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

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

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 $\phi_S(t) = e^{it^T\mu_2 - \frac{1}{2}t^T\Sigma_2 t}$ yields the Riccati equation

$$A\Sigma_1 A = \Sigma_2. \tag{25}$$

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*

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

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

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

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.

$$\begin{aligned}
W_p(\mu,\eta) &\le \left(\int [d(x,z)]^p \, d\zeta(x,z) \right)^{\frac{1}{p}} = \left(\int [d(x,z)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} && (45)\\
&\le \left(\int [d(x,y) + d(y,z)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} && (46)\\
&\le \left(\int [d(x,y)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} + \left(\int [d(y,z)]^p \, d\sigma(x,y,z) \right)^{\frac{1}{p}} \quad \text{(Minkowski)} && (47)\\
&= \left(\int [d(x,y)]^p \, d\pi(x,y) \right)^{\frac{1}{p}} + \left(\int [d(y,z)]^p \, d\xi(y,z) \right)^{\frac{1}{p}} && (48)\\
&= W_p(\mu,\nu) + W_p(\nu,\eta). && (49)
\end{aligned}$$

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,

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

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

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

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

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

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* $1 \leq p \leq q$, $W_p(\mu,\nu) \leq W_q(\mu,\nu) \leq \operatorname{diam}(X)^{\frac{q-p}{q}} W_p^{\frac{p}{q}}(\mu,\nu)$, *where* $\operatorname{diam}(X) = \sup_{x,y\in X} d(x,y)$ *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 $x \mapsto x^{\frac{q}{p}}$,

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

For the second inequality follows from

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

□

This is saying that when X is bounded, all W_p defines equivalent topology. However, we note that W_p are not strongly equivalent ($\exists C_1, C_2 > 0, \forall \mu,\nu, C_1 W_p(\mu,\nu) \leq W_q(\mu,\nu) \leq C_2 W_p(\mu,\nu)$) 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 $q \geq p$, 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:

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

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* $W_1^{\tilde{d}}(\mu,\nu) = \operatorname{TV}(\mu,\nu)$, *i.e. total variation is the 1-Wasserstein distance under* $\tilde{d}$.

Proof. By the variational characterization of total variation distance,

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

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,

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

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

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

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

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

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**:

$$\begin{aligned}
&\sup_{\varphi,\psi}J(\varphi,\psi):=\int\varphi\,d\mu+\int\psi\,d\nu, && (84)\\
&s.t.\ \varphi(x)+\psi(y)\leq c(x,y). && (85)
\end{aligned}$$

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 Fenchel-Rockafeller duality.

Theorem 5 (Fenchel-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 Fenchel 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*,*

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

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 Fenchel conjugates have simple forms).

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

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

Compute the Fenchel 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 Fenchel-Rockafeller duality concludes the proof. □

Remark. *As a corollary of the Fenchel-Rockafeller duality, the infimum in the primal Kantorovich problem is always attained. Actually, the supreme in the dual problem is also always attained. The proof uses the c-transform which will be mentioned in the later context.*

By the complementary slackness, if one solves the optimal φ^*, ψ^*, the optimal coupling is given by

$$\pi^*(x, y) > 0 \Leftrightarrow c(x, y) = \varphi^*(x) + \psi^*(y). \tag{90}$$

Kantorovich-Rubinstein Theorem

Here we provide the proof of the Kantorovich-Rubinstein theorem that provides a useful characterization of the W_1 distance corresponding to the Euclidean distance $d(x,y) = |x-y|$.

Theorem 7 (Kantorovich-Rubinstein Theorem)**.** *For any μ, ν,*

$$W_1(\mu,\nu) = \sup_f \int f\, d(\mu - \nu), \tag{91}$$

where the supreme is taken among all f that are 1-Lipschitz. The supreme is attained by $f = \psi^c$, where

$$\psi^c(x) := \inf_{y\in Y} \{c(x,y) - \psi(y)\} \tag{92}$$

is defined as the c-transform of ψ (in this case $c(x,y) = |x-y|$).

Proof. By Kantorovich duality,

$$W_1(\mu,\nu) = \sup_{\varphi(x)+\psi(y)\le|x-y|} \int \varphi\, d\mu + \int \psi\, d\nu. \tag{93}$$

We denote by $|\varphi|_L$ the Lipschitz constant of function φ. Setting $\varphi = -\psi, |\varphi|_L \le 1$ to see that

$$W_1(\mu,\nu) \ge \sup_{|\varphi|_L\le 1} \int \varphi\, d(\mu - \nu). \tag{94}$$

To prove the inequality in the other direction, we only need to check that the supreme can be attained.

Consider ψ^c:

$$\forall x, x' \in X, \forall y \in Y, \psi^c(x) \le |x-y| - \psi(y) \le |x-x'| + |y-x'| - \psi(y). \tag{95}$$

Taking infimum on both sides w.r.t. $y \in Y$ proves $\psi^c(x) \le |x-x'| + \psi^c(x')$, and exchanging x, x' proves the 1-Lipschitz property. Now we check that

$$\forall\ \varphi(x) + \psi(y) \le |x-y|, \int \varphi\, d\mu + \int \psi\, d\nu \le \int \psi^c\, d(\mu-\nu). \tag{96}$$

since $\varphi \le \psi^c$ while $\psi^c \le -\psi$ (set $y = x$ in the definition). Combining those inequalities yields

$$W_1(\mu,\nu) \le \sup_{|f|_L\le 1} \int f\, d(\mu-\nu) \le W_1(\mu,\nu), \tag{97}$$

which proves the theorem and also we see that the supreme is attained by $f = \psi^c$. □

Wasserstein Geometry and Multimarginal OT (MMOT)

When considering the definition of a line segment in the Wasserstein space between μ_0 and μ_1, there are two natural formulations. The first one is called the **interpolation of metric**, i.e., the line segment is determined by minimizing the interpolated distance

$$\mu_t := \arg\min_{\mu} \left\{(1-t)W_2^2(\mu,\mu_0) + tW_2^2(\mu,\mu_1)\right\},\ t \in [0,1]. \tag{98}$$

The second way is to follow the **geodesic**, i.e., given the optimal transport map $T_{0,1}$ from μ_0 to μ_1, μ_t is derived by directly interpolating the transport map:

$$\mu_t := (tT_{0,1} + (1-t)id)_{\#}\mu_0. \tag{99}$$

In the case of two measures, two formulations coincide.

Proof. We prove that the μ_t defined in the geodesic setting minimizes the optimization problem of the interpolated metric.

$$(1-t)W_2^2(\mu_t,\mu_0) + tW_2^2(\mu_t,\mu_1) = t^2(1-t)W_2^2(\mu_0,\mu_1) + t(1-t)^2W_2^2(\mu_0,\mu_1) = t(1-t)W_2^2(\mu_0,\mu_1). \tag{100}$$

The first equation follows from the fact that $W_2^2(T_{\#}\mu,\mu) = \int |x-T(x)|^2\, d\mu(x)$ so that

$$W_2^2(\mu_t,\mu_0) = t^2 \int |x - T_{0,1}(x)|^2\, d\mu_0(x) = t^2 W_2^2(\mu_1,\mu_0). \tag{101}$$

By geometric mean-square mean inequality and the triangle inequality,

$$\forall \mu,\ t(1-t)W_2^2(\mu_0,\mu_1) \le t(1-t)[W_2(\mu,\mu_0) + W_2(\mu,\mu_1)]^2 \tag{102}$$

$$\le (1-t)W_2^2(\mu,\mu_0) + tW_2^2(\mu,\mu_1). \tag{103}$$

This concludes the proof. □

Remark. *In the two-measure case, computing μ_t for given $t \in [0,1]$ is an easy task. Notice that by Brenier's theorem $T_{0,1} = \nabla f$ for some convex f, so $tT_{0,1} + (1-t)id = \nabla g$ where $g(x) = tf(x) + \frac{1-t}{2}|x|^2$ is still the gradient of a convex function. By Brenier's theorem once again, $tT_{0,1} + (1-t)id$ is actually the optimal transport map from μ_0 to μ_t. In this sense, after solving for $T_{0,1}$ (a single OT problem), we have actually solved out all μ_t.*

The consistency between the geodesic and metric interpolation formulations breaks when the number of measures exceeds three. Consider μ_0, μ_1, μ_2 and we wish to find a triangle in the Wasserstein space with those three vertices. The geodesic formulation finds out optimal transport maps $T_{0,1}, T_{1,2}, T_{0,2}$ and interpolates the transport maps, while the metric interpolation formulation directly interpolates $W_2^2(\mu,\mu_0), W_2^2(\mu,\mu_1), W_2^2(\mu,\mu_2)$. Consider three complex numbers $a = 1, b = e^{i\frac{2}{3}\pi}, c = e^{i\frac{4}{3}\pi}$ such that μ_0 is uniform on $\{a,b\}$, μ_1 is uniform on $\{b,c\}$, μ_2 is uniform on $\{a,c\}$. Under the metric interpolation formulation, the center of the triangle, i.e., $\arg\min_{\mu} \frac{1}{3}W_2^2(\mu,\mu_0) + \frac{1}{3}W_2^2(\mu,\mu_1) +$

$\frac{1}{3}W_2^2(\mu, \mu_2)$, is invariant under any rotations with angle $k\frac{2}{3}\pi, k \in \mathbb{Z}$. Under the geodesic formulation, each point in the triangle consists of a pair of disjoint points which does not exhibit the symmetry. Hence, although we have no idea what those two triangles look like, they must not be the same set.

One practical concern when it comes to the Wasserstein geometry, is to find the "center" of several given measures. In the case of three given measures, we want to solve

$$\min_{\mu} \frac{1}{3}W_2^2(\mu, \mu_0) + \frac{1}{3}W_2^2(\mu, \mu_1) + \frac{1}{3}W_2^2(\mu, \mu_2). \tag{104}$$

By Brenier's theorem, we can write $\mu = (\nabla f)_{\#}\mu_0$ for convex f but, unlike the two-measure case, this is not making our life easier. We get the optimization problem

$$\min_{f \text{ convex}} \frac{1}{3}\int |\nabla f(x) - x|^2 \, d\mu_0(x) + \frac{1}{3}W_2^2((\nabla f)_{\#}\mu_0, \mu_1) + \frac{1}{3}W_2^2((\nabla f)_{\#}\mu_0, \mu_2). \tag{105}$$

The latter two terms are generally not convex in ∇f, so the problem is an non-convex optimization in ∇f. In this case, we shall do some transformations to the original problem by rewriting it as

$$\min_{\mu} \min_{\gamma_0, \gamma_1, \gamma_2, \gamma_i \in \Pi(\mu_i, \mu)} \frac{1}{3}\sum_i \int |x - y|^2 \, d\gamma_i(x, y), \tag{106}$$

explicitly specifying the transport maps used in the Wasserstein distances as γ_i. Then we merge the three constraints on γ_i to a single constraint denoted as a multimarginal coupling

$$\min_{\mu} \min_{\gamma \in \Pi(\mu_0, \mu_1, \mu_2, \mu)} \frac{1}{3}\sum_i \int |x_i - y|^2 \, d\gamma(x_0, x_1, x_2, y), \tag{107}$$

where $\Pi(\mu_0, \mu_1, \mu_2, \mu)$ denotes the collection of probability measures on $(\mathbb{R}^d)^4$ with the corresponding marginals. Notice that we are minimizing w.r.t. μ on the outer layer so we actually have the freedom of choosing μ when doing the minimization. The problem reduces to

$$\min_{\gamma \in \Pi(\mu_0, \mu_1, \mu_2, \cdot)} \int_{(\mathbb{R}^d)^4} \frac{1}{3}\sum_i |x_i - y|^2 \, d\gamma(x_0, x_1, x_2, y). \tag{108}$$

This is an instance of multimarginal optimal transport (MMOT) with cost $c(x_0, x_1, x_2, y) = \frac{1}{3}\sum_i |x_i - y|^2$ and it once again becomes an LP problem (linear in γ).

MMOT are hard to solve numerically (even for 30 marginals) and little is known about if γ can be reduced to a transport map, e.g., $x_1 \mapsto (x_1, T_{1,2}(x_1), ..., T_{1,k}(x_1))$ as a curve, or $(x_1, x_2) \mapsto (x_1, x_2, T_{1,2,3}(x_1, x_2), ..., T_{1,2,k}(x_1, x_2))$ as a surface, etc.

