Problem Set 5

- This problem set will be due on **May 15, 2020**. Please email your solutions to the TA.
- Each problem carries 10 points.
- You must work on the problems by yourself. No collaboration allowed.

1. **(Maximum Eigenvalue of Random Matrices)** In this problem we will use Talagrand's inequality to derive *exponential* concentration of the maximum eigenvalue of a random symmetric matrix around its median value.
   For $1 \leq i \leq j \leq n$ let x_{ij} be independent real variables with $|x_{ij}| \leq 1$, and set $x_{ji} = x_{ij}$. Let X be a real $n \times n$ symmetric matrix with entries $\big(x_{ij}\big)$. Let $\lambda_1(X)$ be the maximum Eigenvalue of X and let M be the median of $\lambda_1(X)$. We will show that

   $$\mathbb{P}(|\lambda_1(X) - M| \geq t) \leq 4\exp(-t^2/32).$$

   (a) Define the sets $A = \{X : \lambda(X) \leq M\}$, $B = \{Y : \lambda(Y) \geq M + t\}$. For any $Y \in B$ let v denote its (unit-norm) eigenvector corresponding to the top eigenvalue. Show that for all $X \in A$:

   $$v^T(Y - X)v \geq t.$$

   (b) Show that there exists an a with $||a|| \leq 1$ such that:

   $$\frac{1}{2\sqrt{2}} v^T(Y - X)v \leq d_a(x, y).$$

   (c) Combine part (a) and (b) and conclude the proof by applying Talagrand's inequality to the sets A and B.

2. **(Upper bound on Expected Norms)** Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ that is sub-Gaussian with parameter σ^2 (this means that $\boldsymbol{v}^T\boldsymbol{X}$ is σ^2-sub-Gaussian for all $\boldsymbol{v} \in \mathbb{R}^d$ s.t. $||\boldsymbol{v}||_2 = 1$). Using **Dudley's entropy integral**, find an upper bound for $\mathbb{E}(||\boldsymbol{X}||_2)$.
   Hint: You may use the fact that the δ-covering number of the Euclidean unit ball in $\mathbb{R}^d$ is upper-bounded by $(1 + \frac{2}{\delta})^d$.

3. **(Gaussian Complexity of ℓ_0-"balls")** Sparsity plays an important role in many classes of high-dimensional statistical models. In this problem, we will compute the Gaussian complexity of an s-sparse ℓ_0-ball intersected with a unit ℓ_2-ball. Consider the set

   $$T^d(s) = \{\theta \in \mathbb{R}^d | ||\theta_0|| \leq s, ||\theta||_2 \leq 1.\}$$

corresponding to all s-sparse vectors contained within the Euclidean unit ball. In this problem, we prove that its Gaussian complexity is upper bounded as

$$\mathcal{G}(T^d(s)) \precsim \sqrt{s \log\left(\frac{ed}{s}\right)}. \tag{1}$$

(a) First show that $\mathcal{G}(T^d(s)) = \mathbb{E}\big[\max_{|S|=s} ||w_S||_2\big]$, where $w_S \in \mathbb{R}^{|S|}$ denotes the subvector of $(w_1, w_2, \ldots, w_d)$ indexed by the subset $S \subseteq \{1, 2, \ldots, d\}$. (b) Next show that for any fixed subset S of cardinality s:

$$\mathbb{P}\big[||w_S||_2 \geq \sqrt{s} + \delta\big] \leq e^{-\delta^2/2}.$$

(c) Use the preceding parts to establish the bound (1).

4. **(Gaussian Complexity of Ellipsoids)** (a) For any set $T \subseteq \mathbb{R}^d$, denote its Rademacher complexity and Gaussian complexity by $\mathcal{R}(T)$ and $\mathcal{G}(T)$ respectively. Show that

$$\sqrt{\frac{2}{\pi}}\mathcal{R}(T) \leq \mathcal{G}(T).$$

(b) Recall that the space $\ell^2(\mathbb{N})$ consists of all real sequences $(\theta_j)_{j=1}^{\infty}$ such that $\sum_j \theta_j^2 < \infty$. Given a non-zero sequence $(\mu_j)_{j=1}^{\infty} \in \ell_2(\mathbb{N})$, consider the associated ellipse

$$\mathcal{E} = \left\{(\theta_j)_{j=1}^{\infty} \,\middle|\, \sum_{j=1}^{\infty} \theta_j^2/\mu_j^2 \leq 1\right\}.$$

Ellipses of this form plays an important role in analyzing the statistical properties of reproducing Kernel Hilbert Spaces (RKHS).

Using the result from part (a), prove that the Gaussian complexity $\mathcal{G}(\mathcal{E})$ of $\mathcal{E}$ satisfies the following bounds

$$\sqrt{\frac{2}{\pi}}\left(\sum_{j=1}^{\infty}\mu_j^2\right)^{1/2} \leq \mathcal{G}(\mathcal{E}) \leq \left(\sum_{j=1}^{\infty}\mu_j^2\right)^{1/2}.$$

5. **(Non-parametric Least-Square Estimation)** Consider the function class $S_{\alpha,\gamma}(C_{\max}, L)$ which we introduced in the notes. Recall that,

$$S_{\alpha,\gamma}(C_{\max}, L) = \{f : [0,1] \to \mathbb{R} : |f^{(j)}|_\infty \leq C_{\max}, \forall 0 \leq j \leq \alpha, \text{ and } \\ |f^\alpha(x) - f^\alpha(y)| \leq L|x-y|^\gamma, \forall x, y \in [0,1].\}$$

It can be shown that for some C (which depends on the parameters), the δ-covering number of $S_{\alpha,\gamma}(C_{\max}, L)$ in the sup-norm may be bounded as follows:

$$\log N(\delta, S_{\alpha,\gamma}(C_{\max}, L), ||\cdot||_\infty) \leq C\left(\frac{1}{\delta}\right)^{1/(\alpha+\gamma)}.$$

Suppose we observe

$$Y_i = f^*(x_i) + \epsilon_i, \quad 1 \leq i \leq n,$$

where $f^* \in S_{\alpha,\gamma}(C_{\max}, L)$, and ϵ_i are i.i.d. standard Gaussians and the x_i's are deterministic points in $[0,1]$. Consider the non-parametric least-square estimator

$$\hat{f} \in \arg\min_{f \in S_{\alpha,\gamma}(C_{\max}, L)} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(x_i)\big)^2.$$

Using the notion of Gaussian complexity of the function class $S_{\alpha,\gamma}(C_{\max}, L)$ and **Dudley's entropy integral**, prove an upper-bound for the mean-squared estimation error:

$$\mathsf{MSE} \equiv \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \big(\hat{f}(x_i) - f^*(x_i)\big)^2\right).$$