PSET 3

Due on **Tuesday Oct 29, 2024 at 11:59 pm ET**

COMPSCI 2243: Algorithms for Data Science, Fall 2024 (Sitan Chen) Version: **1.0**

Submitted by **<name>**

This problem set will cover concepts from the third unit on supervised learning. The questions are meant to be challenging, so as with previous psets, do not feel discouraged if you get stuck and are unable to solve some of them. If you find that you are running low on time to finish all the problems, our recommendation is to try to aim for breadth rather than depth – e.g., it is better to complete a few parts of each of the questions, than to completely solve one of the questions and skip the others.

1 (50 PTS.) LEARNING BOUNDED POLYNOMIALS OVER THE HYPERCUBE

In class, we saw examples where Boolean functions in n variables that are well-approximated by degree-d polynomials can be PAC-learned over the uniform distribution over $\{-1,+1\}^n$ using roughly $n^{O(d)}$ time and samples. In this problem, we will explore the same setting but where the function is *exactly given* by a degree-d polynomial. This might seem like a strictly easier setting, but you will encounter some surprising behavior that, rather remarkably, was only discovered two years ago!

Setup: Let $\epsilon, \delta > 0$, and let $f\colon \{-1,+1\}^n \to \mathbb{R}$ be a function over the Boolean hypercube with Fourier expansion

$$f(x) = \sum_{S \subseteq [n]} \hat{f}[S] \cdot x_S \,. \tag{1}$$

In this question, we will assume that f has degree at most d, that is, $\hat{f}[S] = 0$ for all S satisfying $|S| > d$. Furthermore, we will assume that $|f(x)| \leqslant 1$ for all $x \in \{-1,+1\}^n$. We refer to such functions as ***bounded polynomials***.

Suppose we are given n independent samples $(x_1, y_1), \ldots, (x_N, y_N)$ for $x_i \in \{-1,+1\}^n$ and $y_i = f(x_i)$. In this problem, we will explore how large N must be for PAC learning of bounded polynomials to be possible. Note that one can easily solve this problem with polynomial regression using $n^{O(d)}$ samples, and *a priori*, one might suspect that this is optimal. Here you will show that, remarkably, a refinement of polynomial regression allows one to learn using only $O_d(\log n)$ many samples![1]

You may use the following deep result from functional analysis, the ***Bohnenblust-Hille inequality***, without proof:

Theorem 1.1. *For any function $f : \{-1,+1\}^n \to \mathbb{R}$ of degree at most d, if $r = \frac{2d}{d+1}$, then*

$$\Big(\sum_{S \subseteq [n]} |\hat{f}[S]|^r \Big)^{1/r} \leqslant O_d(\|f\|_\infty) \,, \tag{2}$$

where $\|f\|_\infty \triangleq \max_{x \in \{-1,+1\}^n} |f(x)|$.

1.A. (4 PTS.) Prove Eq. (2) for $r = 2$ (this is not used later but provides good intuition for what Theorem 1.1 is saying).

1.B. (5 PTS.) Let $\beta > 0$ be a parameter to be tuned later. Propose an estimator for $\{\hat{f}[S]\}_{S \subseteq [n]}$ using the samples $(x_1, y_1), \ldots, (x_N, y_N)$. If the estimate for $\hat{f}_S$ is denoted by ϕ_S, give an upper bound on the number of samples N needed to ensure that

$$|\phi_S - \hat{f}[S]| \leqslant \beta \quad \text{for all} \quad S \subseteq [n] \tag{3}$$

with probability at least $1 - \delta$. Denote the event of Eq. (3) by $\mathcal{E}_\beta$.

