# Practice Final - Statistical Learning

Spring 2024 - Yiping

**Name:** ________________________ **NetID:** ________________________

**Please check ☑ your Professor's name:**

- ☐ Professor Yiping Lu

**While you wait, please read and check ☑ the following boxes:**

- ☐ Unless I have extra time with the Moses Center, the time limit is **100 minutes**.
- ☐ I wrote my name and NetID (e.g. ab1234) at the top of this page.
- ☐ I will not detach any pages, especially not the scratch pages at the end.
- ☐ Except for multiple choice questions, I will show my work.
- ☐ If I need more space for an exercise, I will make a note and continue on one of the scratch pages.
- ☐ If I am caught in violation of academic integrity, including but not limited to peaking at another student's work, allowing another student to copy from my work, or speaking with another student, or using unauthorized resources, I will be asked to leave the exam and get a zero.


**Do not start the exam until you are permitted to.**

# Exercise I

Recall that the Rademacher complexity of a class of functions $\mathscr{F}$ is defined as

$$R_n(\mathscr{F}) = \mathbb{E}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right],$$

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

**(a)** Let $f : \mathscr{X} \to \mathbb{R}$ be a function, and let $\mathscr{F} := \{-f, f\}$ be a function class containing only two functions. Upper bound $R_n(\mathscr{F})$ 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:

$$\mathscr{F} = \{x \mapsto w\cdot x : \|w\|_\infty \le B\}.$$

Compute an upper bound on the Rademacher complexity $R_n(\mathscr{F})$. Express your answer as a function of $B$, $k$, $d$, and $n$. Note that this allows us to effectively control the complexity of learning using $L_\infty$ regularization.

**(c)** Consider a prediction problem from $x \in \mathbb{R}$ to $y \in \{0,\ldots,k\}$. For every parameter vector $\theta \in \mathbb{R}^k$, define the prediction function $h_\theta(x) = \sum_{i=1}^{k}\mathbb{I}\{x \ge \theta_i\}$ (monotonically increasing piecewise constant functions). Define the loss function to be $\ell(y,p) = |y-p|$, yielding the following loss class:

$$\mathscr{A} = \{(x,y) \mapsto \ell(y, h_\theta(x)) : \theta \in \mathbb{R}^k\}.$$

Compute an upper bound on the Rademacher complexity of $\mathscr{A}$.

**(d)** Let $\mathscr{F}$ be the class of all continuous functions $f : [0,1] \to [0,1]$ with at most $k$ local maxima. Find an upper bound of the Rademacher complexity of $\mathscr{F}$.

**(e)** Let $X_i$ be independent with support $\{x \in \mathbb{R}^d : \|x\|_2 \le M\}$. Let $\mathscr{F}$ be functions of the form $x \mapsto \langle\theta, x\rangle$ for $\theta \in \Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le r\}$. Give an upper bound on $R_n(\mathscr{F})$.

**(f)** Suppose $k$ is a bounded kernel with $\sup_x \sqrt{k(x,x)} = B < \infty$ and let $\mathscr{F}$ be its RKHS. Let $M > 0$ be fixed. Then for any $S = (X_1,\ldots,X_n)$,

$$\widehat{\mathscr{R}}_S(B_k(M)) \le \frac{MB}{\sqrt{n}}$$

where $B_k(M) = \{f \in \mathscr{F} \mid \|f\|_{\mathscr{F}} \le M\}$.

**Blank page ofr Exercise I.**

## Exercise II

**(a)** For function class
$$\mathscr{F} = \{f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ is } L\text{-Lipschitz}\},$$
show that $\log N(\epsilon, \mathscr{F}, \|\cdot\|_\infty) \lesssim \frac{L}{\epsilon}$.

**(b)** Show the covering number estimation for Sobolev Ellipsoid.

**(c)** Using the Covering Number Bound to show the bound on Rademacher Complexity

**(d)** How does the results informs bounds for non-parametric least square regression? **hint**: using localized complexity

**Blank page ofr Exercise II.**

## Exercise III (Hilbert Embedding of Probability)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with associated RKHS $\mathcal{H}$. Assume that $\mathcal{X}$ is compact. We call $k$ *universal* if it is dense in $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$. That is, for any $\epsilon > 0$ and any continuous function $f : \mathcal{X} \to \mathbb{R}$, there exists a function $h \in \mathcal{H}$ such that $\sup_{x\in\mathcal{X}} |f(x) - h(x)| < \epsilon$.

