Optimal Transport

Data Science 2 AC 209b

Javier Zazo Pavlos Protopapas

February 20, 2019

Abstract

In this section we give an introduction to optimal transport (OT), where we present an operator over discrete and/or continuous measures that fulfill all distance properties (positivity, symmetry and triangularity). For historical reasons, the OT distance is also referred as Earth mover's distance (EMD) in the computer science literature, or Wasserstein distance. Compared to standard divergences such as Kulback-Leibler, OT is well defined when comparing measures of non-overlapping supports, or different number of samples corresponding to empirical measures. We will present the different formulations of OT for discrete and continuous measures, show the metric properties of the operator, and discuss two applications that exploit its properties. The first application considers adding a Wasserstein loss to a classifier to measure semantic relations between classes. The second application discusses the framework of learning a classifier in a target domain where there are no labels available by transporting labels from a source domain where this information is available.

1 Historical overview

The original idea of transport was described by Gaspard Monge in 1781, when he presented his investigations in [1]. He considered the problem of moving dirt from one place (d'eblais) to another (remblais) with minimal effort, by minimizing a displacement cost. In his original formulation, the transported mass from the original source had to match a predefined target mass at destination, requiring non-splitting of the original elements. Figure 1 shows the Monge-assignment problem. In the first case it simply consists of an assignment and two solutions are possible. In the second case the assignment problem is unique.

However, his initial formulation led to a combinatorial problem which was hard to solve, so other authors after him considered more tractable problems. Frank Lauren Hitchcock formulated in 1941 his idea of a transportation problem by enunciating a the factories and warehouses problem (sometimes explained with mines and factories instead). There are n factories that manufacture some product, and m wharehouses, each with some finite capacity that store the product before selling it to the public. There are associated transport costs for

Figure 1: Monge original assignment problem.

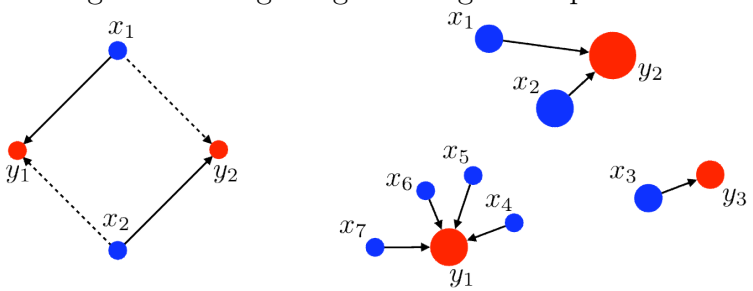
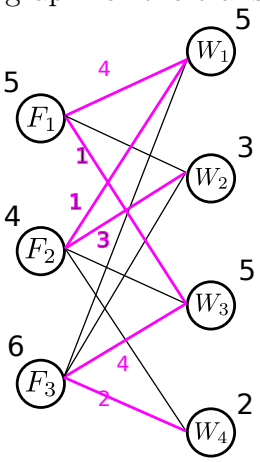


Figure 2: Bipartite graph for the transportation problem.



moving the product from factories to warehouses, and the goal of the problem is to minimize the transportation costs satisfying the constraints.

This formulation leads to a linear program (LP), which can be solved efficiently with specialized techniques such as the simplex methods, or more general, such as interior point methods. Note that the simplex method was proposed by George Dantzig in 1946.

Two of the key ideas of optimal transport distance (which we haven't enunciate yet) is mass conservation and mass tranportation cost. The first one refers to the idea that the goods produce by the factories is not lost or created, which we can enunciate with linear constraints. The second one corresponds to every possible combination of source factory and target warehouse times the transported mass.

As an introductory example, consider 3 factories, each with 5, 4 and 6 units of products; 4 warehouses, with 5, 3, 5 and 2 units of capacity; and a tranportation cost given by

	W_1	W_2	W_3	W_4
F_1	5	4	7	6
F_2	2	5	3	5
F_3	6	3	4	4

The problem can be represented by a bipartite graph in Figure 2, where a conserving tranport plan is depicted (not the optimal plan).

