Notes on OT

Haosheng Zhou

June, 2024

Contents

The notes are based on the materials from *Introduction to Optimal Transport* by Matthew Thorpe, *Course Notes on Computational Optimal Transport* by Gabriel Peyré, and the optimal transport summer school organized by Matt Jacobs and Nicholas Garcia Trillos.

Fundamentals of OT

Optimal matching of Point Clouds

The easiest form of OT is the optimal matching problem. Consider n points in the source space $x_1, ..., x_n \in X$ and the target space $y_1, ..., y_n \in Y$ respectively. Given the cost matrix $C \in \mathbb{R}_+^{n\times n}$, whose entry C_{ij} denotes the cost of matching x_i with y_j. The objective is to look for a permutation $\sigma : [n] \to [n]$ that induces a bijective matching $x_i \to y_{\sigma(i)}$ between those $2n$ points. The permutation shall be optimal in the sense of solving

$$\min_{\sigma} \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}. \tag{1}$$

Intuitively, $x_1, ..., x_n$ can be understood as workers, $y_1, ..., y_n$ can be understood as tasks, and C_{ij} is the cost of letting worker i do task j. One always hopes to find the best work assignment such that the total cost is minimized. Notice that each worker can only take one task and each task can only be assigned to one worker. Obviously, the optimal matching exists but is not unique (e.g., $n = 2$).

An important case would be $X = Y = \mathbb{R}$ and $C_{i,j} = h(x_i - y_j)$ for strictly convex $h \geq 0$, e.g. the power of a norm. In this case, the optimal matching satisfies **monotonicity** condition:

$$\forall (i, j), (x_i - x_j)(y_{\sigma(i)} - y_{\sigma(j)}) \geq 0. \tag{2}$$

To see why it is the case, we prove by contradiction and assume that index pair (i, j) violates the monotonicity condition. Consider another permutation that switches the images of i, j under σ while preserving all other images:

$$\tilde{\sigma}(k) = \begin{cases} \sigma(k) & k \neq i,\ k \neq j \\ \sigma(j) & k = i \\ \sigma(i) & k = j \end{cases}. \tag{3}$$

The strict convexity of h implies

$$\frac{h(x_i - y_{\sigma(i)}) - h(x_i - y_{\sigma(j)})}{y_{\sigma(j)} - y_{\sigma(i)}} > \frac{h(x_j - y_{\sigma(i)}) - h(x_j - y_{\sigma(j)})}{y_{\sigma(j)} - y_{\sigma(i)}}, \tag{4}$$

which directly implies

$$\sum_{k=1}^{n} C_{k,\tilde{\sigma}(k)} \leq \sum_{k=1}^{n} C_{k,\sigma(k)}, \tag{5}$$

meaning that $\tilde{\sigma}$ is a matching with lower cost. Intuitively, the penalty induced by a strictly convex h increases very fast as two points gets farther away, resulting in the monotone behavior of the optimal matching. One might notice that the monotonicity condition is so strong that it directly tells us what the optimal matching looks like. Consider the sorting permutation σ_Y such that $y_{\sigma(1)} \leq ... \leq y_{\sigma(n)}$ and the similar sorting permutation σ_X for $x_1, ..., x_n$. The optimal matching is given by

$$\sigma = \sigma_Y \circ \sigma_X^{-1}. \tag{6}$$

The optimal matching problem for cost matrices induced by strictly convex h reduces to sorting.

Remark. *When h is concave, e.g., $h(x, y) = -|x - y|^2$, consider points $1, 3, 5 \in X$ and $2, 4, 6 \in Y$, the optimal matching sends 1 to 6, 3 to 4 and 5 to 2. The optimal matching encourages a behavior that is totally different from convex h, and it's not a direct generalization of what we talked above.*

For general cost matrix C without special structures, the solution is given by the Hungarian algorithm, whose construction is based on the Kantorovich potentials.

Monge Problem

The optimal matching problem is a special case of OT in the sense that $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ are both empirical measures with the same number of probability masses. In general, the optimal transport problem is provided in the Monge formulation, hoping to transport a measure μ to another measure ν.

Naturally, we first have to define what it means to 'transport' between measures. For a mapping $T : X \to Y$ telling us how each point in X is mapped to a point in Y, we are able to lift it as $T_\# : \mathscr{P}(X) \to \mathscr{P}(Y)$ mapping a measure on X to a measure on Y. $T_\#$ is called the **pushforward** of T, and $\nu = T_\#\mu$ iff

$$\forall B \subset Y \text{ measurable}, \nu(B) = \mu(T^{-1}B). \tag{7}$$

Equivalently,

$$\forall h \in L^1(\nu), \int h(y)\, d\nu(y) = \int h(T(x))\, d\mu(x). \tag{8}$$

$T_\#$ linearizes any map T at the cost of moving from the original space to the space of measures on the original space.

Remark. *Pushforwards often appear in probability theory. Consider random variable $R : \Omega \to \mathbb{R}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The pushforward $R_\#$ has action $\mathbb{Q} = R_\#\mathbb{P}$, where $\mathbb{Q} \in \mathscr{P}(\mathbb{R})$ such that*

$$\forall B \text{ Borel}, \mathbb{Q}(B) = \mathbb{P}(R \in B). \tag{9}$$

It's clear that $\mathbb{Q} = \mathscr{L}(R)$ is the law of R.

Another example illustrates that pushforward is actually just the change of variables. Consider random variable $R \sim \mu, S \sim \nu$, $T_\#\mu = \nu$ iff

$$\forall h \in L^1(\nu), \mathbb{E}h(S) = \mathbb{E}h(T(R)), \tag{10}$$

which implies $S \overset{d}{=} T(R)$.

The Monge problem for given measures μ, ν and given cost function $c : X \times Y \to \mathbb{R}_+$ is given by:

$$\inf_T \int c(x, T(x))\, d\mu(x) \tag{11}$$

$$s.t.\ T_\#\mu = \nu. \tag{12}$$

We are finding a **transport map** T that transports μ to ν. The map is optimal in the sense that the cost of transportation along T is minimized. When μ, ν are both empirical measures with the same number of probability masses, we recover the optimal matching problem.

Remark. *In the language of probability, consider $R \sim \mu, S \sim \nu$, we want to find T subject to $S \overset{d}{=} T(R)$ that minimizes $\mathbb{E}c(R, S)$.*

The Monge formulation of optimal transport is problematic since such T might not exist, e.g., $\mu = \delta_0, \nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Even if such T exists, the infimum might be unattainable. To gain insights into the Monge problem, we consider two important cases: μ, ν are both discrete measures or μ, ν are both absolute continuous measures w.r.t. the Lebesgue measure (density exists) in one dimension.

