Homework 2
CS229T/STATS231 (Fall 2018–2019)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something’s true (by algebra, convexity, etc.). There’s no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived.

Up to two additional points will be awarded for especially well-organized and elegant solutions.

Due date: Nov. 1st, 11pm, Thursday

1. Improved generalization in low error regimes (15 points)

Recall that the finite-sample generalization error bounds from uniform convergence generally are on the order of $1/\sqrt{n}$, whereas in the well-specified case, the asymptotic result shows that the generalization error is on the order of $1/n$ as $n$ goes to infinity. In this problem, we will show that the generalization error bounds can be improved if the hypothesis class is better specified (in the sense that the expected risk minimizer $h^*$ has a small risk $L(h^*)$.) In particular, if the data distribution is realizable in the sense that $L(h^*) = 0$, we will show the generalization error is on the order of $1/n$ (for finite $n$).

To start out, let’s revisit Hoeffding’s inequality, which was used to prove uniform convergence results. Recall that Hoeffding’s inequality states that if $X_1, \ldots, X_n$ are independent random variables such with $\mu = \mathbb{E}[X_i]$, and $a \leq X_i \leq b$ with probability 1 for each $i$, then

$$
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \epsilon\right] \leq \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right). \tag{1}
$$

Since Hoeffding’s inequality only depends on the upper and lower bounds $a$ and $b$ of $X_i$, it can be very loose when the $X_i$ has low variance. For example, compare (i) $X_i = -1$ or $X_i = +1$, each with probability $\frac{1}{2}$; and (ii) $X_i = 0$ with probability $0.98$ and $X_i = -1$ or $X_i = +1$, each with probability $0.01$. Example (ii) should intuitively enjoy a sharper bound because it has smaller variance. In this problem, we will derive better generalization bounds that depend on variance.

Instead of using Hoeffding’s inequality, we will use Bernstein’s inequality. The setup is the same as in Hoeffding’s inequality, except that we also define $\sigma^2 = \text{Var}[X_i]$. The bound is as follows:

$$
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2(\sigma^2 + (b-a)\epsilon/3)}\right). \tag{2}
$$

(An intuitive comparison between (2) and (1): by Bernstein inequality (2), we have that with probability at least $0.9$, $\sum X_i \leq n\mu + O(\sqrt{n}\sigma + |b-a|)$. With Hoeffding inequality (1), we can only get with probability at least $0.9$, $\sum X_i \leq n\mu + O(\sqrt{n}|b-a|)$. Note that $\sigma \leq |b-a|$. Therefore, Bernstein is as strong as Hoeffding inequality in general, and stronger when $\sigma \ll |b-a|$.)

a. (3 points) (risk concentrates for good predictors) We first consider hypotheses with expected risk that is bounded above by a constant $E$. (The existence of a small upper bound $E$ is what makes the problem almost realizable.) Equipped with Bernstein’s inequality, we prove a concentration bound, relating empirical and expected risk for such hypotheses.

Assume our loss function is bounded as follows: $\ell(y, p) \in [0, 1]$. Suppose that we have a fixed predictor $h : \mathcal{X} \to \mathbb{R}$ that achieves expected risk at most $E$; that is, $L(h) \leq E$, where

$$
L(h) \overset{\text{def}}{=} \mathbb{E}_{(x, y) \sim p} [\ell(y, h(x))].
$$

\[1\]We use the term “realizable” or “almost realizable” to informally mean the situations where $L(h^*)$ is zero or very small respectively.
Recall that we defined the empirical risk as the random variable:

\[
\hat{L}(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(y^{(i)}, h(x^{(i)})).
\]

Show that

\[
P[\hat{L}(h) - L(h) \geq \epsilon] \leq \exp \left( \frac{-n\epsilon^2}{2(E + \epsilon/3)} \right).
\]

**Remark 1:** When \( E = 0 \), the exponent behaves like \( O(-n\epsilon) \), which is much better than the usual \( O(-n\epsilon^2) \) when \( \epsilon \) is small.

**Remark 2:** You will get full credits if you only prove a bound of the form \( P[\hat{L}(h) - L(h) \geq c_1\epsilon] \leq \exp \left( \frac{-n\epsilon^2}{c_2E + c_3\epsilon} \right) \) for some universal constant \( c_1, c_2, c_3 > 0 \).

