清華大學
Tsinghua University

Tsinghua University

Convex Optimization: Theory and Application

Introduction to Wasserstein Distance

A reading report

Author
Chenghua Liu, Chenyu Wang, Ziyue Li, Xinran Gu

December 13, 2020

Contents

Chapter 1

Introduction

The concept of **Wasserstein Distance** arose from the optimal mass transport problem, which seeks the most efficient way of transforming one distribution of mass to another relative to a given cost function. The problem was first put forward by Monge and extended by Kantorovich, who proposed a general formulation [1]. Recent years with the burgeoning of computer science, Wasserstein distance gains more attention and enjoys a wide application in machine learning and signal processing. In this report, we first introduce its concepts and properties, and then provide an insight into its cutting-edge applications.

1.1 Basic Concepts

The optimal transport cost between two measures:

$$C(\mu, v) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y) d\pi(x,y) \tag{1.1}$$

where $c(x, y)$ is the cost for transporting one unit of mass from x to y. While we can think of (1.1) as a kind of distance between μ and ν, it does not strictly satisfy the axioms of a distance function. Fortunately, when the cost is defined in terms of a distance, it can be strictly shown that (1.1) is a distance measure. According to [2, p. 150], we give the definition of Wasserstein distance and then specialize it on the real number space.

Definition 1.1.1 (Wasserstein Distance)**.** Let $(\mathcal{X}, d)$ be a Polish metric space, and let $p \in [1, +\infty)$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by the formula

$$\begin{aligned} W_p(\mu,\nu) &:= \left(\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} \mathrm{d}(x,y)^p \; \mathrm{d}\pi(x,y) \right)^{1/p} \\ &= \inf \left\{ \left[\mathbb{E} d(X,Y)^p \right]^{\frac{1}{p}}, \quad \mathrm{law}(X) = \mu, \quad \mathrm{law}(Y) = \nu \right\} \end{aligned} \tag{1.2}$$

In case the above form confuses you, we specialize the p-Wasserstein distance on the space of real numbers:

$$W_p(\mu,\nu) = \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}, \quad p \geq 1 \tag{1.3}$$

Where $\Pi(\mu,\nu)$ denotes the collection of all transference plans from μ to ν, i.e. $\pi \in \Pi(\mu,\nu)$ iff μ is the first marginal and ν is the second.

Theorem 1.1.1. *$W_p(\mu,\nu)$ satisfies the three axioms of distance.*

To prove the above theorem, we first briefly introduce the Gluing Lemma [2, p. 23], which states that if μ_1, μ_2, μ_3 are Borel probability measures on X and $\pi_{1,2} \in \Pi(\mu_1,\mu_2) and \pi_{2,3}(\mu_2,\mu_3)$ are optimal transference plans. Then there exists a Borel probability measure μ on X^3 with marginal distributions $\pi_{1,2}$ to the left $X \times X$ and $\pi_{2,3}$ to the right $X \times X$. By the marginal properties of each measure, the marginal distribution of μ on the first and third X denoted by $\pi_{1,3}$ is a transference plan in $\Pi(\mu_1,\mu_3)$ (not necessarily optimal). μ glues $\pi_{1,2}$ and $\pi_{2,3}$ and that is where the name "Gluing Lemma" comes from.

Proof. Obviously, $W_p(\mu,\nu) = W_p(\nu,\mu)$. Then we apply the Gluing Lemma to prove the triangle inequality,

$$\begin{aligned}
W_p(\mu_1,\mu_3) &\leq \left(\int_{X\times X} d(x,z)^p d\pi_{1,3}(x,z)\right)^{1/p} \\
&= \left(\int_{X\times X\times X} d(x,z)^p d\mu(x,y,z)\right)^{1/p} \\
&\leq \left(\int_{X\times X\times X} (d(x,y)+d(y,z))^p d\mu(x,y,z)\right)^{1/p} \\
&\leq \left(\int_{X\times X\times X} d(x,y)^p d\mu(x,y,z)\right)^{1/p} + \left(\int_{X\times X\times X} d(y,z)^p d\mu(x,y,z)\right)^{1/p} \\
&= \left(\int_{X\times X} d(x,y)^p d\pi_{1,2}(x,y)\right)^{1/p} + \left(\int_{X\times X} d(y,z)^p d\pi_{2,3}(y,z)\right)^{1/p} \\
&= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3)
\end{aligned} \tag{1.4}$$

where the first inequality obtains from the infimum, the third obtains from the Minkovski inequality, and the first and second equalities obtain from marginal properties of measures. So we have completed the proof of triangle inequality.

While the non-negativity of $W_p(\mu,\nu)$ is obvious, $W_p(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$ is not that straightforward. On one hand, there is a probability measure concentrated on the diagonal of $\mathcal{X} \times \mathcal{X}$, whose marginal each space is μ. Since diagonal is the zero set of the metric d, we have $W_p(\mu,\mu)^p \leq \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) = 0$.

On the other hand, $W_p(\mu,\nu) = 0$ implies that there exists $\gamma \in \Pi(\mu,\nu)$ such that $\int_{\mathcal{X}\times\mathcal{Y}} d(x,y)^2 d\gamma(x,y) = 0$. This implies that $\gamma(x,y)$ is concentrated on the diagonal, so that γ is induced by the identity map. In other words,$\nu = id\#\mu = \mu$, where $f\#\mu$ denotes the pushforward of μ regarding a measurable mapping f. ■

1.2 Kantorovich Duality

In this section we introduce the dual formulation of optimal transport problem. For ease of comprehension, we first describe the formulation from an economic perspective and then verify it mathematically. As explained in [2], consider a company that owns bakeries and cafés. If the company were to delivery the bread from the bakeries to cafés on its own, it would like to minimize the transportation cost, which renders the primal optimal transport problem. Now consider a second company who offers to take care of all the transportation, buying bread at the bakeries and selling them to the cafés. Let $\psi(x)$ be the price at which a basket of bread is bought at bakery x, and $\phi(y)$ the price at which it is sold at café y. The core concern for that company is to maximize the total profits:

$$\sup_{(\psi,\phi)} \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) \tag{1.5}$$

On the other hand, in order to be competitive, the price gap should not be greater than the transportation cost , otherwise the buyers would rather purchase directly from the bakery. The constraint can be expressed as:

$$\forall (x,y), \quad \phi(y) - \psi(x) \leq c(x,y) \tag{1.6}$$

where $\psi \in L^1(X,\mu); \phi \in L^1(Y,\nu)$. This is what we call **Kantorovich Duality**.

Theorem 1.2.1 (Weak duality)**.**

$$\sup_{\phi-\psi\leq c} \left\{ \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) \right\} \leq \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \right\} \tag{1.7}$$

Proof.

$$\begin{aligned} \phi(y) - \psi(x) \leq c(x,y) &\Rightarrow \int_{\mathcal{Y}} \phi(y) f(y|x) dy - \int_{\mathcal{Y}} \psi(x) f(y|x) dy \leq \int_{\mathcal{Y}} c(x,y) f(y|x) dy \\ &\Rightarrow \int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(y) f(y|x) dy d\mu(x) - \int_{\mathcal{X}}\int_{\mathcal{Y}} \psi(x) f(y|x) dy d\mu(x) \leq \int_{\mathcal{X}}\int_{\mathcal{Y}} c(x,y) f(y|x) dy d\mu(x) \end{aligned} \tag{1.8}$$

where $f(y|x)$ denotes the conditional probability density function of y given x. Since

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(y) f(y|x) dy d\mu(x) = \int_{\mathcal{Y}} \phi(y) \left(\int_{\mathcal{X}} f(y|x) \right) d\mu(x) dy = \int_{\mathcal{Y}} \phi(y) \frac{d\nu(y)}{dy} dy = \int_{\mathcal{Y}} \phi(y) d\nu(y) \tag{1.9}$$

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \psi(x) f(y|x) dy d\mu(x) = \int_{\mathcal{X}} \psi(x) \left(\int_{\mathcal{Y}} f(y|x) dy \right) d\mu(x) = \int_{\mathcal{X}} \psi(x) d\mu(x) \tag{1.10}$$

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} c(x,y) f(y|x) dy d\mu(x) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \tag{1.11}$$

We have $\forall (\phi,\psi) \quad s.t. \quad \phi(x) - \psi(y) \leq c(x,y), \forall (x,y)$ and $\forall \pi \in \Pi(\mu,\nu)$,

$$\begin{aligned} &\int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) \leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \\ \Rightarrow &\sup_{\phi-\psi\leq c} \left\{ \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) \right\} \leq \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \right\} \end{aligned} \tag{1.12}$$

■

The above theorem implies that the solution to the Kantorovich dual problem gives a lower bound for the minimum transport cost. To dig further about the dual problem, we first discuss the relationship between price functions ψ and ϕ. A function pair (ψ, ϕ) is informally said to be *competitive* if it satisfies (1.6). For a given y, it is in the interest of the company to set the highest possible competitive price $\phi(y)$, i.e. the highest lower bound (the infimum) for for $\psi(x)+c(x,y)$, among all bakeries x. Similarly, for a given x, the price $\psi(x)$ should be the supremum of all $\phi(y)-c(x,y)$. So a price pair (ψ,ϕ) is said to be tight if

$$\phi(y)=\inf_x\left(\psi(x)+c(x,y)\right),\psi(x)=\sup_y\left(\phi(y)-c(x,y)\right) \tag{1.13}$$

For an arbitrary competitive price pair (ψ,ϕ). We can always improve ϕ by replacing it by $\phi_1(y)=\inf_x\left(\psi(x)+c(x,y)\right)$. Then we can also improve ϕ by replacing it by $\phi_1(x)=\sup_y\left(\phi_1(y)-c(x,y)\right)$; then replacing ϕ by $\phi_2(y)=\inf_x\left(\psi_1(x)+c(x,y)\right)$, and so on. Note that this process is stationary, i.e. $\phi_1=\phi_2$, $\psi_1=\psi_2$. The proof is stated as follows:

