COMS 4995-1 S20 Homework 3 (due April 1, 2020)

Instructions

Submit your write-up on Gradescope as a neatly typeset PDF document by 11:00 PM of the due date. Please use TEX, LATEX, or a similar system.

On Gradescope, be sure to select the pages containing your answer for each problem. More details can be found on the Gradescope Student Workflow help page.

(If you don't select pages containing your answer to a problem, you'll receive a zero for that problem.)

Also, please make sure that your "Student ID #" on Gradescope is set to your UNI, using only lowercase letters and numbers (e.g., abc1234).

Finally, please make sure **your name and your UNI** appear prominently on the first page of your write-up.

Problem 1

In this problem, you will derive bounds on l_n^2 covering numbers for linear functions of bounded norm. Recall the *normalized* versions of the l^p norms (for $p \geq 1$):

$$\|\boldsymbol{v}\|_{p,n} := \frac{1}{n^{1/p}}\|\boldsymbol{v}\|_p = \left(\frac{1}{n}\sum_{i=1}^n |v_i|^p\right)^{1/p}.$$

(a) Consider the following greedy algorithm.

- **Input**: vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d \in \mathbb{R}^n$ satisfying $\|\boldsymbol{v}_j\|_{2,n} \leq 1$ for all $j = 1, \ldots, d$; vector $\boldsymbol{u} \in \mathbb{R}^n$ that is promised to satisfy $\boldsymbol{u} = \sum_{j=1}^d \alpha_j \boldsymbol{v}_j$ for some non-negative $\alpha_1, \ldots, \alpha_d$ such that $\sum_{j=1}^d \alpha_j \leq 1$; positive integer k.
- Set $\boldsymbol{v}_0 := \boldsymbol{0} \in \mathbb{R}^n$.
- Set $\boldsymbol{w}_0 := \boldsymbol{0} \in \mathbb{R}^n$, and $k_j := 0$ for all $j = 1, \ldots, d$.
- For $t = 1, \ldots, k$:
  - Pick
  $$j_t \in \operatorname*{arg\,min}_{j \in \{0,\ldots,d\}} \left\| \left(1 - \frac{1}{t}\right)\boldsymbol{w}_{t-1} + \frac{1}{t}\boldsymbol{v}_j - \boldsymbol{u} \right\|_{2,n}^2.$$
  - If $j_t \geq 1$, then set $k_{j_t} := k_{j_t} + 1$.
  - Update:
  $$\boldsymbol{w}_t := \left(1 - \frac{1}{t}\right)\boldsymbol{w}_{t-1} + \frac{1}{t}\boldsymbol{v}_{j_t}.$$
- **Output**: $k_1, \ldots, k_d$

Prove that the algorithm outputs $k_1, \ldots, k_d$ that satisfy $\sum_{j=1}^d k_j \leq k$ and

$$\left\| \boldsymbol{u} - \frac{1}{k}\sum_{j=1}^d k_j \boldsymbol{v}_j \right\|_{2,n} \leq \frac{1}{\sqrt{k}}.$$

Note: The main implication of this algorithm is essentially the same as something you already proved in HW0. This part just asks you to provide an alternative (constructive) proof.

(b) Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be given, and assume, for some $p \in [2, +\infty]$, that $\|\boldsymbol{x}_i\|_p \leq 1$ for all $i = 1, \ldots, n$. Define $q := 1/(1 - 1/p)$ (so that $1/p + 1/q = 1$) and

$$A := \left\{ (\boldsymbol{x}_1^\intercal \boldsymbol{w}, \ldots, \boldsymbol{x}_n^\intercal \boldsymbol{w}) : \boldsymbol{w} \in B_q^d \right\},$$

where $B_q^d := \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_q \leq 1\}$. Also define $\boldsymbol{X} := [\boldsymbol{x}_1|\cdots|\boldsymbol{x}_n]^\intercal \in \mathbb{R}^{n \times d}$, and $\boldsymbol{v}_j := \boldsymbol{X}\boldsymbol{e}_j / \|\boldsymbol{X}\boldsymbol{e}_j\|_{p,n}$ for $j = 1, \ldots, d$. Prove that every $\boldsymbol{a} \in A$ can be written as $\boldsymbol{a} = \sum_{j=1}^d \alpha_j \boldsymbol{v}_j$ for some real numbers $\alpha_1, \ldots, \alpha_d$ such that $\sum_{j=1}^d |\alpha_j| \leq 1$.
Hint: You are likely to need Hölder's inequality, $\boldsymbol{u}^\intercal \boldsymbol{v} \leq \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q$.

