CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma
Scribe: Yizhou Qian

Lecture # 4
Jan 25th, 2021

4.1 Review and overview

In the previous lecture, we discussed properties of *sub-Gaussian* variables and the notion of *concentration inequalities*. In particular, for a *sub-Gaussian* random variable X with proxy variance σ, we have the concentration inequality

$$\Pr[|X - \mu| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{4.1}$$

This gives us a bound for how quickly the random variable X concentrates around its mean.

In this lecture, we will introduce the notion of uniform convergence of the empirical loss to the population loss. Namely, our goal is to obtain bounds of the following form:

$$\Pr\left[\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| \leq \epsilon\right] \geq 1 - \delta. \tag{4.2}$$

In other words, the probability that the difference between our empirical loss and population loss is larger than ϵ is at most δ. We will consider two cases: (i) when the hypothesis class $\mathcal{H}$ is finite, and (ii) when $\mathcal{H}$ is infinite.

4.2 Finite hypothesis class

In this section, assume that $\mathcal{H}$ is finite. The following theorem gives a bound for the excess risk $L(\hat{h}) - L(h^*)$, where $\hat{h}$ and h^* are the minimizers of the empirical loss and population loss respectively.

Theorem 4.1. *Suppose that our hypothesis class $\mathcal{H}$ is finite and that our loss function ℓ is bounded in $[0, 1]$, i.e. $0 \leq \ell((x, y), h) \leq 1$. Then $\forall\delta$ s.t. $0 < \delta < \frac{1}{2}$, with probability at least $1 - \delta$, we have*

$$|L(h) - \hat{L}(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}} \qquad \forall h \in \mathcal{H}. \tag{4.3}$$

As a corollary, we also have

$$L(\hat{h}) - L(h^*) \leq \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}}. \tag{4.4}$$

Proof. We will prove this in two steps:

1. Use concentration inequalities to prove the bound for a fixed $h \in \mathcal{H}$, then
2. Use a union bound across the h's. (Recall that if $E_1, \ldots, E_k$ are a finite set of events, then the union bound states that $\Pr(E_1 \cup \cdots \cup E_k) \leq \sum_{i=1}^{k} \Pr(E_i)$.)

Fix some $\epsilon > 0$. By applying Hoeffding's inequality on the $\ell((x^{(i)}, y^{(i)}), h)$, we know that

$$\Pr\left(|\hat{L}(h) - L(h)| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{4.5}$$

$$= 2\exp\left(-\frac{2n^2\epsilon^2}{n}\right) \tag{4.6}$$

$$= 2\exp(-2n\epsilon^2), \tag{4.7}$$

since we can set $a_i = 0, b_i = 1$. The bound above holds for a single fixed h. To prove a similar inequality that holds for all $h \in \mathcal{H}$, we apply the union bound with $E_h = \{|\hat{L}(h) - L(h)| \geq \epsilon\}$:

$$\Pr\left(\exists h \text{ s.t. } |\hat{L}(h) - L(h)| \geq \epsilon\right) \leq \sum_{h\in\mathcal{H}} \Pr\left(|\hat{L}(h) - L(h)| \geq \epsilon\right) \tag{4.8}$$

$$\leq \sum_{h\in\mathcal{H}} 2\exp(-2n\epsilon^2) \tag{4.9}$$

$$= 2|\mathcal{H}|\exp(-2n\epsilon^2). \tag{4.10}$$

If we take δ such that $2|\mathcal{H}|\exp(-2n\epsilon^2) = \delta$, then it follows that

$$\epsilon = \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}}, \tag{4.11}$$

which proves (4.3). (4.4) follows by the inequality we stated in Lecture 2 (see Section 2.4.1) and taking $\epsilon = \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}}$:

$$\Pr\left(|L(\hat{h}) - L(h^*)| \geq \epsilon\right) \leq \Pr\left(2\sup_{h\in\mathcal{H}}|L(\hat{h}) - L(h^*)| \geq \epsilon\right) \tag{4.12}$$

$$\leq 2|\mathcal{H}|\exp\left(-\frac{n\epsilon^2}{2}\right). \tag{4.13}$$

□

4.2.1 Comparing Theorem 4.1 with standard concentration inequalities

With standard concentration inequalities, we have the following bound that depends on empirical risk:

$$\forall h \in \mathcal{H}, \quad w.h.p. \quad |\hat{L}(h) - L(h)| \leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{4.14}$$

The bound here depends on each h. In contrast, the uniform convergence bound we obtain from (4.11) is uniform over all $h \in \mathcal{H}$:

$$w.h.p., \quad \forall h \in \mathcal{H}, \quad |\hat{L}(h) - L(h)| \leq \tilde{O}\left(\frac{\ln|\mathcal{H}|}{\sqrt{n}}\right), \tag{4.15}$$

if we omit the $\ln(1/\delta)$ factor (we can do this since $\ln(1/\delta)$ is small in general and we take $\delta = \frac{1}{poly(n)}$). Hence, the extra $\ln|\mathcal{H}|$ term that depends on the size of our finite hypothesis family $\mathcal{H}$ can be viewed as a trade-off in order to make the bound uniform.

