INFERENCE IN FUNCTIONAL PROBABILISTIC MODELS

DAESEOK LEE

Abstract.

1. introduciton

Let $\mathcal{X}, \mathcal{Y}$ be metric spaces, $\phi : \mathcal{X} \to \mathcal{Y}$ be a map with a good continuity property, X be a random variable supported in $\mathcal{X}$, and $Y = \phi(X)$. We may ask for a given $y \in \mathcal{Y}$: what is the conditional distribution of X given $Y = y$? Is such a concept even conceivable? If somehow defined well, how can it be computed approximately? Can it be integrated with well-known techniques in probabilistic inferene such as variational inference? In this research, we try to answer these questions. Basically, our answer is based on definition of conditional distribution in terms of "disintegration", explained well in [CP97] and [Tju75].

Before giving the definition, let us guess informally what the definition should be like. First, let's say

$$X \sim \mathcal{N}((0,0), I_2), \phi(x_1, x_2) = \sqrt{x_1^2 + x_2^2}, y = 1$$

If we're asked to describe the conditional distribution $p(\cdot|\phi(X) = y)$, we'd say it is the uniform distribution on the unit circle. This is because it seems plausible that

(1) the distribution is supported on $\{x \in \mathcal{X} | \phi(x) = y\}$
(2) the distribution on the support is roughly proportional to that of the original distribution

The first condition is easily described in the general case. However, the second condition can be quite misleading. If we naively interpret it, we may conclude that the conditional distribution is only determined by the prior distribution of X and the level set $\{x \in \mathcal{X} | \phi(x) = y\}$, which is wrong. In fact, this also depends on how the level set changes near y. After introducing a strategy for calculating disintegration later, we will see a concrete example explaining this.

Therefore, the second condition should be substituted with another one. Let's consider a case where the joint distribution (X, Y) has a probability density function p. It is an easy fact that

$$(2') \ p(x) = \textstyle\int p(y)p(x|y)dy.$$

In fact, this relation is easily generalizable;it is our desired condition. We will see that this is reflected in the definition of disintegration.

Here is summary of the rest of our paper. In the second section, we will present the definition of disintegration and an existence theorem, which are effectively weaker versions of those given in [CP97]. In the third section, we will give a loss function based on Wasserstein distance and prove that under some conditions it is

Date: March 26, 2020.

enough to approximately minimize the loss function in order to obtain an approximate solution of disintegration in the sense of short Wasserstein distance with high probability. Also, we will present a specialized neural network architecture designed to perform well with respect to this loss function and describe the resulting optimization process. In the fourth section, we present another strategy that we call "variational inference with embedding", and present techniques to implement the idea. In the fifth section, we perform an experiment based on mathematically generated examples where we know what the actual disintegration solution looks like, and compare the solution with those approximated by our two strategies given in the preceding two sections. In the sixth section, we survey related previous works. In the last section, we present potential applications of our methods.

2. DISINTEGRATION

The definition and the existence theorem given in this section are far from general, and are rather just specializations of what are given in [CP97].

Definition 2.1. Let $\mathcal{X}, \mathcal{Y}$ be metric spaces, $f : \mathcal{X} \to \mathcal{Y}$ be a Borel measurable function, and P be a Borel measure supported in $\mathcal{X}$. A collection of measures $\{P_y\}_{y \in \mathcal{Y}}$ is called disintegration of P with respect to ϕ, if the following conditions are satisfied.

(1) For all $y \in \mathcal{Y}$,

$$P_y(\phi^{-1}(y)^c) = 0 \tag{2.1}$$

(2) For any Borel set $A \subseteq \mathcal{X}$, the map $y \mapsto P_y(A)$ from $\mathcal{Y}$ to $\mathbb{R}$ is Borel measurable and it satisfies

$$P(A) = \int_{\mathcal{Y}} P_y(A) \phi_* P(dy) \tag{2.2}$$

where $\phi_* P$ refers to the pushforward measure of P induced by ϕ, which satisfies $\phi_* P(B) = P(\phi^{-1}(B))$ for all measurable $B \subseteq \mathcal{Y}$.

