Computational Optimal Transport

Gabriel Peyré
CNRS and DMA, ENS

Marco Cuturi
CREST, ENSAE

Contents

1 Introduction **3**
- 1.1 Notations 4

2 Theoretical Foundations **7**
- 2.1 Histograms and Measures 7
- 2.2 Assignment and Monge Problem 9
- 2.3 Kantorovich relaxation 15
- 2.4 Metric Properties of Optimal Transport 22
- 2.5 Dual Problem 27
- 2.6 Special Cases 35

3 Algorithmic Foundations **42**
- 3.1 The Kantorovich Linear Programs 43
- 3.2 **C**-transforms 45
- 3.3 Complementary Slackness 47
- 3.4 Vertices of the Transportation Polytope 48
- 3.5 A Heuristic Description of the Network Simplex 52
- 3.6 Matching problems 57

4 Entropic Regularization of Optimal Transport **63**
- 4.1 Entropic Regularization 63
- 4.2 Sinkhorn's Algorithm and its Convergence 69

4.3 Speeding-up Sinkhorn's Iterations . . . 81
4.4 Regularized Dual and Log-domain Computations . . . 84
4.5 Regularized Approximations of the Optimal Transport Cost 87
4.6 Generalized Sinkhorn . . . 89

5 Semi-discrete Optimal Transport 92
5.1 c-transform and $\bar{c}$-transform . . . 93
5.2 Semi-discrete Formulation . . . 94
5.3 Entropic Semi-discrete Formulation . . . 97
5.4 Stochastic Optimization Methods . . . 101

6 $\mathcal{W}_1$ Optimal Transport 105
6.1 $\mathcal{W}_1$ on Metric Spaces . . . 106
6.2 $\mathcal{W}_1$ on Euclidean Space . . . 108
6.3 $\mathcal{W}_1$ on a Graph . . . 109

7 Dynamic Formulations 111
7.1 Continuous Formulation . . . 111
7.2 Discretization on Uniform Staggered Grids . . . 115
7.3 Proximal Solvers . . . 116
7.4 Dynamical Unbalanced OT . . . 119
7.5 More General Mobility Functionals . . . 120
7.6 Dynamic Formulation over the Paths Space . . . 121

8 Statistical Divergences 125
8.1 φ-Divergences . . . 125
8.2 Integral Probability Metrics . . . 132
8.3 Wasserstein Spaces are not Hilbertian . . . 138
8.4 Empirical Estimators for OT, MMD and φ-divergences . . . 142
8.5 Entropic Regularization: between OT and MMD . . . 146

9 Variational Wasserstein Problems 148
9.1 Differentiating the Wasserstein Loss . . . 149
9.2 Wasserstein Barycenters, Clustering and Dictionary Learning 154
9.3 Gradient Flows . . . 166
9.4 Minimum Kantorovitch Estimators . . . 173

10 Extensions of Optimal Transport **179**
10.1 Multi-marginal Problems 179
10.2 Unbalanced Optimal Transport 183
10.3 Problems with Extra Constraints on the Couplings 186
10.4 Sliced Wasserstein Distance and Barycenters 187
10.5 Transporting Vectors and Matrices 192
10.6 Gromov-Wasserstein Distances 194

References **202**

Abstract

Optimal Transport (OT) is a mathematical gem at the interface between probability, analysis and optimization. The goal of that theory is to define geometric tools that are useful to compare probability distributions. Let us briefly sketch some key ideas using a vocabulary that was first introduced by Monge two centuries ago: a probability distribution can be thought of as a pile of sand. Peaks indicate where likely observations are to appear. Given a pair of probability distributions—two different piles of sand—there are, in general, multiple ways to morph, *transport* or reshape the first pile so that it matches the second. To every such transport we associate an a "global" cost, using the "local" consideration of how much it costs to move a single grain of sand from one location to another. The goal of optimal transport is to find the least costly transport, and use it to derive an entire geometric toolbox for probability distributions.

Despite this relatively abstract description, optimal transport theory answers many basic questions related to the way our economy works: In the "mines and factories" problem, the sand is distributed across an entire country, each grain of sand represents a unit of a useful raw resource; the target pile indicates where those resources are needed, typically in factories, where they are meant to be processed. In that scenario, one seeks the least costly way to move all these resources, knowing the entire logistic cost matrix needed to ship resources from any storage point to any factory.

Transporting optimally two abstract distributions is also extremely relevant for mathematicians, in the sense that it defines a rich geometric structure on the space of probability distributions. That structure is canonical in the sense that it borrows, in arguably the most natural way, key geometric properties of the underlying "ground" space on which these distributions are defined. For instance, when the underlying space is Euclidean, key concepts such as interpolation, barycenters, convexity or gradients of functions extend very naturally to distributions when endowed with an optimal transport geometry. OT has a rich and varied history. Earlier contributions originated from Monge's work in the 18th century, to be later rediscovered under a different formalism by Tolstoi

in the 1920's, Kantorovich, Hitchcock and Koopmans in the 1940's. The problem was solved numerically by Dantzig in 1949 and others in the 1950's within the framework of linear programming, paving the way for major industrial applications in the second half of the 20th century. OT was later rediscovered under a different light by analysts in the 90's, following important work by Brenier and others, as well as in the computer vision/graphics fields under the name of earth mover's distances. Recent years have witnessed yet another revolution in the spread of OT, thanks to the emergence of approximate solvers that can scale to sizes and dimensions that are relevant to data sciences. Thanks to this newfound scalability, OT is being increasingly used to unlock various problems in imaging sciences (such as color or texture processing), computer vision and graphics (for shape manipulation) or machine learning (for regression, classification and density fitting). This paper reviews OT with a bias toward numerical methods and their applications in data sciences, and sheds lights on the theoretical properties of OT that make it particularly useful for some of these applications. Our focus is on the recent wave of efficient algorithms that have helped translate attractive theoretical properties onto elegant and scalable tools for a wide variety of applications. We also give a prominent place to the many generalizations of OT that have been proposed in but a few years, and connect them with related approaches originating from statistical inference, kernel methods and information theory. A companion website [1] provides bibliographical and numerical resources, and in particular gives access to all the open source software needed to reproduce the figures of this article.

[1] https://optimaltransport.github.io/

1

Introduction

Optimal Transport (OT) has a long and rich history, initiated by Monge in the 18th century [Monge, 1781], then stated in its modern form by Kantorovich [1942], and revitalized in the 90s by a flurry of major mathematical results such as that of Brenier [1991]. Several reference books have been written on this topic, including the two monographs by Villani (2003, 2009), those by Rachev and Rüschendorf [1998a, 1998b] and more recently 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, see 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] trying 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 to explain 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 defined by OT on discrete histograms. Targetting more advanced readers, we also give in the same chapters a deeper and more mathematical exposition using discrete measures in light gray boxes. This corresponds to introducing the support (positions) associated with the bins of the histogram, giving a second important degree of freedom when comparing probability measures (not only variable weights but also variable locations for such weights). 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 (including in particular those with a density w.r.t. a base measure). This is traditionally the default setting for most classic textbooks on OT theory. Chapters 5 to 8 deal with arbitrary measures, and are thus targeting a more mathematically-inclined audience.

1.1 Notations

- $\mathbb{1}_{n,m}$: matrix of $\mathbb{R}^{n\times m}$ with all entries identically set to 1. $\mathbb{1}_n$: vector of ones.
- 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 $(\mathcal{X}, \mathcal{Y})$.
- $\rho_\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}x}$: density with respect to the Lebesgues 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 $\mathbf{U}(\mathbf{a}, \mathbf{b})$.
- $T : \mathcal{X} \to \mathcal{Y}$: Monge map, typically such that $T_\sharp \alpha = \beta$.
- $\mathcal{R}(\alpha, \beta)$: set of admissible dual potentials; for discrete measures $\mathbf{R}(\mathbf{a}, \mathbf{b})$.
- $(\alpha_t)_{t=0}^1$: 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 $f \oplus g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
- $\alpha \otimes \beta$ is the product measure on $\mathcal{X} \times \mathcal{Y}$, *i.e.* $\int_{\mathcal{X}\times\mathcal{Y}} g(x, y)\mathrm{d}(\alpha \otimes \beta)(x, y) \stackrel{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} g(x, y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y)$.
- $\mathbf{a} \otimes \mathbf{b} \stackrel{\text{def.}}{=} \mathbf{a}\mathbf{b}^\top \in \mathbb{R}^{n\times m}$.
- $\mathbf{f} \oplus \mathbf{g} \stackrel{\text{def.}}{=} \mathbf{f}\mathbb{1}_m^\top + \mathbb{1}_n\mathbf{g}^\top \in \mathbb{R}^{n\times m}$, for two vectors $\mathbf{f} \in \mathbb{R}^n$, $\mathbf{g} \in \mathbb{R}^m$.
- $\mathbf{u} \odot \mathbf{v} = (\mathbf{u}_i\mathbf{v}_i) \in \mathbb{R}^n$ for $(\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^n)^2$

2

Theoretical Foundations

This chapter describes the basics of optimal transport, introducing first the notion of optimal couplings between probability vectors $(\mathbf{a}, \mathbf{b})$, then relating this computation to the transport between discrete measures (α, β) defined in embedding spaces $(\mathcal{X}, \mathcal{Y})$, and lastly covering the general setting of arbitrary measures. In a first reading, one 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 and math-inclined reader will be able to grasp more intuition and more general formulation (e.g. in order to move positions of clouds of points or handle measures with continuous densities) using the more general measure setting.

2.1 Histograms and Measures

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

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

A large part of this work focuses exclusively on the study of the geometry induced by optimal transport on the simplex. For more advanced readers, we give a deeper and more mathematical exposition using the formalism of discrete measures, which corresponds to handling both the weights contained in a probability vector and positions (the support of the measure) associated with the bins of the histogram. This formalism can be used implicitly to handle histograms of arbitrary size n and with possibly varying positions. Lastly, the third and most technical layer of exposition deals with arbitrary measures (*i.e.* which need not be discrete, typically continuous with respect to a base measure).

Remark 2.1 (Discrete measures)**.** A discrete measure with weights $\mathbf{a}$ and locations $x_1, \dots, x_n \in \mathcal{X}$ 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 as measure describes a probability measure if, additionally, $\mathbf{a} \in \Sigma_n$, and more generally a positive measure if each of the "weights" described in vector $\mathbf{a}$ is positive itself.

Remark 2.2 (General measures)**.** A convenient feature of OT is that it can deal with discrete and continuous "objects" within the same framework. Such objects only need to be modelled as measures. This corresponds to the notion of Radon measures $\mathcal{M}(\mathcal{X})$ on the space $\mathcal{X}$. The formal definition of that set requires that $\mathcal{X}$ is equipped with a distance, usually denoted d, because one can only access a measure by "testing" (integrating) it against continuous functions, denoted $f \in \mathcal{C}(\mathcal{X})$.

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

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

More general measures, for instance on $\mathcal{X} = \mathbb{R}^d$ (where $d \in \mathbb{N}^*$

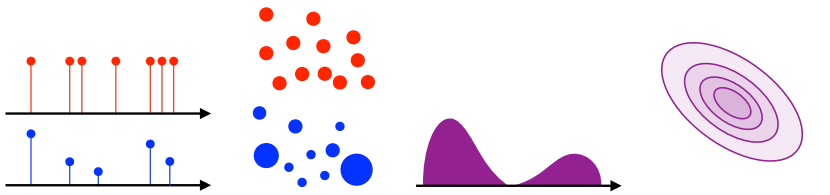


Figure 2.1: Schematic display of discrete distributions $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ (red corresponds to empirical uniform distribution $\mathbf{a}_i = 1/n$, and blue to arbitrary distributions) and densities $\mathrm{d}\alpha(x) = \rho_\alpha(x)\mathrm{d}x$ (in violet), in both 1-D and 2-D. Discrete distributions in 1-D are displayed using vertical segments (with length equal to $\mathbf{a}_i$) and in 2-D using point clouds (radius equal to $\mathbf{a}_i$).

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 needs not to have a density nor be a sum of Diracs) is defined by the fact that it can be integrated agains any continuous function $f \in \mathcal{C}(\mathcal{X})$ and obtain $\int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) \in \mathbb{R}$. If $\mathcal{X}$ is not compact, one should also impose that f has compact support or at least as 0 limit at infinity. Measure as thus in some sense "less regular" than functions, but more regular than distributions (which are dual to smooth functions). For instance, the derivative of a Dirac is not a measure. We denote $\mathcal{M}_+(\mathcal{X})$ the set of all positive measures on $\mathcal{X}$. The set of probability measures is denoted $\mathcal{M}_+^1(\mathcal{X})$, which means that any $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ is positive, and that $\alpha(\mathcal{X}) = \int_{\mathcal{X}} \mathrm{d}\alpha = 1$. Figure 2.1 offers a visualization of the different classes of measures, beyond histograms, considered in this work.

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 per-

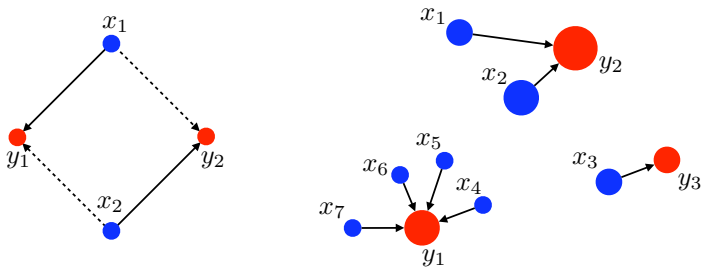


Figure 2.2: (left) blue dots from measure α and red dots from measure β are pairwise equidistant. Hence, either matching $\sigma = (1, 2)$ (full line) or $\sigma = (2, 1)$ (dotted line) is optimal. (right) a Monge map can associate the blue measure α to the red measure β. The weights α_i are displayed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

mutations 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 only be solved if there exist efficient algorithms to optimize that cost function over the set of permutations, which will be the subject of §3.6.

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 4 corners of a 2-dimensional square of side length 1, as represented in the left plot in Figure 2.2. In that case only two assignments exist, and they share the same cost.

Remark 2.4 (Monge Problem between discrete measures)**.** For dis-

crete measures

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

Monge problem [Monge, 1781] seeks for a map that associates to each point x_i a single point y_j, and which must push the mass of α toward the mass of β, which is to say that such a map $T : \{x_1, \dots, x_n\} \to \{y_1, \dots, y_m\}$ 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 $T_\sharp \alpha = \beta$. 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

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

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 an empirical measure to another. This happens when their weight vectors are not compatible, which is always the case when the target measure has more points than the source

measure. For instance, the right plot in Figure 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 some continuous map $T : \mathcal{X} \to \mathcal{Y}$, we define the pushforward operator $T_\sharp : \mathcal{M}(X) \to \mathcal{M}(Y)$. For discrete measures (2.1), the pushforward operation consists simply in moving the positions of all the points in the support of the measure

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

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

Definition 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 $\alpha \in \mathcal{M}(\mathcal{X})$ reads

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

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

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

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. The more general extension $T_\sharp$ can now "move" an entire probability measure on $\mathcal{X}$ towards a new probability measure on $\mathcal{Y}$. The operator $T_\sharp$ "pushes forward" each elementary mass of a measure α on $\mathcal{X}$ by applying the map T to obtain then an elementary mass in $\mathcal{Y}$, to build on aggregate a new measure on $\mathcal{Y}$) written $T_\sharp \alpha$. Note that such a push-forward $T_\sharp : \mathcal{M}_+^1(\mathcal{X}) \to$

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

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

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

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, denoting $y = T(x)$

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

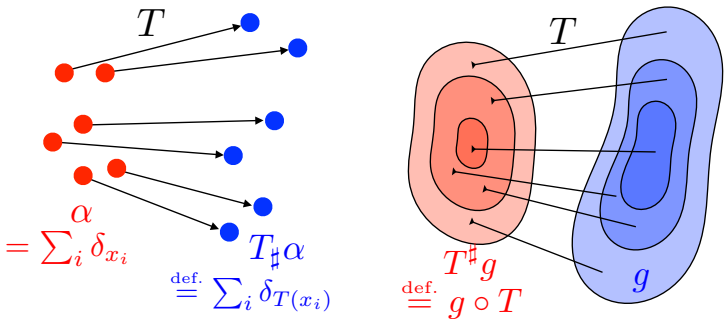
Remark 2.7 (Monge problem between arbitrary measures)**.** Monge problem (2.5) is extended to the setting of two arbitrary probability measures (α, β) on two spaces $(\mathcal{X}, \mathcal{Y})$ as finding a map $T : \mathcal{X} \to \mathcal{Y}$ that minimizes

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

The constraint $T_\sharp\alpha = \beta$ means that T pushes forward the mass of α to β, and makes use of the push-forward operator (2.6).

Remark 2.8 (Push-forward vs. pull-back)**.** The push-forward $T_\sharp$ of measures should not be confounded with the pull-back of function $T^\sharp : \mathcal{C}(\mathcal{Y}) \to \mathcal{C}(\mathcal{X})$ which corresponds to the "warping" of functions. It is the linear map defined, for $g \in \mathcal{C}(\mathcal{Y})$ by $T^\sharp g = g \circ T$. Push-forward and pull-back are actually adjoint one from each others, in the sense that

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



Push-forward of measures Pull-back of functions

Figure 2.3: Comparison of push-forward $T_\sharp$ and pull-back $T^\sharp$.

It is important to realize that even if (α, β) have densities $(\rho_\alpha, \rho_\beta)$, $T_\sharp \alpha$ is not equal to $T^\sharp \rho_\beta$, because of the presence of the Jacobian in (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 illustrate the distinction between these push-forward and pull-back operators.

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 un-specified) probabilized 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

Limitations of the Monge Problem The assignment problem has several limitations in practical settings, also encountered when using the Monge problem. Indeed, because the assignment problem is formulated as a permutation problem, it can only be used to compare two points clouds of the *same* size. A direct generalization to discrete measures with non-uniform weights can be carried out using Monge's formalism of pushforward maps, but that formulation may also be degenerate it there does not exist feasible solutions satisfying the mass conservation constraint (2.4) (see the end of Remark 2.4). Additionally, the assignment Problem (2.5) is combinatorial, whereas the feasible set for the Monge Problem (2.9), consisting in all push-forward measures that satisfy the mass conservation constraint, is *non-convex*. Both are therefore difficult to solve in their original formulation.

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, or transported to one and one location $T(x_i)$ only. Kantorovich proposes instead that the mass at any point x_i be potentially dispatched across several locations. Kantorovich moves away from the idea that mass transportation should be "deterministic" to consider instead a "probabilistic" (or "fuzzy") transportation, which allows what is commonly known now as "mass splitting" from a source towards several targets. This flexibility is encoded using, in place of a permutation σ or a map T, a coupling matrix $\mathbf{P} \in \mathbb{R}_+^{n \times m}$, where $\mathbf{P}_{i,j}$ describes the amount of mass flowing from bin i (or point x_i) towards bin j (or point x_j), x_i towards y_j in the formalism of discrete measures (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, defined by $n + m$ equality constraints, and therefore a convex polytope (the convex hull of a finite set of matrices).

Additionally, whereas the Monge formulation (as illustrated in the right plot of Figure 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 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 be all 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 so-called Birkhoff polytope $\mathbf{U}(\mathbb{1}_n/n, \mathbb{1}_n, n)$. Indeed, for any permutation σ we have $\mathbf{P}_\sigma \mathbb{1} = \mathbb{1}_n$ and ${\mathbf{P}_\sigma}^{\mathrm{T}} \mathbb{1} = \mathbb{1}_n$, whereas $\mathbb{1}_n {\mathbb{1}_n}^{\mathrm{T}}/n^2$ is a valid coupling but not a permutation matrix. Therefore, one has naturally that

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

The following proposition shows that these problems result in fact in the same optimum, namely that one can always find a permutation matrix that minimizes Kantorovich's problem (2.11) between two uniform measures $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, which shows that the Kantorovich relaxation is *tight* when considered on assignment problems. Figure 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) $\mathbf{P}_{\sigma^\star}$, which is a permutation matrix associated to an optimal permutation $\sigma^\star \in \mathrm{Perm}(n)$ for Problem (2.2).

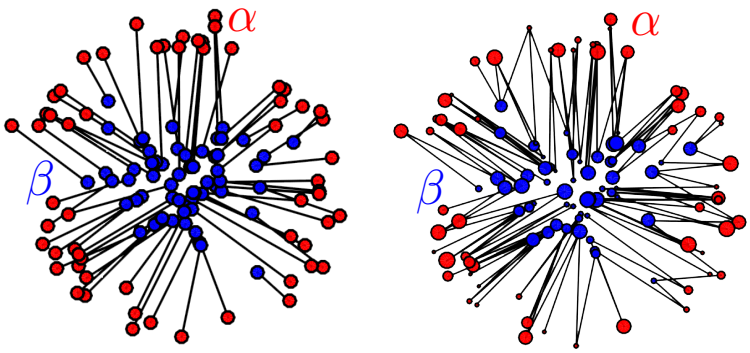


Figure 2.4: Comparison of optimal matching and generic couplings. A black segment between x_i and y_j indicates a non-zero element in the displayed optimal coupling $\mathbf{P}_{i,j}$ solving (2.11). Left: optimal matching, corresponding to the setting of Proposition (2.1) (empirical measures with the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

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 non-empty 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

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

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 matrix $\mathbf{C}$ depends on the support of α and β. The notation $\mathcal{L}_c(\alpha, \beta)$ is however useful in some situation, because it makes explicit the

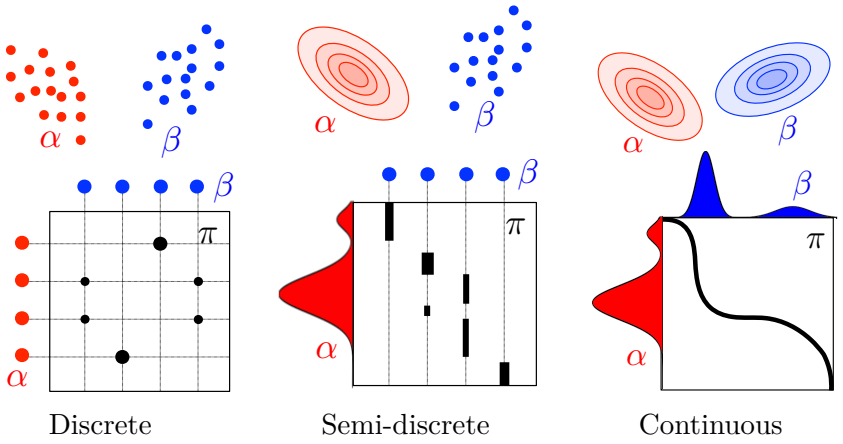


Figure 2.5: Schematic viewed of input measures (α, β) and couplings $\mathcal{U}(\alpha, \beta)$ encountered in the three main scenario for Kantorovich OT. Chapter 5 is dedicated to the semi-discrete setup.

dependency with respect to *both* probability weights and supporting points, the latter being exclusively considered through the cost function c.

Remark 2.12 (Applications of optimal assignment and couplings)**.** The optimal transport 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], 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, 2014, Su et al., 2015], image watermarking [Mathon et al., 2014]. In astrophysics, OT has been used for reconstructing the early universe [Frisch et al., 2002]. Optimal transport has been used for music transcription [Flamary et al., 2016]. It also 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)**.** The definition of $\mathcal{L}_c$ in (2.13) can be extended to arbitrary measures by considering couplings $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ which are joint distributions over the product space. The discrete case is a special situation where one imposes this product measure to be of the form $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}$. In the general case, the mass conservation constraint (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-forward (see Definition 2.1) by the projections $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$. Figure 2.5 shows a schematic visualization of the coupling constraints for different class of problem (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 non-empty (for instance $\alpha \otimes \beta \in \mathcal{U}(\alpha, \beta)$). Figure 2.6 shows examples of discrete and continuous optimal coupling solving (2.15).

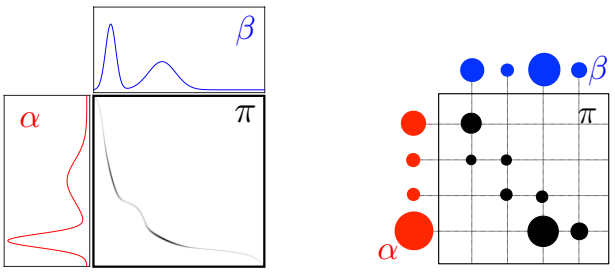


Figure 2.6: Left: "continuous" coupling π solving (2.14) between two 1-D measure with density. The coupling is localized along the graph of the Monge map $(x, T(x))$ (displayed in black). Right: "discrete" coupling T solving (2.11) between two discrete measures of the form (2.3). The non-zero entries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

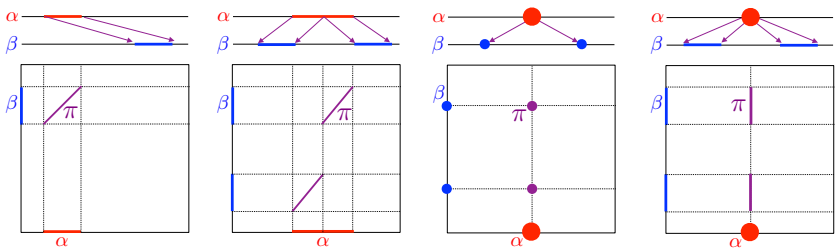


Figure 2.7: Four simple examples of optimal couplings between 1-D distributions, represented as maps above (arrows) and couplings below. Inspired by Levy and Schwindt [2017].

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

Remark 2.14 (Probabilistic interpretation)**.** Kantorovitch's problem can be re-interpreted through the prism of random variables, following Remark 2.9. Indeed, problem (2.15) is equivalent to

$$\min_{(X,Y)} \left\{ \mathbb{E}_{(X,Y)}(c(X,Y)) \;:\; X \sim \alpha, Y \sim \beta \right\}$$

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 introduce 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 meaningful 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

$$\mathrm{W}_p(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \mathrm{L}_{\mathbf{D}^p}(\mathbf{a}, \mathbf{b})^{1/p} \tag{2.16}$$

(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 non-zero element outside the diagonal); by symmetry of $\mathbf{D}^p$, $\mathrm{W}_p(\mathbf{a},\mathbf{b}) = 0$ is itself a symmetric function.

To prove the triangle inequality of Wasserstein distances for arbitrary measures, 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. We define $\bar{\mathbf{b}}_j \overset{\text{def.}}{=} \mathbf{b}_j$ if $\mathbf{b}_j > 0$ and set otherwise $\bar{\mathbf{b}}_j = 1$ (or actually any other value). We then define

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

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

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

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

The triangle inequality follows from

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

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

□

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 to a metric between histograms makes it a method of choice for applications in computer vision and machine learning where one needs 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]. In 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 approximations relying on simplifications of matrix $\mathbf{C}$ such as thresholding [Pele and Werman, 2008, 2009]. More recent applications stress the use of the EMD 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 multi-class classifiers that output bags-of-words Frogner et al. [2015] The review paper Kolouri et al. [2017] present an overview of other applications in signal processing and machine learning.

Remark 2.17 (Wasserstein distance between measures)**.**
Proposition 2.2 generalizes from histogram to 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 $x = y$;

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

Then

$$\mathcal{W}_p(\alpha,\beta) \overset{\text{def.}}{=} \mathcal{L}_{d^p}(\alpha,\beta)^{1/p} \tag{2.17}$$

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

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

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

Definition 2.2 (Weak convergence). $(\alpha_k)_k$ converges weakly to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_k \rightharpoonup \alpha$) if and only if for any continuous function $g \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} g \mathrm{d}\alpha_k \to \int_{\mathcal{X}} g \mathrm{d}\alpha$. This notion of weak convergence corresponds to the convergence in law of random vectors.

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

Remark 2.19 (Translation invariance). A nice feature of the Wasserstein distance over an 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, 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

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

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

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

$$\mathcal{W}_\infty(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \sup_{(x,y)\in \mathrm{Supp}(\pi)} d(x, y), \tag{2.18}$$

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 non-convex 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 Section 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. One has

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \max_{(\mathbf{f},\mathbf{g})\in\mathbf{R}(\mathbf{a},\mathbf{b})} \langle \mathbf{f},\, \mathbf{a}\rangle + \langle \mathbf{g},\, \mathbf{b}\rangle \tag{2.19}$$

where the set of admissible potentials is

$$\mathbf{R}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m \;:\; \forall\, (i,j) \in [\![n]\!] \times [\![m]\!], \mathbf{f} \oplus \mathbf{g} \leq \mathbf{C}\} \tag{2.20}$$

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 that result, namely that the right-hand side of Equation (2.19) is a lower bound on $\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ is discussed in 3.2. For the sake of completeness, let us derive this dual problem with the use of Lagrangian duality. The Lagangian associate to (2.11) reads

$$\min_{\mathbf{P}\geq 0} \max_{(\mathbf{f},\mathbf{g})\in\mathbb{R}^n\times\mathbb{R}^m} \langle \mathbf{C},\, \mathbf{P}\rangle + \langle \mathbf{a} - \mathbf{P}\mathbb{1}_m,\, \mathbf{f}\rangle + \langle \mathbf{b} - \mathbf{P}^\top \mathbb{1}_n,\, \mathbf{g}\rangle. \tag{2.21}$$

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

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

We conclude by remarking that

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

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

The primal-dual optimality relation for the Lagrangian (2.21) allows to locate the support of the optimal transport plan

$$\mathrm{Supp}(\mathbf{P}) \subset \Big\{(i,j) \in [\![n]\!] \times [\![m]\!] \;:\; \mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}\Big\}. \tag{2.22}$$

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) to 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 the following pricing scheme: she will ask money both when collecting a single unit of the resource at a warehouse, and later ask a bit more when delivering that unit of resource to a factory. More precisely, the vendor 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.

Agreement on 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 a large 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.

Of course, the vendor wishes to charge as much as they can for that service. 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 least smaller 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 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 transport, 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 applying these $n \times m$ inequalities that for any transport plan $\mathbf{P}$ (including the optimal one $\mathbf{P}^\star$), the marginal constraints imply

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

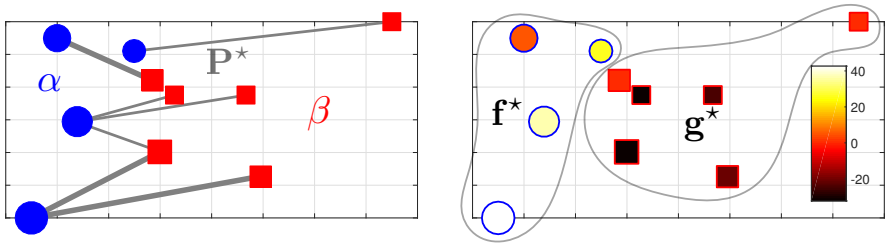


Figure 2.8: Consider in the left plot the optimal transport problem between two discrete measures α and β, represented respectively by blue dots and red squares. The area of these markers is proportional to the weight at each location. That plot also displays the optimal transport $\mathbf{P}^\star$ 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 away a mass at each location in α, or on the contrary to send a mass towards 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.

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

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 (2.19). 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 this problem. On the left, blue dots represent warehouses and red dots stand for factories; the areas of these dots stand for the probability weights $\mathbf{a}, \mathbf{b}$. On the right are pictured the price values obtained by the vendor as a result of optimizing problem 2.19. Prices have been chosen so that their mean is equal to 0. One can clearly see that highest relative prices come from collecting

goods at an isolated warehouse on the lower left of the figure, as well as 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 only be accessed 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

$$\mathcal{L}_c(\alpha,\beta) = \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y), \tag{2.23}$$

where the set of admissible dual potentials is

$$\mathcal{R}(c) \stackrel{\text{def.}}{=} \{(f,g)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y}) \;:\; \forall (x,y), f(x)+g(y)\leq c(x,y)\}\,. \tag{2.24}$$

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

The discrete case (2.19) corresponds to the dual vectors being samples of the continuous potentials, *i.e.* $(\mathbf{f}_i,\mathbf{g}_j) = (f(x_i),g(y_j))$. The primal-dual optimality conditions allow to track the support of optimal plan, and (2.22) is generalized as

$$\mathrm{Supp}(\pi) \subset \{(x,y)\in\mathcal{X}\times\mathcal{Y} \;:\; f(x)+g(y)=c(x,y)\}\,. \tag{2.25}$$

Note that in contrast to the primal problem (2.15), showing the existence of solutions to (2.23) is non-trivial, because the constraint set $\mathcal{R}(c)$ is not compact and the function to minimize non-coercive. Using the machinery of c-transform detailed in Section 5.1, one can however show that optimal (f,g) are necessarily Lipschitz regular, which enable to replace the constraint by a compact one.

Remark 2.23 (Monge-Kantorovich Equivalence – Brenier Theorem)**.** The following celebrated theorem of Brenier [1991] ensures that in $\mathbb{R}^d$ for $p = 2$, if at least one of the two inputs measures has a density, then Kantorovitch and Monge problems are equivalent.

Theorem 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 inputs 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.*

$$\forall\, h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x, y)\mathrm{d}\pi(x, y) = \int_{\mathcal{X}} h(x, T(x))\mathrm{d}\alpha(x). \tag{2.26}$$

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 (2.23) 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 following problem

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

whose dual reads

$$\min_{(\varphi,\psi)} \left\{ \int_{\mathcal{X}} \varphi\mathrm{d}\alpha + \int_{\mathcal{Y}} \psi\mathrm{d}\beta \;:\; \forall (x, y), \quad \varphi(x) + \psi(y) \geq \langle x,\, y\rangle \right\}. \tag{2.27}$$

The relation between these variables and those of (2.24) is $(\varphi, \psi) =$

$(\frac{\|\cdot\|^2}{2} - f, \frac{\|\cdot\|^2}{2} - g)$. One can replace the constraint by

$$\forall\, y, \quad \psi(y) \geq \varphi^*(y) \stackrel{\text{def.}}{=} \sup_x \langle x, y\rangle - \varphi(x). \tag{2.28}$$

Here φ^* is the Legendre transform of φ and is a convex function as a supremum of linear forms (see also (4.50)). Since the objective appearing in (2.29) is linear and the integrating measures positive, one can minimize explicitly with respect to φ and set $\psi = \varphi^*$ in order to consider the unconstraint problem

$$\min_{\varphi} \int_{\mathcal{X}} \varphi \mathrm{d}\alpha + \int_{\mathcal{Y}} \varphi^* \mathrm{d}\beta, \tag{2.29}$$

see also Section 5.1 for a generalization of this idea to generic costs $c(x, y)$. By iterating this argument twice, one can replace φ by φ^{**}, which is a convex function, and thus impose in (2.29) that φ is convex. Condition (2.25) shows that an optimal π is supported on $\{(x, y) \;:\; \varphi(x) + \varphi^*(y) = \langle x, y\rangle\}$ which shows that such an y is optimal for the minimization (2.28) of the Legendre transform, whose optimality condition reads $y \in \partial\varphi(x)$. Since φ is convex, it is differentiable almost everywhere, and since α has a density, it is also differentiable α-almost everywhere. This shows that for each x, the associated y is uniquely defined α-almost everywhere as $y = \nabla\varphi(x)$, and shows that necessarily $\pi = (\mathrm{Id}, \nabla\varphi)_\sharp \alpha$. □

This results shows that in the setting of $\mathcal{W}_2$ with non-singular densities, the Monge problem (2.9) and its Kantorovich relaxation (2.15) are equal (the relaxation is tight). This is the continuous analog of Proposition 2.1 for the assignment case (2.1), which states that the minimum of the optimal transport problem is achieved, when the marginals are equal and uniform, at a permutation matrix (a discrete map). Brenier's theorem, stating that an optimal transport map must be the gradient of a convex function, should be examined under the light that a convex function is the natural generalization of the notion of increasing functions in dimension more than one. Optimal transport can thus plays

an important role to define quantile functions in arbitrary dimensions, which in turn is useful for applications to quantile regression problems 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 2.24 (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

$$\det(\partial^2\varphi(x))\rho_\beta(\nabla\varphi(x)) = \rho_\alpha(x) \tag{2.30}$$

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

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

The convexity constraint forces $\det(\partial^2\varphi(x)) \geq 0$ 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 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 viscosity solution to capture singularity, and even 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 $\det(\varphi''(x)) = f(x)$ 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 a special care, and usual finite differences can be inaccurate. For optimal transport, where f actually depends on $\nabla\varphi$, the discretization of the equation (2.30), 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 points 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 Sections 3,4,7 3 different numerical solvers, we first review special favorable cases where the resolution of the OT problem is easy.

Remark 2.25 (Binary Cost Matrix and 1-Norm)**.** One can easily check that when the cost matrix $\mathbf{C}$ is zero on the diagonal and 1 elsewhere, namely when $\mathbf{C} = \mathbb{1}_{n\times n} - I_n$, the OT distance between $\mathbf{a}$ and $\mathbf{b}$ is equal to the 1-norm of their difference, $\mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\|_1$.

Remark 2.26 (Kronecker Cost Function and Total Variation)**.** In addition to Remark 2.25 above, one can also easily check that this result extends to discrete and discrete measures in the case where $c(x,y)$ is 0 if $x = y$ and 1 when $x \neq y$. The OT distance between two discrete measures α and β is equal to their total variation distance.

Remark 2.27 (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.* $x_1 \leq x_2 \leq \ldots \leq x_n$ and $y_1 \leq y_2 \leq \ldots \leq y_n$, then one has the simple formula

$$\mathcal{W}_p(\alpha,\beta)^p = \sum_{i=1}^p |x_i - y_i|^p, \tag{2.31}$$

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

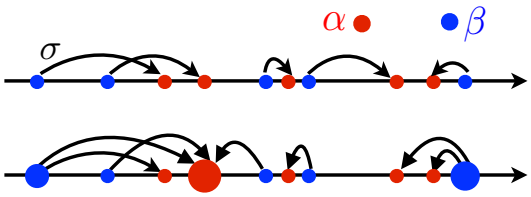


Figure 2.9: 1-D optimal couplings: each arrow $x_i \to y_j$ indicate a non-zero $\mathbf{P}_{i,j}$ in the optimal coupling. Top: empirical measures with same number of points (optimal matching). Bottom: generic case. This corresponds to monotone rearrangements, if $x_i \leq x_{i'}$ are such that $\mathbf{P}_{i,j}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

is only valid locally, in the sense that the order (and those vector representations) might change whenever some of the values change. That formula is a simple consequence of the more general remark given below. Figure 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 possible to leverage this 1-D computation to also compute efficiently OT on the circle, see Delon et al. [2010]. Note that in the case of concave cost of the distance, for instance when $p < 1$, the behaviour of the optimal transport plan is very different, see Delon et al. [2012], which describes an efficient solver in this case.

Remark 2.28 (1-D case – Generic case)**.** For a measure α on $\mathbb{R}$, we introduce the cumulative function

$$\forall\, x \in \mathbb{R}, \quad \mathcal{C}_\alpha(x) \stackrel{\text{def.}}{=} \int_{-\infty}^{x} \mathrm{d}\alpha, \tag{2.32}$$

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

$$\forall\, r \in [0,1], \quad \mathcal{C}_\alpha^{-1}(r) = \min_x \ \{x \in \mathbb{R} \cup \{-\infty\} \ : \ \mathcal{C}_\alpha(x) \geq r\}\,. \tag{2.33}$$

That function is also called the generalized quantile function of α.

For any $p \geq 1$, one has

$$\mathcal{W}_p(\alpha,\beta)^p = \left\|\mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1}\right\|_{L^p([0,1])}^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(r) - \mathcal{C}_\beta^{-1}(r)|^p \mathrm{d}r. \tag{2.34}$$

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

$$\mathcal{W}_1(\alpha,\beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| \mathrm{d}x \tag{2.35}$$

$$= \int_{\mathbb{R}} \left| \int_{-\infty}^x \mathrm{d}(\alpha - \beta) \right| \mathrm{d}x. \tag{2.36}$$

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

$$T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha. \tag{2.37}$$

Figure 2.10 illustrates the computation of 1-D OT through cumulative functions. It also displays displacement interpolations, computed as detailed in (7.7), see also Remark 9.5. For a detailed survey of the properties of optimal transport in 1-D, we refer the reader to [Santambrogio, 2015, Chapter 2].

Remark 2.29 (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

$$T : x \mapsto \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha), \tag{2.38}$$

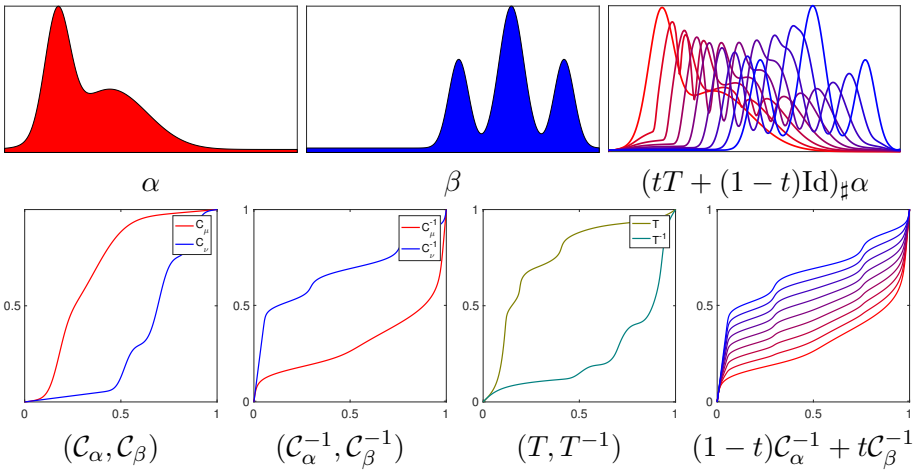


Figure 2.10: Computation of OT and displacement interpolation between two 1-D measures, using cumulant function as detailed in (2.37).

where

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

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 2.23) that T is optimal. Both that map

T and the corresponding potential ψ are illustrated in Figures 2.11 and 2.12

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

$$\mathcal{W}_2^2(\alpha,\beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \tag{2.39}$$

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

$$\mathcal{B}(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta)^2 \overset{\text{def.}}{=} \operatorname{tr}\left(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - 2(\mathbf{\Sigma}_\alpha^{1/2}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{1/2})^{1/2}\right), \tag{2.40}$$

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

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

For 1-D Gaussians, $\mathcal{W}_2$ is thus the Euclidean distance on the 2-D plane $(\mathbf{m}, \sqrt{\mathbf{\Sigma}})$, as illustrated in Figure 2.13. For a detailed treatment of the Wasserstein geometry of Gaussian distributions, we refer to Takatsu [2011].

Remark 2.30 (Distance between Elliptically Contoured Distributions)**.** Gelbrich provides a more general result than that provided in Remark 2.29: 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 (Theorem 2.1 in [1990]). Additionally, the closed form (2.39) 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 as

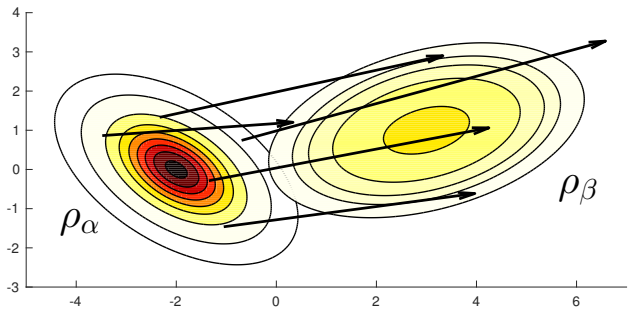


Figure 2.11: Two Gaussians ρ_α and ρ_β, represented using the contour plots of their densities, with respective mean and variance matrices $\mathbf{m}_\alpha = (-2, 0), \mathbf{\Sigma}_\alpha = \frac{1}{2}\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$ and $\mathbf{m}_\beta = (3, 1), \mathbf{\Sigma}_\beta = \begin{pmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$. The arrows originate at random points x taken on the plane and end at the corresponding mappings of those points $T(x) = \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$.

