ can be perturbed using sparse vectors indexed by sets S and S' and still remain feasible.

Proposition 3.5. $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \stackrel{\text{def.}}{=} (\mathbf{f}, \mathbf{g}) + \varepsilon(\mathbb{1}_S, -\mathbb{1}_{S'})$ is dual feasible for a small enough $\varepsilon > 0$ if for all $i \in S$, the fact that (i, j') is balanced implies that $j' \in S'$.

Proof. For any $i \in S$, consider the set $\mathcal{I}_i$ of all $j' \in [\![m]\!]'$ such that (i, j') is inactive, namely such that $\mathbf{f}_i + \mathbf{g}_j < \mathbf{C}_{ij}$. Define $\varepsilon_i \stackrel{\text{def.}}{=} \min_{j \in I_i} \mathbf{C}_{i,j} - \mathbf{f}_i - \mathbf{g}_j$, the smallest margin by which $\mathbf{f}_i$ can be increased without violating the constraints corresponding to $j' \in \mathcal{I}_i$. Indeed, one has that if $\varepsilon \leq \varepsilon_i$ then $\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_j < \mathbf{C}_{i,j}$ for any $j' \in \mathcal{I}_i$. Consider now the set $\mathcal{B}_i$ of balanced edges associated with i. Note that $\mathcal{B}_i = [\![m]\!]' \setminus \mathcal{I}_i$. The assumption above is that $j' \in \mathcal{B}_i \Rightarrow j' \in S'$. Therefore, one has that for $j' \in \mathcal{B}_i$, $\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_j = \mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}$. As a consequence, the inequality $\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_j \leq \mathbf{C}_{i,j}$ is ensured for any $j \in [\![m]\!]'$. Choosing now an increase ε smaller than the smallest possible allowed, namely $\min_{i \in S} \varepsilon_i$, we recover that $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ is dual feasible. □

The main motivation behind the iteration of the network simplex presented in §3.5.1 is to obtain, starting from a feasible primal solution $\mathbf{P}$, a complementary feasible dual pair $(\mathbf{f}, \mathbf{g})$. To reach that goal, $\mathbf{P}$ is progressively modified such that its complementary dual pair reaches dual feasibility. A symmetric approach, starting from a feasible dual variable to obtain a feasible primal $\mathbf{P}$, motivates dual ascent methods. The proposition below is the main engine of dual ascent methods in

the sense that it guarantees (constructively) the existence of an ascent direction for $(\mathbf{f}, \mathbf{g})$ that maintains feasibility. That direction is built, similarly to the network simplex, by designing a candidate primal solution $\mathbf{P}$ whose infeasibility guides an update for $(\mathbf{f}, \mathbf{g})$.

Proposition 3.6. Either $(\mathbf{f}, \mathbf{g})$ is optimal for Problem (3.4) or there exists $S \subset [\![n]\!], S' \subset [\![m]\!]'$ such that $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \overset{\text{def.}}{=} (\mathbf{f}, \mathbf{g}) + \varepsilon(\mathbb{1}_S, -\mathbb{1}_{S'})$ is feasible for a small enough $\varepsilon > 0$ and has a strictly better objective.

Proof. We consider first a complementary primal variable $\mathbf{P}$ to $(\mathbf{f}, \mathbf{g})$. To that effect, let $\mathcal{B}$ be the set of balanced edges, namely all pairs $(i, j') \in [\![n]\!] \times [\![m]\!]'$ such that $\mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}$, and form the bipartite graph whose vertices $\{1, \dots, n, 1', \dots, m'\}$ are linked with edges in $\mathcal{B}$ only, complemented by a source node s connected with *capacitated* edges to all nodes $i \in [\![n]\!]$ with respective capacities $\mathbf{a}_i$, and a terminal node t also connected to all nodes $j' \in [\![m]\!]'$ with edges of respective capacities $\mathbf{b}_j$, as seen in Figure 3.6. The Ford–Fulkerson algorithm (Bertsimas and Tsitsiklis, 1997, p. 305) can be used to compute a maximal flow $\mathbf{F}$ on that network, namely a family of $n + m + |\mathcal{B}|$ nonnegative values indexed by $(i, j') \in \mathcal{B}$ as $f_{si} \leq \mathbf{a}_i, f_{ij'}, f_{j't} \leq \mathbf{b}_j$ that obey flow constraints and such that $\sum_i f_{si}$ is maximal. If the throughput of that flow $\mathbf{F}$ is equal to 1, then a feasible primal solution $\mathbf{P}$, complementary to $\mathbf{f}, \mathbf{g}$ by construction, can be extracted from $\mathbf{F}$ by defining $\mathbf{P}_{i,j} = f_{ij'}$ for $(i, j') \in \mathcal{B}$ and zero elsewhere, resulting in the optimality of $(\mathbf{f}, \mathbf{g})$ and $\mathbf{P}$ by Proposition 3.3. If the throughput of $\mathbf{F}$ is strictly smaller than 1, the labeling algorithm proceeds by labeling (identifying) those nodes reached iteratively from s for which $\mathbf{F}$ does not saturate capacity constraints, as well as those nodes that contribute flow to any of the labeled nodes. Labeled nodes are stored in a nonempty set Q, which does not contain the terminal node t per optimality of $\mathbf{F}$ (see Bertsimas and Tsitsiklis 1997, p. 308, for a rigorous presentation of the algorithm). Q can be split into two sets $S = Q \cap [\![n]\!]$ and $S' = Q \cap [\![m]\!]'$. Because we have assumed that the total throughput is strictly smaller than 1, $S \neq \emptyset$. Note first that if $i \in S$ and (i, j) is balanced, then j' is necessarily in S'. Indeed, since all edges (i, j') have infinite capacity by construction, the labeling algorithm will necessarily reach j' if it includes i in S. By Proposition 3.5, there exists thus a small enough ε to ensure the feasibility of $\tilde{\mathbf{f}}, \tilde{\mathbf{g}}$. One still needs to

prove that ${\mathbb{1}_S}^{\mathrm{T}}\mathbf{a} - {\mathbb{1}_{S'}}^{\mathrm{T}}\mathbf{b} > 0$ to ensure that $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ has a better objective than $(\mathbf{f}, \mathbf{g})$. Let $\bar{S} = [\![n]\!] \setminus S$ and $\bar{S}' = [\![m]\!]' \setminus S'$ and define

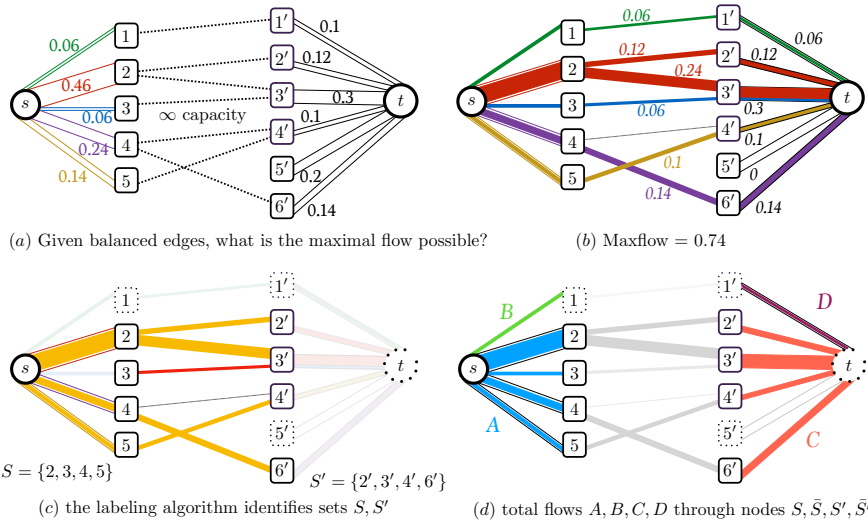
$$A = \sum_{i \in S} f_{si}, \quad B = \sum_{i \in \bar{S}} f_{si}, \quad C = \sum_{j' \in S'} f_{j't}, \quad D = \sum_{j' \in \bar{S}'} f_{j't}.$$

The total maximal flow starts from s and is therefore equal to $A + B$, but also arrives at t and is therefore equal to $C + D$. Flow conservation constraints also impose that the very same flow is equal to $B + C$, therefore $A = C$. On the other hand, by definition of the labeling algorithm, we have for all i in S that $f_{si} < \mathbf{a}_i$, whereas $f_{j't} = \mathbf{b}_j$ for $j' \in \bar{S}'$ because t cannot be in S' by optimality of the considered flow. We therefore have $A < {\mathbb{1}_S}^{\mathrm{T}}\mathbf{a}$ and $C = {\mathbb{1}'_S}^{\mathrm{T}}\mathbf{b}$. Therefore ${\mathbb{1}_S}^{\mathrm{T}}\mathbf{a} - {\mathbb{1}'_S}^{\mathrm{T}}\mathbf{b} > A - C = 0$. □

The dual ascent method proceeds by modifying any feasible solution $(\mathbf{f}, \mathbf{g})$ by any vector generated by sets S, S' that ensure feasibility and improve the objective. When the sets S, S' are those given by construction in the proof of Proposition 3.6, and the steplength ε is defined as in the proof of Proposition 3.5, we recover a method known as the *primal-dual* method. That method reduces to the Hungarian algorithm for matching problems. Dual ascent methods share similarities with the dual variant of the network simplex, yet they differ in at least two important aspects. Simplex-type methods always ensure that the current solution is an *extreme point* of the feasible set, $\mathbf{R}(\mathbf{C})$ for the dual, whereas dual ascent as presented here does not make such an assumption, and can freely produce iterates that lie in the interior of the feasible set. Additionally, whereas the dual network simplex would proceed by modifying $(\mathbf{f}, \mathbf{g})$ to produce a primal solution $\mathbf{P}$ that satisfies linear (marginal constraints) but only nonnegativity upon convergence, dual ascent builds instead a primal solution $\mathbf{P}$ that is always nonnegative but which does not necessarily satisfy marginal constraints.

3.7 Auction Algorithm

The auction algorithm was originally proposed by Bertsekas (1981) and later refined in (Bertsekas and Eckstein, 1988). Several economic



(*a*) Given balanced edges, what is the maximal flow possible? (*b*) Maxflow = 0.74

(*c*) the labeling algorithm identifies sets S, S' (*d*) total flows A, B, C, D through nodes $S, \bar{S}, S', \bar{S}'$

Figure 3.6: Consider a transportation problem involving the marginals introduced first in Figure 3.3, with $n = 5, m = 6$. Given two feasible dual vectors $\mathbf{f}, \mathbf{g}$, we try to obtain the "best" flow matrix P that is complementary to $(\mathbf{f}, \mathbf{g})$. Recall that this means that $\mathbf{P}$ can only take positive values on those edges (i, j') corresponding to indices for which $\mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}$, here represented with dotted lines in plot (a). The best flow that can be achieved with that graph structure can be formulated as a max-flow problem in a capacitated network, starting from an abstract source node s connected to all nodes labeled $i \in [\![n]\!]$, terminating at an abstract terminal node t connected to all nodes labeled j', where $j \in [\![m']\!]$, and such that the capacities of edge $(s, i), (j', t), i \in [\![n]\!], j \in [\![m]\!]$ are respectively $\mathbf{a}_i, \mathbf{b}_j$ and all others infinite. The Ford–Fulkerson algorithm (Bertsimas and Tsitsiklis, 1997, p. 305) can be applied to compute such a max-flow, which, as represented in plot (b), only achieves 0.74 units of mass out of 1 needed to solve the problem. One of the subroutines used by max-flow algorithms, the labeling algorithm (Bertsimas and Tsitsiklis, 1997, p. 308), can be used to identify nodes that receive an unsaturated flow from s (and recursively, all of its successors), denoted by orange lines in plot (c). The labeling algorithm also adds by default nodes that send a positive flow to any labeled node, which is the criterion used to select node 3, which contributes with a red line to $3'$. Labeled nodes can be grouped in sets S, S' to identify nodes which can be better exploited to obtain a higher flow, by modifying $\mathbf{f}, \mathbf{g}$ to obtain a different graph. The proof involves partial sums of flows described in plot (d)

interpretations of this algorithm have been proposed (see *e.g.* Bertsekas (1992)). The algorithm can be adapted for arbitrary marginals, but we present it here in its formulation to solve optimal assignment problems.

Complementary slackness. Notice that in the optimal assignment problem, the primal-dual conditions presented for the optimal transport problem become easier to formulate, because any extremal solution $\mathbf{P}$ is necessarily a permutation matrix $\mathbf{P}_\sigma$ for a given σ (see Equation (3.3)). Given primal $\mathbf{P}_{\sigma^\star}$ and dual $\mathbf{f}^\star, \mathbf{g}^\star$ optimal solutions we necessarily have that

$$\mathbf{f}^\star_i + \mathbf{g}^\star_{\sigma^\star_i} = \mathbf{C}_{i,\sigma^\star_i}.$$

Recall also that, because of the principle of $\mathbf{C}$-transforms enunciated in §3.2, that one can choose $\mathbf{f}^\star$ to be equal to $\mathbf{g}^{\bar{\mathbf{C}}}$. We therefore have that

$$\mathbf{C}_{i,\sigma^\star_i} - \mathbf{g}^\star_{\sigma_i} = \min_j \mathbf{C}_{i,j} - \mathbf{g}^\star_j. \tag{3.7}$$

On the contrary, it is easy to show that if there exists a vector $\mathbf{g}$ and a permutation σ such that

$$\mathbf{C}_{i,\sigma_i} - \mathbf{g}_{\sigma_i} = \min_j \mathbf{C}_{i,j} - \mathbf{g}_j \tag{3.8}$$

holds, then they are both optimal, in the sense that σ is an optimal assignment and $\mathbf{g}^{\bar{\mathbf{C}}}, \mathbf{g}$ is an optimal dual pair.

Partial assignments and ε-complementary slackness. The goal of the auction algorithm is to modify iteratively a triplet $S, \xi, \mathbf{g}$, where S is a subset of $[\![n]\!]$, ξ a partial assignment vector, namely an injective map from S to $[\![n]\!]$, and $\mathbf{g}$ a dual vector. The dual vector is meant to converge toward a solution satisfying an *approximate* complementary slackness property (3.8), whereas S grows to cover $[\![n]\!]$ as ξ describes a permutation. The algorithm works by maintaining the three following properties after each iteration:

(a) $\forall i \in S, \quad \mathbf{C}_{i,\xi_i} - \mathbf{g}_{\xi_i} \leq \varepsilon + \min_j \mathbf{C}_{i,j} - \mathbf{g}_j$ (ε-CS).

(b) The size of S can only increase at each iteration.

(c) There exists an index i such that $\mathbf{g}_i$ decreases by at least ε.

Auction algorithm updates. Given a point j the auction algorithm uses not only the optimum appearing in the usual $\mathbf{C}$-transform but also

a second best,

$$j_i^1 \in \operatorname{argmin}_j \mathbf{C}_{i,j} - \mathbf{g}_j, \quad j_i^2 \in \operatorname{argmin}_{j \neq j_i^1} \mathbf{C}_{i,j} - \mathbf{g}_j,$$

to define the following updates on $\mathbf{g}$ for an index $i \notin S$, as well as on S and ξ:

1. **update g**: Remove to the j_i^1th entry of $\mathbf{g}$ the sum of ε and the difference between the second lowest and lowest adjusted cost $\{\mathbf{C}_{i,j} - \mathbf{g}_j\}_j$,

$$\begin{aligned} \mathbf{g}_{j_i^1} &\leftarrow \mathbf{g}_{j_i^1} - \underbrace{\left((\mathbf{C}_{i,j_i^2} - \mathbf{g}_{j_i^2}) - (\mathbf{C}_{i,j_i^1} - \mathbf{g}_{j_i^1}) + \varepsilon\right)}_{\geq \varepsilon > 0} \\ &= \mathbf{C}_{i,j_i^1} - (\mathbf{C}_{i,j_i^2} - \mathbf{g}_{j_i^2}) - \varepsilon. \end{aligned} \tag{3.9}$$

2. **update S and ξ**: If there exists an index $i' \in S$ such that $\xi_{i'} = j_i^1$, remove it by updating $S \leftarrow S \setminus \{i'\}$. Set $\xi_i = j_i^1$ and add i to S, $S \leftarrow S \cup \{i\}$.

Algorithmic properties. The algorithm proceeds by starting from an empty set of assigned points $S = \emptyset$ with no assignment and empty partial assignment vector ξ, and $\mathbf{g} = \mathbb{0}_n$, terminates when $S = [\![n]\!]$, and loops through both steps above until it terminates. The fact that properties (b) and (c) are valid after each iteration is made obvious by the nature of the updates (it suffices to look at Equation (3.9)). ε-complementary slackness is easy to satisfy at the first iteration since in that case $S = \emptyset$. The fact that iterations preserve that property is shown by the following proposition.

Proposition 3.7. The auction algorithm maintains ε-complementary slackness at each iteration.

Proof. Let $\mathbf{g}, \xi, S$ be the three variables at the beginning of a given iteration. We therefore assume that for any $i' \in S$ the relationship

$$\mathbf{C}_{i,\xi_{i'}} - \mathbf{g}_{\xi_{i'}} \leq \varepsilon + \min_j \mathbf{C}_{i',j} - \mathbf{g}_j$$

holds. Consider now the particular $i \notin S$ considered in an iteration. Three updates happen: $\mathbf{g}, \xi, S$ are updated to $\mathbf{g}^{\mathrm{n}}, \xi^{\mathrm{n}}, S^{\mathrm{n}}$ using indices

j_i^1 and j_i^2. More precisely, $\mathbf{g}^{\mathrm{n}}$ is equal to $\mathbf{g}$ except for element j_i^1, whose value is equal to

$$\mathbf{g}^{\mathrm{n}}_{j_i^1} = \mathbf{g}_{j_i^1} - \Big((\mathbf{C}_{i,j_i^2} - \mathbf{g}_{j_i^2}) - (\mathbf{C}_{i,j_i^1} - \mathbf{g}_{j_i^1})\Big) - \varepsilon \leq \mathbf{g}_{j_i^1} - \varepsilon$$

, ξ^{n} is equal to ξ except for its ith element equal to j_i^1, and S^{n} is equal to the union of $\{i\}$ with S (with possibly one element removed). The update of $\mathbf{g}^{\mathrm{n}}$ can be rewritten

$$\mathbf{g}^{\mathrm{n}}_{j_i^1} = \mathbf{C}_{i,j_i^1} - (\mathbf{C}_{i,j_i^2} - \mathbf{g}_{j_i^2}) - \varepsilon;$$

therefore we have

$$\mathbf{C}_{i,j_i^1} - \mathbf{g}^{\mathrm{n}}_{j_i^1} = \varepsilon + (\mathbf{C}_{i,j_i^2} - \mathbf{g}_{j_i^2}) = \varepsilon + \min_{j \neq j_i^1}(\mathbf{C}_{i,j} - \mathbf{g}_j).$$

Since $-\mathbf{g} \leq -\mathbf{g}^{\mathrm{n}}$ this implies that

$$\mathbf{C}_{i,j_i^1} - \mathbf{g}^{\mathrm{n}}_{j_i^1} = \varepsilon + \min_{j \neq j_i^1}(\mathbf{C}_{i,j} - \mathbf{g}_j) \leq \varepsilon + \min_{j \neq j_i^1}(\mathbf{C}_{i,j} - \mathbf{g}^{\mathrm{n}}_j),$$

and since the inequality is also obviously true for $j = j_i^1$ we therefore obtain the ε-complementary slackness property for index i. For other indices $i' \neq i$, we have again that since $\mathbf{g}^{\mathrm{n}} \leq \mathbf{g}$ the sequence of inequalities holds,

$$\mathbf{C}_{i,\xi^{\mathrm{n}}_{i'}} - \mathbf{g}^{\mathrm{n}}_{\xi^{\mathrm{n}}_{i'}} = \mathbf{C}_{i,\xi_{i'}} - \mathbf{g}_{\xi_{i'}} \leq \varepsilon + \min_j \mathbf{C}_{i',j} - \mathbf{g}_j \leq \varepsilon + \min_j \mathbf{C}_{i',j} - \mathbf{g}^n_j.$$

□

Proposition 3.8. The number of steps of the auction algorithm is at most $N = n\|\mathbf{C}\|_\infty/\varepsilon$.