OT Numerics

There are two general algorithms for solving OT numerically. One is the back-and-forth method, which is fast and accurate but only works for a small dimension d. The other is Sinkhorn, which works for large d, and one can control the trade-off between efficiency and accuracy. Both algorithms start from the Kantorovich dual and tries to learn the Kantorovich potentials. We first state the problem in the general setting without specifying the cost c, but in specific context a c with special structure might be needed to simplify the methods. In OT numerics, μ, ν are typically formed as empirical measures (observed samples from the measures) instead of complete absolute continuous measures.

c-transform

In the proof of the Kantorovich-Rubinstein theorem, we have seen the power of the c-transform and here we provide a more comprehensive introduction to prepare for the numerical methods. For Kantorovich potentials $\varphi : X \to \mathbb{R}, \psi : Y \to \mathbb{R}$, the c-transforms $\varphi^c : Y \to \mathbb{R}, \psi^c : X \to \mathbb{R}$ are defined as

$$\varphi^c(y) := \inf_{x\in X} \left\{c(x,y) - \varphi(x)\right\}, \quad \psi^c(y) := \inf_{y\in Y} \left\{c(x,y) - \psi(y)\right\}. \tag{109}$$

The Kantorovich dual maximizes $\int \varphi \, d\mu + \int \psi \, d\nu$ under the constraint $\varphi(x) + \psi(y) \le c(x,y)$, so the c-transform is naturally defined based on the form of the constraint. As an immediate consequence,

$$\varphi^c \ge \psi, \quad \psi^c \ge \varphi. \tag{110}$$

As a result, we denote

$$J(\varphi) := \int \varphi \, d\mu + \int \varphi^c \, d\nu, \quad I(\psi) := \int \psi^c \, d\mu + \int \psi \, d\nu \tag{111}$$

as two objective functions only in φ and ψ respectively. Clearly,

$$\text{OT cost} = \sup_{\varphi} J(\varphi) = \sup_{\psi} I(\psi) \tag{112}$$

turns the dual problem into an unconstrainted optimization problem for only one of the Kantorovich potentials.

A clever investigation of the double c-transform shows that

$$\varphi^{cc}(x) = \inf_{y\in Y} \left\{c(x,y) - \varphi^c(y)\right\} = \inf_{y\in Y} \left\{c(x,y) - \inf_{z\in X} \left\{c(z,y) - \varphi(z)\right\}\right\} \tag{113}$$

$$= \inf_{y\in Y} \sup_{z\in X} \left\{c(x,y) - c(z,y) + \varphi(z)\right\} \ge \varphi(x), \tag{114}$$

so $\varphi^{cc} : X \to \mathbb{R}, \psi^{cc} : Y \to \mathbb{R}$ and

$$\varphi^{cc} \geq \varphi, \quad \psi^{cc} \geq \psi. \tag{115}$$

If one investigates the triple c-transform,

$$\varphi^{ccc}(y) = \inf_{x\in X} \{c(x, y) - \varphi^{cc}(x)\} = \inf_{x\in X} \sup_{z\in Y} \{c(x, y) - c(x, z) + \varphi^c(z)\} \geq \varphi^c(y). \tag{116}$$

On the other hand,

$$\begin{aligned}\varphi^{ccc}(y) &= \inf_{x\in X} \sup_{z\in Y} \{c(x, y) - c(x, z) + \varphi^c(z)\} = \inf_{x\in X} \sup_{z\in Y} \inf_{w\in X} \{c(x, y) - c(x, z) + c(w, z) - \varphi(w)\} &(117)\\ &\leq \inf_{x\in X} \{c(x, y) - \varphi(x)\} \ \text{(set } w \text{ as } x) = \varphi^c(y). &(118)\end{aligned}$$

This reveals the structure that

$$\varphi^{ccc} = \varphi^c, \quad \psi^{ccc} = \psi^c. \tag{119}$$

Essentially, the only nontrivial transforms are the single and double c-transform.

At this point, we find that replacing Kantorovich potentials with their double c-transforms increases the objective values:

$$J(\varphi^{cc}) \geq J(\varphi), \quad I(\psi^{cc}) \geq I(\psi). \tag{120}$$

Unfortunately, this operation can only be used for once since $J(\varphi^{cccc}) = J(\varphi^{cc})$, and the devil lies in the property $\varphi^{ccc} = \varphi^c$. This puts up a challenge for designing OT numerics that one cannot simply assign ψ as φ^c and alternate between two potentials. One shall realize that this failure is a natural consequence since the c-transform only contains the cost c but contains no information of the two distributions μ, ν! If we have used no information of μ, ν, then we definitely have no idea what the optimal transport map between them looks like.

Inspired by this observation, we realize that "something" has to be done before/after assigning the c-transform such that we can break the property $\varphi^{ccc} = \varphi^c$ and always make progress. This "something" shall change the potentials in a good direction and shall be based on the information of μ and ν. In the following context, we introduce two algorithms with different motivations. The back-and-forth algorithms does a gradient step before applying the c-transform, while the Sinkhorn algorithm changes the notion of c-transform such that $\varphi^{ccc} = \varphi^c$ no longer holds.

Back-and-Forth

The motivation of back-and-forth is to do gradient update for learning the potentials. Let's first focus on learning φ, which requires the functional derivative of $J(\varphi)$ w.r.t. φ. We first calculate the first variation of $J(\varphi)$ w.r.t. φ, denoted as $\delta J(\varphi)(u) := \lim_{t\to 0} \frac{J(\varphi+tu)-J(\varphi)}{t}$ in the Gateaux sense.

$$\begin{aligned} \delta J(\varphi)(u) &= \int u\, d\mu + \lim_{t\to 0} \frac{\int (\varphi+tu)^c - \varphi^c\, d\nu}{t} && (121) \\ &= \int u\, d\mu - \int u(T_\varphi(y))\, d\nu(y), && (122) \end{aligned}$$

where $T_\varphi(y) := \arg\min_{x\in X} \{c(x,y)-\varphi(x)\}$ is the value of x that attains the infimum in $\varphi^c(y)$. Notice that this calculation holds for $c(x,y) = h(x-y)$ where h is strictly convex. Heuristically, the infimum in $(\varphi+tu)^c$ is also attained at $x = T_\varphi(y)$ and the perturbation in the minimizer results in an negligible higher-order term.

Remark. *The first variation provides insights to OT problems. It's clear that if $\varphi^* \in \arg\min_\varphi J(\varphi)$, then $\forall u, \delta J(\varphi^*)(u) = 0$, which implies that*

$$\forall u, \int u\, d\mu = \int u\, d(T_{\varphi^*})_\# \nu. \tag{123}$$

As a result, $\mu = (T_{\varphi^*})_\#\nu$ so T_{φ^*} is a legal transport map between μ and ν. This insight provides a way to extend Brenier's theorem from the quadratic cost to the general $c(x,y) = h(x-y)$ for any strictly convex h.

Although the first variation does not depend on the geometry of the space, gradient does depend on the geometry in the sense that

$$\delta J(\varphi)(u) =: \langle \nabla_{\mathscr{H}} J(\varphi), u\rangle_{\mathscr{H}} \tag{124}$$

for some selected Hilbert space $\mathscr{H}$. After the selection of $\mathscr{H}$, we would derive a formula of the gradient and implement the gradient ascent to iteratively optimize the Kantorovich potentials.

One natural choice of $\mathscr{H}$ would be the L^2 function space, in which case

$$\nabla_{L^2} J(\varphi) = d[\mu - (T_\varphi)_\#\nu], \tag{125}$$

as the likelihood of the measure $\mu - (T_\varphi)_\#\nu$. The gradient has a natural interpretation as closing the gap between μ and the transported ν through the current φ. However, this gradient method has unstable performance numerically, as can be expected, since likelihoods of measures lack regularity and can go wild.

To explain this phenomenon mathematically, we continue to calculate the second variation at u:

$$\delta^2 J(\varphi)(u,u) := \lim_{t\to 0} \frac{\delta J(\varphi+tu)(u) - \delta J(\varphi)(u)}{t}. \tag{126}$$

To get a closed-form solution, we assume a quadratic cost $c(x, y) = \frac{1}{2}|x - y|^2$. Now $T_\varphi(y) - y - \nabla\varphi(T_\varphi(y)) = 0$ and:

$$\begin{aligned}
\delta^2 J(\varphi)(u, u) &= \lim_{t\to 0} \frac{-\int u(T_{\varphi+tu}(y))\, d\nu(y) + \int u(T_\varphi(y))\, d\nu(y)}{t} &(127)\\
&= -\int [\partial_t u(T_{\varphi+tu}(y))]\Big|_{t=0} d\nu(y) &(128)\\
&= -\int [\nabla u(T_\varphi(y))]^T [I - \nabla^2\varphi(T_\varphi(y))] \nabla u(T_\varphi(y))\, d\nu(y) &(129)\\
&= -\int [\nabla u(x)]^T [I - \nabla^2\varphi(x)] \nabla u(x)\, d(T_\varphi)_\# \nu(x). &(130)
\end{aligned}$$

The calculations of $[\partial_t u(T_{\varphi+tu}(y))]\Big|_{t=0}$ follows from the chain rule, and differentiating both sides of $T_{\varphi+tu}(y) - y - \nabla(\varphi + tu)(T_{\varphi+tu}(y)) = 0$ w.r.t. t. The key observation is that $\delta^2 J(\varphi)(u, u)$ contains ∇u, which is not necessary bounded for $u \in L^2$, since ∇u may blow up. From the optimization theory, we quote the following lemma (which is called by Matt the fundamental theorem of optimization):

Lemma 8. *For $F : \mathscr{H} \to \mathbb{R}$, if there exists $L > 0$ such that*

$$\forall a, b \in \mathscr{H}, F(a) \le F(b) + \langle \nabla_{\mathscr{H}} F(b), a - b\rangle_{\mathscr{H}} + \frac{L}{2}\|a - b\|^2_{\mathscr{H}}, \tag{131}$$

then the gradient step $b_{new} = b - \frac{1}{L}\nabla_{\mathscr{H}} F(b)$ decreases F, i.e.,

$$F(\tilde{b}) \le F(b) - \frac{1}{2L}\|\nabla_{\mathscr{H}} F(b)\|^2_{\mathscr{H}}. \tag{132}$$

The lemma tells us a sufficient condition for the gradient step to work is when $\nabla^2 F \le LI$, i.e., the Hessian has a bounded spectrum. **The lack of regularity in ∇u causes the possible blowup of the second variation, which in turn causes the unstable numerical performance of the L^2 gradient method!** Then, shall we choose a very strong Hilbert space $\mathscr{H}$ for u to ensure nice enough regularity? The answer is again no from the lemma, since the progress of the gradient step $\frac{1}{2L}\|\nabla_{\mathscr{H}} F(b)\|^2_{\mathscr{H}}$ would be small for strong Hilbert spaces. **The wisdom is to choose the weakest Hilbert space where gradient methods are stable.** As long as gradient methods are stable, we wish to maximize the progress at each gradient step to have an algorithm that converges faster.

Remark. *If we discretize everything at the very beginning, we would not have noticed the problem caused by ∇u. The wisdom here is that **we shall look at objects in infinite-dimensional spaces first to capture the structures, and always put off discretizeation to the last possible moment**.*

Naturally, we consider the weakest Hilbert space where we have any hope of having a stable gradient method. The Hilbert space shall be a subspace of L^2, and shall also ensure the regularity of ∇u, which is the Sobolev space:

$$H^1 := \left\{u \in L^2 : \nabla u \in L^2\right\}. \tag{133}$$

This Hilbert space has the canonical norm defined as

$$\|u\|_{H^1} := \sqrt{\|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2}. \tag{134}$$

Here, $\mathscr{H}$ is the Hilbert space where the Kantorovich potentials live, but we don't care about constants added or subtracted from the potentials, e.g. φ and $\varphi + C$ has the same objective value $J(\varphi)$ and are essentially same potentials. As a result, we modulo H^1 by additive constants to get the homogeneous Sobolev space $\dot{H}^1$ with the inner product and the norm:

$$\langle u, v\rangle_{\dot{H}^1} := \langle \nabla u, \nabla v\rangle_{L^2}, \quad \|u\|_{\dot{H}^1} := \|\nabla u\|_{L^2}. \tag{135}$$

To clarify, each element u in $\dot{H}^1$ is an equivalent class $[u] := \left\{v \in H^1 : \nabla u = \nabla v\right\}$, whose elements differ from u by an additive constant. Finally, we are able to specify the Hilbert space

$$\mathscr{H} = \dot{H}^1. \tag{136}$$

The next step is to compute the gradient $\nabla_{\dot{H}^1} J(\varphi)$. Using integration by parts,

$$\begin{aligned} \delta J(\varphi)(u) &= \langle \nabla_{\dot{H}^1} J(\varphi), u\rangle_{\dot{H}^1} && (137)\\ &= \langle \nabla\nabla_{\dot{H}^1} J(\varphi), \nabla u\rangle_{L^2} && (138)\\ &= -\langle \Delta[\nabla_{\dot{H}^1} J(\varphi)], u\rangle_{L^2} && (139)\\ &= \int u \, d[\mu - (T_\varphi)_\# \nu]. && (140) \end{aligned}$$

This implies

$$\nabla_{\dot{H}^1} J(\varphi) = -\Delta^{-1}(d[\mu - (T_\varphi)_\# \nu]) = -\Delta^{-1}(\nabla_{L^2} J(\varphi)). \tag{141}$$

The gradient is the negative inverse Laplacian applied to the density gap $d[\mu - (T_\varphi)_\# \nu]$. Compared to the L^2 gradient, an extra inverse Laplacian appears for the purpose of smoothing, which provides stability of the gradient steps.