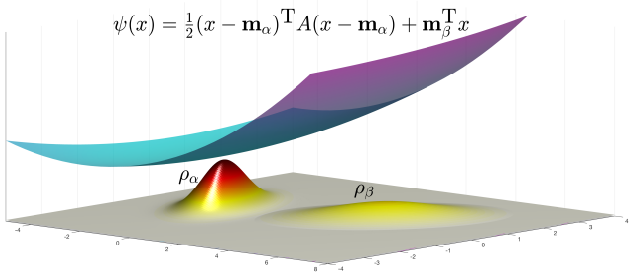


Figure 2.12: Same Gaussians ρ_α and ρ_β as defined in Figure 2.11, represented this time as surfaces. The surface above is the Brenier potential ψ defined up to an additive constant (here $+50$) such that $T = \nabla\psi$. For visual purposes, both Gaussian densities have been multiplied by a 100 factor.

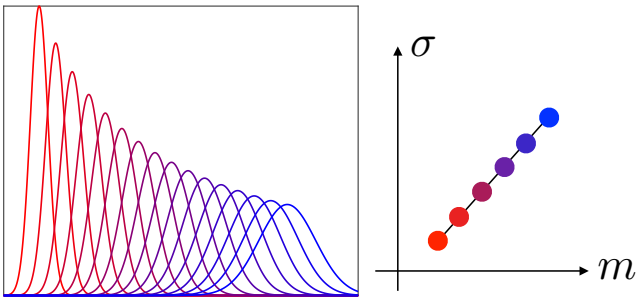


Figure 2.13: Computation of displacement interpolation between two 1-D Gaussians. Denoting $\mathcal{G}_{m,\sigma}(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-m)^2}{2s^2}}$ the Gaussian density, it thus shows the interpolation $\mathcal{G}_{(1-t)m_0+tm_1,(1-t)\sigma_0+t\sigma_1}$.

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

respectively, for the same positive 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 (2.38) and their Wasserstein distance is also given by the expression (2.39). This includes therefore as interesting special cases uniform distributions on ellipses, namely ellipctic shapes.

3

Algorithmic Foundations

This chapter describes algorithmic tools from combinatorial optimization that can be used to solve optimal transport numerically. These tools can only be used on the discrete formulation of optimal transport, as described in the primal problem (2.11) or alternatively its dual (2.19).

The origins of these algorithms can be traced back to the period immediately before [Tolstoı, 1930] and during world war 2, when Hitchcock [1941] and Kantorovich [1942] formalized the generic problem of dispatching available resources towards 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 the problem they had helped define (although the cycle violation method was already hinted 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 optimal transport problems.

The goal of linear programming is to solve optimization problems whose objective 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 however argue 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 stayed 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, Lemma 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 more generally optimization to a new audience [Nocedal and Wright, 1999, §1,p.4] because of its intuitive description.

3.1 The Kantorovich Linear Programs

We have already introduced in Equation (2.11) the primal optimal transport 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; nonnegative constraints on variables. Let $\mathbb{I}_n$ stand for the identity matrix of size n and $\otimes$ be Kronecker's product. The following $(n+m) \times nm$ matrix,

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

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 column-wise) to obtain

the following equivalence:

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

Therefore we can write the original optimal transport problem as:

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

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 our original (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 the rules of 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. We refer the interested reader to [Bertsimas and Tsitsiklis, 1997, §4.10]. To simplify notations, let us 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 $A\mathbf{p} = \mathbf{q}$ is no longer

necessarily enforced, but bears a cost $\mathbf{h}^{\mathrm{T}}(A\mathbf{p}-\mathbf{q})$ 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:

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

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

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

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 that $\underline{z} \leq \bar{z}$, a result usually known as weak duality.

3.2 C-transforms

We present in this section an interesting property of the dual optimal transport problem (3.3) which takes a more important meaning when

used for the semi-discrete optimal transport problem in §5.1. This section builds upon the original formulation (2.19) that splits row and column sum constraints:

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

Consider any dual feasible pair $(\mathbf{f},\mathbf{g})$. If we "freeze" the value of $\mathbf{f}$, we can notice that there is not 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 $(\mathbf{f},\mathbf{f}^{\mathbf{C}}) \in \mathbf{R}(\mathbf{a},\mathbf{b})$ and that $\mathbf{f}^{\mathbf{C}}$ is the largest possible vector such that this constraint is insured. 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 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 value of $\mathbf{g}$ and consider instead the $\bar{\mathbf{C}}$-transform of $\mathbf{g}$, namely vector $\mathbf{g}^{\bar{\mathbf{C}}} \in \mathbb{R}^m$ defined as

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

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

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

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

Proposition 3.1. The following identities 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), (2.19) 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 (2.23) 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 (2.20), each entry of the matrix $\mathbf{C} - \mathbf{f}^\star \oplus \mathbf{g}^\star$ is necessarily non-negative. 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 $\mathbf{P}_{i,j} = 0$. □

3.4 Vertices of the Transportation Polytope

A linear program with a non-empty and bounded feasible set attains its minimum at an 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. Consider their union $V \cup V'$, with $n + m$ nodes, and the set $\mathcal{E}$ of all nm undirected edges $\{\{i, j'\}, i \in [\![n]\!], j \in [\![m]\!]\}$ between them. 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 and sink 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].

Proposition 3.3 (Extremal Solutions)**.** Let $\mathbf{P}$ be an extremal point of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$. Let $F(\mathbf{P}) \subset \mathcal{E}$ be the subset of undirected edges $\{\{i, j'\}, i \in [\![n]\!], j \in [\![m]\!]$ such that $\mathbf{P}_{ij} > 0\}$. Then the graph $G(\mathbf{P}) \stackrel{\text{def.}}{=} (V \cup V', F(\mathbf{P}))$ has no cycles. In particular, $\mathbf{P}$ cannot have more than $n + m - 1$ non-zero entries.

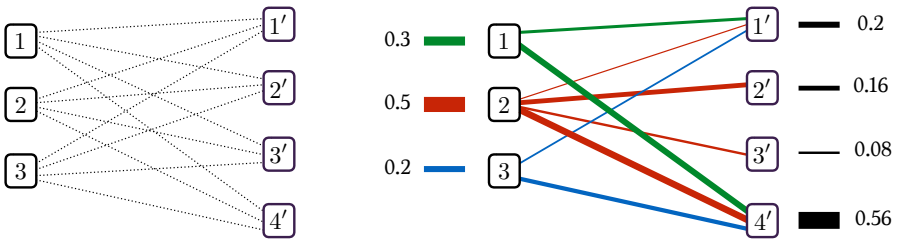


Figure 3.1: The optimal transport problem as a bipartite network flow problem. Here $n = 3, m = 4$. All coordinates of the source histogram, $\mathbf{a}$, are depicted as source nodes on the left labeled $1, 2, 3$ whereas all coordinates of the target histogram $\mathbf{b}$ are labeled as nodes $1', 2', 3', 4'$. The graph is bipartite in the sense that all source nodes are connected to all target nodes, with no additional edges. Each edge $\{i, j'\}$ is attributed a cost corresponding to the entry $\mathbf{C}_{ij}$. A feasible flow is represented on the right. Notice that, however, in light of the result provided in Proposition 3.3, that flow is not extremal since it has at least one cycle given by $((1, 1'), (1', 2), (2, 4'), (4', 1))$.

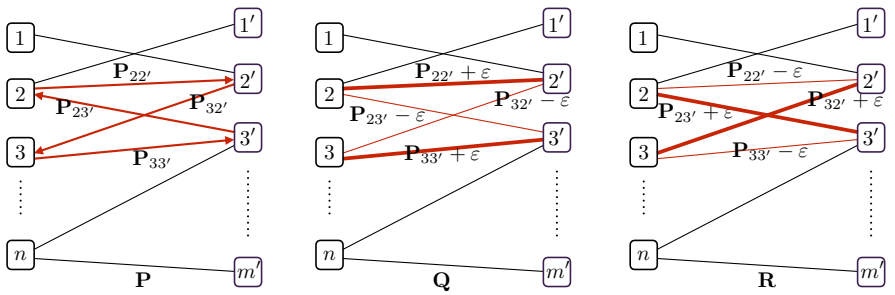


Figure 3.2: A solution $\mathbf{P}$ with a cycle in the graph of its support can be perturbed to obtain two feasible solutions $\mathbf{Q}$ and $\mathbf{R}$ such that $\mathbf{P}$ is their average, therefore disproving that $\mathbf{P}$ is extremal.

Proof. Suppose that $\mathbf{P}$ is an extremal point of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$ and that its corresponding set $F(\mathbf{P})$ of 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 $H = \{\{i_1, j'_1\}, \{j'_1, i_2\}, \{i_2, j'_2\}, \dots, \{i_k, j'_k\}, \{j'_k, i_1\}\}$ forms a subset of F. We construct two feasible matrices $\mathbf{Q}$ and $\mathbf{R}$ such that $\mathbf{P} = (\mathbf{Q}+\mathbf{R})/2$. To do so, we 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 $\varepsilon < \min_{\{i,j'\} \in F} \mathbf{P}_{ij}$. We now form a perturbation E matrix 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. We now define two matrices $\mathbf{Q} = \mathbf{P} + E$ and $\mathbf{R} = \mathbf{P} - E$. Because ε small enough, all elements in $\mathbf{Q}$ and $\mathbf{R}$ are nonnegative. By construction, 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, E is such that $E\mathbb{1}_m = \mathbb{0}_n$ and $E^{\mathrm{T}}\mathbb{1}_n = \mathbb{0}_m$, and we have that $\mathbf{Q}$ and $\mathbf{R}$ have the same marginals as $\mathbf{P}$. 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 $F(\mathbf{P})$ cannot have more than $n + m - 1$ edges, and therefore $\mathbf{P}$ cannot have more than $n + m - 1$ non-zero 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 to initialize any algorithm, 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 row constraint at j, or both if possible. The counters i, j are then updated as follows: i is incremented in the first case, j is in the second, 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 North-West corner solutions that can be produced this way:

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

Note that all NW corner solutions only have by construction up to $n + m - 1$ nonzero elements. The NW corner rule produces a table which is by construction unique for $\mathbf{a}$ and $\mathbf{a}$, but there is an exponential

number of pairs or row/column permutations (σ, σ') that may share 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, Corollary 8.1.4].

3.5 A Heuristic Description of the Network Simplex

Consider a feasible matrix $\mathbf{P}$ whose graph $G(\mathbf{P}) = (V \cup V', F(\mathbf{P}))$ has no cycles. $\mathbf{P}$ has therefore no more than $n+m-1$ non-zero entries, and is a vertex of $\mathbf{U}(\mathbf{a}, \mathbf{b})$ by Proposition 3.3. We start this section with a simple way to check whether $\mathbf{P}$ is optimal. More precisely, we consider the optimality of a feasible *primal-dual* pair:

Proposition 3.4. Let $\mathbf{P}$ and $(\mathbf{f}, \mathbf{g})$ be feasible solutions for the primal (2.23) and dual (2.11) problems. If for any $\{i, j'\} \in F(\mathbf{P})$ one has that $\mathbf{f}_i + \mathbf{g}_j = \mathbf{C}_{i,j}$, 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. □

Given a feasible matrix $\mathbf{P}$, it is therefore sufficient to obtain a dual solution $(\mathbf{f}, \mathbf{g})$ which is feasible and complementary to $\mathbf{P}$, in the sense that $\mathbf{C} - \mathbf{f} \oplus \mathbf{g}$ has non-negative entries *and* pairs of indices $\{i, j'\}$ in $F(\mathbf{P})$ are such that $\mathbf{C}_{i,j} = \mathbf{f}_i + \mathbf{g}_j$, to prove that $\mathbf{P}$ is optimal. Exhibiting whether such a pair $(\mathbf{f}, \mathbf{g})$ exists, and, if not, modify $\mathbf{P}$ to reach that goal is the gist of the network simplex.

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 $F(\mathbf{P})$, $\mathbf{f}_i + \mathbf{g}_j$ is equal to $\mathbf{C}_{i,j}$. Note that this, in itself, does not guarantee that $(\mathbf{f}, \mathbf{g})$ is feasible, namely that $\mathbf{C} - \mathbf{f} \oplus \mathbf{g} \geq 0$.

Let s be the cardinal of $F(\mathbf{P})$. 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

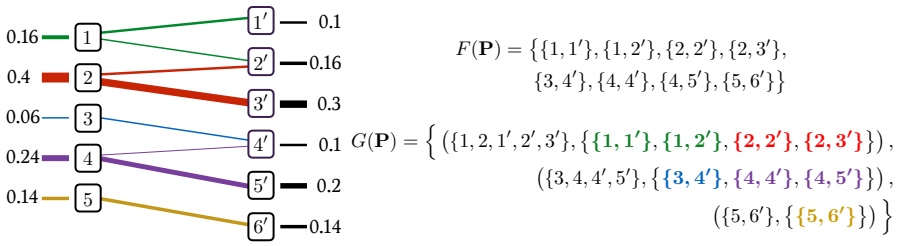


Figure 3.3: A feasible transport $\mathbf{P}$ and its corresponding set of edges $F(\mathbf{P})$ and graph $G(\mathbf{P})$. As can be seen in the picture above, the graph $G(\mathbf{P}) = (\{1,\dots,5,1',\dots,6'\}, F(\mathbf{P}))$ is a forest, meaning that it can be expressed as the union of tree graphs,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

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 $F(\mathbf{P})$ are enumerated as $\{i_1, j_1'\},\dots,\{i_s, j_s'\}$.

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 $i_1,\dots,i_k$ among source nodes and l nodes $j_1',\dots,j_l'$ among target nodes, resulting in $r \stackrel{\text{def.}}{=} k+l$, and $r-1$ edges, corresponding to

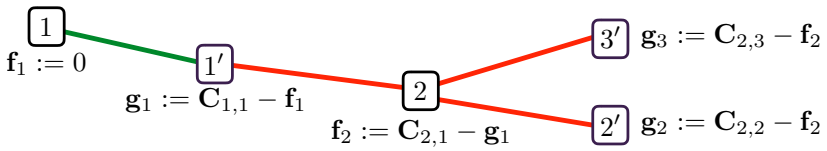


Figure 3.4: The 5 dual variables $\mathbf{f}_1, \mathbf{f}_2, \mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ corresponding to the 5 nodes appearing in the first tree of the graph $G(\mathbf{P})$ illustrated in Figure 3.3 are linked through 4 linear equations that involve corresponding entries in the cost 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 4 remaining dual variables.

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 then, we can traverse the tree using a breadth-first or depth-first search to obtain a sequence of simple variable assignments that determine 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 $\mathbf{P}$.

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 3.4. 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 carry next an iteration of the algorithm itself, which consists first in adding 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 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 has now 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 be all 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 3.3. 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 $\mathbf{P}^{\mathrm{n}}$, equal to $\mathbf{P}$ for all but the following entries:

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

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 as cycle, case *(a)* above, the primal solution remains unchanged. When a cycle is created, case *(b)*, $\mathbf{P}$ is updated to

$\mathbf{P}^{\mathrm{n}}$, and the following equality holds:

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

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

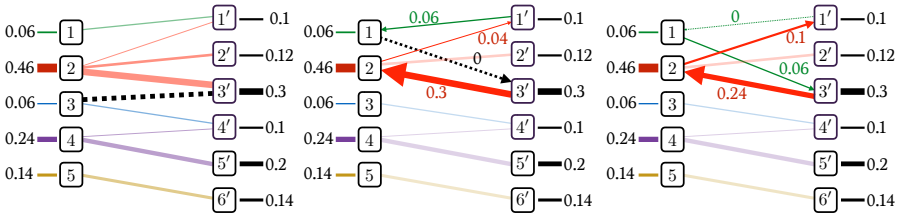
$$\begin{aligned} \sum_{k=1}^{l} \mathbf{C}_{i_k, j_k} - \sum_{k=1}^{l} \mathbf{C}_{i_{k+1}, j_k} &= \mathbf{C}_{i,j} + \sum_{k=2}^{l} \mathbf{f}_{i_k} + \mathbf{g}_{j_k} - \sum_{k=1}^{l} \mathbf{f}_{i_{k+1}} + \mathbf{g}_{j_k} \\ &= \mathbf{C}_{i,j} - (\mathbf{f}_i + \mathbf{g}_j). \end{aligned}$$

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

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

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 north-west 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 $\mathbf{C} - \mathbf{f}^{\star} \oplus \mathbf{g} \geq 0$; 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, build then a complementary pair $\mathbf{f}, \mathbf{g}$ accordingly; return 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].



(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 $\mathbf{P}$.

3.6 Matching problems

The network simplex is meant to be used in the general case where marginals $\mathbf{a}, \mathbf{b}$ have arbitrary values. It is, however, easy to notice that when both $\mathbf{a}$ and $\mathbf{b}$ are equal and both uniform, namely $\mathbf{a} = \mathbb{1}_n/n$, any extremal solution $\mathbf{P}$ is a permutation matrix through Birkhoff's theorem [1946], as pointed out in Proposition 2.1. In that case $G(\mathbf{P})$ is a disjoint set of n atoms, and many of the updates outlined above result in degenerate moves, making the network simplex approach above too general and inefficient to exploit this setting. We introduce two faster alternatives in that case.

3.6.1 Hungarian Algorithm

The Hungarian algorithm precedes the network simplex by quite a few decades, since it can be traced back to work by Jacobi Borchardt and Jocobi [1865] and later König and Egerváry, as recounted by Kuhn [1955]. The Hungarian algorithm is a particular case of a more general class of network flow optimization algorithms known as dual ascent methods, in the sense that they maintain a candidate of dual feasible solutions that is progressively improved. We follow the presentation of [Bertsimas and Tsitsiklis, 1997, §7.7] and start with the following

proposition

Proposition 3.5. A dual pair $\mathbf{f}, \mathbf{g}$ is either optimal for Problem (3.4) or there exists $\varepsilon > 0$ and $S \subset [\![n]\!], S' \subset [\![m]\!]$ such that $\mathbb{1}_S{}^{\mathrm{T}}\mathbf{a} - \mathbb{1}_{S'}{}^{\mathrm{T}}\mathbf{b} > 0$, where $\mathbb{1}_S$ is 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 S'.

3.6.2 Auction Algorithm

The auction algorithm is an alternative to the Hungarian algorithm for the assignment problem that was originally proposed by Bertsekas [1981] and later refined in [Bertsekas and Eckstein, 1988]. Several economic interpretations of this algorithm have been proposed (see *e.g.* Bertsekas [1992]).

Complementary Slackness. Notice that in the optimal assignemnt 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 (3.4)). 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 $\mathbf{f}^{\bar{\mathbf{C}}}$. 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 towards 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^1-th entry of $\mathbf{g}$ the sum of ε and the difference between the second lowest and lowest adjusted cost $\{\mathbf{C}_{i,j} - \mathbf{g}_j\}_j$:

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

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 3.6. 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 i-th 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 3.7. 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 3.8. 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 outputted 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 equality next 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. We explore next 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].

4

Entropic Regularization of Optimal Transport

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

4.1 Entropic Regularization

The discrete entropy of a coupling matrix is defined as

$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def.}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1), \tag{4.1}$$

with an analogous definition for vectors, with the convention that $\mathbf{H}(\mathbf{a}) = -\infty$ if one of the entries $\mathbf{a}_j$ is 0 or negative. The function $\mathbf{H}$ is 1-strongly concave, because its hessian is $\partial^2\mathbf{H}(P) = -\operatorname{diag}(1/\mathbf{P}_{i,j})$ and

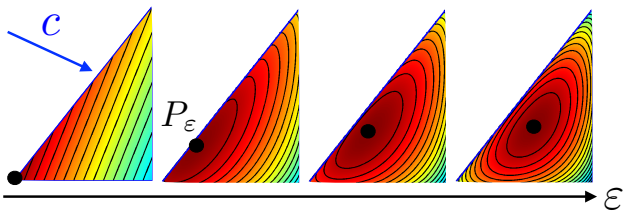


Figure 4.1: Impact of ε on the optimization of a linear function on the simplex, solving $\mathbf{P}_\varepsilon = \operatorname{argmin}_{\mathbf{P}\in\Sigma_3} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon \mathbf{H}(\mathbf{P})$ for a varying ε.

$\mathbf{P}_{i,j} \leq 1$. The idea of the entropic regularization of optimal transport is to use $-\mathbf{H}$ as a regularizing function to obtain approximate solutions to the original transport problem (2.11):

$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon \mathbf{H}(\mathbf{P}). \tag{4.2}$$

Since the objective is a ε-strongly convex function, problem 4.2 has a unique optimal solution. The idea to regularize the optimal transport problem by an entropic term can be traced back to modeling ideas in transportation theory [Wilson, 1969]: Actual traffic patterns in a network do not agree with those predicted by the solution of the optimal transport problem. Indeed, the former are more diffuse than the latter, which tend to rely on a few routes as a result of the sparsity of optimal couplings to the solution of 2.11. To balance for that, researchers in transportation proposed a model, called the "gravity" model [Erlander, 1980], that is able to form a more "blurred" traffic prediction.

Figure 4.1 illustrates the effect of the entropy to regularize a linear program over the simples Σ_3 (which can thus be visualized as a triangle in 2-D). Note how the entropy pushes the original LP solution away from the boundary of the triangle. The optimal $\mathbf{P}_\varepsilon$ progressively moves toward an "entropic center" of the triangle. This is further detailed in the proposition below. The convergence of the solution of that regularized problem towards an optimal solution of the original linear program has been studied by Cominetti and San Martín [1994].

Proposition 4.1 (Convergence with ε)**.** The unique solution $\mathbf{P}_\varepsilon$ of (4.2) converges to the optimal solution with maximal entropy within the set

of all optimal solutions of the Kantorovich problem, namely

$$\mathbf{P}_\varepsilon \overset{\varepsilon\to 0}{\longrightarrow} \operatorname*{argmin}_{\mathbf{P}} \{-\mathbf{H}(\mathbf{P}) \;:\; \mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b}), \langle \mathbf{P},\, \mathbf{C}\rangle = \mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b})\} \tag{4.3}$$

so that in particular

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

One has

$$\mathbf{P}_\varepsilon \overset{\varepsilon\to\infty}{\longrightarrow} \mathbf{a}\mathbf{b}^{\mathrm{T}} = (\mathbf{a}_i\mathbf{b}_j)_{i,j}. \tag{4.4}$$

Proof. We consider a sequence $(\varepsilon_\ell)_\ell$ such that $\varepsilon_\ell \to 0$ and $\varepsilon_\ell > 0$. We denote $\mathbf{P}_\ell$ the solution of (4.2) for $\varepsilon = \varepsilon_\ell$. Since $\mathbf{U}(\mathbf{a},\mathbf{b})$ is bounded, we can extract a sequence (that we do not relabel for sake of simplicity) such that $\mathbf{P}_\ell \to \mathbf{P}^\star$. Since $\mathbf{U}(\mathbf{a},\mathbf{b})$ is closed, $\mathbf{P}^\star \in \mathbf{U}(\mathbf{a},\mathbf{b})$. We consider any $\mathbf{P}$ such that $\langle \mathbf{C},\, \mathbf{P}\rangle = \mathrm{L}_{\mathbf{C}}(\mathbf{a},\mathbf{b})$. By optimality of $\mathbf{P}$ and $\mathbf{P}_\ell$ for their respective optimization problems (for $\varepsilon = 0$ and $\varepsilon = \varepsilon_\ell$), one has

$$0 \leq \langle \mathbf{C},\, \mathbf{P}_\ell\rangle - \langle \mathbf{C},\, \mathbf{P}\rangle \leq \varepsilon_\ell(\mathbf{H}(\mathbf{P}_\ell) - \mathbf{H}(\mathbf{P})). \tag{4.5}$$

Since $\mathbf{H}$ is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle \mathbf{C},\, \mathbf{P}^\star\rangle = \langle \mathbf{C},\, \mathbf{P}\rangle$ so that $\mathbf{P}^\star$ is a feasible point of (4.3). Furthermore, dividing by ε_ℓ in (4.5) and taking the limit shows that $\mathbf{H}(\mathbf{P}) \leq \mathbf{H}(\mathbf{P}^\star)$, which shows that $\mathbf{P}^\star$ is a solution of (4.3). Since the solution $\mathbf{P}^\star_0$ to this program is unique by strict convexity of $-\mathbf{H}$, one has $\mathbf{P}^\star = \mathbf{P}^\star_0$, and the whole sequence is converging. In the limit $\varepsilon \to +\infty$, a similar proof shows that one should rather consider the problem

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} -\mathbf{H}(\mathbf{P})$$

the solution of which is $\mathbf{a}\otimes\mathbf{b}$. □

Formula (4.3) states that for low regularization, the solution converges to the maximum entropy optimal transport coupling. In sharp contrast, (4.4) shows that for large regularization, the solution converges to the coupling with maximal entropy between two prescribed marginals $\mathbf{a},\mathbf{b}$, namely the joint probability between two independent random variables with prescribed distributions. A refined analysis of this convergence is performed in Cominetti and San Martín [1994],

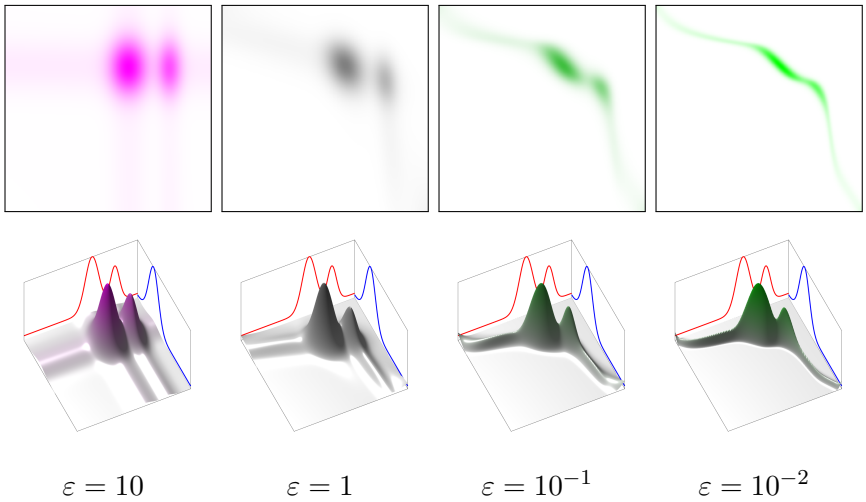


Figure 4.2: Impact of ε on the couplings between two 1-D densities, illustrating Proposition 4.1. Top row: between two 1-D densities. Bottom row: between two 2-D discrete empirical densities with same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

including a first order expansion in ε (resp. $1/\varepsilon$) near $\varepsilon = 0$ (resp $\varepsilon = +\infty$). Figures 4.2 and 4.3 shows visually the effect of these two convergence. A key insight is that, as ε increases, the optimal coupling becomes less and less sparse (in the sense of having entries larger than a prescribed thresholds), which in turn as the effect of both accelerating computational algorithms (as we study in §4.2) but also leading to faster statistical convergence (as exposed in §8.5).

Defining the Kullback-Leibler divergence between couplings as

$$\mathbf{KL}(\mathbf{P}|\mathbf{K}) \overset{\text{def.}}{=} \sum_{i,j} \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}}\right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}, \tag{4.6}$$

the unique solution $\mathbf{P}_\varepsilon$ of (4.2) is a projection onto $\mathbf{U}(\mathbf{a}, \mathbf{b})$ of the Gibbs kernel associated to the cost matrix $\mathbf{C}$ as

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

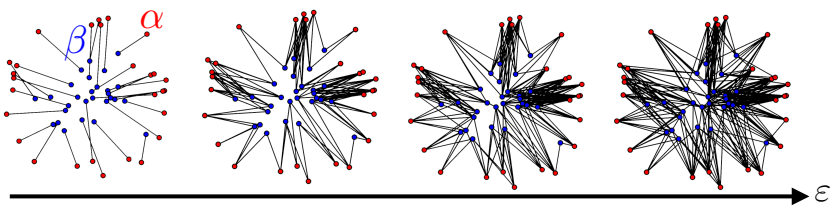


Figure 4.3: Impact of ε on coupling between two 2-D discrete empirical densities with same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

Indeed one has that using the definition above

$$\mathbf{P}_\varepsilon = \mathrm{Proj}^{\mathbf{KL}}_{\mathbf{U}(\mathbf{a},\mathbf{b})}(\mathbf{K}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathbf{KL}(\mathbf{P}|\mathbf{K}). \tag{4.7}$$

Remark 4.1 (Entropic regularization between discrete measures)**.** For discrete measures of the form (2.1), the definition of regularized transport extends naturally to

$$\mathcal{L}^\varepsilon_c(\alpha,\beta) \overset{\text{def.}}{=} \mathrm{L}^\varepsilon_{\mathbf{C}}(\mathbf{a},\mathbf{b}), \tag{4.8}$$

with cost $\mathbf{C}_{i,j} = c(x_i, y_j)$, to emphasize the dependency with respect to the positions (x_i, y_j) supporting the input measures.

Remark 4.2 (General formulation)**.** One can consider arbitrary measures by replacing the discrete entropy by the relative entropy with respect to the product measure $\mathrm{d}\alpha\otimes\mathrm{d}\beta(x,y) \overset{\text{def.}}{=} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$, and propose a regularized counterpart to (2.15) using

$$\mathcal{L}^\varepsilon_c(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{X\times Y} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) \tag{4.9}$$

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence (4.6)

$$\begin{aligned}\mathrm{KL}(\pi|\xi) \overset{\text{def.}}{=} &\int_{\mathcal{X}\times\mathcal{Y}} \log\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\Big)\mathrm{d}\pi(x,y) + \\ &\int_{\mathcal{X}\times\mathcal{Y}} (\mathrm{d}\xi(x,y) - \mathrm{d}\pi(x,y)),\end{aligned} \tag{4.10}$$

and by convention $\mathrm{KL}(\pi|\xi) = +\infty$ if π does not have a density $\frac{\mathrm{d}\pi}{\mathrm{d}\xi}$ with respect to ξ. It is important to realize that the reference measure $\alpha \otimes \beta$ chosen in (4.9) to define the entropic regularizing term $\mathrm{KL}(\cdot|\alpha \otimes \beta)$ plays no specific role, only its support matters, as noted by the following proposition.

Proposition 4.2. For any $\pi \in \mathcal{U}(\alpha, \beta)$, and for any (α', β') with the same support as (α, β) (so that they have both densities with respect to one another) one has

$$\mathrm{KL}(\pi|\alpha \otimes \beta) = \mathrm{KL}(\pi|\alpha' \otimes \beta') + \mathrm{KL}(\alpha' \otimes \beta'|\alpha \otimes \beta).$$

This proposition shows that choosing $\mathrm{KL}(\cdot|\alpha' \otimes \beta')$ in place of $\mathrm{KL}(\cdot|\alpha \otimes \beta)$ in (4.9) results in the same solution.

Formula (4.9) can be re-factored as a projection problem

$$\min_{\pi \in \mathcal{U}(\alpha,\beta)} \mathrm{KL}(\pi|\mathcal{K}) \tag{4.11}$$

where $\mathcal{K}$ is the Gibbs distributions $\mathrm{d}\mathcal{K}(x, y) \overset{\text{def.}}{=} e^{-\frac{\mathrm{c}(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y)$. This problem is often referred to as the "static Schrödinger problem" [Léonard, 2014, Rüschendorf and Thomsen, 1998], since it was initially considered by Schrödinger in statistical physics [Schrödinger, 1931]. As $\varepsilon \to 0$, the unique solution to (4.11) converges to the maximum entropy solution to (2.15), see [Léonard, 2012, Carlier et al., 2017]. §7.6 details an alternate "dynamic" formulation of the Schrödinger problem over the space of paths connecting the points of two measures.

Remark 4.3 (Independence and couplings)**.** A coupling $\pi \in \mathcal{U}(\alpha, \beta)$ describes the distribution of a couple of random variables (X, Y) defined on $(\mathcal{X}, \mathcal{Y})$, where X (resp. Y) has law α (resp. β). Proposition 4.1 caries over for generic (non-necessary discrete) measures, so that the solution π_ε of (4.9) converges to the tensor product coupling $\alpha \otimes \beta$ as $\varepsilon \to +\infty$. This coupling $\alpha \otimes \beta$ corresponds to the random variables (X, Y) being independent. In contrast, as $\varepsilon \to 0$, π_ε convergence to a solution

π_0 of the OT problem (2.15). On $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, if α and β have densities with respect to the Lebesgue measure, as detailed in Remark 2.23, then π_0 is unique and supported on the graph of a bijective Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$. In this case, (X, Y) are in some sense fully dependent, since $Y = T(X)$ and $X = T^{-1}(Y)$. In the simple 1-D case $d = 1$, a convenient way to vizualize the dependency structure between X and Y is to use the copula ξ_π associated to the joint distribution π. The cumulative function defined in (2.32) is extended to couplings as

$$\forall\,(x,y) \in \mathbb{R}^2, \quad \mathcal{C}_\pi(x,y) \overset{\text{def.}}{=} \int_{-\infty}^{x}\int_{-\infty}^{y} \mathrm{d}\pi.$$

The copula is then defined as

$$\forall\,(s,t) \in [0,1]^2, \quad \xi_\pi(s,t) \overset{\text{def.}}{=} \mathcal{C}_\pi(\mathcal{C}_\alpha^{-1}(s), \mathcal{C}_\beta^{-1}(t)).$$

where the pseudo-inverse of a cumulative function is defined in (2.33). For independent variables, $\varepsilon = +\infty$, i.e. $\pi = \alpha \otimes \beta$, one has $\xi_{\pi_{+\infty}}(s,t) = st$. In contrast, for fully dependent variables, $\varepsilon = +\infty$, one has $\xi_{\pi_0}(s,t) = \min(s,t)$. Figure 4.4 shows how entropic regularization generates copula ξ_{π_ε} interpolating between these two extreme cases.

4.2 Sinkhorn's Algorithm and its Convergence

The following proposition shows that the solution of (4.2) has a specific form, which can be parameterized using $n + m$ variables. That parameterization is therefore essentially dual, in the sense that a coupling $\mathbf{P}$ in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ has nm variables but $n + m$ constraints.

Proposition 4.3. The solution to (4.2) is unique and has the form

$$\forall\,(i,j) \in [\![n]\!] \times [\![m]\!], \quad \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \tag{4.12}$$

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

Proof. Introducing two dual variables $\mathbf{f} \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^m$ for each marginal constraint, the Lagrangian of (4.2) reads

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

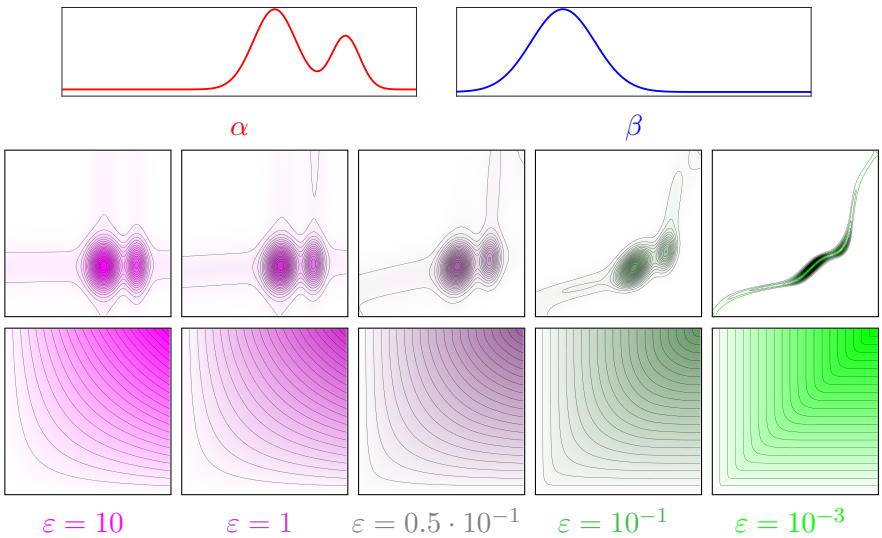


Figure 4.4: Top: evolution with ε of the solution π_ε of (4.9). Bottom: evolution of the copula function ξ_{π_ε}.

Considering first order conditions, we have

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

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

Regularized OT as Matrix Scaling The factorization of the optimal solution exhibited in Equation (4.12) can be conveniently rewritten in matrix form as $\mathbf{P} = \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$. $\mathbf{u}, \mathbf{v}$ must therefore satisfy the following non-linear equations which correspond to the mass conservation constraints inherent to $\mathbf{U}(\mathbf{a}, \mathbf{b})$,

$$\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})\mathbb{1}_m = \mathbf{a}, \quad \text{and} \quad \operatorname{diag}(\mathbf{v})\mathbf{K}^\top \operatorname{diag}(\mathbf{u})\mathbb{1}_n = \mathbf{b}, \quad (4.13)$$

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

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \quad \text{and} \quad \mathbf{v} \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u}) = \mathbf{b} \quad (4.14)$$

where $\odot$ corresponds to entry-wise multiplication of vectors. That problem is known in the numerical analysis community as the matrix scaling problem (see [Nemirovski and Rothblum, 1999] and references therein). An intuitive way to try to solve these equations is to solve them iteratively, by modifying first $\mathbf{u}$ so that it satisfies the left-hand side of Equation (4.14) and then $\mathbf{v}$ to satisfy its right-hand side. These two updates define Sinkhorn's algorithm:

$$\mathbf{u}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \text{and} \quad \mathbf{v}^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(\ell+1)}}, \tag{4.15}$$

initialized with an arbitrary positive vector $\mathbf{v}^{(0)} = \mathbb{1}_m$. The division operator used above between two vectors is to be understood entry-wise. Note that a different initialization will likely lead to a different solution for $\mathbf{u}, \mathbf{v}$, since $\mathbf{u}, \mathbf{v}$ are only defined up to a multiplicative constant (if $\mathbf{u}, \mathbf{v}$ satisfy (4.13) then so do $\lambda\mathbf{u}, \mathbf{v}/\lambda$ for any $\lambda > 0$). It turns out however that these iterations converge (see Remark 4.7 for a justification using iterative projections, and Remark 4.12 for a strict contraction result) and all result in the same optimal coupling $\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$. Figure 4.5, top row, shows the evolution of the coupling $\operatorname{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\operatorname{diag}(\mathbf{v}^{(\ell)})$ computed by Sinkhorn iterations. It evolves from the Gibbs kernel $\mathbf{K}$ towards the optimal coupling solving (4.2) by progressively shifting the mass away from the diagonal.

Remark 4.4 (Historical Perspective)**.** This algorithm was originally introduced with a proof of convergence by Sinkhorn [1964] with later contributions [Sinkhorn and Knopp, 1967, Sinkhorn, 1967]. It was used earlier as a heuristic to scale a matrix so that it fits desired marginals (typically uniform) under the name of Iterative proportional fitting (IPFP) Deming and Stephan [1940] and RAS Bacharach [1965] methods [Idel, 2016], and later extended in infinite dimensions by Ruschendorf [1995]. It was adopted very early as well in the field of economics, precisely to obtain approximate solutions to optimal transport problems, under the name of gravity models [Wilson, 1969, Erlander, 1980, Erlander and Stewart, 1990]. It was rebranded as "softassign" by Kosowsky and Yuille [1994] in the assignment case, namely when $\mathbf{a} = \mathbf{b} = \mathbb{1}_n/n$, and used to solve matching problems in economics

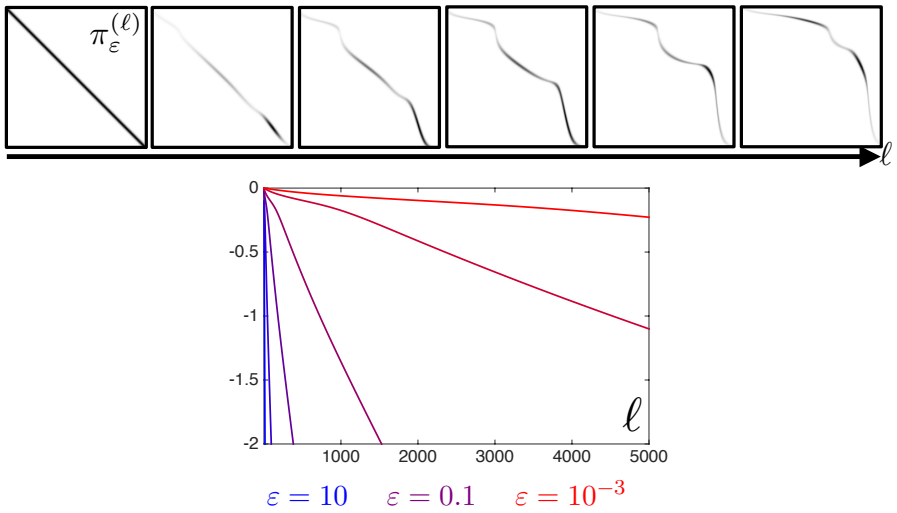


Figure 4.5: Top: evolution of the coupling $\pi_\varepsilon^{(\ell)} = \mathrm{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\,\mathrm{diag}(\mathbf{v}^{(\ell)})$ computed at iteration ℓ of Sinkhorn's iterations, for 1-D densities on $\mathcal{X} = [0,1]$, $c(x,y) = |x-y|^2$, and $\varepsilon = 0.1$. Bottom: impact of ε the convergence rate of Sinkhorn, as measured in term of marginal constraint violation $\log(\|\pi_\varepsilon^{(\ell)}\mathbb{1}_m - \mathbf{b}\|_1)$.

more recently by Galichon and Salanié [2009]. This regularization has received recently renewed attention in data sciences (including machine learning, vision, graphics and imaging) following [Cuturi, 2013], who showed that Sinkhorn's algorithm provides a scalable way to approximate optimal transport, thanks to seamless parallelization when solving several OT problems simultaneously (notably on GPUs, see Remark 4.14) and that it defines, unlike the linear programming formulation, a differentiable loss function (see §4.5). There exist countless extensions and generalizations of the Sinkhorn algorithm (see for instance §4.6). For instance, when $\mathbf{a} = \mathbf{b}$, one can use averaged projection iterations to maintain symmetry Knight et al. [2014].

