CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma Lecture #3
Scribe: Mingda Qiao October 1, 2013

1 Review and Overview

In the first half of this course, the central question that we want to answer is: *Why minimizing the training error often leads to a small testing error?* In the last lecture, we proved the asymptotics of the maximum likelihood estimator (MLE). In particular, as the number of training examples, denoted by n, tends to infinity,

$$L(\hat{\theta}) - L(\theta^*) \approx \frac{p}{2n} + o\left(\frac{1}{n}\right).$$

Here L is the expected loss. $\hat{\theta}$ is the minimizer of the training loss, while θ^* is the ground truth parameter. This result partially explains why the MLE, which has the smallest training loss, is also likely to achieve a small testing error when there are enough training examples.

One limitation of the above result is that it requires well-specifiedness, i.e., the data are distributed precisely according to a particular ground truth parameter θ^* in the parameter space. We would like to prove a more general result in the following form without assuming well-specifiedness.[1]

$$L(\hat{\theta}) - L(\theta^*) \le f(p, n), \forall p, n \ge 1.$$

Another limitation of this asymptotic result is that it ignores the dependence of higher order terms on other hyperparameters (in this case, the dimension p). Consider the following two functions, both of which are of order $\frac{p}{2n} + o\left(\frac{1}{n}\right)$:

$$\frac{p}{2n} + \frac{1}{n^2} \quad \text{vs} \quad \frac{p}{2n} + \frac{p^{100}}{n^2}.$$

Arguably, the first one is a better upper bound since the second bound requires $n > p^{50}$ training examples to be below 1.

From now on, we restrict our attention to the non-asymptotic regime, where n is finite. In this lecture, we preview the form of results that we would like to prove in the following few lectures. Then we introduce the *uniform convergence* framework, which we will instantiate in various settings to prove generalization bounds. We end this lecture by proving uniform convergence on finite hypothesis classes.

2 Notations

In the non-asymptotic setting, we only ignore the absolute constants (also known as universal constants). Specifically, whenever notation $O(X)$ appears in a statement, it means that there exists a universal constant c such that the statement holds if we replace "$O(X)$" by $c \cdot X$. Similarly, we write $A \lesssim B$ as a shorthand for $A \le O(B)$.

In the following, we review a few central notations from previous lectures.

[1] In this case, θ^* needs to be redefined as the minimizer of the expected loss.

- Hypothesis space: $\mathcal{H}$ is a family of hypotheses, i.e., prediction functions.
- Loss function: $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$. This is analogous to the notation $\ell((x, y), \theta)$ for loss functions in the last lecture, yet it accommodates the general case where we cannot naturally parametrize $\mathcal{H}$ by a continuum of parameters.
- Expected loss: $L(h) = \mathbb{E}_{(x,y)\sim P}\left[\ell((x, y), h)\right]$, where P is a data distribution over $\mathcal{X}\times\mathcal{Y}$. Moreover, we define $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} L(h)$ as the minimizer of the expected loss.
- Training loss (also known as empirical risk): $\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\left(x^{(i)}, y^{(i)}\right), h\right)$, where $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$ are n training examples drawn i.i.d. from P.
- Empirical risk minimizer (ERM): $\hat{h} \in \operatorname{argmin}_{h\in\mathcal{H}} \hat{L}(h)$.

Remark 1. *For a fixed data distribution P, there is no randomness in h^*, since h^* is just the minimizer of a deterministic function $L(h)$. On the other hand, $\hat{h}$ is indeed a random variable, as the training loss function $\hat{L}$ is defined based on the training examples.*

The dependence between some of the above concepts is depicted in the following diagram:

$$\begin{tikzcd}
\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n} \arrow[r] \arrow[drr] & \hat{h} \arrow[r] \arrow[dr] & L(\hat{h}) \\
 & & \hat{L}(\hat{h})
\end{tikzcd}$$

3 Objective

Our goal is to prove an upper bound on $L(\hat{h}) - L(h^*)$, the difference between the expected loss of the ERM $\hat{h}$ and that of the optimal hypothesis h^*. In particular, the results that we are going to prove in this and the following lectures are of the following form:

$$\Pr\left[L(\hat{h}) - L(h^*) > \epsilon\right] \le \delta.$$

In words, with probability at least $1 - \delta$, it holds that $L(\hat{h}) - L(h^*) \le \epsilon$.[2]

We can interpret $L(\hat{h}) - L(h^*) > \epsilon$ as a "failure event" since it means that the ERM $\hat{h}$ has an excess risk greater than parameter ϵ, which is undesirable. Therefore, we would like the probability of this failure event to be as small as possible. Moreover, the smaller ϵ is, the stricter we are when evaluating the performance of $\hat{h}$. Thus, we would like parameters ϵ and δ to be as small as possible, given a fixed number of training examples.[3]

Recall the definition of the ERM $\hat{h}$. We have

$$\hat{L}(\hat{h}) = \min_{h\in\mathcal{H}} \hat{L}(h) \le \hat{L}(h^*).$$

In order to prove a high-probability bound of form

$$L(\hat{h}) \le L(h^*) + [\text{extra term}],$$

[2] Here and in the following, the probability is always taken over the randomness in training examples unless otherwise specified.

