CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma — Lecture #7
Scribe: Brian Zhang — October 15, 2018

1 Review and Overview

We will first give a brief review of what has been covered so far.

In the first few lectures, we stated and proved asymptotics on the maximum likelihood estimator. In particular, in the well-specified case, we showed that

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

where $\hat{h}$ is the empirical MLE, h^* is the ground truth function, p is the number of parameters in the model, n is the number of training examples, and L is the expected negative log likelihood (test error).

We then transitioned to non-asymptotics results, and simultaneously we dropped the assumption of well-specificity. We first observed that

$$L(\hat{h}) - L(h^*) \leq 2 \sup_h \left|\hat{L}(h) - L(h)\right|$$

which allows us to use uniform convergence results to establish bounds on the generalization error. Our attention thus turned to bounding the right hand side of the above equation.

In the case of a finite hypothesis class H, using Hoeffding's inequality and then a union bound, we showed that, with probability at least $1 - \delta$,

$$\sup_h \left|\hat{L}(h) - L(h)\right| \lesssim \sqrt{\frac{\log|H| + \log(2/\delta)}{n}}.$$

In the case of an infinite hypothesis class H parametrized by a bounded parameter $\theta \in \mathbb{R}^p$, we concluded that, with probability $1 - O\big(p^{-10}\big)$,

$$\sup_h \left|\hat{L}(h) - L(h)\right| \lesssim \sqrt{\frac{p \log n}{n}}$$

as long as the loss functions are sufficiently nice (Lipschitz and bounded).

Last week, we defined the notion of Rademacher complexity and used it to simplify our discussion of uniform convergence bounds. In particular, we saw that

$$\sup_h \left|\hat{L}(h) - L(h)\right| \lesssim R_S(F) + \sqrt{\frac{\log(2/\delta)}{n}}$$

where $R_S(F)$ is the sample Rademacher complexity on a sample S of size n, and F is a family of loss functions. We note that $R_S(F)$ can be replaced with the expected Rademacher complexity $R_n(F) = \mathbb{E}\, R_S(F)$, but that it is usually easier to reason about $R_S(F)$ because there is one fewer expectation to worry about. If we use the γ-margin loss, then

$$R_S(F) \leq \frac{R_S(H)}{\gamma}.$$

In this lecture, we will focus on bounding $R_S(H)$, which by Talagrand's Lemma and the above discussion will lead to bounds on $L(\hat{h}) - L(h^*)$. We will start with the case of a linear function with bounded parameter, and move to the more general setting of a neural network with a single hidden layer. We note that focusing on $R_S(H)$ instead of $R_S(F)$ carries another advantage: it allows us to ignore the labels and look at only the inputs, which often simplifies the analysis.

2 Rademacher Complexity of Linear Models

Theorem 1. *Let $H = \left\{x \mapsto w^T x : \|w\|_2 \leq B\right\}$. Assume that $\mathbb{E}_{x \sim p} \|x\|^2 \leq C^2$. Then*

$$R_S(H) \leq \frac{B}{n} \sqrt{\sum_i \|x_i\|_2^2} \tag{1}$$

and

$$R_n(H) \leq \frac{BC}{\sqrt{n}}. \tag{2}$$

Proof. We have

$$\begin{aligned} R_S(H) &= \mathbb{E}_{\sigma} \sup_{\|w\|_2 \leq B} \frac{1}{n} \sum_i \sigma_i w^T x_i \\ &= \mathbb{E}_{\sigma} \sup_{\|w\|_2 \leq B} w^T \left(\frac{1}{n} \sum_i \sigma_i x_i \right) \\ &= \frac{B}{n} \mathbb{E}_{\sigma} \left\| \sum_i \sigma_i x_i \right\|_2 \end{aligned}$$

since the L_2-norm is self-dual

$$= \frac{B}{n} \left(\mathbb{E}_{\sigma} \left\| \sum_i \sigma_i x_i \right\|_2^2 \right)^{1/2}$$

by Jensen's inequality: $\mathbb{E}\, z = \mathbb{E} \sqrt{z^2} \leq \sqrt{\mathbb{E}\, z^2}$ for nonnegative z, since $\sqrt{\cdot}$ is concave