Firstly, in the discrete case

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \nu = \sum_{j=1}^{m} b_j \delta_{y_j}. \tag{13}$$

Whenever $T_{\#}\mu = \nu$,

$$\forall j \in [m], b_j = \sum_{i:T(x_i)=y_i} a_i. \tag{14}$$

Such T must be surjective and the masses of μ can merge while the splitting of masses is prohibited, i.e., multiple a_i can be transported to a single b_j but one a_i cannot be transported to multiple b_j. Obviously, when $m > n$ (number of target mass larger than number of source mass), such T does not exist. When $m \leq n$, the compatibility condition above still does not necessarily hold, e.g., μ has masses $\frac{1}{4}, \frac{3}{4}$ while ν has masses $\frac{1}{2}, \frac{1}{2}$. A special case is when $m = n$ and both measures are empirical measures, in which case we have an optimal matching problem.

The case where μ, ν are both absolute continuous measures on $\mathbb{R}$ admits the existence of the optimal transport map. The identification of the optimal transport map requires the following Brenier's theorem (proved later).

Theorem 1 (Brenier)**.** *If $X = Y = \mathbb{R}^d, c(x, y) = |x-y|^2$, and μ is absolute continuous, there exists a unique optimal transport map T. The map is characterized as $T = \nabla\phi$ for convex ϕ such that $T_{\#}\mu = \nu$.*

Remark. *Brenier's theorem can be extended to cost $c(x, y) = h(x - y)$ where $h \in C^1$ is strictly convex, e.g., $c(x, y) = |x - y|^p$ for $p > 1$. The norms are by default Euclidean norm. The transport map being the gradient of a convex function implies the monotonicity of T, which aligns with the one in the optimal matching problem. However, such ϕ generally lacks regularity and its gradient is defined in the almost everywhere sense (Rademacher's theorem).*

Returning to the absolute continuous case, we first try to find transport maps between μ and $U(0, 1)$. Recall the inverse CDF method for sampling, we consider the quantile of μ defined as

$$Q_\mu(r) := \inf\{x : F_\mu(x) \geq r\}, \tag{15}$$

where F_μ is the CDF of μ. Clearly, for $U \sim U(0, 1)$, $Q_\mu(U) \sim \mu$, which implies the following lemma:

Lemma 1. *For any probability measure μ, $(Q_\mu)_{\#}[U(0, 1)] = \mu$.*

In particular, when μ is absolute continuous, F_μ is continuous and $(F_\mu)_{\#}\mu = U(0, 1)$. Clearly,

$$(Q_\nu \circ F_\mu)_{\#}\mu = (Q_\nu)_{\#}(F_\mu)_{\#}\mu = \nu, \tag{16}$$

indicating that $T = Q_\nu \circ F_\mu$ transports μ to ν. Since T is increasing, it must be the gradient of a convex function, and by Brenier's theorem, such T must be the unique optimal transport map. As a result, whenever $X = Y = \mathbb{R}$ and μ is absolute continuous, the optimal transport problem under cost $c(x, y) = |x - y|^2$ is solved. This example shows the power of Brenier's theorem.

At last, we point out that Brenier's theorem allows the derivation of the Monge-Ampere equation for optimal transport. Assume the cost function $c(x, y) = |x - y|^2$ (which will be the cost by default without specification), the optimal transport map admits the representation $T = \nabla\phi$. Assume μ, ν are both absolute continuous with density p_μ, p_ν, then $(\nabla\phi)_{\#}\mu = \nu$ is equivalent to saying

$$\forall h \in L^1(\nu), \int h(y)p_\nu(y)\,dy = \int h(\nabla\phi(x))p_\nu(\nabla\phi(x))\det(\nabla^2\phi(x))\,dx = \int h(\nabla\phi(x))p_\mu(x)\,dx, \tag{17}$$

which implies the **Monge-Ampere equation**

$$p_\nu(\nabla\phi(x))\det(\nabla^2\phi(x)) = p_\mu(x). \tag{18}$$

The solution ϕ characterizes the optimal transport map. Generally, Monge-Ampere equation has the form $\det(\nabla^2 u) = f(x, u, \nabla u)$ and the equation above belongs to this class.

Remark. *The term* $\det(\nabla^2\phi)$ *can be understood as the non-linear Laplacian. Consider the case* $X = Y$*, with a trivial transport map* $T = id$*, clearly* $T = \nabla\phi$*,* $\phi(x) = \frac{1}{2}|x|^2$*. We perturb* ϕ *by* $\varepsilon\psi$ *to get* $\tilde{\phi}(x) = \frac{1}{2}|x|^2 + \varepsilon\psi(x)$*, such that* $\nabla\tilde{\phi}(x) = x + \varepsilon\nabla\psi(x)$*. In this case,* $\det(\nabla^2\tilde{\phi}) = \det(I + \varepsilon\nabla^2\psi)$*. Using* $\det(I + \varepsilon A) = 1 + \varepsilon\mathrm{Tr}(A) + o(\varepsilon)$ $(\varepsilon \to 0)$*, we see that*

$$\det(\nabla^2\tilde{\phi}) = 1 + \varepsilon\Delta\psi + o(\varepsilon). \tag{19}$$

When ϕ *gets perturbed infinitesimally, the first order term in* $\det(\nabla^2\phi)$ *changes by the Laplacian of the perturbation.*

Example: OT for Gaussian

In the case where $X = Y = \mathbb{R}$ and $\mu = N(\mu_1, \sigma_1^2), \nu = N(\mu_2, \sigma_2^2)$, one directly considers the CDF $F_\mu(x) = \Phi(\frac{x-\mu_1}{\sigma_1}), F_\nu(x) = \Phi(\frac{x-\mu_2}{\sigma_2})$ and the optimal transport map is given by the increasing map:

$$T(x) = F_\nu^{-1} \circ F_\mu(x) = \frac{\sigma_2}{\sigma_1}(x - \mu_1) + \mu_2. \tag{20}$$

This is a linear map composed by translations and scalings, indicating that the best operation to take is to translate μ by μ_1 units (get $N(0, \sigma_1^2)$), scale the variance (get $N(0, \sigma_2^2)$), and translate back by μ_2 units (get $\nu = N(\mu_2, \sigma_2^2)$).

