CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma | Lecture 11
Scribe: Jongho Kim, Jamie Kang | October 29th, 2018

1 Overview

This lecture mainly covers

- Recall the statistical theory of GANs from the last lecture
- Introduce Kontorovich-Rubinstein duality of Wasserstein distance and provide the proof

2 Review

Recall the following notations and two definitions from the last lecture:

$z_1, \ldots, z_n \overset{\text{i.i.d}}{\sim} Z$ where $Z \sim \mathcal{N}(0, I_{k\times k})$

$X = G_\theta(Z)$: G_θ is the generator $\mathbb{R}^k \to \mathbb{R}^d$

$x_1, \ldots, x_n \overset{\text{i.i.d}}{\sim} P$: training examples

$\hat{p}$: Uniform distribution over $\{x_1, \ldots, x_n\}$

p_θ : distribution of $X = G_\theta(Z)$

$\hat{p}_\theta$: uniform distribution over $\{G_\theta(z_1), \ldots, G_\theta(z_m)\}$

Definition 1 (Wasserstein Distance).

$$W_1(P, Q) = \sup_{\|f\|_L \le 1} \mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]$$

Here, the supremum is over functions f where f is 1-Lipschitz continuous function with respect to metric d. For simplicity, we denote this as $\|f\|_L \le 1$. Recall that the function f is 1-Lipschitz continuous if

$$|f(x) - f(y)| \le d(x, y) \qquad \forall x, y$$

Most of cases, we let the metric d to be l_2 distance.

Definition 2 ($\mathcal{F}$-Integral Probability Metric ($\mathcal{F}$-IPM)).

$$W_{\mathcal{F}}(P, Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]|$$

where $\mathcal{F}$ is a family of functions (*e.g.*, a family of neural nets $\{f_\phi\}$).

This week (week 6), we would like to address the following question:

$$\underbrace{W_{\mathcal{F}}(\hat{p}, \hat{p}_\theta)}_{\text{training loss}} \quad \text{small} \quad \Longrightarrow \quad W_1(p, p_\theta) \quad \text{small ?}$$

3 Duality of W_1 (Kontorovich-Rubinstein duality)

Theorem 1. Let P, Q be two distributions over $\mathcal{X}$ (assume P and Q have bounded support) and let d be a metric on $\mathcal{X}$. Then, we have

$$\begin{aligned} W_1(P,Q) &= \sup_{\|f\|_L \le 1} \mathbb{E}_{x\sim P}\left[f(x)\right] - \mathbb{E}_{x\sim Q}\left[f(x)\right] \\ &= \inf_{\mathcal{P}} \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[d(x,y)\right] \end{aligned} \tag{1}$$

where $\mathcal{P}$ is coupling of P and Q. A coupling $\mathcal{P}$ is a joint distribution over $\mathcal{X}\times\mathcal{X}$ whose marginals are P and Q. (*i.e.*, If $(x,y)\sim\mathcal{P}$ then $\mathcal{P}_x = P$ and $\mathcal{P}_y = Q$.)

Equivalently, we can think of coupling as if we are sending mass from $\mathcal{X}$ to $\mathcal{X}$. Suppose we have a finite set $\mathcal{X}$ such that $\mathcal{X} = \{v_1,\dots,v_k\}$. Let P and Q be the two distributions over $\mathcal{X}$ as following:

$$P = (p_1,\dots,p_k) \quad \text{where} \quad \textstyle\sum_i p_i = 1 \quad \text{and} \quad p_i \ge 0 \quad \forall i$$

$$Q = (q_1,\dots,q_k) \quad \text{where} \quad \textstyle\sum_i q_i = 1 \quad \text{and} \quad q_i \ge 0 \quad \forall i$$

Let γ to denote the cost of sending a mass from P to Q. Figure 1 shows the relationship between two distributions and γ.

$$\gamma_{ij} = \mathbf{Prob}\left[x = v_i, y = v_j\right] \quad \text{where} \quad (x,y)\sim\mathcal{P}$$

Note that $\sum_{i=1}^k \gamma_{ij} = p_i$ and $\sum_{j=1}^k \gamma_{ij} = q_i$. Using this interpretation, we can think of the Wasserstein distance of P and Q as the cost of transportation. We can write it in terms of following mathematical notation:

$$W_1(P,Q) = \inf_{\mathcal{P}} \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[d(x,y)\right] = \sum_{i,j}\gamma_{ij} d(v_i,v_j)$$

and we want to minimize this cost.

