CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Yining Chen and Melody Guan

Lecture # 13
Nov 05, 2018

1 Review and Overview

Last lecture, we introduced the following dilemma:

If F is of high complexity:

- Generalization is bad, i.e., $W_F(\hat{P}, \hat{P}_\theta) \not\approx W_F(P, P_\theta)$.

If F is of low complexity:

- Generalization is fine.
- However, there is an issue with approximation, i.e., $\exists P, Q$ s.t. $W_F(P, Q) \not\approx W_1(P, Q)$.

In this lecture we will wrap up our discussion about GANs and introduce online learning and bandit problems.

2 Restricted Approximability

Is the approximation issue when F is of low complexity actually a problem? We make the observation that the counterexample P, Q may not be realizable. Namely, even though we know that $\exists Q$ s.t. $W_F(P, Q) \not\approx W_1(P, Q)$, perhaps it is possible that $\forall Q \in G = \{P_\theta\}$, $W_F(P, Q) \approx W_1(P, Q)$? The answer is *yes*, at least in some cases[1]. As such, we can restrict our restrict our attention to evaluating approximability for those distributions that are actually in the generated class.

Concretely, we are looking for statements such as: $\exists$ some fixed c, c',

$$\forall \theta, W_1(P, P_\theta)^c \lesssim W_F(P, P_\theta) \lesssim W_1(P, P_\theta)^{c'}.$$

Here $\lesssim$ may hide not only universal constants but also hyper-parameters such as the dimension of data or the diameter of the distribution.

If this sandwiching[2] statement is true, then the Wasserstein distance is close to the IMP, which we know from lecture 12 to be generalizable if the function class is small.

We now prove a theorem of this flavor.

Assumptions

We make the following three assumptions:

1. $P \in G$.

[1] The answer is probably also true in general but we don't know how to prove the general case yet.

[2] Note that if we only had $W_1(P, P_\theta)^c \lesssim W_F(P, P_\theta)$ or $W_F(P, P_\theta) \lesssim W_1(P, P_\theta)^{c'}$ then we can cheat by scaling up every $f \in F$.

2. $\forall\theta$, P_θ has density p_θ.

3. $\forall\theta_1, \theta_2, \log p_{\theta_1} - \log p_{\theta_2} \in F$.

The last assumption says the discriminator class is powerful enough. In some cases (such as an injective neural network), we can enforce the last assumption. If the network can approximate the logarithm of density then we can add another layer to enable the network to approximate differences between log densities.

Claims

We make the following claims for all $P, Q \in G$:

1. $KL(P||Q) + KL(Q||P) \leq W_F(P, Q)$.

2. If we additionally assume that $\forall f \in F$, f is L-Lipschitz,then $W_F(P, Q) \leq L \cdot W_1(P, Q)$.

3. Suppose χ (the probability space of P, Q) is bounded, i.e., $\forall x \in \chi, ||x||_2 \leq D$, then $W_1(P, Q)^2 \lesssim D^2 \cdot KL(P||Q)$, where $\lesssim$ hides only universal constants.

4. If we additionally assume that $\forall\theta$, the parameterized generator function G_θ is K-Lipschitz, then we have $W_1(P, Q)^2 \lesssim K^2 \cdot KL(P||Q)$.
Comment: Note this bound now depends on the property of the generator. Intuitively, the "nicer" (more Lipschitz) the generator is, the better the bound gets. This bound improves upon the bound in **Claim 3** (proving that $K^2 \lesssim D^2$ is beyond the scope of this lecture).

5. Suppose $\forall x \in \chi, ||x||_2 \leq D$, and $\forall f \in F$, f is L-Lipschitz, then

$$W_1(P, Q)^2 \lesssim D^2 \cdot W_F(P, Q) \lesssim D^2 L \cdot W_1(P, Q).$$

Proof Sketches for Claims

1. Suppose P, Q have densities p, q.

$$\begin{aligned} W_F(P, Q) &= \sup_{f \in F} |\mathbb{E}_P f - \mathbb{E}_Q f| \\ &\geq \mathbb{E}_P[\log p - \log q] - \mathbb{E}_Q[\log p - \log q] \\ &= \mathbb{E}_P[\log \left(\frac{p}{q}\right)] + \mathbb{E}_Q[\log \left(\frac{q}{p}\right)] \\ &= KL(P||Q) + KL(Q||P). \end{aligned}$$

   The inequality is true because $\log p - \log q \in F$.

2. $$W_F(P, Q) = \sup_{f \in F} |\mathbb{E}_P f - \mathbb{E}_Q f| \leq \sup_{f \text{ is } L\text{-Lipschitz}} |\mathbb{E}_P f - \mathbb{E}_Q f| = L \cdot W_1(P, Q).$$

   The inequality is true because F is a family of L-Lipschitz functions. Taking a supremum over F must be upper bounded by taking a supremum over the larger set of all L-Lipschitz functions.