Generally, if $X = Y = \mathbb{R}^d$, OT becomes hard for general distributions but is easy for Gaussians. Assume $\mu = N(\mu_1, \Sigma_1), \nu = N(\mu_2, \Sigma_2)$. We start from the one-dimensional analogue of the optimal transport map:

$$T(x) = A(x - \mu_1) + \mu_2. \tag{21}$$

The translations are kept while the scaling part is replaced with $A \in \mathbb{R}^{d\times d}$ since it still remains unclear how we shall generalize $\frac{\sigma_2}{\sigma_1}$ in multi-dimensional cases. At this point, we shall think about matching the structure $T = \nabla\phi$ for convex ϕ in Brenier's theorem. Clearly, such ϕ has the form

$$\phi(x) = \frac{1}{2}(x - \mu_1)^T A(x - \mu_1) + \mu_2 x. \tag{22}$$

Notice that in order to make sure $T = \nabla\phi$, and ϕ is convex, A has to be a symmetric SPD matrix. Brenier's theorem enables us to put more restrictions on the matrix A.

The final step to determine A is to come back to the relationship $T_\# \mu = \nu$. The calculations can be carried out using the Gaussian characteristic function. Assume $R \sim \mu, S \sim \nu$, the condition is saying $T(R) \overset{d}{=} S$ for some linear transformation T. We use ϕ_R, ϕ_S to denote the characteristic functions respectively, then

$$\phi_S(t) = \phi_{T(R)}(t) = e^{it^T\mu_2}\mathbb{E}e^{it^TA(R-\mu_1)} = e^{it^T\mu_2 - it^TA\mu_1}\phi_R(A^Tt) \tag{23}$$

$$= e^{it^T\mu_2 - it^TA\mu_1}e^{it^TA\mu_1 - \frac{1}{2}t^TA\Sigma_1At} = e^{it^T\mu_2 - \frac{1}{2}t^TA\Sigma_1At}. \tag{24}$$

Comparing with $\phi_S(t) = e^{it^T\mu_2 - \frac{1}{2}t^T\Sigma_2t}$ yields the Riccati equation

$$A\Sigma_1A = \Sigma_2. \tag{25}$$

This Ricatti equation can be solved easily by writing the LHS as a square of a matrix. For SPD matrix A, we use $A^{\frac{1}{2}}$ to denote its unique square root matrix (still SPD and symmetric). As a result, $\Sigma_1^{\frac{1}{2}}A\Sigma_1^{\frac{1}{2}}\Sigma_1^{\frac{1}{2}}A\Sigma_1^{\frac{1}{2}} = \Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}$, which implies $\Sigma_1^{\frac{1}{2}}A\Sigma_1^{\frac{1}{2}} = (\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}$, so

$$A = \Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}. \tag{26}$$

provides the explicit formula of the optimal transport map.

Kantorovich Relaxation

The Monge formulation of OT problem is problematic since it does not allow the splitting of probability masses, e.g., any pushforward of a Dirac point mass must still be a Dirac point mass. The Kantorovich relaxation relaxes the Monge problem by allowing masses to split freely as long as the marginals match the source and target measures. The problem is now given by

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\, d\pi(x,y), \tag{27}$$

where $\Pi(\mu,\nu) = \{\pi \in \mathscr{P}(X\times Y) : \pi \text{ has two marginals } \mu,\nu\}$ is the set of couplings. In other words, by considering the joint distribution, we allow a 'probabilistic' transportation instead of the 'deterministic' transportation induced by T. Such π is called a coupling or a transport plan, which differs from the transport map.

Remark. *Consider $\mu = \delta_0, \nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, then a legal coupling can be given by π, which assigns probability mass $\frac{1}{2}$ to $(0,0)$ and probability mass $\frac{1}{2}$ to $(0,1)$. The existence of coupling is always guaranteed since the product measure $\pi = \mu\otimes\nu$ is a trivial coupling, so one at least does not need to worry about the constraint.*

The following theorem shows that the infimum in the Kantorovich relaxation is always attainable given that the spaces X, Y and the cost c are not too wild.

Theorem 2. *If X, Y are Polish spaces (by default) and $c : X\times Y \to \mathbb{R}_+$ is l.s.c. (lower semi-continuous), then there exists an optimal coupling $\pi \in \Pi(\mu,\nu)$ that attains the infimum.*

Proof. By the inner regularity of Radon measures, $\forall\delta > 0$, there always exists compact $K \subset X, L \subset Y$ such that for $\forall\pi \in \Pi(\mu,\nu)$, $\pi(X\times Y - K\times L) \le 2\delta$. This proves the tightness of $\Pi(\mu,\nu)$ and by Prokhorov's theorem, it's sequentially compact under the weak* topology (under which the convergence is the weak convergence of measures/convergence in distribution).

Let there be a minimizing sequence π_n of the Kantorovich relaxation. The sequential compactness identifies the weak* limit π^*. Since $\Pi(\mu,\nu)$ is weak* closed (prove by definition), $\pi^* \in \Pi(\mu,\nu)$. Followed by l.s.c. and Portmanteau theorem,

$$\lim_{n\to\infty} \int c(x,y)\, d\pi_n(x,y) \ge \int c(x,y)\, d\pi^*(x,y) \tag{28}$$

concludes the proof. □

After arguing the well-posedness of the problem, we return to solving the Kantorovich relaxation. Surprisingly, the objective is a linear function in π and so does the constraint, so this problem is actually a **linear programming** problem on the space of measures. Let's consider the case of transporting two discrete measures, as is often the case in practical applications $\mu = \sum_{i=1}^n a_i\delta_{x_i}, \nu = \sum_{j=1}^m b_j\delta_{y_j}$. The coupling π now reduces to a matrix $P \in \mathbb{R}_+^{n\times m}$ with P_{ij} denoting the amount transporting from x_i to y_j. The conservation of masses requires

$$P\vec{1} = a, \quad P^T\vec{1} = b. \tag{29}$$

The optimization problem is rewritten as

$$\min_P \operatorname{Tr}(P^T C) \tag{30}$$

$$s.t.\ P \in \mathbb{R}_+^{n\times m}, P\vec{1} = a, P^T\vec{1} = b. \tag{31}$$

The LP (linear programming) structure is obvious.

Lastly, we have to make sure that whenever Monge problem has a solution, Kantorovich relaxation always provides exactly the same solution as Monge problem. Whenever $T_{\#}\mu = \nu$ in Monge problem, $\pi = (id, T)_{\#}\mu$ is always a coupling. By definition, $\forall h \in L^1(\pi), \int h(x,y)\, d\pi(x,y) = \int h(x, T(x))\, d\mu(x)$. If $h = h(x)$ has no dependence on y, $\int h(x)\, d(P_1)_{\#}\pi(x) = \int h(x, T(x))\, d\mu(x)$ proves $(P_1)_{\#}\pi = \mu$ (here $P_1(x,y) = x$ is the projection so $(P_1)_{\#}\pi$ is the first marginal of π). Similarly, if $h = h(y)$ has no dependence on x, $\int h(y)\, d(P_2)_{\#}\pi(y) = \int h(T(x))\, d\mu(x) = \int h(y)\, d\nu(y)$ since $T_{\#}\mu = \nu$. This proves $(P_2)_{\#}\pi = \nu$. As a result, $\pi = (id, T)_{\#}\mu \in \Pi(\mu, \nu)$.