Theorem 2.2. *Let $\mathcal{X}$ be a complete seperable metric space, $\mathcal{Y}$ be a separable metric space, $\phi : \mathcal{X} \to \mathcal{Y}$ be a Boreal measurable set and P be any Borel probability measure on $\mathcal{X}$. Then there exists a disintegration of P with respect to ϕ which is unique up to a set of measure zero in $\phi_* P$.*

Here are some comments.

- Actually in [CP97], the measure in 2.2 is not necessarily be the push forward measure, and it is only required that $\phi_* P$ is absolutely continuous with respect to the integrating measure.
- The completeness and separability of $\mathcal{X}$ is only required for it to be a Radon space, on which every Borel probability measure is a Radon measure. A measure is Radon if it is locally finite, inner regular and outer regular.
- The uniqueness in this theorem is only almost surely. However, it should be noted that had the map $y \mapsto P_y$ been continuous on an open set $B \subseteq \mathcal{Y}$ with respect to an appropriate distance in the space of measures, the solution that is continuous on B would have been unique at least on B. Moreover if it is the case $B = \mathbb{R}$, we could conclude that there is a unique continuous solution.

We will take disintegration as definition of conditional distributions for the situation given in the beginning of the first section. However of course, this is only for the special case where the random variable Y is determind as a function of the random variable X. For the general case where (Ω, P) is a probability space and $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ are random variables, we can get the conditional distribution of X given $Y = y$ as follows.

(1) Get $\{P_y\}_{y \in \mathcal{Y}}$, disintegration of P with respect to Y.
(2) Take $\{X_* P_y\}_{y \in \mathcal{Y}}$.

This shows the view of conditional distribution as altering the probability measure on the set of outcomes. In terms of Kolmogorov's definition of conditional expectation, this alteration corresponds to considering a measure μ_y on Ω that satisfies $\mu_y(A) = \mathbb{E}(\mathbf{1}_A|\sigma(Y))(Y^{-1}(y))$ for each measurable $A \subseteq \Omega$. However, here is a serious inherent problem. For each A, the choice of the value $\mathbb{E}(\mathbf{1}_A|\sigma(Y))(Y^{-1}(y))$ can be arbitrary for y in as set of measure 0. We should hope these exceptional sets to be almost identical in order to make μ_y a valid measure at least almost surely , but this is evidently far from being always true. Therefore, conditioning as disintegration requires more useful regularity properties than the Kolmogorov's definition does ([CP97]).

Examples of calculating disintegration are given in [CP97] and [Tju75]. We'd like to conclude this section with a neat connection between disintegration and the concept of conditional probability as usual for the case when there is a probability density function, which will be exploited in the

Theorem 2.3. *Let Y, Z be random variables supported on $\mathcal{Y} \subseteq \mathbb{R}^n, \mathcal{Z} \subseteq \mathbb{R}^m$ respectively, and assume (Y, Z) have a probability density function $p(y, z)$ that satisfies $p(y) > 0$ for each $y \in \mathcal{Y}$. Then probability measures P_y that satisfies*

$$P_y(\{y_0\} \times B) = \begin{cases} \int_B p(y,z)dz/p(y) & y_0 = y \\ 0 & y_0 \neq y \end{cases}$$

makes $\{P_y\}_{y \in \mathcal{Y}}$ into a disintegration of p with respect to the projection $\pi_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}$.

3. A DIRECT LOSS FUNCTION APPROACH

From now on, the main measure of similarity between different probability measures will be Wasserstein distance, which is also called "Earth mover's distance" or "optimal transport distance". For the upcoming proofs, we will use a slightly generalized version for finite measures.

Definition 3.1. Let $\mathcal{X}$ be a metric space and μ_1, μ_2 be finite Borel measures with $\mu_1(\mathcal{X}) = \mu_2(\mathcal{X}) > 0$. Then the Wasserstein distance $W(\mu_1, \mu_2)$ between μ_1 and μ_2 is defined as the infimum of

$$\int_\Omega d(f_1(\omega), f_2(\omega))\mu(d\omega) \tag{3.1}$$

among every choice of finite measure space (Ω, μ) and measurable $f_1, f_2 : \Omega \to \mathcal{X}$ for which $\mu_1 = (f_1)_*\mu$ and $\mu_2 = (f_2)_*\mu$.

