CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Sina Semnani

Lecture #8
October 17, 2018

1 Review and Overview

In the last session we introduced an interesting phenomenon that occures during the training of a two-layer feed-forward neural network and attempted to explain it using Rademacher complexity. Empirically, we observe that increasing the number of neurons in the hidden layer of such networks (i.e. increasing the over-parametrization) reduces the generalization error. We also proved a bound for Rademacher complexity of two-layer neural netwroks that grew with m (number of hidden neurons).

In this session we will prove a stronger bound that will not scale with m. We also apply this bound and discuss how this bound can be useful.[1]

2 Stronger bound for Rademacher complexity

First, let us set up a few notations. Let $\Theta = (w, U)$ denote the parameters of the neural network where $U \in \mathbb{R}^{m\times d}$ and $w \in \mathbb{R}^m$ are the first and second layer weights. $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ for $i \in [n]$ are the input data and labels. In this section we only consider the binary classification problem.

The activation function is defined as

$$\phi(x) = \mathrm{ReLU}(x) = \max\{x, 0\}$$

Given parameters Θ, the neural network computes the function

$$f_\Theta(x) = \sum_{j=1}^{m} w_j \phi\left(u_j^T x\right)$$

where u_j^T is the jth row of matrix U.

Define

$$\mathcal{H}' \triangleq \{f_\Theta : \|w\|_2 \le B_2', \|u_j\|_2 \le B_2 \ \ \forall j = 1, \ldots, m\}$$

In the last lecture we proved that if $\|x_i\|_2 \le C$, then:

$$\mathcal{R}_S(\mathcal{H}') = \frac{2B_2 B_2' C\sqrt{m}}{\sqrt{n}} \tag{1}$$

Theorem 1. *Define* $B(w, U) \triangleq \sum_{j=1}^{m} |w_j| \|u_j\|_2$ *and* $\mathcal{H} \triangleq \{f_\Theta : B(w, U) \le B_1\}$. *If* $\|x_i\|_2 \le C$ *then*

$$\mathcal{R}_S(\mathcal{H}) \le \frac{2B_1 C}{\sqrt{n}} \tag{2}$$

[1] For more details and additional proofs see [1].

Note that this expression might still implicitly depend on m since B_1 can increase with m. We will address this issue later by showing that dividing by γ cancels out the effect of m.

When using bound 2 in practice, we calculate $B(w, U)$ after the neural network is trained, and plug in $B_1 = B(w, U)$ to get the upper bound. Similarly, for bound 1 we plug in B_2 and B_2'. In this sense, for every fixed, trained neural network, bound 2 is at least as strong as bound 1. The following inequalities justify this statement:

$$\begin{aligned} B_1 = B(w, U) &= \sum_{j=1}^{m} |w_j| \, \|u_j\|_2 \\ &\leq \sqrt{\left(\sum_{i=1}^{m} w_j^2\right)\left(\sum_{i=1}^{m} \|u_j\|_2^2\right)} \qquad \text{Cauchy–Schwarz inequality} \\ &\leq \|w\|_2 \sqrt{m} \; \max_j \|u_j\|_2 \\ &\leq B_2' B_2 \sqrt{m} \end{aligned}$$

Therefore

$$\frac{2B_1 C}{\sqrt{n}} \leq \frac{2B_2 B_2' C \sqrt{m}}{\sqrt{n}}.$$

Using Theorem 1, we can prove a generalization bound with the following form:

$$L_\gamma(h) \lesssim \hat{L}_\gamma(h) + \frac{\mathcal{R}_S(\mathcal{H})}{\gamma} + \sqrt{\frac{\log(\frac{2}{\delta})}{n}}$$

Theorem 2. *Fix B_1 and $\gamma > 0$ and define $\mathcal{H}_{B_1} \triangleq \{f_\Theta : B(w, U) \leq B_1\}$. Then with probability $\geq 1 - \delta$:*

$$\forall h \in \mathcal{H}_{B_1} : L_\gamma(h) \lesssim \hat{L}_\gamma(h) + \frac{B_1 C}{\gamma \sqrt{n}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}$$

We can think of $\frac{\gamma}{B_1}$ as the normalized margin.