We have argued that a transport map is a special case of the transport plan, but have not yet proved that an optimal transport plan π^* will degenerate to $(id, T^*)_{\#}\mu$, which is induced by the optimal transport map T if Monge problem admits a solution. For simplicity, we provide the proof only for μ, ν being empirical measures (with the same number of masses and equal splitting of masses). In other words, $\mu = \sum_{i=1}^n \delta_{x_i}, \nu = \sum_{j=1}^n \delta_{y_j}$. In this case, Monge problem degenerates to optimal matching, in which the existence of the optimal transport map is guaranteed. The Kantorovich relaxation is:

$$\min_P \operatorname{Tr}(P^T C) \tag{32}$$

$$s.t.\ P \in \mathscr{B}_n, \tag{33}$$

where $\mathscr{B}_n$ is the collection of all bistochastic matrices (non-negative entries with each row and column adds up to 1). On the other hand, any feasible transport map in the Monge problem must be induced by a permutation σ on $[n]$ so it can be written as a permutation matrix P such that $P_{ij} = 1$ iff $j = \sigma(i)$ and otherwise zero. Denote the collection of all permutation matrices as $\mathscr{P}_n$, so the Monge problem is actually

$$\min_P \operatorname{Tr}(P^T C) \tag{34}$$

$$s.t.\ P \in \mathscr{P}_n. \tag{35}$$

It remains to prove that those two optimization problems have the same optimizer. Clearly, $\mathscr{P}_n \subset \mathscr{B}_n$. It turns out that the connection between $\mathscr{P}_n$ and $\mathscr{B}_n$ is given by the following theorem.

Theorem 3 (Birkhoff, Von Neumann)**.** *Denote* $\operatorname{Extr}(\mathscr{C})$ *as the collection of extremal points of a convex set* $\mathscr{C}$*. The extremal points of* $\mathscr{C}$ *are the points in* $\mathscr{C}$ *that does not admit an nontrivial convex representation using other points in* $\mathscr{C}$*. Then*

$$\operatorname{Extr}(\mathscr{B}_n) = \mathscr{P}_n. \tag{36}$$

Proof. If $P \in \mathscr{B}_n - \mathscr{P}_n$, consider the bipartite graph induced by P (edge (i,j) exists iff $P_{ij} > 0$), which must have a cycle of the shortest length. Using the indices in the cycle, one can always construct two matrices in $\mathscr{B}_n$ distinct from P, whose convex representation provides P. □

The following lemma exploits the LP structure of OT problems.

Lemma 2. *If $\mathscr{C}$ is a compact convex set, then*

$$\mathrm{Extr}(\mathscr{C}) \cap \arg\min_{P \in \mathscr{C}} \mathrm{Tr}(P^T C) \neq \emptyset, \tag{37}$$

meaning that there exists a minimizer as an extremal point at the same time.

Proof. Let $S := \arg\min_{P \in \mathscr{C}} \mathrm{Tr}(P^T C)$ for given cost C. S is a compact convex set, by Krein-Milman theorem, $\mathrm{Extr}(S) \neq \emptyset$. It remains to prove $\mathrm{Extr}(S) \subset \mathrm{Extr}(\mathscr{C})$ (which does not necessarily hold for $S \subset \mathscr{C}$).

For $\forall P \in \mathrm{Extr}(S)$, for any $A, B \in \mathscr{C}$ such that $P = \theta A + (1-\theta)B$ for some $\theta \in [0,1]$, we have $\mathrm{Tr}(P^T C) \leq \mathrm{Tr}(A^T C), \mathrm{Tr}(P^T C) \leq \mathrm{Tr}(B^T C)$. The linearity of trace implies $\mathrm{Tr}(P^T C) = \mathrm{Tr}(A^T C) = \mathrm{Tr}(B^T C)$ so $A, B \in S$. Since $P \in \mathrm{Extr}(S)$, $A = B = P$. This proves that $P \in \mathrm{Extr}(\mathscr{C})$. □

Combining two conclusions by specifying $\mathscr{C} = \mathscr{B}_n$, we get

$$\mathscr{P}_n \cap \arg\min_{P \in \mathscr{B}_n} \mathrm{Tr}(P^T C) \neq \emptyset, \tag{38}$$

which proves that there exists a matrix $P \in \mathscr{P}_n$ (transport map) that also works as a minimizer of the Kantorovich relaxation. This proves that for empirical measures, Kantorovich and Monge formulations provide the same optimizer.

Remark. *Inspired by the proof of the Birkhoff-VNM theorem, since the bipartite graph induced by the optimal P shall be cycle-free, the optimal P has at most $n + m - 1$ non-zero entries for general discrete measures.*

Wasserstein Distance

The Kantorovich relaxation ensures the existence of the optimal coupling between any pair of measures. Naturally, the optimal transport cost measures the effort one has pay transporting one measure to another, which measures the difference between two measures. In general, one considers the case where $X = Y$ and takes $c(x, y) = [d(x, y)]^p$ for some distance d on X. The optimal transport cost under Kantorovich formulation is defined as the power of the p-Wasserstein distance.

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int [d(x, y)]^p \, d\pi(x, y) \right)^{\frac{1}{p}}, 1 \leq p < \infty. \tag{39}$$

We first prove that **p-Wasserstein distance** is actually a distance in the mathematical sense. If $\mu = \nu$, consider π as a measure supported on the diagonal $\Delta = \{(x, x) : x \in X\}$ with the first marginal as μ. It's obvious that

$$\forall h(x, y) = h(y), \int h(y) \, d(P_2)_\# \pi(y) = \int h(x, y) \, d\pi(x, y) = \int h(x, x) \, d\mu(x) = \int h(x) \, d\mu(x). \tag{40}$$

So $(P_2)_\# \pi = \mu = \nu$, $\pi \in \Pi(\mu, \nu)$. Of course $\int [d(x, y)]^p \, d\pi(x, y) = \int [d(x, x)]^p \, d\pi(x, y) = 0$.