Accouting for this constraints, the transportation problem can be formulated as follows:

$$\begin{aligned}
\min_{x_{ij}\geq 0} \quad & 5x_{11} + 4x_{12} + 7x_{13} + 6x_{14} + 2x_{21} + 5x_{22} \\
& +3x_{23} + 2x_{24} + 6x_{31} + 3x_{32} + 4x_{33} + 4x_{34} \\
\text{s.t.} \quad & x_{11} + x_{12} + x_{13} + x_{14} = 5 \\
& x_{21} + x_{22} + x_{23} + x_{24} = 4 \\
& x_{31} + x_{32} + x_{33} + x_{34} = 6 \\
& x_{11} + x_{21} + x_{31} \leq 5 \\
& x_{12} + x_{22} + x_{32} \leq 3 \\
& x_{13} + x_{23} + x_{33} \leq 5 \\
& x_{14} + x_{24} + x_{34} \leq 2
\end{aligned} \tag{1}$$

Here, if x_{ij} belong to the non-negative real numbers, one factory may transport its product to multiple warehouses, and split its mass as well. The solution in that case need not be an integer value.

2 Definitions and notation

Optimal tranport defines a distance between histograms (or measures), and as such, probability mass sums to one. We define the probability simplex as follows:

$$\Delta_n = \left\{ a_i \in \mathbb{R}_+^n \;\middle|\; \sum_{i=1}^{n} a_i = 1 \right\} \tag{2}$$

We refer to a histogram, or a discrete probability distribution, with bold lower case letter, i.e., $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in \Delta_n$. We also consider the space of locations $\mathcal{X}$ such that $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}$. We extend the notion of a discrete probability distribution to discrete measure defining

$$\alpha = \sum_i p_i \delta_{x_i}, \tag{3}$$

where δ_{x_i} refers to the Dirac delta function at location x_i.

We use the set of Radon measures $\mathcal{M}(\mathcal{X})$, which requires that $\mathcal{X}$ is equipped with a distance, which one can integrate against continuous functions f. In the case of a probability density function integrable in the Lebesgue sense, we get for a measure α that

$$\int_{\mathcal{X}} f(x) d\alpha(x) = \int_{\mathcal{X}} f(x) \rho_\alpha(x) dx, \tag{4}$$

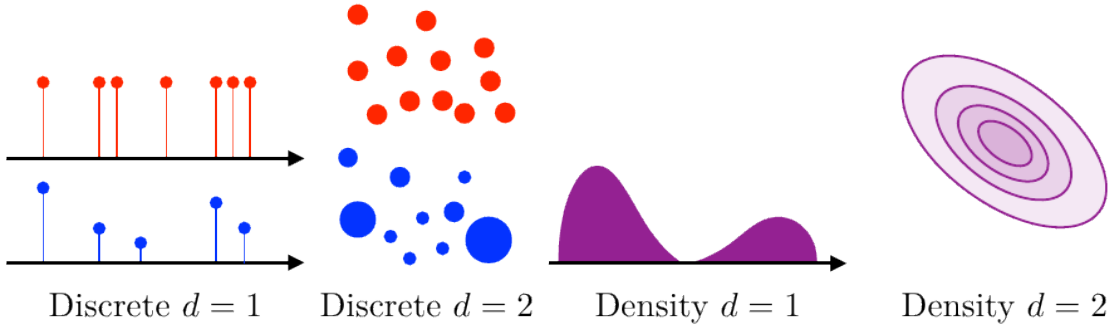
where $\rho_\alpha(x)$ refers to the pdf of the random variable.

We refer to the set of set of positive measures with $\mathcal{M}_+$, such that

$$\int_{\mathcal{X}} f(x) d\alpha(x) \to \mathbb{R}_+, \tag{5}$$

and to the set of probability measures with $\mathcal{M}_+^1$, such that $\int_{\mathcal{X}} d\alpha(x) = 1$. Figure 3 illustrates these ideas.

Figure 3: Discrete and continuous measures for 1 and 2 dimensional $\mathcal{X}$.



3 Assignment and Monge problems

Consider a slightly different problem as the one discussed for the transportation problem by Hitchcock. We have n source elements (agents) and $m = n$ destinations (tasks), and we can to assign each agent to a single task. Of course, there is a cost associated to such an assigment and we are intereseted in minimized such cost. We are looking for a permutation of the set of source elements, such that the cost is minimal:

$$\min_{\sigma \in \text{Perm(n)}} \quad \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}. \tag{6}$$

This is a difficult problem because there are $n!$ possible permutations, and we need a good strategy to find an optimal solution (for n=70, there are approximately 10^{100} possible permuations). This is called the assignment problem, but we will see it is in fact a special case of the more general optimal transport problem.