Define $\varphi(x) = k(\cdot, x)$. (Thus $k(x,z) = \langle \varphi(x), \varphi(z)\rangle$, and $\varphi(x)$ is the representer of evaluation at $x$, i.e., $\langle h, \varphi(x)\rangle = h(x)$ for all $h \in \mathcal{H}$.) Let $\mathcal{P}$ be the collection of distributions on $\mathcal{X}$ for which $\mathbb{E}_P[\sqrt{k(X,X)}] < \infty$.

(a) Using the Riesz representation theorem for Hilbert spaces, argue that the mean mapping $\mu(P) := \mathbb{E}_P[\varphi(X)]$ exists and is a vector in $\mathcal{H}$. *Hint*: Letting $\|\cdot\|$ denote the norm on $\mathcal{H}$, the Riesz representation theorem for Hilbert spaces says that if $L : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional, meaning that $L(f) \le C \cdot \|f\|$ for some constant $C$, then there exists some $h_L \in \mathcal{H}$ such that $L(f) = \langle h_L, f\rangle$ for all $f \in \mathcal{H}$.

(b) Assume that $\mathcal{X}$ is compact and that $k$ is universal. Show that the mean embedding

$$P \mapsto \mathbb{E}_P[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) dP(x)$$

is one-to-one, that is, if $P \ne Q$ then $\mathbb{E}_P[\varphi(X)] \ne \mathbb{E}_Q[\varphi(X)]$.

(c) For distributions $P$ and $Q$, show that

$$\sup_{f\in\mathcal{H}, \|f\|\le 1} \left\{\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\right\} = \sqrt{\mathbb{E}[k(X,X')] + \mathbb{E}[k(Z,Z')] - 2\mathbb{E}[k(X,Z)]},$$

where $X, X' \overset{i.i.d}{\sim} P$ and $Z, Z' \overset{i.i.d}{\sim} Q$.

## Exercise IV (Example of Kernel)

- Let $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a valid kernel function. Define

$$k_{\text{norm}}(x,z) := \frac{k(x,z)}{\sqrt{k(x,x)}\sqrt{k(z,z)}}.$$

Is $k_{\text{norm}}$ a valid kernel? Justify your answer.

- Consider the class of functions

$$\mathscr{H} := \{f : f(0) = 0, f' \in L^2([0,1])\},$$

that is, functions $f : [0,1] \to \mathbb{R}$ with $f(0) = 0$ that are almost everywhere differentiable, where

$$\int_0^1 (f'(x))^2 dx < \infty.$$

On this space of functions, we define the inner product by

$$\langle f, g \rangle = \int_0^1 f'(x) g'(x) dx.$$

Show that $k(x,z) = \min\{x,z\}$ is the reproducing kernel for $\mathscr{H}$, so that it is (i) positive semidefinite and (ii) a valid kernel.

(*My understanding:* By integral by parts, we have $\langle f, g \rangle_{\mathscr{H}} = \langle f, \Delta g \rangle_{\mathscr{L}_2}$ and $\Delta k(\cdot, z) = \delta_z$.)

- Consider the Sobolev space $\mathscr{F}_k$, which is defined as the set of functions that are $(k-1)$-times differentiable and have $k$th derivative almost everywhere on $[0,1]$, where the $k$th derivative is square-integrable. That is, we define

$$\mathscr{F}_k := \left\{ f : [0,1] \mid f^{(k)}(x) \in L^2([0,1]) \right\}.$$

We define the inner product on $\mathscr{F}_k$ by

$$\langle f, g \rangle = \sum_{i=0}^{k-1} f^{(i)}(x) g^{(i)}(x) + \int_0^1 f^{(k)}(x) g^{(k)}(x)\, dx.$$

(a) Find the representer of evaluation for this Hilbert space, that is, find a function $r_x : [0,1] \to \mathbb{R}$ (defined for each $x \in [0,1]$) such that $r_x \in \mathscr{F}_k$ and

$$\langle r_x, f \rangle = f(x)$$

for all $x$.

(b) What is the reproducing kernel $k(x,z)$ associated with this space? (Recall that $k(x,z) = \langle r_x, r_z \rangle$ for an RKHS.)

**Blank page for Exercise III.**

## Exercise V

Explain: Importance weighting, DRO, why localized complexity is better, what is in-context learning, duality of optimal transport

**Blank scratch page.**

**Blank scratch page.**