The following dual characterization is also known.

Theorem 3.2. *$W(\mu,\nu)$ is equal to the supremum of the following value where $f:\mathcal{X}\to\mathbb{R}$ ranges over all 1-Lipschitz functions.*

$$|\mathbb{E}_{x\sim\mu}f(x)-\mathbb{E}_{x\sim\nu}f(x)|$$

For the convergence properties of this distance, see [ACB17]. For the main theorem in this section, we need two lemmas.

Lemma 3.3. *Let λ be a finite measure on $\mathcal{Y}'$ and let $\{\mu_y\}_{y\in\mathcal{Y}'}$ and $\{\nu_y\}_{y\in\mathcal{Y}'}$ be measurable families of measures on $\mathcal{X}$. Then for any $y_0\in\mathcal{Y}$, we have*

$$W(\int_{\mathcal{Y}'}\mu_y\lambda(dy),\int_{\mathcal{Y}'}\nu_y\lambda(dy))\geq\lambda(\mathcal{Y}')(W(\mu_{y_0},\nu_{y_0})-\sup_{y_1\in\mathcal{Y}'}W(\mu_{y_0},\mu_{y_1})-\sup_{y_2\in\mathcal{Y}'}W(\nu_{y_0},\nu_{y_2}))$$

proof of lemma 3.3. We will use theorem 3.2. Let $\epsilon>0$. We can find a 1-Lipschitz function $f:\mathcal{X}\to\mathbb{R}$ such that

$$\mathbb{E}_{x\sim\mu_{y_0}}f(x)-\mathbb{E}_{x\sim\nu_{y_0}}f(x)\geq W(\mu_{y_0},\nu_{y_0})-\epsilon$$

For any $y\in\mathcal{Y}'$, we have

$$\mathbb{E}_{x\sim\mu_y}f(x)-\mathbb{E}_{x\sim\mu_{y_0}}f(x)\geq -W(\mu_{y_0},\mu_y).$$

Likewise, we also have

$$\mathbb{E}_{x\sim\nu_y}f(x)-\mathbb{E}_{x\sim\nu_{y_0}}f(x)\leq W(\nu_{y_0},\nu_y)$$

Combining these, we get

$$\begin{aligned}&\mathbb{E}_{x\sim\int_{\mathcal{Y}'}\mu_y\lambda(dy)}f(x)-\mathbb{E}_{x\sim\int_{\mathcal{Y}'}\nu_y\lambda(dy)}f(x)=\mathbb{E}_{y\sim\lambda}\mathbb{E}_{x\sim\mu_y}f(x)-\mathbb{E}_{y\sim\lambda}\mathbb{E}_{x\sim\nu_y}f(x)\\&=\lambda(\mathcal{Y}')(\mathbb{E}_{x\sim\mu_{y_0}}f(x)-\mathbb{E}_{x\sim\nu_{y_0}}f(x))+\mathbb{E}_{y\sim\lambda}\mathbb{E}_{x\sim\mu_y-\mu_{y_0}}f(x)-\mathbb{E}_{y\sim\lambda}\mathbb{E}_{x\sim\nu_y-\nu_{y_0}}f(x)\\&\geq\lambda(\mathcal{Y}')(W(\mu_{y_0},\nu_{y_0})-\epsilon)-\mathbb{E}_{y\sim\lambda}W(\mu_{y_0},\mu_y)-\mathbb{E}_{y\sim\lambda}W(\nu_{y_0},\nu_y)\\&\geq\lambda(\mathcal{Y}')(W(\mu_{y_0},\nu_{y_0})-\sup_{y_1\in\mathcal{Y}'}W(\mu_{y_0},\mu_{y_1})-\sup_{y_2\in\mathcal{Y}'}W(\nu_{y_0},\nu_{y_2}))-\lambda(\mathcal{Y}')\epsilon\end{aligned}$$