We will skip the formal definition of the Monge problem, but it can be checked in Remark 2.4 of [2]. In summary, it searches for a transport plan such that a source positive measure α transports all of its mass to a destination positive measure β of same mass, while minimizing cost parametrized by a cost function $c(x, y)$. In the following we will restrict ourselves to probabilistic measures of mass 1.

4 Kantorovich relaxation

The core formulation of the optimal transport distance, is to solve the Kantorovich relaxation problem (sometimes called Monge-Kantorovich problem). Assume we have two discrete probability distributions $\mathbf{p} \in \Delta_n$ and $\mathbf{q} \in \Delta_n$ and we want to minimize the transportation cost of $\mathbf{p}$ onto $\mathbf{q}$.

We define a matrix $\mathbf{F} \in \mathbb{R}^{n \times n}$, of positive elements, where each of its components describes the portion of mass that is transported from component i of $\mathbf{p}$ to component j of $\mathbf{q}$. This matrix is called the transport plan, and has identical meaning as the one we introduced for the Hitchcock problem regarding the transportation problem.

Next, we know that the rows of $\mathbf{F}$ represent the amount of goods that are transported from the original sources (factories). Therefore, because of mass conservation, rows have

to add up to the marginal $\mathbf{p}$. Similarly, the columns of $\mathbf{F}$ represent the amount of goods that are transported to the destination nodes (warehouses), and have to sum up to their total capacity (we impose equality in this problem). The first constraint is written $\mathbf{F1} = \mathbf{p}$ and the second $\mathbf{F}^T\mathbf{1} = \mathbf{q}$. Note that it is easy to check that $\mathbf{F}$ itself corresponds to a joint probability distribution of $P(p, q)$ (all of its elements sum up to 1). We define the set of transport plans that satisfy these constraints with $U(\mathbf{p}, \mathbf{q})$:

$$\mathbf{F} \in U(\mathbf{p}, \mathbf{q}) = \{\, \mathbf{F} \in \mathbb{R}_+^{n \times n} \;\mid\; \mathbf{F1} = \mathbf{p} \text{ and } \mathbf{F}^T\mathbf{1} = \mathbf{q} \,\}. \tag{7}$$

$U(\mathbf{p}, \mathbf{q})$ is convex because it's the intersection of affine equalities, and the intersection of the non-negative orthant.

The transportation cost is the same as in the transportation problem given by the objective funtion of (1), and can be written in short with $\mathrm{tr}(\mathbf{FC})$. This results into the discrete Kantorovich problem:

$$\boxed{\begin{aligned} L_{\mathbf{C}}(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{F} \geq 0} \quad & \mathrm{tr}(\mathbf{FC}) \\ & \mathbf{F1} = \mathbf{p}, \quad \mathbf{F}^T\mathbf{1} = \mathbf{q} \end{aligned}} \tag{8}$$

Note that $\mathbf{C}$ is at least a symmetric matrix of positive values, but we will discuss its requirements later when we analyze the metric properties of the Kantorovich problem.

Problem (8) is an LP, and can be solved using general techniques such as the simplex method, interior point methods, or dual descent algorithms. The problem is convex. Solvers include Clp, Gurobi, Mosek, SeDuMi, CPLEX, ECOS, etc. However, specialized solvers for bipartite graphs also exist, which are much more efficient. For instance, there are opensource C implementations for the network simplex, as well as in CPLEX, which have been ported to Python (in the POT library), or Julia.

Finally, a matter of active research is to compute approximate solutions of (8) that scale to high dimensions, using Sinkhorn regularization or smoothed versions of the problem. We refer the reader to [2] for such descriptions and references.

There is a general formulation of (8) for arbitrary measures. In such case, equivalent matrix $\mathbf{C}$ is a cost function $c(x, y)$, i.e.,

$$c(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+, \tag{9}$$

that measures the cost of transport between locations in $\mathcal{X}$ and $\mathcal{Y}$.