Proof. Suppose that the algorithm has not stopped after $T > N$ steps. Then there exists an index j which is not in the image of ξ, namely whose price coordinate $\mathbf{g}_j$ has never been updated and is still $\mathbf{g}_j = 0$. In that case, there cannot exist an index j' such that $\mathbf{g}_{j'}$ was updated n times with $n > \|\mathbf{C}\|_\infty/\varepsilon$. Indeed, if that were the case then for any index i

$$\mathbf{g}_{j'} \leq -n\varepsilon < -\|\mathbf{C}\|_\infty \leq -\mathbf{C}_{i,j} = \mathbf{g}_j - \mathbf{C}_{i,j},$$

which would result in, for all i,

$$\mathbf{C}_{i,j'} - \mathbf{g}_{j'} > \mathbf{C}_{i,j} + (\mathbf{C}_{i,j} - \mathbf{g}_j),$$

which contradicts ε-CS. Therefore, since there cannot be more than $\|C\|_\infty/\varepsilon$ updates for each variable, T cannot be larger than $n\|\mathbf{C}\|_\infty/\varepsilon = N$. □

Remark 3.3. Note that this result yields a naive number of operations of $N^3\|\mathbf{C}\|_\infty/\varepsilon$ for the algorithm to terminate. That complexity can be reduced to $N^3 \log \|\mathbf{C}\|_\infty$ when using a clever method known as ε-scaling, designed to decrease the value of ε with each iteration (Bertsekas, 1998, p. 264).

Proposition 3.9. The auction algorithm finds an assignment whose cost is $n\varepsilon$ suboptimal.

Proof. Let $\sigma, \mathbf{g}^\star$ be the primal and dual optimal solutions of the assignment problem of matrix $\mathbf{C}$, with optimum

$$t^\star = \sum \mathbf{C}_{i,\sigma_i} = \sum_i \min_j \mathbf{C}_{i,j} - \mathbf{g}_j^\star + \sum_j \mathbf{g}_j^\star.$$

Let $\xi, \mathbf{g}$ be the solutions output by the auction algorithm upon termination. The ε-CS conditions yield that for any $i \in S$,

$$\min_j \mathbf{C}_{i,j} - \mathbf{g}_j \geq \mathbf{C}_{i,\xi_i} - \mathbf{g}_{\xi_i} - \varepsilon.$$

Therefore by simple suboptimality of $\mathbf{g}$ we first have

$$\begin{aligned} t^\star &\geq \sum_i \left(\min_j \mathbf{C}_{i,j} - \mathbf{g}_j\right) + \sum_j \mathbf{g}_j \\ &\geq \sum_i -\varepsilon + \left(\mathbf{C}_{i,\xi_i} - \mathbf{g}_{\xi_i}\right) + \sum_j \mathbf{g}_j = -n\varepsilon + \sum_i \mathbf{C}_{i,\xi_j} \geq -n\varepsilon + t^\star. \end{aligned},$$

where the second inequality comes from ε-CS, the next equality by cancellation of the sum of terms in $\mathbf{g}_{\xi_i}$ and $\mathbf{g}_j$, and the last inequality by the suboptimality of ξ as a permutation. □

The auction algorithm can therefore be regarded as an alternative way to use the machinery of $\mathbf{C}$-transforms. Next we explore another approach grounded on regularization, the so-called Sinkhorn algorithm, which also bears similarities with the auction algorithm as discussed in (Schmitzer, 2016b).

Note finally that, on low-dimensional regular grids in Euclidean space, it is possible to couple these classical linear solvers with multiscale strategies, to obtain a significant speed-up (Schmitzer, 2016a; Oberman and Ruan, 2015).

6

$\mathcal{W}_1$ Optimal Transport

This chapter focuses on optimal transport problems in which the ground cost is equal to a distance. Historically, this corresponds to the original problem posed by Monge in 1781; this setting was also that chosen in early applications of optimal transport in computer vision (Rubner *et al.*, 2000) under the name of "earth mover's distances".

Unlike the case where the ground cost is a *squared* Hilbertian distance (studied in particular in Chapter 7), transport problems where the cost is a metric are more difficult to analyze theoretically. In contrast to Remark 2.24 that states the uniqueness of a transport map or coupling between two absolutely continuous measures when using a squared metric, the optimal Kantorovich coupling is in general not unique when the cost is the ground distance itself. Hence, in this regime it is often impossible to recover a uniquely defined Monge map, making this class of problems ill-suited for interpolation of measures. We refer to works by Trudinger and Wang (2001), Caffarelli *et al.* (2002), Sudakov (1979), and Evans and Gangbo (1999) for proofs of existence of optimal $\mathcal{W}_1$ transportation plans and detailed analyses of their geometric structure.

Although more difficult to analyze in theory, optimal transport with a linear ground distance is usually more robust to outliers and noise

than a quadratic cost. Furthermore, a cost that is a metric results in an elegant dual reformulation involving local flow, divergence constraints, or Lipschitzness of the dual potential, suggesting cheaper numerical algorithms that align with *minimum-cost flow* methods over networks in graph theory. This setting is also popular because the associated OT distances define a norm that can compare arbitrary distributions, even if they are not positive; this property is shared by a larger class of so-called *dual norms* (see §8.2 and Remark 10.6 for more details).

6.1 $\mathcal{W}_1$ on Metric Spaces

Here we assume that d is a distance on $\mathcal{X} = \mathcal{Y}$, and we solve the OT problem with the ground cost $c(x, y) = d(x, y)$. The following proposition highlights key properties of the c-transform (5.1) in this setup. In the following, we denote the Lipschitz constant of a function $f \in \mathcal{C}(\mathcal{X})$ as

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

We define Lipschitz functions to be those functions f satisfying $\mathrm{Lip}(f) < +\infty$; they form a convex subset of $\mathcal{C}(\mathcal{X})$.

Proposition 6.1. Suppose $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = d(x, y)$. Then, there exists g such that $f = g^c$ if and only $\mathrm{Lip}(f) \leq 1$. Furthermore, if $\mathrm{Lip}(f) \leq 1$, then $f^c = -f$.

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

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

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

Now, suppose $\mathrm{Lip}(f) \leq 1$, and define $g \stackrel{\text{def.}}{=} -f$. By the Lipschitz property, for all $x, y \in \mathcal{X}$, $f(y) - d(x, y) \leq f(x) \leq f(y) + d(x, y)$.

Applying these inequalities,

$$g^c(y) = \inf_{x\in\mathcal{X}} [d(x,y) + f(x)] \geq \inf_{x\in\mathcal{X}} [d(x,y) + f(y) - d(x,y)] = f(y),$$
$$g^c(y) = \inf_{x\in\mathcal{X}} [d(x,y) + f(x)] \leq \inf_{x\in\mathcal{X}} [d(x,y) + f(y) + d(x,y)] = f(y).$$

Hence, $f = g^c$ with $g = -f$. Using the same inequalities shows

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

This shows $f^c = -f$. □

Starting from the single potential formulation (5.4), one can iterate the construction and replace the couple (g, g^c) by $(g^c, (g^c)^c)$. The last proposition shows that one can thus use $(g^c, -g^c)$, which in turn is equivalent to any pair $(f, -f)$ such that $\mathrm{Lip}(f) \leq 1$. This leads to the following alternative expression for the $\mathcal{W}_1$ distance:

$$\mathcal{W}_1(\alpha,\beta) = \max_f \left\{ \int_{\mathcal{X}} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x)) \;:\; \mathrm{Lip}(f) \leq 1 \right\}. \quad (6.1)$$

This expression shows that $\mathcal{W}_1$ is actually a norm, *i.e.* $\mathcal{W}_1(\alpha,\beta) = \|\alpha - \beta\|_{\mathcal{W}_1}$, and that it is still valid for any measures (not necessary positive) as long as $\int_{\mathcal{X}} \alpha = \int_{\mathcal{X}} \beta$. This norm is often called the Kantorovich and Rubinstein norm (1958).

For discrete measures of the form (2.1), writing $\alpha - \beta = \sum_k \mathbf{m}_k \delta_{z_k}$ with $z_k \in \mathcal{X}$ and $\sum_k \mathbf{m}_k = 0$, the optimization (6.1) can be rewritten as

$$\mathcal{W}_1(\alpha,\beta) = \max_{(\mathbf{f}_k)_k} \left\{ \sum_k \mathbf{f}_k \mathbf{m}_k \;:\; \forall\, (k,\ell), |\mathbf{f}_k - \mathbf{f}_\ell| \leq d(z_k, z_\ell), \right\} \quad (6.2)$$

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

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

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

which is a linear program. Note that furthermore, in this 1-D case, a closed form expression for $\mathcal{W}_1$ using cumulative functions is given in (2.37).

Remark 6.1 ($\mathcal{W}_p$ with $0 < p \leq 1$)**.** If $0 < p \leq 1$, then $\tilde{d}(x,y) \overset{\text{def.}}{=} d(x,y)^p$ satisfies the triangular inequality, and hence $\tilde{d}$ is itself a distance. One can thus apply the results and algorithms detailed above for $\mathcal{W}_1$ to compute $\mathcal{W}_p$ by simply using $\tilde{d}$ in place of d. This is equivalent to stating that $\mathcal{W}_p$ is the dual of p-Hölder functions $\{f \ : \ \mathrm{Lip}_p(f) \leq 1\}$, where

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

6.2 $\mathcal{W}_1$ on Euclidean Spaces

In the special case of Euclidean spaces $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, using $c(x,y) = \|x - y\|$, the global Lipschitz constraint appearing in (6.1) can be made local as a uniform bound on the gradient of f,

$$\mathcal{W}_1(\alpha,\beta) = \max_f \left\{ \int_{\mathbb{R}^d} f(x)(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x)) \ : \ \|\nabla f\|_\infty \leq 1 \right\}. \quad (6.3)$$

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

Considering the dual problem to (6.3), one obtains an optimization problem under fixed divergence constraint

$$\mathcal{W}_1(\alpha,\beta) = \min_s \left\{ \int_{\mathbb{R}^d} \|s(x)\|_2 \, \mathrm{d}x \ : \ \mathrm{div}(s) = \alpha - \beta \right\}, \quad (6.4)$$

which is often called the Beckmann formulation (Beckmann, 1952). Here the vectorial function $s(x) \in \mathbb{R}^2$ can be interpreted as a flow field, describing locally the movement of mass. Outside the support of the two input measures, $\mathrm{div}(s) = 0$, which is the conservation of mass constraint. Once properly discretized using finite elements, Problems (6.3) and (6.4) become nonsmooth convex optimization problems. It is possible to use an off-the-shelf interior points quadratic-cone optimization solver, but as advocated in §7.3, large-scale problems require the use of simpler but more adapted first order methods. One can thus use, for instance,

Douglas–Rachford (DR) iterations (7.14) or the related alternating direction method of multipliers method. Note that on a uniform grid, projecting on the divergence constraint is conveniently handled using the fast Fourier transform. We refer to Solomon *et al.* (2014a) for a detailed account for these approaches and application to OT on triangulated meshes. See also Li *et al.* (2018a), Ryu *et al.* (2017b), and Ryu *et al.* (2017a) for similar approaches using primal-dual splitting schemes. Approximation schemes that relax the Lipschitz constraint on the dual potentials f have also been proposed, using, for instance, a constraint on wavelet coefficients leading to an explicit formula (Shirdhonkar and Jacobs, 2008), or by approximating f by a multilayer neural network with rectified linear activations and clipped weights (Arjovsky *et al.*, 2017).

6.3 $\mathcal{W}_1$ on a Graph

The previous formulations (6.3) and (6.4) of $\mathcal{W}_1$ can be generalized to the setting where $\mathcal{X}$ is a geodesic space, *i.e.* $c(x,y) = d(x,y)$ where d is a geodesic distance. We refer to Feldman and McCann (2002) for a theoretical analysis in the case where $\mathcal{X}$ is a Riemannian manifold. When $\mathcal{X} = [\![1,n]\!]$ is a discrete set, equipped with undirected edges $(i,j) \in \mathcal{E} \subset \mathcal{X}^2$ labeled with a weight (length) $\mathbf{w}_{i,j}$, we recover the important case where $\mathcal{X}$ is a graph equipped with the geodesic distance (or shortest path metric):

$$\mathbf{D}_{i,j} \overset{\text{def.}}{=} \min_{K \geq 0, (i_k)_k : i \to j} \left\{ \sum_{k=1}^{K-1} \mathbf{w}_{i_k, i_{k+1}} \;:\; \forall\, k \in [\![1, K-1]\!], (i_k, i_{k+1}) \in \mathcal{E} \right\},$$

where $i \to j$ indicates that $i_1 = i$ and $i_K = j$, namely that the path starts at i and ends at j.

We consider two vectors $(\mathbf{a}, \mathbf{b}) \in (\mathbb{R}^n)^2$ defining (signed) discrete measures on the graph $\mathcal{X}$ such that $\sum_i \mathbf{a}_i = \sum_i \mathbf{b}_i$ (these weights do not need to be positive). The goal is now to compute $\mathrm{W}_1(\mathbf{a}, \mathbf{b})$, as introduced in (2.17) for $p = 1$, when the ground metric is the graph geodesic distance. This computation should be carried out without going as far as having to compute a "full" coupling $\mathbf{P}$ of size $n \times n$, to rely instead on local operators thanks to the underlying connectivity of the

graph. These operators are discrete formulations for the gradient and divergence differential operators.

A discrete dual Kantorovich potential $\mathbf{f} \in \mathbb{R}^n$ is a vector indexed by all vertices of the graph. The gradient operator $\nabla : \mathbb{R}^n \to \mathbb{R}^{\mathcal{E}}$ is defined as

$$\forall\, (i,j) \in \mathcal{E}, \quad (\nabla \mathbf{f})_{i,j} \overset{\text{def.}}{=} \mathbf{f}_i - \mathbf{f}_j.$$

A flow $\mathbf{s} = (\mathbf{s}_{i,j})_{i,j}$ is defined on edges, and the divergence operator $\mathrm{div} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^n$, which is the adjoint of the gradient ∇, maps flows to vectors defined on vertices and is defined as

$$\forall\, i \in [\![1,n]\!], \quad \mathrm{div}(\mathbf{s})_i \overset{\text{def.}}{=} \sum_{j:(i,j)\in\mathcal{E}} (\mathbf{s}_{i,j} - \mathbf{s}_{j,i}) \in \mathbb{R}^n.$$

Problem (6.3) becomes, in the graph setting,

$$\mathrm{W}_1(\mathbf{a},\mathbf{b}) = \max_{\mathbf{f}\in\mathbb{R}^n} \left\{ \sum_{i=1}^n \mathbf{f}_i(\mathbf{a}_i - \mathbf{b}_i) \;:\; \forall (i,j)\in\mathcal{E},\, |(\nabla\mathbf{f})_{i,j}| \leq \mathbf{w}_{i,j} \right\}. \tag{6.5}$$

The associated dual problem, which is analogous to Formula (6.4), is then

$$\mathrm{W}_1(\mathbf{a},\mathbf{b}) = \min_{\mathbf{s}\in\mathbb{R}_+^{\mathcal{E}}} \left\{ \sum_{(i,j)\in\mathcal{E}} \mathbf{w}_{i,j}\mathbf{s}_{i,j} \;:\; \mathrm{div}(\mathbf{s}) = \mathbf{a} - \mathbf{b} \right\}. \tag{6.6}$$

This is a linear program and more precisely an instance of min-cost flow problems. Highly efficient dedicated simplex solvers have been devised to solve it; see, for instance, (Ling and Okada, 2007). Figure 6.1 shows an example of primal and dual solutions. Formulation (6.6) is the so-called Beckmann formulation (Beckmann, 1952) and has been used and extended to define and study traffic congestion models; see, for instance, (Carlier *et al.*, 2008).

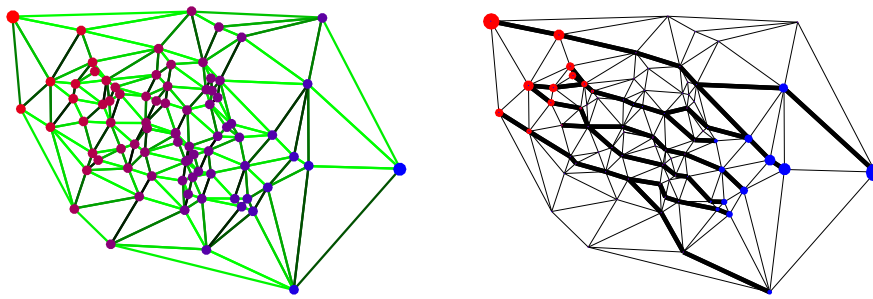


Figure 6.1: Example of computation of $\mathrm{W}_1(\mathbf{a}, \mathbf{b})$ on a planar graph with uniform weights $\mathbf{w}_{i,j} = 1$. Left: potential $\mathbf{f}$ solution of (6.5) (increasing value from red to blue). The green color of the edges is proportional to $|(\nabla \mathbf{f})_{i,j}|$. Right: flow $\mathbf{s}$ solution of (6.6), where bold black edges display nonzero $\mathbf{s}_{i,j}$, which saturate to $\mathbf{w}_{i,j} = 1$. These saturating flow edge on the right match the light green edge on the left where $|(\nabla \mathbf{f})_{i,j}| = 1$.

8

Statistical Divergences

We study in this chapter the statistical properties of the Wasserstein distance. More specifically, we compare it to other major distances and divergences routinely used in data sciences. We quantify how one can approximate the distance between two probability distributions when having only access to samples from said distributions. To introduce these subjects, §8.1 and §8.2 review respectively divergences and integral probability metrics between probability distributions. A divergence D typically satisfies $D(\alpha, \beta) \geq 0$ and $D(\alpha, \beta) = 0$ if and only if $\alpha = \beta$, but it does not need to be symmetric or satisfy the triangular inequality. An integral probability metric for measures is a dual norm defined using a prescribed family of test functions. These quantities are sound alternatives to Wasserstein distances and are routinely used as loss functions to tackle inference problems, as will be covered in §9. We show first in §8.3 that the optimal transport distance is not Hilbertian, *i.e.* one cannot approximate it efficiently using a Hilbertian metric on a suitable feature representation of probability measures. We show in §8.4 how to approximate $D(\alpha, \beta)$ from discrete samples $(x_i)_i$ and $(y_j)_j$ drawn from α and β. A good statistical understanding of that problem is crucial when using the Wasserstein distance in machine learning. Note that this

section will be chiefly concerned with the statistical approximation of optimal transport between distributions supported on *continuous* sets. The very same problem when the ground space is finite has received some attention in the literature following the work of Sommerfeld and Munk (2018), extended to entropic regularized quantities by Bigot *et al.* (2017a).

8.1 φ-Divergences

Before detailing in the following section "weak" norms, whose construction shares similarities with $\mathcal{W}_1$, let us detail a generic construction of so-called divergences between measures, which can then be used as loss functions when estimating probability distributions. Such divergences compare two input measures by comparing their mass *pointwise*, without introducing any notion of mass transportation. Divergences are functionals which, by looking at the pointwise ratio between two measures, give a sense of how close they are. They have nice analytical and computational properties and build upon *entropy functions*.

Definition 8.1 (Entropy function)**.** A function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is an entropy function if it is lower semicontinuous, convex, $\operatorname{dom}\varphi \subset [0, \infty[$, and satisfies the following feasibility condition: $\operatorname{dom}\varphi \cap]0, \infty[\neq \emptyset$. The speed of growth of φ at ∞ is described by

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

If $\varphi'_\infty = \infty$, then φ grows faster than any linear function and φ is said *superlinear*. Any entropy function φ induces a φ-divergence (also known as Ciszár divergence (Ciszár, 1967; Ali and Silvey, 1966) or f-divergence) as follows.

Definition 8.2 (φ-Divergences)**.** Let φ be an entropy function. For $\alpha, \beta \in \mathcal{M}(\mathcal{X})$, let $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^\perp$ be the Lebesgue decomposition[1] of α with respect to β. The divergence $\mathcal{D}_\varphi$ is defined by

$$\mathcal{D}_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + \varphi'_\infty \alpha^\perp(\mathcal{X}) \tag{8.1}$$

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

if α, β are nonnegative and ∞ otherwise.

The additional term $\varphi'_\infty \alpha^\perp(\mathcal{X})$ in (8.1) is important to ensure that $\mathcal{D}_\varphi$ defines a continuous functional (for the weak topology of measures) even if φ has a linear growth at infinity, as this is, for instance, the case for the absolute value (8.8) defining the TV norm. If φ as a superlinear growth, *e.g.* the usual entropy (8.4), then $\varphi'_\infty = +\infty$ so that $\mathcal{D}_\varphi(\alpha|\beta) = +\infty$ if α does not have a density with respect to β.