Since we could have taken $\epsilon>0$ arbitrarily small and found f accordingly, we're done. □

Lemma 3.4. *Let $\mathcal{X}$ be a metric space with diameter $0<D<\infty$. If $\mu_1,\mu_2,\nu_1,\nu_2,\nu_3$ are finite measures on $\mathcal{X}$ supported in $A_1,A_2,B_1,B_2,B_3\subseteq\mathcal{X}$ such that*

(1)
$$\mu_1(\mathcal{X})+\mu_2(\mathcal{X})=\nu_1(\mathcal{X})+\nu_2(\mathcal{X})+\nu_3(\mathcal{X})=1$$
and
$$\mu_1(\mathcal{X})=\nu_1(\mathcal{X})$$

(2)
$$W(\mu_1,\nu_1)\geq a_1 \tag{3.2}$$

(3)
$$\nu_2(\mathcal{X})\leq a_2 \tag{3.3}$$

(4) *There exists $C\subseteq\mathcal{X}$ with $\mu_1(C)\leq a_3$ such that for each $x\in A_1\setminus C$ and each $x'\in B_3$, we have*
$$d(x,x')\geq a_4 \tag{3.4}$$

Then we have

$$W(\mu_1+\mu_2,\nu_1+\nu_2+\nu_3)\geq((a_1/D)-a_2-a_3)/(\frac{1}{D}+\frac{1}{a_4}) \tag{3.5}$$

proof of lemma 3.4. Let (Ω, P) be a measure space and $f, g : \Omega \to \mathcal{X}$ be measurable functions such that $f_*P = \mu_1 + \mu_2$ and $g_*P = \nu_1 + \nu_2 + \nu_3$. By refining (Ω, P, f, g) a little bit if necessary without altering the value 3.1, we may assume that Ω can be subdivided as

$$\Omega = \Omega_{\mu_1} \sqcup \Omega_{\mu_2} = \Omega_{\nu_1} \sqcup \Omega_{\nu_2} \sqcup \Omega_{\nu_3}$$

in such a way that $(f|_{\Omega_{\mu_i}})_*P = \mu_i (i = 1, 2)$ and $(g|_{\Omega_{\nu_j}})_*P = \nu_j (j = 1, 2, 3)$. Let $\Omega_j = \Omega_{\mu_1} \cap \Omega\nu_j$ and $x_j = P(\Omega_j)$ $(j = 1, 2, 3)$.

Since $\mu_1 - (f|_{\Omega_1})_*P$ and $\nu_1 - (g|_{\Omega_1})_*P$ have the same total mass, we can find, for example by taking product measure and projections, a measure space (Ω', P') and measurable $f', g' : \Omega' \to \mathcal{X}$ such that $f'_*P' = \mu_1 - (f|_{\Omega_1})_*P$ and $g'_*P' = \nu_1 - (g|_{\Omega_1})_*P$. Then, let us consider the disjoint union $(\Omega_1 \sqcup \Omega', P \sqcup P', f|_{\Omega_1} \sqcup f', g|_{\Omega_1} \sqcup g')$. This is desgined to satisfy $(f|_{\Omega_1} \sqcup f')_*(P \sqcup P') = \mu_1$ and $(g|_{\Omega_1} \sqcup g')_*(P \sqcup P') = \nu_1$. Therefore, by the definition of Wasserstein distance, we have