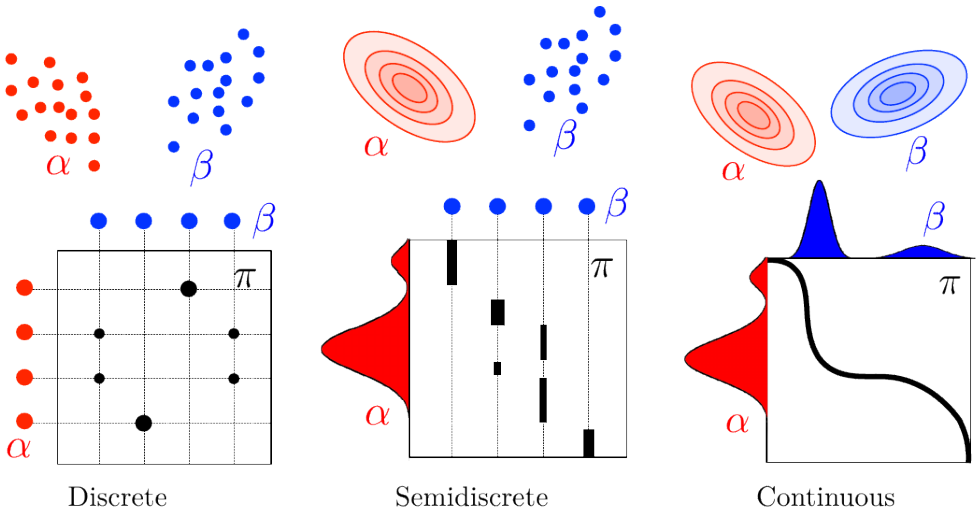
For discrete measures $\alpha = \sum_i p_i \delta_{x_i}$ and $\beta = \sum_i q_i \delta_{y_i}$, the Kantorovich relaxation presents the same form as (8), only that $\mathbf{C}$ is obtained by evaluating $c(x, y)$ in the corresponding locations.

For general probabiliy measures α and β, we define the coupling $\pi \in \mathcal{M}_+^1(\mathcal{X}, \mathcal{Y})$ as the joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$. Same as in the discrete case, the marginalizations of the joint probability measure π has to correspond to α and β. This defines a feasibility region for the joint probability measure:

$$U(\alpha, \beta) = \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X}, \mathcal{Y}) \;\middle|\; P_{\mathcal{X}\sharp}\pi = \alpha \text{ and } P_{\mathcal{Y}\sharp}\pi = \beta \right\}.$$

The $P_{\mathcal{X}\sharp}\pi \in \mathcal{M}(\mathcal{Y})$ is defined as the push operator for a transport plan and marginalizes the joint probability measure π over $\mathcal{Y}$. For a more formal definition see Definition 2.1 in [2].

Figure 4: Transport plans for mixed arbitrary measures.



Finally, the Kantorovich problem for arbitrary measures can be expressed as

$$\mathcal{L}_c(\alpha, \beta) = \min_{\pi \in U(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y). \tag{10}$$

Equivalently, (10) has a probabilistic interpretation where

$$\mathcal{L}_c(\alpha, \beta) = \min_{(X,Y)} \left\{ \mathbb{E}_{(X,Y)}(c(X, Y)) \,\middle|\, X \sim \alpha, Y \sim \beta \right\}. \tag{11}$$

Examples of transport plan problems between discrete, semidiscrete and continuous measures are given in Figure 4

5 Metric properties of the Kantorovich problem

At this point, we consider the metric properties of the discrete optimal transport problem. Note that, when the cost matrix satisfies certain properties, the solution of the Kantorovich problem satisfies positivity, symmetry and triangular inequality and is itself a distance. In the literature, such distance may be refered as Wasserstein distance, OT distance, or Earth mover's distance (EMD).

Theorem 1 (Discrete Wasserstein distance)**.** *Consider* $\mathbf{p}, \mathbf{q} \in \Delta_n$ *and*

$$\mathbf{C} \in \left\{ \mathbf{C} \in \mathbb{R}_+^{n\times n} \,\middle|\, \mathbf{C} = \mathbf{C}^T;\ \mathrm{diag}(\mathbf{C}) = 0;\ \forall (i, j, k)\ C_{i,j} \leq C_{i,k} + C_{k,j};\ C_{i,j} > 0 \text{ for } i \neq j \right\}.$$

Then,

$$W_p(\mathbf{p}, \mathbf{q}) = L_{\mathbf{C}^p}(\mathbf{p}, \mathbf{q})^{1/p}$$

defines a p-Wasserstein *distance on* Δ_n.

Recall $L_{\mathbf{C}}(\mathbf{p}, \mathbf{q})$ from problem (8) for reference.