In the discrete setting, assuming

$$\alpha = \sum_i \mathbf{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_i \mathbf{b}_i \delta_{x_i} \tag{8.2}$$

are supported on the same set of n points $(x_i)_{i=1}^n \subset \mathcal{X}$, (8.1) defines a divergence on Σ_n

$$\mathbf{D}_\varphi(\mathbf{a}|\mathbf{b}) = \sum_{i \in \mathrm{Supp}(\mathbf{b})} \varphi\left(\frac{\mathbf{a}_i}{\mathbf{b}_i}\right) \mathbf{b}_i + \varphi'_\infty \sum_{i \notin \mathrm{Supp}(\mathbf{b})} \mathbf{a}_i, \tag{8.3}$$

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

The proof of the following proposition can be found in Liero *et al.*, 2018, Thm 2.7.

Proposition 8.1. If φ is an entropy function, then $\mathcal{D}_\varphi$ is jointly 1-homogeneous, convex and weakly* lower semicontinuous in (α, β).

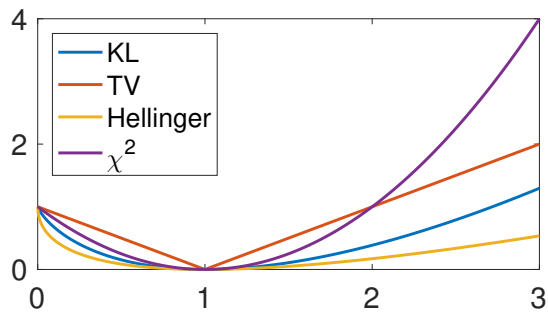


Figure 8.1: Example of entropy functionals.

Remark 8.1 (Dual expression)**.** A φ-divergence can be expressed using the Legendre transform

$$\varphi^*(s) \overset{\text{def.}}{=} \sup_{t \in \mathbb{R}} st - \varphi(t)$$

of φ (see also (4.54)) as

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

see Liero *et al.* (2018) for more details.

We now review a few popular instances of this framework. Figure 8.1 displays the associated entropy functionals, while Figure 8.2 reviews the relationship between them.

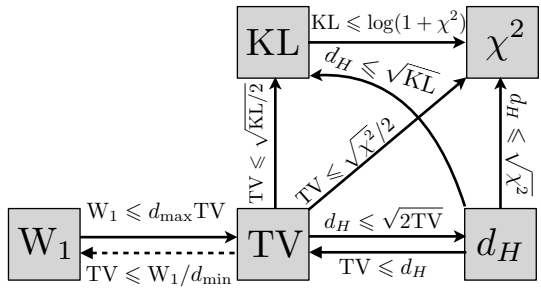


Figure 8.2: Diagram of relationship between divergences (inspired by Gibbs and Su (2002)). For $\mathcal{X}$ a metric space with ground distance d, $d_{\max} = \sup_{(x,x')} d(x,x')$ is the diameter of $\mathcal{X}$. When $\mathcal{X}$ is discrete, $d_{\min} \stackrel{\text{def.}}{=} \min_{x\neq x'} d(x,x')$.

Example 8.1 (Kullback–Leibler divergence)**.** The Kullback–Leibler divergence $\mathrm{KL} \stackrel{\text{def.}}{=} \mathcal{D}_{\varphi_{\mathrm{KL}}}$, also known as the relative entropy, was already introduced in (4.10) and (4.6). It is the divergence associated to the Shannon–Boltzman entropy function φ_{KL}, given by

$$\varphi_{\mathrm{KL}}(s) = \begin{cases} s\log(s) - s + 1 & \text{for } s > 0, \\ 1 & \text{for } s = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{8.4}$$

Remark 8.1 (Bregman divergence)**.** The discrete KL divergence, $\mathbf{KL} \stackrel{\text{def.}}{=} \mathbf{D}_{\varphi_{\mathrm{KL}}}$, has the unique property of being both a φ-divergence and a Bregman divergence. For discrete vectors in $\mathbb{R}^n$, a Bregman divergence (Bregman, 1967) associated to a smooth strictly convex function $\psi : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\mathbf{B}_\psi(\mathbf{a}|\mathbf{b}) \stackrel{\text{def.}}{=} \psi(\mathbf{a}) - \psi(\mathbf{b}) - \langle \nabla\psi(\mathbf{b}), \mathbf{a}-\mathbf{b}\rangle, \tag{8.5}$$

where $\langle \cdot, \cdot \rangle$ is the canonical inner product on $\mathbb{R}^n$. Note that $\mathbf{B}_\psi(\mathbf{a}|\mathbf{b})$ is a convex function of $\mathbf{a}$ and a linear function of ψ. Similarly to φ-divergence, a Bregman divergence satisfies $\mathbf{B}_\psi(\mathbf{a}|\mathbf{b}) \geq 0$ and $\mathbf{B}_\psi(\mathbf{a}|\mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$. The KL divergence is the Bregman divergence for minus the entropy $\psi = -\mathbf{H}$ defined in (4.1)), *i.e.* $\mathbf{KL} = \mathbf{B}_{-\mathbf{H}}$. A Bregman divergence is locally a squared Euclidean distance since

$$\mathbf{B}_\psi(\mathbf{a} + \varepsilon|\mathbf{a} + \eta) = \langle \partial^2 \psi(\mathbf{a})(\varepsilon - \eta), \varepsilon - \eta \rangle + o(\|\varepsilon - \eta\|^2)$$

and the set of separating points $\{\mathbf{a} \ : \ \mathbf{B}_\psi(\mathbf{a}|\mathbf{b}) = \mathbf{B}_\psi(\mathbf{a}|\mathbf{b}')\}$ is a hyperplane between $\mathbf{b}$ and $\mathbf{b}'$. These properties make Bregman divergence suitable to replace Euclidean distances in first order optimization methods. The best know example is mirror gradient descent (Beck and Teboulle, 2003), which is an explicit descent step of the form (9.32). Bregman divergences are also important in convex optimization and can be used, for instance, to derive Sinkhorn iterations and study its convergence in finite dimension; see Remark 4.8.

Remark 8.2 (Hyperbolic geometry of KL)**.** It is interesting to contrast the geometry of the Kullback–Leibler divergence to that defined by quadratic optimal transport when comparing Gaussians. As detailed, for instance, by Costa *et al.* (2015), the Kullback–Leibler divergence has a closed form for Gaussian densities. In the univariate case, $d = 1$, if $\alpha = \mathcal{N}(m_\alpha, \sigma_\alpha^2)$ and $\beta = \mathcal{N}(m_\beta, \sigma_\beta^2)$, one has

$$\mathrm{KL}(\alpha|\beta) = \frac{1}{2}\left(\frac{\sigma_\alpha^2}{\sigma_\beta^2} + \log\left(\frac{\sigma_\beta^2}{\sigma_\alpha^2}\right) + \frac{|m_\alpha - m_\beta|}{\sigma_\beta^2} - 1 \right). \tag{8.6}$$

This expression shows that the divergence between α and β diverges to infinity as σ_β diminishes to 0 and β becomes a Dirac mass. In that sense, one can say that singular Gaussians are infinitely far from all other Gaussians in the KL geometry. That geometry is thus useful when one wants to avoid dealing with singular covariances. To simplify the analysis, one can look at the infinitesimal geometry of KL, which is obtained by performing a Taylor expansion at order 2,

$$\mathrm{KL}(\mathcal{N}(m + \delta_m, (\sigma + \delta_\sigma)^2)|\mathcal{N}(m, \sigma^2)) = \frac{1}{\sigma^2}\left(\frac{1}{2}\delta_m^2 + \delta_\sigma^2\right) + o(\delta_m^2, \delta_\sigma^2).$$

This local Riemannian metric, the so-called Fisher metric, expressed over $(m/\sqrt{2}, \sigma) \in \mathbb{R} \times \mathbb{R}_{+,*}$, matches exactly that of the hyperbolic Poincaré half plane. Geodesics over this space are half circles centered along the $\sigma = 0$ line and have an exponential speed, *i.e.* they only reach the limit $\sigma = 0$ after an infinite time. Note in particular that if $\sigma_\alpha = \sigma_\beta$ but $m_\alpha \neq m_\alpha$, then the geodesic between (α, β) over this hyperbolic half plane does not have a constant standard deviation.

The KL hyperbolic geometry over the space of Gaussian parameters (m, σ) should be contrasted with the Euclidean geometry associated to OT as described in Remark 2.31, since in the univariate case

$$\mathcal{W}_2^2(\alpha, \beta) = |m_\alpha - m_\beta|^2 + |\sigma_\alpha - \sigma_\beta|^2. \tag{8.7}$$

Figure 8.3 shows a visual comparison of these two geometries and their respective geodesics. This interesting comparison was suggested to us by Jean Feydy.

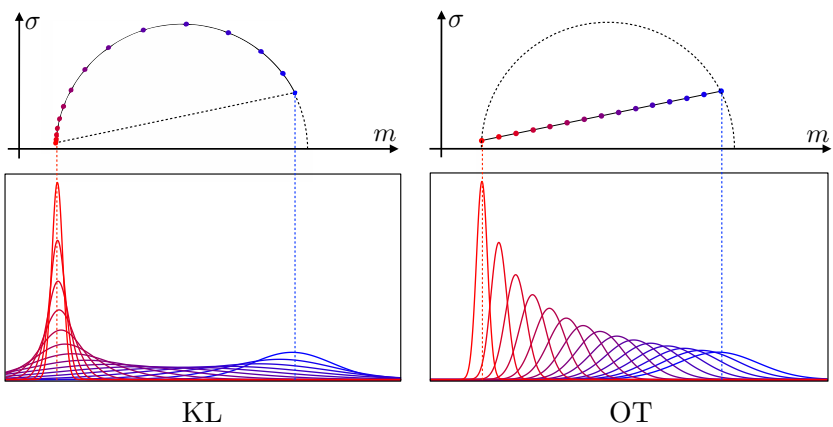


Figure 8.3: Comparisons of interpolation between Gaussians using KL (hyperbolic) and OT (Euclidean) geometries.

Example 8.2 (Total variation)**.** The total variation distance $\mathrm{TV} \overset{\text{def.}}{=} \mathcal{D}_{\varphi_{\mathrm{TV}}}$ is the divergence associated to

$$\varphi_{\mathrm{TV}}(s) = \begin{cases} |s-1| & \text{for } s \geq 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{8.8}$$

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

$$\mathrm{TV}(\alpha|\beta) = \|\alpha - \beta\|_{\mathrm{TV}}, \quad \text{where} \quad \|\alpha\|_{\mathrm{TV}} = |\alpha|(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}|\alpha|(x). \tag{8.9}$$

If α has a density ρ_α on $\mathcal{X} = \mathbb{R}^d$, then the TV norm is the L^1 norm on functions, $\|\alpha\|_{\mathrm{TV}} = \int_{\mathcal{X}} |\rho_\alpha(x)| \mathrm{d}x = \|\rho_\alpha\|_{L^1}$. If α is discrete as in (8.2), then the TV norm is the ℓ^1 norm of vectors in $\mathbb{R}^n$, $\|\alpha\|_{\mathrm{TV}} = \sum_i |\mathbf{a}_i| = \|\mathbf{a}\|_{\ell^1}$.

Remark 8.2 (Strong vs. weak topology)**.** The total variation norm (8.9) defines the so-called "strong" topology on the space of measure. On a compact domain $\mathcal{X}$ of radius R, one has

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

so that this strong notion of convergence implies the weak convergence metrized by Wasserstein distances. The converse is, however, not true, since δ_x does not converge strongly to δ_y if $x \to y$ (note that $\|\delta_x - \delta_y\|_{\mathrm{TV}} = 2$ if $x \neq y$). A chief advantage is that $\mathcal{M}_+^1(\mathcal{X})$ (once again on a compact ground space $\mathcal{X}$) is compact for the weak topology, so that from any sequence of probability measures $(\alpha_k)_k$, one can always extract a converging subsequence, which makes it a suitable space for several optimization problems, such as those considered in Chapter 9.

Example 8.3 (Hellinger)**.** The Hellinger distance $\mathfrak{h} \overset{\text{def.}}{=} \mathcal{D}_{\varphi_H}^{1/2}$ is the square root of the divergence associated to

$$\varphi_H(s) = \begin{cases} |\sqrt{s} - 1|^2 & \text{for } s \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

As its name suggests, $\mathfrak{h}$ is a distance on $\mathcal{M}_+(\mathcal{X})$, which metrizes the strong topology as $\|\cdot\|_{\mathrm{TV}}$. If (α, β) have densities $(\rho_\alpha, \rho_\beta)$ on $\mathcal{X} = \mathbb{R}^d$, then $\mathfrak{h}(\alpha, \beta) = \|\sqrt{\rho_\alpha} - \sqrt{\rho_\beta}\|_{L^2}$. If (α, β) are discrete as in (8.2), then $\mathfrak{h}(\alpha, \beta) = \|\sqrt{\mathbf{a}} - \sqrt{\mathbf{b}}\|$. Considering $\varphi_{L^p}(s) = |s^{1/p} - 1|^p$ generalizes the Hellinger ($p = 2$) and total variation ($p = 1$) distances and $\mathcal{D}_{\varphi_{L^p}}^{1/p}$ is a distance which metrizes the strong convergence for $0 < p < +\infty$.

Example 8.4 (Jensen–Shannon distance)**.** The KL divergence is not symmetric and, while being a Bregman divergence (which are locally

quadratic norms), it is not the square of a distance. On the other hand, the Jensen–Shannon distance $\mathrm{JS}(\alpha,\beta)$, defined as

$$\mathrm{JS}(\alpha,\beta)^2 \overset{\text{def.}}{=} \frac{1}{2}\left(\mathrm{KL}(\alpha|\xi)+\mathrm{KL}(\beta|\xi)\right) \quad \text{where} \quad \xi=\frac{\alpha+\beta}{2},$$

is a distance (Endres and Schindelin, 2003; Österreicher and Vajda, 2003). JS^2 can be shown to be a φ-divergence for $\varphi(s)=t\log(t)-(t+1)\log(t+1)$. In sharp contrast with KL, $\mathrm{JS}(\alpha,\beta)$ is always bounded; more precisely, it satisfies $0\leq \mathrm{JS}(\alpha,\beta)^2\leq \ln(2)$. Similarly to the TV norm and the Hellinger distance, it metrizes the strong convergence.

Example 8.5 (χ^2)**.** The χ^2-divergence $\chi^2 \overset{\text{def.}}{=} \mathcal{D}_{\varphi_{\chi^2}}$ is the divergence associated to

$$\varphi_{\chi^2}(s)=\begin{cases}|s-1|^2 & \text{for } s\geq 0,\\ +\infty & \text{otherwise.}\end{cases}$$

If (α,β) are discrete as in (8.2) and have the same support, then

$$\chi^2(\alpha|\beta)=\sum_i \frac{(\mathbf{a}_i-\mathbf{b}_i)^2}{\mathbf{b}_i}.$$

8.2 Integral Probability Metrics

Formulation (6.3) is a special case of a dual norm. A dual norm is a convenient way to design "weak" norms that can deal with arbitrary measures. For a symmetric convex set B of measurable functions, one defines

$$\|\alpha\|_B \overset{\text{def.}}{=} \max_f \left\{\int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) \;:\; f\in B\right\}. \tag{8.10}$$

These dual norms are often called "integral probability metrics'; see (Sriperumbudur *et al.*, 2012).

Example 8.6 (Total variation)**.** The total variation norm (Example 8.2) is a dual norm associated to the whole space of continuous functions

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

The total variation distance is the only nontrivial divergence that is also a dual norm; see (Sriperumbudur *et al.*, 2009).

Remark 8.3 (Metrizing the weak convergence)**.** By using smaller "balls" B, which typically only contain continuous (and sometimes regular) functions, one defines weaker dual norms. In order for $\|\cdot\|_B$ to metrize the weak convergence (see Definition 2.2), it is sufficient for the space spanned by B to be dense in the set of continuous functions for the sup-norm $\|\cdot\|_\infty$ (i.e. for the topology of uniform convergence); see (Ambrosio *et al.*, 2006, para. 5.1).

Figure 8.4 displays a comparison of several such dual norms, which we now detail.

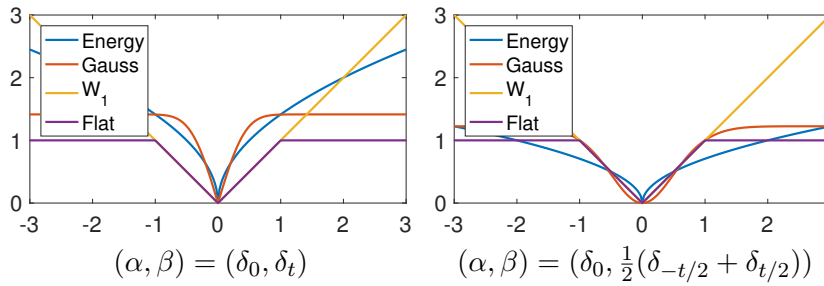


Figure 8.4: Comparison of dual norms.

8.2.1 $\mathcal{W}_1$ and Flat Norm

If the set B is bounded, then $\|\cdot\|_B$ is a norm on the whole space $\mathcal{M}(\mathcal{X})$ of measures. This is not the case of $\mathcal{W}_1$, which is only defined for α such that $\int_{\mathcal{X}} \mathrm{d}\alpha = 0$ (otherwise $\|\alpha\|_B = +\infty$). This can be alleviated by imposing a bound on the value of the potential f, in order to define for instance the flat norm.

Example 8.7 ($\mathcal{W}_1$ norm)**.** $\mathcal{W}_1$ as defined in (6.3), is a special case of dual norm (8.10), using

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

the set of 1-Lipschitz functions.

Example 8.8 (Flat norm and Dudley metric)**.** The flat norm is defined using

$$B = \{f \;:\; \|\nabla f\|_\infty \leq 1 \quad \text{and} \quad \|f\|_\infty \leq 1\}\,. \tag{8.11}$$

It metrizes the weak convergence on the whole space $\mathcal{M}(\mathcal{X})$. Formula (6.2) is extended to compute the flat norm by adding the constraint $|\mathbf{f}_k| \leq 1$. The flat norm is sometimes called the "Kantorovich–Rubinstein" norm (Hanin, 1992) and has been used as a fidelity term for inverse problems in imaging (Lellmann *et al.*, 2014). The flat norm is similar to the Dudley metric, which uses

$$B = \{f \;:\; \|\nabla f\|_\infty + \|f\|_\infty \leq 1\}.$$

8.2.2 Dual RKHS Norms and Maximum Mean Discrepancies

It is also possible to define "Euclidean" norms (built using quadratic functionals) on measures using the machinery of kernel methods and more specifically reproducing kernel Hilbert spaces (RKHS; see (Schölkopf and Smola, 2002) for a survey of their applications in data sciences), of which we recall first some basic definitions.

Definition 8.3. A symmetric function k (resp., φ) defined on a set $\mathcal{X} \times \mathcal{X}$ is said to be positive (resp., negative) definite if for any $n \geq 0$, family $x_1, \ldots, x_n \in \mathcal{Z}$, and vector $r \in \mathbb{R}^n$ the following inequality holds:

$$\sum_{i,j=1}^n r_i r_j k(x_i, x_j) \geq 0, \quad \left(\text{resp.} \quad \sum_{i,j=1}^n r_i r_j \varphi(x_i, x_j) \leq 0\right). \tag{8.12}$$

The kernel is said to be conditionally positive if positivity only holds in (8.12) for zero mean vectors r (i.e. such that $\langle r, \mathbb{1}_n \rangle = 0$).

If k is conditionally positive, one defines the following norm:

$$\|\alpha\|_k^2 \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y). \tag{8.13}$$

These norms are often referred to as "maximum mean discrepancy" (MMD) (see (Gretton *et al.*, 2007)) and have also been called "kernel norms" in shape analysis (Glaunes *et al.*, 2004). This expression (8.13) can be rephrased, introducing two independent random vectors (X, X') on $\mathcal{X}$ distributed with law α, as

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

One can show that $\|\cdot\|_k^2$ is the dual norm in the sense of (8.10) associated to the unit ball B of the RKHS associated to k. We refer to (Berlinet and Thomas-Agnan, 2003; Hofmann *et al.*, 2008; Schölkopf and Smola, 2002) for more details on RKHS functional spaces.