Conversely, if $W_p(\mu, \nu) = 0$, then $\exists \pi^* \in \Pi(\mu, \nu)$, $\int [d(x, y)]^p \, d\pi^*(x, y) = 0$ (inf attainable). This implies π^* is supported on the diagonal Δ due to the positivity of d as a metric. As a result,

$$\forall h, \int h(y, y) \, d\nu(y) = \int h(y, y) \, d\pi^*(x, y) = \int h(x, y) \, d\pi^*(x, y) = \int h(x, x) \, d\pi^*(x, y) = \int h(x, x) \, d\mu(x) \tag{41}$$

proves $\mu = \nu$.

For any $\pi \in \Pi(\mu, \nu)$, consider the map that interchanges components $S(x, y) = (y, x)$. Let's check that $S_\# \pi \in \Pi(\nu, \mu)$. Clearly, $(P_1)_\# S_\# \pi = (P_2)_\# \pi = \nu$ and $(P_2)_\# S_\# \pi = (P_1)_\# \pi = \mu$. In addition, applying $S_\#$ does not change the transport cost.

$$\int [d(x, y)]^p \, dS_\# \pi(x, y) = \int [d(y, x)]^p \, d\pi(x, y) = \int [d(x, y)]^p \, d\pi(x, y) \tag{42}$$

It's clear that we have proved $W_p(\mu, \nu) = W_p(\nu, \mu)$.

The triangle inequality requires more efforts. The difficulty lies in connecting three measures μ, ν, η with two couplings, so we have to refer to the following gluing lemma.

Lemma 3. *Let $\mu \in \mathscr{P}(X), \nu \in \mathscr{P}(Y), \eta \in \mathscr{P}(Z)$ be three measures on the Polish spaces, given $\pi \in \Pi(\mu, \nu), \xi \in \Pi(\nu, \eta)$, there exists $\sigma \in \mathscr{P}(X \times Y \times Z)$ such that $(P_{1,2})_\# \sigma = \pi, (P_{2,3})_\# \sigma = \xi$.*

Proof. Using the disintegration theorem (regular conditional probability), there exists families of measures $\{\pi_y\}_{y \in Y} \subset \mathscr{P}(X)$ (conditional probability measure of π given $y \in Y$), $\{\xi_y\}_{y \in Y} \subset \mathscr{P}(Z)$ (conditional probability measure of ξ

given $y \in Y$). Those conditional measures are identified through

$$\int \left(\int h(x,y)\, d\pi_y(x) \right) d\nu(y) = \int h(x,y)\, d\pi(x,y), \tag{43}$$

$$\int \left(\int h(y,z)\, d\xi_y(z) \right) d\nu(y) = \int h(y,z)\, d\xi(y,z). \tag{44}$$

One can check that $\sigma(x,y,z) = \pi_y(x)\xi_y(z)\nu(y)$ provides the construction. □

Returning to prove the triangle inequality, given μ, ν, η, we identify optimal couplings $\pi \in \Pi(\mu,\nu), \xi \in \Pi(\nu,\eta)$ that attains the infimum in $W_p(\mu,\nu), W_p(\nu,\eta)$ respectively. By the gluing lemma, there exists σ such that $(P_{1,2})_\# \sigma = \pi, (P_{2,3})_\# \sigma = \xi$. Since σ has three marginals as μ, ν, η, we take out the marginals w.r.t. the first and third component $\zeta = (P_{1,3})_\# \sigma \in \Pi(\mu,\eta)$ as the coupling.

$$W_p(\mu,\eta) \le \left(\int [d(x,z)]^p \, d\zeta(x,z) \right)^{\frac{1}{p}} = \left(\int [d(x,z)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} \tag{45}$$

$$\le \left(\int [d(x,y) + d(y,z)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} \tag{46}$$

$$\le \left(\int [d(x,y)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} + \left(\int [d(y,z)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} \quad \text{(Minkowski)} \tag{47}$$

$$= \left(\int [d(x,y)]^p \, d\pi(x,y) \right)^{\frac{1}{p}} + \left(\int [d(y,z)]^p \, d\xi(y,z) \right)^{\frac{1}{p}} \tag{48}$$

$$= W_p(\mu,\nu) + W_p(\nu,\eta). \tag{49}$$

The p-Wasserstein distance between any two measures is always finite (unlike KL divergence), with a physical meaning of the transportation cost. This motivates the research of the Wasserstein geometry and relevant applications as a natural analogue to the finite-dimensional Euclidean spaces.

Example: Wasserstein Distance

When $X = Y = \mathbb{R}$, the optimal transport map is given by $T = Q_\nu \circ F_\mu$, which is also the optimal transport plan given cost function $c(x, y) = |x - y|^p$. Plugging into the definition of the p-Wasserstein distance to get

$$W_p(\mu, \nu) = \left(\int_0^1 |Q_\mu(x) - Q_\nu(x)|^p \, dx \right)^{\frac{1}{p}} = \|Q_\mu - Q_\nu\|_{L^p([0,1])}. \tag{50}$$

Simply speaking, on $\mathbb{R}$, through the mapping $\mu \mapsto Q_\mu$, the Wasserstein distance is isometric to the L^p distance. In particular, let's check what happens when $p = 1$.

$$W_1(\mu, \nu) = \int_0^1 |Q_\mu(x) - Q_\nu(x)| \, dx = \int_0^1 \int_{Q_\mu(x) \wedge Q_\nu(x)}^{Q_\mu(x) \vee Q_\nu(x)} dy \, dx \tag{51}$$

$$= \int_{\mathbb{R}} \int_{F_\mu(y) \wedge F_\nu(y)}^{F_\mu(y) \vee F_\nu(y)} dx \, dy = \int_{\mathbb{R}} |F_\mu(y) - F_\nu(y)| \, dy. \tag{52}$$

The 1-Wasserstein distance is just the area of the difference under the CDF curves.

