Self-study: Optimal Transport

Tianpei Xie

Aug. 17th., 2022

Contents

1 The problem of optimal transport 2
1.1 Probability measures . . . 2
1.2 Optimal assignment problem . . . 2
1.2.1 Problem formulations . . . 2
1.2.2 Push-forward on density . . . 3
1.2.3 Push-forward vs. pull-back . . . 4
1.2.4 Push-forward operator for probablity measures . . . 4
1.3 Kantorovich relaxation . . . 4
1.3.1 Optimal transport as linear programming . . . 4
1.3.2 Kantorovich Relaxation is tight for optimal assignment . . . 5
1.3.3 Generalization to arbitrary measures: infinite-dimension LP . . . 6
1.3.4 Probabilistic interpretation . . . 7
1.4 Dual problem . . . 7
1.4.1 Dual formulation and Kantorovich potentials . . . 7
1.4.2 Unconstrained dual . . . 8
1.4.3 Probabilistic interpretation for dual problem . . . 8
1.5 The newtork flow problem . . . 9

2 C-transform 10
2.1 Properties of c-transform . . . 10
2.2 Dual formulation with c-transform . . . 12

3 Wasserstein distance 13
3.1 Definition . . . 13
3.2 Basic properties of Wasserstein distance . . . 15
3.3 Brenier theorem: Monge-Kantorovich equivalence . . . 16
3.3.1 Probabilistic interpretation for alternative dual problem . . . 17

4 Examples of Wasserstein distance 17
4.1 Wasserstein distance between discrete measures . . . 17
4.2 p-Wasserstein distance on 1-dimensional real line $\mathbb{R}$ . . . 18
4.3 1-Wasserstein distance $\mathcal{W}_1$ . . . 18
4.4 Wasserstein distance between Gaussian measures . . . 21

1 The problem of optimal transport

The problem of **optimal transport** concerns about **moving** *simultaneously several items* (or a continuous distribution thereof) from one configuration onto another with the **least amount of cost** or effort. All major economic problems, in logistics, production planning or network routing, involve moving distributions, and that thread appears in all of the seminal references on optimal transport. Recently, OT is being applied in Machine Learning (ML). Examples include *robust optimization*, *Langevin Monte Carlo (LMC) sampling*, *Generative Adversarial Networks (GANs)*, *Nonnegative Matrix Factorization (NMF)*, semi supervised learning etc. This chapter is mainly based on journals [Peyr and Cuturi, 2019].

1.1 Probability measures

Consider the space of probability simplex $\Delta_n := \{p_1, \dots, p_n : \sum_{i=1}^n p_i = 1, \;\; p_i \geq 0\}$. We can define a **discrete measure** (i.e. distribution of discrete random variables)

$$\alpha := \sum_{i=1}^{n} a_i \delta_{\boldsymbol{x}_i}$$

where $\delta_{\boldsymbol{x}_i}$ is the point mass (Dirac funcation) at $\boldsymbol{x}_i$. $\boldsymbol{a} \in \Delta_n$. α can be used to describe the distribution of the mass in one location.

We also consider the distribution on continous random variable. Specifically, we consider the set of ***Radon measures*** $\mathcal{M}(\mathcal{X})$ on the space $\mathcal{X}$ [Folland, 1999]. Denote $\mathcal{C}(\mathcal{X})$ be the space of all continous functions f on $\mathcal{X}$ with compact support. By **the Riesz representation theorem**, for a positive linear functional I on $\mathcal{C}(\mathcal{X})$, there exists on a unique Radon measure α such that

$$I(f) = \int_{\mathcal{X}} f(x) d\alpha(x) \in \mathbb{R}, \quad \text{for all } f \in \mathcal{C}(\mathcal{X})$$

If $\mathcal{X} := \mathbb{R}^d$, we have Lebesgue measure and $\int_{\mathcal{X}} f(x) d\alpha(x) = \int_{\mathcal{X}} f(x) \rho_\alpha(x) dx, \;\; f \in \mathcal{C}(\mathbb{R}^d)$ for some density $\rho_\alpha = \frac{d\alpha}{dx}$. In this document, we consider only the probablity measure $\int_{\mathcal{X}} d\alpha(x) = 1$ and $\alpha \in \mathcal{M}_+(\mathcal{X})$ is positive measure, i.e. $\alpha \in \mathcal{M}_+^1(\mathcal{X})$.

1.2 Optimal assignment problem

1.2.1 Problem formulations

The **Monge problem** between discrete measures consider the problem of **pushing forward** n piles of mass $\boldsymbol{a} = [a_1, \dots, a_n]$ at one location $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$ to m piles $\boldsymbol{b} = [b_1, \dots, b_m]$ at another location $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_m]$ with *minimized* cost of transportation.

Specifically, given two discrete measures $\alpha = \sum_{i=1}^n a_i \delta_{\boldsymbol{x}_i}$ and $\beta := \sum_{i=1}^m b_i \delta_{\boldsymbol{y}_i}$, consider a map $T : \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \to \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_m\}$. The **optimal assignment problem** is defined by the following optimization problem:

$$\min_T \sum_i c(\boldsymbol{x}_i, T(\boldsymbol{x}_i)) \tag{1}$$

$$\text{s.t. } \boldsymbol{b}_j = \sum_{i:T(\boldsymbol{x}_i)=\boldsymbol{y}_j} \boldsymbol{a}_i, \ \forall\, j, \tag{2}$$

where $c(\boldsymbol{x}, \boldsymbol{y})$ is the cost of transportation between position $\boldsymbol{x}$ and $\boldsymbol{y}$. We can define $\boldsymbol{C}_{n,m} := [C_{i,j}]_{i\in[1:n],j\in[1:m]}$, where $C_{i,j} := c(\boldsymbol{x}_i, \boldsymbol{y}_j) \geq 0$. Note that if $n = m$, the map $T = \sigma \in \mathrm{Perm}(n)$, where $\mathrm{Perm}(n)$ is the set of all permutations for $[1 : n]$. If $\boldsymbol{a} = \boldsymbol{b} = \frac{1}{n}\boldsymbol{1}$ is uniformly distributed, then the constraint (2) can be removed. We have the classical optimal assignment problem [Peyr and Cuturi, 2019]

$$\min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)} \tag{3}$$

The constraint in (2) defines a ***Push-forward operator*** for discrete measures $T_\# \alpha := \sum_{i=1}^{n} a_i \delta_{T(\boldsymbol{x}_i)}$. Here we can define the push-forward operator for general (Radon) measures:

Definition For a *continous* map $T : \mathcal{X} \to \mathcal{Y}$, the ***push-forward operator*** is defined as $T_\# : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ that satisfies

$$\forall\, h \in \mathcal{C}(\mathcal{X}), \quad \int_{\mathcal{Y}} h(\boldsymbol{y})\, d\,(T_\#\alpha)\,(\boldsymbol{y}) = \int_{\mathcal{X}} h(T(\boldsymbol{x}))\, d\alpha(\boldsymbol{x}). \tag{4}$$

$$\text{or equivalently,} \quad (T_\#\alpha)\,(B) := \alpha\,(\{\boldsymbol{x} : T(\boldsymbol{x}) \in B \subset \mathcal{Y}\}) = \alpha(T^{-1}(B)) \tag{5}$$

where the **push-forward measure** $\beta := T_\#\alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$, $T^{-1}(\cdot)$ is the pre-image of T, and $\mathcal{M}(\mathcal{X})$ is the set of *Radon measures* on the space $\mathcal{X}$.

Note that for positive measure $\mathcal{M}_+(\mathcal{X})$, $T(\mathcal{M}_+(\mathcal{X})) \subseteq \mathcal{M}_+(\mathcal{Y})$, i.e. the push-forward operator $T_\#$ preserves positivity and total mass. A push-forward operator is also **linear**: $T_\#(w_1\alpha_1 + w_2\alpha_2) = w_1T_\#(\alpha_1) + w_2T_\#(\alpha_2)$. The operator $T_\#$ pushes forward each elementary mass of a measure α on $\mathcal{X}$ by applying the map T to obtain then an elementary mass in $\mathcal{Y}$.

With the definition of push-forward operator, the **Monge problem between arbitrary measures** can be formulated as

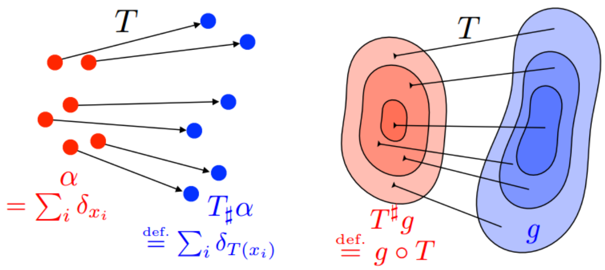
$$\min_T \int_{\mathcal{X}} c(\boldsymbol{x}, T(\boldsymbol{x}))d\alpha(\boldsymbol{x}) \tag{6}$$

$$\text{s.t. } \beta = T_\#\alpha \tag{7}$$

where two arbitrary probability measures (α, β), supported on two spaces $(\mathcal{X}, \mathcal{Y})$ can be linked through a map $T : \mathcal{X} \to \mathcal{Y}$ so that T push forward the mass of α to β. The equality constraint with push-forward operator in fact defines the **mass conservation constraint**.

