Prof. Yiping Lu
IEMS 402 Statistical Learning
November 3, 2024

**Homework 8: Reproducing Kernel Hilbert Space/Robust Learning**

**Question 1. (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 $P \neq Q$ then $\mathbb{E}_P[\varphi(X)] \neq \mathbb{E}_Q[\varphi(X)]$.

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

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

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

**Question 2. (Example of Kernel)**

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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

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

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 $\mathcal{H}$, so that it is (i) positive semidefinite and (ii) a valid kernel.

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

- Consider the Sobolev space $\mathcal{F}_k$, 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

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

We define the inner product on $\mathcal{F}_k$ 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 $r_x \in \mathcal{F}_k$ 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.)

**Question 3. ($\varphi$-divergence DRO and Variance Regularization)** Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $\varphi(1) = 0$. Then the $\varphi$-divergence between distributions $P$ and $Q$ defined on a space $\mathcal{X}$ is

$$D_\varphi(P\|Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ = \int_{\mathcal{X}} \varphi\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x),$$

where $\mu$ is any measure for which $P, Q \ll \mu$, and $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$. Throughout this paper, we use $\varphi(t) = \frac{1}{2}(t-1)^2$, which gives the $\chi^2$-divergence [45]. Given $\varphi$ and a sample $X_1, \ldots, X_n$, we define the local neighborhood of the empirical distribution with radius $\rho$ by

$$\mathcal{P}_n := \left\{\text{distributions } P \text{ such that } D_\varphi\left(P\|\hat{P}_n\right) \le \frac{\rho}{n}\right\},$$

where $\hat{P}_n$ denotes the empirical distribution of the sample, and our choice of $\varphi(t) = \frac{1}{2}(t-1)^2$ means that $\mathcal{P}_n$ consists of discrete distributions supported on the sample $\{X_i\}_{i=1}^n$. We then define the robustly regularized risk

$$R_n(\theta, \mathcal{P}_n) := \sup_{P\in\mathcal{P}_n} \mathbb{E}_P[\ell(\theta, X)] = \sup_P \left\{\mathbb{E}_P[\ell(\theta, X)] : D_\varphi(P\|\hat{P}_n) \le \frac{\rho}{n}\right\}.$$

Using convex duality please show that

$$R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] + \sqrt{\frac{2\rho}{n}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2]}.$$

You can assume strong duality holds.

*Further Reading*: Connection between adversarial training and Wasserstein DRO https://arxiv.org/abs/1710.10571

**Question 4.** (**Derive the dual formulation of the Sinkhorn distance.**) Given two probability vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and a cost matrix $C \in \mathbb{R}^{n\times n}$, the Sinkhorn distance introduces an entropy regularization to the optimal transport problem, *i.e.* the sinkhorn distance is defined as

$$\begin{aligned} \min_{\gamma\in\mathbb{R}^{n\times n}} \quad & \langle \gamma, C\rangle - \epsilon H(\gamma) \\ \text{subject to} \quad & \gamma\mathbf{1} = \mathbf{a}, \qquad \gamma^\top \mathbf{1} = \mathbf{b}, \\ & \gamma \ge 0. \end{aligned}$$

**(a)** Starting from the primal formulation of the entropy-regularized optimal transport problem (Sinkhorn distance), derive its dual form

$$\max_{\mathbf{u},\mathbf{v}\in\mathbb{R}^n} \quad \mathbf{u}^\top \mathbf{a} + \mathbf{v}^\top \mathbf{b} - \epsilon \sum_{i,j} \exp\left(\frac{u_i + v_j - C_{i,j}}{\epsilon}\right).$$

**(b)** Once you know the optimal $u^*$, can you write down the closed-form solution of $v^*$ in terms of $u^*$? (What is the computational cost? [*0pt bouns*])

(*hint*: https://arxiv.org/abs/1306.0895)