ECE 901
Homework 7 - Faster Rates

1. Classification in Threshold Classes. Consider a binary classification problem in which the feature space $\mathcal{X} = [0,1]$, and the Bayes classifier has the form $f^*(x) = \mathbf{1}_{x \geq t}$ for a threshold $t \in (0,1)$.

a. Devise an empirical risk minimization procedure (based on Chernoff's bound) for designing a classifier from n i.i.d. training examples. Show that the excess risk of such a procedure decays like $O(\sqrt{\log n/n})$. Using Chernoff's bound, one can never achieve a rate faster than $n^{-1/2}$. Why?

b. In general, an excess risk decay rate of $n^{-1/2}$ is the best one can achieve in classification, but under certain assumptions on the "noise" it is possible to design classifiers whose risks converge faster to the Bayes risk. Reconsider the threshold problem above, but now assume two conditions:

i. P_X, the distribution of the features, has a density function p_X that is bounded below by a constant $b > 0$.

ii. The conditional probability $\eta(x) := P(Y = 1|X = x)$ satisfies $|\eta(x) - 1/2| \geq c$, for some constant $c > 0$.

Condition (ii) implies that the conditional probability "jumps" across the 1/2-level. Assuming that these two conditions hold, show that it is possible to construct a classifier from the training data whose excess risk is $O(\log n/n)$.

HINTS: Partition the unit interval into m equal sized bins. Use the relative form of Chernoff's bound and condition (i) to show that with very high probability, each bin will contain at least $c_1 n/m$ samples, for some constant $c_1 > 0$. Condition on this event and then study the probabilities $p_i = P(Y = 1|X \in \text{ bin i})$ and their empirical counterparts $\widehat{p}_i$. Using Chernoff's bound and condition (ii) to show that the tests $\widehat{p}_i \geq 1/2$, $i = 1, \ldots, n$, will be correct (i.e., agree with $p_i \geq 1/2$) with very high probability. Combining everything thus far will give you a bound on the estimation error of the form $c_2 m\, e^{-c_3 n/m}$, for some constants $c_2, c_3 > 0$. The approximation error for an m bin histogram is $O(1/m)$. Balance terms to obtain the claimed $O(\log n/n)$ rate of convergence for the excess risk.

2. Complexity regularization in regression. Consider learning under squared error loss. Suppose we have n iid training examples $\{X_i, Y_i\}_{i=1}^n$ and a collection $\mathcal{F}$ of candidate functions mapping $\mathcal{X} = R^d$ to $\mathcal{Y} = R$. Assume that the support of the Y_i and the range of the candidate functions $f \in \mathcal{F}$ is in a known interval of length b. Our empirical and true risks are given by

$$\begin{aligned} \widehat{R}_n(f) &= \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 \\ R(f) &= E[(f(X) - Y)^2] = \frac{1}{n}\sum_{i=1}^{n} E\left[(f(X_i) - Y_i)^2\right] \end{aligned}$$

Notice that we can write $R(f) - \widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n (U_i - E[U_i])$, where $U_i = (f(X_i) - Y_i)^2$. Assume that we have the following concentration inequalities for $\sum_{i=1}^n (U_i - E[U_i])$:

$$P\left(\sum_{i=1}^{n}(U_i - E[U_i]) > \epsilon\right) \leq e^{-\epsilon} \quad \text{and} \quad P\left(\sum_{i=1}^{n}(E[U_i] - U_i) > \epsilon\right) \leq e^{-\epsilon}$$

Select a model from $\mathcal{F}$ according to

$$\widehat{f}_n = \arg\max_{f \in \mathcal{F}} \left\{ \widehat{R}_n(f) + \frac{c(f)\log 2}{n} \right\}$$

where $\{c(f)\}$ are positive numbers satisfying $\sum_{f\in\mathcal{F}} 2^{-c(f)} \leq 1$. Using the concentration inequalities above and mimicking the derivation of the risk bound for complexity regularized *classifier* selection in Lecture 10, prove that

$$E[R(\widehat{f}_n)] \;\leq\; \min_{f\in\mathcal{F}} \left\{ R(f) + \frac{c(f)\log 2}{n} \right\} \;+\; \frac{\log n + b^2 + 1}{n}$$

Compare this result to the bound one obtains using Hoeffding's inequality

$$P\left(\sum_{i=1}^{n} (U_i - E[U_i]) > \epsilon\right) \;\leq\; e^{-2\epsilon^2/n} \quad \text{and} \quad P\left(\sum_{i=1}^{n} (E[U_i] - U_i) > \epsilon\right) \;\leq\; e^{-2\epsilon^2/n}$$