IEMS 402 Statistical Learning - 2025 Winter
HW8

Tuorui Peng[1]

Exercise 1 Hilbert Embedding of Probability

1.(a)

Consider the functional $L: f \mapsto \mathbb{E}\left[f(X)\right]$ which is a bounded linear functional. By Riesz Representation Theorem, there exists a unique $h_L \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H}: \left\langle h_L, f\right\rangle = L(f) = \mathbb{E}_P\left[f(X)\right] = \mathbb{E}_P\left[\left\langle \varphi(X), f\right\rangle\right] = \left\langle \mathbb{E}_P\left[\varphi(X)\right], f\right\rangle$$

i.e. such $\mathcal{H} \ni h_L = \mathbb{E}_P\left[\varphi(X)\right]$.

1.(b)

We prove the contrapositive. Suppose $\mathbb{E}_P\left[\varphi(X)\right] = \mathbb{E}_Q\left[\varphi(X)\right]$, then we consider the following setting: $\forall \varepsilon > 0$, $\forall f \in \mathcal{X}$, $\exists h_{f,\varepsilon} \in \mathcal{H}$ s.t. $\left\|f - h_{f,\varepsilon}\right\|_\infty < \varepsilon$, and we have

$$\begin{aligned}
\mathbb{E}_P\left[h_{f,\varepsilon}(X)\right] =& \mathbb{E}_P\left[\left\langle h_{f,\varepsilon}, \varphi(X)\right\rangle\right] \\
=& \left\langle \mathbb{E}_P\left[\varphi(X)\right], h_{f,\varepsilon}\right\rangle \\
\mathbb{E}_Q\left[h_{f,\varepsilon}(X)\right] =& \mathbb{E}_Q\left[\left\langle h_{f,\varepsilon}, \varphi(X)\right\rangle\right] \\
=& \left\langle \mathbb{E}_Q\left[\varphi(X)\right], h_{f,\varepsilon}\right\rangle
\end{aligned}$$

So we have

$$\begin{aligned}
\left|\mathbb{E}_P\left[f(X)\right] - \mathbb{E}_Q\left[f(X)\right]\right| \leq& 2\varepsilon + \left|\mathbb{E}_P\left[h_{f,\varepsilon}(X)\right] - \mathbb{E}_Q\left[h_{f,\varepsilon}(X)\right]\right| \\
=& 2\varepsilon + \left|\left\langle \mathbb{E}_P\left[\varphi(X)\right] - \mathbb{E}_Q\left[\varphi(X)\right], h_{f,\varepsilon}\right\rangle\right| \\
=& 2\varepsilon
\end{aligned}$$

Note that the above statement is true $\forall \varepsilon > 0$, $\forall f \in \mathcal{C}$, thus proves the contrapositive that $P \overset{\mathrm{d}}{=} Q$ must hold. And we have if $P \overset{\mathrm{d}}{\neq} Q$, then $\mathbb{E}_P\left[\varphi(X)\right] \neq \mathbb{E}_Q\left[\varphi(X)\right]$.

1.(c)

We have for right hand side:

$$\begin{aligned}
\text{R.H.S.} =& \sqrt{\mathbb{E}\left[k(X, X')\right] + \mathbb{E}\left[k(Z, Z')\right] - 2\mathbb{E}\left[k(X, Z)\right]} \\
=& \sqrt{\left\langle \varphi(X), \varphi(X')\right\rangle + \left\langle \varphi(Z), \varphi(Z')\right\rangle - 2\left\langle \varphi(X), \varphi(Z)\right\rangle} \\
=& \sqrt{\left\langle \varphi(X) - \varphi(X'), \varphi(Z) - \varphi(Z')\right\rangle} \\
:=& E
\end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

For left hand side:

$$\sup_{f\in\mathcal{H},\|f\|\leq 1} \left|\mathbb{E}_P\left[f(X)\right]-\mathbb{E}_Q\left[f(X)\right]\right| = \sup_{f\in\mathcal{H},\|f\|\leq 1} \left|\langle\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],f\rangle\rangle\right|$$

which reach maximum when $f\propto\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]:=\alpha\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]$, in which α is taken to make $\|f\|=1$. Thus we have

$$1=\alpha^2\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]\right\rangle=\alpha^2E^2\Rightarrow\alpha=\frac{1}{E}$$

Substitute back to the left hand side, we have

$$\begin{aligned}
\text{L.H.S.}=&\sup_{f\in\mathcal{H},\|f\|\leq 1} \left|\mathbb{E}_P\left[f(X)\right]-\mathbb{E}_Q\left[f(X)\right]\right|\\
=&\left|\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],\frac{1}{E}\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]\right\rangle\right|\\
=&\frac{1}{E}\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]\right\rangle\\
=&\frac{1}{E}E=E=\text{R.H.S.}
\end{aligned}$$