We now prove Theorem 1:

Proof. Define $\hat{u}_j \triangleq \frac{u_j}{\|u_j\|_2}$ as the length-normalized u_j.

$$\begin{aligned}
\mathcal{R}_S(\mathcal{H}) &= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{f_\Theta\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i f_\Theta(x_i)\right] && \text{definition of } R_S(\mathcal{H}) \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{f_\Theta\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{m}w_j\phi\left(u_j^T x_i\right)\right] && \text{definition of } f_\Theta \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{f_\Theta\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{m}w_j\|u_j\|_2\,\phi\left(\hat{u}_j^T x_i\right)\right] && \text{because } \phi(ax)=a\phi(x) \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{f_\Theta\in\mathcal{H}}\sum_{j=1}^{m}w_j\|u_j\|_2\sum_{i=1}^{n}\sigma_i\phi\left(\hat{u}_j^T x_i\right)\right] \\
&\le \frac{1}{n}\mathop{\mathbb{E}}_{\sigma_i}\left[\left(\sup_{f_\Theta\in\mathcal{H}}\sum_{j=1}^{m}|w_j|\,\|u_j\|_2\right)\left(\max_{1\le j\le m}\left|\sum_{i=1}^{n}\sigma_i\phi\left(\hat{u}_j^T x_i\right)\right|\right)\right] \\
&= \frac{1}{n}B_1\mathop{\mathbb{E}}_{\sigma_i}\left[\max_{1\le j\le m}\left|\sum_{i=1}^{n}\sigma_i\phi\left(\hat{u}_j^T x_i\right)\right|\right] \\
&\le \frac{1}{n}B_1\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{\|\hat{u}\|_2=1}\left|\sum_{i=1}^{n}\sigma_i\phi\left(\hat{u}_j^T x_i\right)\right|\right]
\end{aligned}$$

We know that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\sigma_i}\left[\sup_{\|\hat{u}\|_2=1}\left|\sum_{i=1}^{n}\sigma_i\phi\left(\hat{u}_j^T x_i\right)\right|\right] &\le 2n\mathcal{R}_n(\{x\mapsto\phi(u^T x):\|u\|_2\le 1\}) \\
&\le 2n\mathcal{R}_n(\{x\mapsto u^T x:\|u\|_2\le 1\}) \\
&\le 2n\frac{C}{\sqrt{n}} \\
&= 2\sqrt{n}C
\end{aligned}$$

The first inequality is Lemma 1 from scribe note 5 (the two sides are multiplied by n), the second inequality uses Talagrand's lemma and the fact that ϕ is a 1-Lipschitz function. We proved the third inequality in the previous lecture. So

$$\mathcal{R}_S(\mathcal{H}) \le \frac{1}{n}B_1\mathop{\mathbb{E}}_{\sigma_i}\left[\left|\sup_{\|\hat{u}\|_2=1}\sum_{i=1}^{n}\sigma_i\phi(\hat{u}^T x_i)\right|\right] \le \frac{2B_1C}{\sqrt{n}}$$

□

3 Margin-based generalization error

Intuitively, we want to show that if we optimize loss plus a small regularization factor, normalized margin will be large, hence the generalization error will be small. In this section

we prove the theorem for exponential loss instead of logistic loss since it is easier, but the results hold for logistic loss as well.

First, let us define a few notations. Define the λ-regularized exponential loss as

$$L_\lambda(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i f_\Theta(x_i)\right) + \lambda \left\|w\right\|_2^2 + \lambda \left\|U\right\|_F^2$$

where $\|w\|_2^2 + \|U\|_F^2 = \|\Theta\|_2^2$.

Let $\Theta_{\lambda,m}$ be the global optimizer of $L_{\lambda,m}$ (in a neural network with m hidden neurons).

Margin of a parameter is defined to be the following:

$$\gamma(\Theta) = \min_{1\le i\le n} \; y_i f_\Theta(x_i)$$

Note that if a neural network misclassifies some inputs, the margin can be negative.