Remark. *The inverse Laplacian has the interpretation of smoothing by spreading information in the neighborhood. To understand this, consider the Poisson equation $\Delta u = f$, it's generally the case that $u = \Delta^{-1} f$ has a much nicer regularity compared to the potential f.*

At this point, we derived the gradient algorithm for OT that

$$\varphi^{n+1} = \varphi^n - \alpha^n \Delta^{-1}(d[\mu - (T_{\varphi^n})_\# \nu]), \tag{142}$$

where $T_\varphi(y) = \arg\min_x \{c(x, y) - \varphi(x)\}$. When it comes to numerical implementation, μ and ν are typically organized as empirical measures with n and m point masses. In this case, one has to partition the space into grids to calculate the density gap as the difference in the point mass. Notice that if a point $x \in X$ does not receive a point mass, then the density gap is constantly zero so the value $\varphi(x)$ never gets updated. This implies that we only need to maintain and update the value of $\varphi(x)$ for $x \in X$ that receives a point mass (appears in the empirical observation). This enables us to formulate φ without using other function parameterization and approximation tools.

Remark. *To clarify, we consider the case where $X = Y = \mathbb{R}$, where n observations $x_1, ..., x_n$ from μ and m observations $y_1, ..., y_m$ from ν are provided. In this case, $d\mu = \frac{1}{n}\vec{1}_n \in \mathbb{R}^n$ is the pmf of the empirical measure $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ while $d\nu = \frac{1}{m}\vec{1}_m \in \mathbb{R}^m$. Clearly, $(T_\varphi)_\# \nu$ puts point mass $\frac{1}{m}$ at each $T_\varphi(y_j) \in \{x_1, ..., x_n\}$. Hence we partition $\mathbb{R}$ according to n endpoints $x_1, ..., x_n$, and formulate $\varphi \in \mathbb{R}^n$ as the value of φ evaluated at those n points.*

The inverse of the Laplacian has an easy representation under the Fourier transform, i.e., if $\Delta u = f$, then $\hat{u}(\xi) = -\frac{1}{|\xi|^2}\hat{f}(\xi)$. We use FFT to apply a discrete Fourier transform, then apply the inverse Laplacian on the frequency field, and the inverse Fourier transform, to calculate the term $\Delta^{-1}(d[\mu - (T_{\varphi^n})_\# \nu])$. Be careful here that the normal FFT assumes a periodic boundary condition, which does not match with the structure of OT. Ideally, we wish that there's no flux of mass outside of the computational domain, so a **zero Neumann boundary condition** shall be adopted, i.e.,

$$\nabla T_\varphi \cdot \vec{n} = 0. \tag{143}$$

This can be implemented through the cosine transform.

Finally, we emphasize that all the works done above also apply symmetrically for the other Kantorovich potential ψ. Instead of learning only one of the potentials, we prefer learning both potentials at the same time. Numerical tests imply that learning both potentials helps the algorithm converge much faster compared to learning only one of them. Intuitively, both potentials contain features of the optimal transport map, which might be complementary in some cases. This can also be explained mathematically in the remark below.

Remark. *For given $\varphi^c = \psi, \psi^c = \varphi$, clearly T_φ and T_ψ are inverse of each other. As a result,*

$$DT_\varphi \circ T_\psi = (DT_\psi)^{-1}, \quad DT_\psi \circ T_\varphi = (DT_\varphi)^{-1}. \tag{144}$$

A small eigenvalue of DT_ψ is likely to be a large eigenvalue of DT_φ, which can be easily captured by φ but hard for ψ. This implies that singular features of the derivatives of the potentials might exist, which explains why there's a much faster convergence when learning both potentials.

Our strategy goes like this: after updating φ^n, the corresponding ψ is set as the c-transform $(\varphi^n)^c$. After updating ψ^n, the corresponding φ shall be set as the c-transform $(\psi^n)^c$. The algorithm updates both potentials in an alternating way, which is why this algorithm is called back-and-forth. The details are listed in Alg. 1.

When switching between two potentials, we use the c-transform for a reason. Notice that

$$I(\varphi^c) = \int \varphi^{cc}\, d\mu + \int \varphi^c\, d\nu \geq \int \varphi\, d\mu + \int \varphi^c\, d\nu = J(\varphi). \tag{145}$$

Whenever the c-transform is taken and plugged into the objective function of the other potential, it always increases the objective value. As a result, if the gradient steps in the back-and-forth algorithm are guaranteed to increase the

Algorithm 1 Back-and-Forth

```
repeat
    $\varphi^{n+\frac{1}{2}} = \varphi^n - \alpha^n \Delta^{-1}(d[\mu - (T_{\varphi^n})_{\#}\nu])$
    $\psi^{n+\frac{1}{2}} = (\varphi^{n+\frac{1}{2}})^c$
    $\psi^{n+1} = \psi^{n+\frac{1}{2}} - \beta^n \Delta^{-1}(d[\nu - (T_{\psi^{n+\frac{1}{2}}})_{\#}\mu])$
    $\varphi^{n+1} = (\psi^{n+1})^c$
until Enough iterations are done
```

objective value (with an appropriate stepsize), then the algorithm enjoys **monotonic improvements**, i.e.,

$$J(\varphi^n) \leq J(\varphi^{n+\frac{1}{2}}) \leq I(\psi^{n+\frac{1}{2}}) \leq I(\psi^{n+1}) \leq J(\varphi^{n+1}). \tag{146}$$

Although there has been no convergence proof for this algorithm, it demonstrates fast convergence in experiments and the monotonic improvements is guaranteed, despite the high difficulty of implementation.

Sinkhorn

The motivation of Sinkhorn is to change the notion of c-transform. A natural idea would be to add regularization terms and check what new notions of c-transform we are getting. Without changing the optimal coupling in the Kantorovich problem, we consider

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\pi(x,y)\,dx\,dy + \varepsilon \int \left[\pi(x,y)\log\frac{\pi(x,y)}{\pi^*(x,y)} - \pi(x,y)\right] dx\,dy. \tag{147}$$

Here we denote $\pi(x,y)$ as the joint density and $\mu(x),\nu(y)$ the marginal densities without distinguishing them from the measures. Notice that the $-\pi(x,y)$ term integrates to 1 and is added here for the simplicity of subsequent calculations. The remaining term in the regularization is actually the KL-divergence

$$D_{KL}(\pi||\pi^*), \tag{148}$$

which is zero iff $\pi=\pi^*$. In this sense, adding the entropy regularization term does not change the optimizer π^*. However, π^* is what we want to calculate and is unknown to us, so we have to replace it with some coupling in $\Pi(\mu,\nu)$. The most natural coupling that comes to our mind is $\mu\otimes\nu$. So the **entropy-regularized OT** is formed as:

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\pi(x,y)\,dx\,dy + \varepsilon \int \left[\pi(x,y)\log\frac{\pi(x,y)}{\mu(x)\nu(y)} - \pi(x,y)\right] dx\,dy. \tag{149}$$

Likewise, we consider the dual problem. Assume Kantorovich potentials φ,ψ to be the Lagrange multipliers, the Langrangian is

$$\begin{aligned} Q(\pi,\varphi,\psi) = &\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\pi(x,y)\,dx\,dy + \varepsilon \int \left[\pi(x,y)\log\frac{\pi(x,y)}{\mu(x)\nu(y)} - \pi(x,y)\right] dx\,dy & (150)\\ &+ \langle \varphi, \mu - (P_1)_\#\pi\rangle + \langle \psi, \nu - (P_2)_\#\pi\rangle\,. & (151)\end{aligned}$$

Taking derivative w.r.t. π shows that the π that attains the infimum is given by the one that satisfies

$$c(x,y) - \varphi(x) - \psi(y) + \varepsilon\log\frac{\pi(x,y)}{\mu(x)\nu(y)} = 0. \tag{152}$$

We calculate the Lagrange dual function

$$\inf_{\pi} Q(\pi,\varphi,\psi) = \int \varphi(x)\mu(x)\,dx + \int \psi(y)\nu(y)\,dy - \varepsilon\int \mu(x)\nu(y)e^{-\frac{1}{\varepsilon}(c(x,y)-\varphi(x)-\psi(y))}\,dx\,dy. \tag{153}$$

The **Sinkhorn dual** is given by

$$\sup_{\varphi,\psi} D(\varphi,\psi) := \int \varphi(x)\mu(x)\,dx + \int \psi(y)\nu(y)\,dy - \varepsilon\int \mu(x)\nu(y)e^{-\frac{1}{\varepsilon}(c(x,y)-\varphi(x)-\psi(y))}\,dx\,dy. \tag{154}$$

Notice that this problem no longer has hard constraints, which have been replaced with a relaxed softmax penalty term. With a similar argument to the Kantorovich duality, **strong duality** still holds for EOT.

Naturally, we fix φ and find out the optimal ψ. Differentiate the dual objective w.r.t. ψ:

$$\frac{\delta \langle \psi, \nu \rangle}{\delta \psi}(y) = \nu(y), \quad \frac{\delta \int e^{\frac{1}{\varepsilon}[\psi(y)-c(x,y)]}\nu(y)\, dy}{\delta \psi}(y) = e^{-\frac{1}{\varepsilon}c(x,y)}\nu(y)e^{\frac{1}{\varepsilon}\psi(y)}\frac{1}{\varepsilon}. \tag{155}$$

The optimal ψ for given φ satisfies:

$$\nu(y) - \nu(y)\int \mu(x)e^{\frac{1}{\varepsilon}[\varphi(x)+\psi(y)-c(x,y)]}\, dx = 0. \tag{156}$$

Similar to the back-and-forth method, we only need to consider updating the values $\psi(y)$ for $y \in \text{supp}(\nu)$. Solving for ψ gives

$$\psi(y) = \varphi^{c_\varepsilon}(y) := -\varepsilon \log \int \mu(x)e^{\frac{1}{\varepsilon}[\varphi(x)-c(x,y)]}\, dx. \tag{157}$$

This is defined as the c_ε**-transform** of the potential φ, which contains information of μ and is a relaxation of the c-transform φ^c. Similarly, we define the c_ε-transform for ψ as

$$\psi^{c_\varepsilon}(x) := -\varepsilon \log \int \nu(y)e^{\frac{1}{\varepsilon}[\psi(y)-c(x,y)]}\, dy. \tag{158}$$

When μ, ν are empirical measures, the Kantorovich potentials $\varphi(x), \psi(y)$ are only updated at the place where x, y are observed in the empirical samples.

Remark. *When $\varepsilon \to 0$, the order of $\varphi^{c_\varepsilon}(y)$ is asymptotically dominated by $\sup_x \{\varphi(x) - c(x,y)\} = -\varphi^c(y)$. As a result, asymptotic approximation tells*

$$\varphi^{c_\varepsilon}(y) \approx -\varepsilon \log \int \mu(x)e^{-\frac{1}{\varepsilon}\varphi^c(y)}\, dx = -\varepsilon \log e^{-\frac{1}{\varepsilon}\varphi^c(y)} = \varphi^c(y)\ (\varepsilon \to 0). \tag{159}$$

This shows that as $\varepsilon \to 0$, the c_ε-transform degenerates to the c-transform on the support of μ and ν.

Now that c_ε-transform contains information from μ, ν, we are confident that the previous methodology of replacing φ with φ^{cc} shall work. As an analogue, we denote

$$J_\varepsilon(\varphi) := \int \varphi(x)\mu(x)\, dx + \int \varphi^{c_\varepsilon}(y)\nu(y)\, dy - \varepsilon, \quad I_\varepsilon(\psi) := \int \psi^{c_\varepsilon}(x)\mu(x)\, dx + \int \psi(y)\nu(y)\, dy - \varepsilon, \tag{160}$$

as relaxed versions of $J(\varphi)$ and $I(\psi)$. Starting with φ, ψ, we have

$$D(\varphi, \psi) \le D(\varphi, \varphi^{c_\varepsilon}) = J_\varepsilon(\varphi) \le D(\varphi^{c_\varepsilon c_\varepsilon}, \varphi^{c_\varepsilon}) = I_\varepsilon(\varphi^{c_\varepsilon}) \le D(\varphi^{c_\varepsilon c_\varepsilon}, \varphi^{c_\varepsilon c_\varepsilon c_\varepsilon}). \tag{161}$$

Sinkhorn has **monotonic improvements**, thanks to $\varphi^{c_\varepsilon c_\varepsilon} \ge \varphi$, and the fact that $\varphi^{c_\varepsilon c_\varepsilon c_\varepsilon} \ne \varphi^{c_\varepsilon}$ (so that the improvement continues). One can prove that whenever the improvement stops, both potentials must reach optimal simultaneously.