Remark 4.5 (Overall complexity)**.** By doing a careful convergence analysis (assuming $n = m$ for the sake of simplicity), Altschuler et al. [2017] showed that by setting $\varepsilon = \frac{4\log(n)}{\tau}$, $O(\|\mathbf{C}\|_\infty^3 \log(n)\tau^{-3})$ Sinkhorn iterations (with an additional rounding step to com-

pute a valid coupling $\hat{\mathbf{P}} \in \mathbf{U}(\mathbf{a}, \mathbf{b})$) are enough to ensure that $\langle \hat{\mathbf{P}}, \mathbf{C} \rangle \leq \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) + \tau$. This implies that Sinkhorn computes a τ-approximate solution of the unregularized OT problem in $O(n^2 \log(n) \tau^{-3})$ operations. The rounding scheme consists in, given two vectors $\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m$ to carry out the following updates [Altschuler et al., 2017, Alg. 2]:

$$\mathbf{u}' \stackrel{\text{def.}}{=} \mathbf{u} \odot \min\left(\frac{\mathbf{a}}{\mathbf{u} \odot (\mathbf{K}\mathbf{v})}, \mathbb{1}_n\right), \mathbf{v}' \stackrel{\text{def.}}{=} \mathbf{v} \odot \min\left(\frac{\mathbf{b}}{\mathbf{v} \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u}')}, \mathbb{1}_n\right),$$
$$\Delta_{\mathbf{a}} \stackrel{\text{def.}}{=} \mathbf{a} - \mathbf{u}' \odot (\mathbf{K}\mathbf{v}'), \Delta_{\mathbf{b}} \stackrel{\text{def.}}{=} \mathbf{b} - \mathbf{v}' \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u}),$$
$$\hat{\mathbf{P}} \stackrel{\text{def.}}{=} \operatorname{diag}(\mathbf{u}')\mathbf{K}\operatorname{diag}(\mathbf{v}') + \Delta_{\mathbf{a}}(\Delta_{\mathbf{b}})^{\mathrm{T}} / \|\Delta_{\mathbf{a}}\|_1 .$$

This yields a matrix $\hat{\mathbf{P}} \in \mathbf{U}(\mathbf{a}, \mathbf{b})$ such that the 1 norm between $\hat{\mathbf{P}}$ and $\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$ is controlled by the marginal violations of $\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$, namely

$$\left\|\hat{\mathbf{P}} - \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})\right\|_1 \leq \|\mathbf{a} - \mathbf{u} \odot (\mathbf{K}\mathbf{v})\|_1 + \left\|\mathbf{b} - \mathbf{v} \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u})\right\|_1 .$$

Remark 4.6 (Numerical Stability of Sinkhorn Iterations)**.** As we discuss in Remarks 4.12 and 4.13, the convergence of Sinkhorn's algorithm deteriorates as $\varepsilon \to 0$. In numerical practice, however, that slowdown is rarely observed in practice for a simpler reason: Sinkhorn's algorithm will often fail to terminate as soon as some of the elements of the kernel $\mathbf{K}$ or become too negligible to be stored in memory as a positive number, and become instead null. This can then result in a matrix product $\mathbf{K}\mathbf{v}$ or $\mathbf{K}^{\mathrm{T}}\mathbf{u}$ with ever smaller entries that become null and result in a division by 0 in the Sinkhorn update of Equation (4.15). Such issues can be partly resolved by carrying out computations on the multipliers $\mathbf{u}$ and $\mathbf{v}$ in the log domain. That approach is carefully presented in Remark 4.22, and related to a direct resolution of the dual of problem (4.2).

Remark 4.7 (Relation with iterative projections)**.** Denoting

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

the rows and columns constraints, one has $\mathbf{U}(\mathbf{a}, \mathbf{b}) = \mathcal{C}_{\mathbf{a}}^1 \cap \mathcal{C}_{\mathbf{b}}^2$. One can

use Bregman iterative projections [Bregman, 1967]

$$\mathbf{P}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\mathcal{C}^1_{\mathbf{a}}}(\mathbf{P}^{(\ell)}) \quad \text{and} \quad \mathbf{P}^{(\ell+2)} \overset{\text{def.}}{=} \mathrm{Proj}^{\mathbf{KL}}_{\mathcal{C}^2_{\mathbf{b}}}(\mathbf{P}^{(\ell+1)}). \tag{4.16}$$

Since the sets $\mathcal{C}^1_{\mathbf{a}}$ and $\mathcal{C}^2_{\mathbf{b}}$ are affine, these iterations are known to converge to the solution of (4.7), see [Bregman, 1967]. These iterates are equivalent to Sinkhorn iterations (4.15) since defining

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

one has

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

In practice however one should prefer using (4.15) which only requires manipulating scaling vectors and multiplication against a Gibbs kernel, which can often be accelerated (see below Remarks 4.15 and 4.17).

Remark 4.8 (Other regularizations)**.** It is possible to replace the entropic term $-\mathbf{H}(\mathbf{P})$ in (4.2) by any strictly convex penalty $R(\mathbf{P})$, as detailed for instance in [Dessein et al., 2016]. A typical example is the squared ℓ^2 norm

$$R(\mathbf{P}) = \sum_{i,j} \mathbf{P}_{i,j}^2 + \iota_{\mathbb{R}_+}(\mathbf{P}_{i,j}), \tag{4.17}$$

see [Essid and Solomon, 2017]. Another example is the family of Tsallis entropies [Muzellec et al., 2017]. Note however that if the penalty function is defined even when entries of P are non-positive, which is for instance the case for a quadratic regularization (4.17), then one must add back a non-negativity constraint $\mathbf{P} \geq 0$, in addition to the marginal constraints $\mathbf{P}\mathbb{1}_m = \mathbf{a}$ and $\mathbf{P}^\top \mathbb{1}_n = \mathbf{b}$ (one can afford to ignore the nonnegativity constraint using entropy because that penalty incorporates a logarithmic term which forces the entries of $\mathbf{P}$ to stay in the positive orthant). This implies that the set of constraints is no longer affine and iterative Bregman projections do not converge anymore to the solution. A workaround is to use instead Dykstra's algorithm [Dykstra, 1983, 1985, Bauschke and Lewis, 2000], as detailed in [Benamou et al., 2015]. This algorithm uses projections according to the Bregman

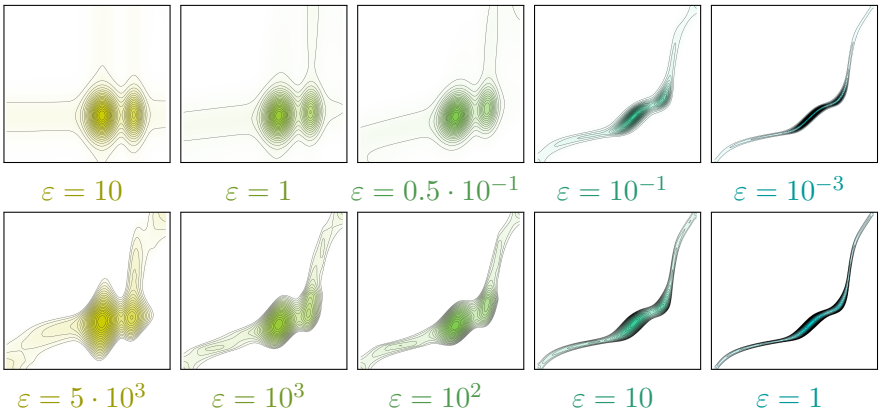


Figure 4.6: Comparison of entropic regularization $R = -\mathbf{H}$ (top row) and quadratic regularization $R = \|\cdot\|^2 + \iota_{\mathbb{R}_+}$ (bottom row). The (α, β) marginals are the same as for Figure 4.4.

divergence associated to R. We refer to Remark 8.1 for more details regarding Bregman divergences. An issue is that in general these projections cannot be computed explicitly. For the squared norm (4.17), this corresponds to computing Euclidean projection on $(\mathcal{C}^1_{\mathbf{a}}, \mathcal{C}^2_{\mathbf{b}})$ (with the extra positivity constraints), which can be solved efficiently using projection algorithms on simplices [Condat, 2015]. The main advantage of the quadratic regularization over entropy is that it produces sparse approximation of the optimal coupling, yet this comes at the expense of a slower algorithm that cannot be parallelized as efficiently as Sinkhorn to compute several optimal transports simultaneously (as discussed in §4.14). Figure 4.6 contrasts the approximation achieved by entropic and quadratic regularizers.

Remark 4.9 (Barycentric projection)**.** The Kantorovitch formulation (2.11) and its entropic regularization (4.2) both yield an optimal coupling $\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})$. In order to define a transportation map $T : \mathcal{X} \to \mathcal{Y}$, in the case where $\mathcal{Y} = \mathbb{R}^d$, one can define the so-called

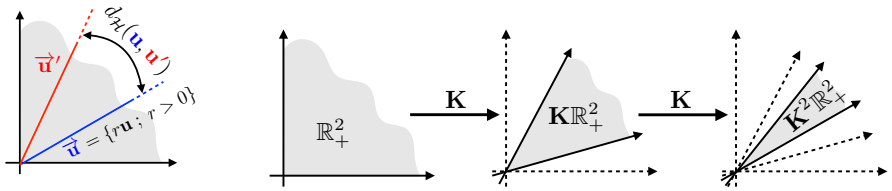


Figure 4.7: Left: the Hilbert metric $d_{\mathcal{H}}$ is a distance over rays in cones (here positive vectors). Right: vizualization of the contraction induced by the iteration of a positive matrix $\mathbf{K}$.

barycentric projection map

$$T : x_i \in \mathcal{X} \longmapsto \frac{1}{\mathbf{a}_i} \sum_j \mathbf{P}_{i,j} y_j \in \mathcal{Y} \tag{4.18}$$

where here the input measures are discrete of the form (2.3). Note that this map is only defined for points $(x_i)_i$ of the support of α. In the case where T is a permutation matrix (as detailed in Proposition 2.1), then T is equal to a Monge map, and as $\varepsilon \to 0$, the barycentric projection progressively converges to that map if it is unique. For arbitrary (not necessarily discrete) measures, solving (2.15) or its regularized version (4.9) defines a coupling $\pi \in \mathcal{U}(\alpha, \beta)$. Note that for $\varepsilon > 0$, because of the entropic regularization, this coupling π has a density $\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}$ with respect to $\alpha \otimes \beta$. A map can thus be retrieved by the formula

$$T : x \in \mathcal{X} \longmapsto \int_{\mathcal{Y}} y \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)} \mathrm{d}\beta(y). \tag{4.19}$$

In the case where, for $\varepsilon = 0$, π is supported on the graph of the Monge map (see Remark 2.23), then using $\varepsilon > 0$ produces a smooth approximation of this map. Such a barycentric projection is useful to apply the OT Monge map to solve problems in imaging, see Figure 9.6 for an application to color modification. It has also been used to compute approximations of geodesic Pincipal Geodesic Analysis (PCA) in the space of probability measures endowed with the Wasserstein metric, see [Seguy and Cuturi, 2015].

Remark 4.10 (Hilbert metric)**.** As initially explained by [Franklin and Lorenz, 1989], the global convergence analysis of Sinkhorn is greatly simplified using Hilbert projective metric on $\mathbb{R}^n_{+,*}$ (positive vectors), defined as

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

It can be shown to be a distance on the projective cone $\mathbb{R}^n_{+,*}/\sim$, where $\mathbf{u}\sim\mathbf{u}'$ means that $\exists r>0,\mathbf{u}=r\mathbf{u}'$ (the vector are equal up to rescaling, hence the naming "projective"). This means that $d_{\mathcal{H}}$ satisfies the triangular inequality and $d_{\mathcal{H}}(\mathbf{u},\mathbf{u}')=0$ if and only if $\mathbf{u}\sim\mathbf{u}'$. This is a projective version of Hilbert's original distance on bounded open convex sets Hilbert [1895]. The projective cone $\mathbb{R}^n_{+,*}/\sim$ is a complete metric space for this distance. By a logarithmic change of variable, the Hilbert metric on the rays of the positive cone is isometric to the variation norm between vectors that are defined up to an additive constant

$$d_{\mathcal{H}}(\mathbf{u},\mathbf{u}')=\big\|\log(\mathbf{u})-\log(\mathbf{u}')\big\|_{\text{var}}\tag{4.20}$$

$$\text{where}\quad\|\mathbf{f}\|_{\text{var}}\overset{\text{def.}}{=}(\max_i\mathbf{f}_i)-(\min_i\mathbf{f}_i).$$

The Hilbert metric was introduced independently by [Birkhoff, 1957] and [Samelson et al., 1957]. They proved the following fundamental theorem, which shows that a positive matrix is a strict contraction on the cone of positive vectors.

Theorem 4.1. Let $\mathbf{K}\in\mathbb{R}^{n\times m}_{+,*}$, then for $(\mathbf{v},\mathbf{v}')\in(\mathbb{R}^m_{+,*})^2$

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

Figure 4.7 illustrates this theorem.

Remark 4.11 (Perron-Frobenius)**.** A typical application of Theorem 4.1 is to provide a quantitative proof of Perron-Frobenius

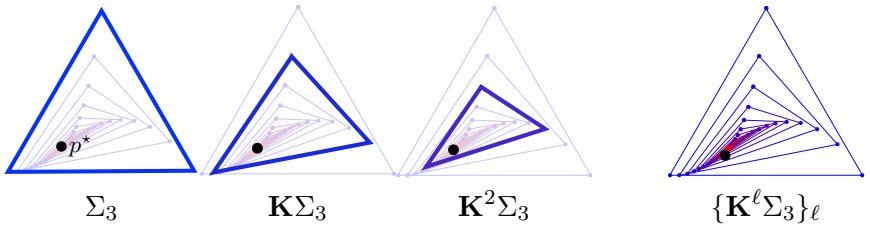


Figure 4.8: Evolution of $\mathbf{K}^\ell\Sigma_3 \to \{p^\star\}$ the invariant probability distribution of $\mathbf{K} \in \mathbb{R}_{+,*}^{3\times 3}$ with $\mathbf{K}^\top \mathbb{1}_3 = \mathbb{1}_3$.

theorem, which, as explained in Remark 4.13, is linked to a local linearization of Sinkhorn's iterates. A matrix $\mathbf{K} \in \mathbb{R}_+^{n\times n}$ with $\mathbf{K}^\top \mathbb{1}_n = \mathbb{1}_n$ maps Σ_n into Σ_n. If furthermore $\mathbf{K} > 0$, then according to Theorem 4.1, it is strictly contractant for the metric $d_{\mathcal{H}}$, hence there exists a unique invariant probability distribution $p^\star \in \Sigma_n$ with $\mathbf{K}p^\star = p^\star$. Furthermore, for any $p_0 \in \Sigma_n$, $d_{\mathcal{H}}(\mathbf{K}^\ell p_0, p^\star) \leq \lambda(\mathbf{K})^\ell d_{\mathcal{H}}(p_0, p^\star)$, i.e. one has linear convergence of the iterates of the matrix toward $p^\star$. This is illustrated on Figure 4.8.

Remark 4.12 (Global convergence)**.** The following theorem, proved by [Franklin and Lorenz, 1989], makes use of this Theorem 4.1 to show the linear convergence of Sinkhorn's iterations.

Theorem 4.2. One has $(\mathbf{u}^{(\ell)}, \mathbf{v}^{(\ell)}) \to (\mathbf{u}^\star, \mathbf{v}^\star)$ and

$$d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) = O(\lambda(\mathbf{K})^{2\ell}), \quad d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) = O(\lambda(\mathbf{K})^{2\ell}). \tag{4.21}$$

One also has

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) &\leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell)}\mathbb{1}_m, \mathbf{a})}{1-\lambda(\mathbf{K})} \\ d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) &\leq \frac{d_{\mathcal{H}}(\mathbf{P}^{(\ell),\top}\mathbb{1}_n, \mathbf{b})}{1-\lambda(\mathbf{K})} \end{aligned} \tag{4.22}$$

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

$$\|\log(\mathbf{P}^{(\ell)}) - \log(\mathbf{P}^\star)\|_\infty \leq d_{\mathcal{H}}(\mathbf{u}^{(\ell)}, \mathbf{u}^\star) + d_{\mathcal{H}}(\mathbf{v}^{(\ell)}, \mathbf{v}^\star) \tag{4.23}$$

where $\mathbf{P}^\star$ is the unique solution of (4.2).

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

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

This shows that

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

where we used Theorem 4.1. This shows (4.21). One also has, using the triangular inequality

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

which gives the first part of (4.22) since $\mathbf{u}^{(\ell)} \odot (\mathbf{K}\mathbf{v}^{(\ell)}) = \mathbf{P}^{(\ell)}\mathbb{1}_m$ (the second one being similar). The proof of (4.23) follows from [Franklin and Lorenz, 1989, Lemma 3] □

The bound (4.22) shows that some error measures on the marginal constraints violation, for instance $\|\mathbf{P}^{(\ell)}\mathbb{1}_m - \mathbf{a}\|_1$ and $\|{\mathbf{P}^{(\ell)}}^{\mathrm{T}}\mathbb{1}_n - \mathbf{b}\|_1$, are useful stopping criteria to monitor the convergence. Note that thanks to (4.20), these Hilbert metric rates on the scaling variable $(\mathbf{u}^{(\ell)}, \mathbf{v}^{(\ell)})$ gives linear rate on the dual variables $(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)}) \overset{\text{def.}}{=} (\varepsilon \log(\mathbf{u}^{(\ell)}), \varepsilon \log(\mathbf{v}^{(\ell)}))$ for the variation norm $\|\cdot\|_{\text{var}}$.

Figure 4.5, bottom row, highlights this linear rate on the constraint violation, and shows how this rate degrades as $\varepsilon \to 0$. These results are proved in [Franklin and Lorenz, 1989] and are tightly

connected to nonlinear Perron-Frobenius Theory [Lemmens and Nussbaum, 2012]. Perron-Frobenius theory corresponds to the linearization of the iterations, see (4.24). This convergence analysis is extended in [Linial et al., 1998], who shows that each iteration of Sinkhorn increases the permanent of the scaled coupling matrix.

Remark 4.13 (Local convergence)**.** The global linear rate (4.23) is often quite pessimistic, typically in $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ for cases where there exists a Monge map when $\varepsilon = 0$ (see Remark 2.7). The global rate is in contrast rather sharp for more difficult situations where the cost matrix $\mathbf{C}$ is close to being random, and in these cases, the rate scales exponentially bad with ε, $1 - \lambda(\mathbf{K}) \sim e^{-1/\varepsilon}$. To obtain a finer asymptotic analysis of the convergence (i.e. if one is interested in high precision solution and performs a large number of iterations), one usually rather studies the local convergence rate. One can write Sinkhorn update as iterations of a fixed-point map $\mathbf{f}^{(\ell+1)} = \Phi(\mathbf{f}^{(\ell)})$ where

$$\Phi \overset{\text{def.}}{=} \Phi_2 \circ \Phi_1 \quad \text{where} \quad \begin{cases} \Phi_1(\mathbf{f}) = \varepsilon \log \mathbf{K}^{\mathrm{T}}(e^{\mathbf{f}/\varepsilon}) - \log(\mathbf{b}), \\ \Phi_2(\mathbf{g}) = \varepsilon \log \mathbf{K}(e^{\mathbf{g}/\varepsilon}) - \log(\mathbf{a}). \end{cases}$$

For optimal $(\mathbf{f}, \mathbf{g})$ solving (4.29), denoting $\mathbf{P} = \operatorname{diag}(e^{\mathbf{f}/\varepsilon})\mathbf{K}\operatorname{diag}(e^{\mathbf{g}/\varepsilon})$ the optimal coupling solving (4.2), one has the following Jacobian

$$\partial\Phi(\mathbf{f}) = \operatorname{diag}(\mathbf{a})^{-1} \circ \mathbf{P} \circ \operatorname{diag}(\mathbf{b})^{-1} \circ \mathbf{P}^{\mathrm{T}}. \tag{4.24}$$

This Jacobian is a positive matrix with $\partial\Phi(\mathbf{f})\mathbb{1}_n = \mathbb{1}_n$, and thus by Perron-Frobenius theorem, it has a single dominant eigenvector $\mathbb{1}_m$ with associated eigenvalue 1. Since $\mathbf{f}$ is defined up to a constant, it is actually the second eigenvalue $1 - \kappa < 1$ which governs the local linear rate, and this shows that for ℓ large enough, $\|\mathbf{f}^{(\ell)} - \mathbf{f}\| = O((1-\kappa)^{\ell})$. Numerically, in "simple cases" (such as when there exists a smooth Monge-map when $\varepsilon = 0$), this rate scales like $\kappa \sim \varepsilon$. We refer to [Knight, 2008] for more details in the bistochastic (assignment) case.

4.3 Speeding-up Sinkhorn's Iterations

The main computational bottleneck of Sinkhorn's iterations is the vector-matrix multiplication against kernels $\mathbf{K}$ and $\mathbf{K}^\top$, with complexity $O(nm)$ if implemented naively. We now detail several important cases where the complexity can be improved significantly.

Remark 4.14 (Parallel and GPU Friendly Computation)**.** The simplicity of Sinkhorn's algorithm yields an extremely efficient approach to compute simultaneously several regularized Wasserstein distances between pairs of histograms. Let N be an integer, $\mathbf{a}_1, \cdots, \mathbf{a}_N$ be histograms in Σ_n and $\mathbf{b}_1, \cdots, \mathbf{b}_N$ histograms in Σ_m. We seek to compute all N approximate distances $\mathrm{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}_1, \mathbf{b}_1), \ldots, \mathrm{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}_N, \mathbf{b}_N)$. In that case, writing $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$ for the $n \times N$ and $m \times N$ matrices storing all histograms, one can notice that all Sinkhorn iterations for all these N pairs can be carried out in parallel, by setting for instance

$$\mathbf{U}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{A}}{\mathbf{K}\mathbf{V}^{(\ell)}} \quad \text{and} \quad \mathbf{V}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{B}}{\mathbf{K}^{\mathrm{T}}\mathbf{U}^{(\ell+1)}}, \tag{4.25}$$

initialized with $\mathbf{V}^{(0)} = \mathbb{1}_{m \times N}$. Here $\frac{\cdot}{\cdot}$ corresponds to entry-wise division of matrices. One can further check that, upon convergence of $\mathbf{V}$ and $\mathbf{U}$, the (row) vector of regularized distances simplifies to

$$(\mathbf{U} \odot \log \mathbf{U})^{\mathrm{T}} (\mathbf{K} \odot \mathbf{C})\, \mathbf{V} + (\mathbf{V} \odot \log \mathbf{V})^{\mathrm{T}} (\mathbf{K} \odot \mathbf{C})^{\mathrm{T}} \mathbf{U} \in \mathbb{R}^N.$$

Note that the basic Sinkhorn iterations described in Equation (4.15) are intrinsically GPU friendly, since they only consist in matrix-vector products, and this was exploited for instance to solve matching problems in Slomp et al. [2011]). However, the matrix-matrix operations presented in Equation (4.25) present even better opportunities for parallelism, which explains the success of Sinkhorn's algorithm to compute OT distances between histograms at large scale.

Remark 4.15 (Speed up for separable kernels)**.** if

$$i = (i_k)_{k=1}^d, j = (j_k)_{k=1}^d, \quad \mathbf{K}_{i,j} = \prod_{k=1}^d \mathbf{K}^k_{i_k, j_k} \tag{4.26}$$

Then one can compute $K\mathbf{u}$ by applying $\mathbf{K}_k$ along each "slice". If $n = m$, complexity $O(n^{1+1/d})$ in place of $O(n^2)$. Typical example is when $c(x,y) = \|x-y\|^2$ for $X = Y = \mathbb{R}^d$ on a regular grid $x_i = (i_1/n_0, \ldots, i_k/n_0)_{i_k=1}^n$, $n = n_0^d$, then $\mathbf{K}^k_{i_k,j_k} = \exp(-|i_k/n_0 - j_k/n_0|^2/\varepsilon)$ is a 1-D Gaussian convolution.

Remark 4.16 (Convolutions approximation)**.** The main computational bottleneck of Sinkhorn iterations (4.15) is the matrix multiplication $\mathbf{Kv}$ (and also its adjoint). Beside using separability (4.26), it is also possible to make use of other special structure of the kernel. The simplest case is for translation invariant kernels $\mathbf{K}_{i,j} = k_{i-j}$, which is typically the case when discretizing the measure on a fixed uniform grid in Euclidean space $\mathcal{X} = \mathbb{R}^d$. Then $\mathbf{Kv} = k \star \mathbf{v}$ is a convolution, and there are several algorithms to approximate the convolution in nearly linear time. The most usual one is by Fourier transform $\mathcal{F}$, assuming for simplicity periodic boundary conditions, because $\mathcal{F}(k \star \mathbf{v}) = \mathcal{F}(k) \odot \mathcal{F}(\mathbf{v})$. This leads however to unstable computations, and is often unacceptable for small ε. Another popular way to speed-up computation is by approximating the convolution using a succession of auto-regressive filter. The most well known approach is the Deriche filtering method Deriche [1993], and we refer to Getreuer [2013] for a comparison of various fast filtering methods.

Remark 4.17 (Geodesic in heat approximation)**.** For non-planar domains, the kernel $\mathbf{K}$ is not a convolution, but in the case where the cost is $\mathbf{C}_{i,j} = d_{\mathcal{M}}(x_i, y_j)^p$ where $d_{\mathcal{M}}$ is a geodesic distance on a surface $\mathcal{M}$ (or a more general manifold), it is also possible to perform fast approximations of the application of $\mathbf{K} = e^{-\frac{d_{\mathcal{M}}}{\varepsilon}}$ to a vector. Indeed, Varadhan formulas Varadhan [1967] assert that this kernel is close to the Laplacian kernel (for $p = 1$) and the heat kernel (for $p = 2$). The first formula of Varadhan states

$$-\frac{\sqrt{t}}{2}\log(\mathcal{P}_t(x,y)) = d_{\mathcal{M}}(x,y) + o(t) \quad \text{where} \quad \mathcal{P}_t \overset{\text{def.}}{=} (\mathrm{Id} - t\Delta_{\mathcal{M}})^{-1} \tag{4.27}$$

where $\Delta_{\mathcal{M}}$ is the Laplace-Beltrami operator associated to the manifold $\mathcal{M}$ (which is negative semi-definite), so that $\mathcal{P}_t$ is an integral kernel and

$g = \int_{\mathcal{M}} \mathcal{P}_t(x,y) f(y) \mathrm{d}y$ is the solution of $g - t\Delta_{\mathcal{M}} g = f$. The second formula of Varadhan states

$$\sqrt{-4t \log(\mathcal{H}_t(x,y))} = d_{\mathcal{M}}(x,y) + o(t) \tag{4.28}$$

where $\mathcal{H}_t$ is the integral kernel defined so that $g_t = \int_{\mathcal{M}} \mathcal{H}_\varepsilon(x,y) f(y) \mathrm{d}y$ is the solution at time t of the heat equation

$$\frac{\partial g_t(x)}{\partial t} = (\Delta_{\mathcal{M}} g_t)(x).$$

The convergence in these formulas (4.27) and (4.28) is uniform on compact manifolds. Numerically, the domain $\mathcal{M}$ is discretized (for instance using finite elements) and $\Delta_{\mathcal{M}}$ is approximated by a discrete Laplacian matrix L. A typical example is when using piecewise linear finite elements, so that L is the celebrated cotangent Laplacian (see [Botsch et al., 2010] for a detailed account for this construction). These formulas can be used to approximate efficiently the multiplication by the Gibbs kernel $\mathbf{K}_{i,j} = e^{-\frac{d(x_i,y_j)^p}{\varepsilon}}$. Equation (4.27) suggests, for the case $p = 1$, to use $\varepsilon = \frac{\sqrt{t}}{2}$ and to replace the multiplication by $\mathbf{K}$ by the multiplication by $(\mathrm{Id} - tL)^{-1}$, which necessitates the resolution of a positive symmetric linear system. Equation (4.28), coupled with R steps of implicit Euler for the stable resolution of the Heat flow, suggests for $p = 2$ to trade the multiplication by $\mathbf{K}$ by the multiplication by $(\mathrm{Id} - \frac{t}{R} L)^{-R}$ for $4t = \varepsilon$, which in turn necessitates R resolutions of linear systems. Fortunately, since these linear systems are supposed to be solved many times during Sinkhorn iterations, one can solve them efficiently by pre-computing a sparse Cholesky factorization. By performing a re-ordering of the rows and columns of the matrix [George and Liu, 1989], one obtains a nearly linear sparsity for 2-D manifolds and thus each iteration of Sinkhorn has linear complexity (the performance degrades with the dimension of the manifold). The use of Varadhan's formula to approximate geodesic distances was initially proposed in [Crane et al., 2013], and its use in conjunction with Sinkhorn iteration in [Solomon et al., 2015].

Remark 4.18 (Extrapolation acceleration)**.** Since Sinkhorn algorithm is a fixed-point algorithm (as shown in Remark 4.13), one can use standard linear or even non-linear extrapolation schemes to enhance the

conditioning of the fixed-point mapping near the solution, and improve the linear convergence rate. This is similar to the Successive overrelaxation method (see for instance [Hadjidimos, 2000]), so that the local linear rate of convergence is improved from $O((1-\kappa)^\ell)$ to $O((1-\sqrt{\kappa})^\ell)$ for some $\kappa > 0$ (see Remark 4.13). We refer to [Peyré et al., 2017] for more details.

4.4 Regularized Dual and Log-domain Computations

The following proposition details the dual problem associated to (4.2).

Proposition 4.4. One has

$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{b}) = \max_{\mathbf{f}\in\mathbb{R}^n,\mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f},\, \mathbf{a}\rangle + \langle \mathbf{g},\, \mathbf{b}\rangle - \varepsilon\langle e^{\mathbf{f}/\varepsilon},\, \mathbf{K}e^{\mathbf{g}/\varepsilon}\rangle. \tag{4.29}$$

The optimal $(\mathbf{f},\mathbf{g})$ are linked to scalings $(\mathbf{u},\mathbf{v})$ appearing in (4.12) through

$$(\mathbf{u},\mathbf{v}) = (e^{\mathbf{f}/\varepsilon}, e^{\mathbf{g}/\varepsilon}). \tag{4.30}$$

Proof. We start from the end of the proof of Proposition 4.3, which links the optimal primal solution $\mathbf{P}$ and dual multipliers $\mathbf{f}$ and $\mathbf{g}$ for the marginal constraints as $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon}$. Substituting in the Lagrangian $\mathcal{E}(\mathbf{P},\mathbf{f},\mathbf{g})$ of Equation (4.2) the optimal $\mathbf{P}$ as a function of $\mathbf{f}$ and $\mathbf{g}$, we obtain that the Lagrange dual function equals

$$\mathbf{f},\mathbf{g} \mapsto \langle e^{\mathbf{f}/\varepsilon},\, (\mathbf{K}\odot\mathbf{C})\, e^{\mathbf{g}/\varepsilon}\rangle - \varepsilon\mathbf{H}(\mathrm{diag}(e^{\mathbf{f}/\varepsilon})\mathbf{K}\,\mathrm{diag}(e^{\mathbf{g}/\varepsilon})). \tag{4.31}$$

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

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

therefore, the first term in (4.31) cancels out with the first term in the entropy above. The remaining times are those displayed in (4.29). □

Remark 4.19 (Dual for generic measures)**.** For generic (non-necessarily discrete) input measures (α,β), the dual problem (4.29)

reads

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

This corresponds to a smoothing of the constraint $\mathcal{R}(c)$ appearing in the original problem (2.23), which is retrieved in the limit $\varepsilon \to 0$. Proving existence (*i.e.* the sup is actually a max) of these Kantorovich potentials (f, g) in the case of entropic transport is less easy than for classical OT (because one cannot use c-transform and potentials are not automatically Lipschitz). Proof of existence can be done using the convergence of Sinkhorn iterations, see Chizat et al. [2017] for more details.

Remark 4.20 (Sinkhorn as a Block Coordinate Ascent on the Dual Problem)**.** A simple approach to solve the unconstrained maximization problem (4.29) is to use an exact *block coordinate ascent* strategy, namely to update alternatively $\mathbf{f}$ and $\mathbf{g}$ to cancel their gradients with respect to the objective of (4.29). Indeed, one can easily notice that, writing $Q(\mathbf{f}, \mathbf{g})$ for the objective of (4.29) that

$$\nabla|_{\mathbf{f}} Q(\mathbf{f}, \mathbf{g}) = \mathbf{a} - e^{\mathbf{f}/\varepsilon} \odot \left(\mathbf{K} e^{\mathbf{g}/\varepsilon}\right), \tag{4.32}$$

$$\nabla|_{\mathbf{g}} Q(\mathbf{f}, \mathbf{g}) = \mathbf{b} - e^{\mathbf{g}/\varepsilon} \odot \left(\mathbf{K}^T e^{\mathbf{f}/\varepsilon}\right). \tag{4.33}$$

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

$$\mathbf{f}^{(\ell+1)} = \varepsilon \log \mathbf{a} - \varepsilon \log\left(\mathbf{K} e^{\mathbf{g}^{(\ell)}/\varepsilon}\right), \tag{4.34}$$

$$\mathbf{g}^{(\ell+1)} = \varepsilon \log \mathbf{b} - \varepsilon \log\left(\mathbf{K}^{\mathrm{T}} e^{\mathbf{f}^{(\ell+1)}/\varepsilon}\right). \tag{4.35}$$

Such iterations are mathematically equivalent to the Sinkhorn iterations (4.15) when considering the primal-dual relations highlighted in (4.30). Indeed, we recover that at any iteration

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

Remark 4.21 (Soft-min rewriting)**.** Iterations (4.34) and (4.35) can be given an alternative interpretation, using the following notation. Given a vector $\mathbf{z}$ of real numbers we write $\min_\varepsilon \mathbf{z}$ for the *soft-minimum* of its coordinates, namely

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

Note that $\min_\varepsilon(\mathbf{z})$ converges to $\min \mathbf{z}$ for any vector $\mathbf{z}$ as $\varepsilon \to 0$. Indeed, $\min_\varepsilon$ can be interpreted as a differentiable approximation of the min function. Using these notations, Equations (4.34) and (4.35) can be rewritten

$$(\mathbf{f}^{(\ell+1)})_i = \min_\varepsilon{}_{(j)} (\mathbf{C}_{ij} - \mathbf{g}_j^{(\ell)}) + \varepsilon \log \mathbf{a}_i, \tag{4.36}$$

$$(\mathbf{g}^{(\ell+1)})_j = \min_\varepsilon{}_{(i)} (\mathbf{C}_{ij} - \mathbf{f}_i^{(\ell)}) + \varepsilon \log \mathbf{b}_j. \tag{4.37}$$

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

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

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

Note that these operations are equivalent to the entropic c-transform introduced in §5.3 (see in particular (5.11)). Using these notations, Sinkhorn's iterates read

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{row}} (\mathbf{C} - \mathbb{1}_n \mathbf{g}^{(\ell)\mathrm{T}}) + \varepsilon \log \mathbf{a}, \tag{4.38}$$

$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{col}} (\mathbf{C} - \mathbf{f}^{(\ell)} \mathbb{1}_m{}^{\mathrm{T}}) + \varepsilon \log \mathbf{b}. \tag{4.39}$$

Note that as $\varepsilon \to 0$, $\min_\varepsilon$ converges to min, but the iterations do not converge anymore in the limit $\varepsilon = 0$, because alternate minimization does not converge for constrained problems (which is the case for the un-regularized dual (2.19)).

Remark 4.22 (Log-domain Sinkhorn)**.** While mathematically equivalent to the Sinkhorn updates (4.15), iterations (4.36) and (4.37) suggest to

use the *log-sum-exp* stabilization trick to avoid underflow for small values of ε. Writing $\underline{z} = \min \mathbf{z}$, that trick suggests to evaluate $\min_\varepsilon \mathbf{z}$ as

$$\min_\varepsilon \mathbf{z} = \underline{z} - \varepsilon \log \sum_i e^{-(\mathbf{z}_i - \underline{z})/\varepsilon}. \tag{4.40}$$

Instead of substracting $\underline{z}$ to stabilize the log domain iterations as in (4.40), one can actually substract the previously computed scalings. This leads to the following stabilized iteration

$$\mathbf{f}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{row}}\left(\mathbf{S}(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)})\right) - \mathbf{f}^{(\ell)} + \varepsilon \log(\mathbf{a}) \tag{4.41}$$

$$\mathbf{g}^{(\ell+1)} = \mathrm{Min}_\varepsilon^{\mathrm{col}}\left(\mathbf{S}(\mathbf{f}^{(\ell+1)}, \mathbf{g}^{(\ell)})\right) - \mathbf{g}^{(\ell)} + \varepsilon \log(\mathbf{b}), \tag{4.42}$$

where we defined

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

In contrast to the original iterations (4.15), these log-domain iterations (4.41) and (4.42) are stable for arbitrary $\varepsilon > 0$, because the quantity $\mathbf{S}(\mathbf{f}, \mathbf{g})$ stays bounded during the iterations. The downside is that it requires nm computations of exp at each step. Computing a $\mathrm{Min}_\varepsilon^{\mathrm{row}}$ or $\mathrm{Min}_\varepsilon^{\mathrm{col}}$ is typically substantially slower than matrix multiplications, and requires computing line by line soft-minima of matrices $\mathbf{S}$. There is therefore no efficient way to parallelize the application of Sinkhorn maps for several marginals simultaneously. In Euclidean domain of small dimension, it is possible to develop efficient multiscale solvers with a decaying ε strategy to significantly speed up the computation using sparse grids [Schmitzer, 2016b].

4.5 Regularized Approximations of the Optimal Transport Cost

The entropic dual (4.29) is a smooth unconstrained concave maximization problem, which approximates the original Kantorovich dual (2.19), as detailed in the following proposition.

Proposition 4.5. Any pair of optimal solutions $(\mathbf{f}^\star, \mathbf{g}^\star)$ to (4.29) are such that $(\mathbf{f}^\star, \mathbf{g}^\star) \in \mathbf{R}(\mathbf{a}, \mathbf{b})$, the set of feasible Kantorovich potentials defined in (2.20). As a consequence, we have that for any ε,

$$\langle \mathbf{f}^\star, \mathbf{a} \rangle + \langle \mathbf{g}^\star, \mathbf{b} \rangle \leq \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}).$$

A chief advantage of the regularized transportation cost $\mathrm{L}_{\mathbf{C}}^{\varepsilon}$ defined in (4.2) is that it is smooth and convex, which makes it a perfect fit to integrate it as a loss function in variational problems (see Chapter 9).

Proposition 4.6. $\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b})$ is a jointly convex function of $\mathbf{a}$ and $\mathbf{b}$, with gradient equal to

$$\nabla \mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) = \begin{bmatrix} \mathbf{f}^\star \\ \mathbf{g}^\star \end{bmatrix},$$

where $\mathbf{f}^\star$ and $\mathbf{g}^\star$ are the optimal solutions of Equation (4.29) chosen so that their coordinates sum to 0.

In [Cuturi, 2013], lower and upper bounds to approximate the Wasserstein distance between two histograms were proposed. These bounds consist in evaluating the primal and dual objectives at the solutions provided by the Sinkhorn algorithm.

Definition 4.1 (Sinkhorn Divergences)**.** Let $\mathbf{f}^\star$ and $\mathbf{g}^\star$ be optimal solutions to (4.29). The Wasserstein distance is approximated using following primal and dual Sinkhorn divergences:

$$\begin{aligned} \mathfrak{P}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) &\stackrel{\text{def.}}{=} \langle \mathbf{C}, \mathbf{P}^\star \rangle = \langle e^{\frac{\mathbf{f}^\star}{\varepsilon}}, (\mathbf{K} \odot \mathbf{C}) e^{\frac{\mathbf{g}^\star}{\varepsilon}} \rangle \\ \mathfrak{D}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) &\stackrel{\text{def.}}{=} \langle \mathbf{f}^\star, \mathbf{a} \rangle + \langle \mathbf{g}^\star, \mathbf{b} \rangle, \end{aligned}$$

where $\odot$ stands for the elementwise product of matrices, and where $\mathbf{P}^\star$ is the solution to (4.2).

Proposition 4.7. The following relationships hold:

$$\mathfrak{D}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) \leq \mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) \leq \mathfrak{P}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}).$$

Furthermore

$$\mathfrak{P}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) - \mathfrak{D}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) = \varepsilon(\mathbf{H}(\mathbf{P}^\star) + 1). \tag{4.43}$$

Proof. Equation (4.43) is obtained by writing that the primal and dual problems have the same values at the optima (see (4.29)), and hence

$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) = \mathfrak{P}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) - \varepsilon \mathbf{H}(\mathbf{P}^\star) = \mathfrak{D}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) - \varepsilon \langle e^{\mathbf{f}^\star/\varepsilon}, \mathbf{K} e^{\mathbf{g}^\star/\varepsilon} \rangle$$

and remarking that $\langle e^{\mathbf{f}^\star/\varepsilon}, \mathbf{K} e^{\mathbf{g}^\star/\varepsilon} \rangle = 1$. □

The relationships given above are only valid upon convergence to the actual solutions of Equation (4.29). We consider next the practical case where the Sinkhorn iterations are terminated after a predetermined number of L iterations, and used to evaluate $\mathfrak{D}_{\mathbf{C}}^{\varepsilon}$ using iterates $\mathbf{f}^{(L)}$ and $\mathbf{g}^{(L)}$ instead of optimal solutions $\mathbf{f}^\star$ and $\mathbf{g}^\star$.

Using notations appearing in Equations (4.41) and (4.42), we thus introduce the following finite step approximation of $\mathrm{L}_{\mathbf{C}}^{\varepsilon}$,

$$\mathfrak{D}_{\mathbf{C}}^{(L)}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \langle \mathbf{f}^{(L)}, \mathbf{a}\rangle + \langle \mathbf{g}^{(L)}, \mathbf{b}\rangle. \tag{4.44}$$

This "algorithmic" Sinkhorn functional lower bound the regularized cost function.

Proposition 4.8 (Finite Sinkhorn Divergences)**.** The following relationship holds

$$\mathfrak{D}_{\mathbf{C}}^{(L)}(\mathbf{a},\mathbf{b}) \leq \mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{b}).$$

Note however that unlike the regularized expression $\mathrm{L}_{\mathbf{C}}^{\varepsilon}$ in (4.29), the finite Sinkhorn divergence $\mathfrak{D}_{\mathbf{C}}^{(L)}(\mathbf{a},\mathbf{b})$ is *not*, in general, a convex function of its arguments (this can be easily checked numerically). $\mathfrak{D}_{\mathbf{C}}^{(L)}(\mathbf{a},\mathbf{b})$ is, however, a differentiable function which can be differentiated using automatic differentiation techniques (see Remark 9.1.3) with respect to any of its arguments, notably $\mathbf{C}, \mathbf{a}$ or $\mathbf{b}$.

4.6 Generalized Sinkhorn

The regularized OT problem (4.2) is a special case of structured convex optimization problem of the form

$$\min_{\mathbf{P}} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j} - \varepsilon \mathbf{H}(\mathbf{P}) + F(\mathbf{P}\mathbb{1}_m) + G(\mathbf{P}^{\mathrm{T}}\mathbb{1}_n). \tag{4.45}$$

Indeed, defining $F = \iota_{\{\mathbf{a}\}}$ and $G = \iota_{\{\mathbf{b}\}}$, where the indicator function of a closed convex set $\mathcal{C}$ is

$$\iota_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if} \quad x \in \mathcal{C}, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.46}$$

one retrieves the hard marginal constraints defining $\mathcal{U}(\mathbf{a},\mathbf{b})$. The proof of Proposition 4.3 carries to this more general problem (4.45), so that the unique solution of (4.45) also has the form (4.12).