Remark 8.4 (Universal kernels)**.** According to Remark 8.3, the MMD norm $\|\cdot\|_k$ metrizes the weak convergence if the span of the dual ball B is dense in the space of continuous functions $\mathcal{C}(\mathcal{X})$. This means that finite sums of the form $\sum_{i=1}^n a_i k(x_i, \cdot)$ (for arbitrary choice of n and points $(x_i)_i$) are dense in $\mathcal{C}(\mathcal{X})$ for the uniform norm $\|\cdot\|_\infty$. For translation-invariant kernels over $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = k_0(x-y)$, this is equivalent to having a nonvanishing Fourier transform, $\hat{k}_0(\omega) > 0$.

In the special case where α is a discrete measure of the form (2.3), one thus has the simple expression

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

In particular, when $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{i=1}^n \mathbf{b}_i \delta_{x_i}$ are supported on the same set of points, $\|\alpha - \beta\|_k^2 = \langle \mathbf{k}(\mathbf{a}-\mathbf{b}), \mathbf{a}-\mathbf{b} \rangle$, so that $\|\cdot\|_k$ is a Euclidean norm (proper if $\mathbf{k}$ is positive definite, degenerate otherwise if $\mathbf{k}$ is semidefinite) on the simplex Σ_n. To compute the discrepancy between two discrete measures of the form (2.3), one can use

$$\|\alpha - \beta\|_k^2 = \sum_{i,i'} \mathbf{a}_i \mathbf{a}_{i'} k(x_i, x_{i'}) + \sum_{j,j'} \mathbf{b}_j \mathbf{b}_{j'} k(y_j, y_{j'}) - 2 \sum_{i,j} \mathbf{a}_i \mathbf{b}_j k(x_i, y_j). \tag{8.14}$$

Example 8.9 (Gaussian RKHS)**.** One of the most popular kernels is the Gaussian one $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$, which is a positive universal kernel on $\mathcal{X} = \mathbb{R}^d$. An attractive feature of the Gaussian kernel is that it is separable as a product of 1-D kernels, which facilitates computations when working on regular grids (see also Remark 4.17). However, an important issue that arises when using the Gaussian kernel is that one needs to select the bandwidth parameter σ. This bandwidth should match the "typical scale" between observations in the measures to be compared. If the measures have multiscale features (some regions may be

very dense, others very sparsely populated), a Gaussian kernel is thus not well adapted, and one should consider a "scale-free" kernel as we detail next. An issue with such scale-free kernels is that they are global (have slow polynomial decay), which makes them typically computationally more expensive, since no compact support approximation is possible. Figure 8.5 shows a comparison between several kernels.

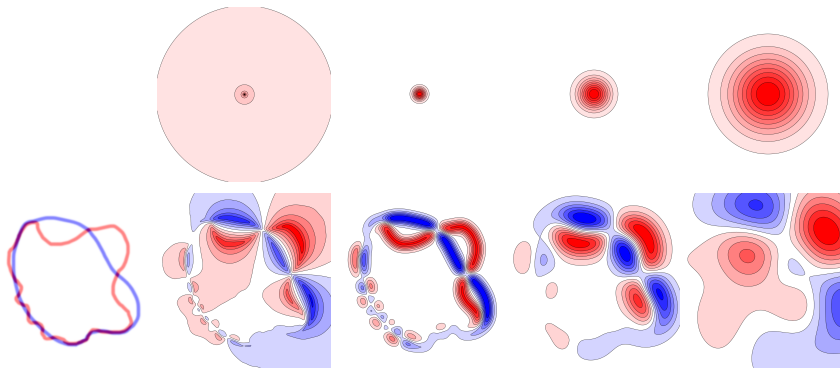


Figure 8.5: Top row: display of ψ such that $\|\alpha - \beta\|_k = \|\psi \star (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{k}_0(\omega)}$, where $k(x, x') = k_0(x - x')$. Bottom row: display of $\psi \star (\alpha - \beta)$. (G,$\sigma$) stands for Gaussian kernel of variance σ^2. The kernel for $\mathrm{ED}(\mathbb{R}^2, \|\cdot\|)$ is $\psi(x) = 1/\sqrt{\|x\|}$.

Example 8.10 ($H^{-1}(\mathbb{R}^d)$)**.** Another important dual norm is $H^{-1}(\mathbb{R}^d)$, the dual (over distributions) of the Sobolev space $H^1(\mathbb{R}^d)$ of functions having derivatives in $L^2(\mathbb{R}^d)$. It is defined using the primal RKHS norm $\|\nabla f\|^2_{L^2(\mathbb{R}^d)}$. It is not defined for singular measures (*e.g.* Diracs) unless $d = 1$ because functions in the Sobolev space $H^1(\mathbb{R}^d)$ are in general not continuous. This H^{-1} norm (defined on the space of zero mean measures with densities) can also be formulated in divergence form,

$$\|\alpha - \beta\|^2_{H^{-1}(\mathbb{R}^d)} = \min_s \left\{ \int_{\mathbb{R}^d} \|s(x)\|_2^2 \, \mathrm{d}x \;:\; \mathrm{div}(s) = \alpha - \beta \right\}, \quad (8.15)$$

which should be contrasted with (6.4), where an L^1 norm of the vector field s was used in place of the L^2 norm used here. The "weighted"

version of this Sobolev dual norm,

$$\|\rho\|^2_{H^{-1}(\alpha)} = \min_{\mathrm{div}(s)=\rho} \int_{\mathbb{R}^d} \|s(x)\|_2^2 \,\mathrm{d}\alpha(x),$$

can be interpreted as the natural "linearization" of the Wasserstein $\mathcal{W}_2$ norm, in the sense that the Benamou–Brenier dynamic formulation can be interpreted infinitesimally as

$$\mathcal{W}_2(\alpha, \alpha + \varepsilon\rho) = \varepsilon \|\rho\|_{H^{-1}(\alpha)} + o(\varepsilon). \tag{8.16}$$

The functionals $\mathcal{W}_2(\alpha, \beta)$ and $\|\alpha - \beta\|_{H^{-1}(\alpha)}$ can be shown to be equivalent (Peyre, 2011). The issue is that $\|\alpha - \beta\|_{H^{-1}(\alpha)}$ is not a norm (because of the weighting by α), and one cannot in general replace it by $\|\alpha - \beta\|_{H^{-1}(\mathbb{R}^d)}$ unless (α, β) have densities. In this case, if α and β have densities on the same support bounded from below by $a > 0$ and from above by $b < +\infty$, then

$$b^{-1/2} \|\alpha - \beta\|_{H^{-1}(\mathbb{R}^d)} \leq W_2(\alpha, \beta) \leq a^{-1/2} \|\alpha - \beta\|_{H^{-1}(\mathbb{R}^d)}; \tag{8.17}$$

see (Santambrogio, 2015, Theo. 5.34), and see (Peyre, 2011) for sharp constants.

Example 8.11 (Negative Sobolev spaces)**.** One can generalize this construction by considering the Sobolev space $H^{-r}(\mathbb{R}^d)$ of arbitrary negative index, which is the dual of the functional Sobolev space $H^r(\mathbb{R}^d)$ of functions having r derivatives (in the sense of distributions) in $L^2(\mathbb{R}^d)$. In order to metrize the weak convergence, one needs functions in $H^r(\mathbb{R}^d)$ to be continuous, which is the case when $r > d/2$. As the dimension d increases, one thus needs to consider higher regularity. For arbitrary α (not necessarily integers), these spaces are defined using the Fourier transform, and for a measure α with Fourier transform $\hat{\alpha}(\omega)$ (written here as a density with respect to the Lebesgue measure $\mathrm{d}\omega$)

$$\|\alpha\|^2_{H^{-r}(\mathbb{R}^d)} \overset{\text{def.}}{=} \int_{\mathbb{R}^d} \|\omega\|^{-2r} |\hat{\alpha}(\omega)|^2 \mathrm{d}\omega.$$

This corresponds to a dual RKHS norm with a convolutive kernel $k(x, y) = k_0(x - y)$ with $\hat{k}_0(\omega) = \pm \|\omega\|^{-2r}$. Taking the inverse Fourier transform, one sees that (up to constant) one has

$$\forall\, x \in \mathbb{R}^d, \quad k_0(x) = \begin{cases} \frac{1}{\|x\|^{d-2r}} & \text{if} \quad r < d/2, \\ -\|x\|^{2r-d} & \text{if} \quad r > d/2. \end{cases} \tag{8.18}$$

Example 8.12 (Energy distance). The energy distance (or Cramer distance when $d = 1$) (Székely and Rizzo, 2004) associated to a distance d is defined as

$$\|\alpha - \beta\|_{\mathrm{ED}(\mathcal{X},d^p)} \overset{\text{def.}}{=} \|\alpha - \beta\|_{k_{\mathrm{ED}}} \quad \text{where} \quad k_{\mathrm{ED}}(x,y) = -d(x,y)^p \tag{8.19}$$

for $0 < p < 2$. It is a valid MMD norm over measures if d is negative definite (see Definition 8.3), a typical example being the Euclidean distance $d(x,y) = \|x - y\|$. For $\mathcal{X} = \mathbb{R}^d$, $d(x,y) = \|\cdot\|$, using (8.18), one sees that the energy distance is a Sobolev norm

$$\|\cdot\|_{\mathrm{ED}(\mathbb{R}^d,\|\cdot\|^p)} = \|\cdot\|_{H^{-\frac{d+p}{2}}(\mathbb{R}^d)}.$$

A chief advantage of the energy distance over more usual kernels such as the Gaussian (Example 8.9) is that it is scale-free and does not depend on a bandwidth parameter σ. More precisely, one has the following scaling behavior on $\mathcal{X} = \mathbb{R}^d$, when denoting $f_s(x) = sx$ the dilation by a factor $s > 0$,

$$\|f_{s\sharp}(\alpha - \beta)\|_{\mathrm{ED}(\mathbb{R}^d,\|\cdot\|^p)} = s^{\frac{p}{2}} \|\alpha - \beta\|_{\mathrm{ED}(\mathbb{R}^d,\|\cdot\|^p)},$$

while the Wasserstein distance exhibits a perfect linear scaling,

$$\mathcal{W}_p(f_{s\sharp}\alpha, f_{s\sharp}\beta)) = s\,\mathcal{W}_p(\alpha,\beta)).$$

Note, however, that for the energy distance, the parameter p must satisfy $0 < p < 2$, and that for $p = 2$, it degenerates to the distance between the means

$$\|\alpha - \beta\|_{\mathrm{ED}(\mathbb{R}^d,\|\cdot\|^2)} = \left\| \int_{\mathbb{R}^d} x(\mathrm{d}\alpha(x) - \mathrm{d}\beta(x)) \right\|,$$

so it is not a norm anymore. This shows that it is not possible to get the same linear scaling under $f_{s\sharp}$ with the energy distance as for the Wasserstein distance.

8.3 Wasserstein Spaces Are Not Hilbertian

Some of the special cases of the Wasserstein geometry outlined earlier in §2.6 have highlighted the fact that the optimal transport distance

can sometimes be computed in closed form. They also illustrate that in such cases the optimal transport distance is a *Hilbertian* metric between probability measures, in the sense that there exists a map ϕ from the space of input measures onto a Hilbert space, as defined below.

Definition 8.4. A distance d defined on a set $\mathcal{Z} \times \mathcal{Z}$ is said to be Hilbertian if there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{Z} \to \mathcal{H}$ such that for any pair z, z' in $\mathcal{Z}$ we have that $d(z, z') = \|\phi(z) - \phi(z')\|_{\mathcal{H}}$.

For instance, Remark 2.30 shows that the Wasserstein metric is a Hilbert norm between univariate distributions, simply by defining ϕ to be the map that associates to a measure its generalized quantile function. Remark 2.31 shows that for univariate Gaussians, as written in (8.7) in this chapter, the Wasserstein distance between two univariate Gaussians is simply the Euclidean distance between their mean and standard deviation.

Hilbertian distances have many favorable properties when used in a data analysis context (Dattorro, 2017). First, they can be easily cast as radial basis function kernels: for any Hilbertian distance d, it is indeed known that $e^{-d^p/t}$ is a positive definite kernel for any value $0 \leq p \leq 2$ and any positive scalar t as shown in (Berg *et al.*, 1984, Cor. 3.3.3, Prop. 3.2.7). The Gaussian ($p = 2$) and Laplace ($p = 1$) kernels are simple applications of that result using the usual Euclidean distance. The entire field of kernel methods (Hofmann *et al.*, 2008) builds upon the positive definiteness of a kernel function to define convex learning algorithms operating on positive definite kernel matrices. Points living in a Hilbertian space can also be efficiently embedded in lower dimensions with low distortion factors (Johnson and Lindenstrauss, 1984), (Barvinok, 2002, §V.6.2) using simple methods such as multidimensional scaling (Borg and Groenen, 2005).

Because Hilbertian distances have such properties, one might hope that the Wasserstein distance remains Hilbertian in more general settings than those outlined above, notably when the dimension of $\mathcal{X}$ is 2 and more. This can be disproved using the following equivalence.

Proposition 8.1. A distance d is Hilbertian if and only if d^2 is negative definite.

Proof. If a distance is Hilbertian, then d^2 is trivially negative definite. Indeed, given n points in $\mathcal{Z}$, the sum $\sum r_i r_j d^2(z_i, z_j)$ can be rewritten as $\sum r_i r_j \|\phi(z_i) - \phi(z_j)\|^2_{\mathcal{H}}$ which can be expanded, taking advantage of the fact that $\sum r_i = 0$ to $-2\sum r_i r_j \langle \phi(z_i), \phi(z_j)\rangle_{\mathcal{H}}$ which is negative by definition of a Hilbert dot product. If, on the contrary, d^2 is negative definite, then the fact that d is Hilbertian proceeds from a key result by Schoenberg (1938) outlined in (Berg *et al.*, 1984, p. 82, Prop. 3.2). □

It is therefore sufficient to show that the squared Wasserstein distance is not negative definite to show that it is not Hilbertian, as stated in the following proposition.

Proposition 8.2. If $\mathcal{X} = \mathbb{R}^d$ with $d \geq 2$ and the ground cost is set to $d(x, y) = \|x - y\|_2$, then the p-Wasserstein distance is not Hilbertian for $p = 1, 2$.

Proof. It suffices to prove the result for $d = 2$ since any counterexample in that dimension suffices to obtain a counterexample in any higher dimension. We provide a nonrandom counterexample which works using measures supported on four vectors $x^1, x^2, x^3, x^4 \in \mathbb{R}^2$ defined as follows: $x^1 = [0, 0], x^2 = [1, 0], x^3 = [0, 1], x^4 = [1, 1]$. We now consider all points on the regular grid on the simplex of four dimensions, with increments of $1/4$. There are $35 = \left(\!\binom{4}{4}\!\right) = \binom{4+4-1}{4}$ such points in the simplex. Each probability vector $\mathbf{a}^i$ on that grid is such that for $j \leq 4$, we have that $\mathbf{a}^i_j$ is in the set $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ and such that $\sum_{j=1}^4 \mathbf{a}^i_j = 1$. For a given p, the 35×35 pairwise Wasserstein distance matrix $\mathbf{D}_p$ between these histograms can be computed. $\mathbf{D}_p$ is not negative definite if and only if its elementwise square $\mathbf{D}_p^2$ is such that $\mathbf{J}\mathbf{D}_p^2\mathbf{J}$ has positive eigenvalues, where $\mathbf{J}$ is the centering matrix $\mathbf{J} = \mathbb{I}_n - \frac{1}{n}\mathbb{1}_{n,n}$, which is the case as illustrated in Figure 8.6. □

8.3.1 Embeddings and Distortion

An important body of work quantifies the hardness of approximating Wasserstein distances using Hilbertian embeddings. It has been shown that embedding measures in ℓ_2 spaces incurs necessarily an important

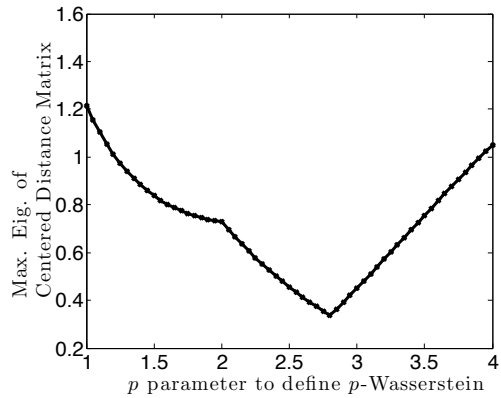


Figure 8.6: One can show that a distance is *not* Hilbertian by looking at the spectrum of the centered matrix $\mathbf{J}\mathbf{D}_p^2\mathbf{J}$ corresponding to the pairwise squared-distance matrix $\mathbf{D}_p^2$ of a set of points. The spectrum of such a matrix is necessarily non-positive if the distance is Hilbertian. Here we plot the values of the maximal eigenvalue of that matrix for points selected in the proof of Proposition 8.2. We do so for varying values of p, and display the maximal eigenvalues we obtain. These eigenvalues are all positive, which shows that for all these values of p, the p-Wasserstein distance is not Hilbertian.

distortion (Naor and Schechtman (2007) and Andoni *et al.* (2018)) as soon as $\mathcal{X} = \mathbb{R}^d$ with $d \geq 2$.

It is possible to embed quasi-isometrically p-Wasserstein spaces for $0 < p \leq 1$ in ℓ_1 (see (Indyk and Thaper, 2003; Andoni *et al.*, 2008; Do Ba *et al.*, 2011)), but the equivalence constant between the distances grows fast with the dimension d. Note also that for $p = 1$ the embedding is true only for discrete measures (*i.e.* the embedding constant depends on the minimum distance between the spikes). A closely related embedding technique consists in using the characterization of $\mathcal{W}_1$ as the dual of Lipschitz functions f (see §6.2) and approximating the Lipschitz constraint $\|\nabla f\|_1 \leq 1$ by a weighted ℓ_1 ball over the wavelets coefficients; see (Shirdhonkar and Jacobs, 2008). This weighted ℓ_1 ball of wavelet coefficients defines a so-called Besov space of negative index (Leeb and Coifman, 2016). These embedding results are also similar to the bound on the Wasserstein distance obtained using dyadic partitions; see (Weed and Bach, 2017, Prop. 1) and also (Fournier and Guillin, 2015). This

also provides a quasi-isometric embedding in ℓ_1 (this embedding being given by rescaled wavelet coefficients) and comes with the advantage that this embedding can be computed approximately in linear time when the input measures are discretized on uniform grids. We refer to (Mallat, 2008) for more details on wavelets. Note that the idea of using multiscale embeddings to compute Wasserstein-like distances has been used extensively in computer vision; see, for instance, (Ling and Okada, 2006; Grauman and Darrell, 2005; Cuturi and Fukumizu, 2007; Lazebnik *et al.*, 2006).

8.3.2 Negative/Positive Definite Variants of Optimal Transport

We show later in §10.4 that the *sliced* approximation to Wasserstein distances, essentially a sum of 1-D directional transportation distance computed on random push-forwards of measures projected on lines, is negative definite as the sum of negative definite functions (Berg *et al.*, 1984, §3.1.11). This result can be used to define a positive definite kernel (Kolouri *et al.*, 2016). Another way to recover a positive definite kernel is to cast the optimal transport problem as a soft-min problem (over all possible transportation tables) rather than a minimum, as proposed by Kosowsky and Yuille (1994) to introduce entropic regularization. That soft-min defines a term whose neg-exponential (also known as a generating function) can be proved to be positive definite (Cuturi, 2012).

8.4 Empirical Estimators for OT, MMD and φ-divergences

In an applied setting, given two input measures $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X})^2$, an important statistical problem is to approximate the (usually unknown) divergence $D(\alpha, \beta)$ using only samples $(x_i)_{i=1}^n$ from α and $(y_j)_{j=1}^m$ from β. These samples are assumed to be independently identically distributed from their respective distributions.

8.4.1 Empirical Estimators for OT and MMD

For the Wasserstein distance $\mathcal{W}_p$ (see 2.18) and MMD norms (see §8.2), a straightforward estimator is the distance itself between the empirical

measures

$$D(\alpha,\beta) \approx D(\hat{\alpha}_n,\hat{\beta}_m) \quad \text{where} \quad \begin{cases} \hat{\alpha}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_i \delta_{x_i}, \\ \hat{\beta}_m \overset{\text{def.}}{=} \frac{1}{m}\sum_j \delta_{y_j}. \end{cases}$$

Note that here both $\hat{\alpha}_n$ and $\hat{\beta}_m$ are random measures, so $D(\hat{\alpha}_n,\hat{\beta}_m)$ is a random number. For simplicity, we assume that $\mathcal{X}$ is compact (handling unbounded domain requires extra constraint on the moments of the input measures).

