ECE 598MR: Statistical Learning Theory

Maxim Raginsky

Homework 1

Assigned September 11, 2014; due September 23, 2014

Note: natural logarithms are used throughout.

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. **Bin packing.** This is a classical application of McDiarmid's inequality. Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$. Each X_i is the size of a package to be shipped. The packages are shipped in bin of size 1, so each bin can hold any set of packages whose sizes sum to at most 1. Let $B_n = f(X_1, \ldots, X_n)$ be the minimal number of bins needed to ship the packages with sizes $X_1, \ldots, X_n$. Computing B_n is a hard combinatorial optimization problem; however, we can say

something about its mean and tail behavior.

(a) Let μ be the common mean of the X_i's. Prove that $\mathbb{E}B_n \geq n\mu$.

(b) Prove that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{B_n}{n} \leq \mu - \varepsilon\right) \leq \exp\left(-2n\varepsilon^2\right).$$

4. **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 $n \geq 1$,

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

(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

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

Consequently, if the number of samples n satisfies

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

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 $Z_i = (X_i, Y_i) = (X_i, f^*(X_i))$, $1 \leq i \leq n$, 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

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

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

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

as well as the event consisting of *misleading* samples

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

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 $\left\{Z^n \in \mathsf{Z}^n : L_P(\widehat{f}_n, f^*) \geq \varepsilon\right\} \subseteq \mathscr{M}$ and conclude from this that

$$P_{f^*}^n\left(Z^n \in \mathsf{Z}^n : L_P(\widehat{f}_n, f^*) \geq \varepsilon\right) \leq P_{f^*}^n\left(Z^n \in \mathscr{M}\right).$$

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

$$P_{f^*}^n\left(Z^n \in \mathscr{M}\right) = P_{f^*}^n\left(\bigcup_{f \in \mathscr{F}_{\text{bad}}} \bigcap_{i=1}^{n} \{f(X_i) = f^*(X_i)\}\right). \tag{3}$$

(d) Finally, prove that (3) implies (2).