Proof. On one hand, we show that $\phi_1\geq\phi_2$:

$$\begin{aligned}\psi_1(x)&=\sup_y\left(\phi_1(y)-c(x,y)\right)\Rightarrow\psi(x)\geq\phi_1(y)-c(x,y)\\ \phi_2(y)&=\inf_x\left(\phi_1(y)+c(x,y)\right)\Rightarrow\phi_1\geq\phi_2\end{aligned} \tag{1.14}$$

We have the last inequality in that $\phi_1(y)$ is a lower bound for $\psi_1(x)+c(x,y)$ and ϕ_2 is the greatest lower bound. One the other hand, $\phi_2\leq\phi_1$ in that:

$$\begin{aligned}&\phi_1(y)=\inf_x\left(\psi(x)+c(x,y)\right)\Rightarrow\phi_1(y)\leq\psi(x)+c(x,y)\\ &\psi_1(x)=\sup_y\left(\phi_1(y)-c(x,y)\right)\Rightarrow\psi(x)\geq\phi_1(y)-c(x,y)\Rightarrow\psi(x)\geq\psi_1(x)\\ \Rightarrow&\phi_2(y)=\inf_x\left(\phi_1(x)+c(x,y)\right)\leq\inf_x\left(\phi(x)+c(x,y)\right)=\phi_1(y)\end{aligned} \tag{1.15}$$

So $\phi_1=\phi_2$. Similarly, $\psi_1=\psi_2$ ■

Based on the above findings, it makes sense to restrict our attention to tight pairs. From (1.13) we can reconstruct ϕ in terms of ψ, so we can just take ϕ as the only unknown in our problem.

However, ψ cannot be just any function: if you take an arbitrary function ψ, compute ϕ by the first formula in (1.13), the second formula may be unbounded. In fact the second formula will hold true if and only if ψ is *c-convex*, which is beyond the range of this report. Readers who are interested can refer to [2, p. 66] for a detailed discussion.

The form of Kantorovich dual problem may be clearer if we focus on the particular case of real numbers and $p=1$, which yields the following theorem:

Theorem 1.2.2 (Kantorovich-Rubinstein Theorem)**.** [3] The type 1 Wasserstein distance between μ and ν admits the dual representation

$$W_1\left(\mu,\nu\right)=\sup_{Lip(\phi)\leq1}\quad\int_{\mathbb{R}^m}\phi\left(y\right)d\nu(y)-\int_{\mathbb{R}^m}\phi(x)d\mu(x) \tag{1.16}$$

where the Lipschitz modulus $Lip(\phi)$ is defined as:

$$Lip(\phi) = \sup_{x \neq x'} \frac{|\phi(x) - \phi(x')|}{\|x - x'\|} \tag{1.17}$$

For a complete proof of the the theorem, one can refer to [2, Chapter 5]. We can easily verify that (ϕ, ϕ) is tight since $Lip(\phi) \leq 1$ yields $\phi(y) \leq \phi(x) + \|y - x\|$. Given a fixed y, the equality can be attained when $x = y$, i.e. $\phi(y) = \inf_x (\phi(x) + \|y - x\|)$, and similarly, $\phi(x) = \sup_y (\phi(y) + \|x - y\|)$.

1.3 Explicit solutions of two cases

In this section we will give two explicit solutions which are important and intuitive.First of all, we give the theorem about optimal transport for common copulas as follow.

Theorem 1.3.1. Let $(X_1, \ldots, X_d)$ and $(Y_1, \ldots, Y_d)$ be random vectors in $\mathbb{R}^d$ with respective marginal distributions (F_i) and (G_i) , for $i = 1, \ldots, d$, and sharing a common copula C. The optimal transport is $\phi^*(\mathbf{x}) = (\phi_1(x_1), \ldots, \phi_d(x_d))$ and $\phi_i(x_i) = G_i^{-1}(F_i(x_i))$.

Proof. First $G_i^{-1}(F_i(X_i))$ gives the correct marginal distribution for Y_i. Second, invoking the invariance property of copulas and the fact that each ϕ_i is increasing, $(\phi_1(X_1), \ldots, \phi_d(X_d))$ also has the correct copula C; hence, the map produces the correct target distribution.

Then the fact that the ϕ_i are increasing functions of the x_i means the derivative of ϕ^* is a nonnegative diagonal matrix, and hence ϕ is the derivative of a convex function and so is an optimal map. ■

1.3.1 One dimension

When we pay attention to the situation in one dimension, there is the following exciting theorem.

Theorem 1.3.2. When $d = 1$, denoting F_X and $F_X^{-1}(q) = \inf\{x : F_X(x) \geq q\}, q \in (0, 1)$, the distribution and quantile functions of X, we have

$$W_p(X, Y) = \left\|F_X^{-1} - F_Y^{-1}\right\|_p = \left(\int_0^1 \left|F_X^{-1}(\alpha) - F_Y^{-1}(\alpha)\right|^p \mathrm{d}\alpha\right)^{1/p}, \quad \mathbf{t}_X^Y = F_Y^{-1} \circ F_X \tag{1.18}$$

In the special case $p = 1$ there is an alternative, often more convenient, formula:

$$W_1(X, Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)| \, \mathrm{d}t = \int_0^1 \left|F_X^{-1}(\alpha) - F_Y^{-1}(\alpha)\right| d\alpha \tag{1.19}$$

In order to proof the above theorem, we first propose and prove the following three lemmas.

Theorem 1.3.3. Let $V : \mathbb{R} \to \mathbb{R}$ be even and convex. For any two collections of real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

$$\inf_{\sigma} \sum_{k=1}^{n} V\left(x_k - y_{\sigma(k)}\right) = \sum_{k=1}^{n} V\left(x_k^* - y_k^*\right) \tag{1.20}$$

where the infimum is taken over all permutations σ of $\{1, \ldots, n\}$ and x_k^* is is the k-th order statistic.

Proof. First, we consider the case n=2,soppозed $x_1 - x_2 = \delta_x,\ y_1 - y_2 = \delta_y$ subject to $0 \le \delta_y \le \delta_x$. So $x_2 - y_2 \le x_1 - y_1$ and $x_1 - y_2 + \delta_y \le x_2 - y_1 - \delta_y$

Because $V(x)$ is the convex function, We have

$V(x_1 - y_1) + V(x_2 - y_2) \le V(x_1 - y_1 + \delta_y) + V(x_2 - y_2 - \delta_y) = V(x_1 - y_2) + V(x_2 - y_1)$

For n>2, it can be easily proved by exchanging two values until that the the form in the formula is reached. ■

Theorem 1.3.4. Given two collections of real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, let μ and ν be the corresponding empirical measures. Then, for any $p \ge 1$

$$W_p^p(\mu, \nu) = \frac{1}{n} \sum_{k=1}^{n} |x_k^* - y_k^*|^p \tag{1.21}$$

In particular,

$$W_\infty(\mu, \nu) = \max_{1 \le k \le n} |x_k^* - y_k^*| \tag{1.22}$$

Proof. By the very definition of the wasserstein distance $W_p(\mu, \nu)$ between μ and ν

$$W_p^p(\mu, \nu) = \inf_{\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y|^p d\pi(x, y) = \inf_{\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - y_j|^p \pi_{ij} \tag{1.23}$$

where the infimum is taken over all probability measures π on the plane $\mathbb{R} \times \mathbb{R}$ with marginals μ and ν, and where we put $\pi_{ij} = \pi\{(x_i, y_j)\}$ (necessarily, π is supported on the points $(x_i, y_j), 1 \le i, j \le n)$. Thus, the second infimum is taken over the set M_n of all $n \times n$ matrices (π_{ij}) with non-negative entries such that, for any $i = 1, \ldots, n$ and any $j = 1, \ldots, n$

$$\sum_{j=1}^{n} \pi_{ij} = \sum_{i=1}^{n} \pi_{ij} = \frac{1}{n} \tag{1.24}$$

Note that M_n represents a convex compact subset of $\mathbb{R}^{n^2}$, and the functional

$$\pi \mapsto \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - y_j|^p \pi_{ij} \tag{1.25}$$

is affine on it. Therefore, this functional attains minimum at one of the extreme points of M_n. Any such point has the form $\pi_{ij} = \frac{1}{n} 1_{\{j=\sigma(i)\}}$, where $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is an arbitrary permutation. Hence,

$$W_p^p(\mu, \nu) = \frac{1}{n} \inf_{\sigma} \sum_{i=1}^{n} \left|x_i - y_{\sigma(i)}\right|^p \tag{1.26}$$

where the infimum is taken over all permutations $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$. Finally, to specify the last infimum, it remains to apply Lemma 1 with the convex function $V(x) = |x|^p$ ■

Theorem 1.3.5. (Elementary Skorokhod Theorem)Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of distribution functions, and assume that $F_n \to F$ weakly, i.e. $\lim_{n\to\infty} F_n(x) = F(x)$ for any point x of continuity of F. Then

$$\lim_{n\to\infty} F_n^{-1}(t) = F^{-1}(t) \tag{1.27}$$

for any point t of continuity of F^{-1}.

Proof. Let t be a point of continuity of F^{-1} and let $T \subset \mathbb{R}$ be the set of all continuity points of all F_n 's (which is dense on the real line). We first show that, given $\varepsilon > 0$, $F_n^{-1}(t) \leq F^{-1}(t)+\varepsilon$ for all n large enough. By Lemma A.3, this is equivalent to $F_n\left(F^{-1}(t)+\varepsilon\right) \geq t$. Choose $x \in T$ such that

$$F^{-1}(t)+\varepsilon > x > F^{-1}(t)+\frac{\varepsilon}{2} \tag{1.28}$$

Then $F_n(x) \geq F\left(F^{-1}(t)+\frac{\varepsilon}{2}\right) - \delta$, for all n large enough with any prescribed $\delta > 0$ Hence, it suffices to prove that $F\left(F^{-1}(t)+\frac{\varepsilon}{2}\right) \geq t+\delta$. But if $0 < \delta < 1-t$, the latter is equivalent to $F^{-1}(t)+\frac{\varepsilon}{2} \geq F^{-1}(t+\delta)$. An appropriate value of δ can be then chosen once F^{-1} is continuous at the point t. By a similar argument, $F_n^{-1}(t) > F^{-1}(t) - \varepsilon$ for all n large enough, concluding therefore the proof of the theorem.