As shown in [Peyré, 2015, Chizat et al., 2017, Karlsson and Ringh, 2016], Sinkhorn iterations (4.15) can hence be extended to this problem, and read

$$\mathbf{u} \leftarrow \frac{\mathrm{Prox}_F^{\mathbf{KL}}(\mathbf{Kv})}{\mathbf{Kv}} \quad \text{and} \quad \mathbf{v} \leftarrow \frac{\mathrm{Prox}_G^{\mathbf{KL}}(\mathbf{K}^{\mathrm{T}}\mathbf{u})}{\mathbf{K}^{\mathrm{T}}\mathbf{u}}. \tag{4.47}$$

where the proximal operator for the KL divergence is

$$\forall\, \mathbf{u} \in \mathbb{R}_+^N, \quad \mathrm{Prox}_F^{\mathbf{KL}}(\mathbf{u}) = \operatorname*{argmin}_{\mathbf{u}' \in \mathbb{R}_+^N} \mathbf{KL}(\mathbf{u}'|\mathbf{u}) + F(\mathbf{u}'). \tag{4.48}$$

For some functions F, G it is possible to prove the linear rate of convergence for iterations (4.47), and these schemes can be generalized to arbitrary measures, see Chizat et al. [2017] for more details.

Iterations (4.47) are thus interesting in the cases where $\mathrm{Prox}_F^{\mathbf{KL}}$ and $\mathrm{Prox}_G^{\mathbf{KL}}$ can be computed in closed form or very efficiently. This is in particular the case for separable functions of the form $F(\mathbf{u}) = \sum_i F_i(\mathbf{u}_i)$ since in this case

$$\mathrm{Prox}_F^{\mathbf{KL}}(\mathbf{u}) = \left(\mathrm{Prox}_{F_i}^{\mathbf{KL}}(\mathbf{u}_i)\right)_i$$

Computing each $\mathrm{Prox}_{F_i}^{\mathbf{KL}}$ is usually simple since it is a scalar optimization problem. Note that, similarly to the initial Sinkhorn algorithm, it is also possible to stabilize the computation using log-domain computations, see Chizat et al. [2017].

This algorithm can be used to approximate the solution to various generalizations of OT, and in particular unbalanced OT problems of the form (10.7) (see §10.2 and in particular iterations (10.9)) and gradient flow problems of the form (9.25) (see §9.3).

Remark 4.23 (Duality and Legendre transform)**.** The dual problem to (4.45) reads

$$\max_{\mathbf{f},\mathbf{g}} \; -F^*(\mathbf{f}) - G^*(\mathbf{g}) - \varepsilon \sum_{i,j} e^{\frac{\mathbf{f}_i + \mathbf{g}_j - \mathbf{C}_{i,j}}{\varepsilon}}. \tag{4.49}$$

so that $(\mathbf{u}, \mathbf{v}) = (e^{\mathbf{f}/\varepsilon}, e^{\mathbf{g}/\varepsilon})$ are the associated scalings appearing in (4.12). Here, F^* and G^* are the Fenchel-Legendre conjugate,

which are convex functions defined as

$$\forall\, \mathbf{f} \in \mathbb{R}^n, \quad F^*(\mathbf{f}) \overset{\text{def.}}{=} \max_{\mathbf{a} \in \mathbb{R}^n} \langle \mathbf{f}, \mathbf{a} \rangle - F(\mathbf{a}). \tag{4.50}$$

The generalized Sinkhorn iterates (4.47) are a special case of Dykstra's algorithm [Dykstra, 1983, 1985] (extended to Bregman divergence [Bauschke and Lewis, 2000, Censor and Reich, 1998], see also Remark 8.1), and is an alternate maximization scheme on the dual problem (4.49).

The formulation (4.45) can be further generalized to more than 2 functions and more than a single coupling

$$\min_{(\mathbf{P}_s)_{s=1}^S} \sum_{i,j,s} \mathbf{C}_{s,i,j} \mathbf{P}_{s,i,j} - \varepsilon \mathbf{H}(\mathbf{P}_s) + \sum_k F_k((\mathbf{P}_s \mathbb{1}_m)_s, (\mathbf{P}_s{}^{\mathrm{T}} \mathbb{1}_n)_s)$$

where $\{\mathbf{C}_s\}_s$ are given cost matrices.

The solution has the form

$$\forall\, (s,i,j) \in [\![1,S]\!] \times [\![n]\!] \times [\![m]\!], \quad \mathbf{P}_{s,i,j} = \mathbf{K}_{s,i,j} \prod_k \mathbf{u}_{k,i} \prod_\ell \mathbf{v}_{k,j} \tag{4.51}$$

where $\mathbf{K}_s = e^{-\mathbf{C}_s/\varepsilon}$ is a Gibbs kernel, and $(\mathbf{u}_k, \mathbf{v}_\ell)_{k,\ell}$ are scaling vectors that need to be computed using Generalized Sinkhorn iterations similar to (4.47), see Chizat et al. [2017]. In the case where the functions F_s are indicator functions, one retrieves the Sinkhorn algorithm (10.2) for the multi-marginal problem, as detailed in §10.1. It is also possible to rewrite the regularized barycenter problem (9.14) this way, and the iterations (9.17) are in fact a special case of this generalized Sinkhorn.

5

Semi-discrete Optimal Transport

This chapter studies methods to tackle the optimal transport problem where one of the two input measures is discrete (a sum of Dirac masses) and the other one is arbitrary (including most importantly the case where it has a density with respect to the Lebesgue measure). When the ambiant space has low dimensions, this problem has a strong geometrical flavor because when representing transport from a continuous density towards a discrete one, one obtains that the support of the density must be split into disjoint cells, which are each mapped to one among all Dirac masses of the discrete measure. When the cost is the squared Euclidean distance, these cells corresponds to an important concept from computational geometry, the so-called Laguerre cells, which are weighted Voronoi cells. This connection allows to borrow tools from computational geometry to obtain fast computational schemes. In high dimensions, the semi-descrite formulation can also be studied as a stochastic programming problem, which can also benefit from a bit of regularization, extending therefore the scope of applications of the entropic regularization scheme presented in Chapter 4. All these constructions rely heavily on the notion of c-transform, this time for general cost functions and not only matrices as in §3.2, which is

a generalization of the Legendre transform from convex analysis, and plays a pivotal role in the theory and algorithms for OT.

5.1 c-transform and $\bar{c}$-transform

Recall that the dual OT problem (2.23) reads

$$\max_{(f,g)} \mathcal{E}(f,g) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y) + \iota_{\mathcal{R}(c)}(f,g)$$

where we used the useful indicator function notation (4.46).

Doing an alternate minimization on either f or g leads to the important notion of c-transform:

$$\forall\, y \in \mathcal{Y}, \quad f^c(y) \stackrel{\text{def.}}{=} \inf_{x\in\mathcal{X}} c(x,y) - f(x), \tag{5.1}$$

$$\forall\, x \in \mathcal{X}, \quad g^{\bar{c}}(x) \stackrel{\text{def.}}{=} \inf_{y\in\mathcal{Y}} c(x,y) - g(y), \tag{5.2}$$

where we denoted $\bar{c}(y,x) \stackrel{\text{def.}}{=} c(x,y)$. Indeed, one can check that

$$f^c \in \operatorname*{argmax}_g \mathcal{E}(f,g) \quad \text{and} \quad g^{\bar{c}} \in \operatorname*{argmax}_f \mathcal{E}(f,g). \tag{5.3}$$

Note that these partial minimizations define maximizers on the support of respectively α and β, while the definitions (5.1) actually define functions on the whole spaces $\mathcal{X}$ and $\mathcal{Y}$. This is thus a way to extend in a canonical way solutions of (2.23) on the whole spaces. When $\mathcal{X} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^p$, then the c-transform (5.1) f^c is the so-called inf-convolution between $-f$ and $\|\cdot\|^p$. The definition of f^c is also often referred to as a "Hopf-Lax formula".

The map $(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mapsto (g^{\bar{c}}, f^c) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})$ replaces dual potentials by "better" ones (improving the dual objective $\mathcal{E}$). Functions that can be written in the form f^c and $g^{\bar{c}}$ are called c-concave and $\bar{c}$-concave functions. In the special case $c(x,y) = \langle x,\, y\rangle$ in $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, this definition coincides with the usual notion of concave functions. Extending naturally Proposition 3.1 to a continuous case, one has the property that

$$f^{c\bar{c}c} = f^c \quad \text{and} \quad g^{\bar{c}c\bar{c}} = g^{\bar{c}}$$

where we denoted $f^{c\bar{c}} = (f^c)^{\bar{c}}$. This invariance property shows that one can only "improve" once the dual potential this way. Alternatively, this

means that alternate maximization does not converge (it immediately enters a cycle), which is classical for functionals involving a non-smooth (a constraint) coupling of the optimized variables. This is in sharp contrast with entropic regularization of OT as exposed in Chapter 4. In this case, because of the regularization, the dual objective (4.29) is smooth, and alternate maximization corresponds to Sinkhorn iterations (4.41) and (4.42). These iterates, written over the dual variables, define entropically-smoothed versions of the c-transform, where min operations are replaced by a "soft-min".

Using (5.3), one can reformulate (2.23) as an unconstrained convex program over a single potential

$$\mathcal{L}_c(\alpha,\beta) = \max_{f\in\mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} f^c(y)\mathrm{d}\beta(y), \tag{5.4}$$

$$= \max_{g\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} g^{\bar{c}}(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y). \tag{5.5}$$

Since one can iterate the map $(f,g)\mapsto(g^{\bar{c}},f^c)$, it is possible to add in these optimization problems the constraint that f is $\bar{c}$-concave and g is c-concave, which is important to ensure enough regularity on these potentials and show for instance existence of solutions to (2.23).

5.2 Semi-discrete Formulation

A case of particular interest is when $\beta=\sum_j \mathbf{b}_j\delta_{y_j}$ is discrete (of course the same construction applies if α is discrete by exchanging the role of (α,β)). One can adapt the definition of the $\bar{c}$ transform (5.1) to this setting by restricting the minimization to the support $(y_j)_j$ of β

$$\forall\,\mathbf{g}\in\mathbb{R}^m,\ \forall\,x\in\mathcal{X},\quad \mathbf{g}^{\bar{c}}(x) \overset{\text{def.}}{=} \min_{j\in[\![m]\!]} c(x,y_j)-\mathbf{g}_j. \tag{5.6}$$

This transform maps a vector $\mathbf{g}$ to a continuous function $\mathbf{g}^{\bar{c}}\in\mathcal{C}(\mathcal{X})$. Note that this definition coincides with (5.1) when imposing that the space $\mathcal{X}$ is equal to the support of β. Figure 5.1 shows some example of such discrete $\bar{c}$-transforms in 1-D and 2-D.

Using this discrete $\bar{c}$-transform, in this semi-discrete case, (5.4) is equivalent to the following finite dimensional optimization

$$\mathcal{L}_c(\alpha,\beta) = \max_{\mathbf{g}\in\mathbb{R}^m} \mathcal{E}(\mathbf{g}) \overset{\text{def.}}{=} \int_{\mathcal{X}} \mathbf{g}^{\bar{c}}(x)\mathrm{d}\alpha(x) + \sum \mathbf{g}_y\mathbf{b}_j \tag{5.7}$$

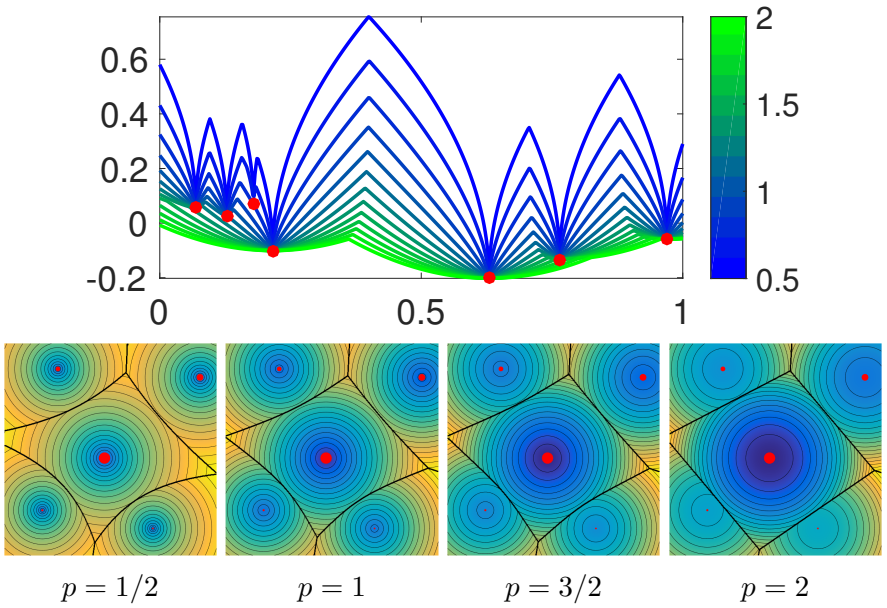


Figure 5.1: Top: examples of semi-discrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c}}$ in 1-D, for ground cost $c(x, y) = |x - y|^p$ for varying p (see colorbar). The red points are at locations $(y_j, -\mathbf{g}_j)_j$. Bottom: examples of semi-discrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c}}$ in 2-D, for ground cost $c(x, y) = \|x - y\|^p$ for varying p. The red points are at locations $y_j \in \mathbb{R}^2$, and their size is proportional to $\mathbf{g}_j$. The regions delimited by bold black curves are the Laguerre cells $(\mathbb{L}_{\mathbf{g}}(y_j))_j$ associated to these points $(y_j)_j$.

The Laguerre cells associated to the dual weights $\mathbf{g}$

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

induce a disjoint decomposition of $\mathcal{X} = \bigcup_j \mathbb{L}_{\mathbf{g}}(y_j)$. When $\mathbf{g}$ is constant, the Laguerre cells decomposition corresponds to the Voronoi diagram partition of the space. Figure 5.1, bottom row, shows examples of Laguerre cells segmentations in 2-D.

This allows one to conveniently rewrite the minimized energy as

$$\mathcal{E}(\mathbf{g}) = \sum_{j=1}^{m} \int_{\mathbb{L}_{\mathbf{g}}(y_j)} \Big(c(x,y_j) - \mathbf{g}_j \Big) \,\mathrm{d}\alpha(x) + \langle \mathbf{g}, \mathbf{b} \rangle. \tag{5.8}$$

The gradient of this function can be computed as follow

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

Once the optimal $\mathbf{g}$ is computed, then the optimal transport map T from α to β is mapping any $x \in \mathbb{L}_{\mathbf{g}}(y_j)$ toward y_j, so it is piecewise constant.

In the special case $c(x,y) = \|x-y\|^2$, the decomposition in Laguerre cells is also known as "power diagram". It can be computed efficiently using computational geometry algorithms, see [Aurenhammer, 1987]. The most widely used algorithm relies on the fact that the power diagram of points in $\mathbb{R}^d$ is equal to the projection on $\mathbb{R}^d$ of the convex hull of the set of points $((y_j, \|y_j\|^2 - \mathbf{g}_j))_{j=1}^m \subset \mathbb{R}^{d+1}$. There are numerous algorithms to compute convex hull, for instance that of Chan [1996] in 2-D and 3-D has complexity $O(P \log(Q))$ where Q is the number of vertices of the convex hull.

The initial idea of a semi-discrete solver for Monge-Ampère equations was proposed by Oliker and Prussner [1989], and its relation to the dual variational problem exposed by Aurenhammer et al. [1998]. A theoretical analysis and its application to the reflector problem in optics is detailed in [Caffarelli et al., 1999]. The semi-discrete formulation was used in [Carlier et al., 2010] in conjunction with a continuation approach based on Knothe's transport. The recent revival of this methods in various fields is due to Mérigot [2011] who proposed a

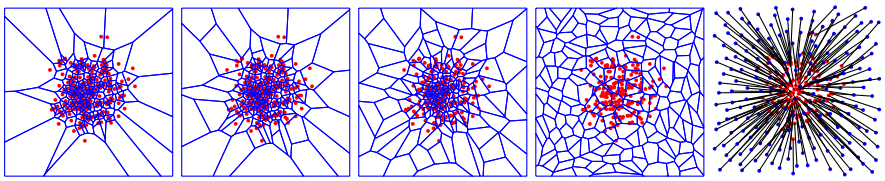


Figure 5.2: Iterations of the semi-discrete OT algorithm minimizing (5.8) (here a simple gradient descent is used). The support $(y_j)_j$ of the discrete measure β is indicated by the red points, while the continuous measure α is the uniform measure on a square. The blue cells display the Laguerre partition $(\mathbb{L}_{\mathbf{g}^{(\ell)}}(y_j))_j$ where $\mathbf{g}^{(\ell)}$ is the discrete dual potential computed at iteration ℓ.

quasi-Newton solver and clarified the link with concepts from computational geometry. We refer to [Levy and Schwindt, 2017] for a recent overview. The use of a Newton solver which is applied to sampling in computer graphics is proposed in [De Goes et al., 2012], see also [Lévy, 2015] for applications to 3-D volume and surface processing. An important area of application of semi-discrete method is for the resolution of incompressible fluid dynamic (Euler's equations) using Lagrangian methods de Goes et al. [2015], Gallouët and Mérigot [2017]. The semi-discrete OT solver enforces incompressibility at each iteration by imposing that the (possibly weighted) points cloud approximates a uniform inside the domain. The convergence (with linear rate) of damped Newton iterations is proved in Mirebeau [2015] for the Monge-Ampère equation, and is refined in Kitagawa et al. [2016] for optimal transport. Semi-discrete OT finds important applications to illumination design, see Mérigot et al. [2017].

5.3 Entropic Semi-discrete Formulation

The dual of the entropic regularized problem between arbitrary measures (4.9) is

$$\begin{aligned}\mathcal{L}_c^\varepsilon(\alpha,\beta) \stackrel{\text{def.}}{=} \max_{(f,g)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\beta(y) \\ - \varepsilon \int_{X\times Y} e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y).\end{aligned} \tag{5.9}$$

This is a smooth unconstrained optimization problem.

Similarly to the unregularized problem (5.1), one can minimize explicitly with respect to f or g in (5.9), and introduce a smoothed c-transform

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

In the case of a discrete measure $\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$, one can restrict its attention to discrete dual potential $\mathbf{g} \in \mathbb{R}^m$, and this corresponds to

$$\forall\, x \in \mathcal{X}, \quad \mathbf{g}^{\bar{c},\varepsilon}(x) \overset{\text{def.}}{=} -\varepsilon \log\left(\sum_{j=1}^m e^{\frac{-c(x,y_j)+\mathbf{g}_j}{\varepsilon}} \mathbf{b}_j\right). \tag{5.10}$$

One defines similarly $\mathbf{f}^{\bar{c},\varepsilon}$ in the case of a discrete measure α. Note that the rewriting (4.38) and (4.39) of Sinkhorn using the soft-min operator $\min_\varepsilon$ corresponds to the alternate computation of entropic smoothed c-transforms

$$\mathbf{f}_i^{(\ell+1)} = \mathbf{g}^{\bar{c},\varepsilon}(x_i) \quad \text{and} \quad \mathbf{g}_j^{(\ell+1)} = \mathbf{f}^{c,\varepsilon}(y_j). \tag{5.11}$$

Instead of maximizing (5.9), one can thus solve the following finite dimensional optimization problem

$$\max_{\mathbf{g}\in\mathbb{R}^n} \mathcal{E}^\varepsilon(\mathbf{g}) \overset{\text{def.}}{=} \int_{\mathcal{X}} \mathbf{g}^{\bar{c},\varepsilon}(x)\mathrm{d}\alpha(x) - \langle \mathbf{g}, \mathbf{b}\rangle. \tag{5.12}$$

Note that this optimization problem is still valid even in the unregularized case $\varepsilon = 0$ and in this case $\mathbf{g}^{\bar{c},\varepsilon=0} = \mathbf{g}^{\bar{c}}$ is the $\bar{c}$-transform defined in (5.6) so that (5.12) is in fact (5.8). The gradient of this functional reads

$$\forall\, j \in [\![m]\!], \quad \nabla\mathcal{E}^\varepsilon(\mathbf{g})_j = -\int_{\mathcal{X}} \chi_j^\varepsilon(x)\mathrm{d}\alpha(x) + \mathbf{b}_j, \tag{5.13}$$

where χ_j^ε is a smoothed version of the indicator χ_j^0 of the Laguerre cell $\mathbb{L}_{\mathbf{g}}(y_j)$

$$\chi_j^\varepsilon(x) = \frac{e^{\frac{-c(x,y_j)+\mathbf{g}_j}{\varepsilon}}}{\sum_\ell e^{\frac{-c(x,y_\ell)+\mathbf{g}_\ell}{\varepsilon}}}.$$

Note once again that this formula (5.13) is still valid for $\varepsilon = 0$. Note also that the family of functions $(\chi_j^\varepsilon)_j$ is a partition of unity, *i.e.* $\sum_j \chi_j^\varepsilon = 1$ and $\chi_j^\varepsilon \geq 0$. Figure 5.3, bottom row, illustrates this.

Remark 5.1 (Second order methods and connection with logistic regression)**.** A crucial aspect of the smoothed semi-dual formulation (5.12) is that it corresponds to the minimization of a smooth function. Indeed, as shown in [Genevay et al., 2016], the Hessian of $\mathcal{E}^\varepsilon$ is upper-bounded by $1/\varepsilon$, so that $\nabla\mathcal{E}^\varepsilon$ is $\frac{1}{\varepsilon}$-Lipschitz continuous. In fact, that problem is very closely related to a multi-class logistic regression problem (see Figure 5.3 for a display of the resulting fuzzy classification boundary) and enjoys the same favorable properties, see [Hosmer Jr et al., 2013], which are generalizations of self-concordance, see [Bach, 2010]. In particular, Newton method converges quadratically, and one can use in practice quasi-Newton techniques, such as L-BFGS, as advocated in [Cuturi and Peyré, 2016]. Note that [Cuturi and Peyré, 2016] studies the more general barycenter problem detailed in §9.2, but is equivalent to this semi-discrete setting when considering only a pair of inputs measures. Note that the use of second order schemes (Newton or L-BFGS) is also advocated in the unregularized case $\varepsilon = 0$ by [Mérigot, 2011, De Goes et al., 2012, Lévy, 2015]. In [Kitagawa et al., 2016, Theorem 5.1], the Hessian of $\mathcal{E}^0(\mathbf{g})$ is shown to be uniformly bounded as long as the volume of the Laguerre cells are bounded by below and α has a continuous density. This allows these authors to show the linear convergence of a damped Newton algorithm with a backtracking to ensure that the Laguerre cells never vanish during the optimization. This result justifies the use of second order methods even in the unregularized case. The intuition is that, while the conditioning of the entropic regularized problem scales like $1/\varepsilon$, when $\varepsilon = 0$, this conditioning is rather driven by m, the number of samples of the discrete distribution (which controls the size of the Laguerre cells). One can also see [Knight and Ruiz, 2013, Sugiyama et al., 2017, Cohen et al., 2017, Allen-Zhu et al., 2017] for alternate methods using second order schemes.

Remark 5.2 (Legendre transforms of OT cost functions)**.** As stated in Proposition 4.6, $\mathrm{L}_{\mathbf{C}}^\varepsilon(\mathbf{a}, \mathbf{b})$ is a convex function of $(\mathbf{a}, \mathbf{b})$ (which is also

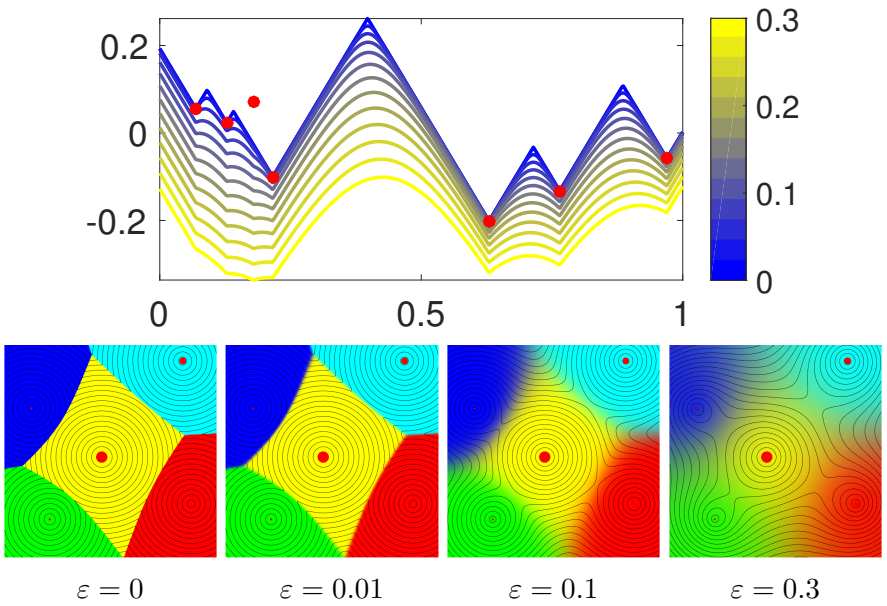


Figure 5.3: Top: examples of entropic semi-discrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c},\varepsilon}$ in 1-D, for ground cost $c(x,y) = |x-y|$ for varying ε (see colorbar). The red points are at locations $(y_j, -\mathbf{g}_j)_j$. Bottom: examples of entropic semi-discrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c},\varepsilon}$ in 2-D, for ground cost $c(x,y) = \|x-y\|$ for varying ε. The black curves are the level sets of the function $\mathbf{g}^{\bar{c},\varepsilon}$, while the colors indicate the smoothed indicator function of the Laguerre cells χ_j^ε. The red points are at locations $y_j \in \mathbb{R}^2$, and their size is proportional to $\mathbf{g}_j$.

true in the unregularized case $\varepsilon = 0$). It is thus possible to compute its Legendre-Fenchel transform, which is defined in (4.50). Denoting $F_{\mathbf{a}}(\mathbf{b}) = \mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$, one has, for a fixed $\mathbf{a}$, following Cuturi and Peyré [2016]

$$F^*_{\mathbf{a}}(\mathbf{g}) = -\varepsilon H(\mathbf{a}) + \sum_i \mathbf{a}_i \mathbf{g}^{\bar{c},\varepsilon}(x_i)$$

Here $\mathbf{g}^{\bar{c},\varepsilon}$ is the entropic-smoothed c-transform introduced in (5.10). In the unregularized case $\varepsilon = 0$, and for generic measures, following Carlier et al. [2015], one has denoting $\mathcal{F}_\beta(\alpha) \overset{\text{def.}}{=} \mathcal{L}_c(\alpha, \beta)$

$$\forall\, f \in \mathcal{C}(\mathcal{X}), \quad \mathcal{F}^*_\beta(f) = \int_{\mathcal{Y}} f^c(y) \mathrm{d}\beta(y),$$

where the c-transform $f^c \in \mathcal{C}(\mathcal{Y})$ of f is defined in §5.1. Note that here, since $\mathcal{M}(\mathcal{X})$ is in duality with $\mathcal{C}(\mathcal{X})$, the Legendre transform is a function of continuous functions. Denoting now $G(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \mathrm{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$, one has, following Cuturi and Peyré [2016]

$$\forall\, (\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m, \quad G^*(\mathbf{f}, \mathbf{g}) = -\varepsilon \log \sum_{i,j} e^{\frac{-\mathbf{C}_{i,j} + \mathbf{f}_i + \mathbf{g}_j}{\varepsilon}}$$

which can be seen as a smoothed version of the Legendre transform of $\mathcal{G}(\alpha, \beta) \overset{\text{def.}}{=} \mathcal{L}_c(\alpha, \beta)$

$$\forall\, (f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}), \quad \mathcal{G}^*(f, g) = \inf_{(x,y) \in \mathcal{X} \times \mathcal{Y}} c(x, y) - f(x) - g(y).$$

5.4 Stochastic Optimization Methods

The semi-discrete formulation (5.8) and its smoothed version (5.12) are appealing because the energies to be minimized are written as an expectation with respect to the probability distribution α

$$\mathcal{E}^{\varepsilon}(\mathbf{g}) = \int_{\mathcal{X}} E^{\varepsilon}(\mathbf{g}, x) \mathrm{d}\alpha(x) = \mathbb{E}_X(E^{\varepsilon}(\mathbf{g}, X))$$

$$\text{where} \quad E^{\varepsilon}(\mathbf{g}, x) \overset{\text{def.}}{=} \mathbf{g}^{\bar{c},\varepsilon}(x) - \langle \mathbf{g},\, \mathbf{b} \rangle$$

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

$$\nabla_{\mathbf{g}} E^{\varepsilon}(x, \mathbf{g}) = (\chi^{\varepsilon}_j(x) - \mathbf{b}_j)_{j=1}^m \in \mathbb{R}^m.$$

One can thus use stochastic optimization methods to perform the maximization, as initially proposed in Genevay et al. [2016]. This allows to obtain provably convergent algorithms without the need to resort to some discretization of α as usually done (either approximating α using sums of Diracs or using quadrature formula for the integrals). The measure α is used as a black box from which one can draw independent samples, which is a natural computational setup for many high dimensional applications in statistics and machine learning. This class of methods has been generalized to the computation of Wasserstein barycenters (as described in Section (9.2)) in Staib et al. [2017].

Stochastic gradient descent (SGD). Initializing $\mathbf{g}^{(0)} = \mathbb{0}_P$, the stochastic gradient descent (SGD – used here as a maximization method), at step ℓ, draws a point $x_\ell \in \mathcal{X}$ (and all the $(x_\ell)_\ell$ are independent) according to distribution α and then update

$$\mathbf{g}^{(\ell+1)} \overset{\text{def.}}{=} \mathbf{g}^{(\ell)} + \tau_\ell \nabla_{\mathbf{g}} E^\varepsilon(\mathbf{g}^{(\ell)}, x_\ell). \tag{5.14}$$

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

$$\tau_\ell \overset{\text{def.}}{=} \frac{\tau_0}{1 + \ell/\ell_0} \tag{5.15}$$

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

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

where $\mathbf{g}^\star$ is a solution of (5.12) and where $\mathbb{E}$ indicates an expectation with respect to the i.i.d. sampling of $(x_\ell)_\ell$ performed at each iteration. Figure 5.4 shows the evolution of the algorithm on a simple 2-D example, where α is the uniform distribution on $[0,1]^2$.

Stochastic Gradient Descent with Averaging (SGA). Stochastic gradient descent is slow because of the fast decay of τ_ℓ toward zero.

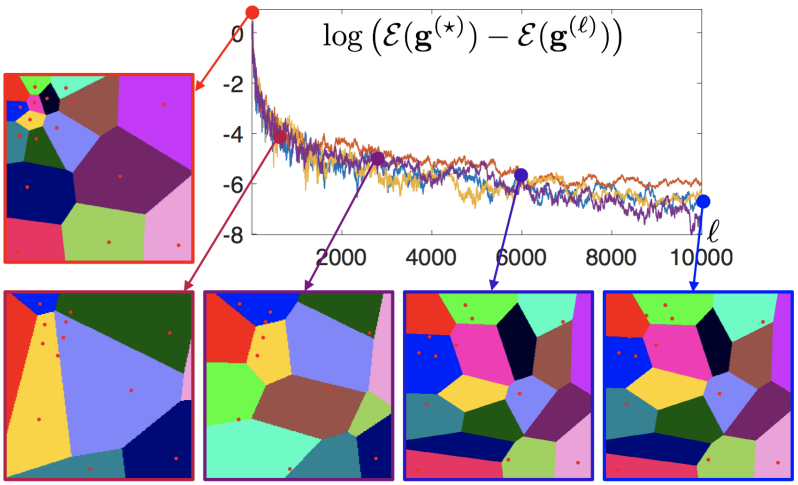


Figure 5.4: Evolution of the energy $\mathcal{E}^\varepsilon(\mathbf{g}^{(\ell)})$, for $\varepsilon = 0$ (no regularization) during the SGD iterations (5.14). Each colored curve show a different randomized run. The images display the evolution of the Laguerre cells $(\mathbb{L}_{\mathbf{g}^{(\ell)}}(y_j))_j$ through the iterations.

To improve somehow the convergence speed, it is possible to average the past iterates, *i.e.* run a "classical" SGD on auxiliary variables $(\tilde{\mathbf{g}}^{(\ell)})_\ell$

$$\tilde{\mathbf{g}}^{(\ell+1)} \overset{\text{def.}}{=} \tilde{\mathbf{g}}^{(\ell)} + \tau_\ell \nabla_{\mathbf{g}} E^\varepsilon(\tilde{\mathbf{g}}^{(\ell)}, x_\ell)$$

where x_ℓ is drawn according to α (and all the $(x_\ell)_\ell$ are independent) and output as estimated weight vector the average

$$\mathbf{g}^{(\ell)} \overset{\text{def.}}{=} \frac{1}{\ell} \sum_{k=1}^{\ell} \tilde{\mathbf{g}}^{(k)}.$$

This defines the Stochastic Gradient Descent with Averaging (SGA) algorithm. Note that it is possible to avoid explicitly storing all the iterates by simply updating a running average as follows

$$\mathbf{g}^{(\ell+1)} = \frac{1}{\ell+1} \tilde{\mathbf{g}}^{(\ell+1)} + \frac{\ell}{\ell+1} \mathbf{g}^{(\ell)}.$$

In this case, a typical choice of decay is rather of the form

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

Notice that the step size now goes much slower to 0 than for (5.15), at rate $\ell^{-1/2}$. Bach [2014] proves that SGA leads to a faster convergence (the constant involved are smaller) than SGD, since on contrast to SGD, SGA is adaptive to the local strong convexity (or concavity for maximization problems) of the functional.

6

$\mathcal{W}_1$ Optimal Transport

This chapter focuses on the setting of optimal transport in which the ground cost is equal to a distance. Historically, this corresponds to the original problem posed by Monge in 1781; it also appears in some of the earliest applications of optimal transport as "earth mover's distances" in computer vision [Rubner et al., 2000].

Unlike the case where transportation cost equals a distance *squared* (studied in particular in Chapter 7), transport with metric costs is more difficult to analyze theoretically. Contrasting with Remark 2.23 for transport with squared ground distances, generally there is no unique optimal Kantorovich coupling 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. We refer to works by Trudinger and Wang [2001], Caffarelli et al. [2002], Sudakov [1979] for proofs of existence of optimal $\mathcal{W}_1$ transportation plans and detailed analyses of their geometric structure.

Nevertheless, transport with linear ground distance is useful to compare histograms, since a non-squared loss is more robust to outliers in noisy data than a quadratic cost. Furthermore, this problem admits an elegant dual reformulation involving local flow or divergence

constraints, 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 Sections 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 inequality next from the identity $|\inf f - \inf g| \leq \sup |f - g|$, the las one 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,

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

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

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

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

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

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

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 $\int_X \alpha = \int_X \beta$. This norm is often called the Kantorovich and Rubinshtein 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 (2.35).

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 $\left\{ f \;:\; \mathrm{Lip}_p(f) \leq 1 \right\}$ 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 Space

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,

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

Here the constraint $\|\nabla f\|_\infty \leq 1$ signifies that the norm of the gradient of f at any point x is upperbounded 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

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

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, problem (6.3) and (6.4) become non-smooth convex optimization problems. It is possible to use off-the-shelf interior points quadratic-cone optimization solver, but as advocated in §7.3, large scale problem requires the use of simpler but more adapted first order method. One can thus use for instance DR iterations (7.14) or the related ADMM method. Note that on 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. [2016] for similar techniques.

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 of $\mathcal{X}$ being a Riemannian manifold. In the discrete setting, this corresponds to graphs, where $\mathcal{X} = [\![1,n]\!]$ is a finite set of indexes, and $(i,j) \in \mathcal{E} \subset \mathcal{X}^2$ is an edge (here the graph is assumed to be undirected) equipped with some weight (length) $\mathbf{w}_{i,j}$. In this case, the geodesic distance is

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

where $i \to j$ indicates that $i_0 = i$ and $i_K = j$.

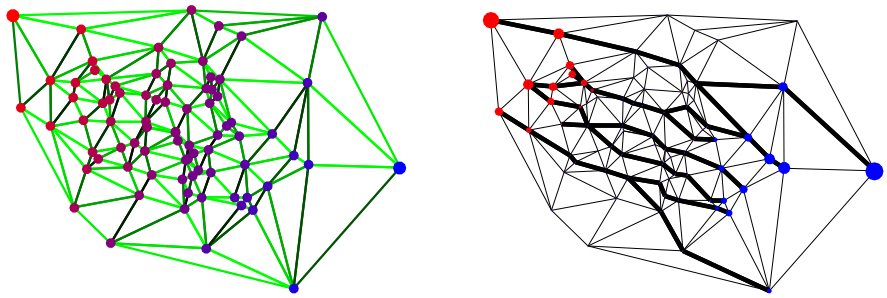
We consider for two vectors $(\mathbf{a},\mathbf{b}) \in (\mathbb{R}^n)^2$ defining discrete probabilities on the graph $\mathcal{X}$, such that $\sum_i \mathbf{a}_i = \sum_i \mathbf{b}_i$ (they do not even need to positive). The goal is now to compute $\mathrm{W}_1(\mathbf{a},\mathbf{b})$, as introduced in (2.16) for $p=1$, when the ground metric is such a geodesic distance. The goal is to achieve this without resorting to the computation of a "full" coupling $\mathbf{P}$ of size $n \times n$, and instead rely on local operators thanks to the underlying connectivity of the graph. These operators are discrete equivalent of the gradient and divergence differential operators.

A discrete dual Kantorovich potential $\mathbf{f} \in \mathbb{R}^n$ is defined on the 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

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



$\mathbf{f}$ $(\mathbf{a}, \mathbf{b})$ and $\mathbf{s}$

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 non-zero $\mathbf{s}_{i,j}$, which saturate to $\mathbf{w}_{i,j} = 1$. These saturating flow edge on the right match the light green edge on the left where $|(\nabla \mathbf{f})_{i,j}| = 1$.

The analogous of formula (6.3), in the graph setting, reads

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

The associated dual formula, which is the analogous of Formula (6.4), now becomes in this setting,

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

This is a linear program, which is a typical instance of min-cost flow problem. 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 model, see for instance Carlier et al. [2008].

7

Dynamic Formulations

This chapter presents the geodesic (also called dynamic) point of view of optimal transport when the cost is a squared geodesic distance. This describes the optimal transport between two measures as a curve in the space of measures minimizing a total length. The dynamic point of view offers an alternate and intuitive interpretation of optimal transport, which not only allows to draw links with fluid dynamics, but also results in an efficient numerical tool to compute OT in small dimensions when interpolating between two densities. The drawback of that approach is that it cannot scale to large-scale sparse measures, and only works in low dimensions on regular domains (because one needs to grid the space) with a squared geodesic cost.

In this chapter, we use the notation (α_0, α_1) in place of (α, β) in agreement with the idea that we start from one measure to reach another one.

7.1 Continuous Formulation

In the case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, and $c(x, y) = \|x - y\|^2$ the optimal transport distance $\mathcal{W}_2^2(\alpha, \beta) = \mathcal{L}_c(\alpha, \beta)$ as defined in (2.15) can be computed by

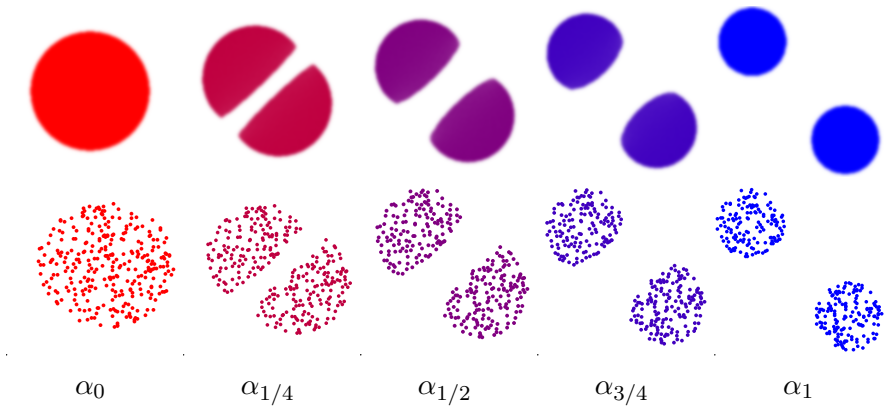


Figure 7.1: Displacement interpolation α_t satisfying (7.2). Top: for two measures (α_0, α_1) with densities with respect to the Lebesgue measure. Bottom: for two discrete empirical measures with the same number of points (bottom).

looking for a minimal length path $(\alpha_t)_{t=0}^1$ between these two measures. This path is described by advecting the measure using a vector field v_t defined at each instant. The vector field v_t and the path α_t must satisfy the conservation of mass formula, resulting in

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(\alpha_t v_t) = 0 \quad \text{and} \quad \alpha_{t=0} = \alpha_0, \alpha_{t=1} = \alpha_1 \tag{7.1}$$

where the equation above should be understood in the sense of distributions on $\mathbb{R}^d$. The infinitesimal length of such a vector field is measured using the L^2 norm associated to the measure α_t, that is defined as

$$\|v_t\|_{L^2(\alpha_t)} = \Big(\int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x)\Big)^{1/2}.$$

This definition leads to the following minimal-path reformulation of $\mathcal{W}_2$, original introduced by Benamou and Brenier [2000]

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \min_{(\alpha_t, v_t)_t \text{ sat. } (7.1)} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x)\mathrm{d}t \tag{7.2}$$

where α_t is a scalar-valued measure and v_t a vector-valued measure. Figure 7.1 shows two examples of such paths of measures.

The formulation (7.2) is a non-convex formulation in the variables $(\alpha_t, v_t)_t$ because of the constraint (7.1) involving the product $\alpha_t v_t$. In-

troducing a vector-valued measure (often called the "momentum")

$$J_t \stackrel{\text{def.}}{=} \alpha_t v_t,$$

Benamou and Brenier showed in their landmark paper [2000] that it is instead convex in the variable $(\alpha_t, J_t)_t$ when writing

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \min_{(\alpha_t, J_t)_t \in \mathcal{C}(\alpha_0, \alpha_1)} \int_0^1 \int_{\mathbb{R}^d} \theta(\alpha_t(x), J_t(x)) \mathrm{d}x \mathrm{d}t \tag{7.3}$$

where we define the set of constraints as

$$\mathcal{C}(\alpha_0, \alpha_1) \stackrel{\text{def.}}{=} \left\{ (\alpha_t, J_t) \;:\; \frac{\partial \alpha_t}{\partial t} + \mathrm{div}(J_t) = 0, \alpha_{t=0} = \alpha_0, \alpha_{t=1} = \alpha_1 \right\}, \tag{7.4}$$

and where $\theta :\to \mathbb{R}^+ \cup \{+\infty\}$ is the following lower-semi continuous convex function

$$\forall\, (a,b) \in \mathbb{R}_+ \times \mathbb{R}^d, \quad \theta(a,b) = \begin{cases} \frac{\|b\|^2}{a} & \text{if} \quad a > 0, \\ 0 \quad \text{if} & (a,b) = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{7.5}$$

This definition might seem complicated, but it is crucial to impose that the momentum $J_t(x)$ should vanish when $\alpha_t(x) = 0$. Note also that (7.3) is written in an informal way as if the measures (α_t, J_t) were density functions, but this is acceptable because θ is a 1-homogeneous function, which can thus be extended in an unambiguous way from density to functions.

