Computational and Statistical Learning theory
Assignment 4

Due: March 2nd
Email solutions to : karthik *at* ttic *dot* edu

Notations/Definitions

Recall the definition of sample based Rademacher complexity :

$$\widehat{\mathcal{R}}_S(\mathcal{F}) := \mathbb{E}_{\epsilon\sim\{\pm 1\}^n}\left[\frac{1}{m}\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\epsilon_i f(x_i)\right]$$

Definition 1. *Given a sample $S = \{x_1, \ldots, x_m\}$, and any $\alpha > 0$, a set $V \subset \mathbb{R}^m$ is said to be an α-cover (in ℓ_p) of function class $\mathcal{F}$ on sample S if*

$$\forall f \in \mathcal{F}, \exists v \in V \text{ s.t. } \left(\frac{1}{m}\sum_{i=1}^{m}|f(x_i) - v_i|^p\right)^{1/p} \leq \alpha$$

Specifically for $p = \infty$ $\left(\frac{1}{m}\sum_{i=1}^{m}|f(x_i) - v_i|^p\right)^{1/p}$ is replaced by $\max_{i\in[m]}|f(x_i) - v_i|$.
Also define

$$\mathcal{N}_p(\mathcal{F}, \alpha, S) := \min\{|V| : V \text{ is an } \alpha\text{-cover of (in } \ell_p) \text{ of} \mathcal{F} \text{ on sample S}\}$$

and

$$\mathcal{N}_p(\mathcal{F}, \alpha, n) := \sup_{x_1,\ldots,x_m} \mathcal{N}_p(\mathcal{F}, \alpha, \{x_1, \ldots, x_m\})$$

Definition 2. *A function $\mathcal{F}$ is said to α-shatter a sample $S = \{x_1, \ldots, x_m\}$ if there exists a sequence of thresholds, $s_1, \ldots, s_m \in \mathbb{R}$ such that*

$$\forall \epsilon \in \{\pm 1\}^m,\ \exists f \in \mathcal{F} \text{ s.t. } \forall i \in [m], \quad \epsilon_i(f(x_i) - s_i) \geq \alpha/2$$

Problems

1. **VC Lemma for Real-valued Function classes :**
   We shall prove that for any function class $\mathcal{F}$ (assume functions in $\mathcal{F}$ are bounded by 1) and scale $\alpha > 0$, the ℓ_∞ covering number at scale α can be bounded using fat shattering dimension at that scale by proving a statement analogous to VC lemma. We shall proceed by first extending the statement to finite (specifically $\{0, \ldots, k\}$) valued function classes and then using this to prove the final bound of form

$$\mathcal{N}_\infty(\mathcal{F}, \alpha, n) \leq \sum_{i=1}^{\mathrm{fat}_\alpha} \binom{n}{i} \left(\frac{1}{\alpha}\right)^i$$

   (a) Let $\mathcal{F}_k \subset \{0, \ldots, k\}^{\mathcal{X}}$ be a function class with $\mathrm{fat}_2(\mathcal{F}_k) = d$, show that

$$\mathcal{N}_\infty(\mathcal{F}_k, 1/2, m) \leq \sum_{i=1}^{d} \binom{m}{i} k^i$$

   Show the above statement using induction on $n + d$ (very similar to first problem on Assignment 2). Hint : In Assignment 2 problem 1 where we used $\mathcal{H}_S^+$ and $\mathcal{H}_S^-$ use instead, for all $i \in \{0, \ldots, k\}$, $\mathcal{F}_i = \{f \in \mathcal{F} : f(x') = i\}$ (note that this is a simple multilable extension and for $k = 1$, $\mathcal{F}_0, \mathcal{F}_1$ are identical to $\mathcal{H}^+, \mathcal{H}^-$). Use the notion of 2-shattering instead of shattering for the VC case and use $1/2$-cover instead of growth function.