Margin of the global optimizer is defined as

$$\gamma_{\lambda,m} = \gamma\left(\frac{\Theta_{\lambda,m}}{\left\|\Theta_{\lambda,m}\right\|_2}\right)$$

Maximum possible margin for a network with m hidden neurons is

$$\gamma^{*,m} \triangleq \max_{\|\Theta\|_2\le 1} \gamma(\Theta)$$

and $\Theta^{*,m} \triangleq \arg\max_{\|\Theta\|_2\le 1} \; \gamma(\Theta)$ is the parameter that achieves that maximum.

Lemma 1 (homogeneity of feed-forward neural networks)**.** $f_{\alpha\Theta}(x) = \alpha^2 f_\Theta(x) \;\; \forall \alpha \in \mathbb{R}$.

Proof.

$$f_{\alpha\Theta}(x) = \sum_{j=1}^{m} \alpha w_j \phi\left(\alpha u_j^T x\right) = \sum_{j=1}^{m} \alpha^2 w_j \phi\left(u_j^T x\right) = \alpha^2 f_\Theta(x)$$

□

Theorem 3. *Assume $\gamma^{*,m} > 0$. Then as $\lambda \to 0$, $\gamma_{\lambda,m} \to \gamma^{*,m}$ and $\Theta_{\lambda,m} \to \Theta^{*,m}$.*

Remark 1. *$\gamma^{*,1} \le \gamma^{*,2} \le \ldots$ i.e. $\gamma^{*,m}$ is non-decreasing in the hidden layer size.*

This is due to the fact that a network with $m+1$ neurons in its hidden layer can simulate a network with m neurons in the hidden layer by setting all additional parameters to zero.

Remark 2. *As we increase m (over-parametrization), the upper bound on the generalization error gets smaller (better).*

$$\begin{aligned}
B_1(w^{*,m}, U^{*,m}) &= \sum_{j=1}^{m} \left|w_j^{*,m}\right| \left\|u_j^{*,m}\right\|_2 \\
&\le \sum_{j=1}^{m} \frac{1}{2}\left(\left|w_j^{*,m}\right|^2 + \left\|u_j^{*,m}\right\|_2^2\right) \\
&= \frac{1}{2}\left(\|w^{*,m}\|_2^2 + \|U^{*,m}\|_F^2\right) \\
&= \frac{1}{2}\|\Theta^{*,m}\|_2^2 \\
&\le \frac{1}{2}
\end{aligned}$$

So we have

$$L_{\gamma^{*,m}}(f_{\Theta^{*,m}}) \leq \hat{L}_{\gamma^{*,m}}(f_{\Theta^{*,m}}) + \frac{C}{2\gamma^{*,m}\sqrt{n}} + \sqrt{\frac{\log(\frac{2}{\delta})}{n}}$$

From the above statement and the previous remark, we conclude this remark.
We now prove Theorem 3:

Proof.

$$\begin{aligned} L_\lambda(\Theta_\lambda) &\leq L_\lambda(\|\Theta_\lambda\|_2 \Theta^*) \\ &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i f_{\|\Theta_\lambda\|_2 \Theta^*}(x_i)\right) + \lambda \|\Theta_\lambda\|_2^2 \|\Theta^*\|_2^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i \|\Theta_\lambda\|_2^2 f_{\Theta^*}(x_i)) + \lambda \|\Theta_\lambda\|_2^2 \end{aligned}$$

$\forall i : y_i \|\Theta_\lambda\|_2^2 f_{\Theta^*}(x_i) \geq \|\Theta_\lambda\|_2^2 \gamma^*$ by definition of $\gamma^* = \min_{1\leq i\leq n} y_i f_{\Theta^*}(x_i)$.
Therefore

$$L(\|\Theta_\lambda\|_2 \Theta^*) \leq \exp(-\|\Theta_\lambda\|_2^2 \gamma^*) + \lambda \|\Theta_\lambda\|_2^2$$

□

We will continue this proof in the next lecture.

References

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