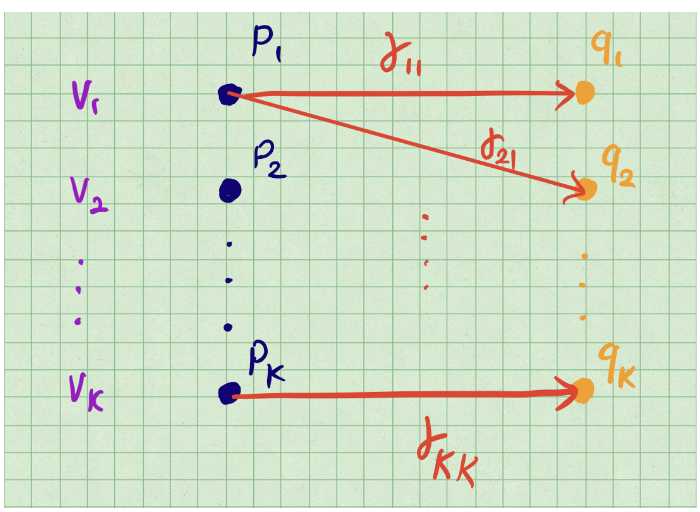

Figure 1: Diagram representation of coupling

Last lecture, we shortly discussed why other distance measures (*e.g.*, Total Variation (TV) distance and Kullback-Leibler(KL) divergence) are not useful in our setting compared to Wasserstein distance. Before we move on to the formal proof of Theorem 1, here we discuss and compare Wasserstein distance with TV distance and KL divergence.

Definition 3 (Total Variation distance).

$$\begin{aligned} \mathrm{TV}(P,Q) &= \sup_{0\le f\le 1} \left|\mathbb{E}_{x\sim P}\left[f(x)\right] - \mathbb{E}_{x\sim Q}\left[f(x)\right]\right| && (\textit{All cases}) \\ &= \frac{1}{2}\int |p(x)-q(x)|\,dx \quad \text{where } P \text{ and } Q \text{ have densities } p \text{ and } q && (\textit{Continuous case}) \end{aligned}$$

Definition 4 (Kullback-Leibler divergence).

$$\begin{aligned} \mathrm{KL}(P,Q) &= \mathbb{E}_{x\sim P}\left[\log\frac{p(x)}{q(x)}\right] = \sum_x P(x)\log\frac{P(x)}{Q(x)} && (\textit{Discrete case}) \\ &= \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx && (\textit{Continuous case}) \end{aligned}$$

Example 1. Suppose P is uniform distribution over sphere and Q is also uniform distribution over sphere which is shifted by ϵ on only one coordinate. Let the metric $d = l_2$. Figure 2 shows two distributions of P and Q.

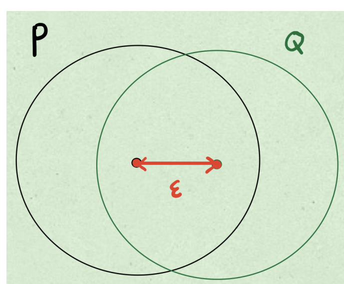

Figure 2: Distributions of P and Q

In this case, $\mathrm{TV}(P,Q) = 1$ because we can construct the function f as following:

$$\begin{aligned} f(x) = 1 \quad &\text{if} \quad X \in \mathrm{support}(P) \\ f(x) = 0 \quad &\text{if} \quad X \in \mathrm{support}(Q) \end{aligned}$$

Note that $\mathrm{KL}(P,Q)$ is not well-defined in this setting because not both P and Q defined on the same probability space. On the other hand, $W_1(P,Q) \le \epsilon$. To see why this is the case, we first construct a coupling as following sampling technique:

$$(x,y)\sim\mathcal{P} \quad \Longleftrightarrow \quad x\sim P, y = x+\epsilon\begin{bmatrix}1\\0\end{bmatrix}$$

Note that the marginal distribution would be P and Q. Here, we achieve

$$\begin{aligned}
W_1(P,Q) &= \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[d(x,y)\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\|x-y\|_2\right] \\
&= \epsilon & \text{since } x-y=\epsilon\cdot\begin{bmatrix}1\\0\end{bmatrix}
\end{aligned}$$

Example 2. We state the following claim:

Claim 1. $\mathrm{TV}(P,Q) = W_1(P,Q)$ with respect to the trivial metric $d(x,y) = \mathbb{I}\{x \neq y\}$ (Note: This claim explains why TV distance does not take into account geometry.)

