Mathematics of Machine Learning
Homework 11 - Solutions

Instructors: Afonso S. Bandeira & Nikita Zhivotovskiy
Course Coordinator: Pedro Abdalla Teixeira

May 21, 2021

Try to solve the questions before looking to the answers. Every item must be proved rigorously. Starred problems are harder.

Problem 1

We say that two unit vectors $u, v \in \mathbb{R}^d$ are ε-almost orthogonal if $|\langle u, v\rangle| \leq \varepsilon$. A set of vectors in which every vector is ε-almost orthogonal to each other is called ε-almost orthogonal set. For example, for $\varepsilon = 0$, we recover the usual notion of orthogonal vectors and the canonical basis is an 0-almost orthogonal set. Clearly, the maximal size of an 0-orthogonal set in $\mathbb{R}^d$ is d. The goal of this exercise is to show that $\varepsilon > 0$, the size of an ε-almost orthogonal set is exponentially large in d and discuss its relevance for frame theory. The proof strategy relies on the probabilistic method, a nonconstructive strategy to show the existence of certain object by using probability.

The following inequality will be useful (you do not need to prove it), if $X_1, \dots, X_n$ are i.i.d Rademacher random variables, i.e, taking values 1 and -1 with probability $\frac{1}{2}$ each, then

$$\mathbb{P}(\sum_{i=1}^{n} X_i \geq t) \leq e^{\frac{-t^2}{2n}}.$$

(a) Let $\varphi \in \mathbb{R}^d$ be a random vector drawn as follows: Each entry is generated independently at random following a Rademacher distribution. Consider N i.i.d copies of the random vector φ, namely $\varphi_1, \dots, \varphi_N$. Prove the following statement: If $N = \lfloor e^{d\varepsilon^2/4} \rfloor$, then the probability that set of normalized vectors $\{\frac{1}{\sqrt{d}}\varphi_1, \dots, \frac{1}{\sqrt{d}}\varphi_N\}$ is an ε-almost orthogonal set is positive (strictly larger than 0).

(b) Conclude that letter "a" implies in the existence of ε-almost orthogonal set with size exponentially large in d.

(c) Give an interpretation of this fact in terms of mutual coherence of frames in the sense of Challenge 11.4 in the lecture notes.

(d) * Improve the estimate of letter "a" from $N = \lfloor e^{d\varepsilon^2/4} \rfloor$ to $N = \lfloor ce^{d\varepsilon^2/2} \rfloor$, where c is an absolute constant less than one. (Hint: Compute the expected number of unordered pair that violates the condition of ε-almost orthogonality, then delete one vector from each pair).

Solution 1

(a) Observe that $\mathbb{P}(\frac{1}{d}|\langle \varphi_i, \varphi_j\rangle| > \varepsilon) = \mathbb{P}(|\sum_{i=1}^d X_i| > \varepsilon d)$, where X_i are also independent Rademacher random variables. By the concentration inequality mentioned above we have

$$\mathbb{P}(|\sum_{i=1}^{d} X_i| > \varepsilon d) \leq 2\mathbb{P}(\sum_{i=1}^{d} X_i > \varepsilon d) \leq 2e^{-\varepsilon^2 \frac{d}{2}}.$$

Therefore, by union bound,

$$\mathbb{P}(\exists i < j \ : \ \frac{1}{d}|\langle \varphi_i, \varphi_j\rangle| > \varepsilon) \leq 2\binom{N}{2} e^{-\varepsilon^2 \frac{d}{2}} < 2\frac{N^2}{2} e^{-\varepsilon^2 \frac{d}{2}} \leq 1.$$

We conclude that the probability the set of vectors $\{\frac{1}{\sqrt{d}}\varphi_1, \dots, \frac{1}{\sqrt{d}}\varphi_N\}$ is ε-almost orthogonal with positive probability.

(b) If the set of vectors with the required property does not exist, then it is an empty event. By construction, the probability of an empty event to occur is zero. The event above has positive probability, so it must be non-empty.

(c) Observe that the set of unit vectors $\frac{1}{\sqrt{d}}\varphi_1, \dots, \frac{1}{\sqrt{d}}\varphi_N$ is ε-almost orthogonal means that they have small worst case coherence. In particular, we prove the existence of N vectors in d dimension with coherence of order $\sqrt{\frac{\log N}{d}}$.

