Stats 231 / CS229T Homework 1 Solutions

Question 1 (Moment generating functions of squares)**:** In this question, we investigate sub-exponential and sub-Gaussian random variables. We let $[t]_+ = \max\{0, t\}$ denote the positive part, and say that $1/0 = +\infty$.

(a) Let Z be $\mathsf{N}(0, \sigma^2)$. Show that

$$\mathbb{E}[e^{\lambda Z^2}] = \frac{1}{\sqrt{[1 - 2\lambda\sigma^2]_+}}.$$

(b) Let X be a mean-zero σ^2-sub-Gaussian random variable. Show that

$$\mathbb{E}[e^{\lambda X^2}] \le \frac{1}{\sqrt{[1 - 2\lambda\sigma^2]_+}} \quad \text{for } \lambda \ge 0.$$

Hint: Introduce an independent Gaussian Z (with some particular variance) and compute $\mathbb{E}[e^{ZX}]$.

(c) Let $Z \sim \mathsf{N}(0, \sigma^2)$. Show that $Z^2 - \mathbb{E}[Z^2]$ is sub-exponential and give sub-exponential parameters for it.

Answer:

(a) We write out the integrals. We have

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z^2}] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(\lambda z^2 - \frac{1}{2\sigma^2} z^2\right) dz \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-\frac{1 - 2\lambda\sigma^2}{2\sigma^2} z^2\right) dz.
\end{aligned}$$

If $2\lambda\sigma^2 \ge 1$, clearly the last integral is $+\infty$. Otherwise, we use that (by the normalization for the Gaussian distribution) $\int e^{-\frac{1}{2\tau^2}z^2} dz = \sqrt{2\pi\tau^2}$, so

$$\int \exp\left(-\frac{1 - 2\lambda\sigma^2}{2\sigma^2} z^2\right) dz = \sqrt{2\pi \frac{\sigma^2}{1 - 2\lambda\sigma^2}}$$

assuming that $2\lambda\sigma^2 < 1$. This is the result.

(b) We assume that $\lambda > 0$ as the result is trivial otherwise. Let $Z \sim \mathsf{N}(0, \sqrt{2\lambda})$. Then

$$\mathbb{E}[e^{ZX}] = \mathbb{E}[e^{\lambda X^2}]$$

by the standard MGF for a Gaussian. Thus we have

$$\mathbb{E}[e^{\lambda X^2}] = \mathbb{E}[e^{ZX}] \stackrel{(i)}{\le} \mathbb{E}\left[\exp\left(\frac{\sigma^2 Z^2}{2}\right)\right] = \frac{1}{\sqrt{\left[1 - 2(\sigma^2/2)\sqrt{2\lambda}^2\right]_+}} = \frac{1}{\sqrt{[1 - 2\lambda\sigma^2]_+}}$$

by part (a).

(c) For $\lambda \in \mathbb{R}$ we have that

$$\mathbb{E}[\exp(\lambda(Z^2 - \mathbb{E}[Z^2]))] = \exp\left(-\frac{1}{2}\log(1 - 2\lambda\sigma^2) - \lambda\sigma^2\right),$$

where we define $\log(t) = -\infty$ for $t \leq 0$. By a Taylor expansion, we have $\log(1-x) = -x - \frac{1}{2}x^2 + O(x^3)$ as $x \to 0$, and moreover, $\log(1-x) \geq -x - x^2$ for $|x| \leq \frac{1}{2}$. Thus we have

$$\begin{aligned}\mathbb{E}[\exp(\lambda(Z^2 - \mathbb{E}[Z^2]))] &= \exp\left(-\frac{1}{2}\log(1 - 2\lambda\sigma^2) - \lambda\sigma^2\right) \\ &\leq \exp\left(\lambda\sigma^2 + \lambda^2\sigma^4 - \lambda\sigma^2\right) = \exp\left(\lambda^2\sigma^4\right) \quad \text{for } |\lambda| \leq \frac{1}{2}.\end{aligned}$$

Recalling the definition of sub-exponential random variables, we say Y is (τ^2, b)-sub-exponential of $\mathbb{E}[e^{\lambda Y}] \leq \exp(\frac{\lambda^2\tau^2}{2})$ for $|\lambda| \leq 1/b$, we obtain that $X = Z^2 - \mathbb{E}[Z^2]$ is $(2\sigma^4, 2)$-sub-exponential.