For such a dual distance that metrizes the weak convergence (see Definition 2.2), since there is the weak convergence $\hat{\alpha}_n \to \alpha$, one has $D(\hat{\alpha}_n,\hat{\beta}_n) \to D(\alpha,\beta)$ as $n \to +\infty$. But an important question is the speed of convergence of $D(\hat{\alpha}_n,\hat{\beta}_n)$ toward $D(\alpha,\beta)$, and this rate is often called the "sample complexity" of D.

Note that for $D(\alpha,\beta) = \|\cdot\|_{\mathrm{TV}}$, since the TV norm does not metrize the weak convergence, $\|\hat{\alpha}_n - \hat{\beta}_n\|_{\mathrm{TV}}$ is not a consistent estimator, namely it does not converge toward $\|\alpha-\beta\|_{\mathrm{TV}}$. Indeed, with probability 1, $\|\hat{\alpha}_n - \hat{\beta}_n\|_{\mathrm{TV}} = 2$ since the support of the two discrete measures does not overlap. Similar issues arise with other φ-divergences, which cannot be estimated using divergences between empirical distributions.

Rates for OT. For $\mathcal{X} = \mathbb{R}^d$ and measure supported on bounded domain, it is shown by (Dudley, 1969) that for $d > 2$, and $1 \leq p < +\infty$,

$$\mathbb{E}(|\mathcal{W}_p(\hat{\alpha}_n,\hat{\beta}_n) - \mathcal{W}_p(\alpha,\beta)|) = O(n^{-\frac{1}{d}}),$$

where the expectation $\mathbb{E}$ is taken with respect to the random samples $(x_i,y_i)_i$. This rate is tight in $\mathbb{R}^d$ if one of the two measures has a density with respect to the Lebesgue measure. This result was proved for general metric spaces (Dudley, 1969) using the notion of covering numbers and was later refined, in particular for $\mathcal{X} = \mathbb{R}^d$ in (Dereich *et al.*, 2013; Fournier and Guillin, 2015). This rate can be refined when the measures are supported on low-dimensional subdomains: Weed and Bach (2017) show that, indeed, the rate depends on the intrinsic dimensionality of the support. Weed and Bach also study the nonasymptotic behavior of that convergence, such as for measures which are discretely approximated (*e.g.* mixture of Gaussians with small variances). It is also possible to prove

concentration of $\mathcal{W}_p(\hat{\alpha}_n, \hat{\beta}_n)$ around its mean $\mathcal{W}_p(\alpha, \beta)$; see (Bolley *et al.*, 2007; Boissard, 2011; Weed and Bach, 2017).

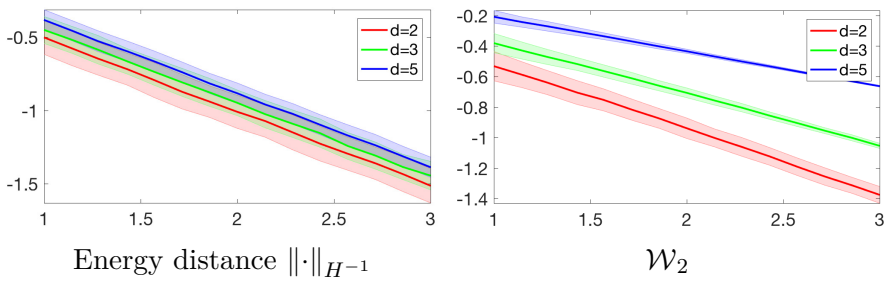
Rates for MMD. For weak norms $\|\cdot\|_k^2$ which are dual of RKHS norms (also called MMD), as defined in (8.13), and contrary to Wasserstein distances, the sample complexity does not depend on the ambient dimension

$$\mathbb{E}(|\|\hat{\alpha}_n - \hat{\beta}_n\|_k - \|\alpha - \beta\|_k|) = O(n^{-\frac{1}{2}});$$

see (Sriperumbudur *et al.*, 2012). Figure 8.7 shows a numerical comparison of the sample complexity rates for Wasserstein and MMD distances. Note, however, that $\|\hat{\alpha}_n - \hat{\beta}_n\|_k^2$ is a slightly biased estimate of $\|\alpha - \beta\|_k^2$. In order to define an unbiased estimator, and thus to be able to use, for instance, SGD when minimizing such losses, one should rather use the unbiased estimator

$$\mathrm{MMD}_k(\hat{\alpha}_n, \hat{\beta}_n)^2 \overset{\text{def.}}{=} \frac{1}{n(n-1)} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{n(n-1)} \sum_{j,j'} k(y_j, y_{j'}) - 2\frac{1}{n^2} \sum_{i,j} k(x_i, y_j),$$

which should be compared to (8.14). It satisfies $\mathbb{E}(\mathrm{MMD}_k(\hat{\alpha}_n, \hat{\beta}_n)^2) = \|\alpha - \beta\|_k^2$; see (Gretton *et al.*, 2012).



Energy distance $\|\cdot\|_{H^{-1}}$ $\mathcal{W}_2$

Figure 8.7: Decay of $\log_{10}(D(\hat{\alpha}_n, \hat{\alpha}'_n))$ as a function of $\log_{10}(n)$ for D being the energy distance $D = \|\cdot\|_{H^{-1}}$ (*i.e.* the H^{-1} norm) as defined in Example 8.12 (left) and the Wasserstein distance $D = \mathcal{W}_2$ (right). Here $(\hat{\alpha}_n, \hat{\alpha}'_n)$ are two independent empirical distributions of α, the uniform distribution on the unit cube $[0,1]^d$, tested for several value of $d \in \{2, 3, 5\}$. The shaded bar displays the confidence interval at $\pm$ the standard deviation of $\log(D(\hat{\alpha}_n, \alpha))$.

8.4.2 Empirical Estimators for φ-divergences

It is not possible to approximate $\mathcal{D}_\varphi(\alpha|\beta)$, as defined in (8.2), from discrete samples using $\mathcal{D}_\varphi(\hat{\alpha}_n|\hat{\beta}_n)$. Indeed, this quantity is either $+\infty$ (for instance, for the KL divergence) or is not converging to $\mathcal{D}_\varphi(\alpha|\beta)$ as $n \to +\infty$ (for instance, for the TV norm). Instead, it is required to use a density estimator to somehow smooth the discrete empirical measures and replace them by densities; see (Silverman, 1986). In a Euclidean space $\mathcal{X} = \mathbb{R}^d$, introducing $h_\sigma = h(\cdot/\sigma)$ with a smooth windowing function and a bandwidth $\sigma > 0$, a density estimator for α is defined using a convolution against this kernel,

$$\hat{\alpha}_n \star h_\sigma = \frac{1}{n} \sum_i h_\sigma(\cdot - x_i). \tag{8.20}$$

One can then approximate the φ divergence using

$$\mathcal{D}^\sigma_\varphi(\hat{\alpha}_n|\hat{\beta}_n) \overset{\text{def.}}{=} \frac{1}{n} \sum_{j=1}^n \varphi\left(\frac{\sum_i h_\sigma(y_j - x_i)}{\sum_{j'} h_\sigma(y_j - y_{j'}),} \right)$$

where σ should be adapted to the number n of samples and to the dimension d. It is also possible to devise nonparametric estimators, which avoids the need to choose a fixed bandwidth σ and instead selects a number k of nearest neighbor. These methods typically make use of the distance between nearest neighbors (Loftsgaarden and Quesenberry, 1965), which is similar to locally adapting the bandwidth σ to the local sampling density. Denoting $\Delta_k(x)$ the distance between $x \in \mathbb{R}^d$ and its kth nearest neighbors among the $(x_i)_{i=1}^n$, a density estimator is defined as

$$\rho^k_{\hat{\alpha}_n}(x) \overset{\text{def.}}{=} \frac{k/n}{|B_d| \Delta_k(x)^r}, \tag{8.21}$$

where $|B_d|$ is the volume of the unit ball in $\mathbb{R}^d$. Instead of somehow "counting" the number of sample falling in an area of width σ in (8.20), this formula (8.21) estimates the radius required to encapsulate k samples. Figure 8.8 compares the estimators (8.20) and (8.21). A typical example of application is detailed in (4.1) for the entropy functional, which is the KL divergence with respect to the Lebesgue measure. We refer to (Moon and Hero, 2014) for more details.

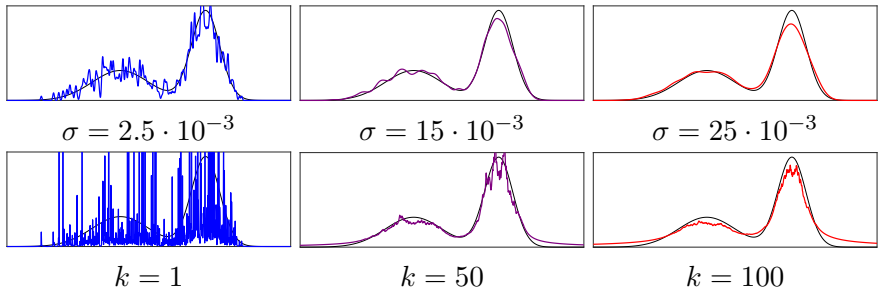


Figure 8.8: Comparison of kernel density estimation $\hat{\alpha}_n \star h_\sigma$ (top, using a Gaussian kernel h) and k-nearest neighbors estimation $\rho^k_{\hat{\alpha}_n}$ (bottom) for $n = 200$ samples from a mixture of two Gaussians.

8.5 Entropic Regularization: Between OT and MMD

Following Proposition 4.7, we recall that the Sinkhorn divergence is defined as

$$\mathfrak{P}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \langle \mathbf{P}^\star, \mathbf{C} \rangle = \langle e^{\frac{\mathbf{f}^\star}{\varepsilon}}, (\mathbf{K} \odot \mathbf{C}) e^{\frac{\mathbf{g}^\star}{\varepsilon}} \rangle,$$

where $\mathbf{P}^\star$ is the solution of (4.2) while $(\mathbf{f}^\star, \mathbf{g}^\star)$ are solutions of (4.30). Assuming $\mathbf{C}_{i,j} = d(x_i, x_j)^p$ for some distance d on $\mathcal{X}$, for two discrete probability distributions of the form (2.3), this defines a regularized Wasserstein cost

$$\mathcal{W}_{p,\varepsilon}(\alpha, \beta)^p \overset{\text{def.}}{=} \mathfrak{P}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b}).$$

This definition is generalized to any input distribution (not necessarily discrete) as

$$\mathcal{W}_{p,\varepsilon}(\alpha, \beta)^p \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)^p \mathrm{d}\pi^\star(x, y),$$

where $\pi^\star$ is the solution of (4.9).

In order to cancel the bias introduced by the regularization (in particular, $\mathcal{W}_{p,\varepsilon}(\alpha, \alpha) \neq 0$), we introduce a corrected regularized divergence

$$\tilde{\mathcal{W}}_{p,\varepsilon}(\alpha, \beta)^p \overset{\text{def.}}{=} 2\,\mathcal{W}_{p,\varepsilon}(\alpha, \beta)^p - \mathcal{W}_{p,\varepsilon}(\alpha, \alpha)^p - \mathcal{W}_{p,\varepsilon}(\beta, \beta)^p.$$

It is proved in (Feydy *et al.*, 2019) that if $e^{-c/\varepsilon}$ is a positive kernel, then a related corrected divergence (obtained by using $\mathrm{L}^\varepsilon_{\mathbf{C}}$ in place of $\mathfrak{P}^\varepsilon_{\mathbf{C}}$) is positive. Note that it is possible to define other renormalization

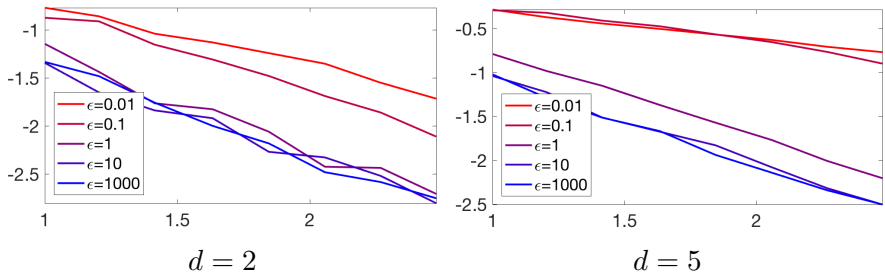


Figure 8.9: Decay of $\mathbb{E}(\log_{10}(\tilde{\mathcal{W}}_{p,\varepsilon}(\hat{\alpha}_n, \hat{\alpha}'_n)))$, for $p = 3/2$ for various ε, as a function of $\log_{10}(n)$ where α is the same as in Figure 8.7.

schemes using regularized optimal transport, as proposed, for instance, by Amari *et al.* (2018).

The following proposition, whose proof can be found in (Ramdas *et al.*, 2017), shows that this regularized divergence interpolates between the Wasserstein distance and the energy distance defined in Example 8.12.

Proposition 8.3. One has

$$\tilde{\mathcal{W}}_{p,\varepsilon}(\alpha,\beta) \overset{\varepsilon\to 0}{\longrightarrow} 2\,\mathcal{W}_p(\alpha,\beta) \quad \text{and} \quad \tilde{\mathcal{W}}_{p,\varepsilon}(\alpha,\beta)^p \overset{\varepsilon\to+\infty}{\longrightarrow} \|\alpha-\beta\|^2_{\mathrm{ED}(\mathcal{X},d)},$$

where $\|\cdot\|_{\mathrm{ED}(\mathcal{X},d)}$ is defined in (8.19).

Figure 8.9 shows numerically the impact of ε on the sample complexity rates. It is proved in Genevay *et al.*, 2019, in the case of $c(x,y) = \|x-y\|^2$ on $\mathcal{X} = \mathbb{R}^d$, that these rates interpolate between the ones of OT and MMD.

9

Variational Wasserstein Problems

In data analysis, common divergences between probability measures (*e.g.* Euclidean, total variation, Hellinger, Kullback–Leibler) are often used to measure an error or a loss in parameter estimation problems. Up to this chapter, we have made the case that the optimal transport geometry has a unique ability, not shared with other information divergences, to leverage physical ideas (mass displacement) and geometry (a cost between observations or bins) to compare measures. These two facts combined make it thus very tempting to use the Wasserstein distance as a loss function. This idea was recently explored for various applied problems. However, the main technical challenge associated with that idea lies in approximating and differentiating efficiently the Wasserstein distance.

In image processing, the Wasserstein distance can be used as a loss to synthesize textures (Tartavel *et al.*, 2016), to account for the discrepancy between statistics of synthesized and input examples. It is also used for image segmentation to account for statistical homogeneity of image regions (Swoboda and Schnörr, 2013; Rabin and Papadakis, 2015; Peyré *et al.*, 2012; Ni *et al.*, 2009; Schmitzer and Schnörr, 2013b; Li *et al.*, 2018b). The Wasserstein distance is also a very natural fidelity term

for inverse problems when the measurements are probability measures, for instance, image restoration (Lellmann *et al.*, 2014), tomographic inversion (Abraham *et al.*, 2017), density regularization (Burger *et al.*, 2012), particle image velocimetry (Saumier *et al.*, 2015), sparse recovery and compressed sensing (Indyk and Price, 2011), and seismic inversion (Métivier *et al.*, 2016). Distances between measures (mostly kernel-based as shown in §8.2.2) are routinely used for shape matching (represented as measures over a lifted space, often called currents) in computational anatomy (Vaillant and Glaunès, 2005), but OT distances offer an interesting alternative (Feydy *et al.*, 2017). To reduce the dimensionality of a dataset of histograms, Lee and Seung have shown that the nonnegative matrix factorization problem can be cast using the Kullback–Leibler divergence to quantify a reconstruction loss (Lee and Seung, 1999). When prior information is available on the geometry of the bins of those histograms, the Wasserstein distance can be used instead, with markedly different results (Sandler and Lindenbaum, 2011; Zen *et al.*, 2014; Rolet *et al.*, 2016).

All of these problems have in common that they require access to the gradients of Wasserstein distances, or approximations thereof. We start this section by presenting methods to approximate such gradients, then follow with three important applications that can be cast as variational Wasserstein problems.

9.1 Differentiating the 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 a Euclidean space. Such a comparison is done through a "loss" or a "fidelity" term, which is the Wasserstein distance in this section. In the simplest scenario, the computation of a suitable parameter θ is obtained by minimizing directly

$$\min_{\theta \in \Theta} \mathcal{E}(\theta) \stackrel{\text{def.}}{=} \mathcal{L}_c(\alpha_\theta, \beta). \tag{9.1}$$

Of course, one can consider more complicated problems: for instance, the barycenter problem described in §9.2 consists in a sum of such terms.

However, most of these more advanced problems can be usually solved by adapting tools defined for the basic case above, either using the chain rule to compute explicitly derivatives or using automatic differentiation as advocated in §9.1.3.

Convexity. The Wasserstein distance between two histograms or two densities is convex with respect to its two inputs, as shown by (2.20) and (2.24), 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 (9.1) remains convex (the first case corresponds to the barycenter problem, the second to one iteration of the dictionary learning problem with a Wasserstein loss (Rolet *et al.*, 2016)). However, for more general parameterizations $\theta \mapsto \alpha_\theta$, Problem (9.1) is in general not convex.

Simple cases. For those simple cases where the Wasserstein distance has a closed form, such as univariate (see §2.30) or elliptically contoured (see §2.31) distributions, simple workarounds exist. They consist mostly in casting the Wasserstein distance as a simpler distance between suitable representations of these distributions (Euclidean on quantile functions for univariate measures, Bures metric for covariance matrices for elliptically contoured distributions of the same family) and solving Problem (9.1) directly on such representations.

In most cases, however, one has to resort to a careful discretization of α_θ to compute a local minimizer for Problem (9.1). Two approaches can be envisioned: Eulerian or Lagrangian. Figure 9.1 illustrates the difference between these two fundamental discretization schemes. At the risk of oversimplifying this argument, one may say that a Eulerian discretization is the most suitable when measures are supported on a low-dimensional space (as when dealing with shapes or color spaces), or for intrinsically discrete problems (such as those arising from string or text analysis). When applied to fitting problems where observations can take continuous values in high-dimensional spaces, a Lagrangian perspective is usually the only suitable choice.

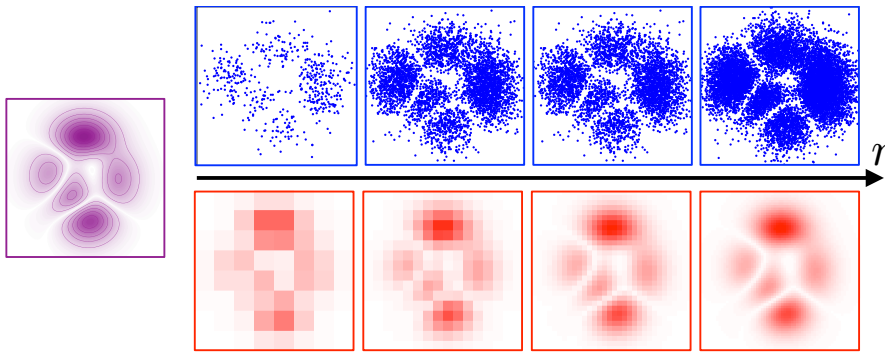


Figure 9.1: Increasing fine discretization of a continuous distribution having a density (violet, left) using a Lagrangian representation $\frac{1}{n}\sum_i \delta_{x_i}$ (blue, top) and an Eulerian representation $\sum_i \mathbf{a}_i \delta_{x_i}$ with x_i representing cells on a grid of increasing size (red, bottom). The Eulerian perspective starts from a pixelated image down to one with such fine resolution that it almost matches the original density. Weights $\mathbf{a}_i$ are directly proportional to each pixel-cell's intensity.