For $X = Y = \mathbb{R}^d$ and Gaussian $\mu = N(\mu_1, \Sigma_1), \nu = N(\mu_2, \Sigma_2)$, we have also derived the optimal transport map $T(x) = A(x - \mu_1) + \mu_2$ where $A = \Sigma_1^{-\frac{1}{2}} (\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}}$. It turns out that when $p = 2$, the Wasserstein distance admits a simple representation:

$$W_2(\mu, \nu) = \sqrt{\int |x - T(x)|^2 \, p_\mu(x) \, dx} \tag{53}$$

$$= \sqrt{\mathbb{E}_{x \sim \mu} \|(I - A)x - (\mu_2 - A\mu_1)\|^2} \tag{54}$$

$$= \sqrt{|\mu_1 - \mu_2|^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}})^{\frac{1}{2}})} \tag{55}$$

$$=: \sqrt{|\mu_1 - \mu_2|^2 + D_B(\Sigma_1, \Sigma_2)^2}. \tag{56}$$

Within the calculations, we use the fact that $(I - A)x \sim N((I - A)\mu_1, (I - A)\Sigma_1(I - A))$ and that if $y \sim N(\mu_y, \Sigma_y)$, then $\mathbb{E}|y|^2 = \mathrm{Tr}(\Sigma_y) + |\mu_y|^2$. Due to the important structure of the 2-Wasserstein distance of Gaussians, we denote the trace term as the square of $D_B(\Sigma_1, \Sigma_2)$, which is the Bures metric, a distance on the space of symmetric SPD matrices, e.g., covariance matrices. In this sense, W_2^2 is the sum of the squared Euclidean distance between mean vectors and the squared Bures distance between covariance matrices. Hence, the 2-Wasserstein distance can be understood as an analogue of the Euclidean distance on the space of measures.

To dig a little deeper into the Bures metric (which also has backgrounds in quantum computing), we provide a proof showing that it's indeed a metric, which is untrivial at the first glance.

Lemma 4. *For Hermitian matrices A, B,*

$$D_B(A, B) = \min_{M \in F(A), N \in F(B)} \|M - N\|_F = \min_{U \in U(n)} \|A^{\frac{1}{2}} - B^{\frac{1}{2}} U\|_F, \tag{57}$$

where $F(A) := \{M \in \mathbb{C}^{n\times n} : A = MM^*\}$, *and* $U(n) := \{U \in \mathbb{C}^{n\times n} : UU^* = U^*U = I\}$.

Proof. By splitting two terms in the norm,

$$\min_{U\in U(n)} \|A^{\frac{1}{2}} - B^{\frac{1}{2}}U\|_F = \mathrm{Tr}(A) + \mathrm{Tr}(B) - \max_{U\in U(n)} \mathrm{Tr}(U^*B^{\frac{1}{2}}A^{\frac{1}{2}} + A^{\frac{1}{2}}B^{\frac{1}{2}}U). \tag{58}$$

Consider the polar decomposition $B^{\frac{1}{2}}A^{\frac{1}{2}} = VP$, where V is unitary and $P = (A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}} \geq 0$.

$$\begin{aligned}\mathrm{Tr}(U^*B^{\frac{1}{2}}A^{\frac{1}{2}} + A^{\frac{1}{2}}B^{\frac{1}{2}}U) &= \mathrm{Tr}(U^*VP + PV^*U) &(59)\\ &= \mathrm{Tr}[(U^*V + V^*U)P] &(60)\\ &= \sum_{j=1}^{n} 2\cos\theta_j P_{jj}. &(61)\end{aligned}$$

The last equation follows from selecting a basis that diagonalizes U^*V. Since it's unitary, the diagonal entries can be written as $e^{i\theta_1}, ..., e^{i\theta_n}$. Clearly, the maximum is attained when $\theta_1 = ... = \theta_n = 0$, i.e., $U = V$. It follows that

$$\max_{U\in U(n)} \mathrm{Tr}(U^*B^{\frac{1}{2}}A^{\frac{1}{2}} + A^{\frac{1}{2}}B^{\frac{1}{2}}U) = 2\mathrm{Tr}(P), \tag{62}$$

which concludes the proof of the last expression.

For the middle one, notice that $A = MM^* = NN^*$ iff M and N differs by a unitary matrix. Since $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$, any matrix in $F(A)$ differs from $A^{\frac{1}{2}}$ by a unitary matrix, which concludes the proof. □

From this lemma, it's obvious that Bures metric is positive and symmetric. For the triangle inequality,

$$\forall U, V \in U(n), D_B(A, C) \leq \|A^{\frac{1}{2}} - C^{\frac{1}{2}}U\|_F \leq \|A^{\frac{1}{2}} - B^{\frac{1}{2}}V\|_F + \|B^{\frac{1}{2}}V - C^{\frac{1}{2}}U\|_F. \tag{63}$$

Take minimum on both sides w.r.t. $U, V \in U(n)$ to conclude.

Remark. *Consider the trivial case where* Σ_1 *is diagonal with diagonal elements* $a_1, ..., a_d$ *and* Σ_2 *is diagonal with diagonal elements* $b_1, ..., b_d$. *Then the Bures metric equals*

$$D_B(\Sigma_1, \Sigma_2) = \sum_{i=1}^{d} (\sqrt{a_i} - \sqrt{b_i})^2, \tag{64}$$

which is the Hellinger square distance in information theory.

Wasserstein Topology

The topology induced by the Wasserstein distance is worth mentioning. Firstly, we prove the lemma indicating the relationship between p-Wasserstein distances when p varies.

Lemma 5. *For $1 \le p \le q$, $W_p(\mu,\nu) \le W_q(\mu,\nu) \le \mathrm{diam}(X)^{\frac{q-p}{q}} W_p^{\frac{p}{q}}(\mu,\nu)$, where $\mathrm{diam}(X) = \sup_{x,y\in X} d(x,y)$ is the diameter of the source space.*

Proof. W_q^q is essentially an expectation when the optimal coupling π^q is given. By Jensen's inequality applied for convex function $x \mapsto x^{\frac{q}{p}}$,

$$W_p^q(\mu,\nu) = \left(\int [d(x,y)]^p \, d\pi^q(x,y)\right)^{\frac{q}{p}} \le \int [d(x,y)]^q \, d\pi^q(x,y) = W_q^q(\mu,\nu). \tag{65}$$

For the second inequality follows from

$$W_q^q(\mu,\nu) \le \int [d(x,y)]^q \, d\pi^p(x,y) \le \mathrm{diam}(X)^{q-p} \int [d(x,y)]^p \, d\pi^p(x,y) = \mathrm{diam}(X)^{q-p} W_p^p(\mu,\nu). \tag{66}$$

□

This is saying that when X is bounded, all W_p defines equivalent topology. However, we note that W_p are not strongly equivalent ($\exists C_1, C_2 > 0, \forall \mu,\nu, C_1 W_p(\mu,\nu) \le W_q(\mu,\nu) \le C_2 W_p(\mu,\nu)$) even when X is bounded. A simple counterexample would be

$$X = [0,1], \quad \mu = \delta_0, \quad \nu_n = \left(1 - \frac{1}{n}\right)\delta_0 + \frac{1}{n}\delta_{\frac{1}{n}}. \tag{67}$$

It's clear that $W_p(\mu,\nu_n) = n^{-(1+\frac{1}{p})} \to 0$ $(n \to \infty)$, so $\frac{W_q(\mu,\nu_n)}{W_p(\mu,\nu_n)} \to \infty$ $(n \to \infty)$ if $q \ge p$, which contradicts with the existence of the constant C_2.