1.2.2 Push-forward on density

Assume that (α, β) have densities $(\rho_\alpha, \rho_\beta)$ with respect to a fixed measure, and $\beta = T_\#\alpha$. In multivariate distribution/density, we see that $T_\#$ acts on a density ρ_α linearly to a density ρ_β as a


T
α
$= \sum_i \delta_{x_i}$
$T_\sharp \alpha$
$\overset{\text{def.}}{=} \sum_i \delta_{T(x_i)}$
T
$T^\sharp g$
$\overset{\text{def.}}{=} g \circ T$
g

Push-forward of measures Pull-back of functions

Figure 1: The comparison between push-forward operator and pull-back operator

change of variable, i.e.

$$\rho_\alpha(\boldsymbol{x}) = \left|\det(T'(\boldsymbol{x}))\right| \rho_\beta(T(\boldsymbol{x})) \tag{8}$$

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

1.2.3 Push-forward vs. pull-back

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

$$\forall(\alpha, g) \in \mathcal{M}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : \quad \int_{\mathcal{Y}} gd\left(T_\#\alpha\right)(\boldsymbol{x}) = \int_{\mathcal{X}} T^\# g d\alpha(\boldsymbol{x}) := \int_{\mathcal{X}} g \circ T d\alpha(\boldsymbol{x}) \tag{9}$$

$$\Leftrightarrow (T_\#\alpha)(A) = \alpha(T^{-1}(A)) \tag{10}$$

Note that even if (α, β) have densities $(\rho_\alpha, \rho_\beta)$ with respect to a fixed measure and $\beta = T_\#\alpha$, the density of push-forward measure is not equal to the pull-back of density, i.e. $\rho_{T_\#\alpha} \neq T^\# \rho_\beta$. The gap is the Jacobian related to change of variables.

1.2.4 Push-forward operator for probablity measures

Radon measures can also be viewed as representing the distributions of random variables. The distribution of random variable X is a Randon measure $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, which is defined by $\alpha(A) := \int_A X(\omega) d\omega = \mathbb{P}(X \in A)$. Following the definition above, the **distribution of** X, α, can be seen as a **push-forward** of $\mathbb{P}$ by X, i.e. $\alpha = X_\#\mathbb{P}$. For a random variable Y so that $Y = T(X)$, the *distribution of* Y *can be seen as a push-forward of the distribution of of* X by T, i.e. $\beta = T_\#\alpha$.

1.3 Kantorovich relaxation

1.3.1 Optimal transport as linear programming

The optimal assignment problem (1) is ***combinatorial*** and for its continous case (6) its feasible set (the set of all push-forward operators) is ***non-convex***. Both are therefore difficult to solve when approached in their original formulation.

Kantorovich comes up with several critical relaxation:

- instead of considering *deterministic transport* T, Kantorovich considers a ***probabilistic transport*** $\boldsymbol{P} \in \mathbb{R}_+^{n\times m}$. This change allows for *mass splitting* from a source toward several targets, i.e. from one-to-one mapping to one-to-many mapping.

- instead of considering *asymmetric* transport i.e. move mass from one location $\mathcal{X}$ to another location $\mathcal{Y}$, Kantorovich considers a **symmetric** transport, i.e. we can change the move direction from $\mathcal{Y}$ to $\mathcal{X}$ and the optimal solution is unchanged.

The ***optimal transport*** under Kantorovich relaxation is formulated as a *linear programming problem*:

$$
\begin{aligned}
\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \langle \boldsymbol{P}\,,\,\boldsymbol{C}\rangle &= \sum_{i,j} C_{i,j}P_{i,j} && (11)\\
\text{s.t. }\ \boldsymbol{P}\boldsymbol{1}_m &= \boldsymbol{a} && (12)\\
\boldsymbol{P}^T\boldsymbol{1}_n &= \boldsymbol{b} && (13)\\
P_{i,j} &\geq 0
\end{aligned}
$$

where $\boldsymbol{C}_{n,m} := [C_{i,j}]_{i\in[1:n],j\in[1:m]}$, $C_{i,j} := c(\boldsymbol{x}_i, \boldsymbol{y}_j) \geq 0$. The **coupling matrix** $\boldsymbol{P} = [P_{i,j}]_{n,m}$ where $P_{i,j}$ describes the amount of mass flowing from bin i toward bin j, or from the mass found at $\boldsymbol{x}_i$ toward $\boldsymbol{y}_j$ in the formalism of discrete measures. Denote the **optimal value** of this problem $L_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) := \min_{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}\,,\,\boldsymbol{C}\rangle$ where $U(\boldsymbol{a},\boldsymbol{b}) := \{\boldsymbol{P}\in\mathbb{R}_+^{n\times m} : \boldsymbol{P}\boldsymbol{1}_m = \boldsymbol{a},\ \boldsymbol{P}^T\boldsymbol{1}_n = \boldsymbol{b}\}$ defines the feasible region.

It is natural to represent the optimal value $L_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b})$ in terms of the discrete measures $\alpha = \sum_{i=1}^n a_i\delta_{\boldsymbol{x}_i}$ and $\beta := \sum_{i=1}^m b_i\delta_{\boldsymbol{y}_i}$, thus the **optimal transport** between α and β is

$$
\mathcal{L}_c(\alpha,\beta) := L_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) = \min_{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}\,,\,\boldsymbol{C}\rangle \qquad (14)
$$

Here we emphasis the cost function $c(x,y)$.

1.3.2 Kantorovich Relaxation is tight for optimal assignment

We can reformulate the original optimal assignment problem (3) into (11). Consider the matrix representation of permuation $\sigma \in \mathrm{Perm}(n)$ as

$$
\begin{aligned}
\boldsymbol{P}_\sigma &= [P_{i,j}]_{n\times n}\\
P_{i,j} &= \begin{cases} \frac{1}{n} & \text{if } j = \sigma(i)\\ 0 & \quad \text{o.w.}\end{cases}
\end{aligned}
$$

Note that $\mathrm{Perm}(n)$ is a symmetric group, usually denoted as S_n. $\boldsymbol{P}_\sigma$ is a coupling matrix as defined in (11) since each row and column only has one number $1/n$, i.e. $\boldsymbol{P}_\sigma\boldsymbol{1}_n = 1/n$ and $\boldsymbol{P}_\sigma^T\boldsymbol{1}_n = 1/n$. We see that (11) can be converted into the classical optimal assignment problem (3). Note that the permutation matrix $\boldsymbol{P}_\sigma$ is a feasible solution, i.e.

$$
L_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) \leq \langle \boldsymbol{P}_\sigma^*\,,\,\boldsymbol{C}\rangle = \min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n}\sum_{i=1}^n C_{i,\sigma(i)}
$$

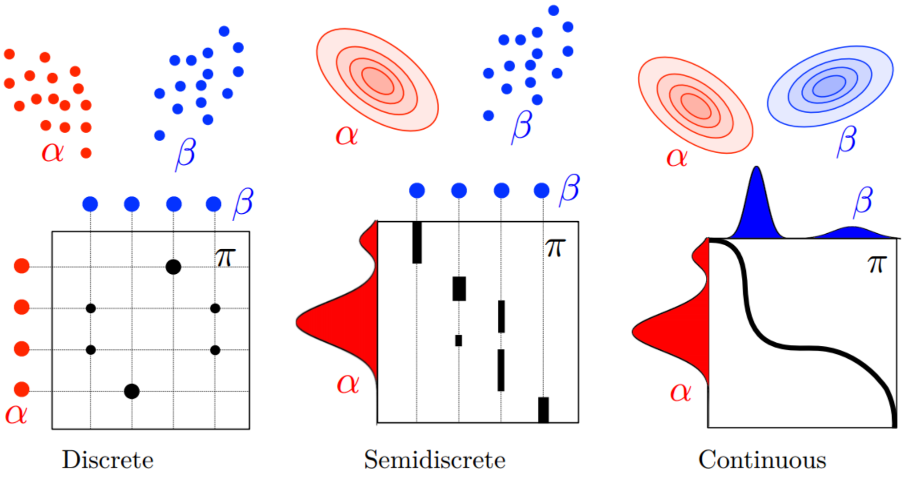


Figure 2: The coupling matrix between discrete measures, between discrete-and-continous measures, and between continuous measures

Now we show that if $\boldsymbol{a} = \boldsymbol{b} = \mathbf{1}_n/n$, then solving LP (11) results in the same optimal solution of (3), i.e. the Kantorvich relaxation is ***tight*** for optimal assignment problem.

Proposition 1.1 *(**Kantorovich for matching**). If $m = n$ and $\boldsymbol{a} = \boldsymbol{b} = \mathbf{1}_n/n$, then there exists an optimal solution for Problem* (11) $\boldsymbol{P}_\sigma$*, which is a permutation matrix associated to an optimal permutation $\sigma \in$ Perm(n) for Problem* (3)*.*

Proof: Note that for finite-dimensional linear programming with non-empty feasible set, the optimal solution exists and it can only be found at the *extreme point* of the non-empty polyhedron $U(\boldsymbol{a}, \boldsymbol{b})$. For $\boldsymbol{a} = \boldsymbol{b} = \mathbf{1}_n/n$, the feasible set $U(\mathbf{1}_n/n, \mathbf{1}_n/n)$ is called the **Birkhoff polytope**. By Birkhoff's theorem [1946], the set of extremal points of $U(\mathbf{1}_n/n, \mathbf{1}_n/n)$ is equal to the set of permutation matrices. Therefore there exists a permutation σ whose matrix representation $\boldsymbol{P}_\sigma$ is the optimal solution. Thus the (3) and (11) are equivalent. ▮

1.3.3 Generalization to arbitrary measures: infinite-dimension LP

We can generalize the Kantorovich relaxation to continous measures. Let us define the set of coupling π as

$$U(\alpha, \beta) := \left\{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : \;\; P_{\mathcal{X}\#}\pi = \alpha, \; P_{\mathcal{Y}\#}\pi = \beta\right\} := \Gamma(\alpha, \beta) \tag{15}$$

Here $P_{\mathcal{X}\#}$ and $P_{\mathcal{Y}\#}$ are the *push-forwards* of the **projections** $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$. This constraint is equivalently enposing that

$$\pi(A \times \mathcal{Y}) = \alpha(A), \;\; \pi(\mathcal{X} \times B) = \beta(B), \quad \text{for any sets } A \subset \mathcal{X}, \; B \subset \mathcal{Y}.$$

The **Kantorovich problem between arbitrary measures** can be formulated as

$$\begin{aligned} \min_{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) \\ \text{s.t. } P_{\mathcal{X}\#}\pi = \alpha, \\ P_{\mathcal{Y}\#}\pi = \beta \end{aligned} \tag{16}$$

And the **optimal transport** is defined as the optimal value $\mathcal{L}_c(\alpha, \beta) := \inf_{\pi \in U(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y)$.