Proof. We need to show positivity, symmetry and triangular inequality. Since $\mathrm{diag}(\mathbf{C}) = 0$, $W_p(\mathbf{p}, \mathbf{p}) = 0$, and $\mathbf{F}^* = \mathrm{diag}(\mathbf{p})$. Because of strict positivity of off-diagonal elements, $W_p(\mathbf{p}, \mathbf{q}) = \mathrm{tr}(\mathbf{CF}) > 0$ for $\mathbf{p} \neq \mathbf{q}$. This shows that $W_p(\mathbf{p}, \mathbf{q}) = 0$ iff $\mathbf{p} \neq \mathbf{q}$.

Since $W_p(\mathbf{p}, \mathbf{q}) = \mathrm{tr}(\mathbf{CF})$, and $\mathbf{C}$ is symmetric, $W_p(\mathbf{p}, \mathbf{q}) = W_p(\mathbf{q}, \mathbf{p})$, which proves simetry.

For triangularity, we will only consider $p = 1$. We define for three vectors $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{t}$ their optimal tansport maps for two pairs as

$$\mathbf{F} = \mathrm{sol}(W_p(\mathbf{p}, \mathbf{q})) \qquad \mathbf{G} = \mathrm{sol}(W_p(\mathbf{q}, \mathbf{t})).$$

Now, we propose a third transport plan between $W_1(\mathbf{p}, \mathbf{t})$. For simplicity, assume $\mathbf{q} > 0$ (a full proof for general p is available in [2]). Define

$$\mathbf{S} = \mathbf{F}\,\mathrm{diag}(1/\mathbf{q})\mathbf{G} \in \mathbb{R}_+^{n \times n}.$$

Note that $\mathbf{F} \in U(\mathbf{p}, \mathbf{t})$ is feasible, i.e.,

$$\mathbf{S1} = \mathbf{F}\,\mathrm{diag}(1/\mathbf{q})\underbrace{\mathbf{G1}}_{\mathbf{q}} = \mathbf{F}\underbrace{\mathrm{diag}(\mathbf{q}/\mathbf{q})}_{\mathbf{1}} = \mathbf{F1} = \mathbf{p}$$
$$\mathbf{S}^T\mathbf{1} = \mathbf{G}^T\,\mathrm{diag}(1/\mathbf{q})\underbrace{\mathbf{F}^T\mathbf{1}}_{\mathbf{q}} = \mathbf{G}^T\underbrace{\mathrm{diag}(\mathbf{q}/\mathbf{q})}_{\mathbf{1}} = \mathbf{G}^T\mathbf{1} = \mathbf{t}$$

Then, we can write

$$\begin{aligned}
W_1(\mathbf{p}, \mathbf{t}) &= \left\{ \min_{\mathbf{F} \in U(\mathbf{p}, \mathbf{q})} \mathrm{tr}(\mathbf{F}, \mathbf{C}) \right\} \leq \mathrm{tr}(\mathbf{C}, \mathbf{S}) \\
&= \sum_{ik} C_{ik} \sum_j \frac{F_{ij} G_{jk}}{q_j} \leq \sum_{ijk} (C_{ij} + C_{jk}) \frac{F_{ij} G_{jk}}{q_j} \\
&= \sum_{ijk} C_{ij} \frac{F_{ij} G_{jk}}{q_j} + \sum_{ijk} C_{jk} \frac{F_{ij} G_{jk}}{q_j} \\
&= \sum_{ij} C_{ij} F_{ij} \underbrace{\sum_k \frac{G_{jk}}{q_j}}_{\mathbf{1}} + \sum_{jk} C_{jk} F_{ij} \underbrace{\sum_i \frac{F_{ij}}{q_j}}_{\mathbf{1}} \\
&= \mathrm{tr}(\mathbf{CF}) + \mathrm{tr}(\mathbf{CG}) = W(\mathbf{p}, \mathbf{q}) + W(\mathbf{q}, \mathbf{t})
\end{aligned}$$

The first inequality comes from suboptimality, and the second from triangular inequality of the cost matrix $\mathbf{C}$ (it is a distance matrix). This concludes the proof. □

There is an equivalent definition of p-Wasserstein distance for arbitrary measures which we give next.