   (b) Using the idea of α-discretizing the output of function class $\mathcal{F}$ we shall conclude the required statement. Do the following :

      i. Create a $\{0, \ldots, k\}$-valued class $\mathcal{G}$ where k is of order $1/\alpha$. Show that covering $\mathcal{G}$ at scale $1/2$ implies we can cover $\mathcal{F}$ at scale α and hence conclude that we can bound $\mathcal{N}_\infty(\mathcal{F}, \alpha, m)$ in terms of covering number at scale $1/2$ for $\mathcal{G}$.

      ii. Show that $\mathrm{fat}_2(\mathcal{G}) \leq \mathrm{fat}_\alpha(\mathcal{F})$

      iii. Combine with the bound on $\mathcal{N}_\infty(\mathcal{G}, 1/2, m)$ from previous sub-problem and conclude that

$$\mathcal{N}_\infty(\mathcal{F}, \alpha, n) \leq \sum_{i=1}^{\mathrm{fat}_\alpha} \binom{n}{i} \left(\frac{1}{\alpha}\right)^i$$

2. **Dudley Vs Pollards' Bounds :**
   In class we saw that Rademacher complexity can be bounded in terms of covering numbers using Pollard's bound, Dudley integral bound and the slightly modified version of Dudley

integral bound as follows :

$$\widehat{\mathcal{R}}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ \alpha + \sqrt{\frac{\mathcal{N}_1(\mathcal{F}, \alpha, m)}{m}} \right\} \leq \inf_{\alpha \geq 0} \left\{ \alpha + \sqrt{\frac{\mathcal{N}_2(\mathcal{F}, \alpha, m)}{m}} \right\} \qquad \text{(Pollard)}$$

$$\widehat{\mathcal{R}}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + 12 \int_{\alpha}^{1} \sqrt{\frac{\mathcal{N}_2(\mathcal{F}, \tau, m)}{m}} d\tau \right\} \qquad \text{(Refined Dudley)}$$

In this problem using some examples we shall compare these bounds.

(a) Class with finite VC subgraph-dimension :
Assume that the VC subgraph-dimension of function class $\mathcal{F}$ is bounded by D. In this case result in problem 1 can be used to bound the covering number of $\mathcal{F}$ in terms of D. Use this bound on covering number and compare Pollard's bound with refined Dudley integral bound by writing down the bounds implied by each one.

(b) Linear class with bounded norm : Linear classes in high dimensional spaces is probably one of the most important and most used function class in machine learning. Consider the specific example where

$$\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\} \quad \text{and} \quad \mathcal{F} = \{\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} : \|\mathbf{w}\|_2 \leq 1\}$$

In class we saw that for any $\epsilon > 0$, $\mathrm{fat}_\epsilon(\mathcal{F}) \leq \frac{4}{\epsilon^2}$. Using this with the result in problem 1 we have that :

$$\mathcal{N}_2(\mathcal{F}, \alpha, m) \leq \mathcal{N}_\infty(\mathcal{F}, \alpha, m) \leq \left(\frac{en}{\epsilon}\right)^{\frac{4}{\epsilon^2}}$$

Use the above bound on the covering number and write down the bound on Rademacher complexity implied by Pollard's bound. Write down the bound on Rademacher complexity implied by the refined version of the Dudley integral bound.

3. **Data Dependent Bound :**
Recall the Rademacher complexity bound we proved in class for functions $\mathcal{F}$ bounded by 1. For any $\delta > 0$ with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}\left[f(x)\right] - \hat{\mathbb{E}}_S[f(x)] \right) \leq 2\mathbb{E}_{S \sim \mathcal{D}^m} \left[\widehat{\mathcal{R}}_S(\mathcal{F}) \right] + \sqrt{\frac{\log(1/\delta)}{m}}$$

Note that we don't know the distribution $\mathcal{D}$. One way we used the above bound was by providing upper bounds on $\widehat{\mathcal{R}}_S(\mathcal{F})$ for any sample of size m and using this instead of $\mathbb{E}_{S \sim \mathcal{D}^m} \left[\widehat{\mathcal{R}}_S(\mathcal{F}) \right]$. But ideally we would like to get tight bounds when the distribution we are faced with is nicer. The aim of this problem is to do this.

Prove that, for any $\delta > 0$ with probability at least $1 - \delta$, over draw of sample $S \sim \mathcal{D}^m$,

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}\left[f(x)\right] - \hat{\mathbb{E}}_S[f(x)] \right) \leq 2\widehat{\mathcal{R}}_S(\mathcal{F}) + K\sqrt{\frac{\log(2/\delta)}{m}}$$

(provide explicit value of constant K above). Notice that in the above bound the expected Rademacher complexity is replaced by sample based one which can be calculated from the training sample.

Hint : Use McDiarmid's inequality on the expected Rademacher complexity.