9.1.1 Eulerian Discretization

A first way to discretize the problem is to suppose that both distributions $\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$ and $\alpha_\theta = \sum_{i=1}^n \mathbf{a}(\theta)_i \delta_{x_i}$ are discrete distributions defined on fixed locations $(x_i)_i$ and $(y_j)_j$. Such locations might stand for cells dividing the entire space of observations in a grid, or a finite subset of points of interest in a continuous space (such as a family of vector embeddings for all words in a given dictionary (Kusner *et al.*, 2015; Rolet *et al.*, 2016)). The parameterized measure α_θ is in that case entirely represented through the weight vector $\mathbf{a} : \theta \mapsto \mathbf{a}(\theta) \in \Sigma_n$, which, in practice, might be very sparse if the grid is large. This setting corresponds to the so-called class of Eulerian discretization methods. In its original form, the objective of Problem (9.1) is not differentiable. In order to obtain a smooth minimization problem, we use the entropic regularized OT and approximate (9.1) using

$$\min_{\theta \in \Theta} \mathcal{E}_E(\theta) \stackrel{\text{def.}}{=} \mathrm{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}(\theta), \mathbf{b}) \quad \text{where} \quad \mathbf{C}_{i,j} \stackrel{\text{def.}}{=} c(x_i, y_j).$$

We recall that Proposition 4.6 shows that the entropic loss function is differentiable and convex with respect to the input histograms, with gradient.

Proposition 9.1 (Derivative with respect to histograms). For $\varepsilon > 0$, $(\mathbf{a}, \mathbf{b}) \mapsto \mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ is convex and differentiable. Its gradient reads

$$\nabla \mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = (\mathbf{f}, \mathbf{g}), \tag{9.2}$$

where $(\mathbf{f}, \mathbf{g})$ is the unique solution to (4.30), centered such that $\sum_i \mathbf{f}_i = \sum_j \mathbf{g}_j = 0$. For $\varepsilon = 0$, this formula defines the elements of the sub-differential of $\mathrm{L}^{\varepsilon}_{\mathbf{C}}$, and the function is differentiable if they are unique.

The zero mean condition on $(\mathbf{f}, \mathbf{g})$ is important when using gradient descent to guarantee conservation of mass. Using the chain rule, one thus obtains that $\mathcal{E}_E$ is smooth and that its gradient is

$$\nabla \mathcal{E}_E(\theta) = [\partial \mathbf{a}(\theta)]^{\top}(\mathbf{f}), \tag{9.3}$$

where $\partial \mathbf{a}(\theta) \in \mathbb{R}^{n \times \dim(\Theta)}$ is the Jacobian (differential) of the map $\mathbf{a}(\theta)$, and where $\mathbf{f} \in \mathbb{R}^n$ is the dual potential vector associated to the dual entropic OT (4.30) between $\mathbf{a}(\theta)$ and $\mathbf{b}$ for the cost matrix $\mathbf{C}$ (which is fixed in a Eulerian setting, and in particular independent of θ). This result enables a simple gradient descent approach to minimize locally $\mathcal{E}_E$.

9.1.2 Lagrangian Discretization

A different approach consists in using instead fixed (typically uniform) weights and approximating an input measure α as an empirical measure $\alpha_\theta = \frac{1}{n}\sum_i \delta_{x(\theta)_i}$ for a point-cloud parameterization map $x : \theta \mapsto x(\theta) = (x(\theta)_i)_{i=1}^n \in \mathcal{X}^n$, where we assume here that $\mathcal{X}$ is Euclidean. Problem (9.1) is thus approximated as

$$\min_\theta \mathcal{E}_L(\theta) \overset{\text{def.}}{=} \mathrm{L}^{\varepsilon}_{\mathbf{C}(x(\theta))}(\mathbb{1}_n/n, \mathbf{b}) \quad \text{where} \quad \mathbf{C}(x)_{i,j} \overset{\text{def.}}{=} c(x(\theta)_i, y_j). \tag{9.4}$$

Note that here the cost matrix $\mathbf{C}(x(\theta))$ now depends on θ since the support of α_θ changes with θ. The following proposition shows that the entropic OT loss is a smooth function of the cost matrix and gives the expression of its gradient.

Proposition 9.2 (Derivative with respect to the cost). For fixed input histograms $(\mathbf{a}, \mathbf{b})$, for $\varepsilon > 0$, the mapping $\mathbf{C} \mapsto \mathcal{R}(\mathbf{C}) \overset{\text{def.}}{=} \mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ is concave and smooth, and

$$\nabla \mathcal{R}(\mathbf{C}) = \mathbf{P}, \tag{9.5}$$

where $\mathbf{P}$ is the unique optimal solution of (4.2). For $\varepsilon = 0$, this formula defines the set of upper gradients.

Assuming $(\mathcal{X}, \mathcal{Y})$ are convex subsets of $\mathbb{R}^d$, for discrete measures (α, β) of the form (2.3), one obtains using the chain rule that $x = (x_i)_{i=1}^n \in \mathcal{X}^n \mapsto \mathcal{F}(x) \overset{\text{def.}}{=} \mathrm{L}_{\mathbf{C}(x)}(\mathbb{1}_n/n, \mathbf{b})$ is smooth and that

$$\nabla \mathcal{F}(x) = \left(\sum_{j=1}^m \mathbf{P}_{i,j} \nabla_1 c(x_i, y_j) \right)_{i=1}^n \in \mathcal{X}^n, \tag{9.6}$$

where $\nabla_1 c$ is the gradient with respect to the first variable. For instance, for $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, for $c(s,t) = \|s-t\|^2$ on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, one has

$$\nabla \mathcal{F}(x) = 2 \left(\mathbf{a}_i x_i - \sum_{j=1}^m \mathbf{P}_{i,j} y_j \right)_{i=1}^n, \tag{9.7}$$

where $\mathbf{a}_i = 1/n$ here. Note that, up to a constant, this gradient is $\mathrm{Id} - T$, where T is the barycentric projection defined in (4.19). Using the chain rule, one thus obtains that the Lagrangian discretized problem (9.4) is smooth and its gradient is

$$\nabla \mathcal{E}_L(\theta) = [\partial x(\theta)]^\top (\nabla \mathcal{F}(x(\theta))), \tag{9.8}$$

where $\partial x(\theta) \in \mathbb{R}^{\dim(\Theta) \times (nd)}$ is the Jacobian of the map $x(\theta)$ and where $\nabla \mathcal{F}$ is implemented as in (9.6) or (9.7) using for $\mathbf{P}$ the optimal coupling matrix between α_θ and β. One can thus implement a gradient descent to compute a local minimizer of $\mathcal{E}_L$, as used, for instance, in (Cuturi and Doucet, 2014).

9.1.3 Automatic Differentiation

The difficulty when applying formulas (9.3) and (9.8) is that one needs to compute the exact optimal solutions $\mathbf{f}$ or $\mathbf{P}$ for these formulas to be valid, which can only be achieved with acceptable precision using a very large number of Sinkhorn iterates. In challenging situations in which the size and the quantity of histograms to be compared are large, the computational budget to compute a single Wasserstein distance is usually limited, therefore allowing only for a few Sinkhorn iterations.

In that case, and rather than approximating the gradient (4.30) using the value obtained at a given iterate, it is usually better to differentiate directly the output of Sinkhorn's algorithm, using reverse mode automatic differentiation. This corresponds to using the "algorithmic" Sinkhorn divergences as introduced in (4.48), rather than the quantity $\mathrm{L}_{\mathbf{C}}^{\varepsilon}$ in (4.2) which incorporates the entropy of the regularized optimal transport, and differentiating it directly as a composition of simple maps using the inputs, either the histogram in the Eulerian case or the cost matrix in the Lagrangian cases. Using definitions introduced in §4.5, this is equivalent to differentiating

$$\mathfrak{D}_{\mathbf{C}}^{(L)}(\mathbf{a}(\theta), \mathbf{b}) \quad \text{or} \quad \mathfrak{D}_{\mathbf{C}(x(\theta))}^{(L)}(\mathbf{a}, \mathbf{b})$$

with respect to θ, in, respectively, the Eulerian and the Lagrangian cases for L large enough.

The cost for computing the gradient of functionals involving Sinkhorn divergences is the same as that of computation of the functional itself; see, for instance, (Bonneel *et al.*, 2016; Genevay *et al.*, 2018) for some applications of this approach. We also refer to (Adams and Zemel, 2011) for an early work on differentiating Sinkhorn iterations with respect to the cost matrix (as done in the Lagrangian framework), with applications to learning rankings. Further details on automatic differentiation can be found in (Griewank and Walther, 2008; Rall, 1981; Neidinger, 2010), in particular on the "reverse mode," which is the fastest way to compute gradients. In terms of implementation, all recent deep-learning Python frameworks feature state-of-the-art reverse-mode differentiation and support for GPU/TPU computations (Al-Rfou *et al.*, 2016; Abadi *et al.*, 2016; Pytorch, 2017), they should be adopted for any large-scale application of Sinkhorn losses. We strongly encourage the use of such automatic differentiation techniques, since they have the same complexity as computing (9.3) and (9.8), these formulas being mostly useful to obtain a theoretical understanding of what automatic differentation is computing. The only downside is that reverse mode automatic differentation is memory intensive (the memory grows proportionally with the number of iterations). There exist, however, subsampling strategies that mitigate this problem (Griewank, 1992).

9.2 Wasserstein Barycenters, Clustering and Dictionary Learning

A basic problem in unsupervised learning is to compute the "mean" or "barycenter" of several data points. A classical way to define such a weighted mean of points $(x_s)_{s=1}^S \in \mathcal{X}^S$ living in a metric space $(\mathcal{X}, d)$ (where d is a distance or more generally a divergence) is by solving a variational problem

$$\min_{x \in \mathcal{X}} \sum_{s=1}^{S} \lambda_s d(x, x_s)^p \tag{9.9}$$

for a given family of weights $(\lambda_s)_s \in \Sigma_S$, where p is often set to $p = 2$. When $\mathcal{X} = \mathbb{R}^d$ and $d(x, y) = \|x - y\|_2$, this leads to the usual definition of the linear average $x = \sum_s \lambda_s x_s$ for $p = 2$ and the more evolved median point when $p = 1$. One can retrieve various notions of means (*e.g.* harmonic or geometric means over $\mathcal{X} = \mathbb{R}_+$) using this formalism. This process is often referred to as the "Fréchet" or "Karcher" mean (see Karcher (2014) for a historical account). For a generic distance d, Problem (9.9) is usually a difficult nonconvex optimization problem. Fortunately, in the case of optimal transport distances, the problem can be formulated as a convex program for which existence can be proved and efficient numerical schemes exist.

Fréchet means over the Wasserstein space. 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

$$\min_{\mathbf{a} \in \Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}_{\mathbf{C}_s}(\mathbf{a}, \mathbf{b}_s), \tag{9.10}$$

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

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

The barycenter problem (9.10) was introduced in a more general form involving arbitrary measures in Agueh and Carlier (2011) following

earlier ideas of Carlier and Ekeland (2010). That presentation is deferred to Remark 9.1. The barycenter problem for histograms (9.10) 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, which by construction must be constrained to share the same row marginal,

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

Although this problem is an LP, its scale forbids the use of generic solvers for medium-scale problems. One can therefore resort to using first order methods such as subgradient descent on the dual (Carlier *et al.*, 2015).

Remark 9.1 (Barycenter of arbitrary measures)**.** Given a set of input measure $(\beta_s)_s$ defined on some space $\mathcal{X}$, the barycenter problem becomes

$$\min_{\alpha\in\mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \mathcal{L}_c(\alpha, \beta_s). \tag{9.11}$$

In the case where $\mathcal{X} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^2$, Agueh and Carlier (2011) show that if one of the input measures has a density, then this barycenter is unique. Problem (9.11) 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 (9.11) is $\delta_{x^\star}$, where $x^\star$ is a Fréchet mean solving (9.9). 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 (9.11) when approximating the input distribution using discrete ones (and thus solving (9.10) in place) is

studied in Carlier *et al.* (2015). Let us also note that it is possible to recast (9.11) as a multimarginal OT problem; see Remark 10.2.

Remark 9.2 (k-means as a Wasserstein variational problem)**.** When the family of input measures $(\beta_s)_s$ is limited to but one measure β, this measure is supported on a discrete finite subset of $\mathcal{X} = \mathbb{R}^d$, and the cost is the squared Euclidean distance, then one can show that the barycenter problem

$$\min_{\alpha \in \mathcal{M}_k^1(\mathcal{X})} \mathcal{L}_c(\alpha, \beta), \tag{9.12}$$

where α is constrained to be a discrete measure with a finite support of size up to k, is equivalent to the usual k-means problem taking β. Indeed, one can easily show that the centroids output by the k-means problem correspond to the support of the solution α and that its weights correspond to the fraction of points in β assigned to each centroid. One can show that approximating $\mathcal{L}_c$ using entropic regularization results in smoothed out assignments that appear in soft-clustering variants of k-means, such as mixtures of Gaussians (Dessein *et al.*, 2017).

Remark 9.3 (Distribution of distributions and consistency)**.** It is possible to generalize (9.11) to a possibly infinite collection of measures. This problem is described by considering a probability distribution M over the space $\mathcal{M}_+^1(\mathcal{X})$ of probability distributions, *i.e.* $M \in \mathcal{M}_+^1(\mathcal{M}_+^1(\mathcal{X}))$. A barycenter is then a solution of

$$\min_{\alpha \in \mathcal{M}_+^1(\mathcal{X})} \mathbb{E}_M(\mathcal{L}_c(\alpha, \beta)) = \int_{\mathcal{M}_+^1(\mathcal{X})} \mathcal{L}_c(\alpha, \beta) \mathrm{d}M(\beta), \tag{9.13}$$

where β is a random measure distributed according to M. Drawing uniformly at random a finite number S of input measures $(\beta_s)_{s=1}^S$ according to M, one can then define $\hat{\beta}_S$ as being a solution of (9.11) for uniform weights $\lambda_s = 1/S$ (note that here $\hat{\beta}_S$ is itself a random measure). Problem (9.11) corresponds to the special case of a "discrete" measure $M = \sum_s \lambda_s \delta_{\beta_s}$. The convergence (in expectation

or with high probability) of $\mathcal{L}_c(\hat{\beta}_S, \alpha)$ to zero (where α is the unique solution to (9.13)) corresponds to the consistency of the barycenters, and is proved in (Bigot and Klein, 2012b; Le Gouic and Loubes, 2016; Bigot and Klein, 2012a). This can be interpreted as a law of large numbers over the Wasserstein space. The extension of this result to a central limit theorem is an important problem; see (Panaretos and Zemel, 2016) and (Agueh and Carlier, 2017) for recent formulations of that problem and solutions in particular cases (1-D distributions and Gaussian measures).

Remark 9.4 (Fixed-point map)**.** When dealing with the Euclidean space $\mathcal{X} = \mathbb{R}^d$ with ground cost $c(x, y) = \|x - y\|^2$, it is possible to study the barycenter problem using transportation maps. Indeed, if α has a density, according to Remark 2.24, one can define optimal transportation maps T_s between α and α_s, in particular such that $T_{s,\sharp}\alpha = \alpha_s$. The average map

$$T^{(\alpha)} \overset{\text{def.}}{=} \sum_{s=1}^{S} \lambda_s T_s$$

(the notation above makes explicit the dependence of this map on α) is itself an optimal map between α and $T^{(\alpha)}_{\sharp}\alpha$ (a positive combination of optimal maps is equal by Brenier's theorem, Remark 2.24, to the sum of gradients of convex functions, equal to the gradient of a sum of convex functions, and therefore optimal by Brenier's theorem again). As shown in (Agueh and Carlier, 2011), first order optimality conditions of the barycenter problem (9.13) actually read $T^{(\alpha^\star)} = \mathbb{I}_{\mathbb{R}^d}$ (the identity map) at the optimal measure $\alpha^\star$ (the barycenter), and it is shown in (Álvarez-Esteban *et al.*, 2016) that the barycenter $\alpha^\star$ is the unique (under regularity conditions clarified in (Zemel and Panaretos, 2018, Theo. 2)) to the fixed-point equation

$$G(\alpha) = \alpha \quad \text{where} \quad G(\alpha) \overset{\text{def.}}{=} T^{(\alpha)}_{\sharp}\alpha, \tag{9.14}$$

Under mild conditions on the input measures, Álvarez-Esteban *et al.* (2016) and Zemel and Panaretos (2018) have shown that $\alpha \mapsto G(\alpha)$

strictly decreases the objective function of (9.13) if α is not the barycenter and that the fixed-point iterations $\alpha^{(\ell+1)} \overset{\text{def.}}{=} G(\alpha^{(\ell)})$ converge to the barycenter $\alpha^\star$. This fixed point algorithm can be used in cases where the optimal transportation maps are known in closed form (*e.g.* for Gaussians). Adapting this algorithm for empirical measures of the same size results in computing optimal assignments in place of Monge maps. For more general discrete measures of arbitrary size the scheme can also be adapted (Cuturi and Doucet, 2014) using barycentric projections (4.19).

Special cases. In general, solving (9.10) or (9.11) is not straightforward, but there exist some special cases for which solutions are explicit or simple.

Remark 9.5 (Barycenter of Gaussians)**.** It is shown in (Agueh and Carlier, 2011) that the barycenter of Gaussians distributions $\alpha_s = \mathcal{N}(\mathbf{m}_s, \mathbf{\Sigma}_s)$, for the squared Euclidean cost $c(x,y) = \|x-y\|^2$, is itself a Gaussian $\mathcal{N}(\mathbf{m}^\star, \mathbf{\Sigma}^\star)$. Making use of (2.41), one sees that the barycenter mean is the mean of the inputs

$$\mathbf{m}^\star = \sum_s \lambda_s \mathbf{m}_s$$

while the covariance minimizes

$$\min_{\mathbf{\Sigma}} \sum_s \lambda_s \mathcal{B}(\mathbf{\Sigma}, \mathbf{\Sigma}_s)^2,$$

where $\mathcal{B}$ is the Bure metric (2.42). As studied in (Agueh and Carlier, 2011), the first order optimality condition of this convex problem shows that $\mathbf{\Sigma}^\star$ is the unique positive definite fixed point of the map

$$\mathbf{\Sigma}^\star = \Psi(\mathbf{\Sigma}^\star) \quad \text{where} \quad \Psi(\mathbf{\Sigma}) \overset{\text{def.}}{=} \sum_s \lambda_s (\mathbf{\Sigma}^{\frac{1}{2}} \mathbf{\Sigma}_s \mathbf{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}},$$

where $\mathbf{\Sigma}^{\frac{1}{2}}$ is the square root of positive semidefinite matrices. This result was known from (Knott and Smith, 1994; Rüschendorf and Uckelmann, 2002) and is proved in (Agueh and Carlier, 2011). While Ψ is not strictly contracting, iterating this fixed-point map,

i.e. defining $\boldsymbol{\Sigma}^{(\ell+1)} \overset{\text{def.}}{=} \Psi(\boldsymbol{\Sigma}^{(\ell)})$ converges in practice to the solution $\boldsymbol{\Sigma}^\star$. This method has been applied to texture synthesis in (Xia *et al.*, 2014). Álvarez-Esteban *et al.* (2016) have also proposed to use an alternative map

$$\bar{\Psi}(\boldsymbol{\Sigma}) \overset{\text{def.}}{=} \boldsymbol{\Sigma}^{-\frac{1}{2}} \Big(\sum_s \lambda_s (\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Sigma}_s \boldsymbol{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}} \Big)^2 \boldsymbol{\Sigma}^{-\frac{1}{2}}$$

for which the iterations $\boldsymbol{\Sigma}^{(\ell+1)} \overset{\text{def.}}{=} \bar{\Psi}(\boldsymbol{\Sigma}^{(\ell)})$ converge. This is because the fixed-point map G defined in (9.14) preserves Gaussian distributions, and in fact,

$$G(\mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})) = \mathcal{N}(\mathbf{m}^\star, \bar{\Psi}(\boldsymbol{\Sigma})).$$

Figure 9.2 shows two examples of computations of barycenters between four 2-D Gaussians.

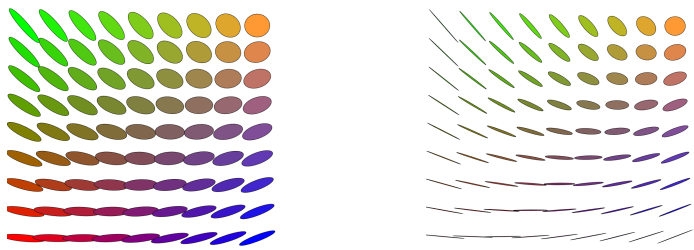


Figure 9.2: Barycenters between four Gaussian distributions in 2-D. Each Gaussian is displayed using an ellipse aligned with the principal axes of the covariance, and with elongations proportional to the corresponding eigenvalues.