$$\leq \frac{B}{n} \left(\mathbb{E}_{\sigma} \left[\sum_i \cancelto{1}{\sigma_i^2} \|x_i\|_2^2 + \cancelto{0}{\sum_{i \neq j} \sigma_i \sigma_j x_i^T x_j} \right] \right)^{1/2}$$

where the second sum vanishes because the σ_is are independent with mean 0

$$= \frac{B}{n} \sqrt{\sum_i \|x_i\|_2^2}$$

This proves Eqn. (1). Eqn. (2) follows from taking an expectation:

$$R_n(H) = \mathbb{E}\, R_S(H) = \frac{B}{n} \mathbb{E} \sqrt{\sum_i \|x_i\|_2^2} \leq \frac{B}{n} \sqrt{\sum_i \mathbb{E} \|x_i\|_2^2} \leq \frac{BC}{\sqrt{n}}$$

where the first inequality is another application of Jensen, and the second follows by definition of C. □

Theorem 2. *Let $H = \{x \mapsto w^T x : \|w\|_1 \leq B\}$. Assume that $\|x_i\|_\infty \leq C$ for all i. Let d be the dimension of the input; i.e. $x_i \in \mathbb{R}^d$ for all i. Then*

$$R_n(H) \leq \frac{2BC\sqrt{\log 2d}}{\sqrt{n}}.$$

We note first the similarities and differences between Theorems 1 and 2. Theorem 1 looks somewhat weaker: we have to make a stronger assumption, namely, $\|x_i\|_\infty \leq C$ for all i, instead of $\mathbb{E}\|x_i\|_\infty \leq C$, and the addition of a factor of $2\sqrt{\log 2d}$. We also remark that it is possible—even likely—that Theorems like 1 and 2 apply for other pairs of norm-dual norm pairs. However, since the most common norms are L_1, L_2, L_∞, we will restrict our attention to these for this class. We will now give a sketch of the proof of Theorem 2. The full proof is left as a homework problem.

Proof Sketch. We will prove a slightly different statement, namely that with probability $1 - d^{-O(1)}$ over the random choice of σ, we have

$$\sup_{\|w\|_1 \leq B} \frac{1}{n} \sum_i \sigma_i w^T x_i \lesssim \frac{2BC\sqrt{\log d}}{\sqrt{n}}. \tag{3}$$

This looks similar enough to our theorem that the theorem sounds believable as long as the tail of the distribution of the LHS is not too heavy.

$$\sup_{\|w\|_1 \leq B} \frac{1}{n} \sum_i \sigma_i w^T x_i = \sup_{\|w\|_1 \leq B} w^T \left(\frac{1}{n} \sum_i \sigma_i x_i \right) = \frac{B}{n} \left\| \sum_i \sigma_i x_i \right\|_\infty$$

Let $v = \sum_i \sigma_i x_i$. We claim that, with probability $1 - d^{-O(1)}$ over the random choice of σ, $\|v\|_\infty \lesssim C\sqrt{n \log d}$, from which (3) follows immediately. To see this, fix an index j. Then $v_j = \sum_i \sigma_i x_{ij}$. Since the x_{ij} are all bounded by C and the σ_i are independent, Hoeffding's inequality implies that

$$\Pr[|v_j| \geq \varepsilon] \leq 2 \exp\left(-O\left(\frac{\varepsilon^2}{nC^2} \right) \right).$$

Taking $\varepsilon = O(C\sqrt{n \log d})$, we conclude that $\Pr[|v_j| \geq \varepsilon] \leq d^{-O(1)}$. Now a union bound over $j = 1, \ldots, d$ gives that $\Pr\left[\|v_j\| \lesssim C\sqrt{n \log d}\right] \leq d^{-O(1)}$, as desired. Notice that the constant hidden by the $O(1)$ can be made small by increasing the constant hidden in the expression for ε; we will thus not concern ourselves with figuring out the constant factors since it is unnecessary. □

3 Rademacher Complexity of Neural Networks

Somewhat surprisingly, it has been empirically observed [1] that, in a neural network with a single hidden layer and no regularization, increasing the number of hidden units improves (or, at least, does not hurt) the generalization loss even past the point when the number of hidden units suffices to achieve zero training error. A recent paper [2] gives some theoretical bounds that justify these practical observations. In the next two lectures, we will see some of these bounds.

