Instructions: Turn in your homework in class on Tuesday 10/3/2017

1. VC classes.
   (a) What is the VC-dimension of origin-centered radial classifiers (i.e. \( h_b(x) = 1\{\|x\|^2 \le b\} \)) in 2-dimensions? Remember to prove that the VC-dimension is \( d \), you must show both the lower and the upper bound.
   (b) What if we also allow for the center of the classifier to move, i.e. \( h_{x_0,b}(x) = 1\{\|x - x_0\|^2 \le b\} \)?
   (c) What is the VC-dimension of convex polyhedral sets in 2 dimensions? Here the function class is \( \mathcal{H} = \{ h_{A,b}(x) = 1\{Ax \le b\} \} \). This class can be equivalently represented by choosing a set of points \( y_1, \ldots, y_T \) and letting \( h_{y_1:T}(x) = 1\{x \in \text{conv}\{y_1, \ldots, y_T\}\} \), which may be easier to reason about.

2. Rademacher calculus. In this problem we will study how the Rademacher complexity of a simple two-layer neural network. Let \( F \subset (\mathbb{R}^d \to \mathbb{R}) \) be a class of sigmoid functions \( f_w(x) = \sigma(\langle w, x \rangle) \) where \( \sigma(a) = \frac{e^a}{1+e^a} \) with the norm constraint \( \|w\|_2 \le 1 \). Fix the “hidden layer” dimension \( d_1 \) and let \( \mathcal{H} \subset (\mathbb{R}^{d_1} \to \mathbb{R}) \) be the analogous sigmoid functions, but with the restriction that \( \sum_{i=1}^{d_1} w_i = 1 \) and \( w_i \ge 0 \) for all \( i \). What is the rademacher complexity of the class of functions \( h(f_1(x), \ldots, f_{d_1}(x)) \) where \( h \in \mathcal{H}, f_1, \ldots, f_{d_1} \in F \)?
   **Bonus (for fun)** Prove that the same bound holds when we just have the restriction that the weight vectors in \( \mathcal{H} \) have \( \|w\|_1 \le 1 \).

3. Goodness-of-Fit Testing. A number of scientific applications involve collecting some data and testing if the data are distributed according to some pre-specified distribution (the hypothesis). A specific formalism of this is goodness-of-fit testing problem, and in this problem we will study a one dimensional problem where we work with cumulative distribution functions. We are given data \( X_1, \ldots, X_n \) iid from a distribution with CDF \( F \) and we want to test:

   \[
   \text{null } H_0 : F = G \quad \text{alternate } H_1 : F \in \Theta_{c} = \{ G' | \|G' - G\|_{\infty} \ge c \}
   \]

   A test statistic \( T : \mathbb{R}^n \to \{0, 1\} \) looks at the data and makes a prediction about whether the null or the alternate is true. We are typically interested in

   \[
   \text{Type I error } \mathbb{P}_{X_1^n \sim G}[T(X_1^n) = 1] \quad \text{Type II error } \sup_{G' \in \Theta_{c}} \mathbb{P}_{X_1^n \sim G'}[T(X_1^n) = 0]
   \]

   In words the Type I errors are when the data really does come from \( G \), but you fail to detect it and the Type II errors are when the data does not come from \( G \) but you think that it does.

   The KS test is a popular test statistic for one-dimensional data. The procedure takes the \( n \) samples \( X_1, \ldots, X_n \sim F \) and forms the empirical cumulative distribution function \( F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \le t\} \). Then we define \( T(X_1^n) = 1\{|F_n - G|_{\infty} > \tau\} \) for some parameter \( \tau \).

   We will give a crude analysis of this procedure.
(a) To analyze the procedure, first establish uniform convergence. Prove that with probability at least \(1 - \delta\)

\[
\sup_t |F_n(t) - F(t)| \leq \sqrt{\frac{2 \log n}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.
\]

Note that a sharper inequality known as the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality is possible, essentially removing the \(\log(n)\) term. This inequality is easier to prove with the tools we currently have.

(b) (Type I Error) How should we set \(\tau\) to ensure that the Type I error is at most \(\alpha\)?

(c) (Type II Error) For this choice of \(\alpha\), how small can \(\epsilon\) be for the Type II error to be at most \(\beta\)?

4. **Histogram Density Estimation.** Let \(X_1, \ldots, X_n\) be an iid sample from a distribution \(P\) with density \(p\) supported on \([0, 1]^d\). In a histogram estimator, we divide \([0, 1]^d\) into hypercubes \(B_1, \ldots, B_N\) of side length \(h\) (the bandwidth) and estimate \(p\) with

\[
\hat{p}_h(x) = \sum_{j=1}^{N} \frac{\hat{\pi}_j}{h^d} 1\{x \in B_j\} \quad \text{where} \quad \hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \in B_j\}.
\]

Assume that \(p \in \mathcal{P} = \{p \mid \|p(x) - p(y)\| \leq L \|x - y\|\}\) upper bound the MSE at a single point \(x\), i.e. 

\[\mathbb{E}_{X^n}(\hat{p}_h(x) - p(x))^2\] as a function of \(h\). What is the optimal choice of \(h\) (as a function of \(n, d, L\)) to minimize the bound and what is the resulting MSE?