CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Weiyun Ma

Lecture #4
October 3, 2018

1 Review and Overview

Last time, we introduced the notion of uniform convergence and showed that it holds for finite hypothesis classes. Specifically, we proved that if $\mathcal{H}$ is a finite hypothesis class with $0 \leq l((x,y),h) \leq 1$, then with probability at least $1-\delta$, we have

$$\forall h \in \mathcal{H}, |\hat{L}(h) - L(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2n}}. \tag{1}$$

In other words, with probability at least $1-\delta$, for *every* hypothesis in $\mathcal{H}$, the difference between its training loss and expected loss is upper bounded by a function that depends on δ, the size of the hypothesis space, and the number of training examples.

Unfortunately, the result above does not generalize to the case when the hypothesis class $\mathcal{H}$ is parameterized by a continuous space, since we can no longer apply union bound over the infinitely many hypotheses in $\mathcal{H}$. However, as we will see in the first part of this lecture, if the continuous parameter space that defines $\mathcal{H}$ is bounded, then we can still achieve a result similar to (1) using a technique called *discretization.*

In the second part of this lecture, we will shift our focus to proving Hoeffding's inequality, which we used last time. Along the way, we will also introduce a few other useful concentration inequalities.

2 Uniform Convergence for Parameterized Hypothesis Classes

In this section, we will prove a uniform convergence result similar to (1) for parameterized hypothesis classes whose parameter spaces are bounded. The high level idea is that we can select a finite subset of representatives from the entire parameter space, and obtain a uniform convergence bound on those representatives by applying Theorem 2 from last lecture. We then show that every other point in the parameter space is close enough to at least one representative, so that uniform convergence can be extended to all points in the parameter space. We call this technique *discretization* of the parameter space.

2.1 Setup

Before we formally demonstrate what discretization is, let us first set up a few notations.

Let $\mathcal{H}$ be a family of hypothesis functions indexed by parameter θ from some subset S of $\mathbb{R}^p$. Formally, we write

$$\mathcal{H} = \{h_\theta : \theta \in S \subseteq \mathbb{R}^p\}.$$

In this section, we focus on the case

$$S = \{\theta \in \mathbb{R}^p : \|\theta\|_2 \le B\},$$

where $B > 0$ is fixed. In other words, we constrain our parameter space S to contain only p-dimensional vectors whose ℓ_2 norms are no greater than some positive threshold B.

Recall that in the parameterized setting, we can write the loss function as $l((x, y), h_\theta) = l((x, y), \theta)$.

The expected loss is defined as $L(\theta) = \mathbb{E}_{(x,y)\sim P}[l((x, y), \theta)]$, where P is a data distribution that is not necessarily the same as the ground truth distribution.

The training loss (or empirical risk) is defined as $\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}[l((x^{(i)}, y^{(i)}), \theta)]$, where $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ are n training examples drawn i.i.d. from P.

2.2 The Uniform Convergence Theorem

We now state the uniform convergence theorem for our parameterized hypothesis class $\mathcal{H}$.

Theorem 1. *Suppose $\ell((x, y), \theta) \in [-M, M]$, and $L(\theta), \hat{L}(\theta)$ are L-Lipschitz with respect to the ℓ_2-norm. Then with probability $\ge 1 - O(p^{-10})$, we have*

$$\forall\theta, |\hat{L}(\theta) - L(\theta)| \lesssim M\sqrt{\frac{p\log(LBn)}{n}}. \tag{2}$$

Here, we give the definition of *L-Lipschitz functions*:

Definition 2. *Let $L \ge 0$ and $\|\cdot\|$ be a norm on the domain D. A function $L : D \to \mathbb{R}$ is said to be L-*Lipschitz *with respect to $\|\cdot\|$ if for all $\theta, \theta' \in D$, we have*

$$|L(\theta) - L(\theta')| \le L\|\theta - \theta'\|.$$

We make a few remarks on Theorem 1:

Remark 3. *The result is not tight. In particular, with a more careful derivation, one can improve the failure probability $O(p^{-10})$ to $O(e^{-p})$.*

Remark 4. *By Theorem 1 and a result (Theorem 2) from last lecture, we have*

$$L(\hat{\theta}) - L(\theta^*) \lesssim M\sqrt{\frac{p\log(LBn)}{n}} \tag{3}$$

with probability $\ge 1 - O(p^{-10})$, where $\theta^* = \operatorname{argmin}_{\theta \in S} L(\theta)$ is the minimizer of the expected loss over the parameter space.

On the other hand, we showed last week that in the asymptotic regime, we have

$$L(\hat{\theta}) - L(\theta^*) \approx \frac{p}{2n} + o\left(\frac{1}{n}\right) \quad \text{as } n \to \infty, \tag{4}$$

where θ^* is the ground truth parameter.