$$\begin{aligned}
a_1 &\leq W(\mu_1, \nu_1) \\
&\leq \int_{\Omega_1 \sqcup \Omega'} d((f|_{\Omega_1} \sqcup f')(\omega), (g|_{\Omega_1} \sqcup g')(\omega))d\omega \\
&= \int_{\Omega_1} d(f(\omega), g(\omega))d\omega + \int_{\Omega'} d(f'(\omega), g'(\omega))d\omega \\
&\leq \int_{\Omega_1} d(f(\omega), g(\omega))d\omega + DP'(\Omega') \\
&= \int_{\Omega_1} d(f(\omega), g(\omega))d\omega + D(x_2 + x_3)
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\int_{\Omega_3} d(f(\omega), g(\omega))d\omega &\geq \int_{\Omega_3 \cap f^{-1}((A_1 \setminus C) \cap B_3)} d(f(\omega), g(\omega))d\omega \\
&\geq P(\Omega_3 \cap f^{-1}((A_1 \setminus C) \cap B_3))a_4 \\
&= P(\Omega_3 \cap (f|_{\Omega_{\mu_1}})^{-1}(A_1 \setminus C) \cap (f|_{\Omega_{\mu_3}})^{-1}(B_3)) \\
&\geq (P(\Omega_3) - (P(\Omega_{\mu_1}) - P((f|_{\Omega_{\mu_1}})^{-1}(A_1 \setminus C))) - (P(\Omega_{\nu_3}) - P((f|_{\Omega_{\nu_3}})^{-1}(B_3))))a_4 \\
&= (x_3 + \mu_1(A_1 \setminus C) + \nu_3(B_3) - \mu_1(\mathcal{X}) - \nu_3(\mathcal{X}))a_4 \geq (x_3 - a_3)a_4
\end{aligned}$$

Summing up, we get

$$\begin{aligned}
\int_{\Omega} d(f(\omega), g(\omega))d\omega &\geq \int_{\Omega_1} d(f(\omega), g(\omega))d\omega + \int_{\Omega_3} d(f(\omega), g(\omega))d\omega \\
&\geq \max(0, a_1 - D(x_2 + x_3)) + \max(0, (x_3 - a_3)a_4)
\end{aligned}$$

where x_2 and x_3 satisfies at least $0 \leq x_2 = P(\Omega_{\mu_1} \cap \Omega_{\nu_3}) \leq P(\Omega_{\nu_3}) = \nu_3(\mathcal{X}) \leq a_2$ and $0 \leq x_3$, hence we get

$$\int_{\Omega} d(f(\omega), g(\omega))d\omega \geq \inf_{0 \leq x_2 \leq a_2, 0 \leq x_3} (\max(0, a_1 - D(x_2 + x_3)) + \max(0, x_3 - a_3)a_4)$$

Let $\delta = ((a_1/D) - a_2 - a_3)/(\frac{1}{D} + \frac{1}{a_4})$. Suppose for some x_2, x_3 such that $0 \leq x_2 \leq a_2$ and $0 \leq x_3$, $\max(0, a_1 - D(x_2 + x_3)) + \max(0, x_3 - a_3)a_4 < \delta$ holds. Then $a_1 < D(x_2 + x_3) + \delta$ and $x_3 < a_3 + (\delta/a_4)$, so $a_2 \geq x_2 = (x_2 + x_3) - x_3 > ((a_1/D) - (\delta/D)) - (a_3 + (\delta/a_4)) = (a_1/D) - a_3 - \delta(\frac{1}{D} + \frac{1}{a_4}) = a_2$, which is a contradiction. Therefore, we always have $\max(0, a_1 - D(x_2 + x_3)) + \max(0, x_3 - a_3)a_4 \geq \delta$, hence

$\int_\Omega d(f(\omega), g(\omega))d\omega \geq \delta$. Since this holds for any choice of (Ω, P, f, g), we can conclude $W(\mu_1 + \mu_2, \nu_1 + \nu_2 + \nu_3) \geq \delta$

□

Now we present the main theorem.

Definition 3.5. A function from a metric space to a space of probability measures will be called W- continuous, uniformly W-continuous,... if it is continuous, uniformly continuous,... when the codomain is given Wasserstein distance.