■

Now let us prove the equation(1.18).

Proof. first we have Dvoretzky–Kiefer–Wolfowitz inequality

$$\Pr\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2} \quad \text{for every } \varepsilon > 0 \tag{1.29}$$

According to the above inequality, if F_n, F, and G_n, G are the associated distribution functions, then necessarily $F_n \to F$ and $G_n \to G$ weakly.

One can approximate μ and ν by such discrete measures μ_n and ν_n in the metric W_p. Hence, by Elementary Skorokhod theorem and Fatou's lemma,

$$\begin{aligned} \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right|^p dt &\leq \liminf_{n\to\infty} \int_0^1 \left|F_n^{-1}(t) - G_n^{-1}(t)\right|^p dt \\ &= \liminf_{n\to\infty} W_p^p\left(\mu_n, \nu_n\right) \\ &= W_p^p(\mu, \nu) \end{aligned} \tag{1.30}$$

On the other hand, the joint distribution F^{-1} of G^{-1} under the Lebesgue measure on (0,1) has μ and ν as marginals. Hence, by Definition there is an opposite inequality

$$W_p^p(\mu, \nu) \leq \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right|^p dt \tag{1.31}$$

■

1.3.2 Gaussian

In this part we will discuss the case which the distribution is Gaussian.

Theorem 1.3.6. when X and Y are Gaussian. If $X \sim N(m_1, \Sigma_1)$ and $Y \sim N(m_2, \Sigma_2)$, then

$$W_2^2(X,Y) = \|m_1 - m_2\|^2 + \operatorname{tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right] \tag{1.32}$$

$$\mathbf{t}_X^Y(x) = m_2 + \Sigma_1^{-1/2}\left[\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right]^{1/2}\Sigma_1^{-1/2}(x - m_1) \tag{1.33}$$

The above theorem can be obtained by the following theorem.

Theorem 1.3.7. Let F and G be known distribution functions of random variables taking values on R^m. A sufficient condition that, among all distributions of (X,Y) on R^{2m} with $E|X-Y|^2$ finite, that of $(X,\phi(X))$ minimizes $E|X-Y|^2$ is that: 1. $\phi(X)$ has distribution G 2. $\partial\phi(x)/\partial x'$ is symmetric and positive semidefinite.

Proof. If $u \in R^m$, then the differential form

$$2du'(u - \phi(u)) = d\zeta$$

is exact because $\partial\phi(x)/\partial x'$ is symmetric by condition (ii) and has continuous elements by assumption. So the line integral along a piecewise smooth path γ from 0 to x in R^m, given by

$$\lambda(x) = \int_\gamma d\zeta$$

depends, as the notation suggests, only on x and not on γ. Similarly, for $v \in R^m$, the differential form

$$-2dv'(\psi(v) - v) = d\eta$$

where

$$\psi(\phi(u)) = u, \qquad \text{for all } u \in R^m$$

is exact. This is, as before, implied by condition (ii) and the well-known result

$$\partial\psi(v)/\partial v' = \left[\partial\phi(u)/\partial u'\right]^{-1}\Big|_{\phi(u)=v}$$

So, the line integral along the piecewise smooth path δ from $\phi(0)$ to y in R^m, given by

$$\mu(y) = \int_\delta d\eta$$

depends only on y and not on δ. We shall define the constant

$$c = |0 - \phi(0)|^2$$

The line integral from $(0,\phi(0))$ to $(x,\phi(x))$ of the differential form $d\zeta + d\eta$ along the path in R^{2m} given by $(u,\phi(u))$, where u moves along path γ in R^m, is not dependent on path γ, and is $|x - \phi(x)|^2 - c$. Thus, for $x \in R^m$

$$\lambda(x) + \mu(\phi(x)) = |x - \phi(x)|^2 - c \tag{1.34}$$

Also, for $(x, y) \in R^{2m}$

$$\lambda(x) + \mu(y) = |\psi(y) - y|^2 + \int_\nu d\zeta - c$$

where ν may be taken as the straight line in R^m from $\psi(y)$ to x. Putting

$$u = \psi(y) + \alpha(x - \psi(y))$$

we have

$$\begin{aligned} \lambda(x) + \mu(y) + c &= |\psi(y) - y|^2 + 2\int_0^1 (x - \psi(y))'(u - \phi(u))d\alpha \\ &= |x - y|^2 + 2\int_0^1 (x - \psi(y))'[(u - \phi(u)) - (u - y)]d\alpha \\ &= |x - y|^2 - 2\int_0^1 \int_0^1 (x - \psi(y))' \left[\partial\phi(v)/\partial v'\right] (x - \psi(y))\alpha d\beta d\alpha \end{aligned}$$

where

$$v = \psi(y) + \beta(u - \psi(y))$$

since the quadratic form in the integral is nonnegative by condition (ii), we deduce that, for all $(x, y) \in R^{2m}$

$$\lambda(x) + \mu(y) \leqslant |x - y|^2 - c \tag{1.35}$$

It will be noticed that $\lambda(x)$ and $\mu(y)$ are the shadow costs in the programming formulation of our problem. Using (1.34) and (1.35) leads to the theorem. If H_0 is any distribution on R^{2m} with the required marginal distributions$F$ and G, and if H is the distribution of $(X, \phi(X))$ where conditions (1) and (2) are satisfied, then by (1.35) and (1.34),

$$\begin{aligned} \underset{H_0}{E}|X - Y|^2 &\geqslant \underset{H_0}{E}[\lambda(X) + \mu(Y)] + c = \underset{H_0}{E}[\lambda(X)] + \underset{H_0}{E}[\mu(Y)] + c \\ &= \underset{H}{E}[\lambda(X)] + \underset{H}{E}[\mu(Y)] + c = \underset{H}{E}[\lambda(X) + \mu(Y)] + c \\ &= \underset{H}{E}\left[\lambda(X) + \mu(\phi(X)\right] + c = \underset{H}{E}|X - \phi(X)|^2 \\ &= \underset{H}{E}|X - Y|^2 \end{aligned}$$

■

It's obvious that $N(m_2, \Sigma_2)$ is the image law of $N(m_1, \Sigma_1)$ with the linear map

$$x \mapsto m_2 + A(x - m_1) \tag{1.36}$$

where

$$A = \Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{-1/2} = A^T$$

To check that this maps $N(m_1, \Sigma_1)$ to $N(m_2, \Sigma_2)$, say in the case $m_1 = m_2 = 0$ for simplicity, one may define the random column vectors $X \sim N(m_1, \Sigma_1)$ and $Y = AX$ and write

$$\begin{aligned} \mathbb{E}\left(YY^T\right) &= A\mathbb{E}\left(XX^T\right)A^T \\ &= \Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{-1/2} \\ &= \Sigma_2 \end{aligned}$$

To check that the map is optimal, one may use,

$$\begin{aligned}E\left(\|X-Y\|_2^2\right) &= E\left(\|X\|_2^2\right)+E\left(\|Y\|_2^2\right)-2E(\langle X,Y\rangle)\\ &= \operatorname{Tr}\left(\Sigma_1\right)+\operatorname{Tr}\left(\Sigma_2\right)-2E(\langle X,AX\rangle)\\ &= \operatorname{Tr}\left(\Sigma_1\right)+\operatorname{Tr}\left(\Sigma_2\right)-2\operatorname{Tr}\left(\Sigma_1 A\right)\end{aligned}$$

and observe that by the cyclic property of the trace,

$$\operatorname{Tr}\left(\Sigma_1 A\right)=\operatorname{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)$$

1.4 Extension: Kullback-Leibler Divergence

1.4.1 Basic concepts

Apart from Wasserstein distance, there are a bunch of measures of discrepancy between probability distributions, Kullback-Leibler divergence, Jensen–Shannon divergence and Réyi divergence, to name but a few. In this sections, we introduce the notable and widely used Kullback-Leibler divergence (abbreviated as KL divergence).

Definition 1.4.1 (KL Divergence)**.** [4] Let $p(x)$ and $q(x)$ be two probability mass functions of a discrete random variable X. That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x)>0$ and $q(x)>0$ for any x in $\mathcal{X}$. KL divergence is defined as follows:

$$D_{KL}(p(x)||q(x))=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)} \tag{1.37}$$

Its continuous version is defined as:

$$D_{KL}(p(x)||q(x))=\int_{-\infty}^{+\infty}p(x)\log\frac{p(x)}{q(x)}dx \tag{1.38}$$

KL divergence is a measure of the information loss when $q(x)$ is used to approximate $p(x)$. Unlike Wasserstein distance, KL divergence is not a distance measure because it is asymmetric and does not satisfy the triangle inequality, which can be easily verified by counterexamples, despite its non-negativity. We list some interesting properties of KL divergence and proof. Since these properties can be easily extended from discrete cases to continuous ones by the definition of integration, we only present the proof for the discrete case.

Theorem 1.4.1 (Non-negativity of KL divergence)**.**

$$D_{KL}(p(x)||q(x))\geq 0, D_{KL}(p(x)||q(x))=0\Leftrightarrow p=q \tag{1.39}$$

Proof. For the discrete case:

$$\begin{aligned}\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)} &= -\sum_{x\in\mathcal{X}}p(x)\log\frac{q(x)}{p(x)}\\ &\geq -\sum_{x\in\mathcal{X}}\log\frac{q(x)}{p(x)}p(x)=0\end{aligned} \tag{1.40}$$

The inequality holds due to the concavity of $logx$. Since $logx$ is strictly concave and $D_{KL}(p(x)||p(x)) = 0$, $D_{KL}(p(x)||q(x)) = 0 \Leftrightarrow p = q$. ■

Theorem 1.4.2. KL divergence is convex in p, q.