The details of Sinkhorn algorithm is provided in Alg. 2.

Algorithm 2 Sinkhorn

```
Require: $\varepsilon > 0$
repeat
    $\varphi^{n+1} = (\psi^n)^{c_\varepsilon} = \arg\max_\varphi D(\varphi, \psi^n)$
    $\psi^{n+1} = (\varphi^{n+1})^{c_\varepsilon} = \arg\max_\psi D(\varphi^{n+1}, \psi)$
until Enough iterations are done
```

When it comes to numerical implementation, we do not approximate φ, ψ but approximate

$$\eta(x) := \mu(x) e^{\frac{1}{\varepsilon}\varphi(x)}, \quad \xi(y) := \nu(y) e^{\frac{1}{\varepsilon}\psi(y)} \tag{162}$$

instead. This greatly simplifies the calculation of the c_ε-transform: if $\psi^{n+1} = (\varphi^{n+1})^{c_\varepsilon}$, then

$$\eta^{n+1}(x) = \frac{\mu(x)}{\int \xi^n(y) e^{-\frac{1}{\varepsilon}c(x,y)}\, dy}, \quad \xi^{n+1}(y) = \frac{\nu(y)}{\int \eta^{n+1}(x) e^{-\frac{1}{\varepsilon}c(x,y)}\, dx}. \tag{163}$$

The denominators are actually convolutions, which are just matrix products in the discrete case.

If μ, ν are empirical measures supported on n, m points respectively, then $\mu, \eta \in \mathbb{R}^n, \nu, \xi \in \mathbb{R}^m$. In this case, we denote $K \in \mathbb{R}^{n\times m}$ such that $K_{ij} := e^{-\frac{1}{\varepsilon}c(x_i, y_j)}$. The Sinkhorn updates reduce to

$$\eta^{n+1} = \frac{\mu}{K\xi^n}, \quad \xi^{n+1} = \frac{\nu}{K^T \eta^{n+1}}, \tag{164}$$

where the division is in the sense of component-wise division. After learning the optimal Kantorovich potentials, we have to go back to the primal EOT problem. Previous calculations imply that the **optimal coupling** is

$$\pi^*(x, y) = e^{\frac{1}{\varepsilon}[\varphi^*(x) + \psi^*(y) - c(x,y)]} \mu(x)\nu(y). \tag{165}$$

In the discrete case, this reduces to

$$\pi^* = \text{diag}(\eta^*) K \text{diag}(\xi^*). \tag{166}$$

Remark. *Sinkhorn can be generalized to solve MMOT (still the product measure as reference). The dual problem has k potentials given k marginals and one still updates one while fixing all the others.*

Applications of OT

Wasserstein Regression

The Wasserstein regression refers to the regression in the space of measures and it provides a way to deal with distributional data (each data point is a measure instead of a vector in the Euclidean space). To understand the power of Wasserstein distance, we first review the traditional transformation method of dealing with distributional data, which first extracts information from distributions, apply approximation tools on the extracted information, and finally go back to distributions.

There are many popular transformation methods, one of which is called the log-quantile transformation. For a given distribution, let f denote its density and Q_f denote its quantile. The transformation starts with

$$\Psi(f)(t) := -\log(f \circ Q_f(t)),\ t \in [0,1]. \tag{167}$$

Each distribution denoted by f is mapped to a function $\Psi(f) : [0,1] \to \mathbb{R}$. Next, we can apply traditional functional data approximation tools for $\Psi(f_1), ..., \Psi(f_n)$ to get the approximation $g : [0,1] \to \mathbb{R}$. Finally, we can go back to distributions by applying

$$\Psi^{-1}(g)(x) := \theta_g e^{-g\circ H_g^{-1}(x)}, \tag{168}$$

where

$$\theta_g := \int_0^1 e^{g(s)}\, ds < \infty, \quad H_g(t) := \theta_g^{-1} \int_0^t e^{g(s)}\, ds. \tag{169}$$

We skip the verifications here but it's easy to check that

$$\int \Psi^{-1}(g)(x)\, dx = \theta_g \int_0^1 e^{-g(y)}\, dH_g(y) = 1 \tag{170}$$

is a legal density.

Remark. *Consider as an example the classification of distributional data points, i.e. given densities $f_1, ..., f_n$, we hope to classify them into two classes. The space of density functions is not even a vector space, but the log-quantile transformed space is a Hilbert space, on which a decision boundary can be figured out. Applying Ψ^{-1} allows us to map the decision boundary in the space of log-quantile transformations to the decision boundary in the space of measures.*

Interesting questions include the stability of the decision boundary, e.g., when points in the log-quantile transformed space are close to the decision boundary, if the points in the space of measures are also close to the decision boundary.

Other examples of transformation methods include the score function and the Langevin dynamics for the purpose of learning a measure and sampling from the measure numerically, etc.

With tools from OT, however, we are able to conduct regression or classification in the native space under the Wasserstein geometry. In Euclidean space, the linear regression problem is a problem to learn $\mathbb{E}(Y|X=x)$ given observations $(X_1,Y_1),...,(X_n,Y_n)$ from the model $y=\beta_0+\beta_1 x+\varepsilon$, where ε is the noise that follows some distribution.

As an analogue, we can formalize the Wasserstein regression problem as observing noisy observations (tuples of measures) $(\mu_1,\nu_1),...,(\mu_n,\nu_n)$ and hoping to recover the relationship between μ and ν. In the ideal general setting without noises, the model is $(\mu,\nu)\sim\mathbb{P}$ for $\mathbb{P}\in\mathscr{P}(\mathscr{P}(\mathbb{R}^d)\times\mathscr{P}(\mathbb{R}^d))$ (a probability distribution on the space of tuples of measures). However, the structure defined by a distribution $\mathbb{P}$ is too general to model, and we hope that some special connections between μ and ν can be assumed. Inspired by the OT problem, we would assume the noiseless model to be

$$\nu=(T_0)_{\#}\mu \tag{171}$$

for some increasing function $T_0:\mathbb{R}^d\to\mathbb{R}^d$. This assumption presents a clever trade-off between generality and interpretability. The pushforward is a lifting of a mapping on the underlying space to the space of measure so it can be interpreted as "ν is a function of μ". When it comes to adding noises to the model, we still use the pushforward, but the pushforward of a random function instead of a deterministic one. The noisy model is given by

$$\nu=(T_\varepsilon)_{\#}(T_0)_{\#}\mu, \tag{172}$$

where ε is a r.v. such that

$$\forall x,\mathbb{E}T_\varepsilon(x)=x. \tag{173}$$

For example, $T_\varepsilon(x)=x+\varepsilon,\varepsilon\sim N(0,\sigma^2)$ is a legal source of the noise. By Brenier's theorem, we restrict to T_0,T_ε as gradients of convex functions (increasing).

After putting up the model, we try to define the conditional expectation $\mathbb{E}(\nu|\mu)$ for $(\mu,\nu)\sim\mathbb{P}$. The main difficulty here is to understand what it means to average a distribution on the space of measures. Recall that the conventional conditional expectation is an orthogonal projection, we can simply replace the L^2 distance with the 2-Wasserstein distance to provide the definition:

$$\forall(\mu,\nu)\sim\mathbb{P},\mathbb{E}(\nu|\mu):=\operatorname*{arg\,min}_{b=T_{\#}\mu}\int W_2^2(\nu,b)\,d\mathbb{P}(\nu|\mu). \tag{174}$$

The definition guarantees "measurability" by requiring $\mathbb{E}(\nu|\mu)$ to be a "function" of μ. If we combine this with the noisy model above, then without any surprise,

$$\forall\nu=(T_\varepsilon)_{\#}(T_0)_{\#}\mu,\ \mathbb{E}(\nu|\mu)=(T_0)_{\#}\mu. \tag{175}$$

At this point, we can formulate the **Wasserstein regression** problem as: given observations $(\mu_1,\nu_1),...,(\mu_n,\nu_n)$ such that $\nu_i=(T_{\varepsilon_i})_{\#}(T_0)_{\#}\mu_i$ for *i.i.d.* r.v. ε_i, we want to learn T_0 from the data.

One natural way to solve this problem is to adopt the least square criterion that

$$M_n(T) := \frac{1}{n}\sum_{i=1}^{n} W_2^2(T_{\#}\mu_i, \nu_i). \tag{176}$$

This quantity works as the mean square loss of T in traditional linear regression, and we aim to find a $\hat{T}$ that minimizes this loss as the least square estimator.

Remark. *When $d = 1$, the Wasserstein distance has a simple formula so the problem reduces to minimizing*

$$M_n(T) := \frac{1}{n}\sum_{i=1}^{n} \|T - Q_{\nu_i} \circ F_{\mu_i}\|_{L^2(\mu_i)}^2, \tag{177}$$

which is quadratic in T. This is a convex optimization problem that can be easily solved.

Gradient Flow and JKO Scheme

For given $F : \mathbb{R}^d \to \mathbb{R}$, we call $X(t)$ a **gradient flow** of F if it's always travelling in the direction of steepest descent, i.e.,

$$X'(t) = -\nabla F(X(t)). \tag{178}$$

Such gradient flow is describing the motion of one particle. With a lot of particles, we need to add a space variable x such that $X(t,x)$ denotes the time t location of the particle that starts at x. All particles move in the direction of $-\nabla F$ and the particles have density $\mu(t,x)$ at time t. Clearly,

$$\partial_t X(t,x) = -\nabla F(X(t,x)), \quad \mu(t,\cdot) = [X(t,\cdot)]_{\#}\mu(0,\cdot). \tag{179}$$

We want to figure out the time evolution of μ.

For any ψ, consider

$$\begin{aligned}
\int \partial_t \mu(t,x)\psi(x)\,dx = \partial_t \int \mu(t,x)\psi(x)\,dx = \partial_t \int \mu(0,x)\psi(X(t,x))\,dx && (180)\\
= \int \mu(0,x)\nabla\psi(X(t,x)) \cdot \partial_t X(t,x)\,dx && (181)\\
= -\int \mu(0,x)\nabla\psi(X(t,x)) \cdot \nabla F(X(t,x))\,dx && (182)\\
= -\int \mu(t,x)\nabla\psi(x) \cdot \nabla F(x)\,dx. && (183)
\end{aligned}$$

Apply integration by parts to get

$$-\int \mu(t,x)\nabla\psi(x) \cdot \nabla F(x)\,dx = \int \psi(x)\mathrm{div}_x[\mu(t,x)\nabla F(x)]\,dx. \tag{184}$$

It follows that

$$\partial_t \mu - \mathrm{div}_x(\mu\nabla F) = 0. \tag{185}$$

This is called the **continuity equation** and it's the same as the Fokker-Planck equation for a deterministic diffusion with drift $-\nabla F$. It's obvious that if $\partial_t X(t,x) = v(X(t,x))$, the particles move according to velocity field v, then the continuity equation becomes

$$\partial_t \mu + \mathrm{div}_x(\mu v) = 0. \tag{186}$$

Now let's consider approximating the gradient flow numerically. We start with considering a single particle and then generalize it to the case of infinitely many particles. Most naturally, we use gradient descent and choose a small time step τ for the time discretization. The gradient flow can be approximated by

$$\frac{X((n+1)\tau,x)-X(n\tau,x)}{\tau}=-\nabla F(X(n\tau,x)). \tag{187}$$

This is the forward Euler scheme of approximating gradient flow but it suffers from the issue of selecting the stepsize τ. A small τ would result in small progress while a large τ might even take the particle to the place where the value of F is larger than the original. A remedy is to use the implicit backward Euler scheme instead, given by

$$\frac{X((n+1)\tau,x)-X(n\tau,x)}{\tau}=-\nabla F(X((n+1)\tau,x)). \tag{188}$$

The implicit scheme is known to be more stable, but as a cost to pay we have to solve an equation in $X((n+1)\tau,x)$ to determine the update. We rewrite the equation as

$$(id+\tau\nabla F)[X((n+1)\tau,x)]=X(n\tau,x). \tag{189}$$

Notice that $id+\tau\nabla F$ can be written as the gradient of $G_\tau(x):=\frac{1}{2}|x|^2+\tau F$. Use the property that $(\nabla G_\tau)^{-1}=\nabla G_\tau^*$ to get (star means Frenchel conjugate)

$$X((n+1)\tau,x)=\nabla G_\tau^*[X(n\tau,x)]. \tag{190}$$

Notice that ∇G_τ^* is the gradient of a convex function, which reminds us of Brenier's theorem, which shows an underlying connection with OT. Typically, ∇G_τ^* is hard to calculate numerically. Instead, we solve the following optimization problem:

$$\min_y\left\{F(y)+\frac{1}{2\tau}|y-X(n\tau,x)|^2\right\}. \tag{191}$$

Taking the derivative w.r.t. y, we see that the optimal y satisfies

$$\nabla F(y)+\frac{1}{\tau}(y-X(n\tau,x))=0. \tag{192}$$

Such optimal y exactly matches $X((n+1)\tau,x)$ in the backward Euler scheme, which is called the **one-particle JKO** scheme.