□

Question 2 (Concentration inequalities)**:** Let X_i be independent random variables with $|X_i| \leq c$ and $\mathbb{E}[X_i] = 0$.

(a) Let $\sigma_i^2 = \mathrm{Var}(X_i)$. Prove that

$$\mathbb{E}[e^{\lambda X_i}] \leq \exp\left(\frac{\sigma_i^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right).$$

(b) Let $h(u) = (1+u)\log(1+u) - u$ and let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Prove *Bennett's inequality*, that is, for any $t \geq 0$ we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right).$$

(c) Under the notation of part (b), prove *Bernstein's inequality*, that is, that for any $t \geq 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \vee \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \leq -t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2 + 2ct/3}\right),$$

where $a \vee b = \max\{a, b\}$.

(d) When is Bernstein's inequality tighter than the Hoeffding's inequality for bounded random variables? Recall that Hoeffding's inequality states (under the above conditions on X_i) that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq \exp\left(-\frac{nt^2}{2c^2}\right).$$

Answer:

(a) Let $\sigma = \sigma_i$ for shorthand and $\mathrm{Var}(X) \le \sigma^2$. We perform a Taylor expansion:

$$\mathbb{E}[e^{\lambda X}] = 1 + \sum_{k=2}^{\infty} \frac{\mathbb{E}[X^k]\lambda^k}{k!} \le 1 + \sum_{k=2}^{\infty} \frac{\mathbb{E}[X^2]c^{k-2}\lambda^k}{k!}$$
$$= 1 + \frac{\sigma^2}{c^2} \sum_{k=2}^{\infty} \frac{c^k \lambda^k}{k!} = 1 + \frac{\sigma^2}{c^2}\left(e^{\lambda c} - 1 - \lambda c\right).$$

Using that $1 + x \le e^x$ for all x gives the result.

(b) Applying the standard Chernoff bound technique, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t\right) \le \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] e^{-\lambda t} \le \exp\left(\frac{n\sigma^2}{c^2}\left(e^{\lambda c} - 1 - \lambda c\right) - \lambda t\right)$$

for all $\lambda \ge 0$, where we have used part (a). Note that $\phi(\lambda) = \frac{n\sigma^2}{c^2}(e^{\lambda c} - 1 - \lambda c) - \lambda t$ is convex in λ, so that differentiating and setting to zero gives us its minimizer. We have

$$\phi'(\lambda) = \frac{n\sigma^2}{c}\left(e^{\lambda c} - 1\right) - t = 0 \;\text{ so }\; e^{\lambda c} = 1 + \frac{ct}{n\sigma^2} \;\text{ or }\; \lambda = \frac{1}{c}\log\left(1 + \frac{ct}{n\sigma^2}\right).$$

Substituting in the preceding display gives

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t\right) \le \exp\left(-\frac{t}{c}\log\left(1 + \frac{ct}{n\sigma^2}\right) + \frac{n\sigma^2}{c^2}\left(\frac{ct}{n\sigma^2} - \log\left(1 + \frac{ct}{n\sigma^2}\right)\right)\right)$$
$$= \exp\left(-\frac{n\sigma^2}{c^2}\left(1 + \frac{ct}{n\sigma^2}\right)\log\left(1 + \frac{ct}{n\sigma^2}\right) + \frac{n\sigma^2}{c^2}\frac{ct}{n\sigma^2}\right) = \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right)$$

as desired.

