Theoretical Foundations for Deep Learning: Problem Set # 2

Instructor: Ankur Moitra

Due: ~~May 7th~~ May 10th, due to student holiday

You can work with other students, but you must write-up your solutions by yourself and indicate at the top who you worked with!

Please submit by May 7th, at 11:59pm, via gradescope.

Problem 1 30 points

In this problem we will take a compression approach to proving generalization bounds. In class, we asserted that the VC-dimension of the class of halfspaces in d dimensions is $d+1$ and we used this fact to give bounds on the generalization error. Instead, let's consider a compression game that works as follows:

(a) Alice receives a set S of n training examples $(x_1, y_1), \cdots, (x_n, y_n)$ that are drawn i.i.d. from a distribution D. We assume the labels are consistent with some unknown halfspace, i.e. $y = \text{sgn}(\bar{w}^T x)$

(b) Alice selects a set $T \subseteq S$ of k examples to send to Bob in such a way that Bob can reconstruct a halfspaces that correctly labels all the examples in Alice's training set, i.e. Bob can find w and b so that for any $(x, y) \in S$

$$y = \text{sgn}(w^T x + b)$$

Find a compression scheme that works with $k = O(d)$. To make things simpler, you may assume that every set of d of the x's are linearly independent and have some separation δ from each other, and you only need your compression scheme to work when this condition holds. Notice that this problem is subtle because it depends on the choice of *how* Bob will find a hypothesis that works on the subset of data that Alice sent. It's *not* true that every halfspace that fits T will fit S too. *Hint:* You may want to think about the problem in low dimensions, like in $\mathbb{R}^2$ first and consider a halfspace that goes through a subset of the data and use an arbitrarily small perturbation. You may assume that Alice sends the points in a particular order, though there is a scheme that does not use order.

The point is you can use certain types of compression schemes to prove generalization bounds: Suppose your compression scheme is deterministic and has the following

property: If T is Alice's compression of a set S then for any S' with $T \subseteq S' \subseteq S$, we have that Alice's compression for S' is unchanged and is still T. With this property in hand, suppose we form a set of $n+1$ samples S chosen i.i.d. from D and let T be its compression. Now apply a random permutation to S. This doesn't change anything because the draws were i.i.d. Think of the first n points as the training set and call it S'. Think of the last example (x, y) as the test point. The key observation is if the test point is not in T, which you can think of as the set of informative examples, then you shouldn't make a mistake on it because you haven't learned anything new from it. Or to put it another way, the halfspace that Bob finds by taking S' compressing it down to T and finding a halfspace that agrees with the compression necessarily fits S, which includes the test point, too.

Problem 2 40 points

In this problem we will explore the task of learning a halfspace with noise. We will work with the *random classification noise* model where we flip the label of each example independently with probability η. In particular, assume there is a distribution D on $\mathbb{R}^d$ and each sample is drawn according to the following procedure: $x \leftarrow D$ and $y = \text{sgn}(\bar{w}^T x + \bar{b})$ with probability $1 - \eta$ and otherwise $y = -\text{sgn}(\bar{w}^T x + \bar{b})$. You can assume that $\eta < 1/2$ and that $\text{sgn}(0) = 1$.

Consider the *Leaky ReLU* loss which is defined as follows. For a given hypothesis (w, b) and a labeled example (x, y) we incur loss

$$L_\lambda(w, b) = |w^T x + b|(1_{-y(w^T x+b)\geq 0} - \lambda)$$

Notice that for $\lambda = 0$ we recover the ReLU loss. Show that for an appropriate choice of $\lambda > 0$, the following properties hold:

(a) For any (w, b),
$$\mathbb{E}_{(x,y)}[L_\lambda(w, b)] \geq 0$$

(b) Furthermore (assuming that no examples are on the decision boundary $\{x|w^T x + b = 0\}$) equality is achieved if and only if (w, b) gets optimal accuracy, i.e.
$$\mathbb{P}_{(x,y)}[\text{sgn}(w^T x + b) = y] = 1 - \eta = \mathbb{P}_{(x,y)}[\text{sgn}(\bar{w}^T x + \bar{b}) = y]$$

(c) How would you design an algorithm for learning a halfspace with optimal agreement in the random classification noise model? *Hint:* Think about how you would use a new training example to take a step in the direction of the minimum of a convex function. You can take for granted that stochastic gradient descent, with an appropriately chosen step size, will converge. And you do not need to worry about proving bounds on the number of iterations or the generalization error, though such things can be done.

(d) What would go wrong if the noise were not random, but adversarial? Is it still true that a (w, b) that minimizes the Leaky ReLU loss necessarily gets optimal

accuracy? Let's be clear about what we mean by adversarial: There is an arbitrary distribution D on (x, y) pairs. We can still define optimal accuracy over a class of hypotheses. And here you want to show that finding the (w, b) that minimizes the Leaky ReLU loss does not necessarily give you optimal accuracy.

Problem 3 30 points

In this problem, we will prove lower bounds for learning junta functions. A k-junta is a function $f : \{\pm 1\}^n \to \{\pm 1\}$ that only depends on k coordinates. In particular, there is a set $S \subseteq [n]$ with $|S| = k$ and if you change any coordinate outside S the function value does not change. Prove that any statistical query algorithm for learning juntas on the uniform distribution must make at least $n^{\Omega(k)}$ queries or make a query with tolerance $\tau = n^{-\Omega(k)}$. We are interested in the setting where k is much smaller than n, and you can assume $k \leq n^{1-\epsilon}$ for some $\epsilon > 0$. As we did in class, since the functions we are trying to learn are $\{\pm 1\}$ valued and the distribution on examples is known, you may assume the queries are all correlational.