Remark 7.1 (Links with McCann's Interpolation)**.** In the case (see Equation (2.26)) where there exists an optimal Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$ with $T_\sharp \alpha_0 = \alpha_1$, then α_t is equal to McCann's interpolation

$$\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha_0. \tag{7.6}$$

In the 1-D case, using Remark 2.28, this interpolation can be computed thanks to the relation

$$\mathcal{C}_{\alpha_t}^{-1} = (1-t)\mathcal{C}_{\alpha_0}^{-1} + t\mathcal{C}_{\alpha_1}^{-1}, \tag{7.7}$$

see Figure 2.10. We refer to Gangbo and McCann [1996] for a detailed review on the Riemannian geometry of the Wasserstein

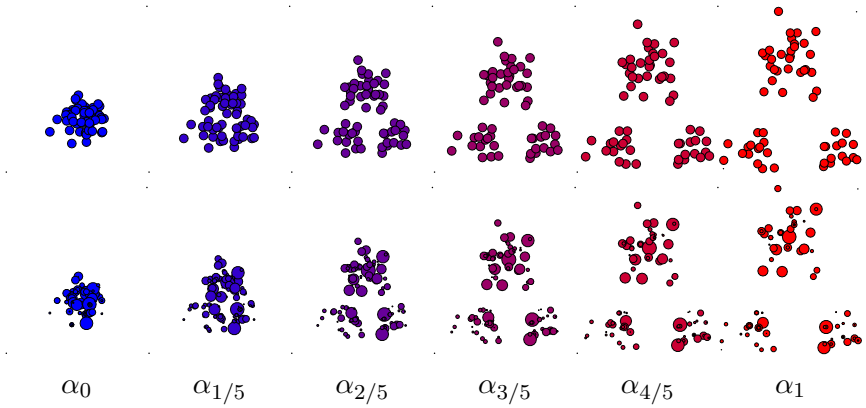


Figure 7.2: Comparison of displacement interpolation (7.8) of discrete measures. Top: point clouds (empirical measures (α_0, α_1) with the same number of points). Bottom: same but with weight. For $0 < t < 1$, the top example corresponds to an empirical measure interpolation α_t with N points, while the bottom one defines a measure supported on $2N - 1$ points.

space. In the case that there is only a coupling π (not necessarily supported on a Monge map), one can compute this interpolant as

$$\alpha_t = P_{t\sharp}\gamma \quad \text{where} \quad P_t : (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto (1-t)x + ty. \tag{7.8}$$

For instance, in the discrete setup (2.3), denoting $\mathbf{P}$ a solution to (2.11), an interpolation is defined as

$$\alpha_t = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(1-t)x_i + ty_j}. \tag{7.9}$$

Such an interpolation is typically supported on $n + m - 1$ points, which is the maximum number of nonzero elements of $\mathbf{P}$. Figure 7.2 shows two example of such displacement interpolation of discrete measures. This construction can be generalized to geodesic spaces $\mathcal{X}$ by replacing P_t by the interpolation along geodesic paths. McCann interpolation finds many applications, for instance color, shape and illumination interpolations in computer graphics [Bonneel et al., 2011].

7.2 Discretization on Uniform Staggered Grids

For simplicity, we describe the numerical scheme in dimension $d = 2$, the extension to higher dimensions is straightforward. We follow the discretization method introduced by Papadakis et al. [2014], which is inspired by staggered grid techniques which are commonly used in fluid dynamics. We discretize time as $t_k = k/T \in [0, 1]$, and assume the space is uniformly discretized at points $x_i = (i_1/n_1, i_2/n_2) \in X = [0, 1]^2$. We use a staggered grid representation, so that α_t is represented using $\mathbf{a} \in \mathbb{R}^{(T+1)\times n_1\times n_2}$ associated to half grid points in time, whereas J is represented using $\mathbf{J} = (\mathbf{J}_1, \mathbf{J}_2)$ where $\mathbf{J}_1 \in \mathbb{R}^{T\times(n_1+1)\times n_2}$ and $\mathbf{J}_1 \in \mathbb{R}^{T\times n_1\times(n_2+1)}$ are stored at half grid point in each space direction. Using this representation, for $(k, i_1, i_2) \in [\![1, T]\!] \times [\![1, n_1]\!] \times [\![1, n_2]\!]$, time derivative is computed as

$$(\partial_t \mathbf{a})_{k,i} \overset{\text{def.}}{=} \mathbf{a}_{k+1,i} - \mathbf{a}_{k,i}$$

and spatial divergence as

$$\operatorname{div}(\mathbf{J})_{k,i} \overset{\text{def.}}{=} \mathbf{J}^1_{k,i_1+1,i_2} - \mathbf{J}^1_{k,i_1,i_2} + \mathbf{J}^2_{k,i_1,i_2+1} - \mathbf{J}^2_{k,i_1,i_2} \tag{7.10}$$

which d are both defined at grid points, thus forming arrays of $\mathbb{R}^{T\times n_1\times n_2}$.

In order to evaluate the functional to be optimized, one needs interpolation operators from mid-grid points to grid points, for all $(k, i_1, i_2) \in [\![1, T]\!] \times [\![1, n_1]\!] \times [\![1, n_2]\!]$,

$$\mathcal{I}_a(\mathbf{a})_{k,i} \overset{\text{def.}}{=} \mathcal{I}(\mathbf{a}_{k+1,i}, \mathbf{a}_{k,i})$$

$$\mathcal{I}_J(\mathbf{J})_{k,i} \overset{\text{def.}}{=} (\mathcal{I}(\mathbf{J}^1_{k,i_1+1,i_2}, \mathbf{J}^1_{k,i_1,i_2}), \mathcal{I}(\mathbf{J}^2_{k,i_1,i_2+1}, \mathbf{J}^2_{k,i_1,i_2})).$$

The simplest choice is to use a linear operator $\mathcal{I}(r, s) = \frac{r+s}{2}$, which is the one we consider next.

The discrete counterpart to (7.3) reads

$$\min_{(\mathbf{a},\mathbf{J})\in\mathbf{C}(\mathbf{a}_0,\mathbf{a}_1)} \Theta(\mathcal{I}_a(\mathbf{a}), \mathcal{I}_J(\mathbf{J})) \tag{7.11}$$

$$\text{where} \quad \Theta(\tilde{\mathbf{a}}, \tilde{\mathbf{J}}) \overset{\text{def.}}{=} \sum_{k=1}^{T} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \theta(\tilde{\mathbf{a}}_{k,i}, \tilde{\mathbf{J}}_{k,i}),$$

and where the constraint now reads

$$\mathbf{C}(\mathbf{a}_0,\mathbf{a}_1) \stackrel{\text{def.}}{=} \{(\mathbf{a},\mathbf{J}) \;:\; \partial_t \mathbf{a} + \mathrm{div}(\mathbf{J}) = 0, (\mathbf{a}_{0,\cdot},\mathbf{a}_{T,\cdot}) = (\mathbf{a}_0,\mathbf{a}_1)\}\,.$$

where $\mathbf{a} \in \mathbb{R}^{(T+1)\times n_1 \times n_2}$, $\mathbf{J} = (\mathbf{J}_1,\mathbf{J}_2)$ with $\mathbf{J}_1 \in \mathbb{R}^{T\times(n_1+1)\times n_2}$, $\mathbf{J}_2 \in \mathbb{R}^{T\times n_1\times(n_2+1)}$. Figure 7.3 shows an example of evolution $(\alpha_t)_t$ approximated using this discretization scheme.

Remark 7.2 (Dynamic Formulation on Graphs)**.** In the case where $\mathcal{X}$ is a graph and $c(x,y) = d_{\mathcal{X}}(x,y)^2$, it is possible to derive faithful discretization methods, which uses a discrete divergence associated to the graph structure in place of the uniform grid discretization (7.10). In order to be ensure that the heat equation has a gradient flow structure (see §9.3 for more details about gradient flows) for the corresponding dynamic Wasserstein distance Maas [2011], Erbar et al. [2017] propose to use a logarithmic mean $\mathcal{I}(r,s)$ (see also Solomon et al. [2016b]).

7.3 Proximal Solvers

The discretized dynamic OT problem (7.11) is challenging to solve because it requires to minimize a non-smooth optimization problem under affine constraints. Indeed, the function θ is convex but non-smooth for measure with vanishing mass $\mathbf{a}_{k,i}$. When interpolating between two compactly supported input $(\mathbf{a}_0,\mathbf{a}_1)$, one typically expects the mass of the interpolated measures $(\mathbf{a}_k)_{k=1}^T$ to vanish also, and the difficult part of the optimization process is indeed to track this evolution of the support. In particular, it is not possible to use standard smooth optimization techniques.

There are several ways to recast (7.11) into a quadratic-cone program, either by considering the dual problem, or simply by replacing the functional $\theta(\mathbf{a}_{k,i},\mathbf{J}_{k,i})$ by a linear function under constraints

$$\Theta(\tilde{\mathbf{a}},\tilde{\mathbf{J}}) = \min_{\tilde{\mathbf{z}}} \left\{ \sum_{k,i} \tilde{\mathbf{z}}_{k,i} \;:\; \forall\,(k,i), (\mathbf{z}_{k,i},\tilde{\mathbf{a}}_{k,i},\tilde{\mathbf{J}}_{i,j}) \in \mathcal{L} \right\},$$

which thus requires the introduction of an extra variable $\tilde{\mathbf{z}}$. Here $\mathcal{L} \stackrel{\text{def.}}{=} \left\{(z,a,J) \in \mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^d \;:\; \|J\|^2 \leq za\right\}$ is a rotated Lorentz quadratic-cone. With this extra variable, it is thus possible to solve the discretized

problem using standard interior point solvers for quadratic cone programs Nesterov and Nemirovskii [1994]. These solvers have fast convergence rates, and are thus capable of computing solution with very high precision. Unfortunately, each iteration is very costly and requires the resolution of a linear system of dimension scaling with the number of discretization points. They are thus not applicable for large scale multi-dimensional problems encountered in imaging applications.

An alternative to these high-precision solvers are low-precision first order methods, which are well-suited for non-smooth but highly structured problems such as (7.11). While this class of solvers is not new, it has recently been revitalized in the fields of imaging and machine learning because they are the perfect fit for these application, where numerical precision is not the driving goal. We refer for instance to the monographs Bauschke and Combettes [2011] for detailed account on these solvers and their use for large scale applications. We here concentrate on a specific solver, but of course many more can be used, and we refer to Papadakis et al. [2014] for a study of several such approaches for dynamical OT. Note that the idea of using first order scheme for dynamical OT was initially proposed in Benamou and Brenier [2000].

The Douglas-Rachford (DR) algorithm Lions and Mercier [1979] is specifically tailored to solve non-smooth structured problems of the form

$$\min_{x\in\mathcal{H}} F(x) + G(x) \tag{7.12}$$

where $\mathcal{H}$ is some Euclidean space, and where $F, G : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are two closed convex functions, for which one can "easily " (*i.e.* in closed form or using a rapidly converging scheme) compute the so-called proximal operator

$$\forall\, x \in \mathcal{H}, \quad \mathrm{Prox}_{\tau F}(x) \overset{\text{def.}}{=} \operatorname*{argmin}_{x'\in\mathcal{H}} \frac{1}{2}\|x - x'\|^2 + \tau F(x) \tag{7.13}$$

for a parameter $\tau > 0$. Note that this corresponds to the proximal map for the Euclidean metric, and that this definition can be extended to more general Bregman divergence in place of $\|x - x'\|^2$, see (4.48) for an example using the **KL** divergence. The iterations of the DR algorithm

define a sequence $(x^{(\ell)}, w^{(\ell)}) \in \mathcal{H}^2$ using a initial $(x^{(0)}, w^{(0)}) \in \mathcal{H}^2$ and

$$\begin{aligned} w^{(\ell+1)} &\overset{\text{def.}}{=} w^{(\ell)} + \alpha(\mathrm{Prox}_{\gamma F}(2x^{(\ell)} - w^{(\ell)}) - x^{(\ell)}), \\ x^{(\ell+1)} &\overset{\text{def.}}{=} \mathrm{Prox}_{\gamma G}(w^{(\ell+1)}). \end{aligned} \tag{7.14}$$

If $0 < \alpha < 2$ and $\gamma > 0$, one can show that $x^{(\ell)} \to z^\star$ a solution of (7.12), see Combettes and Pesquet [2007] for more details. This algorithm is closely related to another popular method, the Alternating Method of Multiplier Gabay and Mercier [1976], Glowinski and Marroco [1975] (see also Boyd et al. [2011] for a review), which can be retrieved by applying DR on a dual problem, see Papadakis et al. [2014] for more details on the equivalence between the two, initially proved by Eckstein and Bertsekas [1992].

There are many way to re-cast problem (7.11) in the form (7.12), and we refer to Papadakis et al. [2014] for an exploration of several possibility. A simple way to achieve this is by setting $x = (\mathbf{a}, \mathbf{J}, \tilde{\mathbf{a}}, \tilde{\mathbf{J}})$, and letting

$$F(x) \overset{\text{def.}}{=} \Theta(\tilde{\mathbf{a}}, \tilde{\mathbf{J}}) + \iota_{\mathbf{C}(\mathbf{a}_0, \mathbf{a}_1)}(\mathbf{a}, \mathbf{J}) \quad \text{and} \quad G(x) = \iota_{\mathcal{D}}(\mathbf{a}, \mathbf{J}, \tilde{\mathbf{a}}, \tilde{\mathbf{J}})$$

$$\text{where} \quad \mathcal{D} \overset{\text{def.}}{=} \left\{ (\mathbf{a}, \mathbf{J}, \tilde{\mathbf{a}}, \tilde{\mathbf{J}}) \;:\; \tilde{\mathbf{a}} = \mathcal{I}_a(\mathbf{a}), \tilde{\mathbf{J}} = \mathcal{I}_J(\mathbf{J}) \right\}.$$

The proximal operator of these two functions can be computed efficiently. Indeed,

$$\mathrm{Prox}_{\tau F}(x) = (\mathrm{Prox}_{\tau\Theta}(\tilde{\mathbf{a}}, \tilde{\mathbf{J}}), \mathrm{Proj}_{\mathbf{C}(\mathbf{a}_0, \mathbf{a}_1)}(\mathbf{a}, \mathbf{J})).$$

The proximal operator $\mathrm{Prox}_{\tau\Theta}$ is computed by solving a cubic polynomial equation at each grid position. The orthogonal projection on the affine constraint $\mathbf{C}(\mathbf{a}_0, \mathbf{a}_1)$ involves the resolution of a Poisson equation, which can be achieved in $O(N \log(N))$ operations using the Fast Fourier Transform, where $N = Tn_1n_2$ is the number of grid points. Lastly, the proximal operator $\mathrm{Prox}_{\tau G}$ is a linear projector, which requires the inversion of a small linear system. We refer to Papadakis et al. [2014] for more details on these computations. Figure (7.3) shows an example of application of this method to compute a dynamical interpolation inside a complicated planar domain. This class of proximal method for dynamical OT has also been used to solve related problem such as mean field games Benamou and Carlier [2015].

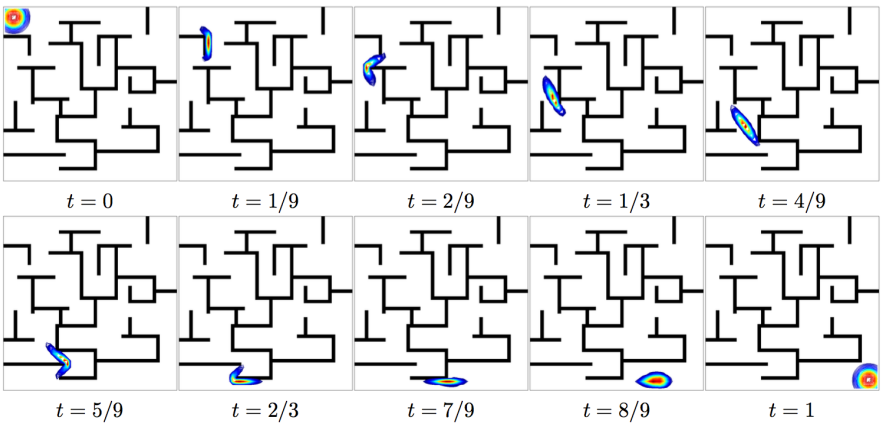


Figure 7.3: Solution α_t of dynamic OT computed with a proximal splitting scheme.

7.4 Dynamical Unbalanced OT

In order to be able to match input measure with different mass $\alpha_0(\mathcal{X}) \neq \alpha_1(\mathcal{X})$ (the so-called "unbalanced" settings, the terminology introduced in Benamou [2003]), and also to cope with local mass variation. Several normalizations or relaxations have been proposed, in particular by relaxing the fixed marginal constraint, see §10.2.

A general methodology consists in introducing a source term $s_t(x)$ in the continuity equation (7.4). We thus consider

$$\bar{\mathcal{C}}(\alpha_0, \alpha_1) \overset{\text{def.}}{=} \left\{ (\alpha_t, J_t, s_t) \;:\; \frac{\partial \alpha_t}{\partial t} + \operatorname{div}(J_t) = s_t, \alpha_{t=0} = \alpha_0, \alpha_{t=1} = \alpha_1 \right\}.$$

The crucial question is how to measure the cost associated to this source term and introduce it in the original dynamic formulation (7.3). Several different proposals have been introduced in the literature, for instance using a L^2 cost Piccoli and Rossi [2014]. In order to avoid having "teleportation of mass" (mass which travels at infinite speed, and suddenly grows in region where there was no mass before), the associated cost should be infinite. It turns out that this can be achieved in a simple convex way, by also allowing s_t to be an arbitrary measure (*i.e.* using a 1-homogeneous cost) by penalizing s_t in the same way as

the momentum J_t,

$$\mathrm{WFR}^2(\alpha_0,\alpha_1) = \min_{(\alpha_t,J_t,s_t)_t \in \bar{\mathcal{C}}(\alpha_0,\alpha_1)} \Theta(\alpha,J,s) \tag{7.15}$$

$$\text{where} \quad \Theta(\alpha,J,s) \overset{\text{def.}}{=} \int_0^1 \int_{\mathbb{R}^d} (\theta(\alpha_t(x),J_t(x)) + \tau\theta(\alpha_t(x),s_t(x)))\,\mathrm{d}x\mathrm{d}t$$

where θ is the convex 1-homogeneous function introduced in (7.5), and τ is a weight controlling the tradeoff between mass transportation and mass creation/destruction. This formulation was proposed independently by several authors Liero et al. [2016], Chizat et al. [2016], Kondratyev et al. [2016]. This "dynamic" formulation has a "static' counterpart, see Remark 10.5. The convex optimization problem (7.15) can be solved using methods similar to those detailed in §7.3. This dynamic formulation resembles "metamorphosis" models for shape registration Trouvé and Younes [2005], and a more precise connection is detailed in Maas et al. [2015, 2016].

As $\tau \to 0$, and if $\alpha_0(\mathcal{X}) = \alpha_1(\mathcal{X})$, then one retrieves the classical OT, $\mathrm{WFR}(\alpha_0,\alpha_1) \to \mathcal{W}(\alpha_0,\alpha_1)$. In contrast, as $\tau \to +\infty$, this distance approaches the Hellinger metric over densities

$$\begin{aligned}\frac{1}{\tau}\mathrm{WFR}(\alpha_0,\alpha_1)^2 &\overset{\tau\to+\infty}{\longrightarrow} \int_{\mathcal{X}} |\sqrt{\rho_{\alpha_0}(x)} - \sqrt{\rho_{\alpha_1}(x)}|^2 \mathrm{d}x \\ &= \int_{\mathcal{X}} |1 - \sqrt{\frac{\mathrm{d}\alpha_1}{\mathrm{d}\alpha_0}(x)}|^2 \mathrm{d}\alpha_0(x).\end{aligned}$$

7.5 More General Mobility Functionals

It is possible to generalize the dynamic formulation (7.3) by considering other "mobility functions" θ in place of the one defined in (7.5). A possible choice for this mobility functional is proposed in Dolbeault et al. [2009]

$$\forall\,(a,b) \in \mathbb{R}_+ \times \mathbb{R}^d, \quad \theta(a,b) = a^{s-p}\,\|b\|^p \tag{7.16}$$

where the parameter should satisfy $p \geq 1$ and $s \in [1,p]$ in order for θ to be convex. Note that this definition should be handled with care in the case $1 < s \leq p$ because θ is not anymore 1-homogeneous, so that

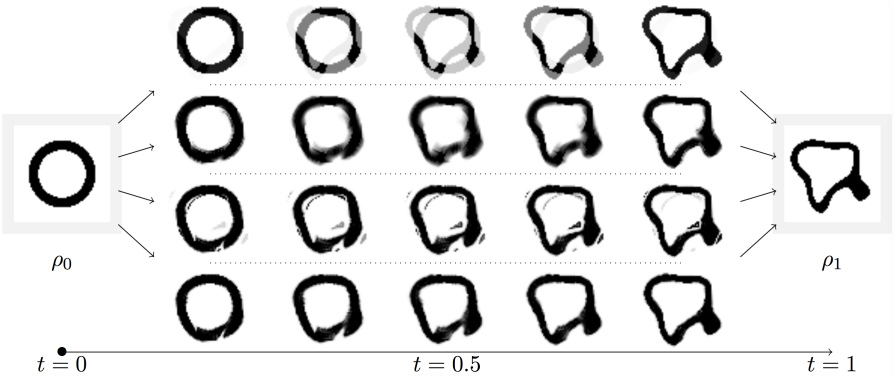


Figure 7.4: Comparison of Hellinger, Wasserstein, and Wasserstein dynamic interpolations with unbalanced OT.

solutions to (7.3) must be constrained to have a density with respect to the Lebesgue measure.

The case $s = 1$ corresponds to the classical OT and the optimal value of (7.3) defines $\mathcal{W}_p(\alpha, \beta)$. In this case, θ is 1-homogeneous, so that solution to (7.3) can be arbitrary measures. The case $(s = 1, p = 2)$ is the initial setup considered in (7.3) to define $\mathcal{W}_2$.

The limiting case $s = p$ is also interesting, because it corresponds to a dual Sobolev norm $W^{-1,p}$ and the value of (7.3) is then equal to

$$\|\alpha - \beta\|^2_{W^{-1,p}(\mathbb{R}^d)} = \min_f \left\{ \int_{\mathbb{R}^d} f \mathrm{d}(\alpha - \beta) \;:\; \int_{\mathbb{R}^d} \|\nabla f(x)\|^q \,\mathrm{d}x \leq 1 \right\}$$

for $1/q + 1/p = 1$. In the limit $p = s \to \infty$, one recovers the $\mathcal{W}_1$ norm. The case $s = p = 2$ corresponds to the Sobolev $H^{-1}(\mathbb{R}^d)$ Hilbert norm defined in (8.16).

7.6 Dynamic Formulation over the Paths Space

There is a natural dynamical formulation of the both classical and entropic regularization (as exposed in Chapter 4), which is based on studying abstract optimization problems on the space $\bar{\mathcal{X}}$ of all possible paths $\gamma : [0, 1] \to \mathcal{X}$ (*i.e.* curves) on the space $\mathcal{X}$. For simplicity, we assume $\mathcal{X} = \mathbb{R}^d$, but this extends to more general spaces such as geodesic spaces and graphs. Informally, the dynamic of "particles" between two

input measures α_0, α_1 at times $t = 0, 1$ is described by a probability distribution $\bar{\pi} \in \mathcal{M}_+^1(\bar{\mathcal{X}})$. Such a distribution should satisfy that the distributions of starting and end points must match (α_0, α_1), which is formally written using push-forward as

$$\bar{\mathcal{U}}(\alpha_0, \alpha_1) \overset{\text{def.}}{=} \left\{ \bar{\pi} \in \mathcal{M}_+^1(\bar{\mathcal{X}}) \;:\; \bar{P}_{0\sharp}\bar{\pi} = \alpha_0, \bar{P}_{1\sharp}\bar{\pi} = \alpha_1 \right\},$$

where, for any path $\gamma \in \bar{\mathcal{X}}$, $P_0(\gamma) = \gamma(0), P_1(\gamma) = \gamma(1)$.

OT over the space of paths. The dynamical version of classical OT (2.15), formulated over the space of paths, then reads

$$\mathcal{W}_2(\alpha_0, \alpha_1)^2 = \min_{\bar{\pi} \in \bar{\mathcal{U}}(\alpha_0, \alpha_1)} \int_{\bar{\mathcal{X}}} \mathcal{L}(\gamma)^2 \mathrm{d}\bar{\pi}(\gamma) \tag{7.17}$$

where $\mathcal{L}(\gamma)$ is the length of a path γ. The connexion between optimal couplings $\pi^\star$ and $\bar{\pi}^\star$ solving respectively (7.17) and (2.15) is that $\bar{\pi}^\star$ only gives mass to geodesics joining pairs of points in proportion prescribed by $\pi^\star$. In the particular case of discrete measures, this means that

$$\pi^\star = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)} \quad \text{and} \quad \bar{\pi}^\star = \sum_{i,j} \mathbf{P}_{i,j} \delta_{\gamma_{x_i, y_j}}$$

where γ_{x_i, y_j} is the geodesic between x_i and y_j. Furthermore, the measures defined by the distribution of the curve points $\gamma(t)$ at time t, where γ is drawn following $\bar{\pi}^\star$, *i.e.*

$$t \in [0,1] \mapsto \alpha_t \overset{\text{def.}}{=} \mathbf{P}_{t\sharp}\bar{\pi}^\star \quad \text{where} \quad P_t(\gamma) = \gamma(t) \in \mathcal{X}, \tag{7.18}$$

is a solution to the dynamical formulation (7.3), *i.e.* it is the displacement interpolation. In the discrete case, one recovers (7.9).

Entropic OT over the space of paths. We now turn to the reinterpretation of entropic OT, defined in Chapter 4, using the space of paths. Similarly to (4.11), this is defined using a Kullback-Leibler projection, but this time of a reference measure over the space of paths $\bar{\mathcal{K}}$ which is the distribution of a reversible Brownian motion, which has a uniform distribution at the initial and final times

$$\min_{\bar{\pi} \in \bar{\mathcal{U}}(\alpha_0, \alpha_1)} \mathrm{KL}(\bar{\pi} | \bar{\mathcal{K}}). \tag{7.19}$$

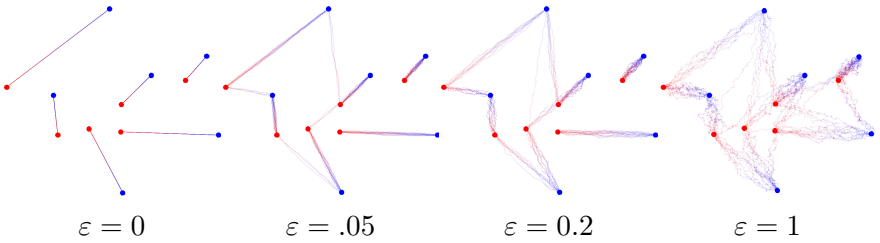


Figure 7.5: Samples from Brownian bridge paths associated to the Schrodinger entropic interpolation (7.20) over path space. Blue corresponds to $t = 0$ and red to $t = 1$.

We refer to the review paper Léonard [2014] for an overview of this problem and an historical account of the work of Schrödinger Schrödinger [1931]. One can show that the (unique) solution $\bar{\pi}^\star_\varepsilon$ to (7.19) converges to a solution of (7.17) as $\varepsilon \to 0$. Furthermore, this solution is linked to the solution of the static entropic OT problem (4.9) using Brownian bridge $\bar{\gamma}^\varepsilon_{x,y} \in \bar{\mathcal{X}}$ (which are similar to fuzzy geodesic, and converge to $\delta_{\gamma_{x,y}}$ as $\varepsilon \to 0$). In the discrete setting, this means that

$$\pi^\star_\varepsilon = \sum_{i,j} \mathbf{P}^\star_{\varepsilon,i,j} \delta_{(x_i,y_j)} \quad \text{and} \quad \bar{\pi}^\star_\varepsilon = \sum_{i,j} \mathbf{P}^\star_{\varepsilon,i,j} \bar{\gamma}^\varepsilon_{x_i,y_j} \tag{7.20}$$

where $\mathbf{P}_{\varepsilon,i,j}$ can be computed using Sinkhorn's algorithm. Similarly to (7.18), one then can define an entropic interpolation as

$$\alpha_{\varepsilon,t} \overset{\text{def.}}{=} \mathbf{P}_{t\sharp} \bar{\pi}^\star_\varepsilon.$$

Since the law $P_{t\sharp}\bar{\gamma}^\varepsilon_{x,y}$ of the position at time t along a Brownian bridge is a Gaussian $\mathcal{G}_{t(1-t)\varepsilon^2}(\cdot - \gamma_{x,y}(t))$ of variance $t(1-t)\varepsilon^2$ centered at $\gamma_{x,y}(t)$, one can deduce that $\alpha_{\varepsilon,t}$ is a Gaussian blurring of a set of traveling diracs

$$\alpha_{\varepsilon,t} = \sum_{i,j} \mathbf{P}^\star_{\varepsilon,i,j} \mathcal{G}_{t(1-t)\varepsilon^2}(\cdot - \gamma_{x_i,y_j}(t)).$$

Another way to describe this entropic interpolation $(\alpha_t)_t$ is using a regularization of the Benamou-Brenier dynamic formulation (7.2),

namely

$$\min_{(\alpha_t,v_t)_t \text{ sat. } (7.1)} \int_0^1 \int_{\mathbb{R}^d} \left(\|v_t(x)\|^2 + \frac{\varepsilon}{4} \|\nabla \log(\alpha_t)(x)\|^2 \right) \mathrm{d}\alpha_t(x)\mathrm{d}t, \tag{7.21}$$

see Gentil et al. [2015], Chen et al. [2016a].

8

Statistical Divergences

§8.1 and 8.2 first review two important classes of "divergences" (not necessarily distances) defined between probability distributions (and sometimes more general classes of measures). 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. These divergences are useful as loss functions to tackle variational problems with applications as diverse as imaging or machine learning, see Chapter 9. §8.4 details how to approximate $D(\alpha,\beta)$ from discrete samples $(x_i)_i$ and $(y_j)_j$ drawn from α and β. This is a important problem for applications in statistics and machine learning.

8.1 φ-Divergences

Before detailing in the following section "weak" norms, whose construction share similarities with $\mathcal{W}_1$, let us detail a generic construction of so-called divergences between measures, which can then be called (and used as) "loss" or "fidelity" functions. 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 are built from *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, $\mathrm{dom}\,\varphi \subset [0, \infty[$ and satisfies the following feasibility condition: $\mathrm{dom}\,\varphi \cap]0, \infty[\neq \emptyset$. The speed of growth of φ at ∞ is described by

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

If $\varphi'_\infty = \infty$, then φ grows faster than any linear function and φ is said *superlinear*. Any entropy function φ induces a φ-divergence (also known as Csiszár 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

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

if α, β are nonnegative and ∞ otherwise.

In definition (8.1), the additional term $\varphi'_\infty \alpha^\perp(\mathcal{X})$ 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 super-linear growth (as for instance for 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

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

[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.

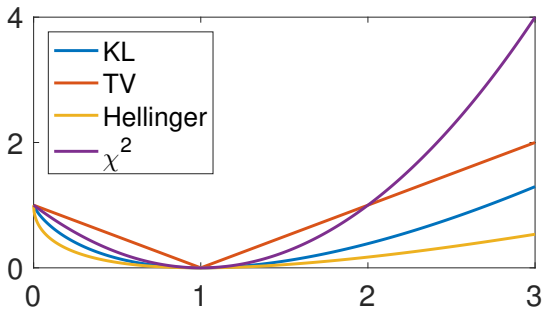


Figure 8.1: Example of entropy functionals.

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., 2015, 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).

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.

Example 8.1 (Kullback-Leibler divergence)**.** The Kullback-Leibler divergence $\mathrm{KL} \overset{\text{def.}}{=} \mathcal{D}_{\varphi_{\mathrm{KL}}}$, 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, $\mathbf{KL} \overset{\text{def.}}{=} \mathbf{D}_{\varphi_{\mathrm{KL}}}$ has the unique property of being both a φ-divergence and a

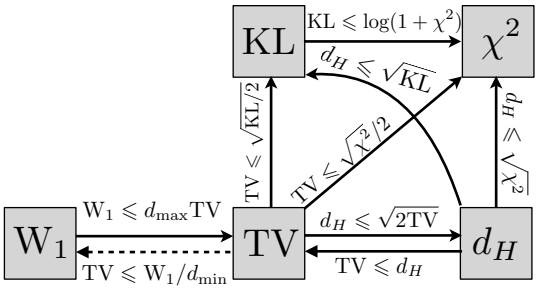


Figure 8.2: Diagram of relationship between divergences (inspired by Gibbs and Su [2002]). For $\mathcal{X}$ a metric space with ground distance d, $d_{\max} = \sup_{(x,x')} d(x,x')$ is the diameter of $\mathcal{X}$. When $\mathcal{X}$ is discrete, $d_{\min} \stackrel{\text{def.}}{=} \min_{x \neq x'}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

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 divergences 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 (9.32)), i.e. $\mathbf{KL} = \mathbf{B}_{-\mathbf{H}}$. A Bregman divergences is locally a squared Euclidean distance since

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

and the set of separating points $\{\mathbf{a} : \mathbf{B}_\psi(\mathbf{a}|\mathbf{b}) = \mathbf{B}_\psi(\mathbf{a}|\mathbf{b}')\}$ is an 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 (9.31). Bregman divergences are also important in convex optimization, and can be used for instance to derive Sinkhorn iterations and study its convergence, see Remark 4.7. While KL is both a φ and a Bregman divergence, these are radically different concepts. In particular, to generalize formula (8.5) to non-discrete measures on some space $\mathcal{X}$, one

needs to fix a reference measure ξ and consider only measures α, β with square integrable densities $\frac{\mathrm{d}\alpha}{\mathrm{d}\xi}, \frac{\mathrm{d}\beta}{\mathrm{d}\xi} \in L^2(\mathrm{d}\xi)$. For some strictly convex $\psi : L^2(\mathrm{d}\xi) \to \mathbb{R}$, one then defines the Bregman divergence as

$$\mathcal{B}_{\psi,\xi}(\alpha|\beta) \stackrel{\text{def.}}{=} \psi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\xi}\right) - \psi\left(\frac{\mathrm{d}\beta}{\mathrm{d}\xi}\right) - \int_{\mathcal{X}} \eta(x)\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\xi}(x) - \frac{\mathrm{d}\beta}{\mathrm{d}\xi}(x)\right)\mathrm{d}\xi(x)$$

where $\eta = \nabla\psi\left(\frac{\mathrm{d}\beta}{\mathrm{d}\xi}\right) \in L^2(\mathrm{d}\xi)$ is the gradient [Rao and Nayak, 1985, Jones and Byrne, 1990].

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 in 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 Gaussian 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 Gaussian-constrained geodesic between (α, β) over this hyperbolic half plane does not have a constant standard deviation.

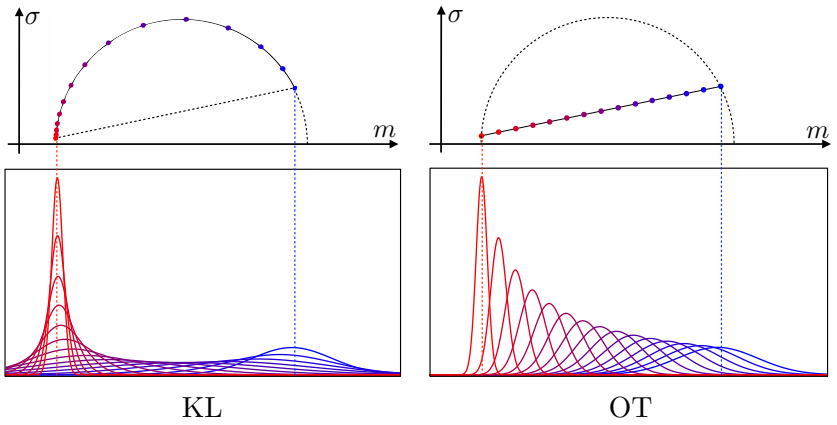


Figure 8.3: Comparisons of interpolation between Gaussians using KL (hyperbolic) and OT (Euclidean) geometries.

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 2.29, 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.

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 8.1 (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

$$\mathcal{W}_1(\alpha,\beta) \leq 2\,\|\alpha-\beta\|_{\mathrm{TV}}$$

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 the weak topology (once again on a compact ground space $\mathcal{X}$) is compact, from any sequence of measure $(\alpha_k)_k$, 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 (ρ_α,ρ_β) 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. 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 indeed a distance [Endres and Schindelin, 2003, Österreicher and Vajda, 2003]. In sharp contrast with KL, $\mathrm{JS}(\alpha,\beta)$ is always bounded,

and similarly to the TV norm and the Hellinger distance, it metrizes the strong convergence.

Example 8.5 (χ^2)**.** The χ^2-divergence $\chi^2 \stackrel{\text{def.}}{=} \mathcal{D}_{\varphi_{\chi^2}}$ 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 "dual norms", which is a convenient way to design "weak" norms that can deal with arbitrary measures. For a symmetric convex set $B \subset \mathcal{C}(\mathcal{X})$ of continuous functions, , one defines

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

These dual norms are often called "Integral Probability Metrics" (IPM), 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 non-trivial divergence that is also a dual norm, see [Sriperumbudur et al., 2009].

By using smaller "balls" B, one defines weaker dual norms, which are thus able to metrize the weak convergence (Definition 2.2). Figure 8.4 displays a comparison of several such dual norms, which we now detail.

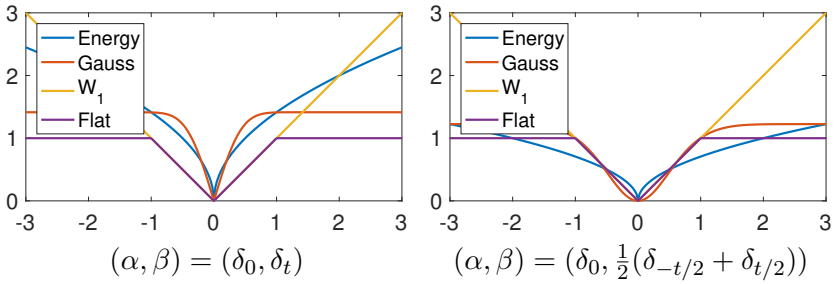


Figure 8.4: Comparison of dual norms.

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 "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 the Maximum Mean Discrepancy

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 $r \in \mathbb{R}^n$ (resp. such that $r^{\mathrm{T}}\mathbb{1}_n = 0$) the following inequality holds

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

In this case, one sets

$$B = \{f \ : \ \|f\|_k \leq 1\} \tag{8.13}$$

where $\|\cdot\|_k$ is a Hilbertian norm defined using a positive definite kernel

$$\|f\|_k^2 \stackrel{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} k(x,y) f(x) f(y) \mathrm{d}\rho(x) \mathrm{d}\rho(y)$$

where ρ is some reference measure, typically Lebesgue's measure on $\mathcal{X} = \mathbb{R}^d$. The resulting dual RKHS norms is often referred to as the "Maximum Mean Discrepancy" (MMD) (see Gretton et al. [2007]), and can be revisited through the prism of Kernel Mean Embeddings (see Muandet et al. [2017] for a review).

The dual norm (8.10) for such a RKHS ball B can be shown to be also defined using a dual kernel k^*, *i.e.*

$$\|\alpha\|_B^2 = \|\alpha\|_{k^*}^2 \stackrel{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{X}} k^*(x,y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y). \tag{8.14}$$

This dual kernel k^* is in some sense the "inverse" of the primal kernel k. For instance, for translation-invariant kernels k on $\mathcal{X} = \mathbb{R}^d$ (*i.e.* convolution) $k(x,y) = k_0(x-y)$, then $k^*(x,y) = k_0^*(x-y)$ where formally $\hat{k}_0^*(\omega) = \hat{k}_0(\omega)^{-1}$ where $\hat{k}_0$ is the Fourier transform.

Expression (8.14) can be re-phrased introducing two independent random vectors (X, X') on $\mathcal{X}$ distributed following α

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

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

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

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 an Euclidean norm (associated to the positive definite matrix $\mathbf{k}^*$) on the simplex Σ_n. To compute the discrepancy between two discrete measures of the form (2.3), one can use

$$\|\alpha - \beta\|_{k^*}^2 = \sum_{i,i'} \mathbf{a}_i \mathbf{a}_{i'} k^*(x_i, x_{i'}) + \sum_{j,j'} \mathbf{b}_j \mathbf{b}_{j'} k^*(y_j, y_{j'}) - 2 \sum_{i,j} \mathbf{a}_i \mathbf{b}_j k^*(x_i, y_j). \tag{8.15}$$

Example 8.9 (Gaussian RKHS). One of the most popular kernel is the Gaussian one, in which case both primal and dual norms have closed form expressions, provided here up to normalization constants:

$$k(x, y) = e^{-\frac{\sigma^2 (x-y)^2}{2}} \quad \text{and} \quad k^*(x, y) = e^{-\frac{(x-y)^2}{2\sigma^2}}$$

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 4.15). 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. Another issue with such kernels is that they are global (have slow polynomial decay) which makes them typically computationally more expensive (no compact support approximation is possible).

Example 8.10 ($H^{-1}(\mathbb{R}^d)$)**.** Another important dual norm is $H^{-1}(\mathbb{R}^d)$, the dual (*i.e.* 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 kernel $\|f\|_k^2 = \|\nabla f\|^2_{L^2(\mathbb{R}^d)}$. 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

$$\|\alpha - \beta\|^2_{H^{-1}(\mathbb{R}^d)} = \min_s \left\{ \int_{\mathbb{R}^d} \|s(x)\|_2^2 \,\mathrm{d}x \;:\; \mathrm{div}(s) = \alpha - \beta \right\}, \quad (8.16)$$

which should be contrasted with (6.4) where a 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

$$\mathcal{W}_2(\alpha, \alpha + \varepsilon\rho) = \varepsilon \,\|\rho\|_{H^{-1}(\alpha)} + o(\varepsilon) \quad (8.17)$$

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

$$b^{-1/2} \,\|\alpha - \beta\|_{H^{-1}(\mathbb{R}^d)} \leq W_2(\alpha, \beta) \leq a^{-1/2} \,\|\alpha - \beta\|_{H^{-1}(\mathbb{R}^d)} \,. \quad (8.18)$$

see [Santambrogio, 2015, Theorem 5.34], see also [Peyre, 2011] for sharp constants.