4. **Learnability and Fat-shattering Dimension** Recall the setting of stochastic optimization problem where objective function is mapping $r : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}$. Sample $S = \{z_1, \ldots, z_m\}$ drawn iid from unknown distribution $\mathcal{D}$ is provided to the learner and the aim of the learner is to output $\hat{h} \in \mathcal{H}$ based on sample that has low expected objective $\mathbb{E}[r(h, z)]$.

   (a) Consider the stochastic optimization problem with r bounded by a, i.e. $|r(h, z)| < a < \infty$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. If function class $\mathcal{F} := \{z \mapsto r(h, z) | h \in \mathcal{H}\}$ has finite fat_α for all $\alpha > 0$, then show that the problem is learnable.

   (b) Conclude that for a supervised learning problem with bounded hypothesis class $\mathcal{H}$ (ie. $\forall x \in \mathcal{X},\ |h(x)| < a$), and loss $\phi : \hat{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$ that is L-Lipschtiz (in first argument), if $\mathcal{H}$ has finite fat_α for all $\alpha > 0$, then the problem is learnable.

   (c) Show a stochastic optimization problem that is learnable even though it has infinite fat_α for all $\alpha \leq 0.1$ (or any other cosntant of your choice). Explicitly write down the hypothesis class, and the learning rule which learns the class, argue that the problem is learnable, and explain why the fat_α is infinite.

   Hint : You can make the learning rule that is successful to even be ERM.

   (d) Prove that for a supervised learning problem with the absolute loss $\phi(\hat{y}, y) = |\hat{y} - y|$, if the fat_α is infinite for some $\alpha > 0$, then the problem is not learnable.

   Hint: as with the binary case, for every m, construct a distribution which is concentrated on a set of points that can be fat-shattered.

Challenge Problems

1. We saw that for any distribution $\mathcal{D}$, the expected Rademacher complexity provided an upper bound on the maximum deviation between mean and average uniformly over function class, specifically we saw that

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(x)] - \hat{\mathbb{E}}[f(x)]\right)\right] \leq 2\mathbb{E}_{S\sim\mathcal{D}^m}\left[\widehat{\mathcal{R}}_S(\mathcal{F})\right]$$

   Prove the (almost) converse that

$$\frac{1}{2}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\widehat{\mathcal{R}}_S(\mathcal{F})\right] \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(x)] - \hat{\mathbb{E}}[f(x)]\right)\right]$$

This basically establishes that Rademacher complexity tightly bounds the uniform maximal deviation for every distribution.

2. The worst case Rademacher complexity is defined as

$$\widehat{\mathcal{R}}_m(\mathcal{F}) = \sup_{S=\{x_1,\ldots,x_m\}} \widehat{\mathcal{R}}_S(\mathcal{F})$$

(ie. supremum over samples of size m).

(a) Prove that for any function class $\mathcal{F}$ and any $\tau > \widehat{\mathcal{R}}_m(\mathcal{F})$, we have that

$$\mathrm{fat}_\tau(\mathcal{F}) \leq \frac{4m\widehat{\mathcal{R}}_m(\mathcal{F})^2}{\tau^2}$$

Hint : First start by proving the statement for larger sample of size $m' = \lceil \frac{m}{\mathrm{fat}_\tau} \rceil \mathrm{fat}_\tau$ by taking fat_τ samples and repeating them appropriate number of times. You will need to start with a shattered set and you will need to use Kintchine's inequality which states that for any n,

$$\mathbb{E}_{\epsilon \sim \mathrm{Unif}\{\pm 1\}^n} \left[\left| \sum_{i=1}^n \epsilon_i \right| \right] \geq \sqrt{\frac{n}{2}}$$

(b) Combine the above with the refined version of Dudley integral bound to prove that

$$\inf_{\alpha \geq 0} \left\{ 4\alpha + 12 \int_\alpha^1 \sqrt{\frac{\mathcal{N}_2(\mathcal{F}, \tau, m)}{m}} d\tau \right\} \leq \widehat{\mathcal{R}}_m(\mathcal{F})\, O(\log^{3/2} m)$$

This shows that the refined dudley integral bound is tight to within log factors of the Rademacher complexity. Thus we have established that in the worst case all the complexity measures for function class like Rademacher complexity, covering numbers and fat shattering dimension all tightly govern the rate of uniform maximal deviation for the function class (all to within log factor).