(c) We ignore the lower (negative) tail as its proof is identical to the positive tail in part (b). We must show

$$-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{\sigma^2}\right) \le -\frac{nt^2}{2\sigma^2 + 2ct/3} \quad\text{or}\quad -\frac{\sigma^2}{c^2} h\left(\frac{ct}{\sigma^2}\right) \le -\frac{t^2}{2\sigma^2 + 2ct/3} \tag{1}$$

for all $t \ge 0$. Letting $u = ct/\sigma^2$, then inequality (1) holds if and only if

$$-\frac{\sigma^2}{ct} h\left(\frac{ct}{\sigma^2}\right) \le -\frac{ct}{2\sigma^2 + 2ct/3} \;\text{ iff }\; -\frac{\sigma^2}{ct} h\left(\frac{ct}{\sigma^2}\right) \le -\frac{\frac{ct}{\sigma^2}}{2 + \frac{2}{3}\frac{ct}{\sigma^2}} \;\text{ iff }\; -\frac{1}{u}h(u) \le -\frac{u}{2 + \frac{2}{3}u},$$

or

$$h(u) \ge \frac{u^2}{2 + \frac{2}{3}u} \quad\text{for all } u \ge 0. \tag{2}$$

At $u = 0$, inequality (2) holds because both sides are zero. If we can show that the derivative of $h(u)$ is larger than that of $u^2/(2 + 2u/3)$ for all $u \ge 0$, this is sufficient.

With that in mind, we have that inequality (2) holds if for all $u \ge 0$, we have

$$\log(1 + u) = h'(u) \ge \frac{2u}{2 + \frac{2}{3}u} - \frac{\frac{2}{3}u^2}{(2 + \frac{2}{3}u)^2} = \frac{4u + \frac{2}{3}u^2}{(2 + \frac{2}{3}u)^2} = \frac{u + \frac{1}{6}u^2}{(1 + \frac{1}{3}u)^2} = \frac{u + \frac{1}{6}u^2}{1 + \frac{2}{3}u + \frac{1}{9}u^2}.$$

Taking second derivatives, we have that it is sufficient that for all $u \geq 0$, we have

$$\frac{1}{1+u} \geq \frac{1+\frac{1}{3}u}{(1+\frac{1}{3}u)^2} - \frac{2(u+\frac{1}{6}u^2)}{3(1+\frac{1}{3}u)^3} = \frac{1}{1+\frac{1}{3}u} - \frac{2(u+\frac{1}{6}u^2)}{3(1+\frac{1}{3}u)^3}$$

or

$$\frac{-2u}{(1+u)(3+u)} \geq -\frac{2(u+\frac{1}{6}u^2)}{3(1+\frac{1}{3}u)^3} \text{ i.e. } \frac{1}{1+u} \leq \frac{1+\frac{1}{6}u}{(1+\frac{1}{3}u)^2} \text{ i.e. } 1+\frac{2}{3}u+\frac{1}{9}u^2 \leq 1+\frac{7}{6}u+\frac{u^2}{6}.$$

The final inequality is clear.

An easier way to do this proof is to simply note that

$$e^\lambda - \lambda - 1 \leq \frac{\lambda^2}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda}{3}\right)^k = \frac{\lambda^2}{2(1-\lambda/3)},$$

then choose $\lambda = \frac{t}{\sigma^2+t/3}$ in the precursor to Bennett's inequality.

(d) We solve

$$\frac{nt^2}{2\sigma^2+2ct/3} = \frac{nt^2}{2c^2} \quad \text{or} \quad 2\sigma^2 + \frac{2ct}{3} = 2c^2$$

for t. Evidently,

$$0 \leq t \leq \frac{3}{c}(c^2-\sigma^2) = 3c - 3\frac{\sigma^2}{c}$$

is sufficient for Bernstein's inequality to be tighter—that is, for small t, it is better to use variance-based-bounds. (Because we have $\sigma^2 \leq c^2$ always.)