When we discuss the topology on the space of measures, the **strong topology** is typically taken as the one induced by the total variation:

$$\mathrm{TV}(\mu,\nu) := \frac{1}{2} \sup_{\|f\|_\infty \le 1} \int f \, d(\mu - \nu). \tag{68}$$

Actually, the total variation distance can also be seen as a Wasserstein distance under a trivial metric $\tilde{d}$.

Lemma 6. *Consider the trivial distance $\tilde{d}(x,y) = \mathbb{I}_{x\neq y}$, then $W_1^{\tilde{d}}(\mu,\nu) = \mathrm{TV}(\mu,\nu)$, i.e. total variation is the 1-Wasserstein distance under $\tilde{d}$.*

Proof. By the variational characterization of total variation distance,

$$\mathrm{TV}(\mu,\nu) = \min_{X\sim\mu, Y\sim\nu} \mathbb{P}(X \neq Y). \tag{69}$$

If one finds trouble proving this equality, check that the total variation coupling

$$\pi(x,y) = \begin{cases} \mu(x) \wedge \nu(y) & \text{if } x = y \\ \frac{[(\mu(x)-\nu(x))\vee 0]\cdot[(\nu(y)-\mu(y))\vee 0]}{\mathrm{TV}(\mu,\nu)} & \text{if } x \neq y \end{cases} \in \Pi(\mu,\nu) \tag{70}$$

attains the minimum. Notice that $\mathbb{E}\tilde{d}(X,Y) = \mathbb{P}(X \neq Y)$ concludes the proof. □

We are not satisfied with the strong topology induced by the total variation distance on the space of measures since the notion of convergence is too strong to be of our interest. Consider a sequence of distinct Dirac point masses δ_{x_n} such that $x_n \to x$ $(n \to \infty)$. Clearly, under the notion of convergence in distribution, $\delta_{x_n} \xrightarrow{d} \delta_x$. However, this is not the case under the strong topology since

$$\forall n, \mathrm{TV}(\delta_{x_n}, \delta_x) = 1. \tag{71}$$

On the other hand, if we equip the space of measures with the **weak* topology** induced by the p-Wasserstein distance (by default $c(x,y) = |x-y|^p$), then

$$W_p(\delta_{x_n}, \delta_x) = |x_n - x| \to 0 \ (n \to \infty). \tag{72}$$

There seems to be a connection between the Wasserstein topology and the convergence in distribution of measures. Before entering into that, we first prove that, as a special case, when X is discrete, two topologies coincide.

Lemma 7. *When X is discrete under metric d, i.e. $d_{min} := \inf_{x,y\in X} d(x,y) < \infty, d_{max} := \sup_{x,y\in X} d(x,y) < \infty$, then the strong topology and the weak topology (induced by W_p for any $p \geq 1$) are equivalent.*

Proof. Followed from the trivial estimate

$$\forall x, y \in X, d_{min} \cdot \tilde{d}(x,y) \leq d(x,y) \leq d_{max} \cdot \tilde{d}(x,y), \tag{73}$$

and the total variation distance as a Wasserstein distance,

$$d_{min} \cdot \mathrm{TV}(\mu,\nu) \leq W_1(\mu,\nu) \leq d_{max} \cdot \mathrm{TV}(\mu,\nu). \tag{74}$$

Since the topology induced by W_p distance are equivalent for $\forall p \geq 1$ when $d_{max} < \infty$, the topology induced by any W_p distance is equivalent to the strong topology on a discrete space. □

It turns out that on a compact set $X \subset \mathbb{R}^d$, the notion of convergence under the topology induced by W_p aligns with the convergence in distribution.

Theorem 4. *If $X \subset \mathbb{R}^d$ is compact, then $\mu_n \xrightarrow{d} \mu$ $(n \to \infty)$ iff $W_p(\mu_n, \mu) \to 0$ $(n \to \infty)$.*

Proof. X is bounded so it suffices to prove for $p = 1$. Here we have to use the Kantorovich-Rubinstein theorem

providing a characterization of the W_1 distance:

$$W_1(\mu,\nu) = \sup_{\varphi} \int \varphi \, d(\mu - \nu), \tag{75}$$

where the supreme is taken among all φ that are 1-Lipschitz. We will skip the proof for now and come back to it when talking about the Kantorovich duality.

If $W_1(\mu_n, \mu) \to 0$, then for any Lipschitz f with Lipschitz constant L_f,

$$\frac{1}{L_f} \int f \, d(\mu_n - \mu) \le W_1(\mu_n, \mu) \to 0. \tag{76}$$

By Portmanteau theorem recognizing all Lipschitz functions as test functions, $\mu_n \xrightarrow{d} \mu$.

Conversely, if $\mu_n \xrightarrow{d} \mu$, there exists subsequence $\{m_k\}$ and a sequence of 1-Lipschitz functions $\{\varphi_{m_k}\}$ such that

$$W_1(\mu_{m_k}, \mu) \le \int \varphi_{m_k} \, d(\mu_{m_k} - \mu) + \frac{1}{k}, \ W_1(\mu_{m_k}, \mu) \to \limsup_{n\to\infty} W_1(\mu_n, \mu) \ (k \to \infty). \tag{77}$$

The sequence of functions is uniformly equicontinuous and uniformly bounded. Arzela-Ascoli theorem identifies a uniform limit φ of a further subsequence of $\{\varphi_{m_k}\}$.

$$\limsup_{n\to\infty} W_1(\mu_n, \mu) \le \limsup_{k\to\infty} \int (\varphi_{m_k} - \varphi) \, d\mu_{m_k} + \int \varphi \, d(\mu_{m_k} - \mu) + \int (\varphi - \varphi_{m_k}) \, d\mu + \frac{1}{k} = 0 \tag{78}$$

by the uniform convergence and $W_1(\mu_n, \mu) \to 0$. □

Remark. *When it comes to the whole Euclidean space $X = \mathbb{R}^d$, we have that $W_1(\mu_n, \mu) \to 0$ iff $\mu_n \xrightarrow{d} \mu$ and $\int |x|^p \, d\mu_n \to \int |x|^p \, d\mu$. Besides the convergence in distribution, one also requires the convergence of the p-th moment. The proof projects μ_n onto a compact set with small changes in the Wasserstein distance and uses the theorem above. Note that the convergence of the p-th moment is necessary. Consider counterexample: $\mu_n = (1-\frac{1}{n})\delta_0 + \frac{1}{n}\delta_n \xrightarrow{d} \mu = \delta_0$, but μ_n has mean 1 while μ has mean 0. This leads to $W_1(\mu_n, \mu) = 1 \not\to 0$.*

From probability theory, it's clear that the topology induced by Levy-Prokhorov metric aligns with the convergence in distribution. As a result, on compact $X \subset \mathbb{R}^d$, the Wasserstein topology is equivalent to the Levy-Prokhorov topology, although both have different motivations.