3. **Bounded Difference Inequality, Stability and Generalization :**
Recall that a function $G : \mathcal{X}^m \mapsto \mathbb{R}$ is said to satisfy the bounded difference inequality if for all $i \in [m]$ and all $x_1, \ldots, x_m, x_i' \in \mathcal{X}$,

$$|G(x_1, \ldots, x_i, \ldots, x_m) - G(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)| \leq c$$

for some $c \geq 0$. In this case the McDiarmid's inequality gave us that for any $\delta > 0$, with probability at least $1 - \delta$,

$$G(x_1, \ldots, x_m) \leq \mathbb{E}\left[G(x_1, \ldots, x_m)\right] + \sqrt{\frac{(mc)^2 \log(1/\delta)}{m}}$$

The bounded difference property turns out to be quiet useful to analyze learning algorithms directly (instead of looking at the uniform deviation over function class).

A proper learning algorithm is $A : \bigcup_{m=1}^{\infty} \mathcal{X}^m \mapsto \mathcal{F}$ is said to be a uniformly β stable is for all $i \in [m]$, and any $x_1, \ldots, x_m, x_i' \in \mathcal{X}$,

$$\sup_x |A(x_1, \ldots, x_i, \ldots, x_m)(x) - A(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)(x)| \leq \beta$$

Assuming functions in $\mathcal{F}$ are bounded by 1 we shall prove that the learning algorithm generalizes well (expected loss is close to empirical loss of the algorithm). Specifically we shall prove that for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(A(S)) \leq \widehat{R}(A(S)) + \beta + 2(m\beta + 1)\sqrt{\frac{2\log(1/\delta)}{m}}$$

where $R(A(S)) = \mathbb{E}_x [A(S)(x)]$ and $\widehat{R}(A(S)) = \frac{1}{m}\sum_{i=1}^{m} A(S)(x_i)$.

(a) First show that $\mathbb{E}_S \left[R(A(S)) - \widehat{R}(A(S)) \right] \leq \beta$.
Hint : Use renaming of variables to first show that for any $i \in [m]$,

$$\mathbb{E}_S [R(A(S))] = \mathbb{E}_{S, x_i'} [A(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)(x_i)]$$

(b) Show that the function $G(S) = R(A(S)) - \widehat{R}(A(S))$ satisfies bounded difference property with $c \leq 2\beta + \frac{2}{m}$. Conclude the required statement using McDiarmid's inequality.

(c) Consider the stochastic convex optimization problem where sample $z = (x, y)$ where y is real valued and x's are from the unit ball in some Hilbert space and hypothesis is weight vectors w from the same Hilbert space with objective

$$r(w, (x, y)) = |\langle w, x \rangle - y| + \lambda \|w\|^2$$

Show that the ERM algorithm is stable for this problem and thus provide a bound for this algorithm.

4. L_1 **Neural Network :**
A k-layer 1-norm neural network is given by function class $\mathcal{F}_k$ which is in turn defined recursively as follows.

$$\mathcal{F}_1 = \left\{ x \mapsto \sum_{j=1}^{d} w_j^1 x_j \;\middle|\; \|w^1\|_1 \leq B_1 \right\}$$

and further for each $2 \leq i \leq k$,

$$\mathcal{F}_i = \left\{ x \mapsto \sum_{j=1}^{d_i} w_j^1 \sigma(f_j(x)) \;\middle|\; \forall j \in [d_i],\ f_j \in \mathcal{F}_{i-1},\ \|w^i\|_1 \leq B_1 \right\}$$

where d_i is the number of nodes in the ith layer of the network. Function $\sigma : \mathbb{R} \mapsto [-1, 1]$ is called the squash function and is generally a smooth monotonic non-decreasing function (typical example is the tanh function). Assume that input space $\mathcal{X} = [0, 1]^d$ and that σ is L-Lipschitz. Prove that

$$\widehat{\mathcal{R}}_S(\mathcal{F}_k) \leq \left(\prod_{i=1}^{k} 2B_i \right) L^{k-1} \sqrt{2T \log d}$$

Notice that the above bound the d_i's don't appear in the bound indicating the number of nodes in intermediate layers don't affect the upper bound on Rademacher complexity.

Hint : prove bound on Rademacher complexity of $\mathcal{F}_i$ recursively in terms of rademacher complexity of $\mathcal{F}_{i-1}$.