□

Question 3: In the realizable setting with binary classification (where the expected risk minimizer $h^\star$ satisfies $L(h^\star) = 0$ for the 0-1 error), we obtained excess risk bounds of $O(1/n)$, but in the unrealizable setting, we had $O(\sqrt{1/n})$. What if the learning problem is *almost* realizable, in that $L(h^\star)$ is small? This problem explores ways to interpolate between $1/n$ and $1/\sqrt{n}$ rates, showing that (roughly) $\sqrt{L(h^\star)/n}+1/n$ rates are possible by developing generalization bounds that depend on the *variance* of losses (recall Question 2).

(a) Assume that the loss function $\ell(y,t)$ takes values in $[0,1]$, where $L(h) = \mathbb{E}[\ell(Y,h(X))]$, and let $\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i,h(X_i))$. Show that for all $\epsilon \geq 0$ we have

$$\mathbb{P}\left(\widehat{L}_n(h) - L(h) \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2(L(h)+\epsilon/3)}\right).$$

(Note that if $L(h) = 0$, this bound scales as $e^{-n\epsilon} \ll e^{-n\epsilon^2}$ for $\epsilon \approx 0$.)

(b) We now show that bad hypotheses usually look pretty bad. Fix any $\varepsilon(h), \epsilon \geq 0$, and assume that

$$L(h) \geq \varepsilon(h) + \epsilon.$$

Show that

$$\mathbb{P}\left(\widehat{L}_n(h) \leq \varepsilon(h)\right) \leq \exp\left(-\frac{n\epsilon^2}{2(\varepsilon(h)+4\epsilon/3)}\right).$$

(c) Assume $\text{card}(\mathcal{H}) < \infty$ and let $h^\star$ satisfy $L(h^\star) = \min_{h\in\mathcal{H}} L(h)$. Using the preceding parts, conclude that if $\widehat{h}_n \in \text{argmin}_{h\in\mathcal{H}} \widehat{L}_n(h)$, then

$$\mathbb{P}\left(L(\widehat{h}_n) - L(h^\star) \geq 2\epsilon\right) \leq \text{card}(\mathcal{H}) \exp\left(-\frac{n\epsilon^2}{2(L(h^\star) + 7\epsilon/3)}\right).$$

Show that this implies (for appropriate numerical constants c_1, c_2) that with probability at least $1 - \delta$, we have

$$L(\widehat{h}_n) \leq L(h^\star) + c_1\sqrt{\frac{L(h^\star)\log\frac{\text{card}(\mathcal{H})}{\delta}}{n}} + c_2\frac{\log\frac{\text{card}(\mathcal{H})}{\delta}}{n}.$$

(d) How does this bound compare with a more naive strategy based on applying Hoeffding's inequality and a union bound?

Answer:

(a) First, we bound the variance of $\ell(Y, h(X))$. We have

$$\text{Var}[\ell(Y, h(X))] \leq \mathbb{E}[\ell(Y, h(X))^2] \leq \mathbb{E}[\ell(Y, h(X))] = \ell(h),$$

where the first inequality is true of all random variables, the second inequality is because $a^2 \leq a$ for $a \in [0, 1]$, and the last inequality is the bound on the expected value given in the problem statement.

Now, $n\widehat{L}_n$ is the sum of n independent copies of $\ell(Y, h(X))$, each of which are bounded in $[0, 1]$ and have variance at most $L(h)$. Therefore, by Bernstein inequality we have:

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

(b) Applying Bernstein's inequality to $-\ell(Y, h'(X))$ gives us the same inequality as for $\ell(Y, h'(X))$, except on the other side of the mean:

$$\mathbb{P}[\widehat{L}_n(h') - L(h') \leq -\epsilon'] \leq \exp\left(\frac{-n\epsilon'^2}{2(L(h') + \epsilon'/3)}\right).$$

Let us set ϵ' to be $L(h') - \varepsilon(h)$. Then we have the bound

$$\mathbb{P}[\widehat{L}_n(h') \leq \varepsilon(h)] \leq \exp\left(\frac{-n(L(h') - \varepsilon(h))^2}{2(L(h') + (L(h') - \varepsilon(h))/3)}\right).$$

We claim that

$$\frac{(L(h') - \varepsilon(h))^2}{L(h') + (L(h') - \varepsilon(h))/3} \geq \frac{\epsilon^2}{L(h') + 4\epsilon/3}$$

for $L(h') \geq \varepsilon(h) + \epsilon$, from which the result would follow.