1.C. (4 PTS.) How small would β have to be, and thus how big would N have to be, for the estimates ϕ_S in Part **1.B.** to satisfy

$$\mathop{\mathbf{E}}_{x \sim \{-1,+1\}^n} [(f(x) - \sum_{S \subseteq [n]} \phi_S x_S)^2] \leqslant \epsilon^2 \,? \tag{4}$$

Next, given a parameter $\eta > 0$, let $\mathcal{S}_{\text{big}}$ denote the set of $S \subseteq [n]$ for which $|\phi_S| \geqslant \eta$, and likewise let $\mathcal{S}_{\text{small}}$ denote the set of $S \subseteq [n]$ for which $|\phi_S| < \eta$.

1.D. (9 PTS.) Use Theorem 1.1 to prove that, conditioned on the event $\mathcal{E}_\beta$ holding,

$$\sum_{S \in \mathcal{S}_{\text{small}}} \hat{f}[S]^2 \leqslant O_d\big((\eta + \beta)^{\frac{2}{d+1}}\big) \,. \tag{5}$$

1.E. (18* PTS.) Use Theorem 1.1 to prove that, conditioned on the event $\mathcal{E}_\beta$ holding,

$$\sum_{S \in \mathcal{S}_{\text{big}}} (\phi_S - \hat{f}[S])^2 \leqslant O_d\big(\beta^2 (\eta - \beta)^{-\frac{2d}{d+1}}\big) \,. \tag{6}$$

[1] Here the notation $f(n) \leqslant O_d(g(n))$ denotes that there is a constant C depending only on d such that $f(n) \leqslant C_d \cdot g(n)$ for n sufficiently large.

1.F. (5 PTS.) By removing some terms from the naive estimator $\sum_{S\subseteq[n]} \phi_S x_S$, propose an alternative estimator f' for f that achieves squared error at most ν, i.e. such that $\mathbf{E}_{x\sim\{-1,+1\}^n}[(f'(x) - f(x))^2] \leqslant \nu$, where

$$\nu = O_d\big(\beta^2(\eta-\beta)^{-\frac{2d}{d+1}} + (\eta+\beta)^{\frac{2}{d+1}}\big)\,. \tag{7}$$

1.G. (5 PTS.) Conclude that there is some $\underline{N} = O_d(\log(n/\delta)/\epsilon^{d+1})$ such that provided $N \geqslant \underline{N}$, there is an algorithm for PAC learning bounded polynomials of degree d to squared error ϵ with probability at least $1-\delta$ using only N samples. Informally explain the intuition for why the Bohnenblust-Hille inequality resulted in such a dramatic savings in the sample complexity dependence on n compared to your answer in Part **1.C.**.

Solution:

1.A.

1.B.

1.C.

1.D.

1.E.

1.F.

1.G.

2 (65 PTS.) PAC LEARNING MLPS WITH VANILLA PCA

In this problem, we will explore a simple approach to obtaining polynomial dependence on $1/\epsilon$ and fixed-parameter tractability for the following special class of one-hidden-layer MLPs:

$$f(x) = \sum_{j=1}^{k} \mathrm{ReLU}(\langle w_j, x\rangle)\,, \tag{8}$$

where $w_1, \ldots, w_k \in \mathbb{R}^d$ are arbitrary, possibly linearly dependent, unit vectors.

Setup: We are given as input $\varepsilon > 0$ and independent samples $(x_1, y_1), \ldots, (x_N, y_N)$ where each $x_i \sim \mathcal{N}(0, I_d)$ and $y_i = f(x_i)$, where f is a fixed function of the form Eq. (8). As we saw in class, polynomial regression yields an algorithm for (improper) PAC learning such function to error ϵ in this setting in $d^{\mathrm{poly}(k/\epsilon)}$ time and samples. In this problem, you will explore a simple algorithm that properly PAC learns such functions in just $(k/\epsilon)^{O(k^2)} + \mathrm{poly}(k, d, 1/\epsilon)$ time and $\mathrm{poly}(k, d, 1/\epsilon)$ samples.

