High Dimensional Statistics **2nd Semester, 2021-2022**

Homework 2 (Deadline: May 22)

1. (10 pts) Let $\{e_k\}_{k=1}^n$ be the n canonical basis vectors of $\mathbb{R}^n$. Suppose we would like to project $\{e_k\}_{k=1}^n$ to a lower dimensional space $\mathbb{R}^k$ ($k < n$) such that the distance between every pair of vectors after projection nearly preserved, i.e.,

$$(1+\delta)\|e_i - e_j\|_2 \leq \|T(e_i) - T(e_j)\|_2 \leq (1+\delta)\|e_i - e_j\|_2, \quad \text{for all } i \neq j \tag{1.1}$$

for a small constant $0 < \delta < 1$, where $T : \mathbb{R}^n \to \mathbb{R}^k$ denotes the projection mapping. Show that no matter what the mapping T is, $k \gtrsim \log n$ is necessary for (1.1) to hold, where the constant hidden in $\gtrsim$ can rely on δ.

Hint: The notion of packing number is helpful.

2. (10 pts) Suppose T is a finite set and $\log \mathbb{E}\left[e^{\lambda X_t}\right] \leq \psi(\lambda)$ for all $\lambda \geq 0$ and $t \in T$, where $\psi(\cdot)$ is a convex function and $\psi(0) = \psi'(0) = 0$. Show that

$$\mathbb{E}\left[\max_{t \in T} X_t\right] \leq \psi^{*-1}(\log |T|),$$

where $\psi^*(x) = \sup\limits_{\lambda \geq 0}\{\lambda x - \psi(\lambda)\}$.

3. (5 pts) Let $g \sim N(0,1)$ and ξ be a Rademacher random variable and assume they are independent. Show that g and $\xi|g|$ have the same distribution.

4. (10 pts) Let $X_1, \cdots, X_n$ be independent, mean zero random vectors in a normed space. Show that

$$\frac{1}{2}\mathbb{E}\left[\left\|\sum_{k=1}^n \varepsilon_k X_k\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{k=1}^n X_k\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{k=1}^n \varepsilon_k X_k\right\|\right].$$

Hint: Mimic the proofs for Lemmas 7.1 and 7.2 in Lecture 7.

5. (10 pts) Define

$$T^n(s) = \{x \in \mathbb{R}^n : \ \|x\|_0 \leq s, \ \|x\|_2 \leq 1\},$$

where $\|x\|_0$ counts the number of non-zero entries in x. Show that the Gaussian complexity of $T^n(s)$, denoted $\mathcal{G}(T^n(s)) = \mathbb{E}\left[\sup_{x \in T^n(s)} \langle g, x\rangle\right]$, $g \sim \mathcal{N}(0, I_n)$, satisfies

$$\mathcal{G}(T^n(s)) \lesssim \sqrt{s \log\left(\frac{en}{s}\right)}.$$

6. (10 pts) Let $W \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ entries. We have established an upper bound of $\mathbb{E}\left[\|W\|_2\right]$ using the one-step discretization bound as well as the Gaussian comparison inequality. In this problem you are required to establish an upper bound of $\mathbb{E}\left[\|W\|_2\right]$ using the Dudley integral.

7. (20 pts) Let $B_1^n = \{x : \ \|x\|_1 \leq 1\}$ be the ℓ_1-norm unit ball. We have already seen that the Gaussian complexity of B_1^n satisfies

$$\mathcal{G}(B_1^n) = \mathbb{E}\left[\sup_{\|x\|_1 \leq 1} \langle g, x\rangle\right] \lesssim \sqrt{\log n}, \quad g \sim \mathcal{N}(0, I_n)$$

based on the duality between ℓ_1-norm and ℓ_∞-norm. In this problem, we attempt to provide a bound for $\mathcal{G}(B_1^n)$ based on the Dudley integral.

- For $\varepsilon > 0$ being sufficiently small, show that the covering of B_1^n under the ℓ_2-norm satisfies

$$\sqrt{\log \mathcal{N}(B_1^n, \|\cdot\|_2, \varepsilon)} \lesssim \min\{\varepsilon^{-1}\sqrt{\log n}, \sqrt{n}\cdot\log(1/\varepsilon)\}.$$

  Hint: Volume argument as presented in Lecture 4 may be useful in providing one bound.

- Using the above result and the Dudley integral to provide a bound for $\mathcal{G}(B_1^n)$.

8. (10 pts) Let $\phi_j : \mathbb{R} \to \mathbb{R}$ $(j = 1, \cdots, n)$ be 1-Lipschitz (i.e., $|\phi_j(t) - \phi_j(s)| \leq |t - s|$). Let w_j $(j = 1, \cdots, n)$ be i.i.d $\mathcal{N}(0, 1)$ random variables. For any $T \subset \mathbb{R}^n$, show that

$$\mathbb{E}\left[\sup_{t=(t_1,\cdots,t_n)\in T} \sum_{j=1}^n w_j \phi_j(t_j)\right] \leq \mathbb{E}\left[\sup_{t=(t_1,\cdots,t_n)\in T} \sum_{j=1}^n w_j t_j\right].$$

What does the above result mean in terms of the Gaussian complexity?

9. (10 pts) Let $f_1, \cdots, f_m$ be a set of functions and $x_1, \cdots, x_n$ be a set of real numbers. Suppose

$$\frac{1}{n}\sum_{k=1}^n (f_i(x_k) - f_j(x_k))^2 \geq \varepsilon^2, \quad \forall\, i, j = 1, \cdots, m.$$

Show that there exist $\ell \asymp \varepsilon^{-4}\log m$ points from $\{x_1, \cdots, x_n\}$ (allowing repetitions), denoted $\{x_1', \cdots, x_\ell'\}$, such that

$$\frac{1}{\ell}\sum_{k=1}^\ell (f_i(x_k') - f_j(x_k'))^2 \geq \varepsilon^2/4, \quad \forall\, i, j = 1, \cdots, m.$$