Proof. For the discrete case, note that $f(p, q) = p \log \frac{p}{q}$ is convex since its Hessian

$$\nabla^2 f(p, q) = \begin{bmatrix} \frac{1}{p} & \frac{-1}{q} \\ \frac{-1}{q} & \frac{p^2}{q} \end{bmatrix} \tag{1.41}$$

is positive semidefinite. Hence $\forall \lambda \in (0, 1), \forall x \in \mathcal{X}$, $(\lambda)f(p_1(x), q_1(x)) + (1-\lambda)f(p_2(x), q_2(x)) \geq f(\lambda p_1(x) + (1-\lambda)p_2(x), \lambda q_1(x) + (1-\lambda)q_2(x))$. Summing both sides up over $\mathcal{X}$, we have $\lambda D_{KL}(p_1(x)||q_1(x)) + (1-\lambda)D_{KL}(p_2(x)||q_2(x)) \geq D_{KL}(\lambda p_1(x) + (1-\lambda)p_2(x)||\lambda q_1(x) + (1-\lambda)q_2(x))$. ■

1.4.2 Comparison with Wasserstein Distance

When considering the advantages of Wasserstein metric compared to KL divergence, then the most obvious one is that Wassterstein distance is a metric whereas KL divergence is not. Also, Wasserstein metric does not require both measures to be on the same probability space, whereas KL divergence requires both measures to be defined on the same probability space. [5][6]Finally, the dissimilarity measure given by Wasserstein distance is more accordant with intuition. Consider two degenerated distributions located at 0 and $\epsilon > 0$ in $\mathbb{R}$:

$$\mu(x) = \begin{cases} 1, & if x = 0 \\ 0, & otherwise \end{cases} \qquad \nu(x) = \begin{cases} 1, & if x = \epsilon \\ 0, & otherwise \end{cases} \tag{1.42}$$

We have $W_1(\mu, \nu) = \epsilon$ and KL divergence goes to $+\infty$. As $\epsilon \to 0$, the two distributions get "closer to each other" and $W_1(\mu, \nu) \to 0$. Despite all the advantages of Wasserstein distance over KL divergence, KL divergence still enjoys a wide application in statistics, machine learning and information theory due to its simplicity for understanding and computation. The central intuition is that the KL divergence effectively measures the average likelihood of observing (infinite) data with the distribution p if the particular model q actually generated the data.For example, it can serve as a loss function for multi-class classification, an objective function to be minimized if you want one distribution gets close to a target distribution.

Chapter 2

Applications

2.1 Wasserstein Discriminant Analysis (WDA)

2.1.1 Introduction

There are many ways for supervised linear dimensionality reduction algorithm[7]. For example, Fisher Discriminant Analysis, which is a kind of popular ways for linear discrimination, trying to firstly project the multi-dimensional data into a line, and then separate the data of different class, as well as making the distances within the same class as short as possible. Moreover, Large Margin Nearest Neighbor Classification (Local-Fisher Discriminant Analysis) is another way of classification by using Mahalanobis distance, which has many benefits such as considering the sample correlation and sample distribution, making the classification to be more reasonable. However, both of these methods only consider one kind of perspective: global or local. To be more precise, Fisher Discriminant Analysis measures all distances between sample data within and between class, while Large Margin Nearest Neighbor only considers samples that lie close to each other.

And these above are some reasons for us to propose this Wasserstein Discriminant Analysis method. As developed from optimal transport problem, it considers both global and local perspectives. Not like FDA where the weights of distances between a sample data in a class and all samples of the other class are the same, WDA relies on an optimal transport matrix $\boldsymbol{T}$ that matches all points in one class to all other points in another class, which lead to the result that these distances are of different weights.

In this section, we firstly have a brief introduction of FDA and LMNN. Then analyze why WDA is better and could achieve both global and local perspectives. Thirdly we give some backgrounds of regularized Wasserstein Distance. After that we describe the process of Wasserstein Discriminant Analysis. Finally we get conclusions.

2.1.2 Background

Fisher Discriminant Analysis

The motivation of Fisher's Discriminant Analysis is that we are trying to find the best projection direction on which training samples of the two class are best separated. To make the within-class samples as close as possible and the between-class samples as far as possible, Fisher's criterion is proposed. The optimal projection direction $\boldsymbol{w}$ is obtained by solving the below optimization problem.

$$\begin{aligned} &max \quad \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} \\ &s.t. \quad \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = c \end{aligned} \tag{2.1}$$

where $\boldsymbol{S}_b$ is the between-class scatter matrix, and $\boldsymbol{S}_w$ is the total within-class scatter matrix.

The optimal problem could be solved using Lagrange function, which is a typical convex optimization problem. The final solution is

$$\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{2.2}$$

where $m_i, i = 1, 2$ is the class mean.

Large Margin Nearest Neighbor

Firstly, we introduce Mahalanobis Distance, which is used in modified KNN (k-nearest neighbour) classification problem. Mahalanobis Distanced is used to measure the covariance distance of data. Assume x, y are sampled from population G, which has a mean vector of μ and covariance matrix of Σ. We define the Mahalanobis Distance of the two data as $d_m^2(x, y) = (x - y)^T \Sigma^{-1}(x - y)$, and the Mahalanobis Distance of data x and population G as $d_m^2(x, \mu_G) = (x - \mu_G)^T \Sigma^{-1}(x - \mu_G)$.

There are several reasons for us to define such a distance metric. Firstly, this measure is scale-invariant, which solves the problem caused by the traditional Euclidean distance that the distance varies with different units, which is always unreasonable. Secondly, it eliminates the influence of sample distribution on the distance measurement. Below is an intuitive example. Assume there are two normal population: $G_1 \sim N(\mu_1, \sigma_1^2), G_2 \sim N(\mu_2, \sigma_2^2)$. If a sample is located in A, which population is it closer to? Although it's closer to population A due to absolute length, it's closer to population B from a probability perspective: it's about 3σ to the left of μ_2, but about 4σ to the right of μ_1. Thirdly, as mentioned above, it takes sample correlation into consideration, which we could see from Σ^{-1} of the definition. It will be helpful if we consider it from the perspective of PCA (Principal Component Analysis).

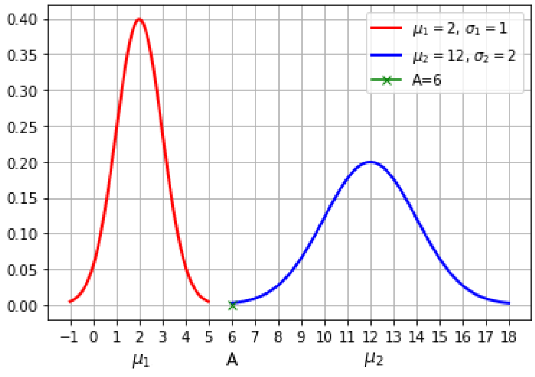

Figure 2.1: An example of sample's distribution-free property of Mahalanobis Distance

The goal of distance metric learning could be classified into two parts: learning a linear transformation $\boldsymbol{x}' = L\boldsymbol{x}$ or learning a metric $M = LL^T$. It could be convenient to transfer that into a convex optimization problem. There are many kind of applications of Mahalanobis metric, such as clustering, online learning and neighbourhood component analysis (NCA).

Now it comes to Large Margin Nearest Neighbor(LMNN), which is a direct use of Mahalabonis metric. We derive the model from two intuitions, one of which is to make label of the input data to be the same with that of its nearest samples, and the other is to separate samples of different class as far as possible. This can be reformulated into a SDP problem,

$$\begin{aligned}
&\text{Minimize } (1-\mu)\textstyle\sum_{i,j\rightsquigarrow i}\left(\vec{x}_i-\vec{x}_j\right)^\top \mathbf{M}\left(\vec{x}_i-\vec{x}_j\right)+\mu\textstyle\sum_{i,j\rightsquigarrow i,l}\left(1-y_{il}\right)\xi_{ijl}\\
&\text{subject to:}\\
&(1)\;\left(\vec{x}_i-\vec{x}_l\right)^\top \mathbf{M}\left(\vec{x}_i-\vec{x}_l\right)-\left(\vec{x}_i-\vec{x}_j\right)^\top \mathbf{M}\left(\vec{x}_i-\vec{x}_j\right)\geq 1-\xi_{ijl}\\
&(2)\;\xi_{ijl}\geq 0\\
&(3)\;\mathbf{M}\succeq 0
\end{aligned}\tag{2.3}$$

where ξ_{ijl} is a non-negative slack variable.

To have a graphical illustration[8], we propose the definition of *impostors* as shown in Figure 2.2. Impostors with different label impose into the border composed by the unit margin and target neighbors of $\boldsymbol{x}_i$. Assume for any sample $(\boldsymbol{x}_i, y_i)$, $\boldsymbol{x}_i$ has the same label with $\boldsymbol{x}_j$ and the different label with $\boldsymbol{x}_l$. Then the impostors satisfy the formula below.

$$\left\|\mathbf{L}\left(\vec{x}_i-\vec{x}_l\right)\right\|^2 \leq \left\|\mathbf{L}\left(\vec{x}_i-\vec{x}_j\right)\right\|^2+1 \tag{2.4}$$

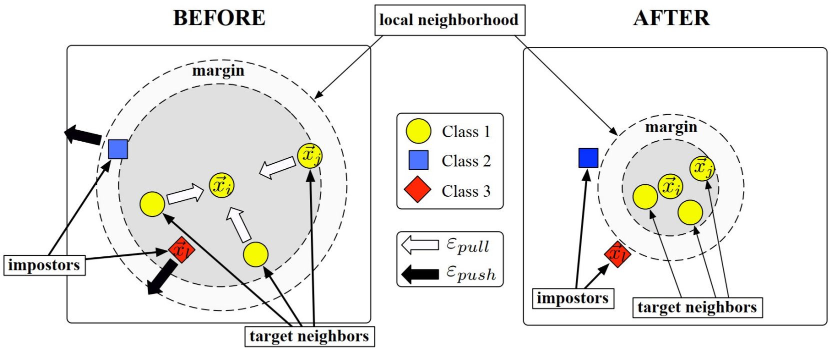

Figure 2.2: Illustration of *impostors*

There are two loss functions denoted as ϵ_{pull} and ϵ_{push}. We set $j \rightsquigarrow i$ as that $\boldsymbol{x}_j$ is the target neighbor of $\boldsymbol{x}_i$. y_{il} equals 1 if $y_i = y_l$ and equals 0 otherwise.