Proof. f is 1-Lipschitz with respect to the metric $d(x,y) = \mathbb{I}\{x \neq y\}$. We claim that

$$|f(x) - f(y)| \leq \mathbb{I}\{x \neq y\} \quad \iff \quad \sup_{x\in\mathcal{X}} f(x) - \inf(f) \leq 1$$

To see this, first note that from $|f(x) - f(y)| \leq I\{x \neq y\}$, we get that $|f(x) - f(y)| \leq 1$ if $x \neq y$. Thus, if the elements obtaining the sup and inf are not the same, then the difference is bounded by 1, otherwise the difference will be 0. Furthermore, if $\sup_{x\in\mathcal{X}} f(x) - \inf(f) \leq 1$, it immediately follows that $|f(x) - f(y)| \leq 1 \ \forall x, y$. This establishes the equivalence.

From here, we denote $\mathbb{E}_{x\sim P}\left[f(x)\right] = \mathbb{E}_{x\sim P}f$ and $\mathbb{E}_{x\sim Q}\left[f(x)\right] = \mathbb{E}_{x\sim Q}f$ for simplicity. Then,

$$\begin{aligned}
W_1(P,Q) &= \sup_{f:\sup f - \inf f \leq 1} \underbrace{\mathbb{E}_{x\sim P}f - \mathbb{E}_{x\sim Q}f}_{\text{shift invariant}} \\
&= \sup_{f:0\leq f\leq 1} \mathbb{E}_{x\sim P}f - \mathbb{E}_{x\sim Q}f \\
&= \max_{\substack{f:0\leq f\leq 1,\\ 0\leq 1-f\leq 1}} \left\{\mathbb{E}_{x\sim P}f - \mathbb{E}_{x\sim Q}f, \quad \mathbb{E}_{x\sim P}(1-f) - \mathbb{E}_{x\sim Q}(1-f)\right\} \\
&= \sup_{0\leq f\leq 1} \left|\mathbb{E}_{x\sim P}f - \mathbb{E}_{x\sim Q}f\right| \\
&= TV(P,Q)
\end{aligned}$$

where the second equality follows from the fact that $\mathbb{E}_{x\sim P}f - \mathbb{E}_{x\sim Q}f$ is shift invariant (*i.e.*, $\mathbb{E}_P[f+c] - \mathbb{E}_Q[f+c] = \mathbb{E}_P f + c - \mathbb{E}_Q f - c = \mathbb{E}_P f - \mathbb{E}_Q f$), and thus we can subtract inf f. The fourth equality is again using the fact that $\mathbb{E}_{x\sim P}(1-f) - \mathbb{E}_{x\sim Q}(1-f)$ is shift invariant and thus we can remove 1 and have the second term become $\mathbb{E}_P(-f) - \mathbb{E}_Q(-f)$.

3.1 Proof of Kontorovich-Rubinstein duality

We use LP duality in fuctional space with f as variable to prove **Theorem 1**.

Assume $\mathcal{X}$ is finite: $\mathcal{X} = \{v_1, ..., v_k\}$ and $f(v_1), ..., f(v_k) \triangleq (f_1, ..., f_k)$ and recall that $P = (p_1, ..., p_k), Q = (q_1, ..., q_k)$. As usual, we assume $p_i > 0, q_i > 0$ for all i. We construct the following linear programming problems:

- **Program (A)**:

$$\begin{aligned}
\underset{f=(f_1,\ldots,f_k)}{\text{maximize}} \quad & \sum_{i=1}^{k} p_i f_i - \sum_{i=1}^{k} q_i f_i && (\mathrm{OPT}_1)\\
\text{subject to} \quad & f(x) - f(y) \le d(x,y) \quad \forall x, y \in \mathcal{X} \\
& (i.e.,\ f(v_i) - f(v_j) \le d(v_i, v_j) \iff f_i - f_j \le d_{ij})
\end{aligned}$$

Note that OPT_1 is the expanded version of the objective function: $\max_f \mathbb{E}_P f - \mathbb{E}_Q f$.

- **Program (P)**:

$$\begin{aligned}
\underset{\gamma}{\text{minimize}} \quad & \sum_{i,j} \gamma_{ij} d_{ij} && (\mathrm{OPT}_2)\\
\text{subject to} \quad & \sum_{j} \gamma_{ij} = p_i \quad \forall i && (\text{dual: } f_i)\\
& \sum_{i} \gamma_{ij} = q_j \quad \forall j && (\text{dual: } -g_j)\\
& \gamma_{ij} \ge 0 \quad \forall i, j && (\text{dual: } w_{ij})
\end{aligned}$$

Goal. We want to prove that $\mathrm{OPT}_1 = \mathrm{OPT}_2$.