---

**b. (5 points) (bad predictors look bad)** Next, we will prove a sort of converse to the above. Consider hypotheses with expected risk that is at least some amount \( E' + \epsilon \). (If such an amount is large, these hypotheses are far from “realizing” the minimum expected problem.) We will show that, as \( \epsilon \) increases from zero, it is increasingly unlikely for the empirical risk of such a hypothesis to fall below the risk threshold \( E' \).

Formally, suppose that instead we now have another fixed predictor \( h' \) with expected risk at least \( E' + \epsilon \):

\[
L(h') \geq E' + \epsilon.
\]

Show that it is unlikely that the empirical risk \( \hat{L}(h') \) is less than \( E' \):

\[
P[\hat{L}(h') \leq E'] \leq \exp \left( \frac{-n\epsilon^2}{2(E' + 4\epsilon/3)} \right).
\]

**Remark 1:** You will get full credits if you only prove a bound of the form \( P[\hat{L}(h') \leq E'] \leq \exp \left( \frac{-n\epsilon^2}{c_1E' + c_2\epsilon} \right) \) for some universal constant \( c_1, c_2 > 0 \).

---

**c. (5 points) (bounding excess risk)** We now bound the excess risk in terms of the smallest valid expected risk bound, \( E = L(h^*) \).

Suppose that our hypothesis class \( \mathcal{H} \) is finite with \( |\mathcal{H}| \) elements. Use the preceding parts to conclude that the empirical risk minimizer \( \hat{h} \) achieves:

\[
P[L(\hat{h}) - L(h^*) \geq 2\epsilon] \leq 2|\mathcal{H}| \exp \left( \frac{-n\epsilon^2}{2(E + 7\epsilon/3)} \right). \tag{3}
\]

**Remark 1:** You will get full credits if you only prove a bound of the form \( P[L(\hat{h}) - L(h^*) \geq c_1\epsilon] \leq c_2|\mathcal{H}| \exp \left( \frac{-n\epsilon^2}{c_3E + c_4\epsilon} \right) \) for some universal constant \( c_1, c_2, c_3 > 0 \).

---

**d. (2 points) (comparison with Hoeffding)** In the previous part, we proved a generalized excess risk bound with the dependencies that we had originally desired. The bound applies beyond the realizable setting, instead depending on the “extent of realizability” \( L(h^*) \).

Compare this bound with the usual bound one obtains with Hoeffding’s inequality, that is,

\[
P[L(\hat{h}) - L(h^*) \geq 2\epsilon] \leq 2|\mathcal{H}| \exp \left( -2n\epsilon^2 \right).
\]
Suppose $\epsilon \leq 0.05$, discuss when the bound above is worse than bound (3). Concretely, find a threshold $\Delta$ (that may depend on $n$ and $\epsilon$) such that when $E \leq \Delta$, then equation (3) is stronger than the bound above. (Your answer for $\Delta$ only has to be accurate up to a universal multiplicative constant factor.)

2. Rademacher complexity on discrete distributions (10 points)

The main point of the question is to show that the Rademacher complexity not only depends on the family of functions but also on the underlying distribution $P$ from which $z_1, \ldots, z_n$ is sampled from. It’s in general a challenging research question to utilize the special properties of the distribution.\footnote{As mentioned in the class, such dependency is implicit in the usual definition.} In this problem, you are asked to bound the Rademacher complexity for a toy case where the distribution $P$ is discrete (part a and d). We will also ask you to build up the tools towards proving the bounds (part b and c) and their implications (part e and f).

Let $P$ be an arbitrary distribution (on $\mathbb{R}^d$) with support size $k$, that is, there exists a set of $k$ distinct points $\{v_1, \ldots, v_k\} \subset \mathbb{R}^d$ on which $P$ is supported. In other words, we have $P(z = v_i) = p_i$, where $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$.

Let $F = \{f : \mathbb{R}^d \rightarrow \{\pm 1\}\}$ be all possible functions with binary outputs, and $z_1, \ldots, z_n$ are drawn i.i.d. from $P$. Note that here we don’t assume any additional properties on $F$.

a. (2 points) (point mass) Suppose that $k = 1$, in which case $P$ is a point mass at some point $v$, i.e. $P(z = v) = 1$. Show that

$$R_n(F) = \mathbb{E} \left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right] \leq \frac{1}{\sqrt{n}}.$$  \hspace{1cm} (4)

b. (3 points) (expected max of sub-gaussian variables) In this part, we develop a general tool that would be useful for part (c) and (d) and many other settings.