The merged loss function is

$$\varepsilon(\mathbf{L}) = (1-\mu)\varepsilon_{pull}(\mathbf{L}) + \mu\varepsilon_{push}(\mathbf{L}) \tag{2.5}$$

where

$$\varepsilon_{\text{pull}}\left(\mathbf{L}\right)=\sum_{j\rightsquigarrow i}\left\|\mathbf{L}\left(\vec{x}_i-\vec{x}_j\right)\right\|^2 \tag{2.6}$$

and

$$\varepsilon_{\text{push}}\left(\mathbf{L}\right)=\sum_{i,j\rightsquigarrow i}\sum_{l}\left(1-y_{il}\right)\left[1+\left\|\mathbf{L}\left(\vec{x}_i-\vec{x}_j\right)\right\|^2-\left\|\mathbf{L}\left(\vec{x}_i-\vec{x}_l\right)\right\|^2\right]_+ \tag{2.7}$$

The problem is transformed to SDP problem as shown in (2.3).

Wasserstein Discriminant Analysis

Recall that optimal transport considers all probabilistic couplings as reflected by the transportation weight T_{ij} that quantifies how important the distance $\|\boldsymbol{Px_i}-\boldsymbol{Px_j}\|$ should be to obtain a good projection matrix $\boldsymbol{P}$. We adopt the ration formulation of FDA to maximize the ratio of the regularized Wasserstein distances between inter class populations and between the intra-class population with itself, when these points are considered in their projected space:

$$\max_{\boldsymbol{P}\in\Delta}\quad \frac{\sum_{c,c^{'}>c}W_{\lambda}(\boldsymbol{PX}^c,\boldsymbol{PX}^{c^{'}})}{\sum_{c}W_{\lambda}(\boldsymbol{PX}^c,\boldsymbol{PX}^c)} \tag{2.8}$$

where $\Delta=\{\boldsymbol{P}=[\boldsymbol{p}_1,\cdots,\boldsymbol{p}_p]|\boldsymbol{p}_i\in\mathbb{R}^d,||\boldsymbol{p}_i||_2=1,\text{ and }\boldsymbol{p}_i^\top\boldsymbol{p}_j=0\text{ for }i\neq j\}$ is the Stiefel manifold, the set of orthogonal $d\times p$ matrices; $\boldsymbol{PX}^c$ is the matrix of projected samples from class c. W_λ is the regularized Wasserstein distance expressed as

$$W_\lambda(\boldsymbol{X},\boldsymbol{Z})=\sum_{i,j}T^*_{i,j}||\boldsymbol{x}_i-\boldsymbol{z}_j||_2^2 \tag{2.9}$$

$T^*_{i,j}$ being the coordinates of the entropic-regularized Optimal Transport (OT) matrix $\boldsymbol{T^*}$.

Relationships Between These Three Methods

When λ is small, WDA boils down to FDA. When λ is large, WDA tries to split apart distributions of classes by maximizing their optimal transport distance. In that process, for a given example $\boldsymbol{x_i}$ in one class, only few components $T_{i,j}$ will be activated so that $\boldsymbol{x}_i$ will be paired with few examples.

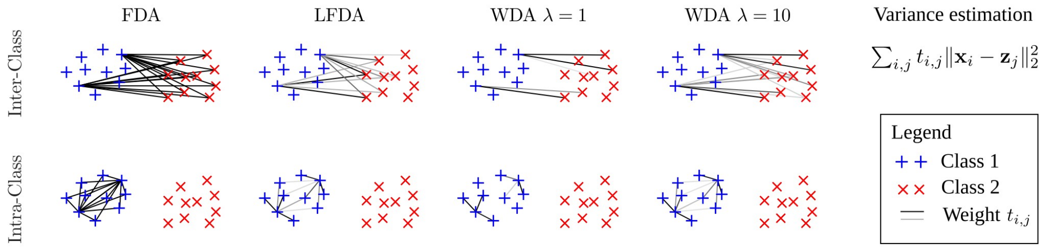

Figure 2.3: Relationships Between These Three Methods

The picture above is a graphical illustration of the relationships between these three methods. Only weights for two samples from class 1 are shown. The color of the link darkens as the weight grows. FDA computes a global variance with uniform weight on all pairwise distances, whereas LFDA focuses only on samples that lie close to each other. WDA relies on an optimal transport matrix $\boldsymbol{T}$ that matches all points in one class to all other points in another class (most links are not visible because they are colored in white as related weights are too small). WDA has both a global (due to matching constraints) and local (due to transportation cost minimization) outlook on the problem, with a tradeoff controlled by the regularization strength λ.

2.1.3 Notations and Regularized Wasserstein Distance

The regularized Wasserstein distance is

$$W_\lambda(\mu, \gamma) :=< \boldsymbol{T}_\lambda, \boldsymbol{M_{X,Z}} > \tag{2.10}$$

where $\boldsymbol{T}_\lambda$ is the solution of an entropy-smoothed optimal transport problem

$$\boldsymbol{T}_\lambda = \text{argmin}_{\boldsymbol{T} \in U_{nm}} \lambda < \boldsymbol{T}, \boldsymbol{M_{X,Z}} > -\Omega(\boldsymbol{T}) \tag{2.11}$$

where $\Omega(\boldsymbol{T})$ is the entropy of $\boldsymbol{T}$ defined as $\Omega(\boldsymbol{T}) := -\sum_{ij} t_{ij} log(t_{ij})$.

The solution of the optimization problem can be expressed as

$$\boldsymbol{T} = \text{diag}(\boldsymbol{u})\boldsymbol{K}\text{diag}(\boldsymbol{v}) = \boldsymbol{u}\mathbf{1}_m^\top \circ \boldsymbol{K} \circ \mathbf{1}_n \boldsymbol{v}^\top \tag{2.12}$$

where $\circ$ stands for elementwise multiplication and $\boldsymbol{K} = e^{-\lambda \boldsymbol{M}}$, and $\boldsymbol{v}^k = \frac{\mathbf{1}_m/m}{\boldsymbol{K}^\top \boldsymbol{u}^{k-1}}, \boldsymbol{u}^k = \frac{\mathbf{1}_n/n}{\boldsymbol{K}\boldsymbol{v}^k}$ with an initialization which will be fixed to $\boldsymbol{u}^0 = \mathbf{1}_n$.

2.1.4 Wasserstein Discriminant Analysis

Define that samples of class c are stored in matrices $\boldsymbol{X}^c$; the number of samples from class c is n_c. Then the Wasserstein Discriminant Analysis optimization problem is

$$\begin{aligned} &\max_{\boldsymbol{P} \in \Delta} \quad J(\boldsymbol{P}, \boldsymbol{T}(\boldsymbol{P})) \\ &\text{s.t.} \quad \boldsymbol{T}(\boldsymbol{P}) = \text{argmin}_{\boldsymbol{T} \in U_{n_c n_{c'}}} \quad E(\boldsymbol{T}, \boldsymbol{P}) \end{aligned} \tag{2.13}$$

where $\boldsymbol{T} = \{\boldsymbol{T}^{c,c'}\}_{c,c'}$ contains all the transport matrices between classes and the inner problem function E is defined as

$$E(\boldsymbol{T},\boldsymbol{P}) = \sum_{c,c>=c'} \lambda < \boldsymbol{T}^{c,c'}, \boldsymbol{M}_{\boldsymbol{PX}^c,\boldsymbol{PX}^{c'}} > -\Omega(\boldsymbol{T}^{c,c'}) \tag{2.14}$$

and the objective function J can be expressed as

$$J(\boldsymbol{P},\boldsymbol{T}(\boldsymbol{P})) = \frac{< \boldsymbol{P}^\top \boldsymbol{P}, \boldsymbol{C}_b >}{< \boldsymbol{P}^\top \boldsymbol{P}, \boldsymbol{C}_w >} \tag{2.15}$$

where $\boldsymbol{C}_b = \sum_{c,c'>c} \boldsymbol{C}_{c,c'}$ and $\boldsymbol{C}_w = \sum_c \boldsymbol{C}_{c,c}$ are the between and within cross-covariance matrices that depend on $\boldsymbol{T}(\boldsymbol{P})$. This optimization problem is a bilevel optimization problem, which could be solved using gradient descent. J is differentiable with respect to $\boldsymbol{P}$, due to the fact that the optimization problems in equation of (2.13) are all strictly convex, making solutions of the problems unique, hence $\boldsymbol{T}(\boldsymbol{P})$ is smooth and differentiable. Thus, using the chain rule, we get

$$\nabla_{\boldsymbol{P}} J(\boldsymbol{P},\boldsymbol{T}(\boldsymbol{P})) = \frac{\partial J(\boldsymbol{P},\boldsymbol{T})}{\partial \boldsymbol{P}} + \sum_{c,c' \geq c} \frac{\partial J(\boldsymbol{P},\boldsymbol{T})}{\partial \boldsymbol{T}^{c,c'}} \frac{\partial \boldsymbol{T}^{c,c'}}{\partial \boldsymbol{P}} \tag{2.16}$$

The first term could be computed efficiently, and the second term is difficult in computing the Jacobian $\frac{\partial \boldsymbol{T}^{c,c'}}{\partial \boldsymbol{P}}$ since the optimal transport matrix is not available as a closed form. We solve this problem using instead an automatic differentiation approach wrapped around the Sinkhorn fixed point iteration algorithm. The process is somewhat complex so we only show the algorithms below.