The problem (16) is a **infinite-dimensional linear programming problem** over 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**, i.e. the feasible set $U(\alpha, \beta) \neq \emptyset$.

1.3.4 Probabilistic interpretation

Let the coupling $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ is the **joint probability measure** of (X, Y) over $\mathcal{X} \times \mathcal{Y}$ and α and β be the distribution of X and Y, respectively, the problem (16) can be formulated as

$$\begin{aligned} \min_{(X,Y)\sim\pi} \; & \mathbb{E}_{(X,Y)} \left[c(X, Y) \right] \\ \text{s.t. } \; & X \sim \alpha, \\ & Y \sim \beta \end{aligned} \tag{17}$$

Note that the push-forward constraint states that the ***marginalization*** of π in $\mathcal{X}$ and $\mathcal{Y}$ is equal to the distribution measure α and β, respectively. That is, given marginal distribution $X \sim \alpha$ and $Y \sim \beta$, we need to find π as the joint distribution of (X, Y) that match each marginal distribution, which is a natural law for joint distribution. In fact, all possible joint distribution π in $\mathcal{X} \times \mathcal{Y}$ that fit the marginal distribution in each random variable is a feasible solution.

1.4 Dual problem

1.4.1 Dual formulation and Kantorovich potentials

The linear programming problems of optimal transport in (11) have **dual problems**. In fact, due to **strong duality** of LP [Bertsimas and Tsitsiklis, 1997], the dual problem and the primal problem share the same optimal value.

The **dual problem** of Kantorovich problem (11) is described as below:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n, \boldsymbol{\mu}\in\mathbb{R}^m} \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle \tag{18}$$

$$\text{s.t. } \lambda_i + \mu_j \le C_{i,j} \quad \forall\, i \in [1:n], j \in [1:m] \tag{19}$$

where $\boldsymbol{\lambda} = [\lambda_i]_n$, $\boldsymbol{\mu} = [\mu_j]_m$ are **dual variables** (slack variables) for marginal distribution constrain $\boldsymbol{a}$ and $\boldsymbol{b}$. We denote $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} := \boldsymbol{\lambda}\mathbf{1}_m^T + \mathbf{1}_n\boldsymbol{\mu}^T \in \mathbb{R}^{n\times m}$ so that the linear constraints is $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} \le \boldsymbol{C}$. Such dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are often referred to as "***Kantorovich potentials***." The feasible set of the dual problem is defined as

$$R(\boldsymbol{C}) := \{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m : \boldsymbol{\lambda} \oplus \boldsymbol{\mu} \le \boldsymbol{C}\} \tag{20}$$

The relationship between (11) and (18) are established via strong duality [Bertsimas and Tsitsiklis, 1997].

The dual problem can be interpreted as a **Optimal Prices problem**: Suppose that the operator does not have the computational means to solve the linear program (11). He decides instead to outsource that task to a vendor. The vendor chooses a pricing scheme with the following structure: the vendor splits the logistic task into that of collecting and then delivering the goods and will apply a collection price λ_i to collect a unit of resource at each warehouse i (no matter where that unit is sent to) and a price μ_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 a_i units at warehouse i and b_j needed at factory j, the vendor asks as a consequence of that pricing scheme a price of $\langle \boldsymbol{\lambda}, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle$ to solve the operators logistic problem.

Note that for any feasilble solution $\boldsymbol{P}$ in (11), we have weak duality inequality

$$\langle \boldsymbol{P}, \boldsymbol{C} \rangle \geq \langle \boldsymbol{\lambda}, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle$$

We can also generalize to **dual problem to arbitrary measures** as dual problem of (16):

$$\mathcal{L}_c(\alpha, \beta) = \max_{(\lambda,\mu)\in\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \lambda(x)d\alpha(x) + \int_{\mathcal{Y}} \mu(y)d\beta(y) \tag{21}$$

$$\text{s.t. } \lambda(x) + \mu(y) \leq c(x,y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, \tag{22}$$

Here, (λ, μ) is a pair of continuous functions and are also called, as in the discrete case, "***Kantorovich potentials***." The feasible region is

$$\mathcal{R}(c) := \{(\lambda, \mu) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : \lambda \oplus \mu \leq c\} \tag{23}$$

where $(\lambda \oplus \mu)(x,y) = \lambda(x) + \mu(y)$.

By the primal-dual optimality conditions , we can track the **support** of the optimal plan π as

$$\text{Supp}(\pi) := \{\pi \in \mathcal{M}_+(\mathcal{X}, \mathcal{Y}) : \pi > 0\} \subset \{(x,y) \in \mathcal{X} \times \mathcal{Y} : \lambda(x) + \mu(y) = c(x,y)\} = \partial\mathcal{R}(c) \tag{24}$$

That is the optimal solution in primal problem (16) can only be found at the **boundary** of the feasible set in dual problem (21), when the equality of constraint is met.

1.4.2 Unconstrained dual

If α and β are probablity measure $\int d\alpha = 1$ and $\int d\beta = 1$, the constrained dual problem (21) can be replaced by an unconstrained one,

$$\mathcal{L}_c(\alpha, \beta) = \max_{(\lambda,\mu)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \lambda(x)d\alpha(x) + \int_{\mathcal{Y}} \mu(y)d\beta(y) + \min_{\lambda\oplus\mu}\{c - \lambda \oplus \mu\} \tag{25}$$

where $(\lambda \oplus \mu)(x,y) = \lambda(x) + \mu(y)$. Here the minimum should be considered as the essential supremum associated to the measure $\alpha \oplus \beta$, i.e., it does not change if λ or μ is modified on sets of zero measure for α and β. It is obtained from the primal problem (16) by adding the redundant constraint $\int d\pi = 1$. An alternative formulation is discussed next chapter via c-transform.

1.4.3 Probabilistic interpretation for dual problem

We see that the objective in (21) can be interpreted via probabilty measure $\alpha = X_{\#}\mathbb{P}$ and $\beta = Y_{\#}\mathbb{P}$,

$$\mathcal{L}_c(\alpha, \beta) = \max_{(\lambda,\mu)\in\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \mathbb{E}_{X\sim\alpha}[\lambda(X)] + \mathbb{E}_{Y\sim\beta}[\mu(Y)] \tag{26}$$

$$\text{s.t. } \lambda(x) + \mu(y) \leq c(x,y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y},$$

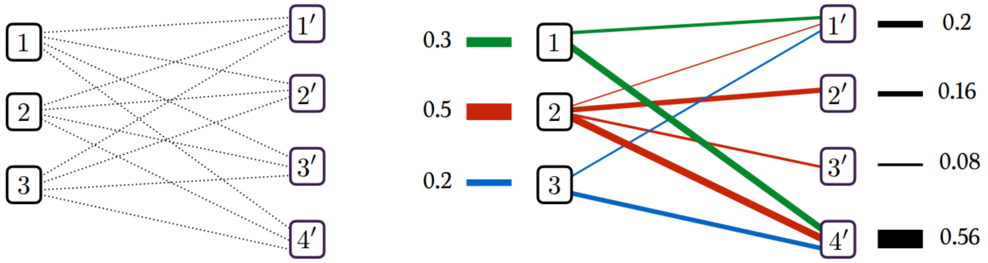


Figure 3.1: The optimal transport problem as a bipartite network flow problem. Here $n = 3, m = 4$. All coordinates of the source histogram, $\mathbf{a}$, are depicted as source nodes on the left labeled $1, 2, 3$, whereas all coordinates of the target histogram $\mathbf{b}$ are labeled as nodes $1',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

Figure 3: The optimal transport problem can be embedded to a graph to become the network flow problem

1.5 The newtork flow problem

The **optimal transport problem** can be interpreted as a ***network flow problem*** when the cost $\boldsymbol{C}$ and the coupling matrix $\boldsymbol{P} = [P_{i,j}]$ is embeded into a **bipartite graph** $\mathcal{G} = (\mathcal{V}_\alpha \cup \mathcal{V}_\beta, \mathcal{E})$. For $i \in \mathcal{V}_\alpha$ and $j \in \mathcal{V}_\beta$, the cost of $C_{i,j} = \infty$ if $(i, j) \notin \mathcal{E}$. The marginal $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^m$ are interpreted as n **sources** and m **sinks**. $P_{i,j}$ is the **network flow** from source i to sink j. For network flow problem $\sum_i^n a_i = \sum_j^m b_j$ for the conservation of mass. The network problem optimizes

$$\min_{\boldsymbol{P} \in \mathbb{R}_+^{n \times m}} \sum_{(i,j) \in \mathcal{E}} C_{i,j} P_{i,j} \tag{27}$$

$$\text{s.t.} \sum_j P_{i,j} = a_i, \forall i \tag{28}$$

$$\sum_i P_{i,j} = b_j, \forall j \tag{29}$$

$$P_{i,j} \geq 0$$

See more variants of network flow problem in [Bertsimas and Tsitsiklis, 1997].

Proposition 1.2 *(The extreme point of feasible region form a tree)*
*Let $\boldsymbol{P} = [P_{i,j}]$ be an extremal point of the polytope $U(\boldsymbol{a}, \boldsymbol{b})$. Let $\mathcal{E}(\boldsymbol{P}) \subset \mathcal{E}$ be the subset of edges $\{(i, j) : P_{i,j} > 0\}$. Then the subgraph $\mathcal{G}(\boldsymbol{P}) := (\mathcal{V}_\alpha \cup \mathcal{V}_\beta, \mathcal{E}(\boldsymbol{P}))$ has **no cycles**. In particular, $\boldsymbol{P}$ cannot have more than $n + m - 1$ nonzero entries.*

2 C-transform

Given the cost function c, we define the **c-transform (c-conjugate)** [Santambrogio, 2015] as an extension to the concept of *conjugate function* introduced in convex optimization [Rockafellar, 1970]. In fact, up to the change of sign, c-concavity has exactly been defined as a generalization of convexity.