Let $X_1, \ldots, X_m$ be sub-Gaussian variables with mean zero and variance proxy $\sigma^2$ (not necessarily independent), show that

$$\mathbb{E} \left[ \max_{1 \leq i \leq m} X_i \right] \leq \sqrt{2 \sigma^2 \log m}. \hspace{1cm} (5)$$

(Hint: one possible way to prove it involves moment generating functions.)

c. (2 points) (Massart’s finite lemma) In this part, we deviate from our general goal (discrete distribution, but no constraint on $F$), and bound the Rademacher complexity of a finite hypothesis class for any distribution $P$. The bound we develop in this part will be useful for our goal though.

Suppose a hypothesis class $G$ (which consists of functions $g : \mathbb{R}^d \rightarrow \{\pm 1\}$) is finite, and let $|G|$ denote its cardinality. Show that there exists a universal constant $C > 0$ such that for any distribution $P$,

$$R_n(G) = \mathbb{E} \left[ \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] \leq C \sqrt{\frac{\log |G|}{n}},$$  \hspace{1cm} (6)

where the expectation is over $z_1, \ldots, z_n$ drawn i.i.d. from $P$ and uniform binary random variable $\sigma_i$’s.
d. (2 points) (general discrete distributions) We now come back to our original setup of discrete distributions. Suppose now \( k > 1 \), show that

\[
R_n(F) = \mathbb{E} \left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right] \leq C \sqrt{\frac{k}{n}}
\]

for some universal constant \( C > 0 \).

e. (1 point) (generalization error bound) Consider the standard classification setting where the label \( y \in \{-1, 1\} \) and let the loss function be the 0-1 loss \( \ell((x, y), h) = 1 \) if \( h(x) \neq y \) and 0 if \( h(x) = y \). Let \( \mathcal{H} = \{h : \mathbb{R}^d \rightarrow \{\pm 1\}\} \) be the hypothesis class, which contains all the binary-output functions.

Let the distribution of \( x \) be an arbitrary distribution with support size \( k \) (and we don’t assume anything on the label distribution conditioned on \( x \)).

Recall that we use \( \hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h) \) to denote the empirical risk and \( L(h) = \mathbb{E}[\ell((x, y), h)] \) to denote the expected risk. Let \( \hat{h} \) be the empirical risk minimizer, and \( h^* \) be the expected risk minimizer. Show that for \( \delta \in (0, 1/3) \), there exists a universal constant \( C > 0 \) such that with probability at least \( 1 - \delta \) over the choices of \( n \) training data points \((x_1, y_1), \ldots, (x_n, y_n)\),

\[
L(\hat{h}) - L(h^*) \leq C \left( \sqrt{\frac{k}{n}} + \sqrt{\frac{\log 1/\delta}{n}} \right).
\]

(Interpretation of the bound: loosely speaking, it says that if we have much more than \( k \) examples, then we should have non-trivial uniform convergence regardless how large the hypothesis class is (i.e., the RHS is much less than 1). There is another argument that can produce the result in a more intuitive way: with much more than \( k \) examples, with a good chance, we see most of the possible data points at least once in the training set. In this case, the ERM can just memorize all the training data and generalize to the test data.)

3. Complexity of hypothesis classes (10 points)

Generalization bounds for the empirical risk minimizer (ERM) depend on the complexity of the hypothesis class that the ERM is defined over. Recall that there are several ways to measure the complexity of \( \mathcal{H} \). In this problem, we will compute the VC dimensions or Rademacher complexity of certain hypothesis classes, in order to develop an intuition for the difficulty of learning these hypothesis classes.

a. (3 points) (two functions) Recall that the Rademacher complexity of a class of functions \( \mathcal{F} \) is defined as

\[
R_n(\mathcal{F}) = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right],
\]

where \( Z_1, \ldots, Z_n \) are drawn i.i.d. from some distribution \( p^* \) and \( \sigma_1, \ldots, \sigma_n \) are Rademacher variables drawn i.i.d. from \( \{-1, 1\} \) with equal probability of \( +1 \) and \( -1 \).