Theorem 3.6. *Let $\mathcal{X}$ be a bounded complete separable metric space, $\mathcal{Y}$ be a separable metric space, P be a Borel probability measure on $\mathcal{X}$, $\phi : \mathcal{X} \to \mathcal{Y}$ be a uniformly continuous funtion, and $\{P_y\}_{y\in\mathcal{Y}}$ be a disintegration of P with respect to ϕ. Suppose the following conditions hold.*

1. *$y \mapsto P_y$ is uniformly W-continuous*
2. *There exists $G_1 : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ with*

$$\lim_{\delta\to 0} G_1(\delta) = 0 \tag{3.6}$$

such that

$$\lim_{\delta\to 0} \phi_* P \{y|\phi_* P(B_y(G_1(\delta))) < \delta\} = 0 \tag{3.7}$$

3. *For any $R > 0$ and $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\sup_{y\in\mathcal{Y}} \phi_* P(B_y(R + \delta) \setminus B_y(R)) \leq \epsilon \tag{3.8}$$

Then given any function $G_2 : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ and $\epsilon_1, \epsilon_2 > 0$, there exists $\delta_1, \delta_2 > 0$ such that whenever a family $\{Q_y\}_{y\in\mathcal{Y}}$ satisfies the following properties,

$$W(P_y, Q_y) \leq \epsilon_2 \tag{3.9}$$

holds for $y \in \mathcal{Y}$ with $\phi_* P$-probability $\geq 1 - \epsilon_1$.

1. *$y \mapsto Q_y$ is uniformly W-continuous, and the degree of continuity is at least as much as that determind by G_2. That is, for any $\epsilon > 0$, for any $y_1, y_2 \in \mathcal{Y}$ with $d_\mathcal{Y}(y_1, y_2) < G_2(\epsilon)$, we have $W(Q_{y_1}, Q_{y_2}) < \epsilon$.*
2.

$$\mathbb{E}_{y\sim\phi_* P}\mathbb{E}_{x\sim Q_y} d_\mathcal{Y}(\phi(x), y) \leq \delta_1 \tag{3.10}$$

3.

$$W(P, \mathbb{E}_{y\sim\phi_* P} Q_y) \leq \delta_2 \tag{3.11}$$

proof of theorem 3.6. Let D be a diameter of $\mathcal{X}$ i.e. for all $x_1, x_2 \in \mathcal{X}$, $d_\mathcal{X}(x_1, x_2) \leq D$.

Choose $R_1 > 0$ such that

1. For any $y_1, y_2 \in \mathcal{Y}$ with $d_\mathcal{Y}(y_1, y_2) \leq R_1$, $W(P_{y_1}, P_{y_2}) \leq \epsilon_2/4$ holds.
2. $R_1 \leq G_2(\epsilon_2/4)$

Let

$$f_1(\delta) = f_2(\delta) = f_3(\delta) = \delta^{1/3} \tag{3.12}$$

and

$$f_4(\delta) = \sup_{y\in\mathcal{Y}} \phi_* P(B_y(R_1 + 2f_3(\delta)) \setminus B_y(R_1)) \tag{3.13}$$

Take $\delta_1 > 0$ such that

(1)

$$2\sqrt{D(f_1(\delta_1)+f_2(\delta_1)+f_4(\delta_1))} < \epsilon_2 \tag{3.14}$$

,which is possible due to the condition 3.8

(2)

$$G_1(\sqrt{D(f_1(\delta_1)+f_2(\delta_1)+f_4(\delta_1)}) < R_1 \tag{3.15}$$

,which is possible since $\lim_{\delta\to 0} G_1(\delta) = 0$.

(3)

(3.16)

$$\phi_*P\left\{y|\phi_*P(B_y(G_1(\sqrt{D(f_1(\delta_1)+f_2(\delta_1)+f_4(\delta_1))}))) < \sqrt{D(f_1(\delta_1)+f_2(\delta_1)+f_4(\delta_1))}\right\} < \epsilon_1$$

,which is possible due to condition 3.7

Now fixing $\delta_2 > 0$, assume that what we want to prove is not true. Later we will choose appropriate δ_2 that leads to a contradiction. If you prefer, you can refer to the later part of the proof to take δ_2 explicitly at this point.

Let $\{Q_y\}_{y\in\mathcal{Y}}$ be any family of measures that satisfies conditions (1)-(3).

