ECE 598MR: Statistical Learning Theory

Maxim Raginsky

Homework 1

Assigned October 6, 2015; due October 15, 2015

---

1. **Convexity.** Let I be an interval of the real line. A function $f : I \to \mathbb{R}$ is called *convex* if

$$f\big(\lambda x + (1-\lambda) y\big) \le \lambda f(x) + (1-\lambda) f(y)$$

for all $\lambda \in [0,1]$ and all $x, y \in I$. Equivalently, f is convex if the straight line segment joining the points $(x, f(x))$ and $(y, f(y))$ for any two $x, y \in I$ lies above the graph of f. Here are some useful facts about convex functions:

- **Second-order condition.** If I is an open interval and f is twice differentiable on I, then it is convex if and only if $f''(x) \ge 0$ for all $x \in I$.
- **Jensen's inequality.** If $f : I \to \mathbb{R}$ is convex, then for any random variable X with values in I,

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)].$$

In this problem, you will get to explore the world of convex functions.

(a) Given a pair $x, y \in I$, consider the function $F_{x,y} : [0,1] \to \mathbb{R}$, defined by

$$F_{x,y}(t) = f\big(x + t(y-x)\big).$$

Prove that f is convex if and only if $F_{x,y}$ is convex for all $x, y \in I$.

(b) Suppose that $g : [0, a] \to \mathbb{R}$ is convex and monotone increasing. Prove that the function $f(x) = g(|x|)$ is convex on the interval $[-a, a]$.

(c) Use convexity to prove the following inequality: for any $a > 0$,

$$e^{ax} \le \cosh a + x \sinh a, \qquad -1 \le x \le 1.$$

(d) Let U be a real-valued random variable. Prove that its logarithmic moment-generating function $\psi(a) = \log \mathbb{E}[e^{aX}]$ is convex on the real line. (You may assume that interchanging derivative and expectation is permissible.)

2. **Generalizing Hoeffding's inequality.** In class, we have proved Hoeffding's inequality that gives an exponential bound on the deviation probability $\mathbb{P}[|X_1 + \ldots + X_n| \geq t]$ for a sum of independent random variables that are bounded and have zero mean. In this problem, you will develop a generalization of Hoeffding's inequality to sums of dependent random variables that satisfy a certain weak orthogonality condition.

   (a) In preparation for the rest of the problem, derive the inequality

   $$\cosh x \leq e^{x^2/2}, \qquad x \in \mathbb{R}$$

   as a consequence of Hoeffding's lemma.

   Hint: Find a suitable bounded random variable U, such that $\cosh x = \mathbb{E}[e^{xU}]$.

   (b) We say that a collection $X_1, \ldots, X_n$ of random variables is a *multiplicative system* if, for any $1 \leq k \leq n$ and any set of k indices $1 \leq i_1 < i_2 < \ldots < i_k \leq n$,

   $$\mathbb{E}[X_{i_1} X_{i_2} \ldots X_{i_k}] = 0.$$

   Prove that if $X_1, \ldots, X_n$ are a multiplicative system, then

   $$\mathbb{E}\left[\prod_{i=1}^{n} (a_i X_i + b_i)\right] = \prod_{i=1}^{n} b_i$$

   for any choice of real constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$.

   (c) Let $U_1, \ldots, U_n$ be n possibly dependent random variables, and let Z be any real-valued random variable jointly distributed with them. For each i, let $X_i = \mathbb{E}[Z|U^i] - \mathbb{E}[Z|U^{i-1}]$ (where $\mathbb{E}[Z|U^0] \equiv \mathbb{E}Z$). Prove that $X_1, \ldots, X_n$ are a multiplicative system.

   (d) Consider a multiplicative system $X_1, \ldots, X_n$, such that $-c_i \leq X_i \leq c_i$ for each i, where $c_i > 0$ are some finite constants. Prove that, for any $t > 0$,

   $$\mathbb{E}\left[\exp\left(t \sum_{i=1}^{n} X_i\right)\right] \leq \prod_{i=1}^{n} \cosh(t c_i).$$

   (e) Now for the final step: prove that if $X_1, \ldots, X_n$ are a multiplicative system of random variables satisfying the boundedness condition of part (c), then

   $$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

3. **Finite concept classes are PAC-learnable.** Consider the concept learning problem in the realizable setting. Let $\mathscr{P}$ be the space of all probability distributions on the feature space X, and let $\mathscr{F}$ be a finite class of binary-valued functions $f : \mathsf{X} \to \{0, 1\}$ (any concept class can be equivalently represented as a function class by associating each concept with its indicator function). Consider the following learning algorithm $\mathscr{A} = \{A_n\}_{n=1}^{\infty}$: for every $n \geq 1$,

   $$\widehat{f}_n = A_n((X_1, Y_1), \ldots, (X_n, Y_n)) = \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - Y_i\big)^2 \tag{1}$$

(if there are several minimizers, an arbitrary rule is used to break ties). This algorithm is called *Empirical Risk Minimization* (ERM). In this problem, you will prove that this algorithm satisfies

$$\bar{r}_{\mathscr{A}}(n,\varepsilon,\mathscr{P}) \le |\mathscr{F}|(1-\varepsilon)^n. \tag{2}$$

Consequently, if the number of samples n satisfies

$$n \ge \frac{\log(|\mathscr{F}|/\delta)}{\varepsilon}, \tag{3}$$

then the target concept f^* can be learned with accuracy ε and confidence $1-\delta$.