Theorem 2 (Wasserstein distance for arbitrary measures)**.** *Consider $\alpha(x) \in \mathcal{M}_+^1(\mathcal{X}), \beta(y) \in \mathcal{M}_+^1(\mathcal{Y})$, $\mathcal{X} = \mathcal{Y}$, and for some $p \geq 1$,*

- $c(x, y) = c(y, x) \geq 0$;
- $c(x, y) = 0$ *if and only if* $x = y$;

- $\forall (x, y, z) \in \mathcal{X}^3, c(x, y) \leq c(x, z) + c(z, y)$

Then,

$$W_p(\alpha, \beta) = \mathcal{L}_{c^p}(\alpha, \beta)^{1/p} \tag{12}$$

defines a p-Wasserstein *distance on* $\mathcal{X}$*.*

Recall, that the Kantorovich problem for arbitrary measures is given by (11).

6 Dual problem

We can use duality theory to find the dual formulation of the Kantorovich problem. If the reader is unfamiliar with duality theory, we provide some brief lecture notes with a basic description and introduction in the subject to further understand the following concepts.

First, because the LP is a linear program, and is feasible for $\mathbf{p} \in \Delta_n$ and $\mathbf{q} \in \Delta_n$, the dual problem is also feasible and strong duality holds.

Theorem 3 (Dual problem of the discrete Kantorovich formulation)**.** *Given* $\mathbf{p} \in \mathbb{R}^n$*,* $\mathbf{q} \in \mathbb{R}^n$ *and* $\mathbf{C} \in \mathbb{R}^{n \times n}$*, the dual of* $L_{\mathbf{C}}(\mathbf{p}, \mathbf{q})$ *has the following form:*

$$\begin{aligned} \max_{\mathbf{r},\mathbf{s}} \quad & \mathbf{p}^T\mathbf{r} + \mathbf{q}^T\mathbf{s} \\ s.t. \quad & \mathbf{r}\mathbf{1}^T + \mathbf{1}^T\mathbf{s} \leq \mathbf{C} \end{aligned} \tag{13}$$

where $\mathbf{r} \in \mathbb{R}^n$*,* $\mathbf{s} \in \mathbb{R}^n$*.*

Proof. We first construct the semilagrangian of the primal problem:

$$J(\mathbf{F}; \mathbf{r}, \mathbf{s}) = \operatorname{tr}(\mathbf{C}\mathbf{F}^T) + \mathbf{r}^T(\mathbf{p} - \mathbf{F}\mathbf{1}) + \mathbf{s}^T(\mathbf{q} - \mathbf{F}^T\mathbf{1}). \tag{14}$$

The dual funcion is the minimum of the previous function, and the dual problem is the maximization of the dual function. This gives the following expression,

$$\max_{\mathbf{r},\mathbf{s}} \mathbf{r}^T\mathbf{p} + \mathbf{s}^T\mathbf{q} + \min_{\mathbf{F} \geq 0} \operatorname{tr}(\mathbf{C}\mathbf{F}^T) - \underbrace{\mathbf{r}^T\mathbf{F}\mathbf{1}}_{\operatorname{tr}(\mathbf{F}^T\mathbf{r}\mathbf{1}^T)} - \underbrace{\mathbf{s}^T\mathbf{F}^T\mathbf{1}}_{\mathbf{F}^T\mathbf{1}\mathbf{s}^T} \tag{15}$$

where $\mathbf{Q} = \mathbf{C} - \mathbf{r}\mathbf{1}^T - \mathbf{1}\mathbf{s}^T$.

The solution to the dual problem can be summarized as follows:

$$\min_{\mathbf{F} \geq 0} \operatorname{tr}(\mathbf{C}\mathbf{F}^T) - \underbrace{\mathbf{r}^T\mathbf{F}\mathbf{1}}_{\operatorname{tr}(\mathbf{F}^T\mathbf{r}\mathbf{1}^T)} - \underbrace{\mathbf{s}^T\mathbf{F}^T\mathbf{1}}_{\mathbf{F}^T\mathbf{1}\mathbf{s}^T} = \begin{cases} 0 & \text{if } \mathbf{Q} \geq 0 \\ -\infty & \text{otherwise,} \end{cases} \tag{16}$$

which finally yields the previous result.

$$\begin{aligned} \max_{\mathbf{r},\mathbf{s}} \quad & \mathbf{r}^T\mathbf{p} + \mathbf{s}^T\mathbf{q} \\ \text{s.t.} \quad & \mathbf{r}\mathbf{1}^T + \mathbf{1}^T\mathbf{s} \leq \mathbf{C}. \end{aligned} \tag{17}$$

□

The dual problem can play an important part in devising algorithms to solve the Kantorovich problem. For example, the network simplex is an algorithm that iteratively finds feasible solutions in the primal problem, and updates the dual variables until they become feasible. Similarly, primal-dual methods exploit the gap between the primal and dual problems to find optimal solutions.