First, we find $Y_1 \subset \mathcal{Y}$ with

$$\phi_*P(Y_1) \geq 1 - f_1(\delta_1) \tag{3.17}$$

and $\{X_y\}_{y\in Y_1}$ with $Q_y(X_y) \geq 1 - f_2(\delta_1)$ such that for all $y \in Y_1$ and $x \in X_y$,

$$d_{\mathcal{Y}}(y, \phi(x)) \leq f_3(\delta_1) \tag{3.18}$$

This is possible since otherwise,

$$\mathbb{E}_{y\sim\phi_*P}\mathbb{E}_{x\sim Q_y} d_{\mathcal{Y}}(y,\phi(x)) > f_1(\delta_1)f_2(\delta_1)f_3(\delta_1) = \delta_1$$

Note that this also implies for $y \sim \phi_*P$ and $x \sim Q_y$, $d_{\mathcal{Y}}(y,\phi(x)) \leq f_3(\delta_1)$ holds except with probability $1-(1-f_1(\delta_1))(1-f_2(\delta_1) < f_1(\delta_1)+f_2(\delta_1)$.

Since for $y \sim \phi_*P$ $W(P_y, Q_y) > \epsilon_2$ with probability $> \epsilon_1$ and because of the inequality 3.16, we can find $y_0 \in \mathcal{Y}$ such that

(1)

(3.19)

$$\phi_*P(B_{y_0}(R_1)) \geq \phi_*P(B_{y_0}(G_1(\sqrt{D(f_1(\delta_1)+f_2(\delta_1)+f_4(\delta_1))}))) \geq \sqrt{D(f_1(\delta_1)+f_2(\delta_1)+f_4(\delta_1))}$$

,where the first inequality is true due to 3.15

(2)

$$W(P_{y_0}, Q_{y_0}) > \epsilon_2 \tag{3.20}$$

Now, let

- $R_2 = R_1 + 2f_3(\delta_1)$
- $A_1 = B_{y_0}(R_1)$, $A_2 = B_{y_0}(R_2) \setminus B_{y_0}(R_1)$, $A_3 = \mathcal{Y} \setminus B_{y_0}(R_2)$
- $\mu_1 = \int_{A_1} Q_y\phi_*P(dy)$, $\mu_2 = \int_{A_2\cup A_3} Q_y\phi_*P(dy)$
- $\nu_1 = \int_{A_1} P_y\phi_*P(dy)$, $\nu_2 = \int_{A_2} P_y\phi_*P(dy)$, $\nu_3 = \int_{A_3} P_y\phi_*P(dy)$

We are going to apply lemma 3.4 to find a lower bound of the Wasserstein distance between $\mu_1 + \mu_2 = \int_{\mathcal{Y}} Q_y\phi_*P(dy)$ and $\nu_1+\nu_2+\nu_3 = \int_{\mathcal{Y}} P_y\phi_*P(dy)$. To do so, we find appropriate positive numbers a_1, a_2, a_3 and a_4 appearing in the statement of the lemma.

(1) For any $y_1, y_2 \in A_1$, we have

- Since $d_{\mathcal{Y}}(y_0, y_1) < R_1$, because of the first condition for R_1,
$$W(P_{y_0}, P_{y_1}) \leq \epsilon_2/4$$
- Since $d_{\mathcal{Y}}(y_0, y_2) < R_1$, because of the second condition for R_1,
$$W(Q_{y_0}, Q_{y_2}) \leq \epsilon_2/4$$

Then applying lemma 3.3, we get

$$\begin{aligned} W(\mu_1, \nu_1) &\geq \phi_* P(A_1)(W(P_{y_0}, Q_{y_0}) - \sup_{y_1 \in A_1} W(P_{y_0}, P_{y_1}) - \sup_{y_2 \in A_1} W(Q_{y_0}, Q_{y_2})) \\ &\geq \phi_* P(A_1)(\epsilon_2 - \epsilon_2/4 - \epsilon_2/4) \\ &\geq \phi_* P(A_1)\epsilon_2/2 \\ &> D(f_1(\delta_1) + f_2(\delta_1) + f_4(\delta_1)) \end{aligned}$$