(c) (Continuing from above.) Prove that $\mathcal{N}(\epsilon, A, l_n^2) \leq (2d+1)^{\lceil 1/\epsilon^2 \rceil}$ for all $\epsilon > 0$.
Hint: Use the result implied by the algorithm from Part (a). It will also be important that $p \geq 2$.

(d) (Continuing from above.) Explain why the d in the bound from Part (c) can be replaced by n when $d \geq n$.

(e) (Continuing from above.) Use the results of the previous parts along with the Discretization Lemma, Dudley's Entropy Integral, or something in-between, to derive a bound on $\mathrm{Rad}_n(A)$.

Problem 2

In this problem, you will prove a *sparsification* result for linear predictors. The idea is similar to the proof technique used in the Schapire-Freund-Bartlett-Lee margin bound.

Let $\ell\colon \mathbb{R} \to \mathbb{R}$ denote a loss function with uniformly bounded second derivatives: $\sup_{z\in\mathbb{R}} \ell''(z) \le \beta$ for some $\beta > 0$. (An example of such a loss function is the *logistic loss* $\ell(z) = \ln(1 + e^{-z})$, which is used in MLE for logistic regression.) Define the risk (i.e., expected loss) of a linear predictor $\boldsymbol{w} \in \mathbb{R}^d$ by

$$\mathcal{R}(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{X},Y)\sim P}[\ell(Y\boldsymbol{X}^\mathsf{T}\boldsymbol{w})]$$

where P is a probability distribution over $[-1,+1]^d \times \{\pm 1\}$.

Give a randomized algorithm (along with proof of correctness) for the following problem.

- **Input**: Minimizer $\boldsymbol{w}^\star \in \mathbb{R}^d$ of $\mathcal{R}$ over $\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_1 \le \|\boldsymbol{w}^\star\|_1\}$; $\epsilon > 0$.
- **Output**: $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ with at most

  $$\left\lceil \frac{\beta\|\boldsymbol{w}^\star\|_1^2}{\epsilon} \right\rceil$$

  non-zero entries such that

  $$\mathcal{R}(\hat{\boldsymbol{w}}) \le \mathcal{R}(\boldsymbol{w}^\star) + \epsilon.$$

The algorithm should produce an output with the desired properties with probability at least $1/2$.

Hint: What is an unbiased estimator of $\boldsymbol{w}^\star$ that has exactly one non-zero entry? Consider picking the coordinate of the non-zero entry according to the probability distribution Q over $\{1,\dots,d\}$ that has $Q(j) \propto |w_j^*|$.

Problem 3

In this problem, you wil prove that the "exponential weights" variant of Boost-by-Majority (BBM) finds a large (well, large-ish) margin hypothesis under the Weak Learning Assumption with parameter $\theta \in (0, 1)$.

Let $h_1, \ldots, h_T \colon \mathcal{X} \to \{\pm 1\}$ be the hypotheses returned by the Weak Learner, and let $g_T \colon \mathcal{X} \to \mathbb{R}$ be the function

$$g_T(x) = \sum_{t=1}^{T} h_t(x), \quad x \in \mathcal{X}.$$

The goal is to guarantee, for some absolute constant $C \geq 1$, that

$$\frac{1}{T} y_i g_T(x_i) \geq \theta / C$$

for all training examples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{\pm 1\}$.[1] The scaling factor $1/T$ is needed so that g_T is a convex combination of $h_1, \ldots, h_T$ (as required by the margin-based generalization bound).

(a) In the "exponential weights" variant of BBM, we use $\phi_T(s_i) := e^{-\eta s_i}$, where $\eta := \frac{1}{2} \ln \frac{1+\theta}{1-\theta}$ is the tuning parameter that minimizes $\phi_0(0)$. Prove that

$$\frac{1}{T} \ln \left(\sum_{i=1}^{n} e^{-\eta y_i g_T(x_i)} \right) \leq \frac{1}{2} \ln(1 - \theta^2) + \frac{\ln(n)}{T}.$$

(b) Using the result from Part (a), prove that

$$\frac{1}{T} y_i g_T(x_i) \geq \frac{\ln \frac{1}{1-\theta^2}}{\ln \frac{1+\theta}{1-\theta}} - \frac{\ln(n)}{\eta T} \quad \text{for all } i = 1, \ldots, n.$$

(c) For what $C \geq 1$ is the stated goal from above achieved as $T \to \infty$?
Hint: Use a Taylor expansion.

(d) (Optional; up to 5 points extra credit.) Explain (and prove) how to modify the BBM algorithm to achieve the stated goal with C arbitrarily close to one (again as $T \to \infty$).

