CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Garrett Thomas, Pengda Liu

Lecture #12
October 31, 2018

1 Review and Overview

Recall the GAN setup: we have independent samples $x_1, \dots, x_n$ drawn from some true (unknown) distibution P. Let $\hat{P}$ be the uniform distribution over these samples.

We assume a latent variable $Z \sim P_Z$, for example $P_Z = \mathcal{N}(0, I)$. Let P_θ be the distribution of $G_\theta(Z)$. Our goal is to learn θ so that P_θ approximates P. For a fixed set of samples $z_1, \dots, z_n \sim P_Z$, we define $\hat{P}_\theta$ to be the uniform distribution over $\{G_\theta(z_1), \dots, G_\theta(z_n)\}$.

In training, we will minimize the integral probability metric

$$W_{\mathcal{F}}(\hat{P}_\theta, \hat{P})$$

where $\mathcal{F} = \{f_\phi : \phi \in \mathbb{R}^p\}$.

Ideally, a small $W_{\mathcal{F}}(\hat{P}_\theta, \hat{P})$ would guarantee a small $W_1(P_\theta, P)$, and conversely a large $W_{\mathcal{F}}(\hat{P}_\theta, \hat{P})$ would guarantee a large $W_1(P_\theta, P)$. That way, we know that the quantity being optimized (the empirical IPM) tells us something about the quantity we really care about (the Wasserstein distance of the true distributions).

Our approach is to relate the following quantities:

$$W_{\mathcal{F}}(\hat{P}_\theta, \hat{P}) \longleftrightarrow W_{\mathcal{F}}(P_\theta, P) \longleftrightarrow W_1(P_\theta, P)$$

The first arrow is a question of *generalization*: is the empirical IPM close to the population IPM? The second arrow is a question of *approximation*: is the population IPM close to the population Wasserstein distance?

We will see that the answers depend on the complexity of the generator class $\mathcal{F}$.

2 What happens when $\mathcal{F}$ is "too complex"?

Lemma 1. *Suppose $\mathcal{F}$ is the set of all 1-Lipschitz functions. (Note this means that $W_{\mathcal{F}} = W_1$.) Assume $n = \text{poly}(d)$. Then there exist distributions P and Q such that $W_1(P, Q) = W_{\mathcal{F}}(P, Q) = 0$ (and therefore $P = Q$), but with high probability, $W_{\mathcal{F}}(\hat{P}, \hat{Q}) \gtrsim 1$, where $\hat{P}$ and $\hat{Q}$ are uniform distributions over fixed sets of independent samples $u_1, \dots, u_n \sim P$ and $v_1, \dots, v_n \sim Q$.*

This lemma implies an undesirable result: if $\mathcal{F}$ is too rich, it is possible to learn the true distribution (in the sense that $P = Q$) without realizing it (in the sense that $W_{\mathcal{F}}(\hat{P}, \hat{Q})$ is large).

Proof of lemma. Let $V = \{\pm\frac{1}{\sqrt{d}}\}^d$ be the vertices of a hypercube in d-dimensions. Let P be a uniform distribution over V, and $P = Q$. Then immediately

$$W_{\mathcal{F}}(P, Q) = W_1(P, Q) = 0$$

Now let $\hat{P} = \text{Uniform}\{u_1, \dots, u_n\}$ and $\hat{Q} = \text{Uniform}\{v_1, \dots, v_n\}$ where $u_1, \dots, u_n, v_1, \dots, v_n$ are sampled independently from P.

Our claim is that random vectors from P have an inner product bounded like $\lesssim \frac{1}{\sqrt{d}}$. More precisely, if u and v are independent samples from P, then for every $c \geq 0$

$$\mathbb{P}\left(|u^\top v| \geq \sqrt{\frac{2c\log d}{d}}\right) \leq 2d^{-c}$$

Proof of claim. Write $u^\top v = \sum_{i=1}^d u_i v_i$. Note that for each u_i, v_i, we have

$$\mathbb{E}[u_i] = \frac{1}{2}\frac{1}{\sqrt{d}} + \frac{1}{2}\left(-\frac{1}{\sqrt{d}}\right) = 0$$

and likewise for v_i. This implies that

$$\mathbb{E}[u^\top v] = \sum_{i=1}^d \mathbb{E}[u_i v_i] = \sum_{i=1}^d \mathbb{E}[u_i]\mathbb{E}[v_i] = 0$$

where we have used the independence of u_i and v_i to factor the expectation. Moreover, since $u_i v_i \in [-\frac{1}{d}, \frac{1}{d}]$, we can apply Hoeffding's inequality to obtain

$$\mathbb{P}(|u^\top v| \geq t) = \mathbb{P}(|u^\top v - \mathbb{E}[u^\top v]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^d \left(\frac{2}{d}\right)^2}\right) = 2\exp\left(-\frac{dt^2}{2}\right)$$