Remark. *The term $\frac{1}{2\tau}|y-X(n\tau,x)|^2$ is the proximal operator that penalizes the next location $X((n+1)\tau,x)$ for being far away from the former location $X(n\tau,x)$. In the context of gradient flows, the proximal operator has a natural interpretation from the backward Euler scheme.*

For infinitely many particles, we generalize the one-particle scheme on knowing that each particle shall follow the one-particle JKO scheme. Instead of focusing on the locations $X((n+1)\tau,x)$ for infinitely many x, we shall instead focus on solving the density $\mu((n+1)\tau,x)$ given $\mu(n\tau,x)$ to describe the time evolution of the population.

The one-particle JKO scheme applies for $X(n\tau, x)$:

$$\forall x \in \mathbb{R}^d, X((n+1)\tau, x) = \underset{T:\mathbb{R}^d\to\mathbb{R}^d}{\arg\min} \left\{ F(T(x)) + \frac{1}{2\tau}|T(x) - X(n\tau, x)|^2 \right\}. \tag{193}$$

Such $T(x)$ is actually just a mapping that maps each location x to the corresponding minimizer in the one-particle JKO scheme. Recall the interpretation of $X(n\tau, x)$ as the time $n\tau$ location of the particle that starts from initial location x. We understand x as the "sample point" in probability theory so that $X(n\tau, \cdot)$ is a random variable that describes all particles' locations at time $n\tau$, which follows $\mu(n\tau, \cdot)$. In addition, with the initial location x as the "sample point", the probability measure on the sample space $\Omega = \mathbb{R}^d$ is given by $\mu(0, \cdot)$. We yield the JKO scheme for all particles:

$$X((n+1)\tau, \cdot) = \underset{Y=T(X_0)}{\arg\min}\, \mathbb{E}_{X_0\sim\mu(0,\cdot)} \left(F(Y) + \frac{1}{2\tau}|Y - X(n\tau, \cdot)|^2 \right). \tag{194}$$

Here T is actually a transport map from $\mu(0, \cdot)$ to $\mu((n+1)\tau, \cdot)$. Given the results from the previous iteration $\mu(n\tau, \cdot)$, we wish to assume $Y = T(X(n\tau, \cdot))$ so that $\mu((n+1)\tau, \cdot)$ only depends on $\mu(n\tau, \cdot)$. In this sense, we can rewrite the problem as

$$\min_{Y=T(X(n\tau,\cdot))} \int \left(F(Y) + \frac{1}{2\tau}|Y - X(n\tau, \cdot)|^2 \right) \mu(0, x)\, dx = \min_T \int \left(F(T(x)) + \frac{1}{2\tau}|T(x) - x|^2 \right) \mu(n\tau, x)\, dx, \tag{195}$$

by absorbing the pushforward by $X(n\tau, \cdot)$. Now $T : \mathbb{R}^d \to \mathbb{R}^d$ denotes a transport map from $\mu(n\tau, \cdot)$ (known) to $\mu((n+1)\tau, \cdot)$ (unknown). This problem is similar to a Monge problem with cost $c(x, T(x)) = F(T(x)) + \frac{1}{2\tau}|T(x)-x|^2$, with the difference that now we also have the freedom to choose the target measure (different from OT which is supervised learning). Denote $\mu := T_\#\mu(n\tau, \cdot)$ to rewrite it as an optimization problem in terms of the measure:

$$\min_\mu \min_{T:T_\#\mu(n\tau,\cdot)=\mu} \int \left(F(T(x)) + \frac{1}{2\tau}|T(x) - x|^2 \right) \mu(n\tau, x)\, dx \tag{196}$$

$$= \min_\mu \left\{ \int F(x)\mu(x)\, dx + \frac{1}{2\tau} \min_{T:T_\#\mu(n\tau,\cdot)=\mu} \int |T(x) - x|^2 \mu(n\tau, x)\, dx \right\} \tag{197}$$

$$= \min_\mu \left\{ \int F(x)\mu(x)\, dx + \frac{1}{2\tau} W_2^2(\mu, \mu(n\tau, \cdot)) \right\}. \tag{198}$$

The optimizer provides $\mu((n+1)\tau, \cdot)$ and this is referred to as the **JKO scheme** in general. The expression is surprisingly simple, with the Euclidean distance in the proximal operator replaced by the Wasserstein distance.

Remark. *All the particles are moving in the direction of the negative gradient of F, so $\int F(x)\mu(x)\, dx$ is the natural objective to minimize. However, only minimizing $\int F(x)\mu(x)\, dx$ aligns with the forward Euler scheme, which leads to the stepsize issue.*

In the context of RL, if we understand $\int F(x)\mu(x)\, dx$ as the expected cost following a certain policy and replace the Wasserstein distance with the KL-divergence, we get the TRPO algorithm with the same philosophy that we shall not allow too much change in the measure so that we do not make unrecoverable catastrophic mistake when exploring.

Benamou-Brenier Formula

Recall the discussion on the continuity equation, if the diffusion of particles is not time-homogeneous, i.e., $\partial_t X(t,x) = v(t, X(t,x))$ with the drift coefficient v to be depending on time t, then the continuity equation becomes $\partial_t \mu(t,x) + \text{div}_x(\mu(t,x)v(t,x)) = 0$. For simplicity, we denote $v_t(\cdot) := v(t,\cdot)$, $\mu_t(\cdot) := \mu(t,\cdot)$ and the sample space $\Omega \subset \mathbb{R}^d$ as the collection of all possible values of x (for the generality of the normal vector on the boundary). If the pair (μ_t, v_t) solves the continuity equation with the no-flux boundary condition

$$\mu_t v_t \cdot n|_{\partial\Omega} = 0, \tag{199}$$

then the **action of the pair** (μ_t, v_t) is defined as

$$A[\mu_t, v_t] := \int_0^1 \int |v_t(x)|^2 \mu_t(x)\, dx\, dt. \tag{200}$$

Remark. *The no-flux boundary condition guarantees the conservation of mass. By the divergence theorem,*

$$\partial_t \int_\Omega \mu_t(x)\, dx = -\int_\Omega \text{div}_x(\mu_t(x)v_t(x))\, dx = \int_{\partial\Omega} \mu_t v_t \cdot n\, dS = 0. \tag{201}$$

Hence μ_t is always a legal density function and the boundary condition is natural.

It's crucial to understand the action of the density-velocity pair. Clearly the term $\int |v_t(x)|^2 \mu_t(x)\, dx$ is accumulating the product of the local mass and the local velocity square, which is proportional to the kinetic energy. Integrated w.r.t. time, we get the cumulative kinetic energy on the time horizon $[0,1]$ induced by the time evolution of mass (described by μ_t) and the time evolution of the velocity field (described by v_t).

Theorem 8 (Benamou-Brenier)**.** *For any μ, ν,*

$$W_2^2(\mu,\nu) = \inf\left\{A[\mu_t, v_t] : \mu_0 = \mu, \mu_1 = \nu, \partial_t \mu_t + \text{div}_x(\mu_t v_t) = 0, \mu_t v_t \cdot n|_{\partial\Omega} = 0\right\}. \tag{202}$$

Proof. Consider the location $X_t(x)$ induced by the diffusion $\partial_t X_t(x) = v_t(X_t(x))$. Clearly $\mu_t = (X_t)_\# \mu_0$ and X_1 is a legal transport map from μ to ν. Using this fact,

$$A[\mu_t, v_t] = \int_0^1 \int |v_t(X_t(x))|^2 \mu_0(x)\, dx\, dt = \int_0^1 \int |\partial_t X_t(x)|^2 \mu_0(x)\, dx\, dt \tag{203}$$

$$= \int \mu_0(x) \int_0^1 |\partial_t X_t(x)|^2\, dt\, dx \geq \int \mu_0(x) \left| \int_0^1 \partial_t X_t(x)\, dt \right|^2 dx \tag{204}$$

$$= \int \mu_0(x) \left|X_1(x) - x\right|^2\, dx \geq W_2^2(\mu,\nu). \tag{205}$$

The first inequality follows from Cauchy-Schwarz and the second from the definition of the Wasserstein distance.

It suffices to construct X that attains the infimum. By Brenier's theorem, when cost is $c(x,y) = |x-y|^2$, the optimal transport map T exists and is unique, and $T = \nabla\phi$ for some convex ϕ. Naturally, we take the geodesic

$X_t(x) = tT(x) + (1-t)x$ such that X_1 coincides with T. In this case, $\mu_t = (X_t)_\# \mu_0$ is induced by X_t, and v_t is the one that solves $\partial_t X_t(x) = v_t(X_t(x))$. The existence of such v_t is guaranteed by the temporal smoothness of μ_t (μ_t is continuous and differentiable in t). The continuity equation and the no flux boundary condition naturally holds when such v_t exists. It remains to check that all inequalities in the calculation of the action are actually equalities. The last equality follows from the optimality of T and

$$\left|\int_0^1 \partial_t X_t(x)\, dt\right|^2 = \left|\int_0^1 (T(x) - x)\, dt\right|^2 = |T(x) - x|^2 = \int_0^1 |\partial_t X_t(x)|^2\, dt. \tag{206}$$

This concludes the proof. □

The Benamou-Brenier formula provides another characterization of the W_2 distance (the optimal transport cost) as the **minimum amount of kinetic energy** required to interpolate μ with ν under some velocity field. Actually, the formula also holds for general W_p distance (with a similar proof), but the definition of the action will then depend on p so one loses the nice interpretation of the optimal transport cost as the kinetic energy.

Example: Heat Equation as Gradient Flow

In the language of the diffusion, gradient flow is actually the special case in which the velocity field is the negative gradient of some potential, i.e., $v = -\nabla F$.

Let's consider a special F given by

$$F(u) = \frac{1}{2}\int |\nabla u|^2 \, dx, \tag{207}$$

where $u = u(t, x) \in C^{1,2}$. We use the notation $u(t, x)$ instead of $X(t, x)$ to focus on the PDE context rather than the particle physics context. Clearly, the gradient of F is the gradient w.r.t. u, which is a function, so we calculate the first variation.

$$\begin{aligned} \delta F(u)(\psi) &= \lim_{\varepsilon \to 0} \frac{F(u + \varepsilon\psi) - F(u)}{\varepsilon} &(208)\\ &= \int \nabla u(x) \cdot \nabla \psi(x) \, dx. &(209) \end{aligned}$$

When writing the gradient according to the first variation, we have to specify a geometry. Without extra information, it's natural to use the geometry induced by the inner product on the Hilbert space $\mathscr{H} = L^2(\mathbb{R}^d)$, which defines the gradient $\nabla_{L^2} F(u)$. By integration by parts,

$$\delta F(u)(\psi) = -\int \Delta u(x)\psi(x) \, dx = \left\langle \nabla_{L^2} F(u), \psi \right\rangle_{L^2}, \tag{210}$$

which implies

$$\nabla_{L^2} F(u) = -\Delta u. \tag{211}$$

The gradient flow induced by such a potential F is exactly the heat equation

$$\partial_t u = \Delta u. \tag{212}$$

Remark. *If one tries to calculate the second variation, it turns out that*

$$\delta^2 F(u)(\psi)(\psi) = \int |\nabla \psi(x)|^2 \, dx \geq 0. \tag{213}$$

The potential F is convex in u.

Wasserstein Gradient Flow

One crucial insight of the Benamou-Brenier formula is the representation of the Wasserstein distance

$$W_2^2(\mu,\nu) = \inf_{\mu_t} \left\{ \int_0^1 \|\partial_t \mu_t\|_{\mu_t}^2 \, dt : \mu_0 = \mu, \mu_1 = \nu \right\}, \tag{214}$$

where

$$\|\partial_t \mu_t\|_{\mu_t}^2 := \inf_{v_t} \left\{ \int_\Omega |v_t|^2 \mu_t \, dx : \partial_t \mu_t + \mathrm{div}_x(\mu_t v_t) = 0, \mu_t v_t \cdot n|_{\partial\Omega} = 0 \right\}, \tag{215}$$

where the infimum w.r.t. μ_t and v_t are written in a specific order. Actually, there is a reason we are doing this: the 2-Wasserstein distance now looks like a metric on a Riemannian manifold (connecting μ and ν using a curve μ_t, the metric is the infimum of the integral along some norm $\|\cdot\|_{\mu_t}$ of the tangent vector $\partial_t \mu_t$), which is defined through the solution to a variational problem, minimizing the energy of the curve. At this point, the formula shows us the interpretation of the Wasserstein space as a Riemannian manifold.

