Prof. Yiping Lu
IEMS 402 Statistical Learning
February 11, 2025

**Homework 7: Complexity of Hypothesis Space**

**Question 1. (Dvoretzky-Kiefer-Wolfowitz inequality via Uniform Bounds)** Let $\mathcal{F} = \{1\{x \leq t\} \mid t \in \mathbb{R}\}$ be the collection of indicator functions for $x \leq t$. Let the $L_2(P)$ metric on $\mathcal{F}$ be defined by $\|f - g\|^2_{L_2(P)} = \int (f(x) - g(x))^2 dP(x)$.

(a) Show that the covering numbers for $\mathcal{F}$ in $L_2(P)$-norm satisfy

$$\sup_P \log N(\mathcal{F}, L_2(P), \epsilon) \leq C \log\left(1 + \frac{1}{\epsilon}\right),$$

where the supremum is over all probability distributions and $C$ is a numerical constant.

(b) Show that $R_n(\mathcal{F}) \leq \frac{C}{\sqrt{n}}$, where $C$ is a universal (numerical) constant.

(c) Prove a (weaker) version of the Dvoretzky-Kiefer-Wolfowitz inequality, that is, that

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |P_n(X \leq t) - P(X \leq t)| \geq \frac{C}{\sqrt{n}} + \epsilon\right) \leq 2e^{-cn\epsilon^2},$$

where $c, C$ are absolute constants. (In fact, $c = 2$ is possible using tools we have covered.)

**Question 2.** Read http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf

**Definition** ($\epsilon$-covering). Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. $\{V_1, \ldots, V_N\}$ is an $\epsilon$-covering of $\Theta$ if $\Theta \subset \bigcup_{i=1}^N \mathcal{B}(V_i, \epsilon)$, or equivalently, $\forall \theta \in \Theta$, $\exists i$ such that $\|\theta - V_i\| \leq \epsilon$.

**Definition** ($\epsilon$-packing). Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. $\{\theta_1, \ldots, \theta_M\}$ is an $\epsilon$-packing of $\Theta$ if $\min_{i \neq j} \|\theta_i - \theta_j\| > \epsilon$ (notice the inequality is strict), or equivalently $\bigcap_{i=1}^M \mathcal{B}(\theta_i, \epsilon/2) = \emptyset$.

**Definition** (Covering number). $N(\Theta, \|\cdot\|, \epsilon) := \min\{n : \exists\, \epsilon\text{-covering over } \Theta \text{ of size } n\}$.

**Definition** (Packing number). $M(\Theta, \|\cdot\|, \epsilon) := \max\{m : \exists\, \epsilon\text{-packing of } \Theta \text{ of size } m\}$.

**Proof**

**(a)** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. Then

$$M(\Theta, \|\cdot\|, 2\epsilon) \overset{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \overset{(b)}{\leq} M(\Theta, \|\cdot\|, \epsilon).$$

**(b)** Prove

$$\frac{\text{vol}(\Theta)}{\text{vol}(B(\epsilon))} \overset{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon) \overset{(b)}{\leq} \frac{\text{vol}(\Theta + B(\frac{\epsilon}{2}))}{\text{vol}(B(\frac{\epsilon}{2}))}.$$

where $B(\epsilon)$ is the norm ball with radius $\epsilon$.

**Question 3. (Covering Number of Sobolev Ellipsoid)** Say that we're interested in functions of the form

$$f = \sum_{j=1}^{\infty} \theta_j \varphi_j : [0, 1] \to \mathbb{R}$$

such that $\sum_{j=1}^{\infty} \theta_j^2 < \infty$ and the $\varphi_j$ form an orthonormal basis for the inner product space, $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ where for any $\psi_i, \psi_j \in \mathcal{F}$,

$$\langle \psi_i, \psi_j \rangle = \int_0^1 \psi_i(x) \psi_j(x)\, dx$$

In other words, for any $\varphi_j, \varphi_k$ in our basis,

$$\int_0^1 \varphi_j(x) \varphi_k(x)\, dx = \mathbb{I}(j = k)$$

We'd like to impose some kind of structure on these functions. One natural way to do so is to assume that "most" of the mass of the sequence is in the early coefficients. Concretely, we may assume that,

$$\sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1,$$

for some sequence of coefficients $\mu_1 \geq \mu_2 \geq \cdots \geq 0$. In particular, consider *Sobolev ellipsoids*, functions for which we have $\mu_j = j^{-2\alpha}$ for some $\alpha > 1/2$. When $\alpha = 1$, these functions should be thought of as a generalization of Lipschitz functions.

Let us return to the task of bounding the metric entropy of the Sobolev ellipsoid. Consider the space of coefficients:

$$\mathcal{E} = \left\{ (\theta_1, \theta_2, \dots) : \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

and

$$\widetilde{\mathcal{E}} = \{\theta \in \mathcal{E} : \theta_j = 0 \text{ for all } j > t\}$$

- *(Bias)* We claim that if we choose $t$ to be the smallest integer such that $\mu_t \leq \delta^2$, then a $\delta$-cover of $\widetilde{\mathcal{E}}$ is a $\sqrt{2}\delta$-cover of $\mathcal{E}$. What is the proper $t$ one should choice?
- *(Covering Number)* The metric entropy is upper bounded as:

$$\log N(\widetilde{\mathcal{E}}; \delta, \|\cdot\|_2) \lesssim \left(\frac{1}{\delta}\right)^{1/\alpha} \log(1/\delta).$$

(*hint*: Using Question 2 and https://www.stat.cmu.edu/~siva/teaching/709/lec3.pdf)

**Question 4.** 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 $Z_1, \dots, Z_n$ are drawn i.i.d. from some distribution $p^*$ and $\sigma_1, \dots, \sigma_n$ are Rademacher variables drawn i.i.d. from $\{-1, 1\}$ with equal probability of $+1$ and $-1$.

**(a)** Let $f : \mathcal{X} \to \mathbb{R}$ 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)** 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 $R_n(\mathcal{F})$.

## References

Northwestern University