(d) In letter "a", we proved that the expected number of pairs violating the ε-almost orthogonality condition is at most $\binom{N}{2}2e^{-d\frac{\varepsilon^2}{2}}$. If we delete one vector from each pair, we will be left with at least $N - \binom{N}{2}2e^{-d\frac{\varepsilon^2}{2}}$, if we choose $N = \frac{1}{2}e^{\frac{d\varepsilon^2}{2}}$ we get $(N-1)e^{\frac{-d\varepsilon^2}{2}} \leq \frac{1}{2}$, so $N - \binom{N}{2}2e^{-d\frac{\varepsilon^2}{2}} = N - N(N-1)e^{-d\frac{\varepsilon^2}{2}} \geq \frac{N}{2}$. Therefore there exits an ε-almost orthogonal set of size $\frac{N}{2} = \left\lfloor \frac{1}{4}e^{-\varepsilon^2 \frac{d}{2}} \right\rfloor$.

Problem 2

The goal of this exercise is to solve the Challenge 12.4 in the lecture notes. Recall that a random graph $G \sim \mathcal{G}(n, p)$ is a graph of n vertices generated by placing each possible edge independently at random with probability p. We use the standard asymptotic notation, $f(n) \ll g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$, also $g(n) \gg f(n)$ if $\lim_{n\to\infty} \frac{g(n)}{f(n)} = \infty$. Moreover $f(n) = o(1)$ if $\lim_{n\to\infty} f(n) = 0$. Finally we define $p := \frac{\lambda \log n}{n}$ for some constant $\lambda > 0$.

(a) Prove that if $\lambda \le 1 - c$, where $c > 0$ is an absolute constant, then the graph G has an isolated vertex with probability $1 - o(1)$. (Hint: Consider a random variable X_n to be the number of isolated vertices in the graph G, you need to prove that the probability of X_n be larger than zero is $1 - o(1)$. Use the inequality $\mathbb{P}(X_n = 0) \le \frac{\operatorname{Var} X_n}{\mathbb{E}X_n^2}$).

(b) Now observe the following: A graph is disconnected if and only if there exits a set of k nodes such that $k \le \left\lfloor \frac{n}{2} \right\rfloor$ and there is no edge connecting the set of k nodes with the complement set of $n - k$ nodes. Use this fact to prove that if $\lambda \ge (1 + c)$ for an absolute constant $c > 0$, then the graph is connected with probability $1 - o(1)$. (Hint: Use union bound twice).

Solution 2

(a) Let $X_n = \sum_{i=1}^n I_i$, where I_i is a indicator function for the event that the vertex i is isolated (zero degree). Note that the probability of a certain vertex i to be isolated is $(1-p)^{n-1}$. So $\mathbb{E}X_n = n(1-p)^{n-1}$. Simple calculus shows that $\mathbb{E}X_n \longrightarrow \infty$ as $n \longrightarrow \infty$. Now we proceed to bound the variance of X_n. Observe that $\operatorname{Var} X_n = n \operatorname{Var} I_1 + n(n-1)\operatorname{Cov}(I_1, I_2)$, where $\operatorname{Cov}(I_1, I_2) := \mathbb{E}[I_1 I_2] - \mathbb{E}I_1\mathbb{E}I_2$. The first term, $\mathbb{E}[I_1 I_2]$, is the probability that node 1 and node 2 are isolated. For this event to occur we need $2n-3$ edges to be absent (not placed), we have

$$\mathbb{E}[I_1 I_2] = (1-p)^{2n-3} \text{ and } \mathbb{E}I_1\mathbb{E}I_2 = (1-p)^{2n-2}.$$

Now we use that $\mathbb{E}X_n^2 \geq (\mathbb{E}X_n)^2$ to write

$$\mathbb{P}(X_n = 0) \leq \frac{n(1-p)^n(1-(1-p)^n) + n(n-1)p(1-p)^{2n}/(1-p)}{n^2(1-p)^{2n}}.$$

By simple calculus we check that $\mathbb{P}(X_n = 0) = o(1)$, equivalently $\mathbb{P}(X_n > 0) = 1 - o(1)$. In particular, the graph is disconnected with high probability.

