ECE 543: Statistical Learning Theory

Maxim Raginsky

Homework 2

Assigned February 8; due February 15

Required reading: lecture notes of Hajek and Raginsky, Chapters 3–5

---

1. **Uniform deviations and Rademacher averages.** Let $\mathscr{F}$ be a class of functions $f : \mathsf{Z} \to [0,1]$. Given a probability distribution $P \in \mathscr{P}(\mathsf{Z})$ and an i.i.d. sample Z^n from P, consider the uniform deviation

$$\Delta_n(Z^n) = \|P_n - P\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - P(f) \right|$$

and the Rademacher average

$$R_n(\mathscr{F}(Z^n)) = \frac{1}{n} \mathbf{E}_{\sigma^n} \left\{ \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| \right\},$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables independent of Z^n. Prove that, for any $t > 0$,

$$\mathbf{P}\left(\Delta_n(Z^n) - 2R_n(\mathscr{F}(Z^n)) \geq \mathbf{E}\left[\Delta_n(Z^n) - 2R_n(\mathscr{F}(Z^n))\right] + t\right) \leq e^{-2nt^2/25}$$

and

$$\mathbf{P}\left(\Delta_n(Z^n) - 2R_n(\mathscr{F}(Z^n)) \geq t\right) \leq e^{-2nt^2/25}.$$

2. **A simple penalized ERM algorithm.** When choosing a hypothesis space, we often face the following dilemma: If the hypothesis space is not “rich" enough, even the best hypothesis from it may have unacceptably high expected risk. On the other hand, if it is too rich, then there may be no guarantee of good behavior of the uniform deviations of empirical means from true means. A great deal of effort in statistical learning theory is devoted to finding ways of coping with this dilemma. One such way is to use multiple hypothesis spaces, run ERM on each of them, and then choose the ERM solution that achieves a good trade-off between the empirical risk and some measure of complexity of the hypothesis space at hand. In this problem, you will investigate a very simple penalized ERM algorithm that chooses between finitely many hypothesis classes, where the complexity of each class is measured in a data-driven way by means of Rademacher averages.

Let $\mathscr{F}_1, \ldots, \mathscr{F}_M$ be a finite collection of hypothesis spaces, where each $\mathscr{F}_m$ is a class of functions from Z into $[0,1]$. Let Z^n be an i.i.d. sample from an unknown disribution $P \in \mathscr{P}(\mathsf{Z})$. For each $m \in [M]$, let

$$\widehat{f}_n^{(m)} := \operatorname*{arg\,min}_{f \in \mathscr{F}_m} P_n(f)$$

be an empirical risk minimizer over $\mathscr{F}_m$, and let $\widehat{f}_n = \widehat{f}_n^{(\widehat{m})}$, where

$$\widehat{m} = \operatorname*{arg\,min}_{m \in [M]} \left[P_n(\widehat{f}_n^{(m)}) + 2R_n(\mathscr{F}_m(Z^n)) \right].$$

Prove that

$$P(\widehat{f}_n) \le \min_{m \in [M]} \left\{ \inf_{f \in \mathscr{F}_m} P_n(f) + 2R_n(\mathscr{F}_m(Z^n)) \right\} + \sqrt{\frac{25 \log\left(\frac{M}{\delta}\right)}{2n}}$$

with probability at least $1 - \delta$.

Hint. You may need to use the bounds from the previous problem separately for each m and then combine them.

3. **Rademacher complexity of L^p balls.** Let $\mathscr{B}_{n,p} := \{v \in \mathbb{R}^n : \|v\|_p \le 1\}$, where $\|v\|_p := \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}$ for $0 < p < \infty$ and $\|v\|_\infty = \max_i |v_i|$.

   (a) Find $R_n(\mathscr{B}_{n,p})$ for $n \ge 1$ and $1 \le p \le \infty$. (**Hint**: For any $y \in \mathbb{R}^n$, $\sup_{v \in \mathscr{B}_{n,p}} \langle y, v \rangle = \|y\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$.)

   (b) Let $\|v\|_0$ denote the number of nonzero coordinates of a vector v. For k with $1 \le k \le n$, let $\mathscr{B}_{n,k,p} = \{v \in \mathbb{R}^n : \|v\|_p \le 1 \text{ and } \|v\|_0 \le k\}$. Find $R_n(\mathscr{B}_{n,k,p})$ for $1 \le k \le n$, and $1 \le p \le \infty$.

4. **Some Rademacher estimates.** Consider the following three classes of binary-valued functions:

   - $\mathscr{F}_1$ is the collection of indicators of all semi-infinite intervals of the form $(-\infty, t]$, $t \in \mathbb{R}$, with the domain $\mathsf{Z} = \mathbb{R}$
   - $\mathscr{F}_2$ is the collection of indicators of all closed intervals of the form $[s, t]$ for $-\infty < s < t < \infty$, with the domain $\mathsf{Z} = \mathbb{R}$
   - $\mathscr{F}_3$ is the collection of all indicators of subsets of $\{1, 2, \ldots\}$ with cardinality of at most k, on the domain $\mathsf{Z} = \{1, 2, \ldots\}$.

   Without relying on VC theory, prove the following:

$$R_n(\mathscr{F}_1) \le 2\sqrt{\frac{\log(n+1)}{n}}, \qquad \forall n$$

$$R_n(\mathscr{F}_2) \le 2\sqrt{\frac{2\log n + \log 2}{n}}, \qquad \forall n$$

$$R_n(\mathscr{F}_3) \le 2\sqrt{\frac{k \log(ne/k)}{n}}, \qquad \forall n \ge k$$

   where for a class of real-valued functions $\mathscr{F}$ on Z we let

$$R_n(\mathscr{F}) := \sup_{z^n \in \mathsf{Z}^n} R_n(\mathscr{F}(z^n)).$$

5. **VC classes.** Prove the following statements.

   (a) Let $\mathscr{C}$ and $\mathscr{C}'$ be two classes of subsets of some feature space X. Suppose that $\mathscr{C} \subseteq \mathscr{C}'$, meaning that if $C \in \mathscr{C}$, then $C \in \mathscr{C}'$ as well. Prove that $V(\mathscr{C}) \leq V(\mathscr{C}')$.

   (b) Let $\mathscr{C}$ be a *finite* class of subsets of X. Prove that $V(\mathscr{C}) \leq \log_2 |\mathscr{C}|$.

   (c) Let X be a *finite* feature space. For a given $k \leq |\mathsf{X}|$, consider the class $\mathscr{F}_k$ of binary-valued functions $f : \mathsf{X} \to \{0,1\}$, such that $\left|\{x \in \mathsf{X} : f(x) = 1\}\right| = k$. Find $V(\mathscr{F}_k)$.