Example 8.11 (Negative Sobolev Spaces)**.** One can generalize this construction by considering Sobolev space $H^{-r}(\mathbb{R}^d)$ of arbitrary negative index, which are dual of the functional Sobolev space $H^r(\mathbb{R}^d)$ of function 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

$$\forall\, x \in \mathbb{R}^d, \quad k_0^*(x) = \begin{cases} \frac{1}{\|x\|^{d-2r}} & \text{if} \quad d/2 < r < d, \\ -\|x\|^{2r-d} & \text{if} \quad r > d. \end{cases} \tag{8.19}$$

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

$$\|\alpha - \beta\|^2_{\mathrm{ED}(\mathcal{X},d^p)} \overset{\text{def.}}{=} \|\alpha - \beta\|_{k^*_{\mathrm{ED}}} \quad \text{where} \quad k^*_{\mathrm{ED}}(x,y) = -d(x,y)^p \tag{8.20}$$

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 (8.19), 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 behaviour 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 exhibit a perfect linear scaling

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

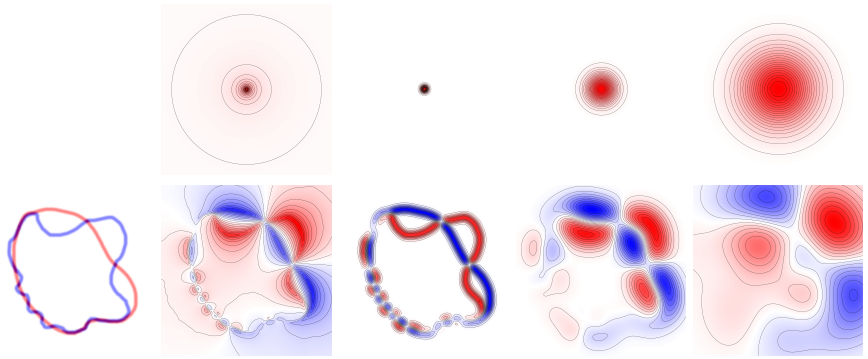


Figure 8.5: Top row: display of ψ such that $\|\alpha - \beta\|_k = \|\psi \star (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{k}_0^*(\omega)}$ where $k^*(x, x') = k_0^*(x - x')$. Bottom row: display of $\psi \star (\alpha - \beta)$. (G,$\sigma$)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

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 be sometimes 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 2.28 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 2.29 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 (RBF) 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, Corr. §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 fundamental result:

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 counter-example in that dimension suffices to obtain a counter-example in any higher dimension. We provide a non-random counterexample which works using measures supported on 4 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 4 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 J is the centering matrix $\mathbf{J} = I_n - \frac{1}{n}\mathbb{1}_{n,n}$, 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 ℓ_2 spaces incurs necessarily an important distortion Naor and Schechtman [2007], Andoni et al. [2017] 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

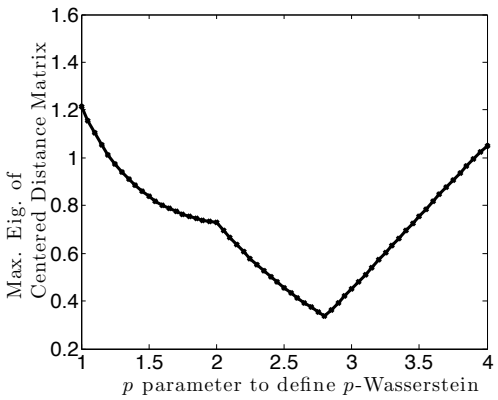


Figure 8.6: Maximal eigenvalue of $\mathbf{J}\mathbf{D}_p^2\mathbf{J}$ for varying values of p.

et al., 2011], but the equivalence constant between the distances growth fast with the dimension d. Note also that also for $p = 1$ the embedding is only true 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 dual of Lipschitz functions f (see §6.2) and approximate 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 multi-scale embedding 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

§10.4 shows 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].

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 (2.17)) 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 $\hat{\alpha}$ and $\hat{\beta}$ are random measures, so $D(\hat{\alpha}, \hat{\beta})$ 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, it does not converges toward $\|\alpha - \beta\|_{\mathrm{TV}}$, indeed, with probability 1,

$\|\hat{\alpha}_n - \hat{\beta}_n\|_{\mathrm{TV}} = 2$ since the support of the discrete measures do 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 $1 \leq p < +\infty$,

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

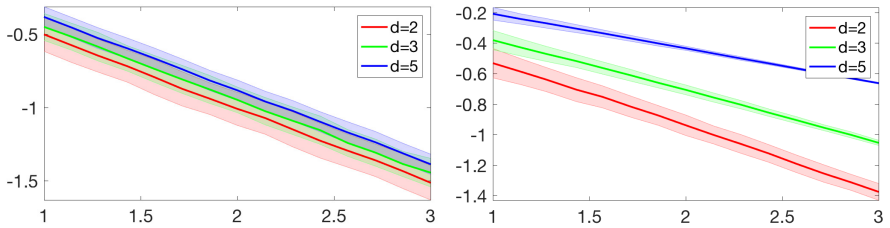
where the expectation $\mathbb{E}$ is taken with respect to the random samples $(x_i, y_i)_i$. This rate is tight in $\mathbb{R}^d$ if one of the two measure has a density with respect to the Lebesgue measure. This result was proved for general metric spaces [Dudley, 1969] using the notion of covering numbers, and later refined, in particular for $\mathcal{X} = \mathbb{R}^d$ in [Dereich et al., 2013, Fournier and Guillin, 2015]

Rate for Wasserstein [Weed and Bach, 2017] with measure supported on low-dimensional sub-domains: the rate depends on the intrinsic dimensionality of the support. Also studies non-asymptotic behavior, such as for measure which are approximated discrete (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 $\|\cdot\|_B^2 = \|\cdot\|_{k^*}^2$ which are dual of RKHS norms (also called MMD), as defined in (8.13), and on contrary to Wasserstein distances, the sample complexity does not depends 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]. 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 stochastic gradient descent when minimizing such losses, one should rather introduce use



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 $\log(D(\hat{\alpha}_n, \alpha))$.

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 (8.15). It satisfies $\mathbb{E}(\mathrm{MMD}_{k^*}(\hat{\alpha}_n, \hat{\beta}_n)^2) = \|\alpha - \beta\|_{k^*}^2$, see [Gretton et al., 2012].

8.4.2 Empirical Estimators for φ-divergences

Is it 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]. On an 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,

$$\hat{\alpha}_n \star h_\sigma = \frac{1}{n} \sum_i h_\sigma(\cdot - x_i). \tag{8.21}$$

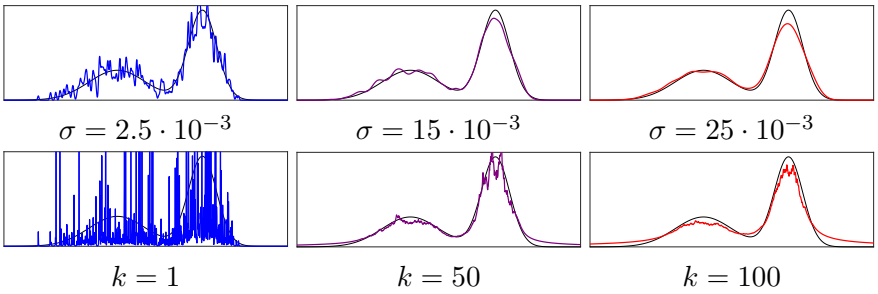


Figure 8.8: Comparison of kernel density estimation $\hat{\alpha}_n \star h_\sigma$ (top, using a Gaussian kernel h) and k-nearest neighbors estimation $\rho^k_{\hat{\alpha}_n}$ (bottom) for $n = 200$ samples from a mixture of two Gaussians.

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 non-parametric 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 k^{th} nearest neighbors among the $(x_i)_{i=1}^n$, a density estimator is defined as

$$\rho^k_{\hat{\alpha}_n}(x) \overset{\text{def.}}{=} \frac{k/n}{|B_d|\Delta_k(x)^r} \tag{8.22}$$

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 (8.21), this formula (8.22) estimate the radius required to encapsulate k samples. Figure 8.8 compares the estimators (8.21) and (8.22). A typical example of application is detailed in (9.32) 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.

8.5 Entropic Regularization: between OT and MMD

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

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

where $\mathbf{P}^\star$ is the solution of (4.2) while $(\mathbf{f}^\star,\mathbf{g}^\star)$ are solutions of (4.29). Assuming $\mathbf{C}_{i,j} = d(x_i,x_j)^p$ for some distance d on $\mathcal{X}$, for two discrete probability distribution of the form (2.3), this defines a regularized Wasserstein cost

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

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

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

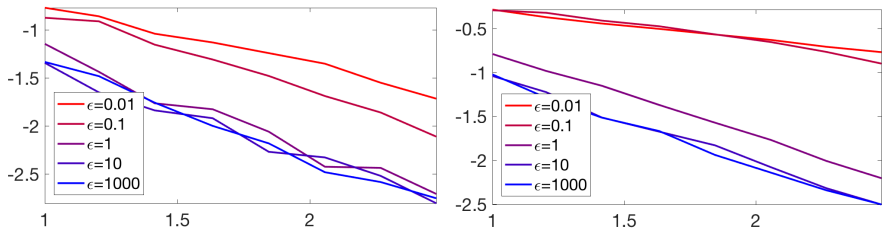
The following proposition, whose proof can be found in Ramdas et al. [2017], shows that this regularized divergence interpolates between the Wasserstein distance and the energy distance defined in Example 8.12.

Proposition 8.3. One has

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

where $\|\cdot\|_{\mathrm{ED}(\mathcal{X},d)}$ is defined in (8.20).

Note that it is possible to define other families of divergence on top of the Sinkhorn's methods, see in particular Amari et al. [2017] for an alternative technique.



$d = 2$ $d = 5$

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.

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; The main technical challenge in these approaches lies in approximating and differentiating efficiently the Wasserstein distance.

In image processing, using the Wasserstein distance as a loss was used to synthetize textures [Tartavel et al., 2016], where the Wasserstein loss was used to account for the discrepancy between statistics of the synthesized and the input exemplar. It was 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]. 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 exposed 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 [1999]. When prior information is available, the Wasserstein distance can be used instead, with markedly different results [Sandler and Lindenbaum, 2011, Zen et al., 2014, Rolet et al., 2016].

Optimization problems that involve Wasserstein distances typically require that we have access to their gradients of approximations thereof. We start this section by presenting methods to approximate such gradients, and 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 an Euclidean space. Such a comparison is done through a "loss" or a "fidelity" term, which, in this section, is the Wasserstein distance. In the simplest scenario, the computation of a suitable parameter θ is obtained by minimizing directly

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

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 basic case: 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 these inputs, as shown by (2.19) and (2.23) 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 $\alpha_1, \dots, \alpha_K$ 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 §2.28) or elliptically contoured (see §2.29) 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 in statistics and machine learning, or when dealing with high-dimensional data, a Lagrangian perspective is usually the only tractable alternative.

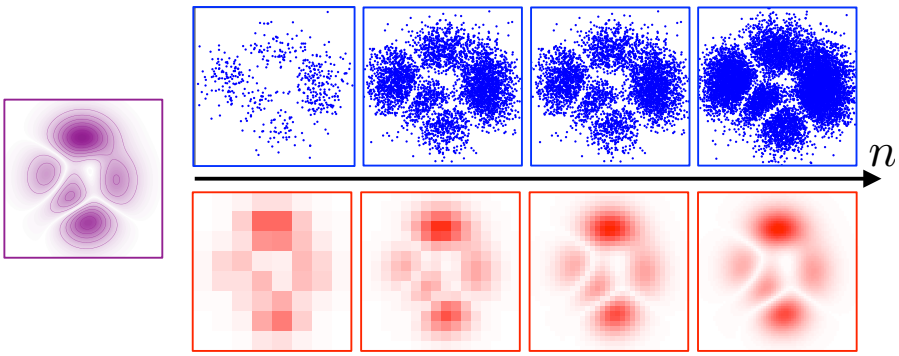


Figure 9.1: Increasing fine discretization of a continuous distribution having a density (violet, left) using a Lagrangian representation $\frac{1}{n}\sum_i \delta_{x_i}$ (blue, top) and an Eulerian representation $\sum_i \mathbf{a}_i \delta_{x_i}$ with x_i representing cells on a grid of increasing 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 fist 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. (9.1) is not differentiable In order to obtain a smooth minimization problem, we use the entropic regularized OT and approximate (9.1) using

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

We recall here Proposition 4.6 which shows that the entropic loss function is differentiable and convex with respect to the input histograms, and give the expression for its 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 (4.29) 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, and the function is differentiable if they are unique.

Note that the zero mean condition on $(\mathbf{f}, \mathbf{g})$ is important when using a 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 (4.29) between $\mathbf{a}(\theta)$ and $\mathbf{b}$ for the cost matrix $\mathbf{C}$ (which is fixed and independent on 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 constant) weights and discretize the measure as empirical measures $\alpha_\theta = \frac{1}{n}\sum_i \delta_{x(\theta)_i}$ for a point-cloud parameterization function $x : \theta \mapsto x(\theta) = (x(\theta)_i)_{i=1}^n \in \mathcal{X}^n$, where we assume here that $\mathcal{X}$ is Euclidean. Problem (9.1) is thus approximated as

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

Note that here the cost matrix $\mathbf{C}(x(\theta))$ now depends on θ since the support of α_θ is intended to move. 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 spikes positions)**.** 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 convex 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 sub-gradients.

Assuming $(\mathcal{X}, \mathcal{Y})$ are (possibly 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 (4.18). 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 $\mathcal{E}_L$.

9.1.3 Automatic Differentiation

The difficulty when applying formulas (9.3) and (9.8) is that one needs to compute the optimal solutions $\mathbf{f}$ or $\mathbf{P}$ for them 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 is 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 (4.29) 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 (4.44), rather than the quantity $\mathrm{L}_{\mathbf{C}}^{\varepsilon}$ in (4.2) which incorporates the entropy of the regularized optimal transport. The cost for computing the gradient of functionals involving Sinkhorn divergences is the same as that of computation the functional itself, see for instance [Bonneel et al., 2016, Genevay et al., 2017a] for some applications of this approach. We also refer to Adams and Zemel [2011] for any early work on differentiating Sinkhorn iterations with respect to the cost matrix (as it is done in the Lagrangian framework), with applications to machine learning over ranking and permutations. We refer to [Griewank and Walther, 2008, Rall, 1981, Neidinger, 2010] for more details on automatic differentiation, and in particular the "reverse mode" which is the fastest way to compute gradients, the efficient computation of Hessian being much more involved. In terms of implementation, all recent "deep-learning" Python frameworks feature state of the art reverse-mode differentiation and support for GPU computations [Al-Rfou et al., 2016, Abadi et al., 2016, pyt, 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 are always at least as fast of the formula (9.3) and (9.8), these formula 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 iteration). There exists however subsampling strategies that mitigates 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 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 "Fréchet" or "Karcher" mean (see Karcher [2014] for an historical account). For a generic distance d, problem (9.9) is usually a difficult non-convex 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

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

This barycenter problem (9.10) was originally introduced by Agueh and Carlier [2011] following earlier ideas of Carlier and Ekeland [2010]. They proved in particular uniqueness of the barycenter for $c(x, y) = \|x - y\|^2$ over $\mathcal{X} = \mathbb{R}^d$, if one of the input measure has a density with respect to the Lebesgue measure (and more generally under the same hypothesis as the one guaranteeing the existence of a Monge map, see Remark 2.23).

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

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

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

The computation of Wasserstein barycenters has found numerous applications in image processing [Rabin et al., 2011], computer graphics [Solomon et al., 2015], statistics [Boissard et al., 2015], Bayesian inference [Srivastava et al., 2015b,a] and machine learning [Cuturi and Doucet, 2014]. For instance, the ability of computing barycenters is the workhorse of clustering methods such as the K-means algorithm [del Barrio et al., 2016, Ho et al., 2017].

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] shows 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 re-cast (9.11) as a multi-marginal OT problem, see Remark 10.2.

Remark 9.2 (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

$$\min_{\alpha \in \mathcal{M}_+^1(\mathcal{X})} \mathbb{E}_M(\mathcal{L}_c(\alpha, \beta)) = \int_{\mathcal{M}_+^1(\mathcal{X})} \mathcal{L}_c(\alpha, \beta) \mathrm{d}M(\beta), \tag{9.12}$$

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 (9.12)) corresponds to the consistency of the barycenters, and is proved in [Bigot and Klein, 2012a, Le Gouic and Loubes, 2016, Bigot and Klein, 2012b]. 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 [Agueh and Carlier, 2017] for a formulation of this problem and its solution in particular cases (1-D distributions and Gaussian measures).

Remark 9.3 (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 2.23, one can define optimal transportation maps T_s between α and α_s, *i.e.* in particular such

that $T_{s,\sharp}\alpha = \alpha_s$. The average map

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

(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 2.23) to the sum of gradients of convex functions, equal to the gradient of a sum of convex functions, therefore optimal by Brenier's theorem again). As shown in [Agueh and Carlier, 2011], first order optimality conditions of the barycenter problem (9.12) actually reads $T^{(\alpha^\star)} = \mathrm{Id}_{\mathbb{R}^d}$ (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 solution to the fixed-point equation

$$G(\alpha) = \alpha \quad \text{where} \quad G(\alpha) \stackrel{\text{def.}}{=} T^{(\alpha)}_\sharp \alpha. \tag{9.13}$$

Under mild conditions on the input measures, it is shown in Álvarez-Esteban et al. [2016] that $\alpha \mapsto G(\alpha)$ strictly decreases the objective function of (9.12) if α is not the barycenter, and that the fixed point iterations $\alpha^{(\ell+1)} \stackrel{\text{def.}}{=} G(\alpha^{(\ell)})$ 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 (4.18).

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 9.4 (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 (2.39), 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 (2.40). 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 following 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 PSD 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 $\mathbf{\Sigma}^{(\ell+1)} \overset{\text{def.}}{=} \Psi(\mathbf{\Sigma}^{(\ell)})$ can be shown to converge to the solution $\mathbf{\Sigma}^\star$ thanks to the more general result provided in [Álvarez-Esteban et al., 2016]. This is because the fixed point map G defined in (9.13) preserves Gaussian distributions, and in fact,

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

This method has been used for application to texture synthesis in [Xia et al., 2014]. Figure 9.2 shows two examples of computations of barycenters between four 2-D Gaussians.

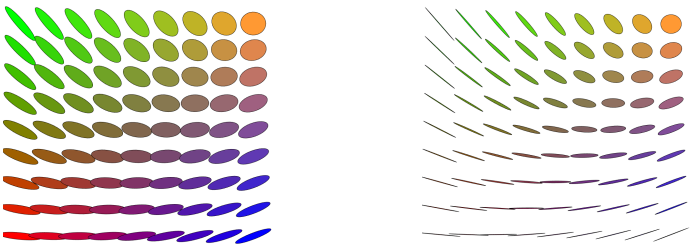


Figure 9.2: Barycenters between 4 Gaussian distributions in 2-D. Each Gaussian is displayed using an ellipse aligned with the principal axes of the covariance, and with elongations proportional to the corresponding eigenvalues.

Remark 9.5 (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 re-arrangement, as detailed in Remark 2.28. 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 (2.31) the barycenters 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 (2.32), and using (2.34), 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 9.6 (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 9.7 (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

$$\min_{\mathbf{a}\in\Sigma_n} \sum_{s=1}^{S} \lambda_s \mathrm{L}^\varepsilon_{\mathbf{C}_s}(\mathbf{a}, \mathbf{b}_s) \tag{9.14}$$

for some $\varepsilon > 0$. This is a smooth convex minimization problem, which can be tackled using gradient descent [Cuturi and Doucet, 2014]. An alternative is to use descent method (typically quasi-Newton) on the semi-dual Cuturi and Peyré [2016], which is useful to integrate additional regularizations on the barycenter (e.g. to impose some smoothness). A simple but effective approach, as remarked in Benamou et al. [2015] is to rewrite (9.14) as a (weighted) KL projection problem

$$\min_{(\mathbf{P}_s)_s} \left\{ \sum_s \lambda_s \varepsilon \mathbf{KL}(\mathbf{P}_s | \mathbf{K}_s) \;:\; \forall\, s, {\mathbf{P}_s}^{\mathrm{T}} \mathbb{1}_m = \mathbf{b}_s,\ \mathbf{P}_1 \mathbb{1}_1 = \ldots = \mathbf{P}_S \mathbb{1}_S \right\} \tag{9.15}$$

where we denoted $\mathbf{K}_s \overset{\text{def.}}{=} e^{-\mathbf{C}_s/\varepsilon}$. Here, the barycenter $\mathbf{a}$ is implicitly encoded in the row marginals of all the couplings $\mathbf{P}_s \in \mathbb{R}^{n\times n_s}$ as $\mathbf{a} = \mathbf{P}_1 \mathbb{1}_1 = \ldots = \mathbf{P}_S \mathbb{1}_S$. As detailed in Benamou et al. [2015], one can generalize Sinkhorn to this problem, which also corresponds to iterative

projection. This can also be seen as a special case of the generalized Sinkhorn detailed in §4.6. The optimal couplings $(\mathbf{P}_s)_s$ solving (9.15) are computed in scaling form as

$$\mathbf{P}_s = \mathrm{diag}(\mathbf{u}_s)\mathbf{K}\,\mathrm{diag}(\mathbf{v}_s), \tag{9.16}$$

and the scalings are sequentially updated as

$$\forall\, s \in [\![1, S]\!], \quad \mathbf{v}_s^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mathbf{b}_s}{\mathbf{K}_s^{\mathrm{T}}\mathbf{u}_s^{(\ell)}}, \tag{9.17}$$

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

$$\text{where} \quad \mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \prod_s (\mathbf{K}_s\mathbf{v}_s^{(\ell+1)})^{\lambda_s}. \tag{9.19}$$

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 (9.16) 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 (9.14))

$$\max_{(\mathbf{f}_s,\mathbf{g}_s)_s} \left\{ \sum_s \lambda_s \left(\langle \mathbf{g}_s,\, \mathbf{b}_s \rangle - \varepsilon \langle \mathbf{K}_s e^{\mathbf{g}_s/\varepsilon},\, e^{\mathbf{f}_s/\varepsilon} \rangle \right) \;:\; \sum_s \lambda_s \mathbf{f}_s = 0 \right\}. \tag{9.20}$$

Proof. Introducing Lagrange multipliers in (9.15) 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 (4.50) of the function $\mathbf{KL}^*(\cdot|\mathbf{K}_s)$. This Legendre transform reads

$$\mathbf{KL}^*(\mathbf{U}|\mathbf{K}) = \sum_{i,j} \mathbf{K}_{i,j}(e^{\mathbf{U}_{i,j}} - 1), \tag{9.21}$$

which shows the desired formula. To show (9.21), 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 (9.20) with respect to each $\mathbf{g}_s$, while keeping all the other variable fixed, is obtained in closed form by (9.17). Minimizing (9.20) with respect to all the $(\mathbf{f}_s)_s$ requires to solve for $\mathbf{a}$ using (9.19) and leads to the expression (9.18).

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 4.17. We refer to [Solomon et al., 2015] for more details and other applications to computer graphics and imaging sciences.

Barycenters have found many applications outside the field of shape analysis. They have been be used for image processing, in particular color modification [Solomon et al., 2015] (see Figure 9.6); Bayesian computations [Srivastava et al., 2015b] to summarize measures; non-linear 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 involved non-convex objective functions which

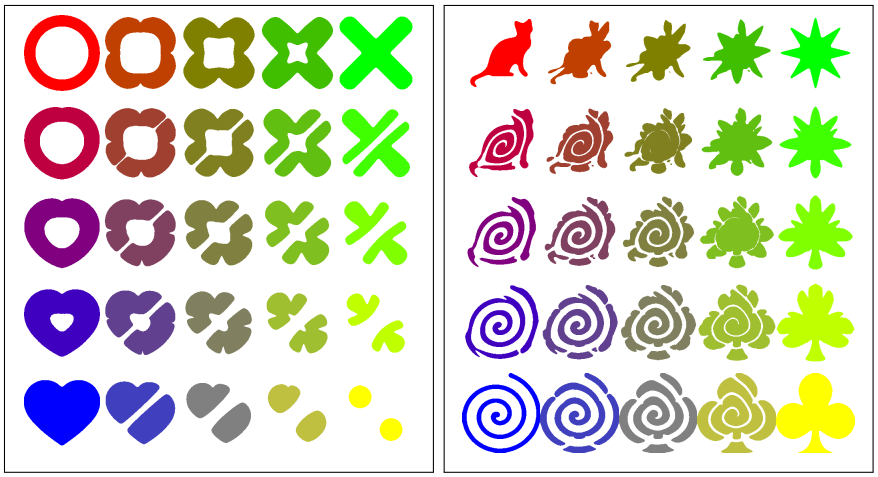


Figure 9.3: Barycenters between 4 input 2-D shapes using entropic regularization (9.14). To display a binary shape, the displayed images shows a thresholded density. The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square.

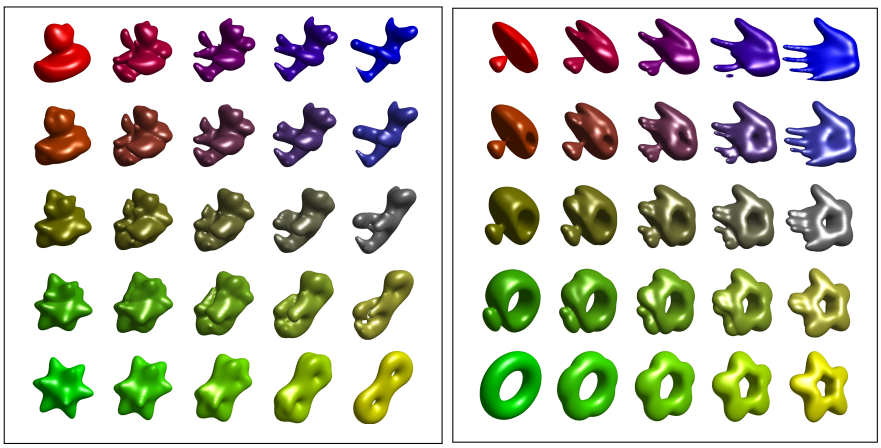


Figure 9.4: Barycenters between 4 input 3-D shapes using entropic regularization (9.14). The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square. Shapes are represented as measures that are uniform within the boundaries of the shape and null outside.

Figure 9.5: Barycenters interpolation between two input measures on surfaces, computed using the geodesic in heat fast kernel approximation (see Remark 4.17). Extracted from [Solomon et al., 2015].

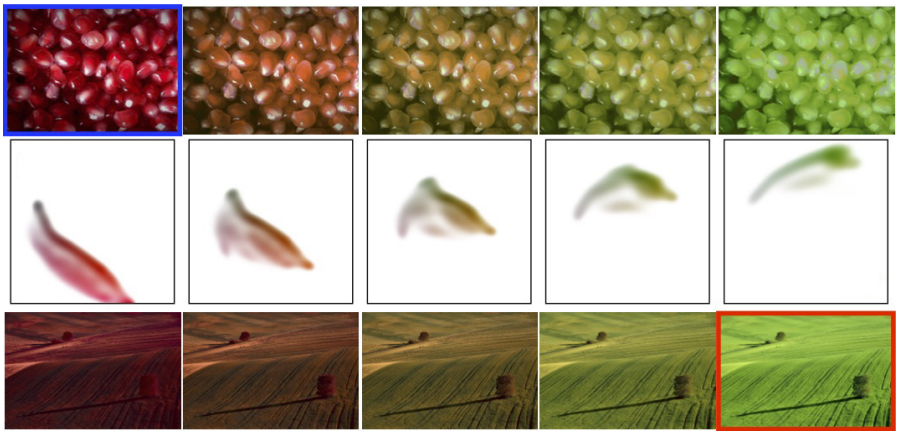


Figure 9.6: Interpolation between the two 3-D color empirical histograms of two input images (here only the 2-D chromatic projection is visualized for simplicity. The modified histogram is then applied to the input images using barycentric projection as detailed in Remark 4.9. Extracted from [Solomon et al., 2015].

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., 2017], and the statistical estimation of template models [Boissard et al., 2015].

Remark 9.8 (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

$$\min_{(\mathbf{b}_u\in\Sigma_{n_u})_{u\in U}} \sum_{(r,s)\in\mathcal{G}} \mathrm{L}_{\mathbf{C}_{r,s}}(\mathbf{b}_r,\mathbf{b}_s) \tag{9.22}$$

where the cost matrices $\mathbf{C}_{r,s}\in\mathbb{R}^{n_r\times n_s}$ needs 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 (9.15), and one extends Sinkhorn iterations (9.17) to this problem (this can also be derived by re-casting this problem in the form of the generalized Sinkhorn algorithm detailed in §4.6). This discrete variational problem (9.22) 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 [2017] for extensions to non-quadratic (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

$$\mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \mathbf{a}^{(\ell)} - \tau\nabla F(\mathbf{a}^{(\ell)}) \tag{9.23}$$

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 non-smooth functions, one can use instead an "implicit" scheme, which is also called the proximal-point algorithm (see for in-

stance Bauschke and Combettes [2011])

$$\mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \mathrm{Prox}_{\tau F}^{\|\cdot\|}(\mathbf{a}^{(\ell)}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathbf{a}} \frac{1}{2}\left\|\mathbf{a}-\mathbf{a}^{(\ell)}\right\|^2 + \tau F(\mathbf{a}). \tag{9.24}$$

Note that this corresponds to the Euclidean proximal operator, already encountered in (7.13). The update (9.23) can be understood as iterating the explicit operator $\mathrm{Id} - \tau\nabla F$, while (9.24) makes use of the implicit operator $(\mathrm{Id}+\tau\nabla F)^{-1}$. For convex F, iterations (9.24) always converge, for any value of $\tau > 0$.

Instead of using the ℓ^2 norm $\|\cdot\|$ in (9.24), if the function F is defined on the simplex of histograms Σ_n, then it makes sense to rather use an optimal transport metric, in order to solve

$$\mathbf{a}^{(\ell+1)} \overset{\text{def.}}{=} \operatorname*{argmin}_{\mathbf{a}} \mathrm{W}_p(\mathbf{a}, \mathbf{a}^{(\ell)})^p + \tau F(\mathbf{a}). \tag{9.25}$$

Remark 9.9 (Wasserstein gradient flows)**.** Equation (9.25) can be generalized to arbitrary measures by defining the iteration

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

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

$$\frac{\partial \alpha_t}{\partial t} = \mathrm{div}(\alpha_t \nabla(F'(\alpha_t))) \tag{9.27}$$

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 the relative entropy with respect to the Lebesgue

measure $\mathcal{L}_{\mathbb{R}^d}$ on $\mathcal{X} = \mathbb{R}^d$

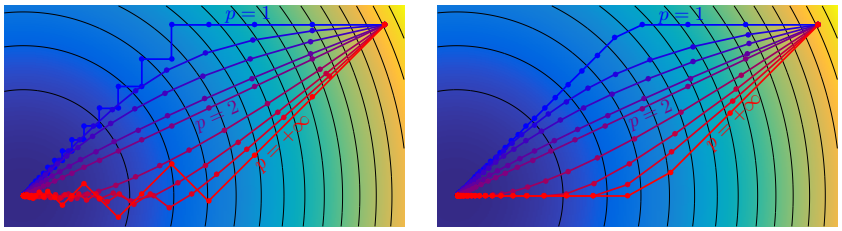
$$H(\alpha) = -\int_{\mathbb{R}^d} \rho_\alpha(x)(\log(\rho_\alpha(x)) - 1)\mathrm{d}x \tag{9.28}$$

(setting $H(\alpha) = -\infty$ when α does not have a density), then (9.27) shows that the gradient flow of this neg-entropy is the linear heat diffusion

$$\frac{\partial \alpha_t}{\partial t} = \Delta \alpha_t \tag{9.29}$$

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 $\int_{\mathbb{R}^d} \left\|\nabla \rho_{\alpha(x)}\right\|^2 \mathrm{d}x$, or, alternatively, as the entropy for the optimal transport flow (somehow an "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 details in the monograph by Ambrosio et al. [2008], 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 non-linear evolution equations such as the porous medium equation [Otto, 2001], total variation flows [Carlier and Poon, 2017], 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 [2010] (see also the review paper [Santambrogio, 2018]).

Remark 9.10 (Gradient flows in metric spaces)**.** The implicit stepping (9.26) is a special case of a general formalism to define gradient flows over general metric spaces $(\mathcal{X}, d)$, where d is a distance, as detailed in Ambrosio et al. [2008]. For some function $F(x)$ defined for $x \in \mathcal{X}$, the implicit discrete minmization step is then defined



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.

as

$$x^{(\ell+1)} \in \operatorname*{argmin}_{x \in \mathcal{X}} d(x^{(\ell)}, x)^2 + \tau F(x). \tag{9.30}$$

The JKO step (9.26) corresponds to the use of the Wasserstein distance on the space of probability distributions. In some cases, one can show that (9.30) admits a continuous flow limit x_t as $\tau \to 0$ and $k\tau = t$. In case that $\mathcal{X}$ also has a Euclidean structure, an explicit stepping is defined by linearizing F

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

In sharp contrast to the implicit formula (9.30) it is usually straightforward to compute but can be unstable. The implicit step is always stable, is also defined for non-smooth 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 (9.31) is unstable for $p = 1$ and $p = +\infty$, and for $p = 1$ it gives axis-aligned steps (coordinate wise descent). In contrast, the implicit scheme (9.30) is stable. Note in particular how, for $p = 1$, when the two coordinates are equal, the following step operates in the diagonal direction.

Remark 9.11 (Lagrangian discretization)**.** The finite-dimensional problem in (9.25) can be interpreted as the Eulerian discretization of a flow over the space of measures (9.26). An alternative way to discretize the problem, using the so-called Lagrangian method, would be 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 (9.25) 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 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 (9.28), for which a consistent estimator is (up to a constant term) can be obtained by summing the logarithms of the distances to nearest neighbors

$$\mathcal{F}(X) = \frac{1}{n}\sum_i \log(d_X(x_i)) \quad \text{where} \quad d_X(x) = \min_{x'\in X, x'\neq x} \|x - x'\|, \tag{9.32}$$

see Beirlant et al. [1997] for a review of non-parametric 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 $t_k = \tau k$

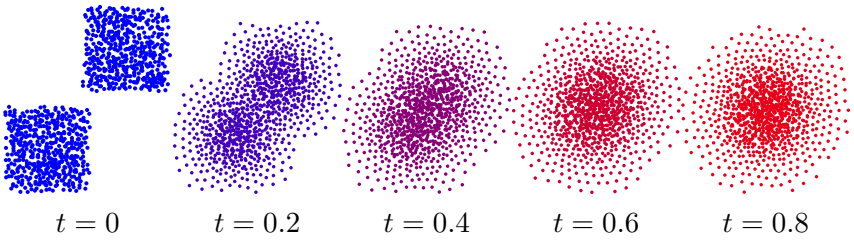


Figure 9.8: Example of gradient flow evolutions using a Lagrangian discretization, for the function $F(\alpha) = \int V \mathrm{d}\alpha - H(\alpha)$, for $V(x) = \|x\|^2$. The entropy is discretized using (9.32). The limiting stationary distribution is a Gaussian.

similarly to (9.23) 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, so 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 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 term of Wasserstein distance, see for instance [Dalalyan and Karagulyan, 2017]. If the function $\mathcal{F}$ is not smooth, one should discretize similarly to (9.24) using implicit Euler steps, i.e. consider

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

In the simplest case of a linear function $F(\alpha) = \int_X V(x) \mathrm{d}\alpha(x)$, then 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).

Remark 9.12 (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 exposed 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 (9.26) 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 an important literature on the numerical resolution of the resulting discretized flow, and we only give 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 2.28) Kinderlehrer and Walkington [1999], Blanchet et al. [2008], Agueh and Bowles [2013], Matthes and Osberger [2014], Blanchet and Carlier [2015]. In higher dimension, it can be tackled using finite elements and finite volume schemes Carrillo et al. [2015], Burger et al. [2010]. Alternative solvers are obtained using Lagrangian schemes (*i.e.* particles systems) Carrillo and Moll [2009], Benamou et al. [2016a], Westdickenberg and Wilkening [2010]. Another direction is to look for discrete flows (typically on discrete grids or graphs) which maintains the properties of their continuous counterparts, and in particular such that the gradient flow of the entropy is the heat equation, see Mielke [2013], Erbar and Maas [2014], Chow et al. [2012], Maas [2011].

An approximate approach to solve the Eulerian discretized problem (9.23) makes use of the entropic regularization. This was initially proposed in Peyré [2015] and refined in Chizat et al. [2017], and theoretically analyzed in Carlier et al. [2017]. Adding an entropic regularization penalty, the problem (9.25) has the form (4.45) when using the identification $G = \iota_{\mathbf{a}^{(\ell)}}$ and F should be replace by τF. One can thus use the iterations (4.47) 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 example of evolution computer with this method. An interesting application of gradient flows is in machine learning, in order to learning the underlying function F that best model some dynamical model of density. This learning can be achieved by solving a smooth non-convex optimization using entropic regularized transport, and automatic differentiation (as advocated in Remark 9.1.3), see Hashimoto et al. [2016].

It should be noted that analyzing the convergence of gradient flows discretized in both time and space is difficult. 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 no evolve, so that the step size should be not too small, as discussed in [Maury and Preux, 2017].

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. Figure 9.10 shows an example of gradient flow corresponding to the celebrated Hele-Shaw model for cells growth [Perthame et al., 2014], which is studied theoretically in [Gallouët and Monsaingeon, 2017, Di Marino and Chizat, 2017]. Such an unbalanced gradient flow can be also approximated using the generalized Sinkhorn algorithm [Chizat et al., 2017].

9.4 Minimum Kantorovitch 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

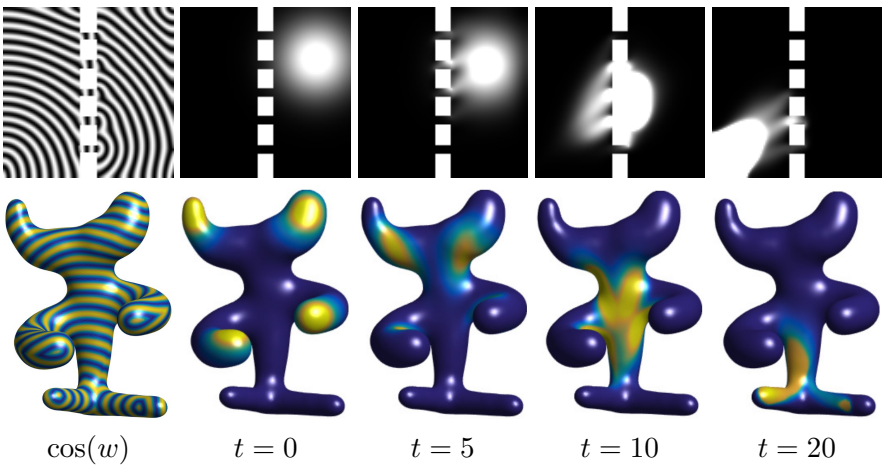


Figure 9.9: Examples of gradient flows evolutions, with drift V and congestion terms (from Peyré [2015]), so that $F(\alpha) = \int_{\mathcal{X}} V(x)\mathrm{d}\alpha(x) + \iota_{\leq \kappa}(\rho_\alpha)$.

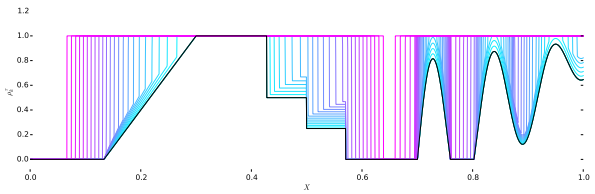


Figure 9.10: Unbalanced OT gradient flow to solve the Hele-Shaw equation (from [Chizat et al., 2017]).

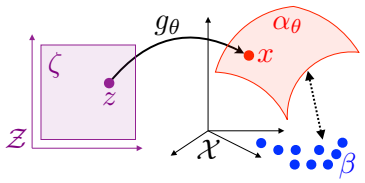


Figure 9.11: Schematic display of the density fitting problem 9.33.

observed empirical input measure β

$$\min_{\theta\in\Theta} \mathcal{L}(\alpha_\theta, \beta) \quad \text{where} \quad \beta = \frac{1}{n}\sum_i \delta_{x_i}, \tag{9.33}$$

where $\mathcal{L}$ is some "loss" function between a discrete and a "continuous" (arbitrary) distribution (see Figure 9.11).

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

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

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

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

This MLE approach is known to lead to optimal estimation procedures in many cases (see for instance Owen [2001]). However, it fails to work when estimating singular distributions, typically when the α_θ does not has a density (so that $\mathcal{L}_{\text{MLE}}(\alpha_\theta, \beta) = +\infty$) or when $(x_i)_i$ are samples from some singular $\bar\beta$ (so that the α_θ should share the same support as β for $\text{KL}(\alpha_\theta|\bar\beta)$ 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 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

$$\mathcal{L}(\alpha,\beta) \overset{\text{def.}}{=} \max_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \left\{ \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{X}} g(x)\mathrm{d}\beta(x) \;:\; (f,g)\in\mathcal{R} \right\}. \tag{9.34}$$

Dual norms exposed in §8.2 correspond to imposing $\mathcal{R} = \{(f,-f) \;:\; f\in B\}$, while optimal transport (2.23) sets $\mathcal{R} = \mathcal{R}(c)$ as defined in (2.24).

For a fixed θ, evaluating the energy to be minimized in (9.33) using such a loss function corresponds to solving a semi-discrete optimal transport, which is the focus of Chapter 5. Minimizing the energy with respect to θ is much more involved, and is typically highly non-convex.

Denoting f_θ a solution to (9.34) when evaluating $\mathcal{E}(\theta) = \mathcal{L}(\alpha_\theta,\beta)$, a sub-gradient is obtained using the formula

$$\nabla\mathcal{E}(\theta) = \int_{\mathcal{X}} [\partial h_\theta(x)]^\top \nabla f_\theta(x)\mathrm{d}\alpha_\theta(x), \tag{9.35}$$

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 semi-discrete problem. As exposed 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 a RKHS Genevay et al. [2016] or a deep-network approximation [Arjovsky et al., 2017]. This however leads to important approximation error 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

$$\mathcal{L}_c(\alpha_\theta, \beta) = \min_{\gamma \in \mathcal{M}(\mathcal{Z}\times\mathcal{X})} \left\{ \int_{\mathcal{Z}\times\mathcal{X}} c(h_\theta(z), x) \mathrm{d}\gamma(z,x) \;:\; \gamma \in \mathcal{U}(\zeta, \beta) \right\}. \tag{9.36}$$

Note that in the semi-discrete case considered here, the objective to be minimized can be actually decomposed as

$$\min_{(\gamma_i)_{i=1}^n} \sum_{i=1}^n \int_{\mathcal{Z}} c(h_\theta(z), x_i) \mathrm{d}\gamma_i(z) \quad \text{where} \quad \sum_{i=1}^n \gamma_i = \zeta, \quad \int_{\mathcal{Z}} \mathrm{d}\gamma_i(z) = \frac{1}{n}, \tag{9.37}$$

where each $\gamma_i \in \mathcal{M}_+^1(\mathcal{Z})$. Once an optimal $(\gamma_{\theta,i})_i$ solving (9.37) 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 (9.35), 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 Kantorovitch Estimators" (MKE), 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. Approximation 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., 2017a,b].

Remark 9.13 (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]). 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]. Another related problem is 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 modelled and solved using OT techniques, see [Courty et al., 2017b,a].

10

Extensions of Optimal Transport

This chapter details several variational problems that are related to (and share the same structure of) the Kantorovitch formulation of optimal transport. The goal is to extend optimal transport to more general settings: several input histograms, un-normalized histograms, more general classes of measures, and transport between measures living in different metric spaces and for which no suitable cost function between the spaces can be defined.