Remark 4.2. There is no standard definition for the term *with high probability* (*w.h.p*). For this class, the term is equivalent to the condition that the probability is higher than $1 - n^{-c}$ for some constant c.

4.2.2 Comparing Theorem 4.1 with asymptotic bounds

We can also compare the bound in Theorem 4.1 with our original asymptotic bound, namely,

$$L(\hat{h}) - L(h^*) \leq \frac{c}{n} + o\left(n^{-1}\right). \tag{4.16}$$

The $o(n^{-1})$ term can vary significantly depending on the problem. For instance, both n^{-2} and $p^{100}n^{-2}$ are $o(n^{-1})$ but the second one converges much more slowly. With the new bound, there are no longer any constants hidden in an $o(n^{-1})$ term (in fact that term is no longer there). However, we now have a slower convergence rate of $O(n^{-1/2})$.

Remark 4.3. $O(n^{-1/2})$ convergence is sometimes known as the *slow rate* while $O(n^{-1})$ convergence is known as the *fast rate*. We were only able to get the slow rate from uniform convergence: we needed asymptotics to get the fast rate. (It is possible to get the fast rate from uniform convergence under certain conditions, e.g. when the population risk on the true h^* is very low.)

4.3 Infinite hypothesis class

Unfortunately, we cannot generalize the results from the previous section directly to the case where the hypothesis class $\mathcal{H}$ is infinite, since we cannot apply the union bound to an infinite number of hypothesis functions $h \in \mathcal{H}$. However, if we consider a *bounded* and *continuous* parameterized space of $\mathcal{H}$, then we can obtain a similar uniform bound by applying a technique called *brute-force discretization*.

For this section, assume that our infinite hypothesis class $\mathcal{H}$ can be parameterized by $\theta \in \mathbb{R}^p$ with $\|\theta\|_2 \leq B$ for some fixed $B > 0$. That is, we have

$$\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, \|\theta\|_2 \leq B\}. \tag{4.17}$$

The intuition behind brute-force discretization is as follows: Let $E_\theta = \{|\widehat{L}(\theta) - L(\theta)| \geq \epsilon\}$ be the "bad" event. We want the bound the probability of any one of these bad events happening (i.e. $\bigcup_\theta E_\theta$). The union bound does not work as we end up with an infinite sum. However, the union bound is very loose: these events can overlap with each other significantly. Instead, we can try to find "prototypical" bad events $E_{\theta_1}, \ldots, E_{\theta_N}$ that are somewhat disjoint so that $\bigcup_\theta E_\theta \approx \bigcup_{i=1}^N E_{\theta_i}$. We can then use the union bound on $\bigcup_{i=1}^N E_{\theta_i}$ to get a non-vacuous upper bound.

We make these ideas precise in the following section.

4.3.1 Discretization of the parameter space by ϵ-covers

We start by defining the notion of an *ϵ-cover* (also *ϵ-net*):

Definition 4.4 (ϵ-cover)**.** Let $\epsilon > 0$. An *ϵ-cover* of a set S with respect to distance metric ρ is a subset $C \subseteq S$ such that $\forall x \in S$, $\exists x' \in C$ such that $\rho(x, x') \leq \epsilon$, or equivalently,

$$S \subseteq \bigcup_{x \in C} \text{Ball}(x, \epsilon, \rho), \quad \text{where} \tag{4.18}$$

$$\text{Ball}(x, \epsilon, \rho) \triangleq \{x' : \rho(x, x') \leq \epsilon\}. \tag{4.19}$$

(We note that in some definitions it is possible for points in C to lie outside of S; we do not worry about this technicality the class.) 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 4.5 (ϵ-cover of ℓ_2 ball)**.** *Let $B, \epsilon > 0$ with $\epsilon \leq B\sqrt{p}$, and let $S = \{x \in \mathbb{R}^p : \|x\|_2 \leq B\}$. Then there exists an ϵ-cover of S with respect to the ℓ_2-norm with at most $\left(\frac{3B\sqrt{p}}{\epsilon}\right)^p$ elements.*

Proof. Set

$$C = \left\{ x \in S : x_i = k_i \frac{\epsilon}{\sqrt{p}}, k_i \in \mathbb{Z}, |k_i| \leq \frac{B\sqrt{p}}{\epsilon} \right\}, \tag{4.20}$$

i.e. C is the set of grid points in $\mathbb{R}^p$ of width $\frac{\epsilon}{\sqrt{p}}$ that are contained in S. See Figure 4.1 for an illustration.

