# Computational and Statistical Learning Theory

## Problem set 3

Due: Monday, November 7th

Please send your solutions to learning-submissions@ttic.edu

**Notation:**

- Input space: $\mathcal{X}$
- Label space: $\mathcal{Y} = \{\pm 1\}$
- Sample: $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X}$
- Hypothesis Class: $\mathcal{H}$
- Risk: $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}_{h(x)\neq y}\right]$
- Empirical Risk: $L_S(h) = \frac{1}{m}\sum_{(x,y)\in S}\mathbf{1}_{h(x)\neq y}$

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

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

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

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

where $d_i$ is the number of nodes in the $i$th 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 $\sigma$ is $L$-Lipschitz. Prove that

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

where $T = 1/|\mathcal{S}| = 1/m$. 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 $\mathcal{F}_{i-1}$.

2. **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.

3. **Boosting and Sparse Linear Predictors**
Suppose there exists $\phi(x) \in \mathbb{R}^d, w^* \in \mathbb{R}^d$ and for all $x, y \sim \mathcal{D}$, we have $y\langle\phi(x), w^*\rangle \geq 1$, and $\|w^*\|_0 \leq s \ll d$.

   (a) Describe a possibly intractable learning rule $\mathcal{A}$ that learns $\mathcal{H}$:

$$\mathcal{H} = \{h = \text{sign}(\langle\phi(x), w\rangle) | w \in \mathbb{R}^d, \|w\|_0 \leq s\}.$$

   How many samples are needed to ensure with probability at least $1-\delta$, $L_{\mathcal{D}}(\mathcal{A}(S)) \leq \epsilon$ based on sample $S$ ?

   (b) Assume $w^* \in [-B, B]^d$ (i.e. each coordinate of $w^*$ has magnitude at most $B$), show that boosting can find a predictor $w$ with $L_{\mathcal{D}}(w) \leq \epsilon$ with probability at least $1-\delta$, what is the sample complexity ?

   (c) Show that without the upper bound on $B$, we cannot ensure boosting work, do this by giving an explicit construction showing boosting fail.

   (d) (Bonus) Show that without the upper bound on $B$, minimizing $\ell_1$ norm also fail, i.e.

$$\hat{w} = \arg\min_{w \in \mathbb{R}^d} \|w\|_1, \text{ s.t. } L_S(w) = 0.$$

   (e) (Optional) If $\langle w, x\rangle \leq B_1$ and $\text{Cov}[x_{\text{supp}(w^*)}] \succeq \lambda I$, then boosting works, where $x_{\text{supp}(w^*)}$ denotes the $s$-dimensional vector that extract the coordinates of $x$ in support of $w^*$.

4. **Dudley Vs Pollards' Bounds (Optional)**
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 (suppose $\mathcal{F} \subset [0,1]^{\mathcal{X}}$) :

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

$$\widehat{\mathcal{R}}_S(\mathcal{F}) \leq 12 \int_0^1 \sqrt{\frac{\mathcal{N}_2(\mathcal{F}, \alpha, m)}{m}} d\alpha \quad \text{(Original Dudley)}$$

$$\widehat{\mathcal{R}}_S(\mathcal{F}) \leq \inf_{\alpha_0 \geq 0} \left\{ 4\alpha_0 + 12 \int_{\alpha_0}^1 \sqrt{\frac{\mathcal{N}_2(\mathcal{F}, \alpha, m)}{m}} d\alpha \right\} \quad \text{(Refined Dudley)}$$

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.