(a) Consider an arbitrary distribution P of the feature X and an arbitrary target function $f^* \in \mathscr{F}$. Given the training set $Z^n = (Z_1,\ldots,Z_n)$, where $Z_i = (X_i, Y_i) = (X_i, f^*(X_i))$, $1 \le i \le n$, and the X_i's are drawn i.i.d. from P, define the *training error* of a function $f : \mathsf{X} \to [0,1]$ by

$$L_{Z^n}(f,f^*) \triangleq \frac{1}{n}\sum_{i=1}^n \big(f(X_i)-Y_i\big)^2 = \frac{1}{n}\sum_{i=1}^n \big(f(X_i)-f^*(X_i)\big)^2.$$

Prove that the hypothesis $\widehat{f}_n$ produced by the ERM algorithm (1) achieves zero training error, i.e., $L_{Z^n}(\widehat{f}_n, f^*) = 0$.

(b) Define the set of *bad* functions

$$\mathscr{F}_{\text{bad}} \triangleq \big\{f \in \mathscr{F} : L_P(f,f^*) \ge \varepsilon\big\},$$

as well as the event consisting of *misleading* samples

$$\mathscr{M} \triangleq \big\{Z^n \in \mathsf{Z}^n : L_{Z^n}(f,f^*) = 0 \text{ for some } f \in \mathscr{F}_{\text{bad}}\big\},$$

where $\mathsf{Z} = \mathsf{X} \times [0,1]$. In other words, a training sample Z^n is misleading if it results in zero training error for some bad function. Prove that $\big\{Z^n \in \mathsf{Z}^n : L_P(\widehat{f}_n, f^*) \ge \varepsilon\big\} \subseteq \mathscr{M}$ and conclude from this that

$$P^n_{f^*}\big(Z^n \in \mathsf{Z}^n : L_P(\widehat{f}_n, f^*) \ge \varepsilon\big) \le P^n_{f^*}\big(Z^n \in \mathscr{M}\big).$$

(c) Prove that the probability of getting a misleading sample is given by

$$P^n_{f^*}\big(Z^n \in \mathscr{M}\big) = P^n_{f^*}\left(\bigcup_{f \in \mathscr{F}_{\text{bad}}} \bigcap_{i=1}^n \big\{f(X_i) = f^*(X_i)\big\}\right). \tag{4}$$

(d) Prove that (4) implies (2).

(e) Compare the sample complexity estimate (3) with what you get using symmetrization and Rademacher averages. Which sample complexity estimate is better? Can you think of a reason why?

4. **An alternative to ERM.** Searching for an empirical risk minimizer in an infinite function class $\mathscr{F}$ may not always be feasible. Let's consider the following alternative procedure that reduces to searching over a *finite* subclass of $\mathscr{F}$. We will be looking at a binary classification problem, so let $\mathscr{F}$ 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 $\mathscr{B}_m$ be the set of all binary strings in $\{0,1\}^m$ of the form

$$b^m(f) = (b_1(f), \ldots, b_m(f)) = \big(f(X_1), \ldots, f(X_m)\big) \tag{5}$$

for some $f \in \mathscr{F}$. For each $b \in \mathscr{B}_m$, pick one $f \in \mathscr{F}$ such that (5) holds. Let $\widehat{\mathscr{F}}_m$ 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

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

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 $\mathscr{F}$, and then performed ERM over this subclass on the second subsample. We will now analyze the classification error of $\widehat{f}_n$.

(a) Let

$$\tilde{f}_m = \operatorname*{arg\,min}_{f \in \widehat{\mathscr{F}}_m} L(f)$$

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

$$L(\widehat{f}_n) - L(\tilde{f}_m) \leq 8\sqrt{\frac{\log|\mathbb{S}_m(\mathscr{F})|}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} \tag{6}$$

with probability at least $1 - \delta/2$, where $\mathbb{S}_m(\mathscr{F})$ is the mth shatter coefficient of $\mathscr{F}$.
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

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

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

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

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

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

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

with probability at least $1-\delta$.

5. **A simple chaining estimate for Rademacher averages.** Consider an arbitrary space Z. The *sup norm* of any $f : \mathsf{Z} \to \mathbb{R}$ is defined as

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

Let $\mathscr{F}$ 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 $\mathscr{F}$) is an *ε-net* for $\mathscr{F}$ (w.r.t. the sup norm) if for any $f \in \mathscr{F}$ there exists at least one $j \in \{1, \ldots, k\}$ such that

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

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

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

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

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

(b) Let

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

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

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

and then use this fact to prove that

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

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

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

for some constant $C > 0$.