Let \( f : \mathcal{X} \rightarrow \mathbb{R} \) be a function, and let \( \mathcal{F} := \{-f, f\} \) be a function class containing only two functions. Upper bound \( R_n(\mathcal{F}) \) using a function of \( n \) and \( \mathbb{E}[f(X)^2] \).
b. (4 points) (sparse features, dense weights) In applications such as natural language processing, we often have sparse feature vectors. Suppose that \( x \in \{0, 1\}^d \) has only \( k \) non-zero entries. For example, in document classification, one feature might be “\( x_{17} = 1 \) iff the document contains the word cat.”

Define the class of linear functions whose coefficients have bounded \( L_\infty \) norm:
\[
\mathcal{F} = \{ x \mapsto w \cdot x : \|w\|_\infty \leq B \}.
\]

Compute an upper bound on the Rademacher complexity \( R_n(\mathcal{F}) \) (as above, each \( Z_i \sim p^* \) where the domain of \( p^* \) is \( \{x \in \{0, 1\}^d : x \) has only \( k \) non-zero entries\}). Express your answer as a function of \( B, k, d, n \). Note that this allows us to effectively control the complexity of learning using \( L_\infty \) regularization.

c. (3 points) (sparse weights, dense features) We now consider the “dual” regime of part (b), where now we assume that the weights are sparse and the features are in general dense. Suppose that \( Z_1, \ldots, Z_n \) are drawn i.i.d. from \( p^* \) where the domain of \( p^* \) is \( \{z \in \mathbb{R}^d : \|z\|_\infty \leq B \} \).

Define the class of \( s \)-sparse linear functions as
\[
\mathcal{F} = \{ x \mapsto w \cdot x : \|w\|_\infty \leq 1, \ w \ has \ at \ most \ s \ non-zero \ entries \}.
\]

Compute an upper bound on the Rademacher complexity \( R_n(\mathcal{F}) \). Show that for some universal constant \( c > 0 \),
\[
R_n(\mathcal{F}) \leq cBs\sqrt{\frac{\log 2d}{n}}.
\] (9)

Note that you are allowed to use any results you have proved in this homework, including bullets in question 2.

d. (4 bonus points) (continuous functions with bounded local maxima) (The solution of this question may require the material we will cover in Week 5. Therefore, we suggest you to work on it after the second lecture of Week 5.)

Let \( \mathcal{F} \) be the class of all continuous functions \( f : [0, 1] \to [0, 1] \) with at most \( k \) local maxima. Prove that the Rademacher complexity of \( \mathcal{F} \) is at most \( O(\sqrt{\frac{k \log n}{n}}) \).

4. Miscellaneous concentration (5 points)

Let us apply concentration inequalities to non-learning problems.

a. (3 points) (diameter of a graph)

Let \( V \) be a set of \( n \) vertices. For each pair of vertices \( i \neq j \), let its edge cost \( C_{ij} \) be drawn independently from Uniform([0, B]).

Define the diameter of the graph to be the longest shortest path:
\[
D_n \overset{\text{def}}{=} \max_{i,j \in V} \min_{\text{path } p \text{ from } i \text{ to } j} \text{cost}(p_{i \to j}),
\] (10)
where the cost of a path \( p_{i \to j} = [i_0, i_1, \ldots, i_L] \) is \( \sum_{j=1}^{L} C_{i_{j-1}i_j} \).

Prove that the diameter concentrates:
\[
\mathbb{P}[|D_n - \mathbb{E}[D_n]| \geq \epsilon] \leq 2 \exp \left( \frac{-4\epsilon^2}{n(n-1)B^2} \right).
\] (11)
b. (2 points) (largest eigenvalue)

Let $p^*$ be a distribution over vectors in $\mathbb{R}^d$ with mean zero, covariance $\Sigma$, and $\|x\|_2 \leq B$ for $x \sim p^*$. Let $x_1, \ldots, x_n$ be drawn i.i.d. from $p^*$. Form the sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top.$$  \hspace{1cm} (12)

Show that the maximum eigenvalue can’t be too much smaller than the true maximum:

$$\mathbb{P}[\lambda_{\max}(\hat{\Sigma}) \leq \lambda_{\max}(\Sigma) - \epsilon] \leq \exp\left(\frac{-n\epsilon^2}{2B^4}\right).$$  \hspace{1cm} (13)