10.1 Multi-marginal Problems

Instead of coupling two input histograms using the Kantorovitch formulation (2.11), one can couple S histograms $(\mathbf{a}_s)_{s=1}^S$, where $\mathbf{a}_s \in \Sigma_{n_s}$, by solving the following multi-marginal problem

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}_s)_s} \langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def.}}{=} \sum_s \sum_{i_s=1}^{n_s} \mathbf{C}_{i_1,\ldots,i_S} \mathbf{P}_{i_1,\ldots,i_S}. \tag{10.1}$$

where the set of valid couplings is

$$\mathbf{U}(\mathbf{a}_s)_s = \left\{ \mathbf{P} \in \mathbb{R}^{n_1 \times \ldots \times n_S} \;:\; \forall\, s, \forall\, i_s, \sum_{\ell \neq s} \sum_{i_\ell=1}^{n_\ell} \mathbf{P}_{i_1,\ldots,i_S} = \mathbf{a}_{s,i_s} \right\}.$$

The entropic regularization scheme (4.2) naturally extends to this setting

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a}_s)_s} \langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon \mathbf{H}(\mathbf{P}).$$

and one can then apply Sinkhorn's algorithm to compute the optimal $\mathbf{P}$ in scaling form

$$\forall\, i=(i_1,\ldots,i_S), \quad \mathbf{P}_i = \mathbf{K}_i \prod_{s=1}^{S} \mathbf{u}_{s,i_s} \quad \text{where} \quad \mathbf{K} \overset{\text{def.}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}},$$

where $\mathbf{u}_s \in \mathbb{R}_+^{n_s}$ are (unknown) scaling vectors, which are iteratively updated, by cycling repeatedly through $s=1,\ldots,S$,

$$\mathbf{u}_{s,i_s} \leftarrow \frac{\mathbf{a}_{s,i_s}}{\sum_{\ell\neq s}\sum_{i_\ell=1}^{n_\ell} \mathbf{K}_i \prod_{r\neq s} \mathbf{u}_{\ell,i_r}} \tag{10.2}$$

Remark 10.1 (General measures)**.** The discrete multimarginal problem (10.1) is generalized to measures $(\alpha_s)_s$ on spaces $(\mathcal{X}_1,\ldots,\mathcal{X}_S)$ by computing a coupling measure

$$\min_{\pi\in\mathcal{U}(\alpha_s)_s} \int_{\mathcal{X}_1\times\ldots\times\mathcal{X}_S} c(x_1,\ldots,x_S)\mathrm{d}\pi(x_1,\ldots,x_S) \tag{10.3}$$

where the set of couplings is

$$\mathcal{U}(\alpha_s)_s \overset{\text{def.}}{=} \left\{ \pi\in\mathcal{M}_+^1(\mathcal{X}_1\times\ldots\times\mathcal{X}_S) \;:\; \forall\, s=1,\ldots,S, P_{s,\sharp}\pi=\alpha_s \right\}$$

where $P_s : \mathcal{X}_1\times\ldots\times\mathcal{X}_S \to \mathcal{X}_s$ is the projection on the s^{th} component, $P_s(x_1,\ldots,x_S)=x_s$, see for instance Gangbo and Swiech [1998a]. We refer to [Pass, 2015, 2012] for a review of the main properties of the multi-marginal OT problem. A typical application of multi-marginal OT is to compute approximation of solutions to quantum chemistry problems, and in particular, for the so-called Density Functional Theory [Cotar et al., 2013]. This problem is obtained when considering the singular Coulomb interaction cost

$$c(x_1,\ldots,x_S) = \sum_{i\neq j} \frac{1}{\|x_i-x_j\|}.$$

Remark 10.2 (Multi-marginal formulation of the barycenter)**.** It is possible to re-cast the linear program optimization (9.11) as an optimization over a single coupling over $\mathcal{X}^{S+1}$ where the first marginal is the barycenter and the other ones are the input measure $(\alpha_s)_{s=1}^S$

$$\min_{\bar{\pi}\in\mathcal{M}_+^1(X^{S+1})} \int_{\mathcal{X}^{S+1}} \sum_{s=1}^S \lambda_s c(x,x_s)\mathrm{d}\bar{\pi}(x_1,\ldots,x_s,x) \tag{10.4}$$

$$\text{subj. to} \quad \forall\, s=1,\ldots,S, \quad P_{s,\sharp}\bar{\pi}=\alpha_s.$$

This stems from the "gluing lemma", which states that given couplings $(\pi_s)_{s=1}^S$ where $\pi_s \in \mathcal{U}(\alpha_s,\alpha)$, one can construct a higher-dimensional coupling $\bar{\pi}\in\mathcal{M}_+^1(X^{K+1})$ with marginals π_s, *i.e.* such that $Q_{s\sharp}\bar{\pi}=\pi_s$ where $Q_s(x_1,\ldots,x_S,x)\overset{\text{def.}}{=}(x_s,x)\in\mathcal{X}^2$. By explicitly minimizing in (10.4) with respect to the first marginal (associated to $x\in\mathcal{X}$), one obtains that solutions α of the barycenter problem (9.11) can be computed as $\alpha = A_{\lambda,\sharp}\pi$ where A_λ is the "barycentric map" defined as

$$A_\lambda : (x_1,\ldots,x_S)\in\mathcal{X}^S \mapsto \operatorname*{argmin}_{x\in\mathcal{X}} \sum_s \lambda_s c(x,x_s)$$

(assuming this map is single-valued), where π is any solution of the multi-marginal problem (10.3) with cost

$$c(x_1,\ldots,x_S)=\sum_\ell \lambda_\ell c(x_\ell, A_\lambda(x_1,\ldots,x_S)). \tag{10.5}$$

For instance, for $c(x,y)=\|x-y\|^2$, one has, removing the constant squared terms,

$$c(x_1,\ldots,x_S)=-\sum_{r\leq s}\lambda_r\lambda_s\langle x_r,\,x_s\rangle,$$

which is a problem studied in Gangbo and Swiech [1998b]. We refer to Agueh and Carlier [2011] for more details. This formula shows that if all the input measures are discrete $\beta_s=\sum_{i_s=1}^{n_s}\mathbf{a}_{s,i_s}\delta_{x_{s,i_s}}$, then the barycenter α is also discrete, and is obtained using the

formula

$$\alpha = \sum_{(i_1,\ldots,i_s)} \pi_{(i_1,\ldots,i_s)} \delta_{A_\lambda(x_{i_1},\ldots,x_{i_s})}.$$

where π is an optimal solution of (10.1) with cost matrix $\mathbf{C}_{i_1,\ldots,i_S} = c(x_{i_1},\ldots,x_{i_s})$ as defined in (10.5). Since π is a nonnegative tensor of $\prod_s n_s$ dimensions obtained as the solution of a linear program with $\sum_s n_s - S + 1$ equality constraints, an optimal solution π with up to $\sum_s n_s - S + 1$ non-zero values can be obtained. A barycenter α with a support of up to $\sum_s n_s - S + 1$ points can therefore be obtained. This direct fact and other considerations in the discrete case can be found in Anderes et al. [2016].

Remark 10.3 (Relaxation of Euler equations)**.** A convex relaxation of Euler equations of incompressible fluid dynamics has been proposed by Brenier [Brenier, 2008, 1990, 1993, 1999, Ambrosio and Figalli, 2009]. Similarly to the setting exposed in §7.6, it corresponds to the problem of finding a probability distribution $\bar{\pi} \in \mathcal{M}_+^1(\bar{\mathcal{X}})$ over the set $\bar{\mathcal{X}}$ of all paths $\gamma : [0,1] \to \mathcal{X}$, which describes the movement of particules in the fluid. This is a relaxed version of the initial partial differential equation model because, as in the Kantorovitch formulation of OT, mass can be split. The evolution with time does not necessarily define anymore a diffemorphism of the underlying space $\mathcal{X}$. The dynamic of the fluid is obtained by minimizing as in (7.17) the energy $\int_0^1 \|\gamma'(t)\|^2 \, \mathrm{d}t$ of each path. The difference with OT over the space of paths is the additional incompressibilty of the fluid. This incompressibilty is taken care of by imposing that the density of particules should be uniform at any time $t \in [0,1]$ (and not just imposed at initial and final times $t \in \{0,1\}$ as in classical OT). Assuming $\mathcal{X}$ is compact and denoting $\rho_{\mathcal{X}}$ the uniform distribution on $\mathcal{X}$, this reads $\bar{P}_{t,\sharp}\bar{\pi} = \rho_{\mathcal{X}}$ where $\bar{P}_t : \gamma \in \bar{\mathcal{X}} \to \gamma(t) \in \mathcal{X}$. One can discretize this problem by replacing a continuous path $(\gamma(t))_{t\in[0,1]}$ by a sequence of S points $(x_{i_1}, x_{i_2}, \ldots, x_{i_S})$ on a grid $(x_k)_{k=1}^n \subset \mathcal{X}$, and $\bar{\Pi}$ is represented by a S-way coupling $\mathbf{P} \in \mathbb{R}^{n^S} \in \mathcal{U}(\mathbf{a}_s)_s$ where the marginals are uniform $\mathbf{a}_s = n^{-1}\mathbb{1}_n$. The cost of the corresponding multi-marginal

problem is then

$$\mathbf{C}_{i_1,\dots,i_S} = \sum_{s=1}^{S-1} \|x_{i_s} - x_{i_{s+1}}\|^2 + R\left\|x_{\sigma(i_1)} - x_{i_S}\right\|^2. \tag{10.6}$$

Here R is a large enough penalization constant, which is here to enforce the movement of particules between initial and final times, which is prescribed by a permutation $\sigma : [\![n]\!] \to [\![n]\!]$. This resulting multi-marginal problem is implemented efficiently in conjunction with Sinkhorn iterations (10.2) using the special structure of the cost, as detailed in Benamou et al. [2015]. Indeed, in place of the $O(n^S)$ cost required to compute the denominator appearing in (10.2), one can decompose it as a succession of S matrix-vector multiplications, hence with a low cost Sn^2. Note that other solvers have been proposed, for instance using the semi-discrete framework exposed in §5.2, see [de Goes et al., 2015, Gallouët and Mérigot, 2017].

10.2 Unbalanced Optimal Transport

A major bottleneck of "classical" optimal transport is that it requires the two input measures (α, β) to have the same total mass. While many workaround have been proposed (including re-normalizing the input measure, or using dual norms such as detailed in 8.2), it is only recently that a satisfying unifying theory have been developed. We only sketch here a simple but important particular case.

Following Liero et al. [2015], to account for arbitrary positive histograms $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$, the initial Kantorovitch formulation (2.11) is "relaxed" by only penalizing marginal deviation using some divergence $\mathbf{D}_\varphi$ (defined in (8.3)). This equivalently corresponds to minimizing an OT distance between approximate measures.

$$\mathrm{L}_{\mathbf{C}}^{\tau}(\mathbf{a}, \mathbf{b}) = \min_{\tilde{\mathbf{a}}, \tilde{\mathbf{b}}} \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) + \tau \mathbf{D}_\varphi(\mathbf{a}, \tilde{\mathbf{a}}) + \tau \mathbf{D}_\varphi(\mathbf{b}, \tilde{\mathbf{b}}) \tag{10.7}$$

$$= \min_{\mathbf{P} \in \mathbb{R}_+^{n \times m}} \langle \mathbf{C}, \mathbf{P} \rangle + \tau_1 \mathbf{D}_\varphi(\mathbf{P}\mathbb{1}_m | \mathbf{a}) + \tau_2 \mathbf{D}_\varphi(\mathbf{P}^\top \mathbb{1}_m | \mathbf{a}). \tag{10.8}$$

where (τ_1, τ_2) controls how much mass variations are penalized as op-

posed to transportation of the mass. In the limit $\tau_1 = \tau_2 \to +\infty$, assuming $\sum_i \mathbf{a}_i = \sum_j \mathbf{b}_j$ (the "balanced" case), one recovers the original optimal transport formulation with hard marginal constraint (2.11).

This formalism recovers many different previous works, for instance introducing for $\mathbf{D}_\varphi$ an ℓ^2 norm [Benamou, 2003] or a ℓ^1 norm as in partial transport [Figalli, 2010, Caffarelli and McCann, 2010]. A case of particular importance is when using $\mathbf{D}_\varphi = \mathbf{KL}$ the Kulback-Leibler divergence, as detailed in Remark 10.5. For this cost, in the limit $\tau_1 = \tau_2 \to 0$, one obtains the so-called squared Hellinger distance (see also Example 8.3)

$$\mathrm{L}_{\mathbf{C}}^{\tau}(\mathbf{a},\mathbf{b}) \overset{\tau\to 0}{\longrightarrow} \mathfrak{h}^2(\mathbf{a},\mathbf{b}) = \sum_i (\sqrt{\mathbf{a}_i} - \sqrt{\mathbf{b}_i})^2.$$

Sinkhorn's iterations (4.15) can be adapted to this problem by making use of the generalized algorithm detailed in §4.6. This means that the solution has the form (4.12) and that the scalings are updated as

$$\mathbf{u} \leftarrow \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}\right)^{\frac{\tau_1}{\tau_1+\varepsilon}} \quad \text{and} \quad \mathbf{v} \leftarrow \left(\frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}}\right)^{\frac{\tau_2}{\tau_2+\varepsilon}}. \tag{10.9}$$

Remark 10.4 (Generic measure)**.** For (α,β) two arbitrary measure, the unbalanced version (also called "log-entropic") of (2.15) read

$$\mathcal{L}_c^{\tau}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \tau\mathcal{D}_\varphi(P_{1,\sharp}\pi|\alpha) + \tau\mathcal{D}_\varphi(P_{2,\sharp}\pi|\beta),$$

where divergences $\mathcal{D}_\varphi$ between measures are defined in (8.1). In the special case $c(x,y) = \|x-y\|^2$, $\mathcal{D}_\varphi = \mathrm{KL}$, $\mathcal{L}_c^{\tau}(\alpha,\beta)^{1/2}$ is the Gaussian-Hellinger distance [Liero et al., 2015], and it is shown to be a distance on $\mathcal{M}_+^1(\mathbb{R}^d)$.

Remark 10.5 (Wasserstein-Fisher-Rao)**.** For the particular choice of cost

$$c(x,y) = -\log\cos(\min(d(x,y)/\kappa, \pi/2))$$

where κ is some cutoff distance, and using $\mathcal{D}_\varphi = \mathrm{KL}$, then

$$\mathrm{WFR}(\alpha,\beta) \overset{\text{def.}}{=} \mathcal{L}_c^\tau(\alpha,\beta)^{\frac{1}{2}}$$

is the so-called Wasserstein-Fisher-Rao or Hellinger-Kantorovitch distance. In the special case $\mathcal{X} = \mathbb{R}^d$, this static (Kantorovitch-like) formulation matches its dynamical counterparts (7.15), as proved independently by Liero et al. [2015], Chizat et al. [2015]. This dynamical formulation is detailed in §7.4.

The barycenter problem (9.11) can be generalized to handle an unbalanced setting by replacing $\mathcal{L}_c$ with $\mathcal{L}_c^\tau$. Figure (10.1) shows the resulting interpolation, providing a good illustration of the usefulness of the relaxation parameter τ. The input measures are mixtures of two Gaussians with unequal mass. Classical OT requires the leftmost bump to be split in two and gives an non-regular interpolation. In sharp contrast, unbalanced OT allows the mass to vary during interpolation, so that the bumps are not split and local modes of the distributions are smoothly matched. Using finite values for τ (recall that OT is equivalent to $\tau = \infty$) is thus important to prevent irregular interpolations that arise because of mass splitting, which happen because of a "hard" mass conservation constraint. The resulting optimization problem can be tackled numerically using entropic regularization and the generalized Sinkhorn algorithm detailed in §4.6.

In practice, unbalanced OT techniques seem to outperform classical OT for applications (such as in imaging or machine learning) where the input data is noisy or not perfectly known. They are also crucial when the signal strength of a measure, as measured by its total mass, must be accounted for, or when normalization is not meaningful. This was the original motivation of Frogner et al. [2015], whose goal was to compare sets of word labels used to describe images. Unbalanced OT and the corresponding Sinkhorn iterations have also been used for application to the dynamics of cells in Schiebinger et al. [2017].

Remark 10.6 (Connexion with Dual norms)**.** A particularly simple setup to account for mass variation is to use dual norms, as detailed in §8.2. By choosing a compact set $B \subset \mathcal{C}(\mathcal{X})$ one obtains a norm defined on the

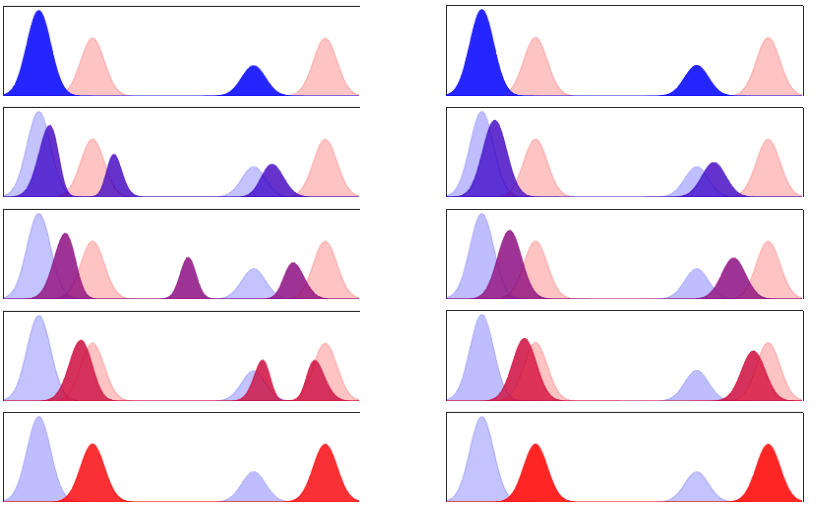


Figure 10.1: Influence of relaxation parameter τ on unbalanced barycenters. Top to bottom displays the evolution of the barycenter between two input measures.

whole space $\mathcal{M}(\mathcal{X})$ (in particular, the measures do not need to be positive). A particular instance of this setting is the flat norm (8.11), which is recovered as a special instance of unbalanced transport, when using $\mathcal{D}_\varphi(\alpha|\alpha') = \|\alpha - \alpha'\|_{\mathrm{TV}}$ to be the total variation norm (8.9), see for instance Hanin [1992], Lellmann et al. [2014]. We also refer to Schmitzer and Wirth [2017] for a general framework to define Wasserstein-1 unbalanced transport.

10.3 Problems with Extra Constraints on the Couplings

Many other OT-like problems have been proposed in the literature. They typically correspond to adding extra constraints $\mathcal{C}$ on the initial OT problem (2.15)

$$\min_{\pi \in \mathcal{U}(\alpha,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \;:\; \pi \in \mathcal{C} \right\}. \tag{10.10}$$

Let us give two representative examples. The optimal transport with capacity constraint [Korman and McCann, 2015] corresponds to im-

posing that the density ρ_π (for instance with respect to the Lebesgue measure) is upper bounded

$$\mathcal{C} = \{\pi \;:\; \rho_\pi \leq \kappa\} \tag{10.11}$$

for some $\kappa > 0$. This constraints rules out any singular and thin coupling that would be localized on a Monge map. The martingale transport problem (see for instance Galichon et al. [2014], Dolinsky and Soner [2014], Tan and Touzi [2013], Beiglböck et al. [2013]), which finds many applications in finance, imposes the so-called martingale constraint on the conditional mean of the coupling, when $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$:

$$\mathcal{C} = \left\{\pi \;:\; \forall\, x \in \mathbb{R}^d, \int_{\mathbb{R}^d} y \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\mathrm{d}\beta(y) = x\right\}. \tag{10.12}$$

Note that this constraints any admissible coupling to be such that its barycentric projection map (4.19) be equal to the identity. For arbitrary (α, β), this set $\mathcal{C}$ is typically empty, but necessary and sufficient conditions exists (α and β should be in "convex order") to ensure $\mathcal{C} \neq \emptyset$ so that (α, β) satisfy a martingale constraint. This constraint can be difficult to enforce numerically when discretizing an existing problem. It also forbids the solution to concentrate on a single Monge-map, and can lead to couplings concentrated on the union of several graphs (a "multi-valued" Monge-map), or even more complicated support sets. Using an entropic penalization as in (4.9), one can solve approximately (10.10) using the Dykstra algorithm as explained in Benamou et al. [2015], which is a generalization of Sinkhorn's algorithm exposed in §4.2. This requires to compute the projection onto $\mathcal{C}$ for the KL divergence, which is straightforward for (10.11), but cannot be done in closed form (10.12) and thus necessitates sub-iterations, see Guo and Obloj [2017] for more details.

10.4 Sliced Wasserstein Distance and Barycenters

One can define a distance between two measures (α, β) defined on $\mathbb{R}^d$ by aggregating 1-D Wasserstein distances between their projections on 1-D lines. This defines

$$\mathrm{SW}(\alpha,\beta)^2 \overset{\text{def.}}{=} \int_{\mathbf{S}^d} \mathcal{W}_2(P_{\theta,\sharp}\alpha, P_{\theta,\sharp}\beta)^2 \mathrm{d}\theta \tag{10.13}$$

where $\mathbf{S}^d = \left\{\theta \in \mathbb{R}^d \;:\; \|\theta\| = 1\right\}$ is the d-dimensional sphere, and $P_\theta : x \in \mathbb{R}^d \to \mathbb{R}$ is the projection. This approach is detailed in Bonneel et al. [2015], following ideas from Marc Bernot. It is related to the problem of Radon inversion over measure spaces Abraham et al. [2017].

Lagrangian discretization and stochastic gradient descent. The advantage of this functional is that 1-D Wasserstein distances are simple to compute, as detailed in §2.6. In the specific case where $m = n$ and

$$\alpha = \sum_{i=1}^N \delta_{x_i} \quad \text{and} \quad \beta = \sum_{i=1}^N \delta_{y_i}, \tag{10.14}$$

this is achieved by simply sorting points

$$\mathrm{SW}(\alpha,\beta)^2 = \int_{\mathbf{S}^d} \left(\sum_{i=1}^n |\langle x_{\sigma_\theta(i)} - y_{\kappa_\theta(i)}, \theta\rangle|^2 \right) \mathrm{d}\theta$$

where $\sigma_\theta, \kappa_\theta \in \mathrm{Perm}(n)$ are the permutation ordering in increasing order respectively $(\langle x_i, \theta\rangle)_i$ and $(\langle y_i, \theta\rangle)_i$.

Fixing the vector y, the function $\mathcal{E}_\beta(x) \overset{\text{def.}}{=} \mathrm{SW}(\alpha,\beta)^2$ is smooth, and one can use this function to define a mapping by gradient descent

$$x \leftarrow x - \tau \nabla \mathcal{E}_\beta(x) \quad \text{where} \tag{10.15}$$

$$\nabla \mathcal{E}_\beta(x)_i = 2 \int_{\mathbf{S}^d} \Big(\langle x_i - y_{\kappa_\theta \circ \sigma_\theta^{-1}(i)}, \theta \rangle \theta \Big) \mathrm{d}\theta.$$

To make the method tractable, one can use a stochastic gradient descent (SGD), replacing this integral with a discrete sum against randomly drawn directions $\theta \in \mathbf{S}^d$ (see §5.4 for more details on SGD). The flow (10.15) can be understood as (Langrangian implementation of) a Wasserstein gradient flow (in the sense of §9.3) of the function $\alpha \mapsto \mathrm{SW}(\alpha,\beta)^2$. Numerically, one finds that this flow has no local minimizer and that it thus converges to $\alpha = \beta$. The usefulness of the Lagrangian solver is that, at convergence, it defines a matching (similar to a Monge map) between the two distributions. This method has been used successfully for color transfer and texture synthesis in [Rabin et al., 2011] and is related to the alternate minimization approach detailed in Pitié et al. [2007].

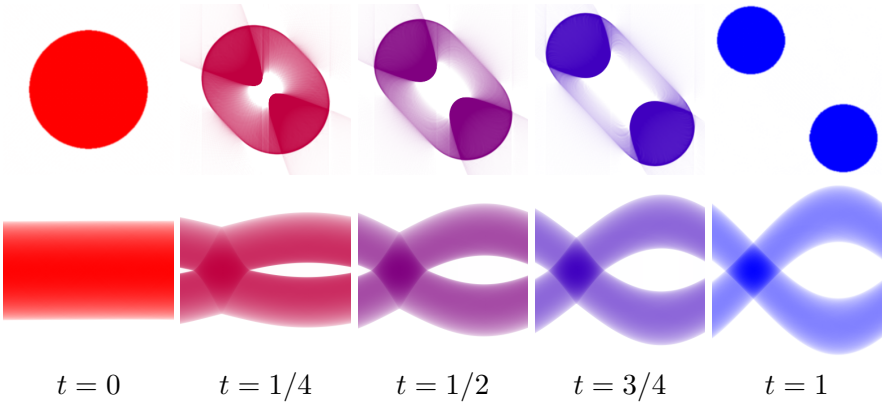


Figure 10.2: Example of sliced barycenters computation using the Radon transform (as defined in (10.20)). Top: barycenters α_t for $S = 2$ two input and weights $(\lambda_1, \lambda_2) = (1-t, t)$. Bottom: their Radon transform $\mathcal{R}(\alpha_t)$ (the horizontal axis being the orientation angle θ).

It is now simple to extend this Lagrangian scheme to compute approximate "sliced" barycenters of measures, by mimicking the Frechet definition of Wasserstein barycenters (9.11) and minimizing

$$\min_{\alpha \in \mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \, \mathrm{SW}(\alpha, \beta_s)^2. \tag{10.16}$$

given a set $(\beta_s)_{s=1}^S$ of fixed input measure. Using a Lagrangian discretization of the form (10.14) for both α and the $(\beta_s)_s$, one can perform the non-convex minimization over the position $x = (x_i)_i$

$$\min_x \mathcal{E}(x) \overset{\text{def.}}{=} \sum_s \lambda_s \mathcal{E}_{\beta_s}(x), \quad \text{and} \quad \nabla \mathcal{E}(x) = \sum_s \lambda_s \nabla \mathcal{E}_{\beta_s}(x), \tag{10.17}$$

by gradient descent using formula (10.15) to compute $\nabla \mathcal{E}_{\beta_s}(x)$ (coupled with a random sampling of the direction θ).

Eulerian discretization and Radon transform. A related way to compute approximated sliced barycenter, without resorting the an iterative minimization scheme, is to use the fact that (10.13) computes a distance between the Radon transforms $\mathcal{R}(\alpha)$ and $\mathcal{R}(\beta)$ where

$$\mathcal{R}(\alpha) \overset{\text{def.}}{=} (P_{\theta,\sharp}\alpha)_{\theta \in \mathbf{S}^d}.$$

A crucial point is that the Radon transform is invertible, and that its inverse can be computed using a filtered backprojection formula. Given a collection of measures $\rho = (\rho_\theta)_{\theta \in \mathbf{S}^d}$, one define the filtered backprojection operator as

$$\mathcal{R}^+(\rho) = C_d \Delta^{\frac{d-1}{2}} \mathcal{B}(\rho) \tag{10.18}$$

where $\xi = \mathcal{B}(\rho) \in \mathcal{M}(\mathbb{R}^d)$ is the measure defined through the relation

$$\forall\, g \in \mathcal{C}(\mathbb{R}^d), \quad \int_{\mathbb{R}^d} g(x) \mathrm{d}\xi(x) = \int_{\mathbf{S}^d} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} g(r\theta + U_\theta z) \mathrm{d}\rho_\theta(r) \mathrm{d}z \mathrm{d}\theta \tag{10.19}$$

where U_θ is any orthogonal basis of $\theta^\perp$, and where $C_d \in \mathbb{R}$ is a normalizing constant which depends on the dimension. Here $\Delta^{\frac{d-1}{2}}$ is a fractional Laplacian, which is the high-pass filter defined over the Fourier domain as $\hat{\Delta}^{\frac{d-1}{2}}(\omega) = \|\omega\|^{d-1}$. The definition of the backprojection (10.19) adds up the contribution of all the measures $(\rho_\theta)_\theta$ by extending each one as being constant in the directions orthogonal to θ. One then has the left-inverse relation $\mathcal{R}^+ \circ \mathcal{R} = \mathrm{Id}_{\mathcal{M}(\mathbb{R}^d)}$, so that $\mathcal{R}^+$ is a valid reconstruction formula.

In order to compute barycenters of input densities, it makes sense to replace formula (9.11) by its equivalent using Radon transform, and thus consider independently for each θ the 1D barycenter problem

$$\rho_\theta^\star \in \operatorname*{argmin}_{(\rho_\theta \in \mathcal{M}_+^1(\mathbb{R}))} \sum_{s=1}^S \lambda_s \mathcal{W}_2(\rho_\theta, P_{\theta,\sharp}\beta_s)^2. \tag{10.20}$$

Each 1-D barycenter problem is easily computed using the monotone rearrangement as detailed in Remark (9.5). The Radon approximation $\alpha_R \overset{\text{def.}}{=} \mathcal{R}^+(\rho^\star)$ of a sliced barycenter solving (9.11) is then obtained by the inverse Radon transform $\mathcal{R}^+$. Note that in general, α_R is not a solution to (9.11) because the Radon transform is not surjective, so that $\rho^\star$, which is obtained as a barycenter of the Radon transforms $\mathcal{R}(\beta_s)$ does not necessarily belong to the range of $\mathcal{R}$. But numerically it seems in practice to be almost the case [Bonneel et al., 2015]. Numerically, this Radon transform formulation is very effective for input measures and barycenters discretized on a fixed grid (e.g. a uniform grid for images), and $\mathcal{R}$ and well as $\mathcal{R}^+$ are computed approximately on this

Figure 10.3: Comparison of barycenters computed using Radon transform (10.20) (Eulerian discretization), Lagrangian discretization (10.17), and Wasserstein OT (computed using Sinkhorn iterations 9.17).

grid using fast algorithms (see for instance Averbuch et al. [2001]). Figure (10.2) illustrates this computation of barycenters (and highlights the way the Radon transforms are interpolated), while Figure (10.3) shows a comparison of the Radon barycenters (10.20) and the ones obtained by Lagrangian discretization (10.17).

Sliced Wasserstein Kernels Beside its computational simplicity, another advantage of the sliced Wasserstein distance is that it is isometric to a Euclidean distance (it is thus an "Hilbertian" metric), as detailed in Remark 2.28 (see in particular formula (2.34)). As highlighted in §8.3, this should be contrasted with the Wasserstein distance $\mathcal{W}_2$ on $\mathbb{R}^d$, which is not Hilbertian in dimension $d \geq 2$. It is thus possible to use this sliced distance to equip the space of distributions $\mathcal{M}_1^+(\mathbb{R}^d)$ with a reproducing kernel Hilbert space structure (as detailed in §8.3). One can for instance use the exponential and energy distance kernels

$$k(\alpha,\beta) = e^{-\frac{\mathrm{SW}(\alpha,\beta)^p}{2\sigma^p}} \quad \text{and} \quad k(\alpha,\beta) = -\mathrm{SW}(\alpha,\beta)^p$$

for $1 \leq p \leq 2$ for the exponential kernels, and $0 < p < 2$ for the energy distance kernels. This means that for any collection $(\alpha_i)_i$ of input measures, the matrix $(k(\alpha_i,\alpha_j))_{i,j}$ is symmetric positive semi-definite. It is possible to use these kernels to perform a variety of machine learning tasks using the "kernel trick", for instance in regression, classification

(SVM and logistic), clustering (K-means) and dimensionality reduction (PCA) [Hofmann et al., 2008]. We refer to Kolouri et al. [2016] for details and applications.

10.5 Transporting Vectors and Matrices

Real valued measures $\alpha \in \mathcal{M}(\mathcal{X})$ are easily generalized to vector valued measures $\alpha \in \mathcal{M}(\mathcal{X}; \mathbb{V})$ where $\mathbb{V}$ is some vector space. For notational simplicity, we assume $\mathbb{V}$ is Euclidean and equipped with some inner product $\langle \cdot, \cdot \rangle$ (typically $\mathbb{V} = \mathbb{R}^d$ and the inner product is the canonical one). Thanks to this inner product, vector valued measures are identified with the dual of continuous functions $g : \mathcal{X} \to \mathbb{V}$, *i.e.* for any such g, one defines its integration against the measure as

$$\int_{\mathcal{X}} g(x) \mathrm{d}\alpha(x) \in \mathbb{R} \tag{10.21}$$

which is a linear operation on g and α. A discrete measure has the form $\alpha = \sum_i \mathbf{a}_i \delta_{x_i}$ where $(x_i, a_i) \in \mathcal{X} \times \mathbb{V}$ and the integration formula (10.21) simply reads

$$\int_{\mathcal{X}} g(x) \mathrm{d}\alpha(x) = \sum_i \langle \mathbf{a}_i,\, g(x_i) \rangle \in \mathbb{R}.$$

Equivalently, if $\mathbb{V} = \mathbb{R}^d$, then such a α can be viewed as a collection $(\alpha_s)_{s=1}^d$ of d "classical" real valued measures (its coordinates), writing

$$\int_{\mathcal{X}} g(x) \mathrm{d}\alpha(x) = \sum_{s=1}^{d} \int_{\mathcal{X}} g_s(x) \mathrm{d}\alpha_s(x),$$

where $g(x) = (g_s(x))_{s=1}^d$ are the coordinates of g in the canonical basis.

Dual norms. It is non-trivial, and in fact in general impossible, to extend OT distances to such a general setting. Even coping with real valued measures taking both positive and negative values is difficult. The only simple option is to consider dual norms, as defined in §8.2. Indeed, formula (6.3) readily extends to $\mathcal{M}(\mathcal{X}; \mathbb{V})$ by considering B to be a subset of $\mathcal{C}(\mathcal{X}; \mathbb{V})$. So in particular, $\mathcal{W}_1$, the flat norm and MMD norms can be computed for vector valued measures.

OT over cone-valued measures. It is possible to define more advanced OT distances when α is restricted to be in a subset $\mathcal{M}(\mathcal{X};\mathcal{V}) \subset \mathcal{M}(\mathcal{X};\mathbb{V})$. The set $\mathcal{V}$ should be a positively 1-homogeneous convex cone of $\mathbb{V}$

$$\mathcal{V} \overset{\text{def.}}{=} \left\{ \lambda u \;:\; \lambda \in \mathbb{R}^+, u \in \mathcal{V}_0 \right\}$$

where $\mathcal{V}_0$ is a compact convex set. A typical example is the set of positive measures where $\mathcal{V} = \mathbb{R}^d_+$. Dynamical convex formulations of OT over such a cone have been proposed, see Zinsl and Matthes [2015]. This has been applied to model the distribution of chemical components. Another important example is the set of positive symmetric matrices $\mathcal{V} = \mathcal{S}^d_+ \subset \mathbb{R}^{d\times d}$. It is of course possible to use dual norms over this space, by treating matrices as vectors, see for instance Ning and Georgiou [2014]. Dynamical convex formulations for OT over such a cone have been provided Chen et al. [2016b], Jiang et al. [2012]. Some static (Kantorovitch-like) formulations have been also proposed Ning et al. [2015], Peyré et al. [2017], but a mathematically sound theoretical framework is still missing. In particular, it is unclear if these static approaches define distances for vector valued measures, and if they relate to some dynamical formulation. Figure 10.4 example of tensor interpolation obtained using the method detailed in Peyré et al. [2017], which proposes a generalization of Sinkhorn algorithms using Quantum relative entropy (10.22) to deal with tensor fields.

OT over positive matrices. A related but quite different setting is to replace discrete measures, *i.e.* histograms $\mathbf{a} \in \Sigma_n$ by positive matrices with unit trace $A \in \mathcal{S}^+_n$ such that $\mathrm{tr}(A) = 1$. The rationale is that the eigenvalues $\lambda(A) \in \Sigma_n$ of A plays the role of an histogram, but one has also to take care of the rotations of the eigenvector, so that this problem is more complicated.

One can extend several divergences introduced in §8.1 to this setting. For instance, the Bures metric (2.40) is a generalization of the Hellinger distance (defined in Remark 8.3), since they are equal on positive diagonal matrices. One can also extend the Kulback-Leibler divergence (4.6) (see also Remark 8.1) is generalized to positive matri-

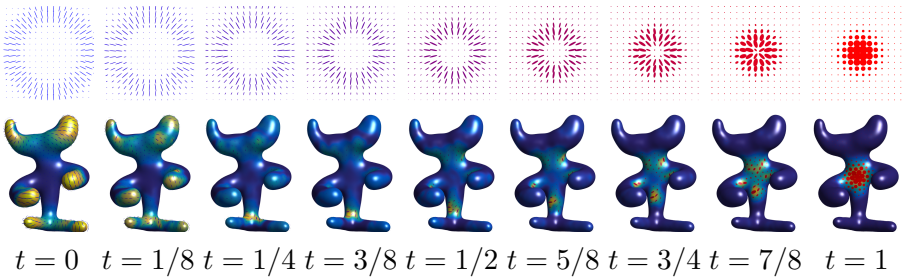


Figure 10.4: Interpolations between two input fields of positive semidefinite matrices (displayed at times $t \in \{0, 1\}$ using ellipses) on some domain (here, a 2-D planar square and a surface mesh), using the method detailed in Peyré et al. [2017]. Unlike linear interpolation schemes, this OT-like method transports the "mass" of the tensors (size of the ellipses) as well as their anisotropy and orientation.

ces as

$$\mathbf{KL}(A|B) \stackrel{\text{def.}}{=} \operatorname{tr}\left(P \log(P) - P \log(Q) - P + Q\right) \tag{10.22}$$

where $\log(\cdot)$ is the matrix logarithm. This matrix $\mathbf{KL}$ is convex with both of its arguments.

It is possible to solve convex dynamic formulations to define OT distances between such matrices Carlen and Maas [2014], Chen et al. [2016b, 2017]. There also exists an equivalent of Sinkhorn's algorithm, which is due to Gurvits Gurvits [2004] and has been extensively studied in Georgiou and Pavon [2015], see also the review paper Idel [2016]. It is known to converge only in some cases, but seems empirically to always work.

10.6 Gromov-Wasserstein Distances

Optimal transport relies on some ground cost in order to compare histograms or measures. A typical setup corresponds to measures defined on the same metric space. We describe here an approach which deals with measures defined on different spaces, but which requires solving a more difficult non-convex optimization problem.

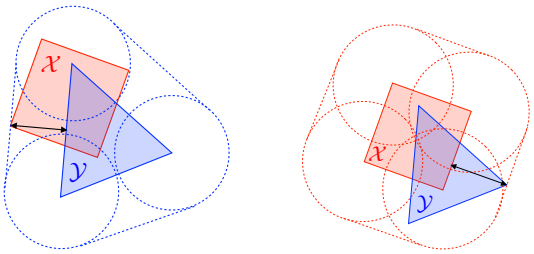


Figure 10.5: Computation of the Hausdorff distance in $\mathbb{R}^2$.

10.6.1 Hausdorff distance

The Hausdorff distance between two sets $A, B \subset \mathcal{Z}$ for some metric $d_{\mathcal{Z}}$ is

$$\mathcal{H}_{\mathcal{Z}}(A,B) \overset{\text{def.}}{=} \max\Big(\sup_{a\in A}\inf_{b\in B} d_{\mathcal{Z}}(a,b), \sup_{b\in B}\inf_{a\in A} d_{\mathcal{Z}}(a,b)\Big).$$

This defines a distance between compact sets $\mathcal{K}(\mathcal{Z})$ of $\mathcal{Z}$, and if $\mathcal{Z}$ is compact, then $(\mathcal{K}(\mathcal{Z}), \mathcal{H}_{\mathcal{Z}})$ is itself compact, see [Burago et al., 2001].

Following [Mémoli, 2011], one remarks that this distance between sets (A, B) can be defined similarly to the Wasserstein distance between measures (which should be somehow understood as "weighted" sets). One replace the measures couplings (2.14) by sets couplings

$$\mathcal{R}(A,B) \overset{\text{def.}}{=} \left\{ R \in \mathcal{X}\times\mathcal{Y} \;:\; \begin{array}{l} \forall\, a\in A, \exists b\in B, (a,b)\in R \\ \forall\, b\in B, \exists a\in A, (a,b)\in R \end{array} \right\}.$$

With respect to Kantorovitch problem (2.15), one should replace integration (since one does not have access to measures) by maximization, and one has

$$\mathcal{H}_{\mathcal{Z}}(A,B) = \inf_{R\in\mathcal{R}(A,B)} \sup_{(a,b)\in R} d(a,b). \tag{10.23}$$

Note that the support of a measure coupling $\pi \in \mathcal{U}(\alpha,\beta)$ is a set coupling between the supports, i.e. $\mathrm{Supp}(\pi) \in \mathcal{R}(\mathrm{Supp}(\alpha), \mathrm{Supp}(\beta))$. The Hausdorff distance is thus connected to the ∞-Wasserstein distance and one has $\mathcal{H}(A,B) \leq \mathcal{W}_\infty(\alpha,\beta)$ for any measure (α,β) whose supports are (A,B).

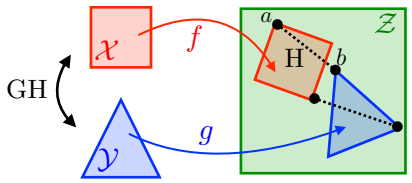


Figure 10.6: The GH approach to compare two metric spaces.

10.6.2 Gromov-Hausdorff distance

The Gromov-Hausdorff distance [Gromov, 2007] (see also [Edwards, 1975]) is a way to measure how two metric spaces $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ by quantifying how much they are isometric one from each other. It is defined as a the minimum Hausdorff distance between every possible isometric embedding of the two spaces in a third one

$$\mathcal{GH}(d_{\mathcal{X}}, d_{\mathcal{Y}}) \overset{\text{def.}}{=} \inf_{\mathcal{Z}, f, g} \left\{ \mathcal{H}_{\mathcal{Z}}(f(\mathcal{X}), g(\mathcal{Y})) \;:\; \begin{array}{c} f : \mathcal{X} \overset{\text{isom}}{\longrightarrow} \mathcal{Z} \\ g : \mathcal{Y} \overset{\text{isom}}{\longrightarrow} \mathcal{Z} \end{array} \right\}.$$

Here, the constraint is that f must be an isometric embedding, meaning that $d_{\mathcal{Z}}(f(x), f(x')) = d_{\mathcal{X}}(x, x')$ for any $(x, x') \in \mathcal{X}^2$ (similarly for g). One can show that $\mathcal{GH}$ defines a distance between compact metric space up to isometry, so that in particular $\mathcal{GH}(d_{\mathcal{X}}, d_{\mathcal{Y}}) = 0$ if and only if there exists an isometry $h : \mathcal{X} \to \mathcal{Y}$, i.e. f is bijective and $d_{\mathcal{Y}}(h(x), h(x')) = d_{\mathcal{X}}(x, x')$ for any $(x, x') \in \mathcal{X}^2$.

Similarly to (10.23) and as explained in [Mémoli, 2011], it is possible to re-write equivalently the Gromov-Hausdorff distance using couplings as follow

$$\mathcal{GH}(d_{\mathcal{X}}, d_{\mathcal{Y}}) = \frac{1}{2} \inf_{R \in \mathcal{R}(\mathcal{X}, \mathcal{Y})} \sup_{((x,y),(x',y')) \in R^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{X}}(y, y')|.$$

For discrete spaces $\mathcal{X} = (x_i)_{i=1}^n, \mathcal{Y} = (y_j)_{j=1}^m$ represented using a distance matrix $\mathbf{D} = (d_{\mathcal{X}}(x_i, x_{i'}))_{i,i'} \in \mathbb{R}^{n \times n}$, $\mathbf{D}' = (d_{\mathcal{Y}}(y_j, y_{j'}))_{j,j'} \in \mathbb{R}^{m \times m}$, one can re-write this optimization using binary matrices $\mathbf{R} \in \{0,1\}^{n \times m}$ indicating the support of the set couplings R as follow

$$\mathrm{GH}(\mathbf{D}, \mathbf{D}') = \frac{1}{2} \inf_{\mathbf{R}\mathbb{1} > 0, \mathbf{R}^\top \mathbb{1} > 0} \max_{(i,i',j,j')} \mathbf{R}_{i,j} \mathbf{R}_{j,j'} |\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|. \quad (10.24)$$

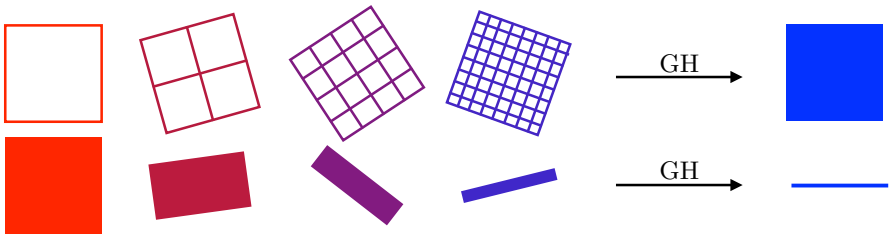


Figure 10.7: GH limit of sequences of metric spaces.