Definition Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, $f : \mathcal{X} \to \mathbb{R}$, the ***c-transform*** of f is defined as

$$f^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - f(x) \tag{30}$$

The function $f^c : \mathcal{Y} \to \mathbb{R}$ is also called the ***c-conjugate function*** of f.

For discrete case, we have C***-transform vector*** for cost matrix $\boldsymbol{C} = [C_{i,j}]_{n \times m}$ and vector $\boldsymbol{f} = [f_1, \dots, f_n] \in \mathbb{R}^n$,

$$\boldsymbol{f}_j^{\boldsymbol{C}} := \min_{i \in [1:n]} C_{i,j} - \boldsymbol{f}_i \tag{31}$$

The vector $\boldsymbol{f}^{\boldsymbol{C}} \in \mathbb{R}^m$ is also called the C***-conjugate vector*** of $\boldsymbol{f}$.

Similarly, since the cost function has two variables, we can define $\bar{c}$**-transform**

Definition Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, $g : \mathcal{Y} \to \mathbb{R}$, the $\bar{c}$***-transform*** of g is defined as

$$g^{\bar{c}}(x) := \inf_{y \in \mathcal{Y}} c(x, y) - g(y) \tag{32}$$

For discrete case, we have $\bar{C}$***-transform vector*** $\boldsymbol{g}^{\bar{\boldsymbol{C}}} \in \mathbb{R}^n$ for cost matrix $\boldsymbol{C} = [C_{i,j}]_{n \times m}$ and vector $\boldsymbol{g} = [g_1, \dots, g_m] \in \mathbb{R}^m$,

$$\boldsymbol{g}_i^{\bar{\boldsymbol{C}}} := \min_{j \in [1:m]} C_{i,j} - \boldsymbol{g}_j \tag{33}$$

Note that c-transform or $\bar{c}$-transform can be obtained by **minimizing the dual objective** (25) with respect to λ or μ, respectively while keeping the other one fixed.

Moreover, we can define c-concave and $\bar{c}$-concave functions

Definition A function $\psi : \mathcal{X} \to \mathbb{R}$ is c**-concave** if there exists some function $\phi : \mathcal{Y} \to \mathbb{R}$ and cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ so that ψ is the $\bar{c}$-transform of ϕ, i.e. $\psi = \phi^{\bar{c}}$. Denote ψ as c-concave($\mathcal{X}$)

Definition A function $\phi : \mathcal{Y} \to \mathbb{R}$ is $\bar{c}$**-concave** if there exists some function $\psi : \mathcal{X} \to \mathbb{R}$ and cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ so that ϕ is the c-transform of ψ, i.e. $\phi = \psi^c$. Denote ϕ as $\bar{c}$-concave($\mathcal{Y}$)

Note that a function is c-concave means that it is the $\bar{c}$-transform of some unknown function, and a function is $\bar{c}$-concave means that is is the c-transform of some unknown function.

If $c = d$ is a distance function and $\mathcal{X} = \mathcal{Y}$, due to symmetry, there is **no need to distinguish** c-transform vs. $\bar{c}$-transform (or $\bar{c}$-concave vs. c-concave).

2.1 Properties of c-transform

c-transform plays an important role in dual representations. There are several properties:

- (for **distance metric** c, **no distinction** between c and $\bar{c}$)
  If c is defined as a distance, $\mathcal{X} = \mathcal{Y}$, due to symmetry $\inf_{y \in \mathcal{Y}} c(x,y) - f(x) = \inf_{y \in \mathcal{Y}} c(x,y) - f(y)$, thus $f^c = f^{\bar{c}}$.

- The following inequality is for piecewise

$$\begin{aligned} i) \quad & \boldsymbol{\lambda}_1 \leq \boldsymbol{\lambda}_2 \qquad && \Leftrightarrow \boldsymbol{\lambda}_1^{\boldsymbol{C}} \geq \boldsymbol{\lambda}_2^{\boldsymbol{C}} \\ ii) \quad & \boldsymbol{\lambda}^{\boldsymbol{C}\bar{\boldsymbol{C}}} \geq \boldsymbol{\lambda}, && \quad \boldsymbol{\mu}^{\bar{\boldsymbol{C}}\boldsymbol{C}} \geq \boldsymbol{\mu} \\ iii) \quad & \boldsymbol{\lambda}^{\boldsymbol{C}\bar{\boldsymbol{C}}\boldsymbol{C}} = \boldsymbol{\lambda}^{\boldsymbol{C}} \end{aligned}$$

  Proof: i) is from the definition of C-transform. Note that $\lambda_{1,j}^{\boldsymbol{C}} = \min_i C_{i,j} - \lambda_{1,i} \geq \min_i C_{i,j} - \lambda_{2,i} = \lambda_{2,j}^{\boldsymbol{C}}$ ii) can be found by applying the definitoin twice:

$$\begin{aligned} \lambda_i^{\boldsymbol{C}\bar{\boldsymbol{C}}} &= \min_j C_{i,j} - \left([\min_{i'} C_{i',j} - \lambda_{i'}]_{i'} \right)_j \\ &= \min_j \left(C_{i,j} - \min_{i'} \left(C_{i',j} - \lambda_{i'} \right) \right) \\ &\text{since } \min_{i'} \left(C_{i',j} - \lambda_{i'} \right) \leq C_{i,j} - \lambda_i \\ &\geq \min_j \left(C_{i,j} - (C_{i,j} - \lambda_i) \right) \\ &= \min_j \left(C_{i,j} - C_{i,j} \right) + \lambda_i = \lambda_i \end{aligned}$$

  Similarly we can proof $\boldsymbol{\mu}^{\bar{\boldsymbol{C}}\boldsymbol{C}} \geq \boldsymbol{\mu}$

  To proof iii), note that by ii) $(\boldsymbol{\lambda}^{\boldsymbol{C}})^{\bar{\boldsymbol{C}}\boldsymbol{C}} \geq \boldsymbol{\lambda}^{\boldsymbol{C}}$; Also by i) and ii) since $\boldsymbol{\lambda}^{\boldsymbol{C}\bar{\boldsymbol{C}}} \geq \boldsymbol{\lambda}$, then $(\boldsymbol{\lambda}^{\boldsymbol{C}\bar{\boldsymbol{C}}})^{\boldsymbol{C}} \leq \boldsymbol{\lambda}^{\boldsymbol{C}}$. Therefore $\boldsymbol{\lambda}^{\boldsymbol{C}\bar{\boldsymbol{C}}\boldsymbol{C}} = \boldsymbol{\lambda}^{\boldsymbol{C}}$. $\blacksquare$

  This proof generalize to aribitrary function f and its c-transform f^c for all x and y

  Proposition 2.1 *Suppose that $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is real valued.*

  1. *For any $f_1 : \mathcal{X} \to \mathbb{R}$ and $f_2 : \mathcal{X} \to \mathbb{R}$, $f_1 \leq f_2, \Leftrightarrow \; f_1^c \geq f_2^c$*
  2. *For any $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, $f^{c\bar{c}} \geq f$, $g^{\bar{c}c} \geq g$ In general, $f^{c\bar{c}}$ is the **smallest** c-**concave function** larger than f*
  3. *$f^{c\bar{c}c} = f^c$ and $g^{\bar{c}c\bar{c}} = g^{\bar{c}}$; in other words, $f^{c\bar{c}} = f$ if and only if f is a c-concave function*

- (c-**concavity** implies a bound on the modulus of **continuity**)

  Proposition 2.2 *If $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a distance, then the function $f : \mathcal{X} \to \mathbb{R}$ is c-**concave** if and only if f is **Lipschitz continuous** with Lipschitz constant less than 1 w.r.t. the distance c. We will denote by Lip_1 the set of these functions. Moreover, for every $f \in Lip_1$, i.e. $\|f\|_L \leq 1$, we have the c-transform of f, $f^c = -f$. [Santambrogio, 2015]*

  Proof: $\Rightarrow$: since f is c-concave, there exists some g so that $f = g^{\bar{c}} = \inf_y c(x,y) - g(y)$. Note that the map $c_y : x \to c(x,y) \in \text{Lip}_1$ due to triangular inequality, and c itself is the distance. Therefore $f \in \text{Lip}_1$ since the infimum of a family of Lipschitz continuous functions with the same constant shares the same modulus of continuity.

  $\Leftarrow$: for $f \in \text{Lip}_1$, we claim that $f(x) = \inf_y c(x,y) + f(y)$. To prove the claim, we see that $\inf_y c(x,y) + f(y) \leq c(x,x) + f(x) = f(x)$ since $c(x,x) = 0$ for definiteness of distance c.

On the other hand, since $f \in \mathrm{Lip}_1$, therefore $f(x) - f(y) \leq |f(x) - f(y)| \leq c(x, y)$. We have $f(x) \leq c(x, y) + f(y)$ for all y. Taking infimum over y at both sides, $f(x) \leq \inf_y c(x, y) + f(y)$. Thus we prove that $f(x) = \inf_y c(x, y) + f(y)$. By definition of c-transform and c-concave, we have that $f = (-f)^{\bar{c}}$ so f is c-concave; Moreover, applying $-f \to f$ in the last formula, $f^{\bar{c}} = -f$, due to symmetry of c, $f^c = -f$. ∎

2.2 Dual formulation with c-transform

We present in this section an important property of the dual optimal transport problem (**??**). Note that by strong duality

$$L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \max_{R(\boldsymbol{C})} \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle$$

If we freeze the value of $\boldsymbol{\lambda}$, we can notice that there is no better vector solution for $\boldsymbol{\mu}$ than the ***C-transform vector*** of $\boldsymbol{\lambda}$, denoted $\boldsymbol{\lambda}^{\boldsymbol{C}} \in \mathbb{R}^m$.