Lemma 1 ($\mathrm{OPT}_1 \le \mathrm{OPT}_2$).
Proof. Let f, γ be optimal solutions to **(A)** and **(P)** respectively. Then,

$$\begin{aligned}
\mathrm{OPT}_1 &= \sum_i (p_i - q_i) f_i \\
&= \sum_i (\sum_j \gamma_{ij}) f_i - \sum_i (\sum_j \gamma_{ji}) f_i && (\text{by definition of } p_i \text{ and } q_i)\\
&= \sum_i (\sum_j \gamma_{ij}) f_i - \sum_j (\sum_i \gamma_{ij}) f_j \\
&= \sum_{i,j} \gamma_{ij} (f_i - f_j) \\
&\le \sum_{i,j} \gamma_{ij} d_{ij} = \mathrm{OPT}_2.
\end{aligned}$$

Next, we construct the dual of **(P)** using dual variables specified in parentheses in **(P)**:

- **Program (D)**:

$$\begin{aligned}
\underset{f,g}{\text{maximize}} \quad & \sum_i p_i f_i - \sum_i q_i g_i && (\mathrm{OPT}_3)\\
\text{subject to} \quad & f_i - g_j + w_{ij} = d_{ij} \quad \forall i, j
\end{aligned}$$

Note that the constraint can be re-written as $f_i - g_j \le d_{ij} \quad \forall i, j$ because w_{ij} is non-negative by definition of dual variable.

Lemma 2 ($\mathrm{OPT}_1 \geq \mathrm{OPT}_3$).
Proof. The main idea here is to use the fact that d is a metric. Let f, g, γ be optimal solutions to **(P)** and **(D)**. Then, by *complementary slackness*[1],

$$f_i - g_j = d_{ij} \quad \text{if} \quad \gamma_{ij} > 0 \quad \forall i, j.$$

First, we show that $g_j - g_t \leq d_{jt} \quad \forall j, t$ by following

$$\begin{aligned} \forall j, \text{ pick } i \text{ such that } \gamma_{ij} > 0 \quad &\Longrightarrow \quad f_i - g_j = d_{ij} && (1) \\ \text{Also, by constraint: } \forall i, t \quad & f_i - g_t \leq d_{it} && (2) \end{aligned}$$

$(1) - (2)$ yields: $g_j - g_t \leq d_{it} - d_{ij} \leq d_{jt}$ where the last inequality is from triangle inequality as d is a metric. Therefore, g is a feasible solution of **(A)**. Note that from **(D)**'s constraint, $f_i - g_i \leq d_{ii}$ and $d_{ii} = 0$ since it's a metric. Therefore, $f_i \leq g_i$ for all i. Using this, we obtain:

$$\begin{aligned} \mathrm{OPT}_1 &\geq \sum_i p_i g_i - \sum_i q_i g_i && \text{(feasibility of } g\text{)} \\ &\geq \sum_i p_i f_i - \sum_i q_i g_i && (f_i \leq g_i) \\ &= \mathrm{OPT}_3. \end{aligned}$$

Finally, by **Lemma 1** and **Lemma 2** and the fact that $\mathrm{OPT}_2 = \mathrm{OPT}_3$[2], we conclude

$$\mathrm{OPT}_1 = \mathrm{OPT}_2 = \mathrm{OPT}_3.$$

4 Plans for future lectures

Recall that the main question we aim to address is:

$$\underbrace{W_{\mathcal{F}}(\hat{p}, \hat{p}_\theta)}_{\text{training loss}} \quad \text{small} \quad \Longrightarrow \quad W_1(p, p_\theta) \quad \text{small ?}$$

Our plan for addressing this question is following:

$$\textit{Plan:} \quad W_{\mathcal{F}}(\hat{p}, \hat{p}_\theta) \quad \xleftrightarrow{\text{generalization}} \quad W_{\mathcal{F}}(p, p_\theta) \quad \xleftrightarrow{\text{approx.} W_1 \approx W_{\mathcal{F}}} \quad W_1(p, p_\theta).$$

Hence, in the next lecture we will show three observations:

1. If $\mathcal{F}$ is complex $\Longrightarrow$ generalization is **bad**
2. If $\mathcal{F}$ has small complexity $\Longrightarrow$ generalization is **good**
3. If $\mathcal{F}$ has small complexity $\Longrightarrow$ approximation **may not be good**

and explore ways to overcome this dilemma.

[1] If you are not familiar with this part, please refer to Chapter 5 of *Convex Optimization* [**?**] and/or EE 364A course lecture slide 17 on `http://web.stanford.edu/class/ee364a/lectures/duality.pdf`

[2] Since OPT_3 is dual of OPT_2 and *strong duality* holds for linear programming, $\mathrm{OPT}_2 = \mathrm{OPT}_3$.