University of Salzburg | Lecturer: Roland Kwitt

Machine Learning (911.236)

Exercise sheet **F**

Exercise 1. 6 P.

We are going to do a (step-by-step) proof of an (empirical) Rademacher complexity bound for a **two-layer neural network**, i.e., functions of the form

$$f_\theta = \langle \mathbf{w}, \phi(\mathbf{U}\mathbf{x}) \rangle$$

with $\mathbf{U} \in \mathbb{R}^{m \times d}$, $\mathbf{w} \in \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^d$. Here, $\phi(a) = \max(a, 0)$ will be the **ReLU** activation function (which operates component-wise on each dimension of $\mathbf{U}\mathbf{x}$). In particular, our hypothesis class $\mathcal{H}$ is

$$\mathcal{H} = \{f_\theta : \|\mathbf{w}\|_2 \leq B, \forall j \in \{1, \ldots, m\} : \|\mathbf{u}_j\|_2 \leq C\}$$

where $\boldsymbol{\theta} = (\mathbf{U}, \mathbf{w})$ subsumes the parameters of the two-layer neural network. Also note that $\mathbf{u}_j$ denotes the j-th column of $\mathbf{U}$. You will start with the definition of empirical Rademacher complexity

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{H}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\theta \in \mathcal{H}} \sum_{i=1}^n \sigma_i f_\theta(\mathbf{x}_i)\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\theta \in \mathcal{H}} \sum_{i=1}^n \sigma_i \langle \mathbf{w}, \phi(\mathbf{U}\mathbf{x}_i)\rangle\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\mathbf{U}:\|\mathbf{u}_j\|_2 \leq B} \sup_{\|\mathbf{w}\|_2 \leq C} \sum_{i=1}^n \sigma_i \langle \mathbf{w}, \phi(\mathbf{U}\mathbf{x}_i)\right]
\end{aligned}$$

In the next step(s), you (1) write the summation term as an inner product of $\mathbf{w}$ and $\sum_i \ldots$ and (2) bound that inner product (and thus the whole expression) using the Cauchy-Schwartz inequality (as we did in the previous PS). This should allow you to eliminate one of the supremums. Next, (3) use $\|\mathbf{v}\|_2 \leq \sqrt{m}\|\mathbf{v}\|_\infty$ for $\mathbf{v} \in \mathbb{R}^m$ to bound the $\|\cdot\|_2$ norm (via the $\|\cdot\|_\infty$ norm) and (4) replace $\|\cdot\|_\infty$ by its definition. Overall, completing steps (1)–(4) will get you to

$$\cdots \leq \frac{B\sqrt{m}}{n}\mathbb{E}_\sigma\left[\sup_{\|\mathbf{u}\|_2 \leq C}\left|\sum_{i=1}^n \sigma_i \phi(\mathbf{u}^\top \mathbf{x}_i)\right|\right] \tag{1}$$

Lemma 1. *Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ be Rademacher variables (i.e., $\sigma_i \sim$ Uniform$(\{\pm 1\})$), then*

$$\mathbb{E}_\sigma\left[\sup_\theta \left|\sum_{i=1}^n \sigma_i f_\theta(\mathbf{x}_i)\right|\right] \leq 2\,\mathbb{E}_\sigma\left[\sum_{i=1}^n \sigma_i f_\theta(\mathbf{x}_i)\right]$$

In step (5), use Lemma 1 to bound Eq. (1).

Lemma 2 (Contraction). *For each $i \in \{1, \ldots, m\}$ let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function; namely, for all $\alpha, \beta \in \mathbb{R}$ we have $|\phi_i(\alpha) - \phi_i(\beta)| \leq \rho|\alpha - \beta|$. For $\mathbf{a} \in \mathbb{R}^m$ let $\phi(\mathbf{a})$ denote $(\phi(a_1), \ldots, \phi(a_m))$ and let $\phi \circ A = \{\phi(\mathbf{a}) : \mathbf{a} \in A\}$. Then*

$$\hat{\mathcal{R}}_S(\phi \circ A) \leq \rho \hat{\mathcal{R}}_S(A) \enspace .$$

Finally, (6) use Lemma 2, knowing that ReLU is 1-Lipschitz, and (7) complete the bound by using our result empirical Rademacher complexity bound for linear classes from the previous PS. For completeness, there we had $\mathcal{G} = \{\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{w}\rangle : \|\mathbf{w}\|_2 \leq 1\}$ and $\hat{\mathcal{R}}_S(\mathcal{G}) \leq \max_i \|\mathbf{x}_i\|_2 1/\sqrt{n}$.

Total #points: 6 P.