We are satisfied with the representation of W_2^2 but not the one for $\|\cdot\|_{\mu_t}$, which comes from an optimization problem in v_t. For given μ_t, if v_t is the minimizer under constraints $\partial_t \mu_t + \mathrm{div}_x(\mu_t v_t) = 0, \mu_t v_t \cdot n|_{\partial\Omega} = 0$, the perturbed version $\tilde{v}_t = v_t + \varepsilon \frac{w}{\mu_t}$ is also feasible given that $\mathrm{div}_x(w) = 0, w \cdot n|_{\partial\Omega} = 0$. As a result,

$$\int_\Omega |v_t|^2 \mu_t \, dx \leq \int_\Omega |\tilde{v}_t|^2 \mu_t \, dx. \tag{216}$$

Set $\varepsilon \to 0$ to see that

$$\int_\Omega v_t(x) \cdot w(x) \, dx \geq 0. \tag{217}$$

Setting w as $-w$ ($\tilde{v}_t$ still feasible) yields

$$\int_\Omega v_t(x) \cdot w(x) \, dx = 0. \tag{218}$$

The perturbation tells us that the optimal v_t lies in the space

$$V = \{w : \mathrm{div}_x(w) = 0, w \cdot n|_{\partial\Omega} = 0\}^\perp. \tag{219}$$

Since w is divergence-free, its Helmholtz decomposition only contains the gradient term.

$$V = \{\nabla \psi | \psi : \Omega \to \mathbb{R}\}. \tag{220}$$

Clearly, the representation of the minimizer $v_t = \nabla \psi_t$ results in

$$\|\partial_t \mu_t\|_{\mu_t}^2 = \int_\Omega |\nabla \psi_t|^2 \mu_t \, dx, \tag{221}$$

where ψ_t is the solution to the PDE with zero Neumann boundary conditions:

$$\begin{cases} \partial_t \mu_t + \operatorname{div}_x(\mu_t \nabla \psi_t) = 0 \\ \mu_t \nabla \psi_t \cdot n|_{\partial\Omega} = 0 \end{cases} . \tag{222}$$

So far, we have stated a simpler representation of $\|\partial_t \mu_t\|_{\mu_t}^2$, which has something to do with the PDE above.

This directly motivates the following definition as the **Wasserstein inner product**: for $f, g : \Omega \to \mathbb{R}$ such that $\int_\Omega f = \int_\Omega g = 0$, the Wasserstein inner product at measure μ is defined as

$$\langle f, g \rangle_\mu := \int_\Omega \nabla \psi_f(x) \cdot \nabla \psi_g(x) \; \mu(x) \, dx, \tag{223}$$

where the potentials ψ_f, ψ_g are determined through solving the PDE:

$$\begin{cases} f + \operatorname{div}_x(\mu \nabla \psi_f) = 0 \\ \mu \nabla \psi_f \cdot n|_{\partial\Omega} = 0 \end{cases} . \tag{224}$$

Remark. *Intuitively speaking, we treat f, g as $\partial_t \mu_t$ so the conditions $\int_\Omega f = \int_\Omega g = 0$ are necessary since $\int f(x)\, dx = \partial_t \int \mu_t(x)\, dx = \partial_t 1 = 0$.*

With the geometry on W_2 to be induced by an inner product, we can finally define gradients on the Wasserstein space. For a given function $J : \mathscr{P}(\Omega) \to \mathbb{R}$, denote its first variation at μ as $\delta J(\mu)(\eta)$, then

$$\langle \nabla_{W_2} J(\mu), \eta \rangle_\mu = \delta J(\mu)(\eta) \tag{225}$$

defines the Wasserstein gradient $\nabla_{W_2} J(\mu)$.

The following lemma calculates the Wasserstein gradient for a family of important examples of J.

Lemma 9. *If $J(\mu) = \int U(\mu(x))\, dx$ for some $U : \mathbb{R} \to \mathbb{R}$ where $\mu(x)$ is the density function of μ, then*

$$\nabla_{W_2} J(\mu) = -\operatorname{div}_x(\mu U''(\mu) \nabla \mu). \tag{226}$$

Proof. Since $\Omega = \mathbb{R}$ has no boundary, all boundary terms vanish. Calculate the first variation

$$\delta J(\mu)(\eta) = \int U'(\mu(x)) \eta(x) \, dx. \tag{227}$$

Then let's match with the Wasserstein inner product. By intergation by parts,

$$\int U'(\mu(x))\eta(x)\,dx = \int \nabla\psi_J(x)\nabla\psi_\eta(x)\;\mu(x)\,dx = -\int \psi_J\;\mathrm{div}_x(\nabla\psi_\eta\mu)\,dx \tag{228}$$

where

$$\nabla_{W_2}J(\mu) + \mathrm{div}_x(\mu\nabla\psi_J) = 0, \tag{229}$$

$$\eta + \mathrm{div}_x(\mu\nabla\psi_\eta) = 0. \tag{230}$$

Plug in to get

$$\int U'(\mu(x))\eta(x)\,dx = \int \psi_J\eta\,dx, \quad \psi_J = U'\circ\mu. \tag{231}$$

Finally, we use again the PDE to get

$$\nabla_{W_2}J(\mu) = -\mathrm{div}_x(\mu\nabla\psi_J) = -\mathrm{div}_x(\mu U''(\mu)\nabla\mu). \tag{232}$$

□

Similar to the gradient flow on Euclidean spaces, we define the **Wasserstein gradient flow** induced by J as the PDE

$$\partial_t\mu_t = -\nabla_{W_2}J(\mu_t). \tag{233}$$

At this point it should be clear that the JKO scheme introduced in the previous context is actually a numerical algorithm for solving Wasserstein gradient flows.

It's crucial to realize that a lot of PDEs are actually Wasserstein gradient flows. For example, if we set $U(x) = x\log x$, then $J(\mu)$ is the negative entropy of μ. From simple calculations,

$$\nabla_{W_2}J(\mu) = -\mathrm{div}_x(\nabla\mu) = -\Delta\mu. \tag{234}$$

The Wasserstein gradient flow induced by $U(x) = x\log x$ (J as the negative entropy) is the heat equation with zero Neumann boundary condition. In this sense, the JKO scheme

$$\mu_{(n+1)\tau} = \arg\min_{\mu}\left\{J(\mu) + \frac{1}{2\tau}W_2^2(\mu,\mu_{n\tau})\right\} \tag{235}$$

provides a numerical algorithm solving the heat equation.

Remark. *PDEs like $\partial_t u = \Delta(u^m)$ are called porous medium equation/fast diffusion equation based on the value of m. This family of PDEs are Wasserstein gradient flows with $U(x) = \frac{x^m}{m-1}$ so they can also be solved through the JKO scheme. The key lies in recognizing that a certain PDE is a Wasserstein gradient flow.*

Remark. *The linear functional derivative w.r.t. a measure* $\frac{\delta J(\mu)}{\delta\mu}(x)$ *is defined as the one such that*

$$\delta J(\mu)(\eta) =: \left\langle \frac{\delta J(\mu)}{\delta\mu}, \eta \right\rangle_{L^2}. \tag{236}$$

In the example above, it's clear that

$$\frac{\delta J(\mu)}{\delta\mu} = U' \circ \mu. \tag{237}$$

If J *is the negative entropy, then*

$$\frac{\delta J(\mu)}{\delta\mu}(x) = \log\mu(x) + 1. \tag{238}$$

The L-derivative is given by

$$\partial_\mu J(\mu)(x) = \nabla_x \frac{\delta J(\mu)}{\delta\mu}(x) = \nabla\log\mu(x), \tag{239}$$

which is the score function. The Wasserstein gradient flow provides another way to formalize the derivative w.r.t. a measure, and is consistent with the L-derivative.

Adversarial Training: an example in Linear Regression

Neural networks, even if restricted to the supervised learning tasks, are still vulnerable to adversarial attacks. For example, if one has a trained image classification model and adds small noise to a picture in a certain category, human eyes cannot distinguish the difference while it might greatly perturb the classification and provides ridiculous predictions. Surprisingly, those adversarial attacks turn out to be transferrable in the sense that the same methodology would work for models trained for different purposes. That means, adversarial training is necessary to improve the robustness of the models.

Denote x, X as features, y, Y as labels, θ as the model parameter and l as the loss function. The original training task can be described as

$$\min_{\theta} \mathbb{E}_{(X,Y)\sim\mu} l((X,Y),\theta). \tag{240}$$

Here μ is typically a measure on a low-dimensional manifold in a high-dimensional space and the loss measures the accuracy of the model prediction compared to the true label. When taking into account the robustness, one does the **distributionally robust optimization (DRO)**

$$\min_{\theta} \sup_{\tilde{\mu}:D(\mu,\tilde{\mu})\leq\varepsilon} \mathbb{E}_{\tilde{Z}\sim\tilde{\mu}} l(\tilde{Z},\theta), \tag{241}$$

where $z = (x, y)$ denotes the concatenation of the feature and the label. The adversarial first perturbs the distribution μ a little bit and then we train the model to guarantee the worst-case performance.

Let's first check a toy example with linear regression. We specify the following setting

$$z = (x,y) \in \mathbb{R}^{d-1}\times\mathbb{R}, \quad \theta\in\mathbb{R}^{d-1}, \quad l(\theta,z) = |y-\theta^T x|^2, \quad D = W_2^2. \tag{242}$$

Firstly, start with the supremum in DRO

$$\sup_{\tilde{\mu}:W_2^2(\mu,\tilde{\mu})\leq\varepsilon} \mathbb{E}_{\tilde{Z}\sim\tilde{\mu}} l(\tilde{Z},\theta) = \sup_{\pi:(P_1)_\#\pi=\mu, \int |z-\tilde{z}|^2\,d\pi(z,\tilde{z})\leq\varepsilon} \int l(\tilde{z},\theta)\,d\pi(z,\tilde{z}). \tag{243}$$

This is a constrained optimization problem. We leave the constraint $(P_1)_\#\pi = \mu$ unchanged while turning $\int |z-\tilde{z}|^2\,d\pi(z,\tilde{z}) \leq \varepsilon$ into the dual. Write down the Langrangian for $\beta \geq 0$:

$$Q(\pi,\beta) = \int l(\tilde{z},\theta)\,d\pi(z,\tilde{z}) + \beta\left(\varepsilon - \int |z-\tilde{z}|^2\,d\pi(z,\tilde{z})\right). \tag{244}$$

Calculate the dual objective

$$\sup_{\pi:(P_1)_\#\pi=\mu} Q(\pi,\beta) = \sup_{\pi:(P_1)_\#\pi=\mu} \left\{\int [l(\tilde{z},\theta) - \beta|z-\tilde{z}|^2]\,d\pi(z,\tilde{z}) + \beta\varepsilon\right\}. \tag{245}$$

Here the coupling π starts from μ but can send the masses to anywhere. Maximizing w.r.t. such π is equivalent to maximizing pointwisely for each z, i.e., sending each z to the $\tilde{z}$ that maximizes the objective. For fixed z,

$$T(z) := \arg\max_{\tilde{z}} \left\{ l(\tilde{z}, \theta) - \beta |z - \tilde{z}|^2 \right\} \tag{246}$$

defines the best place we shall send z to. On the other hand,

$$\psi_\beta^\theta(z) := \max_{\tilde{z}} \left\{ l(\tilde{z}, \theta) - \beta |z - \tilde{z}|^2 \right\} \tag{247}$$

defines the best cost at z if we stick to the transport map T. As a result, the dual problem is

$$\inf_{\beta \geq 0} \left\{ \int \psi_\beta^\theta(z) \, d\mu(z) + \beta\varepsilon \right\} \tag{248}$$

and strong duality holds due to the convex-concave minimax theorem.