Follwing the definition of C-transform vector (31), we see that $\lambda_i + \lambda_j^{\boldsymbol{C}} \leq C_{i,j}$ for all i, j. Therefore $(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{\boldsymbol{C}}) \in R(\boldsymbol{C})$ is a feasible solution of dual problem. Moreover, $\boldsymbol{\lambda}^{\boldsymbol{C}}$ is the largest possible μ to satisfies the constraint. (since $\boldsymbol{\lambda} \oplus \boldsymbol{\lambda}^{\boldsymbol{C}} = \min_i \boldsymbol{C} \leq \boldsymbol{C}$). Therefore

$$\langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle \leq \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \left\langle \boldsymbol{\lambda}^{\boldsymbol{C}}\,,\, \boldsymbol{b} \right\rangle$$

Thus the dual problem (18) is equivalent to an **unconstrained optimization problem**

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \left\langle \boldsymbol{\lambda}^{\boldsymbol{C}}\,,\, \boldsymbol{b} \right\rangle \tag{34}$$

or using $\bar{C}$-transform vector

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^m} \left\langle \boldsymbol{\mu}^{\bar{\boldsymbol{C}}}\,,\, \boldsymbol{a} \right\rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle \tag{35}$$

For probability measure we can write the dual formulation as

$$\begin{aligned} \mathcal{L}_c(\alpha, \beta) &:= \max_{\lambda \in \mathcal{C}(\mathcal{X})} \mathbb{E}_{X \sim \alpha}\left[\lambda(X)\right] + \mathbb{E}_{Y \sim \beta}\left[\lambda^c(Y)\right] && (36) \\ &= \max_{\lambda \in c\text{-concave}(\mathcal{X})} \mathbb{E}_{X \sim \alpha}\left[\lambda(X)\right] + \mathbb{E}_{Y \sim \beta}\left[\lambda^c(Y)\right] && (37) \end{aligned}$$

3 Wasserstein distance

3.1 Definition

We can show that if $c(x, y)$ is a proper distance measure, then the optimal transport $\mathcal{L}_c(\alpha, \beta)$ defines a divergence measure between the probability distribution α and β.

Here we can define the distance between between histograms supported on these bins

Proposition 3.1 *We suppose $n = m$ and that for some $p \geq 1$, $\boldsymbol{C} = \boldsymbol{D}^p = [D^p_{i,j}]_{n\times n} \in \mathbb{R}^{n\times n}$, where $\boldsymbol{D} \in \mathbb{R}^{n\times n}_+$ is a distance on $[1:n]$, i.e.*

1. *$\boldsymbol{D} \in \mathbb{R}^{n\times n}_+$ is symmetric;*
2. *$D_{i,j} = 0$ iff $i = j$;*
3. *(Triangular inequality): $D_{i,j} \leq D_{i,k} + D_{k,j}$, for all $i, j, k \in [1:n]$*

*Then the <u>p- **Wasserstein distance**</u> $W_p(\boldsymbol{a}, \boldsymbol{b}) := L_{\boldsymbol{D}^p}(\boldsymbol{a}, \boldsymbol{b})^{\frac{1}{p}}$ is a distance between $\boldsymbol{a}, \boldsymbol{b} \in \Delta_n$, i.e.*

1. *(Symmetric): $W_p(\boldsymbol{a}, \boldsymbol{b}) = W_p(\boldsymbol{b}, \boldsymbol{a})$*
2. *(Definiteness): $W_p(\boldsymbol{a}, \boldsymbol{b}) = 0$ iff $\boldsymbol{a} = \boldsymbol{b}$*
3. *(Triangular inequality): $W_p(\boldsymbol{a}, \boldsymbol{b}) \leq W_p(\boldsymbol{a}, \boldsymbol{c}) + W_p(\boldsymbol{c}, \boldsymbol{b})$, for all $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \Delta_n$.*

Proof: It is easy to show the symmetricity since the distance is symmetric thus the objective function is symmetric. The definiteness is shown by observing that $C_{i,i} = 0$, so the objective function $\sum_{i,j} C_{i,j}P_{i,j} = \sum_{i>j} C_{i,j}(P_{i,j} + P_{j,i}) > 0$, since $\sum_j P_{i,j} = a_i > 0$. If $\boldsymbol{a} = \boldsymbol{b}$, we choose $\boldsymbol{P} = \operatorname{diag}(\boldsymbol{a}) \in U(\boldsymbol{a}, \boldsymbol{a})$. We can show that $\operatorname{diag}(\boldsymbol{a}) = \boldsymbol{P}^*$ since $\langle \operatorname{diag}(\boldsymbol{a})\ , \boldsymbol{C}\rangle = \sum_{i>j} C_{i,j}(P_{i,j} + P_{j,i}) = 0$, when $P_{i,j} = P_{j,i} = 0$. Also $\langle \boldsymbol{P}\ , \boldsymbol{C}\rangle \geq 0$ for all $\boldsymbol{a}$ $\boldsymbol{b}$ due to non-negativity of both $\boldsymbol{C}$ and $\boldsymbol{P}$. Therefore, $\langle \operatorname{diag}(\boldsymbol{a})\ , \boldsymbol{C}\rangle = W_p(\boldsymbol{a}, \boldsymbol{a}) = 0$.

To proof the triangular inequality, let $\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{c})$ and $\boldsymbol{Q} \in U(\boldsymbol{c}, \boldsymbol{b})$ be the optimal solution for $W_p(\boldsymbol{a}, \boldsymbol{c})$ and $W_p(\boldsymbol{c}, \boldsymbol{b})$. To avoid issues that may arise from null coordinates in $\boldsymbol{c}$, we define a vector $\tilde{\boldsymbol{c}}$ such that $\tilde{c}_j := c_j$ if $c_j > 0$, and $\tilde{c}_j := 1$ otherwise, to write

$$\boldsymbol{S} := \boldsymbol{P}\operatorname{diag}(1/\tilde{\boldsymbol{c}})\,\boldsymbol{Q} \in \mathbb{R}^{n\times n}_+,$$

We can show that $\boldsymbol{S} \in U(\boldsymbol{a}, \boldsymbol{b})$.

$$\begin{aligned}
\boldsymbol{S}\mathbf{1}_n &= \boldsymbol{P}\operatorname{diag}(1/\tilde{\boldsymbol{c}})\,\boldsymbol{Q}\mathbf{1}_n \\
&= \boldsymbol{P}\operatorname{diag}(1/\tilde{\boldsymbol{c}})\,\boldsymbol{c} \qquad\qquad \text{since } \boldsymbol{Q} \in U(\boldsymbol{c}, \boldsymbol{b}) \\
&= \boldsymbol{P}\mathbf{1}_n \qquad\qquad \text{since each element is } c_i/c_i = 1 \\
&= \boldsymbol{a} \qquad\qquad \text{since } \boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{c})
\end{aligned}$$

Similarly, $\boldsymbol{S}^T\mathbf{1}_n = \boldsymbol{b}$.

Therefore,

$$
\begin{aligned}
W_p(\boldsymbol{a},\boldsymbol{b}) &= L_{\boldsymbol{D}^p}(\boldsymbol{a},\boldsymbol{b})^{\frac{1}{p}} \\
&\leq \langle \boldsymbol{S}\,,\,\boldsymbol{D}^p\rangle^{\frac{1}{p}} \quad (\text{ optimality of LP}) \\
&= \left(\sum_{i,j} D^p_{i,j}\sum_k \frac{P_{i,k}Q_{k,j}}{c_k}\right)^{\frac{1}{p}} = \left(\sum_{i,j,k} D^p_{i,j}\frac{P_{i,k}Q_{k,j}}{c_k}\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i,j,k} (D_{i,k}+D_{k,j})^p\,\frac{P_{i,k}Q_{k,j}}{c_k}\right)^{\frac{1}{p}} \quad \text{(by triangular inequality of } D) \\
&\leq \left(\sum_{i,j,k} D^p_{i,k}\frac{P_{i,k}Q_{k,j}}{c_k}\right)^{\frac{1}{p}} + \left(\sum_{i,j,k} D^p_{k,j}\frac{P_{i,k}Q_{k,j}}{c_k}\right)^{\frac{1}{p}} \quad \text{(by Minkowski's inequality)} \\
&= \left(\sum_{i,k} D^p_{i,k}P_{i,k}\sum_j \frac{Q_{k,j}}{c_k}\right)^{\frac{1}{p}} + \left(\sum_{j,k} D^p_{k,j}Q_{k,j}\sum_i \frac{P_{i,k}}{c_k}\right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i,k} D^p_{i,k}P_{i,k}\frac{c_k}{c_k}\right)^{\frac{1}{p}} + \left(\sum_{j,k} D^p_{k,j}Q_{k,j}\frac{c_k}{c_k}\right)^{\frac{1}{p}} \quad \text{since } \boldsymbol{Q}\in U(\boldsymbol{c},\boldsymbol{b}) \text{ and } \boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{c}) \\
&= \left(\sum_{i,k} D^p_{i,k}P_{i,k}\right)^{\frac{1}{p}} + \left(\sum_{j,k} D^p_{k,j}Q_{k,j}\right)^{\frac{1}{p}} \\
&= W_p(\boldsymbol{a},\boldsymbol{c}) + W_p(\boldsymbol{c},\boldsymbol{b}). \quad \text{since } \boldsymbol{Q},\boldsymbol{P} \text{ are optimal solutions.}
\end{aligned}
$$

which concludes the proof. ▮

Note that for $1 > p \geq 0$, $\boldsymbol{D}^p$ is a distance. It implies that while for $p \geq 1$, $\mathcal{W}_p(\alpha,\beta)$ is a distance, in the case $0 < p \leq 1$, it is actually $\mathcal{W}_p(\alpha,\beta)^p$ which defines a distance on the simplex.