Kantorovich Duality

When numerically solving the Kantorovich problem, it's hard to start with the primal problem:

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\, d\pi(x,y) \tag{79}$$

due to the characterization of the coupling. Instead, one considers the dual problem and approximates the Kantorovich potentials. We first heuristically derive the dual problem and then talk about how to prove relevant results.

In the context of optimization, the dual problem refers to the one concerning the Lagrange dual function. The Kantorovich problem is an LP with linear constraints $(P_1)_{\#}\pi = \mu, (P_2)_{\#}\pi = \nu$. It seems hard to write down the Lagrange multiplier function since our optimizer is a measure. However, inspired by the Riesz–Markov–Kakutani representation theorem, the Lanrange multipliers shall be denoted as two integrable functions $\varphi : X \to \mathbb{R}, \psi : Y \to \mathbb{R}$ such that $\langle \varphi, \mu\rangle := \int \varphi\, d\mu, \langle \psi, \nu\rangle := \int \psi\, d\nu$ are defined. In this sense, we write down the Lagrangian:

$$Q(\pi,\varphi,\psi) = \int c(x,y)\, d\pi(x,y) + \langle \varphi, \mu - (P_1)_{\#}\pi\rangle + \langle \psi, \nu - (P_2)_{\#}\pi\rangle\,. \tag{80}$$

The dual function is defined as:

$$\begin{aligned} J(\varphi,\psi) &= \inf_{\pi} Q(\pi,\varphi,\psi) && (81)\\ &= \int \varphi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) + \inf_{\pi}\left\{\int [c(x,y) - \varphi(x) - \psi(y)]\, d\pi(x,y)\right\} && (82)\\ &= \begin{cases} \int \varphi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) & \text{if } c(x,y) \geq \varphi(x) + \psi(y) \\ -\infty & \text{else} \end{cases}. && (83) \end{aligned}$$

As a result, we derived the **Kantorovich dual problem**:

$$\sup_{\varphi,\psi} J(\varphi,\psi) := \int \varphi\, d\mu + \int \psi\, d\nu, \tag{84}$$

$$s.t.\ \varphi(x) + \psi(y) \leq c(x,y). \tag{85}$$

Due to the LP nature of the primal problem, it's not surprising at all that **strong duality** holds, i.e. the primal and the dual problem has the same optimal value of the objective function. This is called the Kantorovich duality and φ, ψ are called Kantorovich potentials.

Due to optimization theory, the weak duality always holds, i.e. the optimal objective value of the dual is always less than the optimal objective value of the primal. The difficulty lies in proving the converse. In this situation, we need to borrow tools from the optimization theory, known as the Frenchel-Rockafeller duality.

Theorem 5 (Frenchel-Rockafeller Duality)**.** *E is a normed vector space, with $\Theta, \Sigma : E \to \mathbb{R} \cup \{\infty\}$ to be convex*

functions. Assume that $\exists z_0 \in E, \Theta(z_0) < \infty, \Sigma(z_0) < \infty$ *and* Θ *is continuous at* z_0*, then*

$$\inf_{z\in E} \{\Theta(z) + \Sigma(z)\} = \max_{z^*\in E^*} \{-\Theta^*(-z^*) - \Sigma^*(z^*)\}, \tag{86}$$

where $\Theta*, \Sigma* : E^* \to \mathbb{R} \cup \{\infty\}$ *are Frenchel conjugates. Moreover, the maximum on the RHS can be attained.*

Proof. The proof can be found everywhere so we only provide a sketch. Let $A := \text{epi}(\Theta), B := \text{hypo}(M - \Sigma) \subset E \times \mathbb{R}$ where $M := \inf_{z\in E} \{\Theta(z) + \Sigma(z)\}$. Both sets are convex and non-empty so there exists a hyperplane $H = \{(x, t) \in E \times \mathbb{R} : f(x) + kt = \alpha, f \in E^*\}$ that separates the disjoint convex open set $C = A^\circ$ and convex B. Prove that $k \neq 0$ (the hyperplane is not parallel to the last dimension), which implies that $z^* = \frac{f}{k}$ attains the maximum on the RHS. □

At this point, we provide the proof of the Kantorovich duality.

Theorem 6 (Kantorovich Duality)**.** *For Polish spaces* X, Y *and l.s.c. cost* c*,*

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x, y)\, d\pi(x, y) = \sup_{\varphi(x)+\psi(y)\le c(x,y)} \int \varphi\, d\mu + \int \psi\, d\nu. \tag{87}$$

Proof. One side of the inequality is obvious due to weak duality. We only prove that the RHS is larger than the LHS. For simplicity, we only prove for compact X, Y and continuous c (can be relaxed, needed for the Riesz–Markov–Kakutani representation theorem to hold such that the Frenchel conjugates have simple forms).

In this case, $E = C_c(X \times Y)$ is equipped with the sup norm, consider convex functions

$$\Theta(u) = \begin{cases} 0 & \text{if } u(x, y) \ge -c(x, y) \\ +\infty & \text{else} \end{cases}, \ \Sigma(u) = \begin{cases} \int \varphi\, d\mu + \int \psi\, d\nu & \text{if } u(x, y) = \varphi(x) + \psi(y) \\ +\infty & \text{else} \end{cases}. \tag{88}$$

Compute the Frenchel conjugates on the collection of all Radon measures $E^* = \mathscr{M}(X \times Y)$:

$$\Theta^*(-\pi) = \begin{cases} \int c(x, y)\, d\pi(x, y) & \text{if } \pi \in \mathscr{M}_+(X \times Y) \text{ (positive)} \\ +\infty & \text{else} \end{cases}, \Sigma^*(\pi) = \begin{cases} 0 & \text{if } \pi \in \Pi(\mu, \nu) \\ +\infty & \text{else} \end{cases} \tag{89}$$

The Frenchel-Rockafeller duality concludes the proof. □

Remark. *As a corollary of the Frenchel-Rockafeller duality, the infimum in the primal Kantorovich problem is always attained. Actually, the supreme in the dual problem is also always attained.*