3. Recall that the definition of $W_1(P,Q)$ is $\sup_{f \text{ is } 1-\text{Lipschitz}} |\mathbb{E}_P f - \mathbb{E}_Q f|$. Suppose there is a 1-Lipschitz function f^* such that

$$|\mathbb{E}_P f^* - \mathbb{E}_Q f^*| = \sup_{f \text{ is } 1-\text{Lipschitz}} |\mathbb{E}_P f - \mathbb{E}_Q f|.$$

Such f^* may not exist, so in a more rigorous proof we will have to consider a sequence of functions and take the limit. For the sake of simplicity, in this proof sketch, we assume such f^* exists. Similarly, suppose $\exists x_1, x_2 \in \chi$ such that

$$f^*(x_1) = \sup_{x \in \chi} f^*(x),$$

$$f^*(x_2) = \inf_{x \in \chi} f^*(x).$$

Since f^* is 1-Lipschitz on χ,

$$\sup_{x \in \chi} f^*(x) - \inf_{x \in \chi} f^*(x) \leq |x_1 - x_2| \leq 2D.$$

Recall that $\mathbb{E}_P f - \mathbb{E}_Q f$ is shift-invariant. We can shift f^* so that its range lies in $[0, 2D]$ without changing the value of $\mathbb{E}_P f^* - \mathbb{E}_Q f^*$:

$$W_1(P,Q) = |\mathbb{E}_P f^* - \mathbb{E}_Q f^*| \leq \sup_{0 \leq f \leq 2D} |\mathbb{E}_p f - \mathbb{E}_Q f|.$$

Recall that definition of Total Variation distance is $TV(P,Q) = \sup_{0 \leq f \leq 1} |\mathbb{E}_p f - \mathbb{E}_Q f|$, so

$$W_1(P,Q) \leq \sup_{0 \leq f \leq 2D} |\mathbb{E}_p f - \mathbb{E}_Q f| \leq 2D \cdot TV(P,Q).$$

By Pinsker's Inequality,

$$TV(P,Q) \lesssim \sqrt{KL(P||Q)}.$$

Therefore

$$W_1(P,Q) \lesssim 2D \cdot \sqrt{KL(P||Q)} \iff W_1(P,Q)^2 \lesssim D^2 \cdot KL(P||Q).$$

4. To prove this claim, we need two more theorems:

Theorem 1 (Bobkov-Gotze Theorem) *Suppose distribution P satisfies that for all 1-Lipschitz functions f, $x \sim P$,$f(x)$ is a sub-Gaussian distribution with variance proxy σ^2, then*

$$\forall Q, W_1(P,Q)^2 \leq 2\sigma^2 KL(P||Q).$$

Theorem 2 (Gaussian Concentration Theorem) *Suppose $Z \sim N(0, I_{k \times k})$ and $g : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz, then $g(Z)$ is sub-Gaussian with variance proxy L^2.*

To prove **Claim 4**, we let $x \sim P = G_\theta(Z)$. Since P is K-Lipschitz, for any 1-Lipschitz function f, $f \circ G_\theta(Z)$ is K-Lipschitz. By Gaussian Concentration Theorem, $f(x) = f(G_\theta(Z))$ is sub-Gaussian with variance proxy K^2. By Bobkov-Gotze Theorem,

$$\forall Q, W_1(P,Q)^2 \leq 2K^2 \cdot KL(P||Q) \Rightarrow W_1(P,Q)^2 \lesssim K^2 \cdot KL(P||Q).$$

5. The proof of **Claim 5** is immediate from **Claims 1, 2, 3**.

3 General Overview of Online Learning and Bandit Problems

Online learning problems have several distinctive features:

1. The data is *sequential*. At each time step, the algorithm receives some data and makes a prediction.
2. The data may be *adversarially* given. We cannot assume that each sample is drawn independently from some distribution.
3. The feedback is *limited*. In particular, in bandit problems, the algorithm can only know whether its prediction is right or wrong, but is not given any further information.

The high-level framework for bandit problems can be seen as a two-player game. In each round, the learner, or player, makes a prediction; then the environment, or adversary, provides some feedback.

One example of an online learning problem is spam detection. It can be formulated as an online classification (or regression) task:

Algorithm 1 Spam Detection

```
1: procedure SPAM DETECTION
2:     for t ← 1 to T do
3:         Learner receives some email x_t ∈ X.
4:         Learner predicts ŷ_t ∈ Y (whether it is spam). (ŷ_t = A(x_1, ..., x_t, y_1, ..., y_{t-1}).)
5:         Learner sees true label y_t ∈ Y (maybe adversarially chosen).
6:         Learner suffers loss l(y_t, ŷ_t).
```

How should we evaluate an algorithm for an online learning problem? We can sum up the losses at each time step: $\sum_{t=1}^{T} l(y_t, \hat{y}_t)$, but the issue is that the environment can simply return $y_t \neq \hat{y}_t$. The algorithm would get loss $\Theta(T)$ and there is nothing better it can do. Instead, we compare how well the algorithm does with a baseline: Let H be a set of models $X \to Y$. We define

$$\text{Regret} = \sum_{t=1}^{T} l(y_t, \hat{y}_t) - \min_{h \in H} \sum_{t=1}^{T} l(y_t, h(x_t)).$$

The second term is the best loss in hindsight, i.e., the minimum loss of a single hypothesis in H, which is chosen after seeing the entire history of samples. This is a more reasonable measure of the performance of the algorithm.