We can generalize the definition of **Wasserstein distance to aribitrary measure** $\mathcal{X} = \mathcal{Y}$ that defines a distance d.

Proposition 3.2 *We suppose $\mathcal{X} = \mathcal{Y}$ and that for some $p \geq 1$, $c(x,y) = d(x,y)^p$, where d is a distance on $\mathcal{X}$, i.e.*

1. *$d(x,y) = d(y,x)$ for all $x,y \in \mathcal{X}$;*
2. *$d(x,y) = 0$ iff $x = y$;*
3. *(Triangular inequality): $d(x,y) \leq d(x,z) + d(z,y)$, for all $x,y,z \in \mathcal{X}$*

*Then the **p-Wasserstein distance** on $\mathcal{X}$, $\mathcal{W}_p(\alpha,\beta) := \mathcal{L}_{d^p}(\alpha,\beta)^{\frac{1}{p}}$ is a valid distance between $\alpha,\beta \in \mathcal{M}^1_+(\mathcal{X})$, i.e.*

1. *(Symmetric): $\mathcal{W}_p(\alpha,\beta) = \mathcal{W}_p(\beta,\alpha)$ for all $\alpha,\beta \in \mathcal{M}(\mathcal{X})$*
2. *(Definiteness): $\mathcal{W}_p(\alpha,\beta) = 0$ iff $\alpha = \beta$*
3. *(Triangular inequality): $\mathcal{W}_p(\alpha,\beta) \leq \mathcal{W}_p(\alpha,\gamma) + \mathcal{W}_p(\gamma,\beta)$, for all $\alpha,\beta,\gamma \in \mathcal{M}^1_+(\mathcal{X})$.*

3.2 Basic properties of Wasserstein distance

Note that

- Wasserstein distance $\mathcal{W}_p(\alpha, \beta)$ is a distance between two Radon measures α, β on $\mathcal{X}$.
- The general optimal transport problem in (16) and (11) does not require that $c = d$ is distance matrix. That is the **Wasserstein distance** is the optimal value of a **special case** in **optimal transport problem**.
- Wasserstein distance, or Optimal Transport (OT), $\mathcal{W}_p(\alpha, \beta)$ depends on the *distance definition* d on the ***base*** **measure** $\mathcal{X}$. In other word, OT can be seen as automatically "**lifting**" a *ground metric* in $\mathcal{X}$ to a *metric* between ***measures*** on $\mathcal{X}$
- One of most **important** propery of Wasserstein distance is that it is a ***weak distance***, i.e. it allows one to compare singular distributions (for instance, discrete ones) whose **supports *do not overlap*** and to quantify the spatial shift between the supports of two distributions.

  In fact, $\mathcal{W}_p$ is a way to quantify the ***weak convergence*** or *convergence in distribution (in law)* [Villani, 2009]:

  Definition (**Weak Convergence or Convergence in distribution**) On a compact domain $\mathcal{X}$, $(\alpha_k)_k$ converges **weakly** to α in $\mathcal{M}_+^1(\mathcal{X})$ (denoted $\alpha_n \overset{d}{\to} \alpha$) if and only if *for any* ***continuous*** *function* $g \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} g d\alpha_k \to \int_{\mathcal{X}} d\alpha$. One needs to add additional decay conditions on g on noncompact domains.

  This notion of weak convergence corresponds to the **convergence in the distribution** of random vectors. Note the any random variable X_n is a continous function on Ω, and its distribution is the push-forward measure $\alpha_n = X_{n\#}\mathbb{P}$. Therefore, $\alpha_n \overset{d}{\to} \alpha$ is equivalent to $X_n \overset{d}{\to} X$.

  This convergence can be shown to be equivalent to $\mathcal{W}_p(\alpha_n, \alpha) \to 0$ [Villani, 2009, Theorem 6.8]. Thus we can also write the weak convergance as $\alpha_n \overset{\mathcal{W}_p}{\to} \alpha$.
- (**Translation**): A nice feature of the Wasserstein distance over a Euclidean space $\mathcal{X} = \mathbb{R}^d$ for the ℓ^2-distance ground cost $c(x, y) = \|x - y\|_2^2$ is that one can ***factor out*** *translations.* Define the translation operator $T_\tau : x \to x - \tau$. Then

$$\mathcal{W}_p(T_{\tau_1\#}\alpha,\, T_{\tau_2\#}\beta)^2 = \mathcal{W}_p(\alpha, \beta)^2 - 2\,\langle \tau_1 - \tau_2\,,\, \boldsymbol{m}_\alpha - \boldsymbol{m}_\beta \rangle + \|\tau_1 - \tau_2\|_2^2 \tag{38}$$

  where $\boldsymbol{m}_\alpha := \int_{\mathcal{X}} x d\alpha(x) \in \mathbb{R}^d$ is the barycenter of the measure α (in prob. the mean of random variable X). This interpretation is intuitive, since the translation operator moves the barycenter of the mass, thus the distance need to take into account of the difference between two barycenters.

  In particular, this implies the nice **decomposition** of the distance as

$$\mathcal{W}_p(\alpha, \beta)^2 = \mathcal{W}_p(\tilde{\alpha},\, \tilde{\beta})^2 + \|\boldsymbol{m}_\alpha - \boldsymbol{m}_\beta\|_2^2 \tag{39}$$

  where $\tilde{\alpha} = T_{\boldsymbol{m}_\alpha\#}\alpha$ and $\tilde{\beta} = T_{\boldsymbol{m}_\beta\#}\beta$ are the centered measures with zero mean.

3.3 Brenier theorem: Monge-Kantorovich equivalence

Brenier theorem shows that in $\mathbb{R}^d$ for $p = 2$, (i.e. $\mathcal{W}_2$) if at least one of the two input measures has a *density*, and for measures with *second order moments*, then the ***Kantorovich and Monge problems are equivalent***. That is the non-convex optimal assignment problem in (6) and the convex optimal transport problem in (16) have the same optimal solution and the Kantorovich relaxation is **tight**. Note that we have already shown in section 1.3.2 that in discrete case, these two problems are equivalent when the permutation matrix is the optimal coupling matrix. Brenier's theorem is for arbitrary measures.

Theorem 3.3 ***(Brenier)***
In the case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|_2^2$, if at least one of the two input measures (denoted α) has a density ρ_α with respect to the Lebesgue measure, then the optimal π in the Kantorovich formulation (16) *is* ***unique*** *and is* ***supported*** *on the graph $(\boldsymbol{x}, T(\boldsymbol{x}))$ of a* ***"Monge map"*** *$T : \mathbb{R}^d \to \mathbb{R}^d$. This means that $\pi = (Id, T)_{\#}\alpha$, i.e.*

$$\forall\, h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x, y) d\pi(x, y) = \int_{\mathcal{X}} h(x, T(x)) d\alpha(x). \tag{40}$$

Furthermore, this map T is uniquely defined as the ***gradient*** *of a* ***convex function*** *ϕ, $T(x) = \nabla\phi(x)$, where ϕ is the* ***unique*** *(up to an additive constant)* ***convex*** *function such that $(\nabla\phi)_{\#}\alpha = \beta$. This convex function is related to the* ***dual potential*** *$\lambda(\cdot)$ solving* (21) *as $\phi(x) = \frac{1}{2}\|x\|_2^2 - \lambda(x)$, $Id(\cdot)$ is an identity mapping.*

Proof: We sketch the main ingredients of the proof; more proof can be found in [Santambrogio, 2015, Figalli and Glaudo, 2021].

We mark that the $c(x, y) = \|x - y\|_2^2$ has decomposition $c(x, y) = \|x\|_2^2 + \|y\|_2^2 - 2\langle x\,, y\rangle$. Therefore, $\int c(x, y)d\pi(x, y) = C_{\alpha,\beta} - 2\int \langle x\,, y\rangle\, d\pi(x, y)$ where the constant is $C_{\alpha,\beta} = \int \|x\|_2^2\, d\alpha(x) + \int \|y\|_2^2\, d\beta(y)$. Instead of solving the primal problem (16), one solves

$$\begin{aligned} \max_{\pi\in\mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})} & \int_{\mathcal{X}\times\mathcal{Y}} \langle x\,, y\rangle\, d\pi(x, y) \\ \text{s.t. } & P_{\mathcal{X}\#}\pi = \alpha, \\ & P_{\mathcal{Y}\#}\pi = \beta \end{aligned}$$

This problem has a dual problem

$$\begin{aligned} \min_{(\phi,\psi)\in\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} & \int_{\mathcal{X}} \phi(x)d\alpha(x) + \int_{\mathcal{Y}} \psi(y)d\beta(y) \\ \text{s.t. } & \phi(x) + \psi(y) \geq \langle x\,, y\rangle\,, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, \end{aligned}$$

The relation between (ϕ, ψ) and (λ, μ) is that $\phi(\cdot) = \frac{1}{2}\|\cdot\|_2^2 - \lambda(\cdot)$ and $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2 - \mu(\cdot)$, respectively. One can replace the constraint above by

$$\psi(y) \geq \phi^*(y) := \sup_x \langle x\,, y\rangle - \phi(x) \tag{41}$$

ϕ^* is the **Legendre transform** (or, Fenchel's ***conjugate function***) of ϕ and is a **convex function** as a supremum of linear forms [Rockafellar, 1970]. Since the objective appearing in dual form above is linear and the integrating measures positive, one can ***minimize*** explicitly with respect to ψ and set $\psi = \phi^*$ in order to consider the unconstrained problem

$$\min_{\phi \in \mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} \phi(x) d\alpha(x) + \int_{\mathcal{Y}} \phi^*(y) d\beta(y) \tag{42}$$

since the objective function in (42) is the lower bound of the original dual objective above with $\psi = \phi^*$ as a feasible solution, maximizing this problem will be equivalent to maximizing the original dual objective. By iterating this argument twice, one can replace ϕ by ϕ^{**}, which is a **convex function**, and thus impose in (42) that $\underline{\phi \text{ is convex}}$. Therefore

$$\begin{aligned}\inf_{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} - \langle x\,,\, y\rangle \, d\pi(x,y) &\geq \sup_{\phi(x)+\psi(y) \geq \langle x\,,\, y\rangle} \int_{\mathcal{X}} -\phi(x) d\alpha(x) + \int_{\mathcal{Y}} -\psi(y) d\beta(y) \\ &\geq \sup_{\phi \text{ convex}} \int_{\mathcal{X}} -\phi(x) d\alpha(x) + \int_{\mathcal{Y}} -\phi^*(y) d\beta(y)\end{aligned}$$

Note that the support of π is bounded by the equality condition

$$\mathrm{Supp}(\pi) \subset \{(x,y) \in \mathcal{X} \times \mathcal{Y} : \lambda(x) + \phi^*(y) = \langle x\,,\, y\rangle\}\,, \tag{43}$$

which shows that such a y is optimal for the minimization (41) of the Legendre transform, whose optimality condition reads $y \in \partial\phi(x)$, the ***sub-differential*** of ϕ [Rockafellar, 1970].

