ECE 543: Statistical Learning Theory

Maxim Raginsky

March 5, 2021

Homework 2, v2

Assigned March 4; due March 16, 2021

Note: natural logarithms are used throughout.

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

$$\Delta_n(Z^n) = \|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$$

and the Rademacher average

$$R_n(\mathcal{F}(Z^n)) = \mathbf{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i) \right| \right]$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of n i.i.d. Rademacher random variables independent of Z^n. Prove that, for any $t > 0$,

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

and

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

2. **A simple covering number bound on Rademacher averages.** Consider an arbitrary space Z. The *sup norm* of any $f : \mathsf{Z} \to \mathbb{R}$ is defined as

$$\|f\|_\infty := \sup_{z \in \mathsf{Z}} |f(z)|.$$

Let $\mathcal{F}$ be a class of real-valued functions on Z. Given an $\varepsilon > 0$, we say that a finite set of functions $\{f_1, \ldots, f_k\}$ (not necessarily in $\mathcal{F}$) is an *ε-net* for $\mathcal{F}$ (w.r.t. the sup norm) if for any $f \in \mathcal{F}$ there exists at least one $j \in \{1, \ldots, k\}$ such that

$$\|f - f_j\|_\infty \equiv \sup_{z \in \mathsf{Z}} |f(z) - f_j(z)| \leq \varepsilon.$$

The ε-*covering number* of $\mathcal{F}$ w.r.t. the sup norm, denoted by $N_\infty(\mathcal{F}, \varepsilon)$, is the cardinality of a minimal ε-net of $\mathcal{F}$. If $\mathcal{F}$ does not admit an ε-net, then we set $N_\infty(\mathcal{F}, \varepsilon) = +\infty$.

(a) Suppose that all the functions in $\mathcal{F}$ are uniformly bounded, i.e., there exists some $L > 0$, such that $\|f\|_\infty \leq L$ for all $f \in \mathcal{F}$. Prove that

$$R_n(\mathcal{F}) \leq \inf_{\varepsilon > 0} \left(\varepsilon + 2L \sqrt{\frac{\log N_\infty(\mathcal{F}, \varepsilon)}{n}} \right).$$

[The logarithm of the covering number is called the ε-*entropy* of $\mathcal{F}$ and denoted by $H_\infty(\mathcal{F}, \varepsilon)$.]

(b) Let

$$\mathsf{Z} = \left\{ z = (z^{(1)}, \ldots, z^{(d)}) \in \mathbb{R}^d : \|z\|_1 = \sum_{j=1}^{d} |z^{(j)}| \leq 1 \right\}$$

and let $\mathcal{F}$ consist of all functions of the form $f(z) = f_w(z) = \langle w, z \rangle$ for all $w \in \mathbb{R}^d$ with $\|w\|_\infty = \max_{1 \leq j \leq d} |w^{(j)}| \leq 1$. Prove that

$$N_\infty(\mathcal{F}, \varepsilon) \leq \left(\frac{2}{\varepsilon} \right)^d,$$

and then use this fact to prove that

$$R_n(\mathcal{F}) = O\left(\sqrt{\frac{d \log n}{n}} \right).$$

(c) Suppose that $\mathcal{F}$ is such that $H_\infty(\mathcal{F}, \varepsilon) \leq C\varepsilon^{-1/\alpha}$ for some constants $C > 0$ and $\alpha > 0$. (For example, if $\mathcal{F}$ is the class of all differentiable functions $f : [0,1] \to [0,1]$ with $|f'| \leq 1$, then the above bound holds with $\alpha = 1$.) Use the result of part (a) to prove that

$$R_n(\mathcal{F}) \leq C n^{-\frac{\alpha}{2\alpha+1}}$$

for some constant $C > 0$.

3. **An alternative to ERM.** Searching for an empirical risk minimizer in an infinite function class $\mathcal{F}$ may not always be feasible. Let's consider the following alternative procedure that reduces to searching over a *finite* subclass of $\mathcal{F}$. We will be looking at a binary classification problem, so let $\mathcal{F}$ be a class of functions $f : \mathsf{X} \to \{0, 1\}$, where X is some feature space. Given a training sample $\{(X_i, Y_i)\}_{i=1}^n$ from an unknown probability distribution P on $\mathsf{X} \times \{0, 1\}$, we carry out the following two-step procedure:

   - Pick some $m < n$. Let $\mathcal{B}_m$ be the set of all binary strings in $\{0,1\}^m$ of the form

   $$b^m(f) = (b_1(f), \ldots, b_m(f)) = (f(X_1), \ldots, f(X_m)) \tag{1}$$

   for some $f \in \mathcal{F}$. For each $b \in \mathcal{B}_m$, pick one $f \in \mathcal{F}$ such that (1) holds. Let $\widehat{\mathcal{F}}_m$ denote the (finite) set of all such f's. Note that this is a random set, since it depends on the sample $(X_1, \ldots, X_m)$.