Remark 9.6 (1-D cases)**.** For 1-D distributions, the $\mathcal{W}_p$ barycenter can be computed almost in closed form using the fact that the transport is the monotone rearrangement, as detailed in Remark 2.30. The simplest case is for empirical measures with n points, *i.e.* $\beta_s = \frac{1}{n} \sum_{i=1}^n \delta_{y_{s,i}}$, where the points are assumed to be sorted $y_{s,1} \leq y_{s,2} \leq \ldots$. Using (2.33) the barycenter α_λ is also an

empirical measure on n points

$$\alpha_\lambda = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{\lambda,i}} \quad \text{where} \quad x_{\lambda,i} = A_\lambda(x_{s,i})_s,$$

where A_λ is the barycentric map

$$A_\lambda(x_s)_s \overset{\text{def.}}{=} \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{s=1}^{S} \lambda_s |x - x_s|^p.$$

For instance, for $p = 2$, one has $x_{\lambda,i} = \sum_{s=1}^{S} \lambda_s x_{s,i}$. In the general case, one needs to use the cumulative functions as defined in (2.34), and using (2.36), one has

$$\forall\, r \in [0,1], \quad \mathcal{C}^{-1}_{\alpha_\lambda}(r) = A_\lambda(\mathcal{C}^{-1}_{\alpha_s}(r))_{s=1}^{S},$$

which can be used, for instance, to compute barycenters between discrete measures supported on less than n points in $O(n\log(n))$ operations, using a simple sorting procedure.

Remark 9.7 (Simple cases). Denoting by $T_{r,u} : x \mapsto rx+u$ a scaling and translation, and assuming that $\alpha_s = T_{r_s,u_s,\sharp}\alpha_0$ is obtained by scaling and translating an initial template measure, then a barycenter α_λ is also obtained using scaling and translation

$$\alpha_\lambda = T_{r^\star,u^\star,\sharp}\alpha_0 \quad \text{where} \quad \begin{cases} r^\star = (\sum_s \lambda_s/r_s)^{-1}, \\ u^\star = \sum_s \lambda_s u_s. \end{cases}$$

Remark 9.8 (Case $S = 2$). In the case where $\mathcal{X} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^2$ (this can be extended to geodesic spaces), the barycenter between $S = 2$ measures (α_0, α_1) is the McCann interpolant as already introduced in (7.6). Denoting $T_\sharp\alpha_0 = \alpha_1$ the Monge map, one has that the barycenter α_λ reads $\alpha_\lambda = (\lambda_1 \mathrm{Id} + \lambda_2 T)_\sharp\alpha_0$. Formula (7.9) explains how to perform the computation in the discrete case.

Entropic approximation of barycenters. One can use entropic smoothing and approximate the solution of (9.10) using

$$\min_{\mathbf{a}\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}^{\varepsilon}_{\mathbf{C}_s}(\mathbf{a}, \mathbf{b}_s) \tag{9.15}$$

for some $\varepsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent (Cuturi and Doucet, 2014; Gramfort *et al.*, 2015). An alternative is to use descent methods (typically quasi-Newton) on the semi-dual (Cuturi and Peyré, 2016), which is useful to integrate additional regularizations on the barycenter, to impose, for instance, some smoothness w.r.t a given norm. A simpler yet very effective approach, as remarked by Benamou *et al.* (2015) is to rewrite (9.15) as a (weighted) KL projection problem

$$\min_{(\mathbf{P}_s)_s} \left\{ \sum_s \lambda_s \varepsilon \mathbf{KL}(\mathbf{P}_s | \mathbf{K}_s) \;:\; \forall\, s, {\mathbf{P}_s}^{\mathrm{T}} \mathbb{1}_m = \mathbf{b}_s,\ \mathbf{P}_1 \mathbb{1}_1 = \cdots = \mathbf{P}_S \mathbb{1}_S, \right\} \tag{9.16}$$

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 $\mathbf{a} = \mathbf{P}_1 \mathbb{1}_1 = \cdots = \mathbf{P}_S \mathbb{1}_S$. As detailed by Benamou *et al.* (2015), one can generalize Sinkhorn to this problem, which also corresponds to iterative projections. This can also be seen as a special case of the generalized Sinkhorn detailed in §4.6. The optimal couplings $(\mathbf{P}_s)_s$ solving (9.16) are computed in scaling form as

$$\mathbf{P}_s = \mathrm{diag}(\mathbf{u}_s)\mathbf{K}\,\mathrm{diag}(\mathbf{v}_s), \tag{9.17}$$

and the scalings are sequentially updated as

$$\forall\, s \in [\![1, S]\!], \quad \mathbf{v}_s^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}_s}{\mathbf{K}_s^{\mathrm{T}} \mathbf{u}_s^{(\ell)}}, \tag{9.18}$$

$$\forall\, s \in [\![1, S]\!], \quad \mathbf{u}_s^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{a}^{(\ell+1)}}{\mathbf{K}_s \mathbf{v}_s^{(\ell+1)}}, \tag{9.19}$$

$$\text{where} \quad \mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \prod_s (\mathbf{K}_s \mathbf{v}_s^{(\ell+1)})^{\lambda_s}. \tag{9.20}$$

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

Proposition 9.1. The optimal $(\mathbf{u}_s, \mathbf{v}_s)$ appearing in (9.17) 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 (9.15)):

$$\max_{(\mathbf{f}_s,\mathbf{g}_s)_s} \left\{ \sum_s \lambda_s \left(\langle \mathbf{g}_s,\, \mathbf{b}_s \rangle - \varepsilon \langle \mathbf{K}_s e^{\mathbf{g}_s/\varepsilon},\, e^{\mathbf{f}_s/\varepsilon} \rangle \right) \;:\; \sum_s \lambda_s \mathbf{f}_s = 0 \right\}. \quad (9.21)$$

Proof. Introducing Lagrange multipliers in (9.16) 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, to obtain

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

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

$$\mathbf{KL}^*(\mathbf{U}|\mathbf{K}) = \sum_{i,j} \mathbf{K}_{i,j} (e^{\mathbf{U}_{i,j}} - 1), \quad (9.22)$$

which shows the desired formula. To show (9.22), 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 (9.21) with respect to each $\mathbf{g}_s$, while keeping all the other variables fixed, is obtained in closed form by (9.18). Minimizing (9.21)

with respect to all the $(\mathbf{f}_s)_s$ requires us to solve for $\mathbf{a}$ using (9.20) and leads to the expression (9.19).

Figures 9.3 and 9.4 show applications to 2-D and 3-D shapes interpolation. Figure 9.5 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 4.19. We refer to (Solomon *et al.*, 2015) for more details and other applications to computer graphics and imaging sciences.

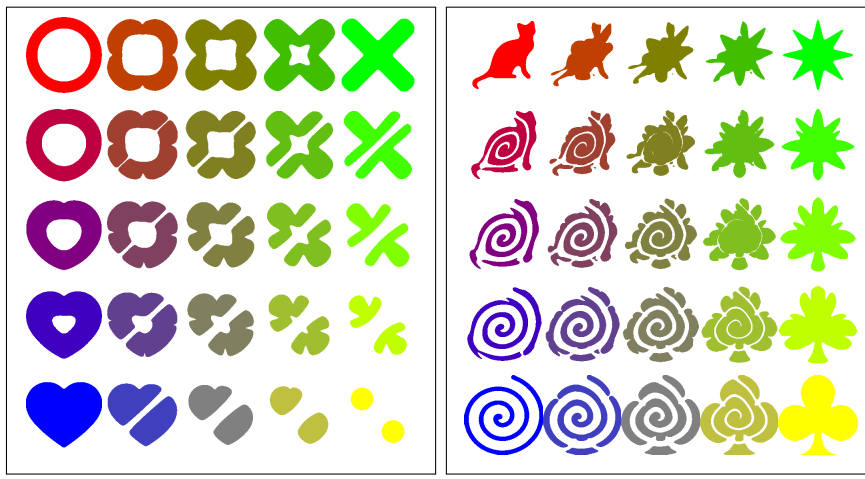


Figure 9.3: Barycenters between four input 2-D shapes using entropic regularization (9.15). To display a binary shape, the displayed images shows a thresholded density. The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square.

The efficient computation of Wasserstein barycenters remains at this time an active research topic (Staib *et al.*, 2017b; Dvurechenskii *et al.*, 2018). Beyond their methodological interest, Wasserstein barycenters have found many applications outside the field of shape analysis. They have been used for image processing (Rabin *et al.*, 2011), in particular color modification (Solomon *et al.*, 2015) (see Figure 9.6); Bayesian computations (**srivastava2015scalable**; Srivastava *et al.*, 2015) to summarize measures; and nonlinear dimensionality reduction, to express an input measure as a Wasserstein barycenter of other known measures (Bonneel *et al.*, 2016). All of these problems result in in-

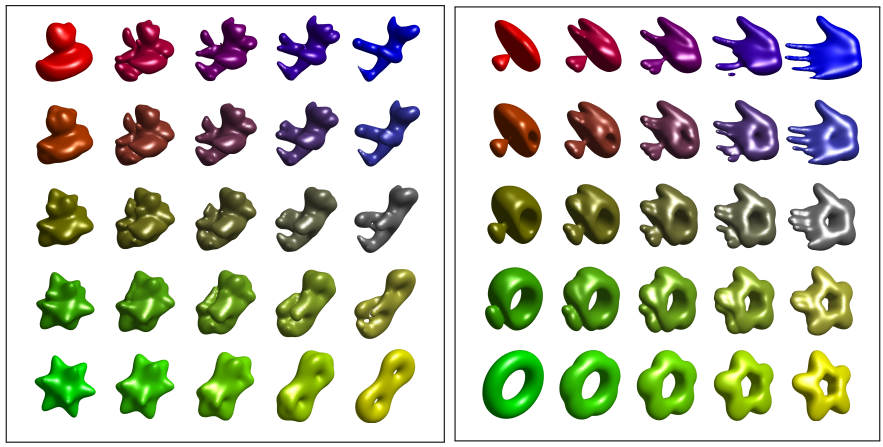


Figure 9.4: Barycenters between four input 3-D shapes using entropic regularization (9.15). The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square. Shapes are represented as measures that are uniform within the boundaries of the shape and null outside.

volved nonconvex objective functions which can be accurately optimized using automatic differentiation (see Remark 9.1.3). Problems closely related to the computation of barycenters include the computation of principal components analyses over the Wasserstein space (see, for instance, (Seguy and Cuturi, 2015; Bigot *et al.*, 2017b)) and the statistical estimation of template models (Boissard *et al.*, 2015). The ability to compute barycenters enables more advanced clustering methods such as the k-means on the space of probability measures (Barrio *et al.*, 2016; Ho *et al.*, 2017).

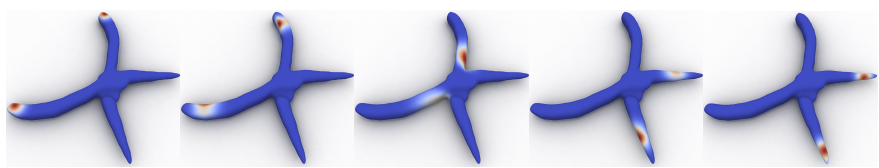


Figure 9.5: Barycenters interpolation between two input measures on surfaces, computed using the geodesic in heat fast kernel approximation (see Remark 4.19). Extracted from (Solomon *et al.*, 2015).

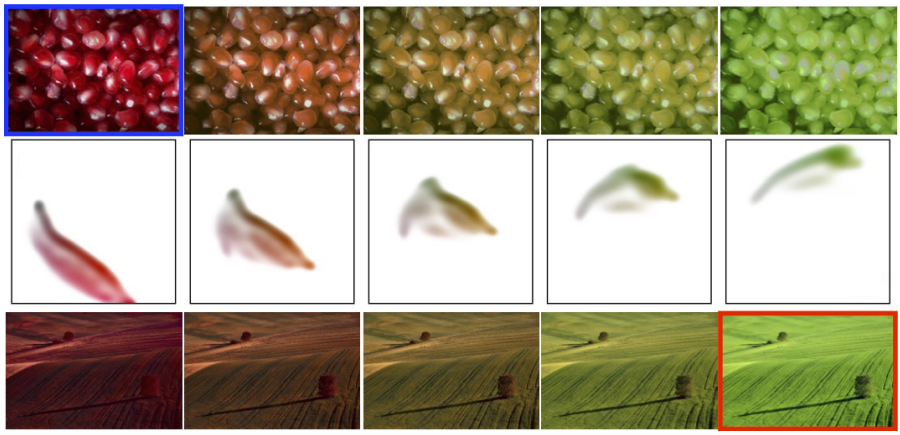


Figure 9.6: Interpolation between the two 3-D color empirical histograms of two input images (here only the 2-D chromatic projection is visualized for simplicity). The modified histogram is then applied to the input images using barycentric projection as detailed in Remark 4.11. Extracted from (Solomon *et al.*, 2015).

Remark 9.9 (Wasserstein propagation)**.** As studied in Solomon *et al.* (2014b), it is possible to generalize the barycenter problem (9.10), where one looks for distributions $(\mathbf{b}_u)_{u\in U}$ at some given set U of nodes in a graph $\mathcal{G}$ given a set of fixed input distributions $(\mathbf{b}_v)_{v\in V}$ on the complementary set V of the nodes. The unknown are determined by minimizing the overall transportation distance between all pairs of nodes $(r,s)\in\mathcal{G}$ forming edges in the graph

$$\min_{(\mathbf{b}_u\in\Sigma_{n_u})_{u\in U}} \sum_{(r,s)\in\mathcal{G}} \mathrm{L}_{\mathbf{C}_{r,s}}(\mathbf{b}_r,\mathbf{b}_s), \tag{9.23}$$

where the cost matrices $\mathbf{C}_{r,s}\in\mathbb{R}^{n_r\times n_s}$ need to be specified by the user. The barycenter problem (9.10) is a special case of this problem where the considered graph $\mathcal{G}$ is "star shaped," where U is a single vertex connected to all the other vertices V (the weight λ_s associated to $\mathbf{b}_s$ can be absorbed in the cost matrix). Introducing explicitly a coupling $\mathbf{P}_{r,s}\in\mathbf{U}(\mathbf{b}_r,\mathbf{b}_s)$ for each edge $(r,s)\in\mathcal{G}$, and using entropy regularization, one can rewrite this problem similarly as in (9.16), and one extends Sinkhorn iterations (9.18) to this

problem (this can also be derived by recasting this problem in the form of the generalized Sinkhorn algorithm detailed in §4.6). This discrete variational problem (9.23) on a graph can be generalized to define a Dirichlet energy when replacing the graph by a continuous domain (Solomon *et al.*, 2013). This in turn leads to the definition of measure-valued harmonic functions which finds application in image and surface processing. We refer also to Lavenant (2017) for a theoretical analysis and to Vogt and Lellmann (2018) for extensions to nonquadratic (total-variation) functionals and applications to imaging.

9.3 Gradient Flows

Given a smooth function $\mathbf{a} \mapsto F(\mathbf{a})$, one can use the standard gradient descent

$$\mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \mathbf{a}^{(\ell)} - \tau \nabla F(\mathbf{a}^{(\ell)}), \tag{9.24}$$

where τ is a small enough step size. This corresponds to a so-called "explicit" minimization scheme and only applies for smooth functions F. For nonsmooth functions, one can use instead an "implicit" scheme, which is also called the proximal-point algorithm (see, for instance, Bauschke and Combettes (2011))

$$\mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Prox}_{\tau F}^{\|\cdot\|}(\mathbf{a}^{(\ell)}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathbf{a}} \frac{1}{2} \left\| \mathbf{a} - \mathbf{a}^{(\ell)} \right\|^2 + \tau F(\mathbf{a}). \tag{9.25}$$

Note that this corresponds to the Euclidean proximal operator, already encountered in (7.13). The update (9.24) can be understood as iterating the explicit operator $\mathrm{Id} - \tau \nabla F$, while (9.25) makes use of the implicit operator $(\mathrm{Id} + \tau \nabla F)^{-1}$. For convex F, iterations (9.25) always converge, for any value of $\tau > 0$.

If the function F is defined on the simplex of histograms Σ_n, then it makes sense to use an optimal transport metric in place of the ℓ^2 norm $\|\cdot\|$ in (9.25), in order to solve

$$\mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathbf{a}} \mathrm{W}_p(\mathbf{a}, \mathbf{a}^{(\ell)})^p + \tau F(\mathbf{a}). \tag{9.26}$$

Remark 9.10 (Wasserstein gradient flows). Equation (9.26) can be generalized to arbitrary measures by defining the iteration

$$\alpha^{(\ell+1)} \overset{\text{def.}}{=} \operatorname*{argmin}_{\alpha} \mathcal{W}_p(\alpha, \alpha^{(\ell)})^p + \tau F(\alpha) \tag{9.27}$$

for some function F defined on $\mathcal{M}_+^1(\mathcal{X})$. This implicit time stepping is a useful tool to construct continuous flows, by formally taking the limit $\tau \to 0$ and introducing the time $t = \tau\ell$, so that $\alpha^{(\ell)}$ is intended to approximate a continuous flow $t \in \mathbb{R}_+ \mapsto \alpha_t$. For the special case $p = 2$ and $\mathcal{X} = \mathbb{R}^d$, a formal calculus shows that α_t is expected to solve a PDE of the form

$$\frac{\partial \alpha_t}{\partial t} = \operatorname{div}(\alpha_t \nabla(F'(\alpha_t))), \tag{9.28}$$

where $F'(\alpha)$ denotes the derivative of the function F in the sense that it is a continuous function $F'(\alpha) \in \mathcal{C}(\mathcal{X})$ such that

$$F(\alpha + \varepsilon\xi) = F(\alpha) + \varepsilon \int_{\mathcal{X}} F'(\alpha)\mathrm{d}\xi(x) + o(\varepsilon).$$

A typical example is when using $F = -H$, where $H(\alpha) = \mathrm{KL}(\alpha|\mathcal{L}_{\mathbb{R}^d})$ is the relative entropy with respect to the Lebesgue measure $\mathcal{L}_{\mathbb{R}^d}$ on $\mathcal{X} = \mathbb{R}^d$

$$H(\alpha) = -\int_{\mathbb{R}^d} \rho_\alpha(x)(\log(\rho_\alpha(x)) - 1)\mathrm{d}x \tag{9.29}$$

(setting $H(\alpha) = -\infty$ when α does not have a density), then (9.28) shows that the gradient flow of this neg-entropy is the linear heat diffusion

$$\frac{\partial \alpha_t}{\partial t} = \Delta \alpha_t, \tag{9.30}$$

where Δ is the spatial Laplacian. The heat diffusion can therefore be interpreted either as the "classical" Euclidian flow (somehow performing "vertical" movements with respect to mass amplitudes) of the Dirichlet energy $\int_{\mathbb{R}^d} \|\nabla \rho_\alpha(x)\|^2 \mathrm{d}x$ or, alternatively, as the entropy for the optimal transport flow (somehow a "horizontal" movement with respect to mass positions). Interest in Wasserstein

gradient flows was sparked by the seminal paper of Jordan, Kinderlehrer and Otto (Jordan *et al.*, 1998), and these evolutions are often called "JKO flows" following their work. As shown in detail in the monograph by Ambrosio *et al.* (2006), JKO flows are a special case of gradient flows in metric spaces. We also refer to the recent survey paper (Santambrogio, 2017). JKO flows can be used to study in particular nonlinear evolution equations such as the porous medium equation (Otto, 2001), total variation flows (Carlier and Poon, 2019), quantum drifts (Gianazza *et al.*, 2009), or heat evolutions on manifolds (Erbar, 2010). Their flexible formalism allows for constraints on the solution, such as the congestion constraint (an upper bound on the density at any point) that Maury *et al.* used to model crowd motion (Maury *et al.*, 2010) (see also the review paper (Santambrogio, 2018)).