Comparing the generalization bound (3) with (4), we see that although the rate in (3) has a worse dependency on n, (3) does not contrain how large n has to be. Moreover, (4) requires well-specification, i.e. it assumes that the data is from the ground truth distribution, while (3) and Theorem 1 do not need any assumption on the data distribution.

2.2.1 Discretization of the Parameter Space by ε-Covers

To prove Theorem 1, we need a measure to discretize the parameter space. We start by defining an *ε-cover*:

Definition 5. *Let S be a set with metric ρ, and let $\varepsilon > 0$. An* ε-cover *of S with respect to ρ is a subset $C \subseteq S$ such that $\forall x \in S$, $\exists x' \in C$ such that $\rho(x, x') \leq \varepsilon$, or equivalently,*

$$S \subseteq \bigcup_{x \in C} \mathrm{Ball}(x, \varepsilon, \rho).$$

The following lemma tells us that our parameter space $S = \{\theta \in \mathbb{R}^p : \|\theta\|_2 \leq B\}$ has an ε-cover with not too many elements:

Lemma 6. *Let $B, \varepsilon > 0$ and $S = \{x \in \mathbb{R}^p : \|x\|_2 \leq B\}$. Then there exists an ε-cover of S with respect to the ℓ_2-norm with at most $\left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p$ elements.*

Proof. We set

$$C = \{x \in S : x_i = k_i \tfrac{\varepsilon}{\sqrt{p}}, k_i \in \mathbb{Z}, |k_i| \leq \tfrac{B\sqrt{p}}{\varepsilon}\},$$

i.e. C is the set of grid points of the grid in $\mathbb{R}^p$ of width $\frac{\varepsilon}{\sqrt{p}}$ that are contained in S. See Figure 1 for an illustration.

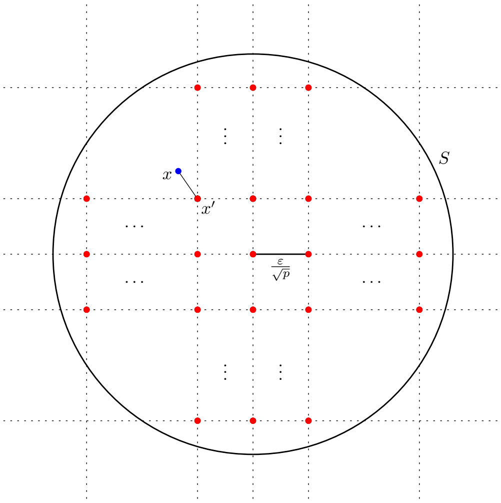

Figure 1: Our chosen ε-cover (shown in red) of S. For $x \in S$, we choose the grid point x' such that $\|x - x'\|_2 \leq \varepsilon$.

Notice that $\forall x \in S$, there exists a grid point $x' \in C$ such that $|x_i - x'_i| \leq \frac{\varepsilon}{\sqrt{p}}$ for each i. Therefore,

$$\|x - x'\|_2 = \sqrt{\sum_{i=1}^{p} |x_i - x'_i|^2} \leq \sqrt{p \cdot \frac{\varepsilon^2}{p}} = \varepsilon.$$

This verifies that C is an ε-cover of S with respect to the ℓ_2-norm.

We now bound the size of C. Since each k_i in the definition of C has at most $2\frac{B\sqrt{p}}{\varepsilon}+1$ choices, we have

$$|C| \leq \left(\frac{2B\sqrt{p}}{\varepsilon}+1\right)^p \leq \left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p.^{[1]}$$

□

Remark 7. *In fact, we can prove a stronger version of Lemma 6 that there exists an ε-cover of S with at most $\left(\frac{3B}{\varepsilon}\right)^p$ elements. We will leave this as a homework exercise.*

2.2.2 Proof of Theroem 1

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let $\delta, \varepsilon > 0$ be parameters to be specified later. Let C be the ε-cover of our parameter space S with respect to the ℓ_2-norm constructed in Lemma 6. By part (2) of Theorem 2 from last time, we have

$$\forall \theta \in C, |\hat{L}(\theta) - L(\theta)| \leq \delta \tag{5}$$

with probability $\geq 1 - 2|C|\exp(-\frac{n\delta^2}{2M^2})$. [2] We refer to this event as E.