Since ϕ is convex, it is *differentiable* 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 $y = \nabla\phi(x)$, and it shows that necessarily $\pi = (\mathrm{Id}, \nabla\phi)_{\#}\alpha$. ▮

Brenier's theorem, stating that an optimal transport map must be the **gradient of a convex function**, provides a useful generalization of the notion of increasing functions in dimension more than one. This is the main reason why optimal transport can be used to define **quantile functions** in *arbitrary dimensions*, which is in turn useful for applications to ***quantile regression*** *problems.*

3.3.1 Probabilistic interpretation for alternative dual problem

From the proof, we see that when $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|_2^2$ there exists an alternative dual problem (42). The probabilistic interpretation for (42) is

$$\min_{\phi \in \mathcal{C}(\mathcal{X})} \mathbb{E}_{X\sim\alpha}\left[\phi(X)\right] + \mathbb{E}_{Y\sim\beta}\left[\phi^*(Y)\right] \tag{44}$$

4 Examples of Wasserstein distance

4.1 Wasserstein distance between discrete measures

- (**discrete measures with** ℓ_1 as ground metric)
  For two discrete measures $\alpha = \sum_{i=1}^n a_i \delta_{\boldsymbol{x}_i}$ and $\beta := \sum_{i=1}^n b_i \delta_{\boldsymbol{y}_i}$,

  $$\mathcal{W}_1(\alpha, \beta) = \|\boldsymbol{a} - \boldsymbol{b}\|_1\,, \tag{45}$$

  i.e. $L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_1$ where $\boldsymbol{C} = \mathbf{1}\mathbf{1}^T - \boldsymbol{I}$.

- (**discrete measures with binary cost** as ground metric)
  For two discrete measures $\alpha \in \mathcal{M}_+(\mathcal{X})$ and $\beta \in \mathcal{M}_+(\mathcal{X})$, and $c(x, y)$ is 0 if $x = y$ and 1 when $x \neq y$,

  $$\mathcal{W}_1(\alpha, \beta) = \mathrm{TV}(\alpha, \beta)$$

  where $\mathrm{TV}(\alpha, \beta)$ is the total variation distance between two measures.

- (**discrete empirical measures with ℓ_p** as ground metric)
  Consider two empirical measures on $\mathcal{X} = \mathbb{R}$, $\alpha_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\beta_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$. 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

  $$\mathcal{W}_p(\alpha_n, \beta_n)^p = \frac{1}{n}\sum_{i}^{n} |x_i - y_i|^p = \|\boldsymbol{x} - \boldsymbol{y}\|_p^p := (d_p(\boldsymbol{x}, \boldsymbol{y}))^p \tag{46}$$

  which is the ℓ^p norm between two vectors of ordered values of α_n and β_n. That statement is valid only ***locally***, in the sense that the *order* (and those vector representations) *might change* whenever some of the values change.

4.2 p-Wasserstein distance on 1-dimensional real line $\mathbb{R}$

- (**Wasserstein distance on real line $\mathbb{R}$**)
  Define the ***cumulative distribution function*** $F_\alpha(t) := \int_{-\infty}^{t} d\alpha := \mathbb{P}(X \in (-\infty, t])$ for measure $\alpha \in \mathcal{M}(\mathbb{R})$ and corresponding random variable X. Then the pseudoinverse $F_\alpha^{-1} : [0, 1] \to \mathbb{R} \cup \{-\infty\}$

  $$F_\alpha^{-1}(r) := \min_{x} \{x \in \mathbb{R} \cup \{-\infty\} : \; F_\alpha(x) \geq r\} \tag{47}$$

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

  Then

  $$\mathcal{W}_p(\alpha, \beta)^p = \left\| F_\alpha^{-1} - F_\beta^{-1} \right\|_p^p := \int_0^1 \left\| F_\alpha^{-1}(r) - F_\beta^{-1}(r) \right\|_p^p dr \tag{48}$$

  where the norm $\left\| F_\alpha^{-1} - F_\beta^{-1} \right\|_p$ is L^p-norm on functionals.

  This means that through map $\alpha \to F_\alpha^{-1}$, 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 1-dimensional real line $\mathbb{R}$ is a *Hilbertian metric***.

4.3 1-Wasserstein distance $\mathcal{W}_1$

- $\mathcal{W}_1$ is also called the **earth movers distances (EMD)**.
- ($\mathcal{W}_1$ on $\mathbb{R}$ is a **norm** and is equal to **Total Variation distance**)

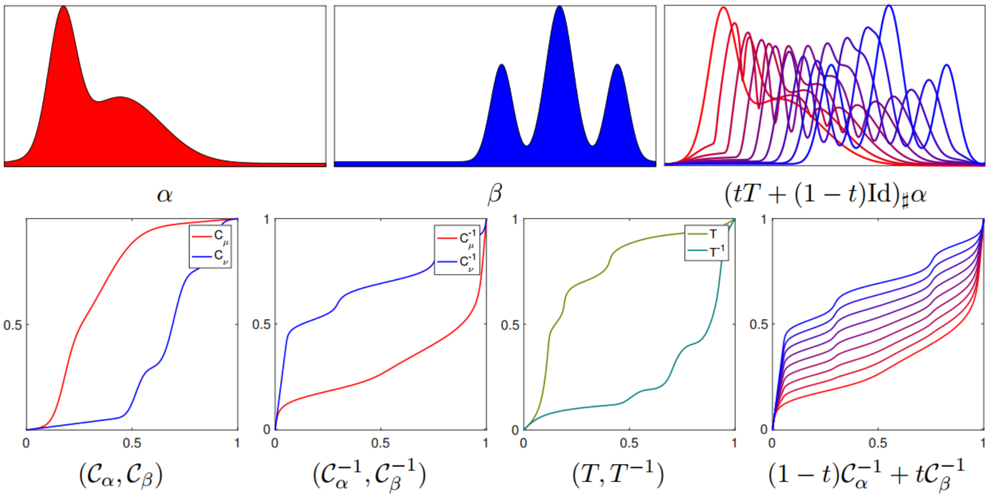


Figure 2.11: Computation of OT and displacement interpolation between two 1-D measures, using cumulant function as detailed in (2.39).

Figure 4: The interpolation between two 1-D measures, using cumulant function and quantile functions

Following (48), when $p = 1$, it further simplifies as

$$\mathcal{W}_1(\alpha, \beta) = \|F_\alpha - F_\beta\|_1 := \int_{-\infty}^{\infty} \|F_\alpha(x) - F_\beta(x)\|_1 \, dx \tag{49}$$

$$= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{x} d(\alpha - \beta) \right| \tag{50}$$

which shows that $\mathcal{W}_1$ on $\mathbb{R}$ is a **norm**. An optimal Monge map T such that $T_\# \alpha = \beta$ is then defined by

$$T = F_\beta^{-1} \circ F_\alpha \tag{51}$$

- (p-**Wasserstein distance is related to Lipschitz continous function f**)
  As a result of **Proposition** 2.2, The 1-Wasserstein distance can also be written as

$$\mathcal{W}_1(\alpha, \beta) := \inf_{\pi \in U(\alpha,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{X}} d_p(x, y) d\pi(x, y) \right\}$$

$$= \sup_{f \in \mathrm{Lip}_1} \left\{ \int_M f(x) d(\alpha - \beta)(x) \right\} \tag{52}$$

This results can be extended to $\mathcal{W}_p$

$$\mathcal{W}_p(\alpha, \beta) = \sup_{f \in \mathrm{Lip}_p} \left\{ \int_M f(x) d(\alpha - \beta)(x) \right\} \tag{53}$$

where

$$\mathrm{Lip}_p = \left\{ f : \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)^p} \leq 1 \right\}$$

This is equivalent to stating that $\mathcal{W}_p$ is the dual of p-Hölder functions.

If $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, the global Lipschitz constraint appearing in (52) can be made local as a uniform bound on the ***gradient** of* f,

$$\begin{aligned}\mathcal{W}_1(\alpha, \beta) &= \sup_{\|\nabla f\|_\infty \le 1} \left\{ \int_M f(x) d(\alpha - \beta)(x) \right\} && (54)\\ &= \inf_s \left\{ \int_{\mathbb{R}^d} \|s(x)\|_2 \, dx : \mathrm{div}(s) = \alpha - \beta \right\} && (55)\end{aligned}$$

The latter is an optimization problem under fixed ***divergence*** constraint, which is often called the ***Beckmann formulation***. Here the vectorial function $s(\boldsymbol{x}) \in \mathbb{R}^2$ can be interpreted as a flow field, describing locally the movement of mass. Outside the support of the two input measures, $\mathrm{div}(s) = 0$, which is the conservation of mass constraint. Once properly discretized using finite elements, Problems (54) and (55) become nonsmooth convex optimization problems.