[3] As a general rule of thumb, we can make δ inverse polynomially small without using too many training examples, while the cost of minimizing ϵ is generally higher.

it remains to argue that $\hat{L}(\hat{h}) \approx L(\hat{h})$ and $\hat{L}(h^*) \approx L(h^*)$ (up to a small additive error) with high probability.

Proving $\hat{L}(h^*) \approx L(h^*)$ is relatively simple: $\hat{L}(h^*) = \frac{1}{n}\sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), h^*)$ is the average of n i.i.d. random variables, each with expectation $L(h^*)$, so we can prove a bound of the following form by applying standard concentration inequalities:

$$\Pr\left[\left|\hat{L}(h^*) - L(h^*)\right| \leq \epsilon\right] \geq 1 - \delta.$$

In fact, this argument holds for *any* fixed hypothesis $h \in \mathcal{H}$, as long as h does not depend on the training examples.

The difficulty is in proving $\hat{L}(\hat{h}) \approx L(\hat{h})$ since $\hat{h}$ is indeed a random variable that depends on the training examples (see Remark 1). Instead of a concentration bound for a single fixed hypothesis h, we need a stronger concentration property that holds for *every* hypothesis in $\mathcal{H}$ simultaneously. This is where the notion of uniform convergence comes into play.

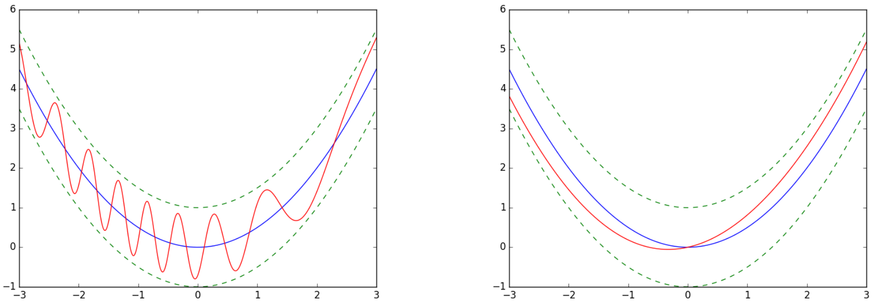
4 Uniform Convergence

Uniform convergence is a property of the hypothesis class $\mathcal{H}$ of the following form:

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

In words, it states that with probability at least $1 - \delta$ over the random draw of training data, the training loss is pointwise close to the expected loss, up to an additive error of at most ϵ.

In general, if we parametrize the hypothesis space by $\mathbb{R}$, we would expect the picture of training loss and expected loss to be as in Figure 1(a) if uniform convergence holds. It turns out that for particular learning tasks, the training loss exhibits a nicer landscape: it is not only pointwise close to the expected loss but also of the same shape, which is informally depicted in Figure 1(b). (See [**?**, **?**] for some of the recent work along this line of research.) Nevertheless, as only uniform convergence is concerned in this lecture, we do not distinguish these two different landscapes.



(a)

(b)

Figure 1: Two different empirical risk landscapes. The blue line and the red line denote the expected and training losses. The dashed green lines denote the expected loss ± ϵ.

4.1 Uniform Convergence Implies Generalization

Before proving uniform convergence for specific hypothesis classes, we first demonstrate how it implies generalization, i.e., an upper bound on $L(\hat{h}) - L(h^*)$. We can write $L(\hat{h}) - L(h^*)$ as

$$\begin{aligned} L(\hat{h}) - L(h^*) &= \left[L(\hat{h}) - \hat{L}(\hat{h})\right] + \left[\hat{L}(\hat{h}) - \hat{L}(h^*)\right] + \left[\hat{L}(h^*) - L(h^*)\right] \\ &\le \left|L(\hat{h}) - \hat{L}(\hat{h})\right| + 0 + \left|\hat{L}(h^*) - L(h^*)\right| \le 2 \sup_{h\in\mathcal{H}} \left|\hat{L}(h) - L(h)\right|. \end{aligned} \tag{2}$$

Here the second step applies the definition of $\hat{h}$, and the third step follows directly from the definition of supremum.

By Equation (2), we have the following implication

$$\forall h \in \mathcal{H}, \left|\hat{L}(h) - L(h)\right| \le \epsilon \implies \sup_{h\in\mathcal{H}} \left|\hat{L}(h) - L(h)\right| \le \epsilon \implies L(\hat{h}) - L(h^*) \le 2\epsilon. \tag{3}$$

Therefore, if we could prove uniform convergence (1) for hypothesis class $\mathcal{H}$, we have

$$\Pr\left[L(\hat{h}) - L(h^*) \le 2\epsilon\right] \ge \Pr\left[\forall h \in \mathcal{H}, \left|\hat{L}(h) - L(h)\right| \le \epsilon\right] \ge 1 - \delta,$$

a generalization bound that we desire.