Now for any $\theta \in S$, we can pick $\theta_0 \in C$ such that $\|\theta - \theta_0\|_2 \leq \varepsilon$. Using the L-Lipschitz assumption on the functions L and $\hat{L}$, we have

$$|L(\theta) - L(\theta_0)| \leq L\|\theta - \theta_0\|_2 \leq L\varepsilon,$$

$$|\hat{L}(\theta) - \hat{L}(\theta_0)| \leq L\|\theta - \theta_0\|_2 \leq L\varepsilon.$$

Therefore, by (5) conditioned on E, we have

$$|\hat{L}(\theta) - L(\theta)| \leq |\hat{L}(\theta) - \hat{L}(\theta_0)| + |\hat{L}(\theta_0) - L(\theta_0)| + |L(\theta_0) - L(\theta)| \leq 2L\varepsilon + \delta. \tag{6}$$

It remains to choose suitable parameters δ and ε to get the desired bound in Theorem 1 while making the failure probability small. We set

$$\delta = 10\sqrt{\frac{M^2 p \log(LBn)}{n}}, \text{ and } \varepsilon = \frac{\delta}{2L}.$$

In this way, we see from (6) that

$$|\hat{L}(\theta) - L(\theta)| \leq 2L\varepsilon + \delta = 2\delta \lesssim M\sqrt{\frac{p\log(LBn)}{n}}.$$

[1] We assume that $\varepsilon \leq B\sqrt{p}$. Otherwise, S is contained in the ball centered at the origin with radius ε.

[2] Here, we have an additional multiplicative factor of $\frac{1}{4M^2}$ in the exponent, as compared to part (2) of Theorem 2 from last time. The reason is that in our Theorem 1, the loss function $l((x,y),\theta)$ takes value in the interval $[-M,M]$ (whose width is $2M$), while in Theorem 2 of last time, the range is $[0,1]$. Thus the factor of $\frac{1}{4M^2}$ pops out as we apply Hoeffding's inequality.

Finally, we estimate the failure probability for this choice of δ and ε. We have

$$\begin{aligned}
|C|\exp\left(-\frac{n\delta^2}{M^2}\right) &\le \left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p \exp\left(-\frac{n(100M^2 p\log(LBn))}{nM^2}\right) \\
&= \exp\left(p\log\left(\frac{6LB\sqrt{p}}{\delta}\right) - 100p\log(LBn)\right) \\
&= \exp\left(p\log(6LB\sqrt{n}) - p\log(10M\sqrt{\log LBn}) - 100p\log(LBn)\right) \\
&= \exp\left(p\left(-\frac{199}{2}\log n - \frac{1}{2}\log\log n + \log 6 - \log(10M) - 99\log(LB) - \frac{1}{2}\log\log(LB)\right)\right) \\
&\le \exp\left(p\left(\log 6 - \log(10M) - 99\log(LB) - \frac{1}{2}\log\log(LB)\right)\right) \\
&\le O(e^{-p}),
\end{aligned}$$

as long as M, L, B are greater than some small constant. Hence we have actually proven the stronger bound on the failure probability in Remark 3. □

3 Concentration Inequalities

In this section, we will look at a few tools for bounding the tail distribution of a random variable. These bounds can be very useful in estimating the failure probability of a certain event of interest. For example, as we saw from the last lecture, we can use Hoeffding's inequality to bound the failure probability of uniform convergence for finite hypothesis classes. Let us recall the statement of Hoeffding's inequality:

Proposition 8 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables such that $a_i \le X_i \le b_i$ almost surely for each $i \in [n]$. Let*

$$\hat{\mu} = \sum_{i=1}^n X_i, \mu = \mathbb{E}\left[\sum_{i=1}^n X_i\right].$$

Then for any $t > 0$, we have

$$\Pr[|\hat{\mu} - \mu| \ge t] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

In order to prove Hoeffding's inequality, we will first look at two other basic concentration inequalities: Markov's inequaility and Chebyshev's inequality. We will then apply these simpler concentration inequalities to the moment generating function of the random variable $\hat{\mu}$ to obtain a heuristic argument for Hoeffding's inequality. A good reference for our discussion here is Chapter 2 of [1].

3.1 Markov's Inequality and Chebyshev's Inequality

Proposition 9 (Markov's inequality)**.** *Let X be a non-negative random variable. Then for any $t > 0$, we have*

$$\Pr[X \ge t] \le \frac{\mathbb{E}[X]}{t}.$$

Proposition 10 (Chebyshev's inequality)**.** *Let Z be a random variable. Then for any $t > 0$, we have*

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{Var(Z)}{t^2}.$$

Here is a proof of Chebyshev's inequality using Markov's inequality:

Proof. Note that

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] = \Pr[|Z - \mathbb{E}[Z]|^2 \geq t^2].$$

Now applying Markov's inequality to the non-negative random variable $(Z - \mathbb{E}[Z])^2$, we get

$$\Pr[|Z - \mathbb{E}[Z]|^2 \geq t^2] \leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{t^2} = \frac{\mathrm{Var}(Z)}{t^2}.$$

□

In fact, for any positive integer k, we have

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] = \Pr[|Z - \mathbb{E}[Z]|^k \geq t^k].$$

The same procedure can be applied to non-decreasing functions other than polynomials. For example, for $\lambda \geq 0$, we have

$$\Pr[Z - \mathbb{E}[Z] \geq t] = \Pr[e^{\lambda(Z - \mathbb{E}[Z])} \geq e^{\lambda t}] \leq \frac{\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}]}{e^{\lambda t}}, \tag{7}$$

where the last inequality again follows from Markov's inequality.