Throughout this section, we will use the following notation:

- $\Theta = (w, U)$ are the parameters of the model
- $f_\Theta(x) = w^T \phi(Ux)$ is the model function
- m is the number of hidden units (i.e. the number of rows of U, and the number of elements of w)
- $\phi(x) = \max(x, 0)$ is the (element-wise) ReLU function
- $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ is the training set

The main theorem we will seek to prove has the form

Theorem 3 (Theorem 3.1 in [2], informally)**.** *Suppose we train a two-layer neural network to minimize a logistic loss function. Then the generalization error decreases as the training error increases.*

The proof and precise statement of this result will be deferred to the next lecture. We will first warm up by showing a simpler statement.

Theorem 4 (Special case of Theorem 43, in Section 6.4 of Percy Liang's notes)**.** *Let*

$$H = \left\{ f_\Theta : \|w\|_2 \leq B' \text{ and } \|u_i\| \leq B \text{ for all } i \right\}$$

where u_i is the ith column of U. Then

$$R_n(H) \leq 2BB'C\sqrt{\frac{m}{n}}.$$

Proof. Let $g_U(x) \triangleq \phi(Ux)$ for matrices U. Then $f_\Theta(x) = w^T g_U(x)$. Thus:

$$\begin{aligned} R_S(H) &= \mathop{\mathbb{E}}_{\sigma} \sup_{\Theta} \frac{1}{n} \sum_i \sigma_i w^T g_U(x_i) \\ &= \mathop{\mathbb{E}}_{\sigma} \sup_{\Theta} w^T \left(\frac{1}{n} \sum_i \sigma_i g_U(x_i) \right) \\ &= \mathop{\mathbb{E}}_{\sigma} \sup_{\Theta} \|w\|_2 \left\| \frac{1}{n} \sum_i \sigma_i g_U(x_i) \right\|_2 \end{aligned}$$

since we can always pick w to point in the same direction as $\frac{1}{n} \sum_i \sigma_i g_U(x_i)$

$$\begin{aligned} &= B' \mathop{\mathbb{E}}_{\sigma} \sup_{U : \|u_j\|_2 \leq B \, \forall j} \left\| \frac{1}{n} \sum_i \sigma_i g_U(x_i) \right\|_2 \\ &\leq B' \sqrt{m} \mathop{\mathbb{E}}_{\sigma} \max_j \sup_{\|u_j\|_2 \leq B} \left| \frac{1}{n} \sum_i \sigma_i \phi(u_j^T x_i) \right| \end{aligned}$$

by symmetry

$$= B' \sqrt{m} \mathop{\mathbb{E}}_{\sigma} \sup_{\|u\|_2 \leq B} \left| \frac{1}{n} \sum_i \sigma_i \phi(u^T x_i) \right|.$$

The expectation looks very suspiciously like $R_S\big(\{x \mapsto \phi(u^T x) : \|u\|_2 \le B\}\big)$; we only differ from the definition of Rademacher complexity by an absolute value sign. In fact, some sources, including the original paper on Rademacher complexity [3] define it with the absolute value sign. An analogue of Talagrand's Lemma for this definition is stated in Theorem 12 (Section 3.1) of [3], in which we lose an extra factor of 2. Using this, since ϕ is 1-Lipschitz, we can continue

$$\begin{aligned}
R_S(H) &\le 2B'\sqrt{m}\,\mathop{\mathbb{E}}_{\sigma} \sup_{\|u\|_2 \le B} \left| \frac{1}{n} \sum_i \sigma_i u^T x_i \right| \\
&= 2B'\sqrt{m}\,\mathop{\mathbb{E}}_{\sigma} \sup_{\|u\|_2 \le B} \frac{1}{n} \sum_i \sigma_i u^T x_i \\
&= 2B'\sqrt{m} R_S\big(\{x \mapsto u^T x : \|u\|_2 \le B\}\big) \\
R_n(H) &\le 2B'\sqrt{m} R_n\big(\{x \mapsto u^T x : \|u\|_2 \le B\}\big) \le 2BB'C\sqrt{\frac{m}{n}}
\end{aligned}$$

where the last line follows from Theorem 1. □

Theorem 4 is weaker than the theorem we would like to prove, since there is a $\sqrt{m}$ in the numerator which we'd like to get rid of. We will refine this bound in the next lecture.

References

[1] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[2] C. Wei, J. D. Lee, Q. Liu, and T. Ma. On the Margin Theory of Feedforward Neural Networks. *ArXiv e-prints*, October 2018.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.