for any $t \geq 0$. Taking $t = \sqrt{\frac{2c\log d}{d}}$, we obtain

$$\mathbb{P}\left(|u^\top v| \geq \sqrt{\frac{2c\log d}{d}}\right) \leq 2\exp\left(-\frac{2dc\log d}{2d}\right) = 2d^{-c}$$

as stated. □

By a union bound over all n^2 pairs (u_i, v_j), it follows that

$$\mathbb{P}\left(\forall i,j,\ |u_i^\top v_j| \leq \sqrt{\frac{2c\log d}{d}}\right) \geq 1 - 2n^2 d^{-c}$$

for any $c \geq 0$. Then if $n \leq d^{O(1)}$ (polynomial sample complexity), we have with high probability that

$$\forall i,j,\ |u_i^\top v_j| \leq O\left(\sqrt{\frac{\log d}{d}}\right)$$

which implies

$$\|u_i - v_j\|_2^2 = \|u_i\|_2^2 + \|v_j\|_2^2 - 2u_i^\top v_j \geq 2 - O\left(\sqrt{\frac{\log d}{d}}\right) \gtrsim 1$$

Now let Γ be a coupling of $\hat{P}$ and $\hat{Q}$. Then with high probability, $(x, y) \sim \Gamma$ satisfy $\|x - y\|_2 \gtrsim 1$, so

$$\mathbb{E}_{(x,y)\sim\Gamma}\|x - y\|_2 \gtrsim 1$$

Conditioned on $\|u_i - v_j\|_2^2 \gtrsim 1$ for all i, j (which is a high probability event), for any Γ,

$$W_{\mathcal{F}}(\hat{P}, \hat{Q}) = W_1(\hat{P}, \hat{Q}) = \inf_\Gamma \mathbb{E}_{(x,y)\sim\Gamma}\|x - y\|_2 \gtrsim 1$$

which proves the lemma. □

3 What happens when $\mathcal{F}$ is "too simple"?

3.1 Good generalization

Theorem 1. *(Heuristical) For any fixed P, Q, with high probability over the randomness of $\hat{P}, \hat{Q}$, we have $W_{\mathcal{F}}(\hat{P}, \hat{Q}) - W_{\mathcal{F}}(P, Q) \lesssim R_n(\mathcal{F})$*

Remark 1. *This theorem is not enough, we need a theorem that is more uniform convergence, that is we have bounded difference for any fixed P and any Q.*

Definition 1. *Let $G = \{P_\theta\}$ be all possible generated distributions, assume $0 \in \mathcal{F}$. For any $P \in G$, define*

$$R_n(\mathcal{F}, P) = \mathbb{E}_{z_1,\dots,z_n \sim P}[\mathbb{E}_\sigma[\frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(z_i)]]$$

and define

$$R_n(\mathcal{F}, G) = \sup_{P\in G} R_n(\mathcal{F}, P)$$

Recall that over the randomness of training examples, we have $\forall\theta, \hat{L}(\theta) \approx L(\theta)$ thus we can apply something similar here by redefining the training error as $\mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P}, \hat{Q})]$. More specifically, we introduce the following theorem.

Theorem 2. *Assume that $\forall f \in \mathcal{F}, ||f||_\infty \le M$, then for fixed $P \in G$ with probability $\ge 1-\delta$, over the randomness of $\hat{P}$, $\forall Q \in G$, we have*

$$|W_{\mathcal{F}}(P, Q) - \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P}, \hat{Q})]| \lesssim R_n(\mathcal{F}, G) + M\sqrt{\frac{\log 2/\delta}{n}}$$

Proof.

$$W_{\mathcal{F}}(P, Q) - \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P}, \hat{Q})] = \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(P, Q) - W_{\mathcal{F}}(\hat{P}, \hat{Q})]$$

$$(\textit{triangle inequality}\ ) \le \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(P, \hat{P}) + W_{\mathcal{F}}(\hat{P}, \hat{Q}) + W_{\mathcal{F}}(\hat{Q}, Q) - W_{\mathcal{F}}(\hat{P}, \hat{Q})]$$

$$= \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(P, \hat{P}) + W_{\mathcal{F}}(\hat{Q}, Q)] = W_{\mathcal{F}}(P, \hat{P}) + \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{Q}, Q)]$$

Thus we have

$$W_{\mathcal{F}}(P, Q) - \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P}, \hat{Q})] \le W_{\mathcal{F}}(P, \hat{P}) + \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{Q}, Q)] \tag{1}$$

We bound those two terms individually. By lemma 1 in lecture note 5:

$$\mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{Q}, Q)] = \mathbb{E}_{z_1,\dots,z_n\sim\hat{Q}}\Big[\sup_{f\in\mathcal{F}}\big|\mathbb{E}_{z\sim Q}[f(z)] - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\big|\Big] \le R_n(\mathcal{F}, Q) \tag{2}$$