- Compute

$$\widehat{f}_n = \arg\min_{f \in \widehat{\mathcal{F}}_m} \frac{1}{n-m} \sum_{i=m+1}^{n} \mathbf{1}_{\{f(X_i) \neq Y_i\}}.$$

  This will be our actual classifier.

What we have done is split the original training sample into two subsamples, used the first subsample to extract a finite subclass of $\mathcal{F}$, and then performed ERM over this subclass on the second subsample. We will now analyze the classification error of $\widehat{f}_n$.

(a) Let

$$\tilde{f}_m = \arg\min_{f \in \widehat{\mathcal{F}}_m} L(f)$$

be the best classifier in $\widehat{\mathcal{F}}_m$. Note that this is a random object, since it depends on the random set $\widehat{\mathcal{F}}_m$. Prove that

$$L(\widehat{f}_n) - L(\tilde{f}_m) \leq 8\sqrt{\frac{\log |\mathbb{S}_m(\mathcal{F})|}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} \tag{2}$$

with probability at least $1 - \delta/2$, where $\mathbb{S}_m(\mathcal{F})$ is the mth shatter coefficient of $\mathcal{F}$.
Hint: Use the fact that $\widehat{f}_n$ is a solution of an ERM problem over the second subsample, add and subtract appropriate empirical quantities, and then apply the Finite Class Lemma.

(b) Observe that

$$\begin{aligned} L(\tilde{f}_m) - L^*(\mathcal{F}) &\leq \sup_{f,f' \in \mathcal{F}:\, b^m(f) = b^m(f')} \left| L(f) - L(f') \right| \\ &\leq \sup_{f,f' \in \mathcal{F}:\, b^m(f) = b^m(f')} \mathbf{P}\left[f(X) \neq f(X') \right] \\ &\leq \sup_{A \in \mathcal{A}} \left| P(X \in A) - \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{X_i \in A\}} \right|, \end{aligned}$$

where $\mathcal{A}$ is the class of all sets of the form $\{x \in \mathsf{X} : f(x) \neq f'(x)\}$ for all pairs $f, f' \in \mathcal{F}$. Use this to prove that

$$L(\tilde{f}_m) - L^*(\mathcal{F}) \leq C\sqrt{\frac{V(\mathcal{A})}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

with probability at least $1 - \delta/2$, where $C > 0$ is an absolute constant. (It is not hard to show that $V(\mathcal{A}) \leq 4V(\mathcal{F})$ — you don't have to do this.)

(c) Finally, use parts (a)–(b) to prove that

$$L(\widehat{f}_n) - L^*(\mathcal{F}) \leq 8\sqrt{\frac{\log |\mathbb{S}_m(\mathcal{F})|}{n-m}} + C\sqrt{\frac{V(\mathcal{A})}{m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

with probability at least $1 - \delta$.

4. **A simple penalized ERM scheme.** When choosing a suitable 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 population 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 $\mathcal{F}_1, \ldots, \mathcal{F}_M$ be a finite collection of hypothesis spaces, where each $\mathcal{F}_m$ consists of functions f from some space Z into $[0,1]$. Let Z^n be an i.i.d. sample from an unknown distribution P over Z. For each $m = 1, \ldots, M$ let

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

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

$$\widehat{m} = \operatorname*{arg\,min}_{1 \le m \le M} \left\{ P_n(\widehat{f}_n^{(m)}) + 2R_n(\mathcal{F}_m(Z^n)) \right\}.$$

   Prove that

$$P(\widehat{f}_n) \le \min_{1 \le m \le M} \left\{ \inf_{f \in \mathcal{F}_m} P_n(f) + 2R_n(\mathcal{F}_m(Z^n)) \right\} + \sqrt{\frac{25 \log\left(\frac{M}{\delta}\right)}{n}}$$

   with probability at least $1 - \delta$.