The initial motivation of the GH distance is to define and study limits of metic spaces, as illustrated in Figure 10.7, and we refer to [Burago et al., 2001] for details. There is an explicit description of the geodesics for the GW distance [Chowdhury and Mémoli, 2016], which is very similar to the one of Gromov-Wasserstein space detailed in Remark 10.8.

The underlying optimization problem (10.24) is highly non-convex, and computing the global minimum is untractable. It has been approached numerically using approximation schemes, and has found applications in vision and graphics for shape matching [Mémoli and Sapiro, 2005, Bronstein et al., 2006].

It often desirable to "smooth" the definition of the Hausdorff distance by replacing the maximization by an integration. This in turn necessitates the introduction of measures, and it is one of the motivations for the definition of the Gromov-Wasserstein distance in the next section.

10.6.3 Gromov-Wasserstein Distance

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

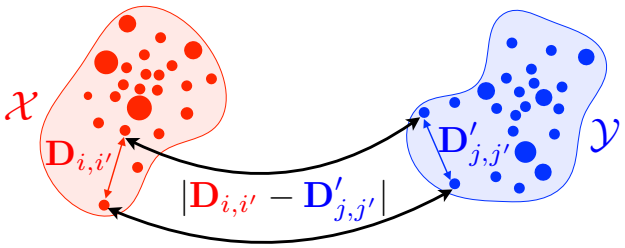


Figure 10.8: The GW approach to compare two metric measure spaces.

Wasserstein problem reads

$$\mathrm{GW}((\mathbf{a},\mathbf{D}),(\mathbf{b},\mathbf{D}'))^2 \stackrel{\text{def.}}{=} \tag{10.25}$$
$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}) \stackrel{\text{def.}}{=} \sum_{i,j,i',j'} |\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|^2 \mathbf{P}_{i,j}\mathbf{P}_{i',j'}.$$

This problem is similar to the GH problem (10.24) when replacing maximization by a sum, and set couplings by measure couplings. This is a non-convex problem, which can be recast as a Quadratic Assignment Problem (QAP) [Loiola et al., 2007] and is in full generality NP-hard to solve for arbitrary inputs. It is in fact equivalent to a graph matching problem [Lyzinski et al., 2016] for a particular cost.

One can show that GW satisfies the triangular inequality, and in fact it defines a distance between metric spaces equipped with a probability distribution (here assumed to be discrete in definition (10.25)) up to isometries preserving the measures. This distance was introduced and studied in details by Memoli in [Mémoli, 2011]. An in-depth mathematical exposition (in particular, its geodesic structure and gradient flows) is given in Sturm [2012]. See also Schmitzer and Schnörr [2013a] for applications in computer vision. This distance is also tightly connected with the Gromov-Hausdorff distance [Gromov, 2001] between metric spaces, which have been used for shape matching Mémoli [2007], Bronstein et al. [2010].

Remark 10.7 (Gromov-Wasserstein distance)**.** The general setting corresponds to computing couplings between metric measure spaces $(\mathcal{X}, d_\mathcal{X}, \alpha_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y}, \alpha_\mathcal{Y})$ where $(d_\mathcal{X}, d_\mathcal{Y})$ are distances

and $(\alpha_{\mathcal{X}}, \alpha_{\mathcal{Y}})$ are measures on their respective spaces. One defines

$$\mathcal{GW}((\alpha_{\mathcal{X}}, d_{\mathcal{X}}), (\alpha_{\mathcal{Y}}, d_{\mathcal{Y}}))^2 \stackrel{\text{def.}}{=} \min_{\pi \in \mathbf{U}(\alpha_{\mathcal{X}}, \alpha_Y)} \int_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y'). \tag{10.26}$$

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

Remark 10.8 (Gromov-Wasserstein geodesics)**.** The space of metric spaces (up to isometries) endowed with this $\mathcal{GW}$ distance (10.26) has a geodesic structure. Sturm [2012] shows that the geodesic between $(\mathcal{X}_0, d_{\mathcal{X}_0}, \alpha_0)$ and $(\mathcal{X}_1, d_{\mathcal{X}_1}, \alpha_1)$ can be chosen to be $t \in [0, 1] \mapsto (\mathcal{X}_0 \times \mathcal{X}_1, d_t, \pi^\star)$ where $\pi^\star$ is a solution of (10.26) and for all $((x_0, x_1), (x_0', x_1')) \in (\mathcal{X}_0 \times \mathcal{X}_1)^2$,

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

This formula allows one to define and analyze gradient flows which minimize functionals involving metric spaces, see Sturm [2012]. It is however difficult to handle numerically, because it involves computations over the product space $\mathcal{X}_0 \times \mathcal{X}_1$. A heuristic approach is used in Peyré et al. [2016] to define geodesics and barycenters of metric measure spaces while imposing the cardinality of the involved spaces and making use of the entropic smoothing (10.27) detailed below.

10.6.4 Entropic Regularization

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

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \mathcal{E}_{\mathbf{D}, \mathbf{D}'}(\mathbf{P}) - \varepsilon \mathbf{H}(\mathbf{P}). \tag{10.27}$$

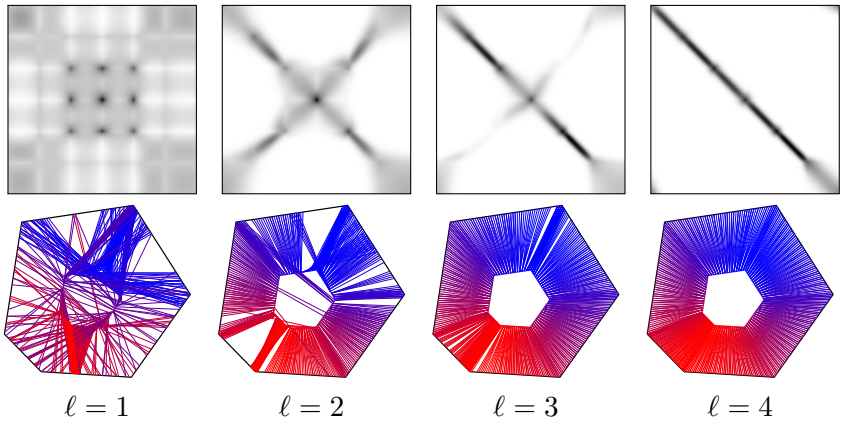


Figure 10.9: Iterations of the entropic GW algorithm (10.28) between two shapes $(x_i)_i$ and $(y_j)_j$ in $\mathbb{R}^2$, initialized with $\mathbf{P}^{(0)} = \mathbf{a} \otimes \mathbf{b}$. The distance matrices are $\mathbf{D}_{i,i'} = \|x_i - x_{i'}\|$ and $\mathbf{D}'_{j,j'} = \|y_j - y_{j'}\|$. Top row: coupling $P^{(\ell)}$ displayed as a 2-D image. Bottom row: matching induced by $P^{(\ell)}$ (each point x_i is connected to the 3 y_j with the 3 largest values among $\{P^{(\ell)}_{i,j}\}_j$). The shapes have the same size, but for displaying purpose, the inner shape $(x_i)_i$ has been reduced in size.

As proposed initially in [Gold and Rangarajan, 1996, Rangarajan et al., 1999], and later revisited in [Solomon et al., 2016a] for applications in graphics, one can use iteratively Sinkhorn's algorithm to progressively compute a stationary point of (10.27). Indeed, successive linearizations of the objective function lead to consider the succession of updates

$$\mathbf{P}^{(\ell+1)} \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}^{(\ell)} \rangle - \varepsilon H(\mathbf{P}) \quad \text{where} \tag{10.28}$$

$$\mathbf{C}^{(\ell)} \overset{\text{def.}}{=} \nabla \mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}^{(\ell)}) = -\mathbf{D}\mathbf{P}^{(\ell)}\mathbf{D}',$$

which can be interpreted as a mirror-descent scheme [Solomon et al., 2016a]. Each update can thus be solved using Sinkhorn iterations (4.15) with cost $\mathbf{C}^{(\ell)}$. Figure 10.9 displays the evolution of the algorithm. Figure (10.10) illustrates the use of this entropic Gromov-Wasserstein to compute soft maps between domains.

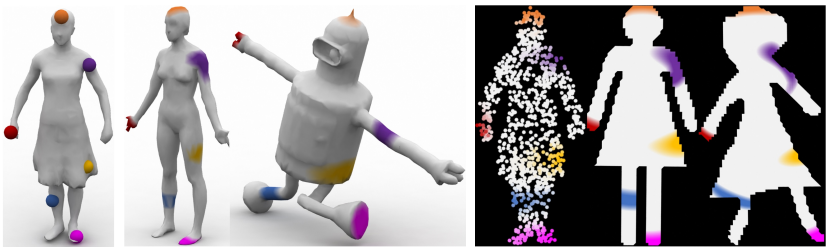


Figure 10.10: Example of fuzzy correspondences computed by solving GW problem (10.27) with Sinkhorn iterations (10.28). Extracted from [Solomon et al., 2016a].

References

Pytorch library. *http://pytorch.org/*, 2017.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Isabelle Abraham, Romain Abraham, Maıtine Bergounioux, and Guillaume Carlier. Tomographic reconstruction from a few views: a multi-marginal optimal transport approach. *Applied Mathematics & Optimization*, 75(1): 55–73, 2017.

Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

Martial Agueh and Malcolm Bowles. One-dimensional numerical algorithms for gradient flows in the p-Wasserstein spaces. *Acta Applicandae Mathematicae*, 125(1):121–134, 2013.

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM J. on Mathematical Analysis*, 43(2):904–924, 2011.

Martial Agueh and Guillaume Carlier. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *Comptes Rendus Mathematique*, 2017.

Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermüller, Dzmitry Bahdanau, Nicolas Ballas, ..., and Ying Zhang. Theano: A python framework for fast computation of mathematical expressions. *CoRR*, abs/1605.02688, 2016.

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

Zeyuan Allen-Zhu, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Much faster algorithms for matrix scaling. *arXiv preprint arXiv:1704.02315*, 2017.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *arXiv preprint arXiv:1705.09634*, 2017.

Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. *arXiv preprint arXiv:1709.10219*, 2017.

Luigi Ambrosio and Alessio Figalli. Geodesics in the space of measure-preserving maps and plans. *Arch. Ration. Mech. Anal.*, 194(2):421–462, 2009. ISSN 0003-9527.

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

Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete Wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016.

Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 343–352. Society for Industrial and Applied Mathematics, 2008.

Alexandr Andoni, Assaf Naor, and Ofer Neiman. Snowflake universality of Wasserstein spaces. *Annales scientifiques de l'École normale supérieure*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

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

Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

Amir Averbuch, RR Coifman, DL Donoho, Moshe Israeli, and Johan Walden. Fast slant stack: A notion of radon transform for data in a cartesian grid which is rapidly computible, algebraically exact, geometrically faithful and invertible. *SIAM Scientific Computing*, 2001.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Francis R Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

Michael Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310, 1965.

Alexander Barvinok. *A Course in Convexity*. Graduate studies in mathematics. American Mathematical Society, 2002. ISBN 9780821829684.

Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12): 1298–1302, 2006.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer-Verlag, New York, 2011.

Heinz H Bauschke and Adrian S Lewis. Dykstra's algorithm with Bregman projections: a convergence proof. *Optimization*, 48(4):409–427, 2000.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31 (3):167–175, 2003.

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

Mathias Beiglböck, Pierre Henry-Labordère, and Friedrich Penkner. Model-independent bounds for option prices?a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(05):851–868, 2003.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Jean-David Benamou and Guillaume Carlier. Augmented lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1):1–26, 2015.

Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampere equation. *Journal of Computational Physics*, 260:107–126, 2014.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge–Ampère operator. *Numerische Mathematik*, 134(3):611–636, 2016a.

Jean-David Benamou, Francis Collino, and Jean-Marie Mirebeau. Monotone and consistent discretization of the Monge-Ampere operator. *Mathematics of computation*, 85(302):2743–2775, 2016b.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146*, 2017.

Dimitri P Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

Dimitri P Bertsekas. Auction algorithms for network flow problems: A tutorial introduction. *Computational optimization and applications*, 1(1):7–66, 1992.

Dimitri P Bertsekas. *Network optimization: continuous and discrete models*. Athena Scientific Belmont, 1998.

Dimitri P Bertsekas and Jonathan Eckstein. Dual coordinate step methods for linear network flow problems. *Mathematical Programming*, 42(1):203–243, 1988.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

Jérémie Bigot and Thierry Klein. Consistent estimation of a population barycenter in the Wasserstein space. *Preprint arXiv:1212.2562*, 2012a.

Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *arXiv preprint arXiv:1212.2562*, 2012b.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex pca. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucumán Rev. Ser. A*, 5:147–151, 1946.

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

Adrien Blanchet and Guillaume Carlier. Optimal transport and Cournot-Nash equilibria. *Mathematics of Operations Research*, 41(1):125–145, 2015.

Adrien Blanchet, Vincent Calvez, and José A Carrillo. Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model. *SIAM Journal on Numerical Analysis*, 46(2):691–721, 2008.

Emmanuel Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance. *Electronic Journal of Probability*, 16:2296–2333, 2011.

Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

Franccois Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3):541–593, 2007.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Transactions on Graphics (TOG)*, 30(6):158, 2011.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics (Proc. SIGGRAPH 2016)*, 35(4):71:1–71:10, 2016. .

CW Borchardt and CGJ Jocobi. De investigando ordine systematis aequationum differentialium vulgarium cujuscunque. *Journal für die reine und angewandte Mathematik*, 64:297–320, 1865.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. *Polygon Mesh Processing*. Taylor & Francis, 2010.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. Preprint 1705.07642, Arxiv, 2017.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7 (3):200–217, 1967.

Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. of the AMS*, 2:225–255, 1990.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

Yann Brenier. The dual least action problem for an ideal, incompressible fluid. *Archive for Rational Mechanics and Analysis*, 122(4):323–351, 1993. ISSN 0003-9527.

Yann Brenier. Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations. *Communications on Pure and Applied Mathematics*, 52(4):411–452, 1999. ISSN 1097-0312.

Yann Brenier. Generalized solutions and hydrostatic approximation of the Euler equations. *Phys. D*, 237(14-17):1982–1988, 2008. ISSN 0167-2789.

Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006.

Alexander M Bronstein, Michael M Bronstein, Ron Kimmel, Mona Mahmoudi, and Guillermo Sapiro. A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *Inter. Journal on Computer Vision*, 89(2-3):266–286, 2010.

Richard A Brualdi. *Combinatorial matrix classes*, volume 108. Cambridge University Press, 2006.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society Providence, RI, 2001.

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

Martin Burger, José Antonio Carrillo de la Plata, and Marie-Therese Wolfram. A mixed finite element method for nonlinear diffusion equations. *Kinetic and Related Models*, 3(1):59–83, 2010.

Martin Burger, Marzena Franek, and Carola-Bibiane Schönlieb. Regularised regression and density estimation based on optimal transport. *Appl. Math. Res. Express*, 2:209–253, 2012.

Luis Caffarelli. The Monge-Ampere equation and optimal transportation, an elementary review. *LECTURE NOTES IN MATHEMATICS-SPRINGER VERLAG*, pages 1–10, 2003.

Luis Caffarelli, Mikhail Feldman, and Robert McCann. Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *Journal of the American Mathematical Society*, 15(1):1–26, 2002.

Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Ann. of Math.*, 171(2):673–730, 2010. ISSN 0003-486X.

Luis A Caffarelli, Sergey A Kochengin, and Vladimir I Oliker. Problem of reflector design with given far-field scattering data. In *Monge Ampère Equation: Applications to Geometry and Optimization: NSF-CBMS Conference on the Monge Ampère Equation, Applications to Geometry and Optimization, July 9-13, 1997, Florida Atlantic University*, volume 226, page 13. American Mathematical Soc., 1999.

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2492–2500. Curran Associates, Inc., 2012.

Eric A Carlen and Jan Maas. An analog of the 2-Wasserstein metric in noncommutative probability under which the fermionic Fokker–Planck equation is gradient flow for the entropy. *Communications in Mathematical Physics*, 331(3):887–926, 2014.

Guillaume Carlier and Ivar Ekeland. Matching for teams. *Econom. Theory*, 42(2):397–418, 2010. ISSN 0938-2259. .

Guillaume Carlier and Clarice Poon. On the total variation Wasserstein gradient flow and the TV-JKO scheme. *arXiv preprint arXiv:1703.00243*, 2017.

Guillaume Carlier, Chloé Jimenez, and Filippo Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM J. Control Optim.*, 47(3):1330–1350, 2008.

Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

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

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

José A Carrillo and J Salvador Moll. Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms. *SIAM Journal on Scientific Computing*, 31(6):4305–4329, 2009.

José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17:233–258, 1 2015. ISSN 1991-7120.

Yair Censor and Simeon Reich. The Dykstra algorithm with Bregman projections. *Communications in Applied Analysis*, 2:407–419, 1998.

Thierry Champion, Luigi De Pascale, and Petri Juutinen. The ∞-wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

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

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016a.

Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: A quantum mechanical approach. *Preprint arXiv:1610.03041*, 2016b.

Yongxin Chen, Wilfrid Gangbo, Tryphon T Georgiou, and Allen Tannenbaum. On the matrix Monge-Kantorovich problem. *Preprint arXiv:1701.02826*, 2017.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. Unbalanced optimal transport: Geometry and Kantorovich formulation. Preprint 1508.05216, Arxiv, 2015.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, pages 1–44, 2016.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *to appear in Mathematics of Computation*, 2017.

Shui-Nee Chow, Wen Huang, Yao Li, and Haomin Zhou. Fokker-Planck equations for a free energy functional or Markov process on a graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012. ISSN 0003-9527. .

Samir Chowdhury and Facundo Mémoli. Constructing geodesics on the space of compact metric spaces. *arXiv preprint arXiv:1603.02385*, 2016.

Imre Ciszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. *arXiv preprint arXiv:1704.02310*, 2017.

Patrick L Combettes and Jean-Christophe Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564 –574, 2007. ISSN 1932-4553. .

Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, 1994.

Laurent Condat. Fast projection onto the simplex and the ℓ_1 ball. *Proc. Math. Program., Ser. A*, pages 1–11, 2015.

Sueli IR Costa, Sandra A Santos, and João E Strapasson. Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

Codina Cotar, Gero Friesecke, and Claudia Klüppelberg. Density functional theory and optimal transportation with Coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013. ISSN 1097-0312.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Thomas Corpetti. Optimal transport for data fusion in remote sensing. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 3571–3574. IEEE, 2016.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *arXiv preprint arXiv:1705.08848*, 2017a.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017b.

Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Trans. Graph.*, 32(5):152:1–152:11, October 2013.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 2292–2300, 2013.

Marco Cuturi and David Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15:533–564, 2014.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of ICML*, volume 32, pages 685–693, 2014.

Marco Cuturi and Kenji Fukumizu. Kernels on structured objects through nested histograms. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 329–336. MIT Press, 2007.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016. .

Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

George B. Dantzig. Programming of interdependent activities: II mathematical model. *Econometrica*, 17(3/4):200–211, 1949.

George B Dantzig. Application of the simplex method to a transportation problem. *Activity analysis of production and allocation*, 13:359–373, 1951.

George B. Dantzig. *Reminiscences About the Origins of Linear Programming*, pages 78–86. Springer Berlin Heidelberg, 1983.

George B. Dantzig. Linear programming. In J. K. Lenstra, A. H. G. Rinnooy Kan, and A. Schrijver, editors, *History of Mathematical Programming: A Collection of Personal Reminiscences*, pages 257–282. Elsevier Science Publishers, 1991.

Jon Dattorro. *Convex optimization & Euclidean distance geometry.* Meboo publishing, 2017.

Fernando De Goes, Katherine Breeden, Victor Ostromoukhov, and Mathieu Desbrun. Blue noise through optimal transport. *ACM Transactions on Graphics (TOG)*, 31(6):171, 2012.

Fernando de Goes, Corentin Wallez, Jin Huang, Dmitry Pavlov, and Mathieu Desbrun. Power particles: An incompressible fluid solver based on power diagrams. *ACM Trans. Graph.*, 34(4):50:1–50:11, July 2015.

Eustasio del Barrio, JA Cuesta-Albertos, C Matrán, and A Mayo-Íscar. Robust clustering tools based on optimal transportation. *arXiv preprint arXiv:1607.01179*, 2016.

Julie Delon. Midway image equalization. *Journal of Mathematical Imaging and Vision*, 21(2):119–134, 2004.

Julie Delon, Julien Salomon, and Andrei Sobolevski. Fast transport optimization for Monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7):2239–2258, 2010.

Julie Delon, Julien Salomon, and Andrei Sobolevski. Local matching indicators for transport problems with concave costs. *SIAM Journal on Discrete Mathematics*, 26(2):801–827, 2012.

Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals Mathematical Statistics*, 11(4):427–444, 1940.

Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 49, pages 1183–1203. Institut Henri Poincaré, 2013.

Rachid Deriche. *Recursively implementating the Gaussian and its derivatives.* PhD thesis, INRIA, 1993.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the Rot mover's distance. *arXiv preprint arXiv:1610.06447*, 2016.

Simone Di Marino and Lenaic Chizat. A tumor growth model of Hele-Shaw type as a gradient flow. *Arxiv*, 2017.

Khanh Do Ba, Huy L Nguyen, Huy N Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, 48(2):428–442, 2011.

Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

Yan Dolinsky and H Mete Soner. Martingale optimal transport and robust hedging in continuous time. *Probability Theory and Related Fields*, 160 (1-2):391–427, 2014.

Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Arnaud Dupuy, Alfred Galichon, and Yifei Sun. Estimating matching affinity matrix under low-rank constraints. *Arxiv:1612.09585*, 2016.

Richard L Dykstra. An algorithm for restricted least squares regression. *J. Amer. Stat.*, 78(384):839–842, 1983.

Richard L Dykstra. An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Ann. Probab.*, 13(3):975–984, 1985. ISSN 0091-1798.

Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

David A Edwards. The structure of superspace. In *Studies in topology*, pages 121–133. Elsevier, 1975.

Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Matthias Erbar. The heat equation on manifolds as a gradient flow in the Wasserstein space. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46(1):1–23, 2010.

Matthias Erbar and Jan Maas. Gradient flow structures for discrete porous medium equations. *Discrete Contin. Dyn. Syst.*, 34(4):1355–1374, 2014.

Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *arXiv preprint arXiv:1707.06859*, 2017.

Sven Erlander. *Optimal Spatial Interaction and the Gravity Model*, volume 173. Springer-Verlag, 1980.

Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*. Vsp, 1990.

Montacer Essid and Justin Solomon. Quadratically-regularized optimal transport on graphs. *arXiv preprint arXiv:1704.08200*, 2017.

Mikhail Feldman and Robert McCann. Monge's transport problem on a Riemannian manifold. *Trans. AMS*, 354(4):1667–1697, 2002.

Jean Feydy, Benjamin Charlier, Francois-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *Proc. MICCAI'17*, 2017.

Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

Rémi Flamary, Cédric Févotte, Nicolas Courty, and Valentin Emiya. Optimal spectral transportation with application to music transcription. In *Advances in Neural Information Processing Systems*, pages 703–711, 2016.

Lester Randolph Ford and Delbert Ray Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

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

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

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

Uriel Frisch, Sabino Matarrese, Roya Mohayaee, and Andrei Sobolevski. A reconstruction of the initial conditions of the universe by optimal mass transportation. *Nature*, 417(6886):260–262, 2002.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

Alfred Galichon and Bernard Salanié. Matching with trade-offs: Revealed preferences over competing characteristics. Technical report, Preprint SSRN-1487307, 2009.

Alfred Galichon, Pierre Henry-Labordere, and Nizar Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *The Annals of Applied Probability*, 24(1):312–336, 2014.

Thomas O Gallouët and Quentin Mérigot. A lagrangian scheme à la brenier for the incompressible euler equations. *Foundations of Computational Mathematics*, pages 1–31, 2017.

Thomas O Gallouët and Leonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

Wilfrid Gangbo and Andrzej Swiech. Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on pure and applied mathematics*, 51(1):23–45, 1998a.

Wilfrid Gangbo and Andrzej Swiech. Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on pure and applied mathematics*, 51(1):23–45, 1998b.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1): 185–203, 1990.

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

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. Preprint 1706.00292, Arxiv, 2017a.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. GAN and VAE from an optimal transport point of view. Preprint 1706.01807, Arxiv, 2017b.

Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. *arXiv preprint arXiv:1510.08230*, 2015.

Alan George and Joseph WH Liu. The evolution of the minimum degree ordering algorithm. *Siam review*, 31(1):1–19, 1989.

Tryphon T Georgiou and Michele Pavon. Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56(3):033301, 2015.

Pascal Getreuer. A survey of Gaussian convolution algorithms. *Image Processing On Line*, 2013:286–310, 2013.

Ugo Gianazza, Giuseppe Savaré, and Giuseppe Toscani. The Wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation. *Archive for Rational Mechanics and Analysis*, 194(1):133–220, 2009. ISSN 0003-9527. .

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

Roland Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 9 (R2):41–76, 1975.

Steven Gold and Anand Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Trans. on PAMI*, 18(4):377–388, April 1996.

Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1458–1465. IEEE, 2005.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Andreas Griewank. Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation. *Optimization Methods and software*, 1(1):35–54, 1992.

Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

Mikhail Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Progress in Mathematics. Birkhäuser, 2001.

Mikhail Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Springer Science & Business Media, 2007.

Gaoyue Guo and Jan Obloj. Computational methods for martingale optimal transport problems. *arXiv preprint arXiv:1710.07911*, 2017.

Leonid Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69(3):448–484, 2004.

Cristian E Gutiérrez. *The Monge-Ampere equation*. Springer, 2016.

Jorge Gutierrez, Julien Rabin, Bruno Galerne, and Thomas Hurtut. Optimal patch assignment for statistically constrained texture synthesis. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 172–183. Springer, 2017.

A Hadjidimos. Successive overrelaxation (SOR) and related methods. *Journal of Computational and Applied Mathematics*, 123(1):177–199, 2000.

Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

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

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pages 2417–2426, 2016.

David Hilbert. Über die gerade linie als kürzeste verbindung zweier punkte. *Mathematische Annalen*, 46(1):91–96, 1895.

Frank L Hitchcock. The distribution of a product from several sources to numerous localities. *Studies in Applied Mathematics*, 20(1-4):224–230, 1941.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. *arXiv preprint arXiv:1706.03883*, 2017.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 2008.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870, 2016.

Martin Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

Piotr Indyk and Eric Price. K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 627–636. ACM, 2011.

Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

Xianhua Jiang, Lipeng Ning, and Tryphon T Georgiou. Distances and Riemannian metrics for multivariate spectral densities. *IEEE Transactions on Automatic Control*, 57(7):1723–1735, July 2012. ISSN 0018-9286. .

William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

Lee K Jones and Charles L Byrne. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE transactions on Information Theory*, 36(1):23–30, 1990.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

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

LV Kantorovich and G.S. Rubinshtein. On a space of totally additive functions. *Vestn Lening. Univ.*, 13:52–59, 1958.

Hermann Karcher. Riemannian center of mass and so called Karcher mean. *arXiv preprint arXiv:1407.2087*, 2014.

Johan Karlsson and Axel Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *arXiv preprint arXiv:1612.02273*, 2016.

Sanggyun Kim, Rui Ma, Diego Mesa, and Todd P Coleman. Efficient bayesian inference methods via convex optimization and optimal transport. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2259–2263. IEEE, 2013.

David Kinderlehrer and Noel J Walkington. Approximation of parabolic equations using the Wasserstein metric. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(04):837–852, 1999.

Jun Kitagawa, Quentin Mérigot, and Boris Thibert. A Newton algorithm for semi-discrete optimal transport. *arXiv preprint arXiv:1603.05579*, 2016.

Philip A Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013.

Philip A Knight, Daniel Ruiz, and Bora Uccar. A symmetry preserving algorithm for matrix scaling. *SIAM journal on Matrix Analysis and Applications*, 35(3):931–955, 2014.

Martin Knott and Cyril S Smith. On a generalization of cyclic monotonicity and distances among random vectors. *Linear algebra and its applications*, 199:363–371, 1994.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164, 2016.

Tjalling C Koopmans. Optimum utilization of the transportation system. *Econometrica: Journal of the Econometric Society*, pages 136–146, 1949.

Jonathan Korman and Robert McCann. Optimal transportation with capacity constraints. *Transactions of the American Mathematical Society*, 367(3): 1501–1521, 2015.

Bernhard Korte and Jens Vygen. *Combinatorial optimization*. Springer, 2012.

JJ Kosowsky and Alan L Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural networks*, 7(3):477–490, 1994.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

Rongjie Lai and Hongkai Zhao. Multi-scale non-rigid point cloud registration using robust sliced-Wasserstein distance via Laplace-Beltrami eigenmap. *arXiv preprint arXiv:1406.3758*, 2014.

Hugo Lavenant. Harmonic mappings valued in the Wasserstein space. *Preprint cvgmt 3649*, 2017.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, pages 1–17, 2016.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

William Leeb and Ronald Coifman. Hölder–Lipschitz norms and their duals on spaces with semigroups, with applications to earth mover's distance. *Journal of Fourier Analysis and Applications*, 22(4):910–953, 2016.

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

Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

Christian Léonard. From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst. A*, 34(4): 1533–1574, 2014.

Bruno Lévy. A numerical algorithm for l^2 semi-discrete optimal transport in 3d. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6): 1693–1715, 2015.

Bruno Levy and Erica Schwindt. Notions of optimal transport theory and how to implement them on a computer. *arXiv:1710.02634*, 2017.

Peihua Li, Qilong Wang, and Lei Zhang. A novel earth mover's distance methodology for image matching with Gaussian mixture models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1696, 2013.

Wuchen Li, Penghang Yin, and Stanley Osher. A fast algorithm for unbalanced l^1 Monge–Kantorovich problem. *CAM report*, 2016.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *ArXiv e-prints*, 2015.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: The Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

Haibin Ling and Kazunori Okada. Diffusion distance for histogram comparison. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 246–253. IEEE, 2006.

Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Tr. on PAMI*, 29(5):840–853, 2007.

Nathan Linial, Alex Samorodnitsky, and Avi Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 644–652. ACM, 1998.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *Siam J. Numer. Anal.*, 16:964–979, 1979.

Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. A survey for the quadratic assignment problem. *European J. Operational Research*, 176(2):657–690, 2007.

Vince Lyzinski, Donniell E Fishkind, Marcelo Fiori, Joshua T Vogelstein, Carey E Priebe, and Guillermo Sapiro. Graph matching: Relax at your own risk. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):60–73, 2016.

Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011. .

Jan Maas, Martin Rumpf, Carola Schönlieb, and Stefan Simon. A generalized model for optimal transport of images including dissipation and density modulation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49 (6):1745–1769, 2015.

Jan Maas, Martin Rumpf, and Stefan Simon. Generalized optimal transport with singular sources. *arXiv preprint arXiv:1607.01186*, 2016.

Yasushi Makihara and Yasushi Yagi. Earth mover's morphing: Topology-free shape morphing using cluster-based EMD flows. In *Asian Conference on Computer Vision*, pages 202–215. Springer, 2010.

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

Benjamin Mathon, Francois Cayre, Patrick Bas, and Benoit Macq. Optimal transport for secure spread-spectrum watermarking of still images. *IEEE Transactions on Image Processing*, 23(4):1694–1705, 2014.

Daniel Matthes and Horst Osberger. Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation? *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(3):697–726, 2014.

Bertrand Maury and Anthony Preux. Pressureless Euler equations with maximal density constraint: a time-splitting scheme. *Topological Optimization and Optimal Transport: In the Applied Sciences*, 17:333, 2017.

Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Mathematical Models and Methods in Applied Sciences*, 20(10):1787–1821, 2010.

Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Facundo Mémoli. On the use of Gromov–Hausdorff distances for shape comparison. In *Symposium on Point Based Graphics*, pages 81–90. 2007.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4): 417–487, 2011.

Facundo Mémoli and Guillermo Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*, 5(3):313–347, 2005.

Quentin Mérigot. A multiscale approach to optimal transport. *Comput. Graph. Forum*, 30(5):1583–1592, 2011.

Quentin Mérigot, Jocelyn Meyron, and Boris Thibert. Light in power: A general and parameter-free algorithm for caustic design. *arXiv preprint arXiv:1708.04820*, 2017.

Ludovic Métivier, Romain Brossier, Quentin Merigot, Edouard Oudet, and Jean Virieux. An optimal transport approach for seismic tomography: Application to 3D full waveform inversion. *Inverse Problems*, 32(11):115008, 2016.

Alexander Mielke. Geodesic convexity of the relative entropy in reversible Markov chains. *Calculus of Variations and Partial Differential Equations*, 48(1-2):1–31, 2013.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jean-Marie Mirebeau. Discretization of the 3D Monge-Ampere operator, between wide stencils and power diagrams. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(5):1511–1523, 2015.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted Boltzmann machines. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3718–3726. 2016.

Kevin Moon and Alfred Hero. Multivariate f-divergence estimation with confidence. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2014.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Oleg Museyko, Michael Stiglmayr, Kathrin Klamroth, and Günter Leugering. On the application of the Monge–Kantorovich problem to image registration. *SIAM Journal on Imaging Sciences*, 2(4):1068–1097, 2009.

Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *AAAI*, pages 2387–2393, 2017.

Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1. *SIAM J. Comput.*, 37(3):804–826, 2007.

Richard D Neidinger. Introduction to automatic differentiation and MATLAB object-oriented programming. *SIAM Review*, 52(3):545–563, 2010.

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

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

Kangyu Ni, Xavier Bresson, Tony Chan, and Selim Esedoglu. Local histogram based segmentation using the Wasserstein distance. *International journal of computer vision*, 84(1):97–111, 2009.

Lipeng Ning and Tryphon T Georgiou. Metrics for matrix-valued measures via test functions. In *53rd IEEE Conference on Decision and Control*, pages 2642–2647. IEEE, 2014.

Lipeng Ning, Tryphon T Georgiou, and Allen Tannenbaum. On matrix-valued Monge–Kantorovich optimal mass transport. *IEEE transactions on automatic control*, 60(2):373–382, 2015.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer-Verlag, 1999.

Adam M Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

Vladimir Oliker and Laird D Prussner. On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ and its discretizations, I. *Numerische Mathematik*, 54(3):271–293, 1989.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

Dean S Oliver. Minimization for conditional simulation: Relationship to optimal transport. *Journal of Computational Physics*, 265:1–15, 2014.

Ferdinand Österreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in partial differential equations*, 26 (1-2):101–174, 2001.

Art B Owen. *Empirical likelihood*. Wiley Online Library, 2001.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

Brendan Pass. On the local structure of optimal measures in the multi-marginal optimal transportation problem. *Calc. Var. Partial Differential Equations*, 43(3-4):529–536, 2012. ISSN 0944-2669.

Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. *Computer Vision–ECCV 2008*, pages 495–508, 2008.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *Computer vision, 2009 IEEE 12th international conference on*, pages 460–467. IEEE, 2009.

Benoît Perthame, Fernando Quirós, and Juan Luis Vázquez. The Hele-Shaw asymptotics for mechanical models of tumor growth. *Archive for Rational Mechanics and Analysis*, 212(1):93–127, 2014. ISSN 0003-9527.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Gabriel Peyré, Jalal Fadili, and Julien Rabin. Wasserstein active contours. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2541–2544. IEEE, 2012.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

Gabriel Peyré, Lenaic Chizat, Francois-Xavier Vialard, and Justin Solomon. Quantum entropic regularization of matrix-valued optimal transport. *to appear in European Journal of Applied Mathematics*, 2017.

Rémi Peyre. Comparison between w_2 distance and h^{-1} norm, and localisation of Wasserstein distance. *arXiv preprint arXiv:1104.4631*, 2011.

Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

Franccois Pitié, Anil C Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1):123–137, 2007.

Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *Proc. SSVM'15*, pages 256–269, 2015.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998a.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume II: Applications*, volume 2. Springer Science & Business Media, 1998b.

Louis B Rall. *Automatic differentiation: Techniques and applications*. Springer, 1981.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Anand Rangarajan, Alan L Yuille, Steven Gold, and Eric Mjolsness. Convergence properties of the softassign quadratic assignment algorithm. *Neural Comput.*, 11(6):1455–1474, August 1999.

C Rao and T Nayak. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, 31 (5):589–593, 1985.

Sebastian Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 630–638, 2016.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Ludger Ruschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, 23(4):1160–1174, 1995.

Ludger Rüschendorf and Wolfgang Thomsen. Closedness of sum spaces and the generalized Schrodinger problem. *Theory of Probability and its Applications*, 42(3):483–494, 1998.

Ludger Rüschendorf and Ludger Uckelmann. On the n-coupling problem. *Journal of multivariate analysis*, 81(2):242–258, 2002.

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

Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhauser, 2015.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

Filippo Santambrogio. Crowd motion and population dynamics under density constraints. *GMT preprint 3728*, 2018.

Louis-Philippe Saumier, Boualem Khouider, and Martial Agueh. Optimal transport for particle image velocimetry. *Communications in Mathematical Sciences*, 13(1):269–296, 2015.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *bioRxiv*, page 191056, 2017.

Bernhard Schmitzer. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016a.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016b.

Bernhard Schmitzer and Christoph Schnörr. Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. *Journal of mathematical imaging and vision*, 46(1):143–159, 2013a.

Bernhard Schmitzer and Christoph Schnörr. Object segmentation by shape matching with Wasserstein modes. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 123–136. Springer, 2013b.

Bernhard Schmitzer and Benedikt Wirth. A framework for Wasserstein-1-type metrics. *arXiv preprint arXiv:1701.01945*, 2017.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 38:522–356, 1938.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002.

Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.

Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015.

Sameer Shirdhonkar and David W Jacobs. Approximate earth mover's distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly*, 74:402–405, 1967.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.

Marcos Slomp, Michihiro Mikamo, Bisser Raytchev, Toru Tamaki, and Kazufumi Kaneda. Gpu-based softassign for maximizing image utilization in photomosaics. *International Journal of Networking and Computing*, 1(2): 211–229, 2011.

Justin Solomon, Leonidas Guibas, and Adrian Butscher. Dirichlet energy for analysis and synthesis of soft maps. In *Computer Graphics Forum*, volume 32, pages 197–206. Wiley Online Library, 2013.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *Transaction on Graphics (proc. SIGGRAPH)*, 33(4), 2014a.

Justin Solomon, Raif Rustamov, Guibas Leonidas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 306–314, 2014b.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (Proc. SIGGRAPH 2015)*, 34(4): 66:1–66:11, 2015. .

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (Proc. SIGGRAPH 2016)*, 35(4):72:1–72:13, 2016a.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Continuous-flow graph transportation distances. *arXiv preprint arXiv:1603.06927*, 2016b.

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

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

Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. WASP: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015a.

Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in Wasserstein space. *arXiv preprint arXiv:1508.05880*, 2015b.

Matthew Staib, Sebastian Claici, Justin Solomon, and Stefanie Jegelka. Parallel streaming Wasserstein barycenters. *arXiv preprint arXiv:1705.07443*, 2017.

Leen Stougie. A polynomial bound on the diameter of the transportation polytope. Technical report, TU/e, Technische Universiteit Eindhoven, Department of Mathematics and Computing Science, 2002.

Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. Preprint 1208.0434, arXiv, 2012.

Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11): 2246–2259, 2015.

Vladimir N Sudakov. *Geometric problems in the theory of infinite-dimensional probability distributions*. Number 141. American Mathematical Soc., 1979.

Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Tensor balancing on statistical manifold. *arXiv preprint arXiv:1702.08142*, 2017.

Paul Swoboda and Christoph Schnörr. Convex variational image restoration with histogram priors. *SIAM Journal on Imaging Sciences*, 6(3):1719–1735, 2013.

Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 2004.

Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

Xiaolu Tan and Nizar Touzi. Optimal transportation under controlled stochastic dynamics. *Ann. Probab.*, 41(5):3201–3240, 2013.

Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4): 1726–1755, 2016.

Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation l^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2):187–210, 2017.

AN Tolstoı. Methods of finding the minimal total kilometrage in cargo transportation planning in space. *TransPress of the National Commissariat of Transportation*, pages 23–55, 1930.

AN Tolstoı. Metody ustraneniya neratsionalâĂŹnykh perevozok pri planirovanii [russian; methods of removing irrational transportation in planning]. *Sotsialisticheskiı Transport*, 9:28–51, 1939.

Alain Trouvé and Laurent Younes. Metamorphoses through Lie group action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005. ISSN 1615-3375.

Neil S Trudinger and Xu-Jia Wang. On the monge mass transfer problem. *Calculus of Variations and Partial Differential Equations*, 13(1):19–31, 2001.

Marc Vaillant and Joan Glaunès. Surface matching via currents. In *Information processing in medical imaging*, pages 1–5. Springer, 2005.

Sathamangalam R Srinivasa Varadhan. On the behavior of the fundamental solution of the heat equation with variable coefficients. *Comm. on Pure and Applied Math.*, 20(2):431–455, 1967.

Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

Cedric Villani. *Optimal transport: old and new*, volume 338. Springer Verlag, 2009.

Thomas Vogt and Jan Lellmann. Measure-valued variational models with applications to diffusion-weighted imaging. *arXiv preprint arXiv:1710.00798*, 2017.

Fan Wang and Leonidas J Guibas. Supervised earth mover's distance learning and its computer vision applications. *ECCV2012*, pages 442–455, 2012.

Wei Wang, John A Ozolek, Dejan Slepcev, Ann B Lee, Cheng Chen, and Gustavo K Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE transactions on medical imaging*, 30(3):621–631, 2011.

Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

Michael Westdickenberg and Jon Wilkening. Variational particle schemes for the porous medium equation and for the system of isentropic Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44 (1):133–166, 2010.

Alan Geoffrey Wilson. The use of entropy maximizing models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

Gui-Song Xia, Sira Ferradans, Gabriel Peyré, and Jean-Franccois Aujol. Synthesizing and mixing stationary Gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508, 2014.

Gloria Zen, Elisa Ricci, and Nicu Sebe. Simultaneous ground metric learning and matrix factorization with earth mover's distance. *Proc. ICPR'14*, pages 3690–3695, 2014.

Lei Zhu, Yan Yang, Steven Haker, and Allen Tannenbaum. An image morphing technique based on optimal mass preserving mapping. *IEEE Transactions on Image Processing*, 16(6):1481–1495, 2007.

Jonathan Zinsl and Daniel Matthes. Transport distances and geodesic convexity for systems of degenerate diffusion equations. *Calculus of Variations and Partial Differential Equations*, 54(4):3397–3438, 2015.