To show this, consider the function $f(L; E') := \frac{(L-E')^2}{L+(L-E')/3}$. It suffices to show that the function $f(\cdot; E')$ is monotonically increasing on $[E', \infty)$. Taking the derivative with respect to L:

$$\begin{aligned}
\frac{d}{dL}\frac{(L - E')^2}{L + (L - E')/3} &= \frac{2(L - E')(L + (L - E')/3) - (4/3)(L - E')^2}{(L + (L - E')/3)^2} && (3)\\
&= \frac{L - E'}{(L + (L - E')/3)^2}\left(2(L + (L - E')/3) - 4(L - E')/3\right) && (4)\\
&= \frac{L - E'}{(L + (L - E')/3)^2}\left(4L/3 + 2E'/3\right), && (5)
\end{aligned}$$

which is positive whenever $L > E'$. This proves the claim of monotonicity of $f(\cdot; E')$ on $[E', \infty)$ and thus the desired result.

(c) Consider the following two events:

(a) $\widehat{L}_n(h^*) \geq L(h^*) + \epsilon$.

(b) $\widehat{L}_n(h) \leq L(h^*) + \epsilon$ for all h with $L(h) \geq E + 2\epsilon$.

Let P and Q be the probabilities of these two events holding, respectively. The probability of either event happening is at most $P + Q$. Observe that if $L(\hat{h}) \geq L(h^\star) + 2\epsilon$ happens, then one of the two events must have happened. Therefore, $\mathbb{P}[L(\widehat{h}_n) - L(h^\star) \geq 2\epsilon] \leq P + Q$.

P is easy to bound: just apply the result from (a) to get that

$$P \leq \exp\left(\frac{-n\epsilon^2}{L(h^\star) + \epsilon/3}\right).$$

To bound Q, first note that for any h such that $L(h) \geq L(h^\star) + 2\epsilon$, the probability that $\hat{L}(h) \leq L(h^\star) + \epsilon$ for any given h can be bounded by $\exp\left(\frac{-n\epsilon^2}{L(h^\star)+7\epsilon/3}\right)$ by applying part (b) with $\varepsilon(h) = L(h^\star) + \epsilon$ (since we have the bound $L(h) \geq L(h^\star) + 2\epsilon = \varepsilon(h) + \epsilon$ in this case). Then, there are at most $|\mathcal{H}| - 1$ such h (since there are at most $|\mathcal{H}|$ hypotheses total and at least one of them — namely, h^* — has $L(h) < L(h^\star) + 2\epsilon$). Therefore, we have,

$$Q \leq (|\mathcal{H}| - 1) \exp\left(\frac{-n\epsilon^2}{L(h^\star) + 7\epsilon/3}\right).$$

Combining these gives

$$\mathbb{P}[L(\hat{h}) - E \geq 2\epsilon] \leq P + Q \leq |\mathcal{H}| \exp\left(\frac{-n\epsilon^2}{L(h^\star) + 7\epsilon/3}\right), \tag{6}$$

which yields the desired result.

(d) The realizable case gives a bound of $|\mathcal{H}| \exp(-c_1 n\epsilon)$, and the regular Hoeffding bound gives a bound of $|\mathcal{H}| \exp(-c_2 n\epsilon^2)$. The bound in part (c) is in some sense an interpolation between them: when E is small compared to ϵ then the bound behaves like the bound in the realizable case; when E is large compared to ϵ, it behaves like the regular Hoeffding bound. The range of value of ϵ for which we get the same behavior as the realizable case depends on "how close" to realizable we are.