[1] This is what we mean by "large-ish margin". Think of it as a constant factor approximation to the margin of θ that is guaranteed by the Von Neumann Min-Max Theorem.

Problem 4

In this problem, you will prove a margin bound for weighted majority vote hypotheses that achieve a large margin on *all* training examples. The bound will have a better dependence on the sample size n than what we had proved in lecture (which applies to hypotheses that might have fail to achieve a large margin on an arbitrary constant fraction of training examples).

Modify the proof of the Schapire-Freund-Bartlett-Lee margin bound to prove the following claim.

Let $\mathcal{H}$ denote a finite hypothesis class comprised of $\{\pm 1\}$-valued functions on an input space $\mathcal{X}$, and let $\mathcal{F}$ be the convex hull of $\mathcal{H}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an iid sample from P over $\mathcal{X} \times \{\pm 1\}$, and fix any $\theta \in (0, 1)$. With probability at least $1 - \delta$, for any $f \in \mathcal{F}$ such that

$$\Pr_{(X,Y)\sim P_n}[Yf(X) \leq \theta] = 0,$$

we have

$$\Pr_{(X,Y)\sim P}[Yf(X) \leq 0] \leq O\left(\frac{\frac{\log n}{\theta^2} \cdot \log|\mathcal{H}| + \log(1/\delta)}{n}\right).$$

Hint: Bernstein's inequality may be useful.

Problem 5

In this problem, you will prove margin bounds using Rademacher complexity.

Let $\mathcal{H}$ denote a (possibly infinite) hypothesis class comprised of $\{\pm 1\}$-valued functions on an input space $\mathcal{X}$, and let $\mathcal{F}$ be the convex hull of $\mathcal{H}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid random variables with distribution P over $\mathcal{X} \times \{\pm 1\}$, and let P_n denote the empirical distribution on these examples.

(a) Prove that for any $\delta \in (0,1)$ and any $\theta \in (0,1)$, with probability at least $1-\delta$,

$$\Pr_{(X,Y)\sim P}[Yf(X) \le 0] \le \Pr_{(X,Y)\sim P_n}[Yf(X) \le \theta] + O\left(\frac{1}{\theta} \cdot \mathrm{Rad}_n(\mathcal{H}(X_{1:n})) + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

for all $f \in \mathcal{F}$.

(b) Assume that $\mathcal{H} = \{h_1, \ldots, h_N\}$ is finite. Define a feature map $\boldsymbol{\varphi}\colon \mathcal{X} \to \{\pm 1/\sqrt{N}\}^N$ by

$$\boldsymbol{\varphi}(x) := \frac{1}{\sqrt{N}}(h_1(x), \ldots, h_N(x)) \quad \forall x \in \mathcal{X}.$$

Prove that for any $\delta \in (0,1)$ and any $\theta \in (0,1)$, with probability at least $1-\delta$,

$$\Pr_{(X,Y)\sim P}[Y\boldsymbol{\varphi}(X)^\mathsf{T}\boldsymbol{w} \le 0] \le \Pr_{(X,Y)\sim P_n}[Y\boldsymbol{\varphi}(X)^\mathsf{T}\boldsymbol{w} \le \theta] + O\left(\sqrt{\frac{1}{\theta^2 n} + \frac{\log(1/\delta)}{n}}\right)$$

for all $\boldsymbol{w} \in \mathbb{R}^N$ with $\|\boldsymbol{w}\|_2 \le 1$.

(c) (Optional; up to 5 points extra credit.) Determine and prove an analogous l^p norm version of the claim from Part (b) for arbitrary $p \ge 2$. (What should the feature map be? What should the norm bound on the weight vectors be? What is the final bound?)

Hints: For Part (a), you will need to combine a few ideas from the course so far.

- Use McDiarmid's inequality (as in HW1 Problem 2(a)).
- Upper-bound the zero-one loss $\ell_{0/1}(s) := \mathbb{1}_{\{s \le 0\}}$ by the following $1/\theta$-Lipschitz loss:

$$\tilde{\ell}^\theta(s) := \begin{cases} 1 & \text{if } s \le 0; \\ 1 - s/\theta & \text{if } 0 \le s \le \theta; \\ 0 & \text{if } s \ge \theta. \end{cases}$$

- Upper-bound $\tilde{\ell}^\theta(s)$ by the "margin loss" $\ell^\theta(s) := \mathbb{1}_{\{s \le \theta\}}$.
- Use properties of Rad_n such as the Contraction Lemma.

Parts (b) and (c) will need similar ideas, along with techniques for dealing with bounded linear functions. For these parts, your proofs can just give the main ideas that are different from what is needed in Part (a).