The last inequality follows from 3.19 and 3.14. Thus, we can take

$$a_1 > D(f_1(\delta_1) + f_2(\delta_1) + f_4(\delta_1)) \tag{3.21}$$

(2) Because of the definition 3.13, we have $\nu_2(\mathcal{X}) = \phi_* P(A_2) = \phi_* P(B_{y_0}(R_1 + 2f_3(\delta_1))) - \phi_* P(B_{y_0}(R_1)) \leq f_4(\delta_1)$, so we can take

$$a_2 = f_4(\delta_1) \tag{3.22}$$

(3) By a previous construction, for $y \sim (\phi_* P)|_{A_1}$ and $x \sim Q_y$, except with measure $f_1(\delta_1) + f_2(\delta_1)$ we have $d_{\mathcal{Y}}(y, \phi(x)) \leq f_3(\delta_1)$ hence $d_{\mathcal{Y}}(y_0, \phi(x)) \leq d_{\mathcal{Y}}(y_0, y) + d_{\mathcal{Y}}(y, \phi(x)) \leq R_1 + f_3(\delta_1)$, which means for $x \sim \mu_1$ except with measure $f_1(\delta_1) + f_2(\delta_1)$

$$d_{\mathcal{Y}}(y_0, \phi(x)) \leq R_1 + f_3(\delta_1)$$

holds. On the other hand for $x' \sim \nu_1$, we always have

$$d_{\mathcal{Y}}(y_0, \phi(x')) \geq R_2 = R_1 + 2f_3(\delta_1)$$

For these x and x', we have $d_{\mathcal{Y}}(\phi(x), \phi(x')) \geq d_{\mathcal{Y}}(y_0, \phi(x')) - d_{\mathcal{Y}}(y_0, \phi(x)) \geq f_3(\delta_1)$. Using the uniform continuity of ϕ, we find $a_4 > 0$ such that

$$d_{\mathcal{X}}(x, x') < a_4 \Rightarrow d_{\mathcal{Y}}(\phi(x), \phi(x')) < f_3(\delta_1) \tag{3.23}$$

Then by letting

$$a_3 = f_1(\delta_1) + f_2(\delta_1) \tag{3.24}$$

the last condition of lemma 3.4 is satisfied.

By plugging these a_1, a_2, a_3, a_4 into the equation 3.5, we get

$$W(\mathbb{E}_{y \sim \phi_* P} Q_y, P) = W(\mu_1 + \mu_2, \nu_1 + \nu_2 + \nu_3) \geq ((a_1/D) - a_2 - a_3)/(\frac{1}{D} + \frac{1}{a_4}) > 0$$

Back to the beginning of the proof, letting δ_2 to be the half of this positive number, we find a contradiction. □

Basically this theorem allows us, when some technical conditions hold, to directly attack the loss function defined as

$$\mathcal{L}(\{Q_y\}_{y \in \mathcal{Y}}) = W(P, \mathbb{E}_{y \sim \phi_* P} Q_y) + \lambda \mathbb{E}_{y \sim \phi_* P} \mathbb{E}_{x \sim Q_y} d_{\mathcal{Y}}(\phi(x), y) \tag{3.25}$$

in order to obtain an approximate solution of disintegration, when the sense of approximation is to have small Wasserstein distance with high probability. Some notes on the conditions:

- Of course, P doesn't have to be supported on a set of full dimension .

- With compactness of $\mathcal{Y}$, the uniform W-continuity of $y \mapsto P_y$ can be relaxed to W-continuity.
- With compactness of $\mathcal{Y}$, the condition 3.8 automatically holds.
- The uniform W-continuity regulation on $y \mapsto Q_y$ has a special case of being L-Lipschitz for a given $L > 0$.

From this result, a straightforward implementation strategy can be considered where the Wasserstein distance is calculated either as in [ACB17] using theorem 3.2 or as in [Cut13] using an additional regularizing term.

4. Neural network architecture for the loss function approach

5. Variational inference with embedding

6. Strategies for regularity of embedding

7. Experiments

8. Related works

item

9. Potential applications

[Pea00]

References

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[CP97] Joseph T Chang and David Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[Pea00] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[Tju75] Tue Tjur. *A constructive definition of conditional distributions.* Institute of Mathematical Statistics, University of Copenhagen, 1975.