□

Question 4 (VC Dimension)**:**

(a) Let $\mathcal{X} = \mathbb{R}^2$ and consider the hypothesis class of indicators for convex polygons, that is,

$$\mathcal{H} = \{h_C(x) = \mathbf{1}\{x \in C\} : C \text{ is a convex polygon}\}.$$

What is $\mathsf{VC}(\mathcal{H})$?

(b) A decision tree T is a binary tree that classifies points in $\mathbb{R}^d$. Each internal node (non-leaf node) v in T has an attribute $j_v \in \{1, 2, \ldots, d\}$ and a threshold $t_v \in \mathbb{R}$. Each leaf node is labeled with one of the two classes, +1 or -1. Given a point $x \in \mathbb{R}^d$, we start from the root, and every time we encounter an internal node v, we check the condition $\mathbf{1}\{x_{j_v} \geq t_v\}$. We go to the left child if the condition is not met, and the right child otherwise. We repeat such process until we reach a leaf node, and classifies the point according to the label of the node.

Show that the VC dimension of the hypothesis class corresponding to all depth-k decision trees defined above is $\Omega(2^k \log d)$.

Answer:

(a) For $p = 1$, we can only shatter 2 points. For other p, $\mathcal{H}$ has infinite VC dimension, for $p \geq 2$. Consider any n distinct points on the p-dimensional sphere $x_1, \ldots, x_n$ so that $\|x_i\|_2 = 1$. We can assign positive labels to any subset of $m \leq n$ points $x_1, \ldots, x_m$, by using the hypothesis $h(x) = \mathbb{I}[x \in \text{Convex-hull}(x_1, \ldots, x_m)] \in \mathcal{H}$. To show that any other point $x_j, j > m$ with norm 1 will be assigned a negative label, we can show that $\|x\|_2 < 1$ for $x \in \text{Convex-hull}(x_1, \ldots, x_m)$, $x \notin \{x_{1:m}\}$. For a non-vertex point in the convex-hull $x = \sum_{i=1}^m \theta_i x_i$, for $1 > \theta_i \geq 0$ and $\sum_{i=1}^m \theta_i = 1$, then

$$\|x\|_2^2 = \sum_{i=1}^m \sum_{j=1}^m \theta_i \theta_j x_i \cdot x_j < \sum_{i=1}^m \sum_{j=1}^m \theta_i \theta_j = 1$$

The strict inequality comes from distinctness as well as x being a non-vertex. An intuitive argument is sufficient for the problem as well.

(b) We first prove the case of $k = 1$, where we get to split the points (once) along a certain axis. In this case, suppose we have $n = \lfloor \log_2 d \rfloor$ points. We associate each of the d dimensions j with a subset $S(j)$ of the n points, and let $x_j^{(i)} = 1$ if $i \in S(j)$ and -1 otherwise. In this way, each desired labeling $S(j)$ can be achieved using the condition $\mathbb{I}[x_j \geq 0]$. Since we have at least 2^n dimensions, we can achieve all labelings.

For the general case, we show that increasing the depth of the tree by 1 allows us to at least double the number of points we can shatter. If this is true, by induction we can shatter at least $2^k \lfloor \log_2 d \rfloor$ points.

Suppose depth-k trees can shatter an n-element set A. Without loss of generality, $x_1 > 0$ for all $x \in A$, which can be achieved by shifting the points. Depth-k trees can also shatter $A' = \{(-x_1, x_2, \ldots, x_d) : x \in A\}$. Thus, the set $B = A \cup A'$ can be shattered by depth-$(k+1)$ trees as follow: at the root, let the condition be $\mathbb{I}[x_1 \geq 0]$, splitting the points into A and A', and we can shatter both sets with depth-k trees by assumption.