Algorithm: *Projected gradient algorithm for WDA*

```
Require: $\prod_\Delta$: projection on the Stiefel manifold
  • Initialize $k = 0$, $\boldsymbol{P}^0$
  • repeat
      – compute all the $\boldsymbol{T}^{c,c'}$ as given in the equation of (2.13) by means of Sinkhorn-Knopp algorithm with automatic differentiation
      – compute $\boldsymbol{C}_b = \sum_{c,c'>c} \boldsymbol{C}_{c,c'}$ and $\boldsymbol{C}_w = \sum_c \boldsymbol{C}_{c,c}$
      – compute $\boldsymbol{P}^k$
      – compute $\frac{\partial \boldsymbol{T}^k}{\partial \boldsymbol{P}}$ using automatic differentiation
      – compute gradient $\boldsymbol{G}^k = \nabla_{\boldsymbol{P}} J(\boldsymbol{P}^k, \boldsymbol{T}(\boldsymbol{P}^k))$ using all above elements
      – compute descent direction $\boldsymbol{D}^k = \prod_\Delta(\boldsymbol{P}^k - \boldsymbol{G}) - \boldsymbol{P}^k$
      – linesearch on the step-size $\alpha_k$
      – $\boldsymbol{P}^{k+1} \leftarrow \prod_\Delta(\boldsymbol{P}^k + \alpha_k \boldsymbol{D}^k)$
```

- $k \leftarrow k+1$

- **until convergence**

2.1.5 Conclusion and Further Discussions

As shown above, WDA allows construction of both a local and global interaction of the empirical distributions to compare. Globality naturally results from the Wasserstein distance, which is a metric on probability measures, and as such it measures discrepancy between distributions at whole level. Locality comes as a specific feature of regularized Wasserstein distance.

WDA encompasses Fisher Discriminant analysis in the limit case where λ approaches 0. Beyond this limit case and when $\lambda > 0$, the smoothed optimal transport matrix $\boldsymbol{T}$ promotes cross-covariance matrices that are estimated from local relations as illustrated in Fig. 2.3.

Moreover, WDA has relations to other information-theoretic discriminant analysis. As it provides a useful way to compare probability measures, it has been widely used in machine learning problems as well. In the future, we may consider stochastic versions of the same approach in order to enhance further the ability of the method to handle high-dimensional problems.

2.2 Wasserstein distance and federated learning

Federated learning (abbreviated as FL), also known as federated optimization, allows multiple parties to collaboratively train a model without data sharing. With the increasing awareness of privacy preserving and extensive use of intelligent devices, it has become a heated research topic in recent years. Readers who are interested may refer to for a comprehensive literature review[9]. One challenge that distinguishes federated learning from distributed machine learning is the data heterogeneity of data distributions among users, which can be quantified by Wasserstein distance.[10]

Consider a typical FL training objective:

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{2.17}$$

where f_i represents expected loss over the data distribution of user i, i.e.

$$f_i(w) := \mathbb{E}_{(x,y) \sim p_i} \left[l_i(w; x, y) \right] \tag{2.18}$$

where $l_i(w; x, y)$ measures the error of model w in predicting the true label $y \in \mathcal{Y}_i$ given the input $x \in \mathcal{X}_i$, and p_i is the distribution over $\mathcal{X}_i \times \mathcal{Y}_i$. In convergence of analysis of federate algorithms like FedAvg[11], typically require some regularity condition of first order and seconder order gradient for f_i. Consider the following regularity condition:

Assumption 2.2.1. For every $i \in \{1, ..., n\}$, and $z := (x, y)$, $l_i(z; w)$ is L smooth:

$$\left\| \nabla_w l\left(z_1; w\right) - \nabla_w l\left(z_2; w\right) \right\| \leq L d\left(z_1, z_2\right) \tag{2.19}$$

where $d(z_1, z_2)$ is a distance measure, typically $d(z_1, d_2) = |z_1 - z_2|$.

Theorem 2.2.2. With (2.19), an upper bound for the diversity of gradients based on Wasserstein distance:

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(w) - \nabla f(w)\|^2 \leq \gamma_G^2 = L^2 \frac{1}{n}\sum_{i=1}^{n} W_1\left(p_i, p\right)^2 \tag{2.20}$$

Proof.

$$\begin{aligned} ||\nabla f_i(w) - \nabla f(w)|| &= ||\mathbb{E}_{Z\sim p_i}[\nabla_w l(z; w)] - \mathbb{E}_{Z\sim p_i}[\nabla_w l(z; w)]|| \\ &= ||\mathbb{E}_{(Z_1,Z_2)\sim\Pi_i}[\nabla_w l(z_1; w) - \nabla_w l(z_2; w)]|| \\ &\leq \mathbb{E}_{(Z_1,Z_2)\sim\Pi_i}||\nabla_w l(z_1; w) - \nabla_w l(z_2; w)|| \\ &\leq L\mathbb{E}_{(Z_1,Z_2)\sim\Pi_i}||d(z_1, z_2)|| \end{aligned} \tag{2.21}$$

The inequality holds with all Π_i with margins p and p_i, so

$$||\nabla f_i(w) - \nabla f(w)|| \leq \inf_{\Pi_i} \mathbb{E}_{(Z_1,Z_2)\sim\Pi_i}||d(z_1, z_2)|| = W_1(p_i, p) \tag{2.22}$$

Applying the convexity of $f(x) = x^2$, we have (2.20). ■

2.3 Wasserstein-Distance-Based Gaussian Mixture Reduction

2.3.1 Introduction

The main target of this task in [12] is to conduct Gaussian Mixture Reduction (GMR) to deal with the problem of exponentially increasing number of components in Gaussian Mixtures (GMs). GMR is a common practice in many signal processing applications to cut the computational cost at the price of minimal information loss. Existing GMR algorithms include top-down greedy approaches and clustering, the central to which is a probabilistic similarity measure. Frequently used measures including Kullback-Leibler divergence (KLD) and the Bhattacharyya distance are employed to identify the the similar components in GMs. These similar components are then replaced by a single Gaussian component parameterized by a moment-matching approach.

The Wasserstein distance (WD) is a promising candidate for this application to measure the similarity between probability distributions. Compared with currently used KLD, WD enumerates the geometric shape difference between two probability density functions (PDFs) by minimizing the cost of transferring one PDF to another. KLD, however, as a nonsymmetric nonmetric probabilistic similarity measure, fails to capture the geometric shape information. Therefore, WD is able to deliver a superior similarity measure for our GMR task, which can also be seen from latter simulation experiments.

2.3.2 Algorithm Details

In accordance to traditional methods, two main steps are utilized in the WD-based GMR: *Wasserstein-Based Greedy GMR* and *Wasserstein-Based Clustering GMR*. We use $p(x)$ in this section to denote a GM with M Gaussian components:

$$p(x) = \sum_{i=1}^{M} \omega_i p_i(x) \tag{2.23}$$

where $x \in \mathbb{R}^n$ is a random variable, $p_i(x) = \mathcal{N}(x; \mu_i, \Sigma_i)$ is the ith Gaussian component of $p_i(x)$ with mean $\mu_i \in \mathbb{R}^n$ and covariance matrix $\Sigma_i \in \mathbb{R}^{n\times n}$, and ω_i is the normalized weight of the ith component. We aim at reducing it to a GM $q(x) \sum_{i=1}^{N} \omega_i^R q_i(x), q_i(x) = \mathcal{N}(x; \mu_i^R, \Sigma_i^R)$ with $N < M$ components and least information loss. The two steps are introduced in detail as following:

- **Step One:** *Wasserstein-Based Greedy GMR*

  In this greedy approach, the two closest components in the GM are selected and merged in each iteration, which makes GM components number lower by 1. Such operation is done successively until the components number equals to N, giving rise to the required reduced GM. Since the closest components are merged in each step, it can be categorized as a greedy algorithm.

  For *Wasserstein-Based Greedy GMR*, "closest components" are defined according to WD. The mean and covariance of the merged component are generated by minimizing the Wasserstein-based weighted average distance between original components and the merged one. Details will be elaborated later.

- **Step Two:** *Wasserstein-Based Clustering GMR*

  This step is based on an initial candidate of $q(x)$, which is the output of *Wasserstein-Based Greedy GMR*. Since the choice of initial $q(x)$ can affect the outcome of the clustering GMR, the first step can be viewed as a preparation to the second one. For each iteration, every component of $p(x)$ is associated with one component of $q(x)$ with the closest WD, forming a cluster. Then the weight and distribution of each component in $q(x)$ are replaced with formulas derived from cluster components. Such operations are done until the algorithm converge to a stable $q(x)$.

  Instead of using the frequently used moment-matching approach to calculate moments of cluster center, a Wasserstein-based method is utilized here to capture the components shape information. It can be viewed as a multi-components version of the minimization problem in step one, which will be discussed in detail later.

The mathematical formulas of components center derivation mentioned in **Step One** and **Step Two** are shown as following:

As a review, Wasserstein distance D_{ij} between Gaussian components $p_i(x)$ and $p_j(x)$ is

$$D_{ij} = D_{W_2}^2(p_i(x), p_j(x)) = \operatorname{tr}\{\Sigma_i + \Sigma_j - 2(\Sigma_i^{\frac{1}{2}} \Sigma_j \Sigma_i^{\frac{1}{2}})^{\frac{1}{2}}\} + \|\mu_i - \mu_j\|_2^2 \tag{2.24}$$

The optimization problem we encounter in the above steps can be formulated as:

$$\min_{\mu_i^R, \Sigma_i^R} L_i = \min_{\mu_i^R, \Sigma_i^R} \sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} D_{W_2}^2(p_{l_i(k)}(x), q_i(x)) \tag{2.25}$$

where $\tilde{\omega}_{l_i(k)} = \omega_{l_i(k)} / \sum_{k=1}^{K} \omega_{l_i(k)}$, $l_i(k)$ is the kth components of cluster i. We assume that no cluster is empty (which can be achieved by initialization with greedy GMR). It can be seen that the optimization problem in **Step One** is a special case of the one above when $K_i = 2$, $l_i(1) = i$ and $l_i(2) = j$.

