Foundations of Deep Learning – Homework Assignment #4

Adi Album & Tomer Epshtein

Part 2: (2)

Question:

Let $H = \{h_\theta : \mathcal{X} \to \mathcal{Y} :\ \theta \in \mathbb{R}^p, \|\theta\|_\infty \leq 0.5\}$ be a hypothesis space corresponding to a neural network with p parameters bounded in $[-0.5, 0.5]$. For any subset $\Theta \subseteq \{\theta \in \mathbb{R}^p : \|\theta\|_\infty \leq 0.5\}$, denote $\mathcal{H}_\Theta \coloneqq \{h_\theta : \theta \in \Theta\}$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ and a training set $S = \{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y}$, the Radamacher complexity of $\mathcal{H}_\Theta$ is defined to be:

$$R(\ell \circ \mathcal{H}_\Theta \circ S) \coloneqq \frac{1}{m} \mathbb{E}_\xi \left[\sup_{v \in \ell \circ \mathcal{H}_\Theta \circ S} \sum_{i=1}^m \xi_i v_i \right]$$

Where:

- ξ is short for $\xi_1, \ldots, \xi_m \overset{i.i.d}{\sim} \begin{cases} +1, & w.p\ 0.5 \\ -1, & w.p\ 0.5 \end{cases}$
- $\ell \circ \mathcal{H}_\Theta \circ S = \left\{ \left(\ell(y_1, h(x_1)), \ell(y_2, h(x_2)), \ldots, \ell(y_m, h(x_m)) \right) : h \in \mathcal{H}_\Theta \right\} \subseteq \mathbb{R}^m$

Assume that

$$\mathbb{E}_S[R(\ell \circ \mathcal{H}_\Theta \circ S)] = Volume(\Theta) \coloneqq \int_{\theta \in \mathbb{R}^p} 1_{[\theta \in \Theta]} d\theta$$

Assume also that the implicit regularization of optimization leads to solutions with high $\|\cdot\|_\infty$, i.e., to:

$$h_{\hat\theta} \in \mathcal{H}\ , \hat\theta \in argmax_{\theta \in \mathbb{R}^p, \|\theta\|_\infty \leq 0.5} \|\theta\|_\infty\ s.t\ \theta\ minimizes\ training\ loss$$

Derive a generalization bound for $\mathcal{H}$ that takes advantage of our knowledge on the implicit regularization, i.e. under which learned solutions with high $\|\cdot\|_\infty$ ensure small generalization gap.

Proof:

Define $\Theta^{(c)} \coloneqq \left\{ \theta \in \mathbb{R}^p : \frac{1}{2} - c <= \|\theta\|_\infty \leq \frac{1}{2} \right\}$ for $c \in \left[0, \frac{1}{2}\right]$.

Let $\epsilon > 0$ be a small positive real value such that for some $t \in \mathbb{N}$: $t\epsilon = \frac{1}{2}$. We define a series $\Theta_1, \Theta_2, \Theta_3, \ldots$ by:

- For all $i \leq t$: $\Theta_i \coloneqq \Theta^{(i \cdot \epsilon)}$
- For all $i > t$: $\Theta_i \coloneqq \Theta$

We have $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$. For each $i \in \mathbb{N}$, Θ_i defines $\mathcal{H}_{\Theta_i}$ and this produces a series of subsets: $\mathcal{H}_{\Theta_1} \subseteq \mathcal{H}_{\Theta_2} \subseteq \cdots$

1. For every $k \in \mathbb{N}$ such that $k\epsilon \leq \frac{1}{2}$:

$$\mathbb{E}_S\left[R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right)\right] = Volume(\Theta_k) = \left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p$$

$R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right)$ is a non-negative random variable, so by Markov's Inequality:
$\forall a > 0$:

$$\Pr_S\left(R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) \geq a\right) \leq \frac{\mathbb{E}_S\left[R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right)\right]}{a} = \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{a} \overset{?}{\leq} \frac{\delta}{2}$$

We want to choose a such that $(?)$ holds
I.e.

$$\frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{a} \overset{?}{\leq} \frac{\delta}{2}$$

Choose:

$$a = \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}$$

So we have

$$\Pr_S\left(R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) \geq \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right) \leq \frac{\delta}{2}$$

Or by viewing complement:

$$\Pr_S\left(R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right) > 1 - \frac{\delta}{2}$$

2. By proposition proved in class,
   $\forall k \in \mathbb{N}, \delta \in (0,1)$ w.p. $\geq 1-\frac{\delta}{2}$ over $S \sim D^m$:

$$h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_\mathrm{k} \circ S) + \sqrt{\frac{2 \cdot \ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}$$

   Let $\delta \in (0,1)$, $k \in \mathbb{N}$.

$$\Pr\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}\right) \geq 1 - \frac{\delta}{2}$$

3. Reminder: For any two probabilistic events A and B we have

$$\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$$

Let $k \in \mathbb{N}$ such that $k\epsilon \leq \frac{1}{2}$. Let's look at:

$$\Pr\left(\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}\right) \cap \left(R\left(\ell \circ \mathcal{H}_{\Theta_\mathrm{k}} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right)\right)$$

- On the one hand:

$$\Pr\left(\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}\right) \cap \left(R\left(\ell \circ \mathcal{H}_{\Theta_\mathrm{k}} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right)\right) \overset{(3)}{\geq}$$

$$\geq \Pr\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}\right) + \Pr\left(R\left(\ell \circ \mathcal{H}_{\Theta_\mathrm{k}} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right) - 1 \overset{(1)\&(2)}{\geq}$$

$$\geq \left(1 - \frac{\delta}{2}\right) + \left(1 - \frac{\delta}{2}\right) - 1 = 1 - \delta$$

- On the other hand:

$$\Pr\left(\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}\right) \cap \left(R\left(\ell \circ \mathcal{H}_{\Theta_\mathrm{k}} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right)\right) \leq$$

$$\leq \Pr\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/4} + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}\right)$$

Bringing it all together, we have:

$\forall \delta \in (0,1), \forall \epsilon > 0$ *and* $k \in \mathbb{N}$ such that $k\epsilon \leq \frac{1}{2}$, w.p. $\geq 1 - \delta$

$$h \in \mathcal{H}_{\Theta_k}: L_D(h) - L_S(h) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/4} + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3}k^2\frac{2}{\delta}\right)}{m}}$$

Therefor, solutions with high $\|\cdot\|_\infty$ are hypothesis' $h \in \mathcal{H}_k$ with small k, yielding small generalization gaps.