(b) We assume without loss of generality that n is even (just to simplify the notation) and write

$$\begin{aligned}
\mathbb{P}(G \text{ is disconnected}) &= \mathbb{P}(\cup_{k=1}^{\frac{n}{2}} \text{some set of } k \text{ nodes is disconnected}) \\
&\leq \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} \mathbb{P}(\text{a specific set of } k \text{ nodes is disconnected}) \\
&= \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} (1-p)^{k(n-k)}.
\end{aligned}$$

The rest of the proof is dedicated to show that the last term above vanishes. We choose $n^* := \lfloor n(1-\lambda^{-1}) \rfloor$ such that $\lambda(n-n^*) > n$. By the numeric inequality $1 - x < e^{-x}$ we obtain

$$S := \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} (1-p)^{k(n-k)} \leq \sum_{k=1}^{\frac{n}{2}} \binom{n}{k} n^{-\lambda k(n-k)/n}.$$

We split the summation in two parts. The first one is

$$\sum_{k=1}^{n^*} \binom{n}{k} n^{-\lambda k(n-k)/n} \leq \sum_{k=1}^{n^*} n^{-k(-1+\lambda(n-n^*)/n)} \leq \frac{n^{\lambda(n-n^*)/n-1}}{1 - n^{\lambda(n-n^*)/n-1}} = o(1).$$

In the second part, we use the estimate $\binom{n}{k} \leq (\frac{en}{k})^k$.

$$\sum_{k=n^*+1}^{\frac{n}{2}} \binom{n}{k} n^{-\lambda k(n-k)/n} \leq \sum_{k=n^*+1}^{\frac{n}{2}} (\frac{en^{1-\lambda(n-k)/n}}{n^*+1})^k \leq \sum_{k=n^*+1}^{\frac{n}{2}} (\frac{en^{-\lambda/2}}{1-\lambda^{-1}})^k.$$

Finally, notice that the right hand side is a geometric progression with $(\frac{en^{-\lambda/2}}{1-\lambda^{-1}}) := \delta < 1$. The right hand side is $\frac{\delta^{n^*}}{1-\delta}$, since $n^* \gg 1$, we obtain that the right hand side term is $o(1)$. We conclude that $S = o(1)$.

Problem 3

Let $\mathcal{A}$ and $\mathcal{B}$ be two family of events with finite VC-dimension. Prove the following facts:

(a) If $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, then $\mathcal{S}_{\mathcal{C}}(n) \le \mathcal{S}_{\mathcal{A}}(n) + \mathcal{S}_{\mathcal{B}}(n)$.

(b) If $\mathcal{C} = \mathcal{A} \cap \mathcal{B}$, then $\mathcal{S}_{\mathcal{C}}(n) \le \mathcal{S}_{\mathcal{A}}(n)\mathcal{S}_{\mathcal{B}}(n)$.

Solution 3

(a) For notation simplicity, let $x = (x_1, \dots, x_n)$ and $h_C(x) := (\mathbb{1}_{x_1 \in C}, \dots, \mathbb{1}_{x_n \in C})$. Observe that the set $\{h(x) : x \in \mathcal{C}^n\} = \{h(x) : x \in \mathcal{A}^n \cup \mathcal{B}^n\}$, so by the sub-additivity of the supremum we have

$$\sup_{x \in \mathcal{C}^n} |h(x) : x \in \mathcal{C}| \le \sup_{x \in \mathcal{A}^n} |h(x) : x \in \mathcal{A}| + \sup_{x \in \mathcal{B}^n} |h(x) : x \in \mathcal{B}|.$$

It immediately follows that $\mathcal{S}_{\mathcal{C}}(n) \le \mathcal{S}_{\mathcal{A}}(n) + \mathcal{S}_{\mathcal{B}}(n)$.

(b) Let $x = (x_1, \dots, x_n)$. Let $h_1, \dots, h_k$ be a collection of binary vectors formed by indicator functions of events in $x_i \in \mathcal{A}$. By definition of growth function, $k \le \mathcal{S}_{\mathcal{A}}(n)$. For every h_i, we have at most $\mathcal{S}_{\mathcal{B}}(n)$ binary vectors are formed by indicator functions of events $x_i \in \mathcal{B}$ (again by definition of growth function). Therefore the maximum number of binary vectors with indicator function of events in $x_i \in \mathcal{A} \cap \mathcal{B}$ with $i \in [n]$ is at most

$$k\mathcal{S}_{\mathcal{B}} \le \mathcal{S}_{\mathcal{A}}(n)\mathcal{S}_{\mathcal{B}}(n).$$

Observe the bound holds for arbitrary $x_1, \dots, x_n$, then we can conclude the proof.