- (**Probabilistic interpretation of dual form** of $\mathcal{W}_1$)
  The following theorm is a special caes of general dual representation (36) and **Proposition** 2.2

  Theorem 4.1 ***(Kantorovich-Rubenstein duality)***
  When the probability space Ω is a metric space, and α and β are probability measures on Ω, then for any fixed $K > 0$,

$$\mathcal{W}_1(\alpha, \beta) = \frac{1}{K} \sup_{\|f\|_L \le K} \left\{ \mathbb{E}_{X \sim \alpha}\left[f(X)\right] - \mathbb{E}_{Y \sim \beta}\left[f(Y)\right] \right\} \quad (56)$$

  *where $f : \mathcal{X} \to \mathbb{R}$ is **Lipschitz continous** function with **Lipschitz norm** $\|f\|_L \le K$.*

  Note here the we use $\mu(y) = \inf_x d(x, y) - \lambda(x)$ and since $\|\lambda\|_L \le 1$, this implies $g(y) = -f(y)$. This theorem in the basis for Wasserstein GAN algorithm using 1-Wasserstein distance $\mathcal{W}_1$ [Arjovsky et al., 2017].

- (**Concentration inequality of p-Wasserstein distance**)

  Definition [Wainwright, 2019] For a given metric d, the *probability measure* $\mathbb{P}$ is said to satisfy a d-**transportation cost inequality** with parameter $\gamma > 0$ if

$$\mathcal{W}_p(\mathbb{Q}, \mathbb{P}) \le \sqrt{2\gamma \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)}, \quad (57)$$

  for all *probability measures* $\mathbb{Q}$.

  Definition [Wainwright, 2019] The concentration function $\alpha : [0, \infty) \to \mathbb{R}_+$ associated with metric measure space $(\mathbb{P}, \mathcal{X}, d)$ is given by

$$\alpha_{\mathbb{P},(\mathcal{X},d)}(\epsilon) := \sup_{A \subseteq \mathcal{X}} \left\{ 1 - \mathbb{P}\{x \in \mathcal{X} : d(x, A) < \epsilon\} \middle| \mathbb{P}\{A\} \ge \frac{1}{2} \right\}, \quad (58)$$

  where the supremum is taken over all measurable subsets $A \subseteq \mathcal{X}$.

Theorem 4.2 *(From transportation cost to concentration [Wainwright, 2019]) Consider a metric measure space $(\mathbb{P}, \mathcal{X}, d)$, and suppose that $\mathbb{P}$ satisfies the d-transportation cost inequality* (57)*. Then its concentration function satisfies the bound*

$$\alpha_{\mathbb{P},(\mathcal{X},d)}(t) \leq 2 \exp\left(-\frac{t^2}{2\gamma}\right). \tag{59}$$

Moreover, for any $X \sim \mathbb{P}$ and any L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the concentration inequality

$$\mathbb{P}\left\{\left|f(X) - \mathbb{E}\left[f(X)\right]\right| \geq t\right\} \leq 2 \exp\left(-\frac{t^2}{2\gamma L^2}\right) \tag{60}$$

4.4 Wasserstein distance between Gaussian measures

- Given $\alpha = \mathcal{N}(\boldsymbol{m}_\alpha, \boldsymbol{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\boldsymbol{m}_\beta, \boldsymbol{\Sigma}_\beta)$ are two Guassian measures on $\mathbb{R}^d$, the push-forward operator T can be defined as a linear map

  $$\begin{aligned} T(\boldsymbol{x}) &= \boldsymbol{m}_\beta + \boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right) \qquad (61)\\ \text{where } \boldsymbol{A} &:= \boldsymbol{\Sigma}_\alpha^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\boldsymbol{\Sigma}_\beta\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{\Sigma}_\alpha^{-\frac{1}{2}} \end{aligned}$$

  where $\boldsymbol{A}$ is symmetric and positive definite. We can show that $T_{\#}\alpha = \beta$ and T is the optimal transport map.

  Proof: Note that the density after push-forward

  $$\begin{aligned} \rho_\beta(T(\boldsymbol{x})) &= \det\left(2\pi\left|\boldsymbol{\Sigma}_\beta\right|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(T(\boldsymbol{x}) - \boldsymbol{m}_\beta\right)^T \boldsymbol{\Sigma}_\beta^{-1}\left(T(\boldsymbol{x}) - \boldsymbol{m}_\beta\right)\right\}\\ &= \det\left(2\pi\left|\boldsymbol{\Sigma}_\beta\right|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)\right)^T \boldsymbol{\Sigma}_\beta^{-1}\left(\boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)\right)\right\}\\ &= \det\left(2\pi\left|\boldsymbol{\Sigma}_\beta\right|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)^T \left(\boldsymbol{A}^T\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{A}\right)\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)\right\}\\ &= \det\left(2\pi\left|\boldsymbol{\Sigma}_\beta\right|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)^T \tilde{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)\right\}\\ \text{where } \tilde{\boldsymbol{\Sigma}}^{-1} &= \left(\boldsymbol{\Sigma}_\alpha^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\boldsymbol{\Sigma}_\beta\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}}\left(\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\boldsymbol{\Sigma}_\beta\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\right)^{-1}\left(\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\boldsymbol{\Sigma}_\beta\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{\Sigma}_\alpha^{-\frac{1}{2}}\right)\\ &= \boldsymbol{\Sigma}_\alpha^{-1}\\ \rho_\beta(T(\boldsymbol{x})) &= \det\left(2\pi\left|\boldsymbol{\Sigma}_\beta\right|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)^T \boldsymbol{\Sigma}_\alpha^{-1}\left(\boldsymbol{x} - \boldsymbol{m}_\alpha\right)\right\} \end{aligned}$$

  and since T is a linear map, the Jacobian

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

  which means $\rho_\alpha = \left|\det T'(x)\right| \rho_\beta(T(\boldsymbol{x}))$, therefore $T_{\#}\alpha = \beta$.

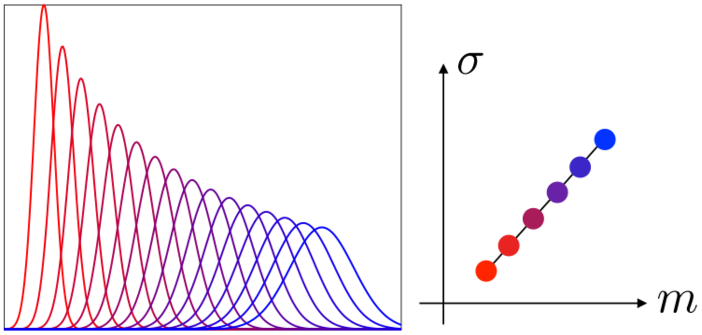


Figure 2.14: Computation of displacement interpolation between two 1-D Gaussians. Denoting $\mathcal{G}_{m,\sigma}(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-m)^2}{2s^2}}$ the Gaussian density, it thus shows the interpolation $\mathcal{G}_{(1-t)m_0+tm_1,(1-t)\sigma_0+t\sigma_1}$.

Figure 5: Computation of displacement interpolation between two 1-D Gaussians.

Also note that T is the *gradient* of the *convex* (quadratic) function $\phi(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_\alpha)^T \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{m}_\alpha) + \langle \boldsymbol{m}_\beta \, , \, \boldsymbol{x} \rangle$, which is the potential function of the dual. By Brenier's theorem, that T is optimal.
∎

Therefore, we can find the **2-Wasserstein distance between two Gaussian measures** as

$$\begin{aligned} \mathcal{W}_2(\alpha, \beta)^2 &= \mathcal{W}_2(\mathcal{N}(\boldsymbol{m}_\alpha, \boldsymbol{\Sigma}_\alpha), \mathcal{N}(\boldsymbol{m}_\beta, \boldsymbol{\Sigma}_\beta))^2 \\ &= \|\boldsymbol{m}_\alpha - \boldsymbol{m}_\beta\|_2^2 + \mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2 \end{aligned} \tag{62}$$

Note that this uses the *translation interpolation* formula in (39). The second term is called ***Bures metric***. It is defined as

$$\mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2 = \operatorname{tr}\left(\boldsymbol{\Sigma}_\alpha + \boldsymbol{\Sigma}_\beta - 2\left(\boldsymbol{\Sigma}_\alpha^{\frac{1}{2}} \boldsymbol{\Sigma}_\beta \boldsymbol{\Sigma}_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \right). \tag{63}$$

Bures metric is the distance between two positive definite matrices. $\mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2$ is also **convex** with respect to both its arguments.

If both Gaussian measuers have independent coordinates $\boldsymbol{\Sigma}_\alpha = \operatorname{diag}(\lambda_{\alpha,1}, \ldots, \lambda_{\alpha,n})$ and $\boldsymbol{\Sigma}_\beta = \operatorname{diag}(\lambda_{\beta,1}, \ldots, \lambda_{\beta,n})$, then the Bures metric is the **Hellinger distance**

$$\mathcal{B}(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta) = \left\| \sqrt{\boldsymbol{\lambda}_\alpha} - \sqrt{\boldsymbol{\lambda}_\beta} \right\|_2. \tag{64}$$

For a detailed treatment of the Wasserstein geometry of Gaussian distributions, we refer to [Takatsu, 2011], and for additional considerations on the Bures metric the reader can consult the very recent references [Malago et al., 2018, Bhatia et al., 2019].

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* 2021.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Luigi Malago, Luigi Montrucchio, and Giovanni Pistone. Wasserstein riemannian geometry of positive definite matrices. *arXiv preprint arXiv:1801.09269*, 2018.

Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 55. Springer, 2015.

Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005–1026, 2011.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.