Also, by theorem 2 in lecture note 6, with probability $\ge 1-\delta$

$$W_{\mathcal{F}}(P, \hat{P}) \lesssim R_n(\mathcal{F}, P) + M\sqrt{\frac{\log 2/\delta}{n}} \tag{3}$$

Then by (1),(2),(3) we have

$$W_{\mathcal{F}}(P, Q) - \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P}, \hat{Q})] \lesssim R_n(\mathcal{F}, Q) + R_n(\mathcal{F}, P) + M\sqrt{\frac{\log 2/\delta}{n}} \lesssim R_n(\mathcal{F}, G) + M\sqrt{\frac{\log 2/\delta}{n}}$$

Similarly, we can show that

$$\mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P},\hat{Q})] - W_{\mathcal{F}}(P,Q) \lesssim R_n(\mathcal{F},G) + M\sqrt{\frac{\log 2/\delta}{n}}$$

Thus

$$|W_{\mathcal{F}}(P,Q) - \mathbb{E}_{\hat{Q}}[W_{\mathcal{F}}(\hat{P},\hat{Q})]| \lesssim R_n(\mathcal{F},G) + M\sqrt{\frac{\log 2/\delta}{n}}$$

□

3.2 (Maybe) bad approximation

We introduce the following lemma regarding the approximation quality of $\mathcal{F}$ with small complexity.

Lemma 2. *Assume P uniform over $\{\pm\frac{1}{\sqrt{d}}\}^d$, suppose $R_n(\mathcal{F},G) \le \frac{c}{\sqrt{n}}$ for some constant c, then $\forall \epsilon > 1/poly(d)$, there is Q such that $W_{\mathcal{F}}(P,Q) \le \epsilon$ but $W_1(P,Q) \gtrsim 1$.*

Proof. Take $m \ge \frac{c}{\epsilon^2}$ and $Q = \hat{P}^m$ uniform over $\{x_1,...,x_m\}$ where each $x_i \sim P$. Then by (3), we have, if we pick some large enough m, ignoring the $M\sqrt{\frac{\log 2/\delta}{m}}$ term, we have

$$W_{\mathcal{F}}(P,Q) = W_{\mathcal{F}}(P,\hat{P}) \le R_m(\mathcal{F},Q) \le \frac{c}{\sqrt{m}} \le \epsilon$$

We also have

$$W_1(P,Q) = W_1(P,\hat{P}^m) = \inf_p \mathbb{E}_{(x,y)\sim p}[||x-y||_2]$$

Since $\epsilon \ge 1/poly(d)$, we only require $m = poly(d)$. Furthermore, we note that $E_{x\sim P}[\|x - x_i\|_2^2] = 2$, and $\|x - x_i\|_2^2$ is the sum of d independent random variables. Therefore, $\Pr_{x\sim P}(\|x - x_i\|_2 \le 1)$ is exponentially small in d (see Lemma 3 for a formal statement). Therefore, we can union bound over all $m = poly(d)$ choices of i to get that

$$\Pr_{x\sim P}(\forall i, ||x - x_i||_2 \ge 1) \ge \frac{1}{2}$$

which gives that for any coupling p of P and Q,

$$\Pr_{x\sim P}(\Pr_{y\sim p(y|x)}(||x-y||_2 \ge 1)) \ge \frac{1}{2}$$

Thus

$$W_1(P,Q) = \inf_p \mathbb{E}_{(x,y)\sim p}[||x-y||_2] \ge \frac{1}{2}$$

□

3.2.1 Points are "widespread" in high dimension

Lemma 3. *Let $x \sim \mathrm{Unif}(\{\pm\frac{1}{\sqrt{d}}\}^d)$ and y be an arbitrary fixed vector in $\{\pm\frac{1}{\sqrt{d}}\}^d$. Then we have*

$$\Pr(\|x-y\|_2 \le 1) \le \exp(-d/8).$$

Proof. By the fact that $\|x-y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle = 2 - 2\langle x, y\rangle$, we have

$$\Pr(\|x-y\|_2 \leq 1) = \Pr(\langle x, y\rangle \geq 1/2).$$

Now, regardless of the value of y, each variable $x_i y_i$ is uniformly distributed in $\{\pm\frac{1}{d}\}$, and is thus mean-zero and sub-Gaussian with variance proxy $\frac{1}{d^2}$. As the coordinates are further independent, we get that $\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i$ is mean-zero with sub-Gaussian with variance proxy $1/d$. Applying the sub-Gaussian concentration, we get

$$\Pr(\langle x, y\rangle \geq 1/2) \leq \exp\left(-\frac{(1/2)^2}{2 \cdot 1/d}\right) = \exp(-d/8).$$

□