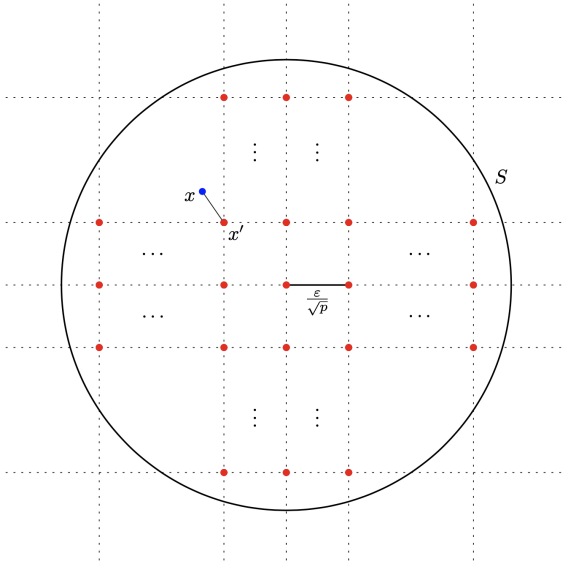

Figure 4.1: Our chosen ϵ-cover (shown in red) of S. For $x \in S$, we choose the grid point x' such that $\|x - x'\|_2 \leq \epsilon$.

We claim that C is an ϵ-cover of S with respect to the ℓ_2-norm: $\forall x \in S$, there exists a grid point $x' \in C$ such that $|x_i - x_i'| \leq \frac{\epsilon}{\sqrt{p}}$ for each i. Therefore,

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

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

$$|C| \leq \left(\frac{2B\sqrt{p}}{\epsilon} + 1\right)^p \leq \left(\frac{3B\sqrt{p}}{\epsilon}\right)^p. \tag{4.21}$$

□

Remark 4.6. If $\epsilon > B\sqrt{p}$, then S is contained in the ball centered at the origin with radius ϵ and the ϵ-cover has size 1.

Remark 4.7. We can actually prove a stronger version of Lemma 4.5: there exists an ϵ-cover of S with at most $\left(\frac{3B}{\epsilon}\right)^p$ elements. We will be using this version of the lemma in the proof below. (We will leave the proof of this stronger version as a homework exercise.)

4.3.2 Uniform convergence bound for infinite $\mathcal{H}$

Definition 4.8 (κ-Lipschitz functions)**.** Let $\kappa \geq 0$ and $\|\cdot\|$ be a norm on the domain D. A function $L : D \to \mathbb{R}$ is said to be *κ-Lipschitz* with respect to $\|\cdot\|$ if for all $\theta, \theta' \in D$, we have

$$|L(\theta) - L(\theta')| \leq \kappa\|\theta - \theta'\|.$$

Assume that our infinite hypothesis class $\mathcal{H}$ can be parameterized by $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, \|\theta\|_2 \leq B$. We have the following uniform convergence theorem for our infinite hypothesis class $\mathcal{H}$:

Theorem 4.9. *Suppose $\ell((x, y), \theta) \in [0, 1]$, and $\ell((x, y), \theta)$ is κ-Lipschitz in θ with respect to the ℓ_2-norm for all (x, y). Then with probability $\geq 1 - O(\exp(-\Omega(p)))$, we have*

$$\forall\theta, \quad |\hat{L}(\theta) - L(\theta)| \leq O\left(\sqrt{\frac{p \max(\ln(\kappa Bn), 1)}{n}}\right). \tag{4.22}$$

Proof of Theorem 4.9. Fix parameters $\delta, \epsilon > 0$ (we will specify their values later). Let C be the ϵ-cover of our parameter space S with respect to the ℓ_2-norm constructed in Lemma 4.5. Define event $E = \left\{\forall\theta \in C,\ |\hat{L}(\theta) - L(\theta)| \leq \delta\right\}$. By Theorem 4.1, we have $\Pr(E) \geq 1 - 2|C|\exp(-2n\delta^2)$.

