MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2018**

Problem Set 3

Due Date: **Sat, Nov. 10 2018, 11:59 pm** *(online)*

Instructions: There are **2 problems** in total in this problem set. The breakdown of individual scores per sub-problem are provided. Use the provided LaTeX template to typeset your report. Provide sufficient explanations in all solutions but avoid proving lecture or out-of-scope material (unless explicitly asked to). An 8-page submission maximum is allowed (do not change the font or margin of the template).

What to submit: Submit your report **online through Stellar** by the due date/time. Submission must be a single pdf in LaTeX format. Include code files separately, if applicable.

Policies: Collaborative reports are not allowed. Even if you discuss problems with classmates, you are expected to write and submit **individual reports**.

Problem 1 [60 points]

1.1 [10pts] Consider a class $\mathcal{F} = \{x \mapsto \mathbf{I}\{a \le x \le b\} : a, b \in \mathbb{R}\}$ on $\mathbb{R}$. Prove an $O\left(\sqrt{\frac{\log n}{n}}\right)$ upper bound for uniform deviations over $\mathcal{F}$:

$$\mathbb{E} \max_{f \in \mathcal{F}} \left[\mathbb{E} f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right] \le c \sqrt{\frac{\log n}{n}}$$

for some constant c. *Hint*: analyze Rademacher averages.

1.2 [10pts] Consider the zero-one loss function $\ell(f(x), y) = \mathbf{I}\{f(x) \neq y\}$, where y takes values in $\{\pm 1\}$ and f is also $\{\pm 1\}$-valued. Given a class $\mathcal{F}$ of such functions, consider the loss class

$$\ell \circ \mathcal{F} = \{(x, y) \mapsto \mathbf{I}\{f(x) \neq y\} : f \in \mathcal{F}\}.$$

Show that Rademacher averages of $\mathcal{F}$ and $\ell \circ \mathcal{F}$ coincide up to a multiplicative constant 2. *Hint*: write indicator loss $\mathbf{I}\{a \neq b\} = (1 - ab)/2$ for $a, b \in \{\pm 1\}$.

1.3 [10pts] Use 1.1 and 1.2 to argue that ERM $\widehat{f}_n$ over $\mathcal{F}$ with respect to zero-one loss enjoys

$$\mathbb{E}\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) \le \epsilon$$

as soon as (ignoring log factors) $n \ge \tilde{O}(\epsilon^{-2})$.

1.4 [10pts] Let $p \geq 1$. Show that Rademacher averages of unit ℓ_p-norm ball B_p^n scale as $O(n^{-1/p})$. *Hint*: use definition of dual norm.

1.5 [10pts] Use convexity of "max" to show that Rademacher averages are always non-negative.

1.6 [10pts] Use concavity of "min" to show that for an empirical minimizer $\widehat{f}_n$ over class $\mathcal{F}$,

$$\mathbb{E}_{\mathscr{S}}\widehat{\mathbf{L}}(\widehat{f}_n) \leq \mathbf{L}(f_{\mathcal{F}}),$$

where $f_{\mathcal{F}}$ is a minimizer of expected loss in $\mathcal{F}$.

Problem 2 [30 points] (Generalization error on finite hypotheses space) Recall Hoeffding's inequality: if $U_1, \ldots, U_n, U$ are i.i.d. real random variables with values in $[a, b]$, then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}U \geq t\right) \leq \exp\left\{-\frac{2nt^2}{(a-b)^2}\right\}$$

and

$$P\left(\mathbb{E}U - \frac{1}{n}\sum_{i=1}^{n} U_i \geq t\right) \leq \exp\left\{-\frac{2nt^2}{(a-b)^2}\right\}$$

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = [-M, M]$ for some $0 < M < \infty$ and consider a training set of n points sampled i.i.d from a fixed probability distribution P. Consider the square loss function ℓ and a hypothesis space comprised of N distinct functions, $\mathcal{F} = \{f_1, \ldots, f_N\}$ which are uniformly bounded, i.e. $\sup_{x\in\mathcal{X}} |f(x)| \leq C$ for all $f \in \mathcal{F}$. Recall that $\mathbf{L}(f) = \mathbb{E}\ell(Y, f(X))$ is the expected risk, and $\widehat{\mathbf{L}}(f)$ the empirical risk $\widehat{\mathbf{L}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$.

In this problem, we will derive learning guarantees in high probability *without* having to go through Rademacher averages.

2.1 [10pts] By applying Hoeffdings's inequality, derive an explicit bound on the probability

$$\Pr\left(\max_{f\in\mathcal{F}} \left|\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right| \geq \epsilon\right) \quad \forall \epsilon > 0. \tag{3.1}$$

2.2 [10pts] Let $\widehat{f}_n$ be the minimizer of the empirical risk on $\mathcal{F}$. Show that (3.1) implies that for any $0 < \delta \leq 1$, with probability at least $1 - \delta$, we have

$$\mathbf{L}(\widehat{f}_n) \leq \widehat{\mathbf{L}}(\widehat{f}_n) + \epsilon(n, N, \delta) \tag{3.2}$$

for some suitable function $\epsilon(n, N, \delta)$.

2.3 [10pts] Let $f_{\mathcal{F}}$ be the minimizer of the expected risk on $\mathcal{F}$. Show that (3.1) also implies that with probability at least $1 - \delta$ we have

$$\mathbf{L}(\widehat{f}_n) - \mathbf{L}(f_{\mathcal{F}}) \leq 2\epsilon(n, N, \delta). \tag{3.3}$$

Hint: add and subtract a few terms as we did in class and study the two difference-components of the expression individually.