Furthermore, the dual variables corresponding to the Kantorovich problem also have an interpretable function, as they can be related to the price payed by the transport plan to move the goods in $\mathbf{p}$, and to collect the goods in $\mathbf{q}$ amounts at the warehouses. This constitutes an alternative view to the *mass* transportation explanation that is available from the primal formulation.

Finally, to conclude this section and for the sake of completeness, we provide the dual formulation for arbitrary measures.

Theorem 4 (Dual of the Kantorovich problem with arbitrary measures)**.** *Given $\alpha \in \mathcal{M}(\mathcal{X})$, $\beta \in \mathcal{M}(\mathcal{Y})$ and $c(x, y)$ a distance function, one has*

$$\mathcal{L}_c(\alpha, \beta) = \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y) \tag{18}$$

where the set of admissible dual potentials is

$$\mathcal{R}(c) = \{\, (f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \;\mid\; \forall (x, y), f(x) + g(y) \leq c(x, y) \,\}\,. \tag{19}$$

In this case we search for a pair of continuous functions f and g that satisfy the cost constraint, and maximize the revenue for transporting the α and β masses. The interpretation of f and g can also be understood from the perspective of transportation prices.

7 Special cases

There are certain special cases where Wassertein has a very clear interpretation, or a closed form solution.

In the discrete case where $\mathbf{C} = \mathbf{1}\mathbf{1}^T - \mathbf{I}$, the solution to the Kantorovich problem reduces to a l1-norm computation, i.e., $L_{\mathbf{C}} = \|\mathbf{p} - \mathbf{q}\|_1$.

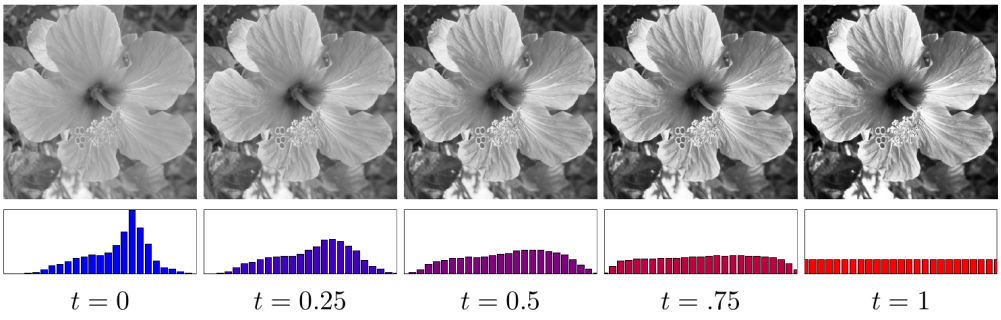
For discrete measures with equiprobable elements in 1D case, i.e., $\mathcal{X} = \mathbb{R}$, and $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_i \delta_{y_i}$, there is also closed form solution. In this case, we can assume that the locations of the discrete measure elements are ordered, i.e., $x_1 \leq x_2, \ldots \leq x_n$ and $y_1 \leq y_2, \ldots \leq y_n$. Then, the Wasserstein loss value has the simple formula

$$W_p(\mathbf{p}, \mathbf{q})^p = \sum_{i=1}^{n} |x_i - y_i|^p, \tag{20}$$

where p refers to the l-norm of two vectors.

Discrete optimal transport for 1D elements has an interesting application for histogram equalization. Assume we have two images with different light profiles and we want to interpolate an image between both inputs. The idea is as follows, convert the images into vectors,

Figure 5: Histogram equalization where t parametrizes the displacement interpolation between the histograms.



where each pixel corresponds to a luminosity value. Then, sort the pixels with a permuation such that they are ordered in increasing values of luminosity, and do this procedure for both source and target images. Then, create a new vector that is the convex interpolation between the previous sorted images, and undo the permutation. This will result into a new image that has transported luminosity from the target image to the source image along the geodesics defined by the Wasserstein distance. The results of this procedure are presented in Figure 5.