At this point, we rewrite **DRO dual** as

$$\min_\theta \inf_{\beta \geq 0} \left\{ \int \psi_\beta^\theta(z) \, d\mu(z) + \beta\varepsilon \right\}, \tag{249}$$

which is the best we can do for a general model. Here we use the specific setting of linear regression to get something interpretable. We explicitly calculate $T(z)$ and $\psi_\beta^\theta(z)$:

$$T(z) = (\tilde{x}, \tilde{y}), \quad \begin{cases} \tilde{y} - \theta^T \tilde{x} = \beta(\tilde{y} - y) \\ (\tilde{y} - \theta^T \tilde{x})\theta + \beta(\tilde{x} - x) = 0 \end{cases} \Rightarrow T(z) = \left(x - \frac{y - \theta^T x}{\beta - |\theta|^2 - 1}\theta, y + \frac{y - \theta^T x}{\beta - |\theta|^2 - 1} \right). \tag{250}$$

Hence,

$$\psi_\beta^\theta(z) = (y - \theta^T x)^2 \frac{\beta}{\beta - |\theta|^2 - 1}. \tag{251}$$

Finally, we solve

$$\inf_{\beta \geq 0} \left\{ \frac{\beta}{\beta - |\theta|^2 - 1} \int (y - \theta^T x)^2 \, d\mu(z) + \beta\varepsilon \right\} \tag{252}$$

to see the minimizer

$$\beta^* = 1 + |\theta|^2 + \sqrt{\frac{(|\theta|^2 + 1)\mathbb{E}_{Z \sim \mu} l(\theta, Z)}{\varepsilon}}. \tag{253}$$

The DRO problem for linear regression finally becomes:

$$\min_\theta \left\{ \sqrt{\mathbb{E}_{Z \sim \mu} l(\theta, Z)} + \sqrt{\varepsilon(1 + |\theta|^2)} \right\}. \tag{254}$$

Besides the risk of the original linear regression problem, the robust version also penalizes for a large $|\theta|$ (ℓ_2 regularization sense). Actually, for the purpose of simplicity, we are taking all norms to be the Euclidean norm. Instead, one can check that if the norm in the cost function $c(x,y) = \|x-y\|$ in the definition of the Wasserstein distance is taken as any norm, a similar conclusion still holds for the DRO problem as

$$\min_{\theta} \left\{ \sqrt{\mathbb{E}_{Z\sim\mu} l(\theta, Z)} + \sqrt{\varepsilon} \|(\theta, -1)\|_* \right\}, \tag{255}$$

where $\|\cdot\|_*$ is the dual norm. If we take $c(x,y) = \|x-y\|_\infty$ (W_∞ sense), then the DRO problem becomes

$$\min_{\theta} \left\{ \sqrt{\mathbb{E}_{Z\sim\mu} l(\theta, Z)} + \sqrt{\varepsilon} \|(\theta, -1)\|_1 \right\}, \tag{256}$$

which aligns with the square-root LASSO. From this perspective, we can interpret square-root LASSO as the robust version of linear regression against W_∞ adversarial attacks (which was not mentioned in the original paper).

Adversarial Attack: the Model-Agnostic Case

Now let's try to understand DRO in the model-agnostic case. In particular, we consider the abstract **non-parametric DRO** problem that does not even require knowledge on the form of the model:

$$\inf_{f\in\mathscr{F}}\sup_{\tilde{\mu}}\mathbb{E}_{\tilde{Z}\sim\tilde{\mu}}l(f,\tilde{Z})-C(\mu,\tilde{\mu}),\tag{257}$$

where $\mathscr{F}$ is a function space in which the model f lives. The loss function still depends on the true model f and the data $\tilde{Z}$ from the perturbed measure $\tilde{\mu}$. Instead of formalizing a constrained optimization, we put the constraints as regularization terms, requiring the perturbation in $\tilde{\mu}$ from μ to be not too crazy.

In a general classification problem, in particular, the data $z=(x,y)\in\mathbb{R}^d\times[k]$ is given, with the label y indicating the category of the feature x out of k possible categories. $\mathscr{F}$ is understood as the set of all measurable soft classifiers such that $\forall f\in\mathscr{F}$, $f=(f_1,...,f_k)$ where $f_l:\mathbb{R}^d\to[0,1]$ and $\sum_l f_l\equiv 1$. Intuitively speaking, $f_l(x)$ is the model predicted confidence level (probability) of the feature x belonging to category l, so the probabilities always add up to one. When it comes to the loss function, we take zero-one loss for simplicity:

$$l(f,z)=1-f_y(x).\tag{258}$$

The loss penalizes predictions far away from the true label. Ideally, a model shall output $f_y(x)=1$ since feature x belongs to category y. The penalty in the zero-one loss increases as the probability $f_y(x)$ decreases. Finally, $C(\mu,\tilde{\mu})$ is taken as:

$$C(\mu,\tilde{\mu}):=\inf_{\pi\in\Pi(\mu,\tilde{\mu})}\int C_z(z,\tilde{z})\,d\pi(z,\tilde{z}),\tag{259}$$

where

$$C_z(z,\tilde{z}):=c(x,\tilde{x})\mathbb{I}_{y=\tilde{y}}+\infty\mathbb{I}_{y\neq\tilde{y}}.\tag{260}$$

Note that the perturbations shall only happen on the features but not the labels. C is the optimal transport cost induced by C_Z, a "distance" between z and the perturbed $\tilde{z}$, which is again induced by the "distance" on the feature space denoted as c.

Remark. *In the context of image classification, given a picture of a cat, the adversary is allowed to change the image a little bit, but is not allowed to change the label of the picture from "cat" to something like "dog". This matches the problem setting in reality since the changes in the label are mostly crazy changes that largely confuses the model and are not considered small perturbations in the data.*

Since we don't allow perturbations in the labels, only the x-marginals of μ and $\tilde{\mu}$ are of our interests. Naturally, we define

$$\mu_i(\cdot):=\mu(\cdot\times\{i\}),\quad\tilde{\mu}_i(\cdot):=\tilde{\mu}(\cdot\times\{i\}),\quad C(\mu_i,\tilde{\mu}_i):=\inf_{\pi\in\Pi(\mu_i,\tilde{\mu}_i)}\int c(x_i,\tilde{x}_i)\,d\pi(x_i,\tilde{x}_i).\tag{261}$$

Here $C(\mu_i, \tilde{\mu}_i)$ is the OT cost from μ to $\tilde{\mu}$, restricted on label i. This implies an additive form of $C(\mu, \tilde{\mu})$ that

$$C(\mu, \tilde{\mu}) = \sum_{i=1}^{k} C(\mu_i, \tilde{\mu}_i). \tag{262}$$

At this point, by adding assumptions, we get the **DRO for classification problem**:

$$\inf_{f \in \mathscr{F}} \sup_{\tilde{\mu}_1, \dots, \tilde{\mu}_k} \mathbb{E}_{\tilde{Z} \sim \tilde{\mu}} l(f, \tilde{Z}) - \sum_{i=1}^{k} C(\mu_i, \tilde{\mu}_i). \tag{263}$$

In the following context, we perform calculations and show its connection with OT. Firstly, notice that strong duality holds (convex-concave structure), so the problem becomes

$$\inf_{\tilde{\mu}_1, \dots, \tilde{\mu}_k} \sup_{f \in \mathscr{F}} \mathbb{E}_{\tilde{Z} \sim \tilde{\mu}} f_{\tilde{Y}}(\tilde{X}) + \sum_{i=1}^{k} C(\mu_i, \tilde{\mu}_i) = \inf_{\tilde{\mu}_1, \dots, \tilde{\mu}_k} \sup_{f \in \mathscr{F}} \sum_{i=1}^{k} \int f_i(x)\, d\tilde{\mu}_i(x) + \sum_{i=1}^{k} C(\mu_i, \tilde{\mu}_i). \tag{264}$$

The supremum can be explicitly solved here. There exists a reference measure $\nu := \frac{\sum_{i=1}^{k} \tilde{\mu}_i}{k}$ for all $\tilde{\mu}_i$ such that

$$\sup_{f \in \mathscr{F}} \sum_{i=1}^{k} \int f_i(x)\, d\tilde{\mu}_i(x) = \sup_{f \in \mathscr{F}} \sum_{i=1}^{k} \int f_i(x) \frac{d\tilde{\mu}_i}{d\nu}(x)\, d\nu(x) = \int \sup_{f \in \mathscr{F}} \sum_{i=1}^{k} f_i(x) \frac{d\tilde{\mu}_i}{d\nu}(x)\, d\nu(x). \tag{265}$$

Clearly, since $f_i \geq 0, \sum_i f_i \equiv 1$ are weights, the supremum in $\sup_{f \in \mathscr{F}} \sum_{i=1}^{k} f_i(x) \frac{d\tilde{\mu}_i}{d\nu}(x)$ is attained when f puts all its probability mass on the component that attains $\max_{i \in [k]} \frac{d\tilde{\mu}_i}{d\nu}(x)$ pointwisely for each x. As a result,

$$\sup_{f \in \mathscr{F}} \sum_{i=1}^{k} \int f_i(x)\, d\tilde{\mu}_i(x) = \int \max_{i \in [k]} \left\{ \frac{d\tilde{\mu}_i}{d\nu}(x) \right\} d\nu(x). \tag{266}$$

Now we want to get rid of the reference measure ν. The best way is to introduce a new measure Λ such that $\forall i, \tilde{\mu}_i \leq \Lambda$, in the sense of measure domination on any measurable sets (NOT absolute continuity!).

$$\int \max_{i \in [k]} \left\{ \frac{d\tilde{\mu}_i}{d\nu}(x) \right\} d\nu(x) = \int \max_{i \in [k]} \left\{ \frac{d\tilde{\mu}_i}{d\nu}(x) \right\} \frac{d\nu}{d\Lambda}(x)\, d\Lambda(x) = \int \max_{i \in [k]} \left\{ \frac{d\tilde{\mu}_i}{d\Lambda}(x) \right\} d\Lambda(x). \tag{267}$$

Nicely, $\max_{i \in [k]} \left\{ \frac{d\tilde{\mu}_i}{d\Lambda}(x) \right\} \leq 1$, with the upper bound 1 to be tight, so

$$\int \max_{i \in [k]} \left\{ \frac{d\tilde{\mu}_i}{d\Lambda}(x) \right\} d\Lambda(x) = \inf_{\Lambda} \int 1\, d\Lambda(x) = \inf_{\Lambda} \Lambda(\mathscr{X}), \tag{268}$$

where $\mathscr{X}$ is the whole feature space from which x takes values.

The DRO for classification problem becomes the **generalized barycenter problem**:

$$\inf_{\tilde{\mu}_1,\ldots,\tilde{\mu}_k,\Lambda} \Lambda(\mathscr{X}) + \sum_{i=1}^{k} C(\mu_i,\tilde{\mu}_i) \tag{269}$$

$$s.t.\ \forall i, \tilde{\mu}_i \le \Lambda. \tag{270}$$

Recall that the typical barycenter problem has the form $\inf_\mu \sum_i C(\mu_i,\mu)$, minimizing the (weighted) sum of distances from each measure to the barycenter. However, in our problem, we minimize the total mass of the generalized barycenter Λ that dominates all x-marginals of the perturbed measure $\tilde{\mu}$. Intuitively speaking, we want to find the smallest Λ that covers all different parts of $\tilde{\mu}$ so we are comparing "some parts" of the Barycenter Λ with each $\tilde{\mu}_i$. Since the "some parts" of Λ varies as i varies, this is not the same as a conventional Wasserstein barycenter but a more broadly defined one.

It turns out that the generalized barycenter problem can be converted into an MMOT problem to be solved numerically. The MMOT has k marginals so one might expect to meet with practical difficulties when k is above like 20. Interestingly, this is NOT the case in practice. If the cost c is taken as the most natural one, e.g.,

$$c_\varepsilon(x,\tilde{x}) := \infty \mathbb{I}_{|x-\tilde{x}|>\varepsilon}. \tag{271}$$

Then for $A \subset [k]$ as a subset of labels, we might define

$$C_{A,\varepsilon}(\{x_i\}_{i\in A}) := \inf_{\tilde{x}} \sum_{i\in A} c_\varepsilon(x_i,\tilde{x}). \tag{272}$$

Such $C_{A,\varepsilon}(\{x_i\}_{i\in A})$ is finite (and zero) iff $\exists \tilde{x}$, such that $\forall i \in A$, $x_i \in B_\varepsilon(\tilde{x})$, i.e. when one can find a feature that is close to all features x_i that belong to the labels in A. In the example of image classification, if we take A as "animals", this is saying that we can find an image that looks similar to any pictures of animals. That certain image is probably also an image of animals that has labels in A.

As a result, when ε is not too big, it's typically hard to find some neighborhood that contains a large number of features with different labels, i.e. the size of A is not that large in practice. In this case, the number of marginals in MMOT can be reduced to the size of A instead of the large k, which remains numerically feasible to solve. In the image classification example, we have altogether $k = 100$ labels. If the criterion of "similar" is not too trivial, let's say we get 50 different values of A, e.g., "animals", "human beings", "scenery", etc. However, within each A like "animals", there are not too many different labels, e.g., only "cat" and "dog". Within "human beings", e.g., there are only "elderly", "adult" and "kid". Within "scenery", e.g., there are only "city", "rural". As a result, the MMOT on $k = 100$ marginals (infeasible) can be decomposed into 50 smaller MMOT problems within each value of A, having only two or three marginals (feasible).