Remark 9.11 (Gradient flows in metric spaces)**.** The implicit stepping (9.27) is a special case of a more general formalism to define gradient flows over metric spaces $(\mathcal{X}, d)$, where d is a distance, as detailed in (Ambrosio *et al.*, 2006). For some function $F(x)$ defined for $x \in \mathcal{X}$, the implicit discrete minmization step is then defined as

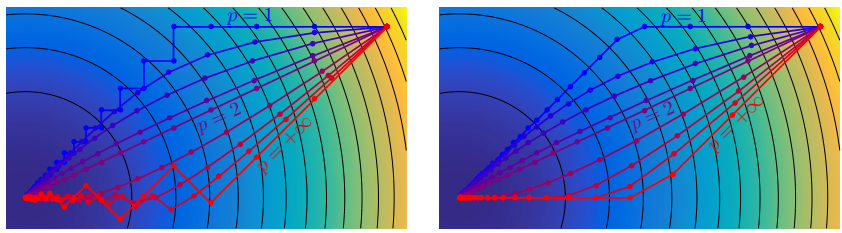
$$x^{(\ell+1)} \in \operatorname*{argmin}_{x \in \mathcal{X}} d(x^{(\ell)}, x)^2 + \tau F(x). \tag{9.31}$$

The JKO step (9.27) corresponds to the use of the Wasserstein distance on the space of probability distributions. In some cases, one can show that (9.31) admits a continuous flow limit x_t as $\tau \to 0$ and $k\tau = t$. In the case that $\mathcal{X}$ also has a Euclidean structure, an explicit stepping is defined by linearizing F

$$x^{(\ell+1)} = \operatorname*{argmin}_{x \in \mathcal{X}} d(x^{(\ell)}, x)^2 + \tau \langle \nabla F(x^{(\ell)}),\, x \rangle. \tag{9.32}$$

In sharp contrast to the implicit formula (9.31) it is usually straightforward to compute but can be unstable. The implicit step is always stable, is also defined for nonsmooth F, but is usually not accessible in closed form. Figure 9.7 illustrates this concept on

the function $F(x) = \|x\|^2$ on $\mathcal{X} = \mathbb{R}^2$ for the distances $d(x,y) = \|x-y\|_p = (|x_1-y_1|^p + |x_2-y_2|^p)^{\frac{1}{p}}$ for several values of p. The explicit scheme (9.32) is unstable for $p=1$ and $p=+\infty$, and for $p=1$ it gives axis-aligned steps (coordinatewise descent). In contrast, the implicit scheme (9.31) is stable. Note in particular how, for $p=1$, when the two coordinates are equal, the following step operates in the diagonal direction.



Explicit Implicit

Figure 9.7: Comparison of explicit and implicit gradient flow to minimize the function $f(x) = \|x\|^2$ on $\mathcal{X} = \mathbb{R}^2$ for the distances $d(x,y) = \|x-y\|_p$ for several values of p.

Remark 9.12 (Lagrangian discretization using particles systems)**.** The finite-dimensional problem in (9.26) can be interpreted as the Eulerian discretization of a flow over the space of measures (9.27). An alternative way to discretize the problem, using the so-called Lagrangian method using particles systems, is to parameterize instead the solution as a (discrete) empirical measure moving with time, where the locations of that measure (and not its weights) become the variables of interest. In practice, one can consider a dynamic point cloud of particles $\alpha_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ indexed with time. The initial problem (9.26) is then replaced by a set of n coupled ODE prescribing the dynamic of the points $X(t) = (x_i(t))_i \in \mathcal{X}^n$. If the energy F is finite for discrete measures, then one can simply define $\mathcal{F}(X) = F(\frac{1}{n}\sum_{i=1}^n \delta_{x_i})$. Typical examples are linear functions $F(\alpha) = \int_{\mathcal{X}} V(x)\mathrm{d}\alpha(x)$ and quadratic interactions $F(\alpha) =$

$\int_{\mathcal{X}^2} W(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$, in which case one can use respectively

$$\mathcal{F}(X) = \frac{1}{n}\sum_i V(x_i) \quad \text{and} \quad \mathcal{F}(X) = \frac{1}{n^2}\sum_{i,j} W(x_i,x_j).$$

For functions such as generalized entropy, which are only finite for measures having densities, one should apply a density estimator to convert the point cloud into a density, which allows us to also define function $\mathcal{F}(x)$ consistent with F as $n \to +\infty$. A typical example is for the entropy $F(\alpha) = H(\alpha)$ defined in (9.29), for which a consistent estimator (up to a constant term) can be obtained by summing the logarithms of the distances to nearest neighbors

$$\mathcal{F}(X) = \frac{1}{n}\sum_i \log(d_X(x_i)) \quad \text{where} \quad d_X(x) = \min_{x'\in X, x'\neq x} \|x-x'\|\,; \tag{9.33}$$

see Beirlant *et al.* (1997) for a review of nonparametric entropy estimators. For small enough step sizes τ, assuming $\mathcal{X} = \mathbb{R}^d$, the Wasserstein distance $\mathcal{W}_2$ matches the Euclidean distance on the points, *i.e.* if $|t-t'|$ is small enough, $\mathcal{W}_2(\alpha_t,\alpha_{t'}) = \|X(t)-X(t')\|$. The gradient flow is thus equivalent to the Euclidean flow on positions $X'(t) = -\nabla\mathcal{F}(X(t))$, which is discretized for times $t_k = \tau k$ similarly to (9.24) using explicit Euler steps

$$X^{(\ell+1)} \overset{\text{def.}}{=} X^{(\ell)} - \tau\nabla\mathcal{F}(X^{(\ell)}).$$

Figure 9.8 shows an example of such a discretized explicit evolution for a linear plus entropy functional, resulting in a discretized version of a Fokker–Planck equation. Note that for this particular case of linear Fokker–Planck equation, it is possible also to resort to stochastic PDEs methods, and it can be approximated numerically by evolving a single random particle with a Gaussian drift. The convergence of these schemes (so-called Langevin Monte Carlo) to the stationary distribution can in turn be quantified in terms of Wasserstein distance; see, for instance, (Dalalyan and Karagulyan, 2017). If the function $\mathcal{F}$ is not smooth, one should discretize

similarly to (9.25) using implicit Euler steps, *i.e.* consider

$$X^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Prox}^{\|\cdot\|}_{\tau\mathcal{F}}(X^{(\ell)}) \overset{\text{def.}}{=} \operatorname*{argmin}_{Z\in\mathcal{X}^n} \frac{1}{2}\left\|Z - X^{(\ell)}\right\|^2 + \tau\mathcal{F}(Z).$$

In the simplest case of a linear function $F(\alpha) = \int_X V(x)\mathrm{d}\alpha(x)$, the flow operates independently over each particule $x_i(t)$ and corresponds to a usual Euclidean flow for the function V, $x_i'(t) = -\nabla V(x_i(t))$ (and is an advection PDEs of the density along the integral curves of the flow).

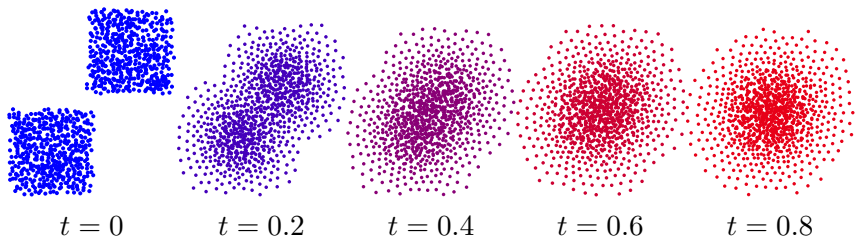


Figure 9.8: Example of gradient flow evolutions using a Lagrangian discretization, for the function $F(\alpha) = \int V\mathrm{d}\alpha - H(\alpha)$, for $V(x) = \|x\|^2$. The entropy is discretized using (9.33). The limiting stationary distribution is a Gaussian.

Remark 9.13 (Geodesic convexity)**.** An important concept related to gradient flows is the convexity of the functional F with respect to the Wasserstein-2 geometry, *i.e.* the convexity of F along Wasserstein geodesics (*i.e.* displacement interpolations as shown in Remark 7.1). The Wasserstein gradient flow (with a continuous time) for such a function exists, is unique, and is the limit of the discrete stepping (9.27) as $\tau \to 0$. It converges to a fixed stationary distribution as $t \to +\infty$. The entropy is a typical example of geodesically convex function, and so are linear functions of the form $F(\alpha) = \int_{\mathcal{X}} V(x)\mathrm{d}\alpha(x)$ and quadratic interaction functions $F(\alpha) = \int_{\mathcal{X}\times\mathcal{X}} W(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$ for convex functions $V : \mathcal{X} \to \mathbb{R}$, $W : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Note that while linear functions are convex in the classical sense, quadratic interaction functions might fail to be. A typical example is $W(x,y) = \|x-y\|^2$, which is a negative semi-

definite kernel (see Definition 8.3) and thus corresponds to $F(\alpha)$ being a concave function in the usual sense (while it is geodesically convex). An important result of McCann (1997) is that generalized "entropy" functions of the form $F(\alpha) = \int_{\mathbb{R}^d} \varphi(\rho_\alpha(x))\mathrm{d}x$ on $\mathcal{X} = \mathbb{R}^d$ are geodesically convex if φ is convex, with $\varphi(0) = 0$, $\varphi(t)/t \to +\infty$ as $t \to +\infty$ and such that $s \mapsto s^d \varphi(s^{-d})$ is convex decaying.

There is important literature on the numerical resolution of the resulting discretized flow, and we give only a few representative publications. For 1-D problems, very precise solvers have been developed because OT is a quadratic functional in the inverse cumulative function (see Remark 2.30): Kinderlehrer and Walkington (1999), Blanchet *et al.* (2008), Agueh and Bowles (2013), Matthes and Osberger (2014), and Blanchet and Carlier (2015). In higher dimensions, it can be tackled using finite elements and finite volume schemes: Carrillo *et al.* (2015) and Burger *et al.* (2010). Alternative solvers are obtained using Lagrangian schemes (*i.e.* particles systems): Carrillo and Moll (2009), Benamou *et al.* (2016a), and Westdickenberg and Wilkening (2010). Another direction is to look for discrete flows (typically on discrete grids or graphs) which maintain some properties of their continuous counterparts; see Mielke (2013), Erbar and Maas (2014), Chow *et al.* (2012), and Maas (2011).

An approximate approach to solve the Eulerian discretized problem (9.24) relying on entropic regularization was initially proposed in Peyré (2015), refined in Chizat *et al.* (2018b) and theoretically analyzed in Carlier *et al.* (2017). With an entropic regularization, Problem (9.26) has the form (4.49) when setting $G = \iota_{\mathbf{a}^{(\ell)}}$ and replacing F by τF. One can thus use the iterations (4.51) to approximate $\mathbf{a}^{(\ell+1)}$ as proposed initially in Peyré (2015). The convergence of this scheme as $\varepsilon \to 0$ is proved in Carlier *et al.* (2017). Figure 9.9 shows an example of evolution computed with this method. An interesting application of gradient flows to machine learning is to learn the underlying function F that best models some dynamical model of density. This learning can be achieved by solving a smooth nonconvex optimization using entropic regularized transport and automatic differentiation (see Remark 9.1.3); see Hashimoto *et al.* (2016).

Analyzing the convergence of gradient flows discretized in both time and space is difficult in general. Due to the polyhedral nature of the linear program defining the distance, using too-small step sizes leads to a "locking" phenomena (the distribution is stuck and does not evolve, so that the step size should be not too small, as discussed in (Maury and Preux, 2017)). We refer to (Matthes and Osberger, 2014; Matthes and Osberger, 2017) for a convergence analysis of a discretization method for gradient flows in one dimension.

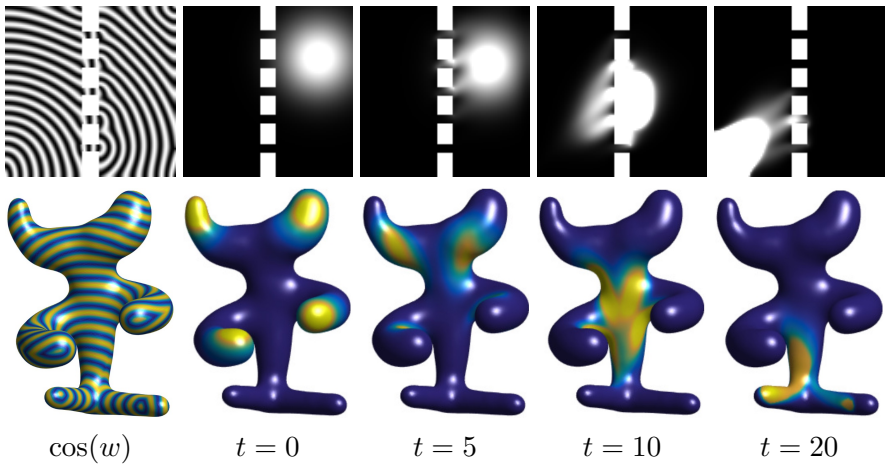


Figure 9.9: Examples of gradient flows evolutions, with drift V and congestion terms (from Peyré (2015)), so that $F(\alpha) = \int_{\mathcal{X}} V(x)\mathrm{d}\alpha(x) + \iota_{\leq\kappa}(\rho_\alpha)$.

It is also possible to compute gradient flows for unbalanced optimal transport distances as detailed in §10.2. This results in evolutions allowing mass creation or destruction, which is crucial to model many physical, biological or chemical phenomena. An example of unbalanced gradient flow is the celebrated Hele-Shaw model for cell growth (Perthame *et al.*, 2014), which is studied theoretically in (Gallouët and Monsaingeon, 2017; Di Marino and Chizat, 2017). Such an unbalanced gradient flow also can be approximated using the generalized Sinkhorn algorithm (Chizat *et al.*, 2018b).

9.4 Minimum Kantorovich Estimators

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

$$\min_{\theta \in \Theta} \mathcal{L}(\alpha_\theta, \beta) \quad \text{where} \quad \beta = \frac{1}{n} \sum_i \delta_{x_i}, \tag{9.34}$$

where $\mathcal{L}$ is some "loss" function between a discrete and a "continuous" (arbitrary) distribution (see Figure 9.10).

In the case where α_θ as a density $\rho_\theta \overset{\text{def.}}{=} \rho_{\alpha_\theta}$ 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

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

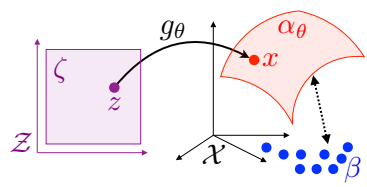


Figure 9.10: Schematic display of the density fitting problem 9.34.

This MLE approach is known to lead to optimal estimation procedures in many cases (see, for instance, Owen (2001)). However, it fails to work when estimating singular distributions, typically when the α_θ does not have 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 $\text{KL}(\alpha_\theta | \bar{\beta})$ 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 2.1. 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 a 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

$$\mathcal{L}(\alpha,\beta) \overset{\text{def.}}{=} \max_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \left\{ \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{X}} g(x)\mathrm{d}\beta(x) \;:\; (f,g)\in\mathcal{R} \right\}. \tag{9.35}$$

Dual norms shown in §8.2 correspond to imposing

$$\mathcal{R} = \{(f,-f) \;:\; f \in B\},$$

while optimal transport (2.24) sets $\mathcal{R} = \mathcal{R}(c)$ as defined in (2.25).

For a fixed θ, evaluating the energy to be minimized in (9.34) using such a loss function corresponds to solving a semidiscrete optimal transport, which is the focus of Chapter 5. Minimizing the energy with respect to θ is much more involved and is typically highly nonconvex.

Denoting f_θ a solution to (9.35) when evaluating $\mathcal{E}(\theta) = \mathcal{L}(\alpha_\theta,\beta)$, a subgradient is obtained using the formula

$$\nabla\mathcal{E}(\theta) = \int_{\mathcal{X}} [\partial h_\theta(x)]^\top \nabla f_\theta(x)\mathrm{d}\alpha_\theta(x), \tag{9.36}$$

where $\partial h_\theta(x) \in \mathbb{R}^{\dim(\Theta)\times d}$ is the differential (with respect to θ) of $\theta \in \mathbb{R}^{\dim(\Theta)} \mapsto h_\theta(x)$, while $\nabla f_\theta(x)$ is the gradient (with respect to x) of f_θ. This formula is hard to use numerically, first because it requires first computing a continuous function f_θ, which is a solution to a semidiscrete problem. As shown in §8.5, for OT loss, this can be achieved

using stochastic optimization, but this is hardly applicable in high dimension. Another option is to impose a parametric form for this potential, for instance expansion in an RKHS (Genevay *et al.* (2016)) or a deep-network approximation ((Arjovsky *et al.*, 2017)). This, however, leads to important approximation errors that are not yet analyzed theoretically. A last issue is that it is unstable numerically because it requires the computation of the gradient ∇f_θ of the dual potential f_θ.

For the OT loss, an alternative gradient formula is obtained when one rather computes a primal optimal coupling for the following equivalent problem:

$$\mathcal{L}_c(\alpha_\theta, \beta) = \min_{\gamma \in \mathcal{M}(\mathcal{Z}\times\mathcal{X})} \left\{ \int_{\mathcal{Z}\times\mathcal{X}} c(h_\theta(z), x) \mathrm{d}\gamma(z,x) \;:\; \gamma \in \mathcal{U}(\zeta, \beta) \right\}. \tag{9.37}$$

Note that in the semidiscrete case considered here, the objective to be minimized can be actually decomposed as

$$\min_{(\gamma_i)_{i=1}^n} \sum_{i=1}^n \int_{\mathcal{Z}} c(h_\theta(z), x_i) \mathrm{d}\gamma_i(z) \quad \text{where} \quad \sum_{i=1}^n \gamma_i = \zeta, \quad \int_{\mathcal{Z}} \mathrm{d}\gamma_i(z) = \frac{1}{n}, \tag{9.38}$$

where each $\gamma_i \in \mathcal{M}_+^1(\mathcal{Z})$. Once an optimal $(\gamma_{\theta,i})_i$ solving (9.38) is obtained, the gradient of $\mathcal{E}(\theta)$ is computed as

$$\nabla \mathcal{E}(\theta) = \sum_{i=1}^n \int_{\mathcal{Z}} [\partial h_\theta(z)]^\top \nabla_1 c(h_\theta(z), x_i) \mathrm{d}\gamma_i(z),$$

where $\nabla_1 c(x,y) \in \mathbb{R}^d$ is the gradient of $x \mapsto c(x,y)$. Note that as opposed to (9.36), this formula does not involve computing the gradient of the potentials being solutions of the dual OT problem.

The class of estimators obtained using $\mathcal{L} = \mathcal{L}_c$, often called "minimum Kantorovich estimators," was initially introduced in (Bassetti *et al.*, 2006); see also (Canas and Rosasco, 2012). It has been used in the context of generative models by (Montavon *et al.*, 2016) to train restricted Boltzmann machines and in (Bernton *et al.*, 2017) in conjunction with approximate Bayesian computations. Approximations of these computations using Deep Network are used to train deep generative models for both GAN (Arjovsky *et al.*, 2017) and VAE (Bousquet *et al.*, 2017); see also (Genevay *et al.*, 2018; Genevay *et al.*, 2017; Salimans

et al., 2018). Note that the use of Sinkhorn divergences for parametric model fitting is used routinely for shape matching and registration, see (Gold *et al.*, 1998; Chui and Rangarajan, 2000; Myronenko and Song, 2010; Feydy *et al.*, 2017).

Remark 9.14 (Metric learning and transfer learning)**.** Let us insist on the fact that, for applications in machine learning, the success of OT-related methods very much depends on the choice of an adapted cost $c(x, y)$ which captures the geometry of the data. While it is possible to embed many kinds of data in Euclidean spaces (see, for instance, (Mikolov *et al.*, 2013) for words embedding), in many cases, some sort of adaptation or optimization of the metric is needed. Metric learning for supervised tasks is a classical problem (see, for instance, (Kulis, 2012; Weinberger and Saul, 2009)) and it has been extended to the learning of the ground metric $c(x, y)$ when some OT distance is used in a learning pipeline (Cuturi and Avis, 2014) (see also Zen *et al.* 2014; Wang and Guibas 2012; Huang *et al.* 2016). Let us also mention the related inverse problem of learning the cost matrix from the observations of an optimal coupling $\mathbf{P}$, which can be regularized using a low-rank prior (Dupuy *et al.*, 2016). Related problems are transfer learning (Pan and Yang, 2010) and domain adaptation (Glorot *et al.*, 2011), where one wants to transfer some trained machine learning pipeline to adapt it to some new dataset. This problem can be modeled and solved using OT techniques; see (Courty *et al.*, 2017b; Courty *et al.*, 2017a).