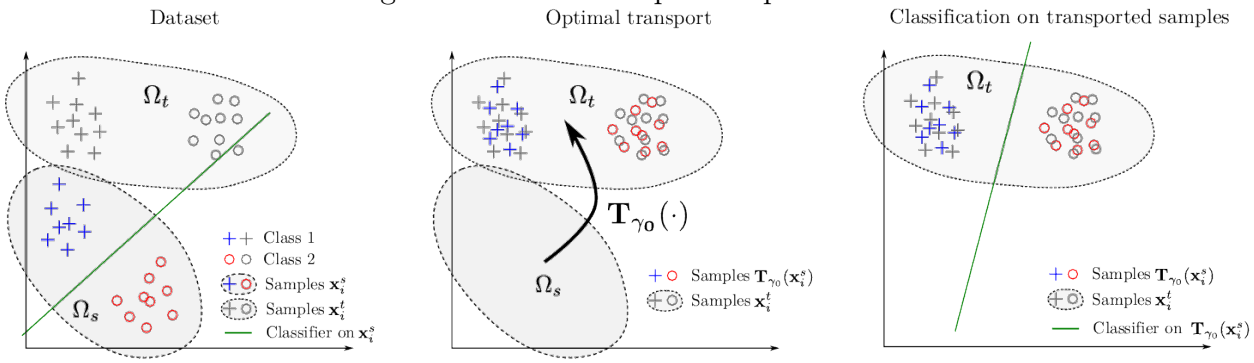
Transport of color from an image to second one is also possible, and is analyzed in [3]. Shape preservations is more convoluted than in the previous case, where both images are the same with different illumination patters. Nonetheless, color transfer it can be achieved with proper weighting and analysis. The authors precompute the images to identify clusters where there are high and low variations of color and compute a graph of indexes and weights for the images. Then, they propose a regularized OT problem where they penalize high variations (through a divergence, similar to a total variation loss), and the pixels transportation cost is weighted according to the cluster weight they belong to. In that sense, they are able to transport color, and preserve edges and shapes. We refer the reader to the author's original publication to observe their outcome results.

8 Applications

8.1 Classification with Wasserstein loss function

A first application to consider is classification when there exists a semantic relation between the classes of a classification problem. In such setting, certain classes may be closely related and in order to learn more appropriately, these classes should not be penalize as strongly as others withe less semantic relation. The reason is that these related examples may be confused easily because of their close relationship in the metric space, and the classifier should not give these examples the same weight as mistakes of more unrelated classes. Wasserstein loss allows to define these relations on the space of class labels for some metric, and learn a classifier that weights these relations. We refer the description of the techniques and employed analysis to the original paper [4].

Figure 6: Domain adaptation procedure.



8.2 Domain adaptation using optimal transport

In this application we consider a classification task in a target space where there are no available labels. However, we assume there exists a source domain with a different joint probability distribution (between features and labels), where such labels are available. The goal is to learn a transport plan that maps objects from the source space to the target space, and then train a classifier in the trasported domain.

The idea is described in Figure 6 and we refer the reader to the original publication [5] to further complete the techniques and ideas presented in class.

8.3 Other problems or applications of interest

1. **Approximate methods** to scale problem dimensions, such as Sinkhorn or smooth OT.
2. **Ground metric learning** allows to learn the cost matrix from data, potentially improving performance compared to a p-Wasserstein loss as we have seen in examples.
3. Barycenter estimation: for clustering, or interpolation between histograms.
4. Transfer learning.
5. Unbalanced optimal transport.
6. Wasserstein discriminant analysis.
7. Etc.

9 Acknowledgments

All figures except Figure 6 are borrowed from [2]. Figure 6 is borrowed from [5].

References

[1] Gaspard Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de lAcadmie Royale des Sciences*, p. 666704, 1781.

[2] Gabriel Peyré and Marco Cuturi, "Computational optimal transport," 2018, arXiv:1803.00567.

[3] Julien Rabin, Sira Ferradans, and Nicolas Papadakis, "Adaptive color transfer with relaxed optimal transport," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 4852–4856.

[4] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio, "Learning with a wasserstein loss," 2015, arXiv:1506.05439.

[5] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, "Optimal transport for domain adaptation," 2015, arXiv:1507.00504.