With multi-components, the last item in the trace of WD formula is approximated and L_i has the following approximation:

$$L_i \approx \widehat{L_i} = \sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} \{ \mathrm{tr}\{ (\Sigma_{l_i(k)}^{\frac{1}{2}} - (\Sigma_i^R)^{\frac{1}{2}})^2 \} + \| \mu_{l_i(k)} - \mu_i \|_2^2 \} \tag{2.26}$$

With the form of weighted sum of quadratic terms, it is convex in μ_i and Σ_i. Thus, the global minimum can be achieved by taking first order derivatives:

$$\frac{\partial \widehat{L_i}}{\partial \mu_i^R} = 2 \sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} (\mu_i^R - \mu_{l_i(k)}) \tag{2.27}$$

$$\begin{aligned} \frac{\partial \widehat{L_i}}{\partial \Sigma_i^R} &= \sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} \{ 2((\Sigma_i^R)^{\frac{1}{2}} - \Sigma_{l_i(k)}^{\frac{1}{2}}) \frac{1}{2} (\Sigma_i^R)^{-\frac{1}{2}} \} \\ &= \mathbf{I} - \sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} \Sigma_{l_i(k)}^{\frac{1}{2}} (\Sigma_i^R)^{-\frac{1}{2}} \end{aligned} \tag{2.28}$$

Set them to zero and we get

$$\mu_i^{R*} = \sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} \mu_{l_i(k)} \tag{2.29}$$

$$\begin{aligned} \Sigma_i^{R*} &= (\sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} \Sigma_{l_i(k)}^{\frac{1}{2}})(\sum_{k=1}^{K_i} \tilde{\omega}_{l_i(k)} \Sigma_{l_i(k)}^{\frac{1}{2}}) \\ &= \sum_{k=1}^{K_i} \sum_{k'=1}^{K_i} \tilde{\omega}_{l_i(k)} \tilde{\omega}_{l_i(k')} \Sigma_{l_i(k)}^{\frac{1}{2}} \Sigma_{l_i(k')}^{\frac{1}{2}} \end{aligned} \tag{2.30}$$

They are utilized for parameters update in cluster GMR of **Step Two**.

For the case in **Step One**, when $K_i = 2$, an accurate solution can be derived instead of the approximation above:

$$\mu_i^{R*} = \tilde{\omega}_i \mu_i + \tilde{\omega}_j \mu_j \tag{2.31}$$

$$\Sigma_i^{R*} = \tilde{\omega}_i^2 \Sigma_i + \tilde{\omega}_j^2 \Sigma_j + \tilde{\omega}_i \tilde{\omega}_j ((\Sigma_i \Sigma_j)^{\frac{1}{2}} + (\Sigma_j \Sigma_i)^{\frac{1}{2}}) \tag{2.32}$$

Utilizing the above μ_i^{R*}, Σ_i^{R*}, and combining it with $\omega_i^{R*} = \sum_{i=1}^{K_i} \omega_{l_i(k)}$, we derive the parameter update method for both steps.

The detailed algorithms are illustrated as the following:

Algorithm 1: *Wasserstein-Based Greedy GMR*

```
While $M > N$ Do
    • Find closest Gaussian components pair in terms of WD.
    • Merge the two components into a new one with parameters $\omega_i^{R*}$, $\mu_i^{R*}$ and $\Sigma_i^{R*}$.
    • $N = N - 1$
End While
```

Algorithm 2: *Wasserstein-Based Clustering GMR*

```
Initialize $q(x)$ with results of Algorithm 1.
While $q_{new} \neq q_{old}$ Do
    • Associate each $p_i(x)$ to the closest reduced component cluster $j$ in terms of WD.
    • Update parameters for each cluster center with moments of the associated original components.
    • Derive $q_{new}(x) = \sum_{i=1}^{N} \omega_i^R \mathcal{N}(x; \mu_i^R, \Sigma_i^R)$
End While
```

2.3.3 Simulation Experiments

Two one-dimensional simulation scenarios are conducted as the way in [12].

In the first scenario, GMs with components number $M = 10$ to $M = 100$ are considered to be reduced to GM with $N = 5$ components. Parameters of the original GMs are drawn randomly from the following ranges: $\omega_i \in [0.05, 0.5]$, $\mu_i \in [0, 3]$ and $\Sigma_i \in [0.09^2, 0.5^2]$. In our experiment, they are sampled from uniform distributions. The GMR is conducted using the proposed WD-based clustering GMR (Wb-GMR) and traditional KLD-based GMR (Kb-GMR). Two metrics including WD and time cost are calculated for each M. Results can be seen in Figure 2.4.

The WD between the original and the reduced GMs with Wb-GMR is consistently lower than the case with Kb-GMR, indicating that Wb-GMR is able to generate reduced GMs with more similar geometric shape as the original one. As for computational cost, Wb-GMR is slightly higher than Kb-GMR.

As a supplementary issue, the method of WD estimation between two GMs used in this experiment is elaborated as below utilizing the method proposed in [13] (since exact WD form is not available for GMs).

Suppose two GMs μ_0, μ_1 take the form $\mu_i = \sum_{j=1}^{N_i} p_i^j v_i^j, i = 0, 1$. Our framework is built on the following discrete Optimal Mass Transport (OMT) problem

$$\min_{\pi \in \Pi(p_0, p_1)} \sum_{i,j} c(i,j)\pi(i,j) \tag{2.33}$$

where $c(i, j)$ is the square of WD $D^2_{W_2}(v^i_0, v^j_1)$ and $\Pi(p_0, p_1)$ denotes the space of joint distributions between p_0 and p_1.

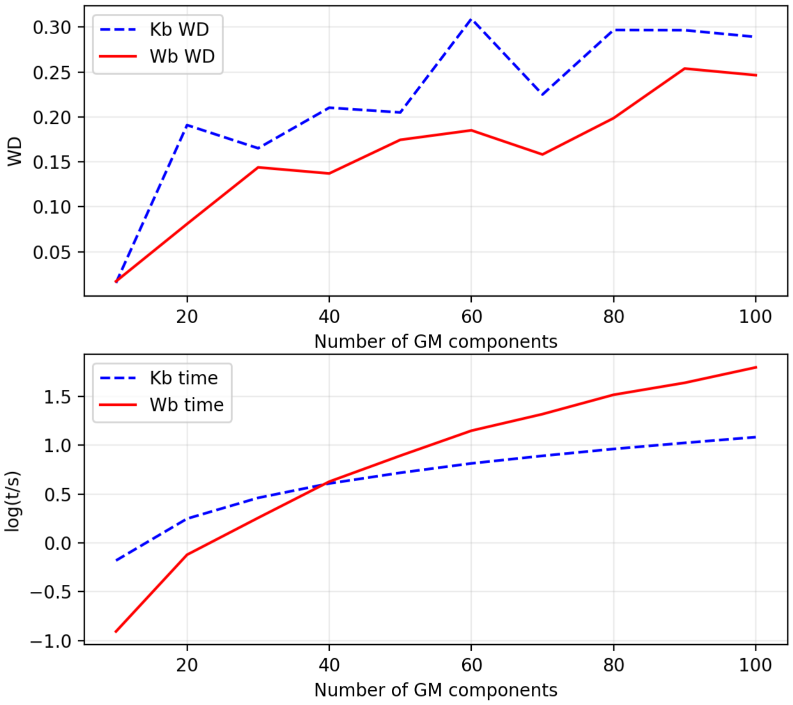

Figure 2.4: Comparison of Kb-GMR and Wb-GMR in terms of WD (above) and computational cost (below)

Given values of $c(i, j)$ and linear constraints according to properties of $\pi(p_0, p_1)$, the OMT problem can be solved by standard linear programming. The approximate WD between two GMs is derived as:

$$d(\mu_0, \mu_1) = \sqrt{\sum_{i,j} c(i, j)\pi^*(i, j)} \tag{2.34}$$

which is utilized in metric calculation in this experiment.

In the second scenario, a 1-D GM with 10 components is used to illustrate WD-based GMR's shape maintaining property. For better comparison, parameters of this GM are set the same as in [12]:

$$\begin{cases} \omega = [0.03, 0.18, 0.12, 0.19, 0.02, 0.16, 0.06, 0.1, 0.08, 0.06] \\ \mu = [1.45, 2.20, 0.67, 0.48, 1.49, 0.91, 1.01, 1.42, 2.77, 0.89] \\ \Sigma = [0.0487, 0.0305, 0.1171, 0.0174, 0.0295, 0.0102, 0.0323, 0.0380, 0.0115, 0.0679] \end{cases} \tag{2.35}$$

The original GM and reduced GMs by Wb-GMR and Kb-GMR are shown in Figure 2.5:

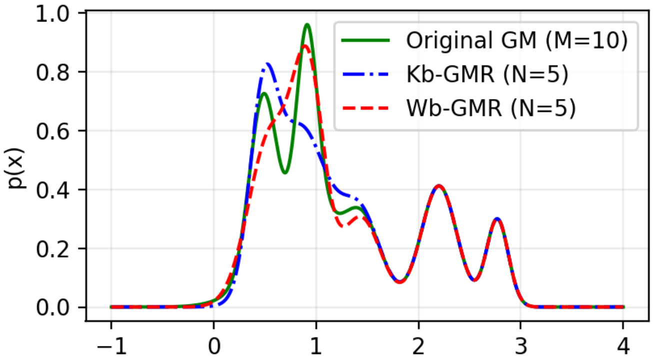

Figure 2.5: Distributions of original GM and Kb-GMR/Wb-GMR reduced form

It can be seen intuitively that the Wb-GMR result is closer to the original form in terms of geometric shape. Codes for this section can be found in `Wb-GMR.py`.

2.4 Sliced Wasserstein Distance for Learning Gaussian Mixture Models

2.4.1 Introduction

The target of this task in [14] is to estimate the parameters of a finite Gaussian Mixture Model (GMM). Existing methods are based on minimizing the negative log-likelihood (NLL) of the data with respect to parameters and the EM algorithm is the prominent way in practice. Problems related to this method include:

1. Studies show that likelihood function has very bad local maxima. The fact that EM only guarantees convergence to a stationary point of the target function makes it sometimes leads to arbitrarily worse log-likelihood values;
2. EM algorithms is sensitive to the choice of initial parameters.