Remark. *When solving MMOT in practice, the wisdom is to find special structures to reduce the number of marginals!*

Parameter Inference and Denoising

Let's consider another application of OT in statistics. We first propose the problem within the framework of decision theory. Consider parameter θ taking values in the parameter space Θ in the Bayesian setting, i.e., $\theta \sim \pi$ follows the prior distribution π. We observe samples generated by the model (likelihood) $Z|_\theta \sim p_\theta(\cdot)$ and the goal is to find a decision rule $\delta(z)$ that maps each observation z_i to its parameter θ_i, in the optimal sense, that minimizes the mean square risk

$$\min_\delta \mathbb{E}_{(\theta,Z)}|\delta(Z)-\theta|^2. \tag{273}$$

From statistics, the minimizer must be the posterior mean

$$\delta^*(Z) = \overline{\theta}(Z) := \mathbb{E}(\theta|Z). \tag{274}$$

However, calculating the posterior requires knowledge of the prior, and in most cases we don't have prior knowledge. In this case, we are assuming that there exists an **underlying unknown prior** π.

Nicely, if the distributions belong to the exponential family, then the posterior mean can be directly computed from the marginal of Z without even knowing the prior. Otherwise, one has to use empirical Bayes methods in statistics to solve this problem. In some cases, however, an issue called overshrinking appears, i.e., $(\overline{\theta})_{\#}\mu_Z \neq \pi$. For example, consider prior π to be the uniform distribution on the unit circle C, and $p_\theta(\cdot)$ to be the two-dimensional Gaussian $N(\theta, \sigma^2 I)$. Simple calculations show that

$$p(\theta|z) \propto \mathbb{I}_{\theta\in C} \cdot e^{-\frac{|z-\theta|^2}{2\sigma^2}}. \tag{275}$$

The posterior mean can be represented by integrals along curves:

$$\overline{\theta}(z) = \frac{\int_C \theta e^{-\frac{|z-\theta|^2}{2\sigma^2}}\,d\theta}{\int_C e^{-\frac{|z-\theta|^2}{2\sigma^2}}\,d\theta} = \left(\frac{\int_0^{2\pi} e^{-\frac{|z-(\cos\eta,\sin\eta)|^2}{2\sigma^2}}\cos\eta\,d\eta}{\int_0^{2\pi} e^{-\frac{|z-(\cos\eta,\sin\eta)|^2}{2\sigma^2}}\,d\eta}, \frac{\int_0^{2\pi} e^{-\frac{|z-(\cos\eta,\sin\eta)|^2}{2\sigma^2}}\sin\eta\,d\eta}{\int_0^{2\pi} e^{-\frac{|z-(\cos\eta,\sin\eta)|^2}{2\sigma^2}}\,d\eta}\right). \tag{276}$$

Writing as double integrals,

$$|\overline{\theta}(z)|^2 = \frac{\int_0^{2\pi}\int_0^{2\pi} e^{-\frac{|z-(\cos\eta,\sin\eta)|^2}{2\sigma^2}} e^{-\frac{|z-(\cos\mu,\sin\mu)|^2}{2\sigma^2}}\cos(\eta-\mu)\,d\eta\,d\mu}{\int_0^{2\pi}\int_0^{2\pi} e^{-\frac{|z-(\cos\eta,\sin\eta)|^2}{2\sigma^2}} e^{-\frac{|z-(\cos\mu,\sin\mu)|^2}{2\sigma^2}}\,d\eta\,d\mu} < 1. \tag{277}$$

We see that the posterior mean always lies inside the unit circle, so $\delta(z)$ never takes values on the unit circle, which has disjoint support compared to the prior π. This problem does not appear when the support of the prior is the whole $\mathbb{R}^d$ but has a great impact when the support of the prior is a low-dimensional manifold. Easy remedies include projecting $\delta(z)$ back onto the unit circle, but as a cost, the good properties of the posterior mean are lost.

Instead, we formalize the **OT-based denoising problem** by adding the hard constraint

$$\delta_{\#}\mu_Z = \pi. \tag{278}$$

In other words, we are first restricting the space of decision rules and then find the rule that minimizes the risk from this space. This approach avoids taking projections onto $\text{supp}(\pi)$ after calculating the posterior mean and still guarantees the optimality of the decision rule at the same time. For the sake of the completeness, we state the constrained optimization problem

$$\min_{\delta} \mathbb{E}_{(\theta,Z)}|\delta(Z)-\theta|^2 \tag{279}$$

$$s.t.\ \delta_{\#}\mu_Z = \pi. \tag{280}$$

The first step is to get rid of θ in the objective by noticing

$$\begin{aligned}\mathbb{E}_{(\theta,Z)}|\delta(Z)-\theta|^2 &= \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)+\overline{\theta}(Z)-\theta|^2 &(281)\\ &= \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)|^2+\mathbb{E}_{(\theta,Z)}|\overline{\theta}(Z)-\theta|^2+2\mathbb{E}_{(\theta,Z)}\left\langle\delta(Z)-\overline{\theta}(Z),\overline{\theta}(Z)-\theta\right\rangle &(282)\\ &= \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)|^2+\mathbb{E}_{(\theta,Z)}|\overline{\theta}(Z)-\theta|^2, &(283)\end{aligned}$$

where the second term does not contain δ. As a result, the optimization problem reduces to

$$\min_{\delta} \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)|^2 \tag{284}$$

$$s.t.\ \delta_{\#}\mu_Z = \pi. \tag{285}$$

The objective can be recognized as an expected cost with $c(z,\delta(z)) = |\delta(z)-\overline{\theta}(z)|^2$ and δ being a transport map from μ_Z to π. This is actually an OT problem if the prior π is given, whereas we are assuming an unknown prior π, which differs from a traditional OT problem. Consider the regularized version through W_2^2 on the parameter space:

$$\min_{\delta} \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)|^2+\tau W_2^2(\delta_{\#}\mu_Z,\pi). \tag{286}$$

However, the problem that π is unknown remains unsolved, and some modifications on the regularization are required. The wisdom is to shift the regularization effect from the parameter space to the space of observations by noticing that if $\delta_{\#}\mu_Z = \pi$ holds, then

$$\mu_Z = \mu_\delta \tag{287}$$

shall also hold, where

$$\mu_\delta(\cdot) := \int p_\theta(\cdot)\, d(\delta_{\#}\mu_Z)(\theta). \tag{288}$$

Now that μ_δ does not contain the unknown π and we aim to solve the transformed problem:

$$\min_\delta \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)|^2+\tau W_2^2(\mu_Z,\mu_\delta). \tag{289}$$

Although the transformed problem has a different objective from the original problem, under the identifiability condition, i.e., $\int p_\theta(\cdot)\,d\pi_1(\theta) = \int p_\theta(\cdot)\,d\pi_2(\theta)$ implies $\pi_1=\pi_2$, as $\tau\to\infty$, the minimizer δ_τ^* converges to the minimizer δ^* of the OT-denoising problem

$$\min_\delta \mathbb{E}_{(\theta,Z)}|\delta(Z)-\overline{\theta}(Z)|^2 \tag{290}$$

$$s.t.\ \delta_\# \mu_Z = \pi. \tag{291}$$

The change of the objective does not interfere with the limit of the solution, which is the only thing we care about for solving OT-denoising. Consequently, we can focus on solving the transformed problem, which can be written in the Kantorovich formulation:

$$\min_{\gamma\in\Gamma}\int |\theta-\overline{\theta}(z_1)|^2+\tau|z_3-z_4|^2\,d\gamma(z_1,\theta,z_3,z_4), \tag{292}$$

$$\Gamma := \left\{\gamma : (P_1)_\#\gamma=\mu_Z, (P_4)_\#\gamma=\mu_Z, (P_3)_\#\gamma=\int p_\theta(\cdot)\,d(P_2)_\#\gamma(\theta)\right\}. \tag{293}$$

One can prove that if γ^* solves this MMOT, then $(P_{1,2})_\#\gamma^*$ can be written as $(id\times\delta_\tau^*)_\#\mu_Z$, where δ_τ^* is a solution to the transformed problem.

Remark. *The Monge formulation can be understood as searching for a non-randomized decision rule, while the Kantorovich formulation can be can be understood as searching for a randomized decision rule. Here one can prove that the optimal randomized decision rule exists and must have a non-randomized version, which is the type of conclusions often seen in decision theory.*

The MMOT problem has a special constraint on the second and the third marginals of the coupling, but one can always do a simulation to figure out the third marginal based on the second marginal.

Momentum Training on Wasserstein Space

Adding momentum results in the training to be more likely to follow past movements. When considering momentum in Wasserstein space, we have to talk about the geometry of the lines instead of line segments. First consider the geodesic formulation. Given μ_0, μ_1, now consider

$$\mu_t = [tT + (1-t)id]_{\#}\mu_0,\ t \in \mathbb{R}, \tag{294}$$

where $T = \nabla f$ is the optimal transport map for f convex. In this case,

$$\mu_t = (\nabla_x G)_{\#}\mu_0,\ G(x) = \frac{1-t}{2}|x|^2 + tf, \tag{295}$$

so G is no longer guaranteed convex when $t \notin [0,1]$. This implies that the geodesic formulation is bad at dealing with the geometry of lines.

Instead, we consider the metric interpolation approach.

$$\mu_t = \arg\min_{\mu} \left\{ \frac{1-t}{2} W_2^2(\mu, \mu_0) + \frac{t}{2} W_2^2(\mu, \mu_1) \right\}, \tag{296}$$

which has unique solution for $\forall t \in \mathbb{R}$. To see why this is the case, let's assume $t > 1$ and consider $\mu = (\nabla f)_{\#}\mu_1$ for some convex f. At the first glance, the problem

$$\min_{f \text{ convex}} \left\{ \frac{1-t}{2} W_2^2((\nabla f)_{\#}\mu_1, \mu_0) + \frac{t}{2} W_2^2((\nabla f)_{\#}\mu_1, \mu_1) \right\} \tag{297}$$

doesn't seem attractive. The term $W_2^2((\nabla f)_{\#}\mu_1, \mu_1)$ is nice (convex in ∇f), while $W_2^2((\nabla f)_{\#}\mu_1, \mu_0)$ is hard to deal with, as mentioned when we discuss the Wasserstein geometry. However, different from the case of the line segments, $W_2^2((\nabla f)_{\#}\mu_1, \mu_0)$ now has a negative coefficient, and

$$-W_2^2((\nabla f)_{\#}\mu_1, \mu_0) \tag{298}$$

is -1 **convex** in ∇f, i.e., $\frac{1}{2}\int |\nabla f|^2 \, d\mu_1 - W_2^2((\nabla f)_{\#}\mu_1, \mu_0)$ is convex in ∇f.

Remark. *The philosophy here is that the term $W_2^2((\nabla f)_{\#}\mu_1, \mu_0)$ is "bad" with positive coefficients but "good" with negative coefficients.*

Rewrite the objective

$$\frac{|1-t|}{2} \left[-W_2^2((\nabla f)_{\#}\mu_1, \mu_0) + \frac{t}{|1-t|} \int |\nabla f(x) - x|^2 \, d\mu_1(x) \right] \tag{299}$$

$$= \frac{|1-t|}{2} \Big[\frac{1}{2} \int |\nabla f|^2 \, d\mu_1 - W_2^2((\nabla f)_{\#}\mu_1, \mu_0) + \left(\frac{t}{|1-t|} - \frac{1}{2} \right) \int |\nabla f|^2 \, d\mu_1 \tag{300}$$

$$- \frac{2t}{|1-t|} \int \langle \nabla f(x), x \rangle \, d\mu_1(x) + \frac{t}{|1-t|} \int |x|^2 \, d\mu_1 \Big], \tag{301}$$

which is now convex in ∇f thanks to the -1 convexity. A similar conclusion holds when $t < 0$. This implies that the metric interpolation problem is still a convex optimization problem for ∇f when considering the geometry of lines.

Open Problems

A collection of open problems from the summer school:

1. Can we obtain an equivalent MMOT for the barycenter problem when there are negative weights $\lambda_i < 0$. (suspected to be no)

2. Numerics for the OT-denoising MMOT problem.

3. Analyze convergence of back-and-forth via continuous-time PDE analogue, get rate of convergence.

4. Cross entropy loss for the adversarial training problem (no hard constraint), MMOT formulation?

5. Choose a metric for $(\tilde{\mu}_1, ..., \tilde{\mu}_k, \Lambda)$ to turn adversarial training problem into a PDE which is a gradient flow.

6. Inverse OT, with constraints on the cost, e.g., Lagrangian structure.

7. Special cases of measures for quick calculations of Wasserstein distance.