2.A. (5 PTS.) Given the samples $(x_1, y_1), \ldots, (x_N, y_N)$, propose an estimator $\hat{M}$ whose expected value is given by the matrix $M \triangleq \sum_{i=1}^{k} w_i w_i^\top$.

So we don't have to worry about matrix concentration in this homework, in the rest of the parts you may assume that your estimator $\hat{M}$ from Part **2.A.** satisfies

$$\|\hat{M} - M\|_{\mathsf{op}} \leqslant \eta \tag{9}$$

for some sufficiently small $\eta > 0$, which it turns out holds with high probability with $N = \mathrm{poly}(kd/\eta)$ samples.

Let W be the subspace spanned by the top k singular vectors of $\hat{M}$. Let $\Pi^\perp$ denote the projection to the orthogonal complement of W.

2.B. (10 PTS.) For every $j \in [k]$, let $r_j = \Pi^\perp w_j$. Prove that

$$r_j^\top \hat{M} r_j \leqslant \eta \|r_j\|^2\,. \tag{10}$$

(**Hint:** Look at the $(k+1)^{\mathsf{th}}$ singular value of $\hat{M}$.)

2.C. (10 PTS.) Prove that for all $j \in [k]$,

$$r_j^\top M r_j \geqslant \|r_j\|^4 \tag{11}$$

and conclude that $\|r_j\|^2 \leqslant 2\eta$.

2.D. (10 PTS.) Prove that for any vectors $u, v \in \mathbb{R}^d$,

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}(0, I_d)} [(\mathrm{ReLU}(\langle u, x\rangle) - \mathrm{ReLU}(\langle v, x\rangle))^2] \leqslant \|u - v\|^2\,. \tag{12}$$

2.E. (10 PTS.) Use Parts **2.C.** and **2.D.** to conclude that there exist vectors $\tilde{w}_1, \ldots, \tilde{w}_k$ in W, each of norm at most 1, such that

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}(0, I_d)} \left[\left(f(x) - \sum_{j=1}^{k} \mathrm{ReLU}(\langle \tilde{w}_j, x\rangle) \right)^2 \right] \leqslant \eta \cdot \mathrm{poly}(k)\,. \tag{13}$$

Let $(B_W)^k$ be the space of k-tuples of vectors in W each of norm at most 1. Let $\mathcal{N}$ be an *η-net* over this space, that is, a discrete set of points in $(B_W)^k$ such that for any $(\tilde{w}_1, \ldots, \tilde{w}_k) \in (B_W)^k$, there exists $(w'_1, \ldots, w'_k) \in \mathcal{N}$ such that $\|\tilde{w}_j - w'_j\|_2 \leqslant \eta$ for all $j \in [k]$. You may use without proof the fact that there is a (computationally efficient) construction of $\mathcal{N}$ of size at most $(1/\eta)^{O(k^2)}$.

2.F. (5 PTS.) Explain informally why one should expect $\mathcal{N}$ to have size exponential in k^2.

2.G. (5 PTS.) Prove that there exists $(w'_1, \ldots, w'_k) \in \mathcal{N}$ such that

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}(0, I_d)} \left[\left(f(x) - \sum_{j=1}^{k} \mathrm{ReLU}(\langle w'_j, x\rangle) \right)^2 \right] \leqslant \eta \cdot \mathrm{poly}(k)\,. \tag{14}$$

2.H. (5 PTS.) Explain informally how to combine the preceding steps to obtain an algorithm with sample complexity $\mathrm{poly}(k, d, 1/\epsilon)$ and runtime $(k/\epsilon)^{O(k^2)} + \mathrm{poly}(k, d, 1/\epsilon)$ for PAC learning one-hidden-layer MLPs of the form Eq. (8).

2.I. (5 PTS.) Explain informally why the algorithm in Part **2.H.** does not contradict the CSQ lower bound of $d^{\Omega(k)}$ from class.

Solution:

2.A.

2.B.
2.C.
2.D.
2.E.
2.F.
2.G.
2.H.
2.I.