Since NLL minimization is equivalent to minimizing KLD between data distribution and GMM in the limit in terms of sample numbers, replacing KLD with WD is a potential way to deal with the problems above. More specifically, the sliced p-Wasserstein distance (sliced p-WD) is utilized to overcome computational burden in high dimensional cases. Details about the sliced p-WD are elaborated in the following section.

2.4.2 Formulas Derivation and Analysis

The optimization problem of GMM parameters learning can be formulated as:

$$\inf_{I_x} D^p_{W_p}(I_x, I_y) \tag{2.36}$$

where $I_y = \frac{1}{N}\sum_{n=1}^{N}\phi(y - y_n)$ is the empirical distribution of sample set Y and

$$I_x(x) = \sum_k \frac{\alpha_k}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_k|}} exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\} \tag{2.37}$$

is the distribution of parameterized GMM I_x with $\sum_k \alpha_k = 1$.

For this task, p-WD has several benefits over the commonly used KLD:

1. The GMM model I_x is continuous and smooth in its parameters and is locally Lipschitz. Thus, $D_{W_P}(I_x, I_y)$ is always continuous and differentiable, which is not the case for KLD.
2. WD is more suitable to scenarios where the distributions are supported by low dimensional manifolds due to its *Lagrangian* nature. In this case, the cost function derived by KLD can be difficult to optimize since they compare distributions at fixed spatial coordinates.
3. The WD landscape is more flat and thus less sensitive to the starting point, making it possible for gradient descent to work. As the practical case in [14], both methods lead to the same global optimum, while the WD landscape makes it easier to converge to the optimal point regardless of the starting point as illustrated in Figure 2.6

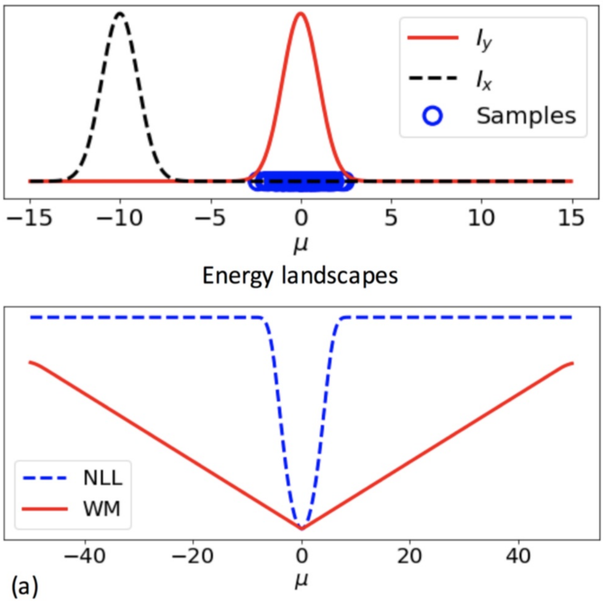

Figure 2.6: Energy Landscapes for NLL and WD

However, the calculation of WD can be very computational expensive for high-dimensional data. To deal with the problem, the sliced WD is utilized instead of the original WD:

$$\begin{aligned}\inf_{\mu_k,\Sigma_k,\alpha_k} SW_p^p(I_x, Iy) &= \inf_{\mu_k,\Sigma_k,\alpha_k} \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}I_x(.,\theta), \mathcal{R}I_y(.,\theta))d\theta \\ &= \inf_{\mu_k,\Sigma_k,\alpha_k} \int_{\mathbb{S}^{d-1}} \inf_{f(.,\theta)} \int_{\mathbb{R}} |f(t,\theta) - t|^p \mathcal{R}I_x(t,\theta)dtd\theta\end{aligned} \tag{2.38}$$

where $\mathcal{R}I(t,\theta) := \int_{\mathbb{R}^d} I(x)\delta(t - x \dots \theta)dx$ is the Randon transform mapping d-dimensional input to one-dimensional output with parameter $\theta \in \mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$, and $f(.,\theta)$ is the optimal transport map between $\mathcal{R}I_x(.,\theta)$ and $\mathcal{R}I_y(.,\theta)$.

The optimization problem is solved by gradient based methods with RMSProp optimizer, using the summation over L random projections $\theta_l \in \mathbb{S}^{d-1}$. Algorithm details are elaborated as below:

Algorithm: *Sliced Wasserstein Distance for GMM Parameters Fitting*

- Generate L random samples $\theta_1, ..., \theta_L$ from $\mathcal{S}^{d-1}$;
- For each $l \in \{1, ..., L\}$, calculate $f(t, \theta_l) = \mathcal{R}J_y^{-1}(\mathcal{R}J_x(t, \theta_l), \theta_l)$ where $\mathcal{R}J_x(., \theta_l)$ is the CDF of $\mathcal{R}I_x(., \theta_l)$;
- Update GMM parameters with RMSProp optimizer and gradients derived by differentiating 2.38;
- Project Σ_ks to semidefinite cone and renormalize α_ks.

The superiority of WD over KLD to be less sensitive to initial points can be seen in the following numerical experiment result by [14] in Figure 2.7:

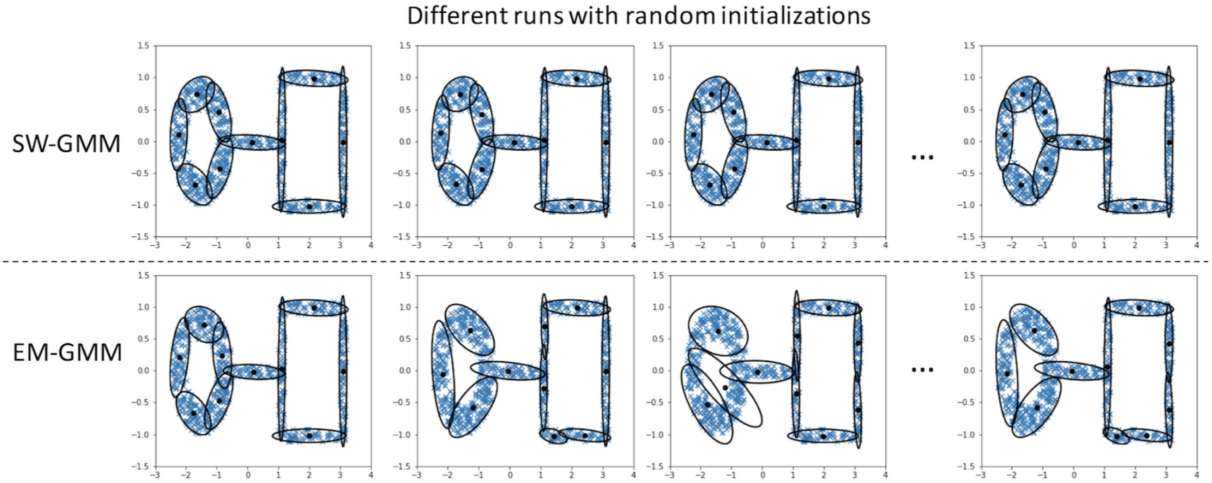

Figure 2.7: Results of 100 runs of EM-GMM and SW-GMM fitting a model with 10 components to the ring-line-square dataset

It can be seen that SW-GMM has a consistent result regardless of initial points, which is not the case for EM-GMM, giving supporting evidence for the robustness of the proposed sliced-WD based GMM fitting method.

Bibliography

[1] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. "Optimal Mass Transport: Signal processing and machine-learning applications". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 43–59. DOI: 10.1109/MSP.2017.2695801 (page 1).

[2] Cédric Villani. "Optimal transport – Old and new". In: vol. 338. Jan. 2008. DOI: 10.1007/978-3-540-71050-9 (pages 1–5).

[3] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*. 2019. arXiv: 1908.08729 [stat.ML] (page 4).

[4] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining concepts and techniques, third edition*. 2012. URL: http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1 (page 10).

[5] Lucas Roberts (https://stats.stackexchange.com/users/150025/lucas-roberts). *What is the advantages of Wasserstein metric compared to Kullback-Leibler divergence?* Cross Validated. URL:https://stats.stackexchange.com/q/295729 (version: 2019-10-27). eprint: https://stats.stackexchange.com/q/295729. URL: https://stats.stackexchange.com/q/295729 (page 11).

[6] antike (https://stats.stackexchange.com/users/187743/antike). *What is the advantages of Wasserstein metric compared to Kullback-Leibler divergence?* Cross Validated. URL:https://stats.stackexchange.com/q/351153 (version: 2018-06-13). eprint: https://stats.stackexchange.com/q/351153. URL: https://stats.stackexchange.com/q/351153 (page 11).

[7] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. "Wasserstein discriminant analysis". In: *Machine Learning* 107.12 (2018), pp. 1923–1945. ISSN: 0885-6125 1573-0565. DOI: 10.1007/s10994-018-5717-1 (page 12).

[8] Kilian Q Weinberger and Lawrence K Saul. "Distance metric learning for large margin nearest neighbor classification." In: *Journal of Machine Learning Research* 10.2 (2009) (page 14).

[9] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song,

Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. *Advances and Open Problems in Federated Learning*. 2019. arXiv: 1912.04977 [cs.LG] (page 18).

[10] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. *Personalized Federated Learning: A Meta-Learning Approach*. 2020. arXiv: 2002.07948 [cs.LG] (page 18).

[11] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. *Communication-Efficient Learning of Deep Networks from Decentralized Data*. 2017. arXiv: 1602.05629 [cs.LG] (page 18).

[12] Akbar Assa and Konstantinos N Plataniotis. "Wasserstein-distance-based Gaussian mixture reduction". In: *IEEE Signal Processing Letters* 25.10 (2018), pp. 1465–1469 (pages 19, 22, 23).

[13] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. "Optimal transport for Gaussian mixture models". In: *IEEE Access* 7 (2018), pp. 6269–6278 (page 22).

[14] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. "Sliced wasserstein distance for learning gaussian mixture models". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3427–3436 (pages 24–26).