Exercise 2 Example of Kernel

2.(a)

We can verify the condition for k_{norm} to be a kernel function by checking the positive semi-definiteness and symmetry easily:

$$\begin{aligned}
k_{\text{norm}}(x,z)=&k_{\text{norm}}(z,x)\\
\forall x_1^m,\,\alpha_1^n,\,\sum_{i,j=1}^n k_{\text{norm}}(x_i,x_j)\alpha_i\alpha_j=&\sum_{i,j=1}^n k(x_i,z_i)\frac{\alpha_i}{\sqrt{k}(x_i,x_i)}\frac{\alpha_j}{\sqrt{k(x_j,x_j)}}\geq 0
\end{aligned}$$

2.(b)

We prove the reproducing property by checking the following:

$$\begin{aligned}
\forall f,\,\forall x\,:\,\langle k(x,\,\cdot\,),f\rangle=&\int_0^1 k'(x,z)f'(z)\,\mathrm{d}z\\
=&\int_0^1 \mathbf{1}_{[0,x]}(z)\mathbf{1}_{[0,x]}(z)f'(z)\,\mathrm{d}z\\
=&\int_0^x f'(z)\,\mathrm{d}z\\
=&f(x)
\end{aligned}$$

And we can easily verify the symmetry and positive semi-definiteness of $k(\,\cdot\,,\,\cdot\,) = \,\cdot\, \wedge \,\cdot\,$ as follows:

$$\begin{aligned}
k(x,z) =& x\wedge z = z\wedge x = k(z,x)\\
\forall g:\int_0^1\int_0^1 g(x)g(z)k(x,z)\,\mathrm{d}x\,\mathrm{d}z =& \int_0^1\int_0^1 g(x)g(z)\left\langle \mathbf{1}_{[0,x]},\mathbf{1}_{[0,z]}\right\rangle\,\mathrm{d}x\,\mathrm{d}z\\
=& \int_0^1\int_0^1 \left\langle g(x)\mathbf{1}_{[0,x]},g(z)\mathbf{1}_{[0,z]}\right\rangle\,\mathrm{d}x\,\mathrm{d}z\\
=& \left\|\int_0^1 g(x)\mathbf{1}_{[0,x]}\,\mathrm{d}x\right\|^2 \geq 0.
\end{aligned}$$

2.(c)

WLOG take $f^{(i)}(0)=0$ for all $i\leq k-1$. So that $\langle f,g\rangle = \int_0^1 f^{(k)}(x)g^{(k)}(x)\,\mathrm{d}x$.

Using similar integration by parts idea, we should have: for each given x, the function $k(x,\,\cdot\,)$ satisfies

$$\begin{aligned}
g(x) &= \langle k(x,\,\cdot\,),g\rangle\\
&= \int_0^1 k^{(k)}(x,z)g^{(k)}(z)\,\mathrm{d}z
\end{aligned}$$

By the Taylor expansion of g at 0 with integraion remainders, i.e.

$$g(x) = \int_0^x \frac{g^{(k)}(z)}{(k-1)!}(x-z)^{k-1}\,\mathrm{d}z$$

we have

$$k^{(k)}(x,z) = \frac{(x-z)^{k-1}}{(k-1)!}\mathbf{1}_{[0,x]}(z) = \frac{(x-z)_+^{k-1}}{(k-1)!}$$

and thus

$$\begin{aligned}
k(x,z) &= \langle k(x,\,\cdot\,),k(z,\,\cdot\,)\rangle\\
&= \int_0^1 k^{(k)}(x,u)k^{(k)}(z,u)\,\mathrm{d}u\\
&= \int_0^1 \frac{(x-u)_+^{k-1}}{(k-1)!}\frac{(z-u)_+^{k-1}}{(k-1)!}\,\mathrm{d}u
\end{aligned}$$

Exercise 3 φ-divergence DRO and Variance Regularization

Optimzation problem is formalized as:

$$\sup_{P\in\mathcal{P}_n}\mathbb{E}_P\left[\ell(\theta,X)\right],\quad s.t.\, D_\varphi(P\|\hat{P}_n)\leq\frac{\rho}{n}$$

Since the empirical distribution $\hat{P}_n$ is a Dirac measure, the optimizer would also be a Dirac measure supported on $\mathrm{supp}(X_1^n)$. We denoted this PMF as:

$$P:\, X=X_i,\quad w.p.\, p_i$$

So the optimization problem can be reformulated as:

$$\sup_{p_i\geq 0,\,\sum_{i=1}^n p_i=1} \sum_{i=1}^n p_i\ell(\theta,X_i),\quad s.t.\, D_\varphi(P\|\hat{P}_n)\leq \frac{\rho}{n}$$