4.2 Finite Hypothesis Classes

Now we prove that uniform convergence indeed holds for finite hypothesis classes. Recall that n is the number of examples drawn i.i.d. from the data distribution, and the probability is always taken over the randomness in training examples.

Theorem 2. *If $\mathcal{H}$ is finite and $\ell((x, y), h) \in [0, 1]$, we have the following statements:*

(1) For any fixed $h \in \mathcal{H}$ and $\epsilon > 0$,

$$\Pr\left[\left|\hat{L}(h) - L(h)\right| \le \epsilon\right] \ge 1 - 2e^{-2n\epsilon^2}.$$

(2) For any $\epsilon > 0$,

$$\Pr\left[\forall h \in \mathcal{H}, \left|\hat{L}(h) - L(h)\right| \le \epsilon\right] \ge 1 - 2\left|\mathcal{H}\right| e^{-2n\epsilon^2}.$$

(3) With probability at least $1 - \delta$, for any $h \in \mathcal{H}$,

$$\left|\hat{L}(h) - L(h)\right| \le \sqrt{\frac{\ln\left|\mathcal{H}\right| + \ln\frac{2}{\delta}}{2n}}.$$

(4) With probability at least $1 - \delta$,

$$L(\hat{h}) - L(h^*) \lesssim \sqrt{\frac{\ln\left|\mathcal{H}\right| + \ln\frac{1}{\delta}}{n}}.$$

The proof of the theorem relies on the following concentration inequality, which is a quantitative version of the central limit theorem.

Lemma 3 (Hoeffding's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely for each $i \in [n]$. Then, for any $\epsilon > 0$,*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \leq \epsilon\right] \geq 1 - \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Now we are ready to prove Theorem 2. We first prove Statement (1) using Hoeffding's inequality, then we show that each statement directly implies the next.

Proof of Theorem 2. Statement (1) follows from Hoeffding's inequality by taking $X_i = \ell((x^{(i)}, y^{(i)}), h)$, $a_i = 0$ and $b_i = 1$ for each $i \in [n]$ in Lemma 3. Then, Statement (2) follows from a union bound:

$$\begin{aligned}\Pr\left[\exists h \in \mathcal{H}, \left|\hat{L}(h) - L(h)\right| > \epsilon\right] &\leq \sum_{h \in \mathcal{H}} \Pr\left[\left|\hat{L}(h) - L(h)\right| > \epsilon\right] && \text{(union bound)}\\ &\leq \sum_{h \in \mathcal{H}} 2e^{-2n\epsilon^2} = 2\left|\mathcal{H}\right| e^{-2n\epsilon^2}. && \text{(Statement (1))}\end{aligned}$$

Statement (3) follows from plugging $\epsilon = \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2n}}$ into Statement (2). Finally, Statement (3) and the implication in Equation (3) imply Statement (4). □

5 Digression: The PAC Learning Framework

Probably approximately correct (PAC) learning is a theoretical framework of machine learning proposed by Valiant [?]. One of the key definitions in PAC learning is the notion of PAC learning algorithms.

Definition 4 (PAC learning algorithm)**.** *Algorithm $\mathcal{A}$ is a PAC learning algorithm for hypothesis class $\mathcal{H}$, if for any distribution P over $\mathcal{X} \times \mathcal{Y}$, $\epsilon > 0$ and $\delta \in (0, 1)$,*

$$\hat{h} = \mathcal{A}\left(\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right)$$

satisfies

$$\Pr\left[L(\hat{h}) - L(h^*) > \epsilon\right] \leq \delta,$$

and $\mathcal{A}$ runs in polynomial time with respect to $\text{size}(\mathcal{X})$, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

Remark 5. *Informally, $\text{size}(\mathcal{X})$ is the number of bits to describe an element of $\mathcal{X}$. For example, $\text{size}(\mathcal{X}) = \log_2|\mathcal{X}|$ if $\mathcal{X}$ is finite. In the general case where $\mathcal{X}$ is parametrized by real numbers, this definition requires $\mathcal{X}$ to be discretized first.*

Remark 6. *Definition 4 implicitly requires the number of training examples, n, to be polynomial in $\text{size}(\mathcal{X})$, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. Otherwise, $\mathcal{A}$ does not have enough time to read its entire input.*

One limitation of Valiant's framework is that it requires the algorithm to work on *every* data distribution P, which turns out to be overly ambitious and thus unrealistic. In contrast, recent research on learning theory usually requires certain assumptions on the data distribution, e.g., the distribution is Gaussian or realizable (i.e., the hypothesis class contains the function to be learned).

Thus, it is worth thinking about the role of assumptions on learning theory research. Suppose we could prove the following three theorems:

- **Theorem A.** Statement P implies Statement Q.
- **Theorem B.** Under certain assumptions, P implies Q.
- **Theorem C.** Under certain assumptions, P implies a statement stronger than Q.

Theorem B is definitely the weakest among these three, yet Theorems A and C are incomparable. It is often a matter of taste whether one prefers A or C. In deep learning theory, however, Statement Q is often vacuous for most practical uses. In this case, we had better prove a result similar to Theorem C, and then find out how the assumptions can be weakened and eventually removed.