Now for any $\theta \in S$, we can pick some $\theta_0 \in C$ such that $\|\theta - \theta_0\|_2 \leq \epsilon$. Since L and $\widehat{L}$ are κ-Lipschitz functions (this follows from the Lipschitzness of ℓ), we have

$$|L(\theta) - L(\theta_0)| \leq \kappa\|\theta - \theta_0\|_2 \leq \kappa\epsilon, \text{ and} \tag{4.23}$$

$$|\hat{L}(\theta) - \hat{L}(\theta_0)| \leq \kappa\|\theta - \theta_0\|_2 \leq \kappa\epsilon. \tag{4.24}$$

Therefore, conditional on E, we have

$$|\hat{L}(\theta) - L(\theta)| \leq |\hat{L}(\theta) - \hat{L}(\theta_0)| + |\hat{L}(\theta_0) - L(\theta_0)| + |L(\theta_0) - L(\theta)| \leq 2\kappa\epsilon + \delta. \tag{4.25}$$

It remains to choose suitable parameters δ and ϵ to get the desired bound in Theorem 4.9 while making the failure probability small. First, set $\epsilon = \delta/(2\kappa)$ so that conditional on E,

$$|\hat{L}(\theta) - L(\theta)| \leq 2\delta. \tag{4.26}$$

If we set $\delta = \sqrt{\frac{c_0 p \max(1, \ln(\kappa Bn))}{n}}$ with $c_0 = 36$ (see Remark 4.10 for some intuition), then by

Remark 4.7,

$$\ln|C| - 2n\delta^2 \le p\ln\left(\frac{6B\kappa}{\delta}\right) - 2n\delta^2 \tag{4.27}$$

$$\le p\ln\left(\frac{6B\kappa\sqrt{n}}{\sqrt{c_0 p\max(1,\ln(\kappa Bn))}}\right) - 2n\frac{c_0 p}{n}\ln(\kappa Bn) \qquad \text{(dfn of } \delta\text{)} \tag{4.28}$$

$$\le p\ln\left(\frac{B\kappa\sqrt{n}}{\sqrt{p}}\right) - 72p\ln(\kappa Bn) \qquad (\max(1,\ln(\kappa Bn)) \ge 1, c_0 = 36) \tag{4.29}$$

$$\le p\ln(B\kappa n) - 72p\ln(B\kappa n) \qquad (\sqrt{n/p} \le n) \tag{4.30}$$

$$\le -p, \tag{4.31}$$

since $\ln(B\kappa n) \ge 1$ for large enough n. Therefore, with probability greater than $1-2|C|\exp(-2n\delta^2) = 1 - 2\exp(\ln|C| - 2n\delta^2) \ge 1 - O(e^{-p})$, we have

$$|\hat{L}(\theta) - L(\theta)| \le 2\delta = O\left(\sqrt{\frac{p}{n}\max(1,\ln(\kappa Bn))}\right). \tag{4.32}$$

□

Remark 4.10. Here is the intuition for the choice of δ: The event E happens with probability $1 - 2|C|\exp(-2n\delta^2) = 1-2\exp(\ln|C|-2n\delta^2)$. From Remark 4.7, we know that $\ln|C| \le p\ln(3B/(\delta/2))$. If we ignore the log term and assume $\ln|c| \le p$, then this would give us the high probability bound we want:

$$2|C|\exp(-2n\delta^2) = 2\exp(\ln|C| - 2n\delta^2) \le 2\exp(p-2p) = 2\exp(-p). \tag{4.33}$$

(At the same time, we see from (4.26) that this choice of δ gives $|\hat{L}(\theta) - L(\theta)| \le 2\sqrt{\frac{p}{n}}$, which is roughly the bound we want.)

Since we cannot drop the log term in the inequality, we need to make δ a little big bigger. δ in the proof was chosen with this intuition in mind to make the subsequent chain of logic work.

Remark 4.11. We bounded the generalization error $|\hat{L}(\theta) - L(\theta)|$ by $\delta + 2\epsilon\kappa \le \sqrt{\frac{\ln|C|}{n}} + 2\epsilon\kappa$. The term $2\epsilon\kappa$ represents the error from our brute-force discretization. It is not a problem is because we can always choose ϵ small enough without worrying about the growth of the first term $\sqrt{\frac{\ln|C|}{n}}$. This in turn is because $\ln|C| \approx p\ln\epsilon^{-1}$, which is very insensitive to ϵ, even if we let $\epsilon = \frac{1}{poly(n)}$. We also observe that both $\sqrt{\frac{\ln|C|}{n}}$ and $\sqrt{\frac{p}{n}}$ are bounds that depend on the "size" of our hypothesis class, in terms of either its total size or dimensionality. This possibly explains why one may need more training samples when the hypothesis class is larger.

In the future lectures, we will also discuss more fine-grained measures of the "complexity" of the hypothesis class $\mathcal{H}$.