□

Question 5 (Rademacher complexity)**:** In many applications, for example, in natural language processing (NLP), one has very sparse feature vectors in very high dimensions. Suppose that we know that any feature vector $x \in \{0, 1\}^d$ satisfies $\|x\|_1 \leq k$, i.e. there are at most k non-zeros.

(a) Give an example application and data representation where such characteristics might hold.

You decide to use a linear classifier for this "sparse x" problem, where you represent the classifier by a weight vector $w \in \mathbb{R}^d$ so that $f(x) = w^\top x$, and you restrict your classifiers to be in a particular norm ball $\{w : \|w\| \leq B\}$.

(b) Is using the ℓ_1-norm ball, i.e. $\mathcal{F} = \{x \mapsto f(x) = w^\top x : \|w\|_1 \leq B\}$ likely to be a good idea? In a sentence or two, explain why or why not. (No need for serious mathematical derivations.)

(c) You decide instead to use dense feature vectors, restricting w to an ℓ_∞ norm ball, i.e.

$$\mathcal{F} := \{f \mid f(x) = w^\top x, \|w\|_\infty \leq B\}.$$

Give an upper bound on $R_n(\mathcal{F})$, which should depend on k (the number of non-zeros), n, B, and d.

Answer:

(a) In document classification, considering the binary features $x \in \mathbb{R}^p$ where $x_i = 0$ if and only if the document contains the i-th word. Since typically a document can only have a very small fraction of words in dictionary, in this case, the features are sparse for the samples.

(b) It depends. Learning linear classifiers with l_1 constrained w typically results in sparse weights w. In document classification, sparse weighting of all dictionary words can get you extraction of the key words that separate the two class of documents. Yet, it suffers the risk that when B is too small, since already the feature space is sparse, a too sparse weighting of w can lose important and useful features that discriminate the two classes.

(c) First, note that $\|x\|_1 \leq k$. One can compute the Rademacher complexity:

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{w:\|w\|_\infty \leq B} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle w, Z_i\rangle\right] = \frac{1}{n}\mathbb{E}\left[\sup_{w:\|w\|_\infty \leq B}\left\langle w, \sum_{i=1}^n \sigma_i Z_i\right\rangle\right] = \frac{B}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \sigma_i Z_i\right\|_1\right]$$

where the last equality follows from Hölder's inequality, or the fact that $\|\cdot\|_1$ is the dual norm of $\|\cdot\|_\infty$. Each Z_i has at most k 1's, so there are a total of at most kn 1's, spread across the d dimensions. Let a_j be total number of 1's among $Z_1, Z_2, \ldots, Z_n$ in the j-th dimension, so $a_1 + a_2 + \ldots + a_d \leq nk$. Moreover, the expected value of the j-th dimension of $\sum_{i=1}^n \sigma_i Z_i$ can be upper bounded by

$$\mathbb{E}\left[|\sum_{l=1}^{a_j}\sigma_l|\right] \leq \sqrt{\mathbb{E}\left[\left(\sum_{l=1}^{a_j}\sigma_l\right)^2\right]} = \sqrt{\mathbb{E}\left[\sum_{l_1,l_2=1}^{a_j}\sigma_{l_1}\sigma_{l_2}\right]} = \sqrt{a_i}$$

where the first inequality follows from the bound $\mathbb{E}[|A|] \leq \sqrt{\mathbb{E}[A^2]}$ (due to Jensen's inequality). Since the 1-norm is the sum of absolute values of each dimension, by linearity of expectation,

$$R_n(\mathcal{F}) \quad \leq \quad \frac{B}{n}\sum_{i=1}^d \sqrt{a_i} \leq \frac{B}{n}\sqrt{\sum_{i=1}^d a_i \sum_{i=1}^d 1} \leq \frac{B\sqrt{kd}}{\sqrt{n}}$$

where the second step uses Cauchy-Schwartz.

$\square$