One may attempt to prove Hoeffding's inequality using Chebyshev's inequality as follows. Assume that $a_i = 0$, $b_i = 1$ for each i. Let $Z = \hat{\mu}$. Since $X_1, \ldots, X_n$ are independent, we have

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \leq n.$$

Now Chebyshev's inequality tells us

$$\Pr[|\hat{\mu} - \mu| \geq t] \leq \frac{\mathrm{Var}(Z)}{t^2} \leq \frac{n}{t^2}.$$

If we set $t = c\sqrt{n}$, then the above becomes

$$\Pr[|\hat{\mu} - \mu| \geq c\sqrt{n}] \leq \frac{1}{c^2}.$$

However, Hoeffding's inequality in this case reads as

$$\Pr[|\hat{\mu} - \mu| \geq c\sqrt{n}] \leq 2e^{-2c^2}.$$

So this attempt does not give us what we want.

3.2 Moment Generating Function

Based on (7), we may also try to prove Hoeffding's inequality using the moment generating function of our random variable $Z = \hat{\mu} = \sum_{i=1}^{n} X_i$. Let us first recall the definition:

Definition 11. *The* moment generating function *of a random variable X, denoted M_X, is given by*

$$M_X(\lambda) = \mathbb{E}[e^{\lambda X}].$$

Using Taylor expansion, the above definition can be rewritten as

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[1 + \lambda X + \frac{(\lambda X)^2}{2!} + \cdots\right] = 1 + \lambda \mathbb{E}[X] + \frac{\lambda^2}{2}\mathbb{E}[X^2] + \cdots .$$

This applied to the random variable $X - \mathbb{E}[X]$ gives

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = 1 + \lambda\mathbb{E}[X - \mathbb{E}[X]] + \frac{\lambda^2}{2}\mathbb{E}[(X-\mathbb{E}[X])^2] + \cdots = 1 + \lambda^2\mathrm{Var}(X) + \cdots . \quad (8)$$

Back to our case $Z = \sum_{i=1}^{n} X_i$. Since $X_1, \ldots, X_n$ are independent, we have

$$\mathbb{E}[e^{\lambda(\mathbb{Z}-\mathbb{E}[Z])}] = \mathbb{E}[e^{\lambda(X_1-\mathbb{E}[X_1])} \cdots e^{\lambda(X_n-\mathbb{E}[X_n])}] = \mathbb{E}[e^{\lambda(X_1-\mathbb{E}[X_1])}] \cdots \mathbb{E}[e^{\lambda(X_n-\mathbb{E}[X_n])}].$$

Hence by (7), we have the following bound:

$$\Pr[Z - \mathbb{E}[Z] \geq t] \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda(X_1-\mathbb{E}[X_1])}] \cdots \mathbb{E}[e^{\lambda(X_n-\mathbb{E}[X_n])}]}{e^{\lambda t}},$$

i.e.,

$$\log \Pr[Z - \mathbb{E}[Z] \geq t] \leq \inf_{\lambda \geq 0} \sum_{i=1}^{n} \log \mathbb{E}[e^{\lambda(X_1-\mathbb{E}[X_1])}] - \lambda t. \quad (9)$$

Now in the case $X_i \in [0,1]$ for each $i \in [n]$, we give a heuristic argument for Hoeffding's inequality. In this case, $\mathrm{Var}(X_i) \leq 1$ for each $i \in [n]$, so from (8) we get

$$\mathbb{E}[e^{\lambda(X_i-\mathbb{E}[X_i])}] = 1 + \frac{\lambda^2}{2} + \text{higher order terms},$$

for each $i \in [n]$. If we can get rid of these higher order terms in some way, we get

$$\mathbb{E}[e^{\lambda(X_i-\mathbb{E}[X_i])}] = 1 + \frac{\lambda^2}{2}$$

for each $i \in [n]$. Thus the right hand side of (9) now becomes

$$\inf_{\lambda \geq 0} \left(n \log\left(1 + \frac{\lambda^2}{2}\right) - \lambda t\right).$$

When λ is small, $\log\left(1 + \frac{\lambda^2}{2}\right)$ can be approximated by $\frac{\lambda^2}{2}$. So the above becomes

$$\inf_{\lambda \geq 0} \frac{n\lambda^2}{2} - \lambda t = -\frac{t^2}{2n}.$$

The "result" above is not quite the same as Hoeffding's inequality, but has the same magnitude. Hopefully some modification of this argument could give us an actual proof. We will continue the discussion next time.

References

[1] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.