Lagrangian:

$$\mathcal{L}(\vec{p};\lambda,\mu)=\sum_{i=1}^n p_i\ell(\theta,X_i)+\lambda\left(\sum_{i=1}^n np_i^2-1-\frac{2\rho}{n}\right)+\mu\left(\sum_{i=1}^n p_i-1\right),\quad \lambda\leq 0$$

which is maximized w.r.t. $\vec{p}$ when

$$p_i^*=-\frac{\ell(X_i,\theta)+\mu}{2\lambda n}$$

and gives dual problem:

$$\theta_D(\lambda,\mu)=-\sum_{i=1}^n\frac{(\ell(X_i,\theta)+\mu)^2}{4\lambda n}-\lambda(1+\frac{2\rho}{n})-\mu,\quad \lambda\geq 0$$

Which is minimized w.r.t. λ,μ:

$$0=\begin{cases}\dfrac{\partial}{\partial\lambda}\theta_D= & \sum_{i=1}^n\dfrac{(\ell(X_i,\theta)+\mu)^2}{4\lambda^2 n}-(1+\dfrac{2\rho}{n})\\ \dfrac{\partial}{\partial\mu}\theta_D= & -\sum_{i=1}^n\dfrac{(\ell(X_i,\theta)+\mu)}{2\lambda n}-1\end{cases}\Rightarrow\begin{cases}\mu=\sqrt{\dfrac{var_{\hat{P}_n}(\ell(X,\theta))}{2\rho/n}}-\mathbb{E}_{\hat{P}_n}\left[\ell(X,\theta)\right]\\ \lambda=-\dfrac{1}{2}\sqrt{\dfrac{var_{\hat{P}_n}(\ell(X,\theta))}{2\rho/n}}\end{cases}$$

and gives (with strong duality):

$$\begin{aligned}R_n(\theta,\mathcal{P}_n)=\sup_{P\in\mathcal{P}_n}\mathbb{E}_P\left[\ell(\theta,X)\right]&=\inf_{\lambda\leq 0,\mu}\theta_D(\lambda,\mu)\\ &=\mathbb{E}_{\hat{P}_n}\left[\ell(X,\theta)\right]+\sqrt{\frac{2\rho\mathrm{var}_{\hat{P}_n}(\ell(X,\theta))}{n}}\end{aligned}$$

And we revisit the solution p_i^*:

$$p_i^*=-\frac{\ell(X_i,\theta)+\mu}{2\lambda n}=\frac{1}{n}+\frac{\sqrt{2\rho}(\ell(X_i,\theta)-\mathbb{E}_{\hat{P}_n}\left[\ell(X,\theta)\right])}{n\sqrt{n}\sqrt{\mathrm{var}_{\hat{P}_n}(\ell(X,\theta))}}$$

could satisfy $p_i^*\geq 0$ if the empirical variance is large enough & ρ is chosen small enough.

Exercise 4 Derive the dual formulation of the Sinkhorn distance

4.(a)

The Lagrangian is given by:

$$\mathcal{L}(\gamma;u,v)=\langle\gamma,C\rangle-\varepsilon H(\gamma)+u'(\gamma\mathbf{1}-a)+v'(\gamma^T\mathbf{1}-b)$$

miimizing w.r.t. γ gives the optimality condition:

$$\frac{\partial\mathcal{L}}{\partial\gamma_{ij}}=C_{ij}+\varepsilon(\log\gamma_{ij}+1)+u_i+v_j=0\Rightarrow\gamma_{ij}=\exp\left(-\frac{C_{ij}+u_i+v_j}{\varepsilon}-1\right)$$

And we have dual problem:

$$\begin{aligned}
\inf_{u,v} \mathcal{L}(\gamma; u, v) =& \inf_{u,v} \langle \gamma, C \rangle - \varepsilon H(\gamma) + u'(\gamma \mathbf{1} - a) + v'(\gamma^T \mathbf{1} - b) \\
=& \inf_{u,v} \quad -u'a - v'b - \varepsilon \sum_{i,j} \exp\left(-\frac{C_{ij} + u_i + v_j}{\varepsilon} - 1 \right)
\end{aligned}$$

with re-parametrization $u \mapsto -u - \varepsilon, \quad v \mapsto -v$ and omitting constant, we have dual problem:

$$\inf_{u,v} \quad u'a + v'b - \varepsilon \sum_{i,j} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon} \right)$$

4.(b)

Once the optimal u^* is known, it's left to solve for v^*:

$$\begin{aligned}
&\inf_{v} \quad v'b - \varepsilon \sum_{i,j} \exp\left(\frac{u_i^* + v_j - C_{ij}}{\varepsilon} \right) v \\
\Rightarrow & v_j^* = \frac{\varepsilon}{b_j} \sum_{i} \exp\left(\frac{u_i^* + v_j^* - C_{ij}}{\varepsilon} \right) \Rightarrow v^*, \quad \forall j
\end{aligned}$$

which can be solved by fixed point iteration.