36-708 Statistical Methods for Machine Learning Homework #1 Solutions

February 1, 2019

Problem 1 [15 pts.]

Let $X_1, \ldots, X_n \sim P$ where $X_i \in [0,1]$ and P has density p. Let $\widehat{p}$ be the histogram estimator using m bins. Let $h = 1/m$. Recall that the L_2 error is $\int (\widehat{p}(x) - p(x))^2 = \int \widehat{p}^2(x)dx - 2\int \widehat{p}(x)p(x)dx + \int p^2(x)dx$. As usual, we may ignore the last term so we define the loss to be

$$L(h) = \int \widehat{p}^2(x)dx - 2\int \widehat{p}(x)p(x)dx.$$

(a) Suppose we used the direct estimator of the loss, namely, we replace the integral with the average to get

$$\widehat{L}(h) = \int \widehat{p}^2(x)dx - \frac{2}{n}\sum_i \widehat{p}(X_i).$$

Show that this fails in the sense that it is minimized by taking $h = 0$.

(b) Recall that the leave-one-out estimator of the risk is

$$\widehat{L}(h) = \int \widehat{p}^2(x)dx - \frac{2}{n}\sum_i \widehat{p}_{-(i)}(X_i),$$

Show that

$$\widehat{L}(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_j Z_j^2$$

where Z_j is the number of observations in bin j.

Solution.

Define

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \in B_j) \quad \text{and} \quad Z_j = n\widehat{\theta}_j$$

for $j = 1, \ldots, m$.

(a) **(7 pts.)**

$$\begin{aligned}
\widehat{L}(h) &= \int_0^1 \widehat{p}^2(x)dx - \frac{2}{n}\sum_i \widehat{p}(X_i) \\
&= \int_0^1 \left(\sum_{j=1}^m \frac{\widehat{\theta}_j}{h}\mathbb{1}(x \in B_j)\right)^2 dx - \frac{2}{n}\sum_{i=1}^n\sum_{j=1}^m \frac{\widehat{\theta}_j}{h}\mathbb{1}(X_i \in B_j) \\
&= \frac{1}{h^2}\int_0^1 \left(\sum_{k=1}^m\sum_{j=1}^m \widehat{\theta}_j\widehat{\theta}_k\mathbb{1}(x \in B_j \cap B_k)\right)dx - \frac{2}{h}\sum_{j=1}^m \widehat{\theta}_j \cdot \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \in B_j) \\
&= \frac{1}{h^2}\int_0^1 \left(\sum_{j=1}^m \widehat{\theta}_j^2\mathbb{1}(x \in B_j)\right)dx - \frac{2}{h}\sum_{j=1}^m \widehat{\theta}_j^2 \\
&= \frac{1}{h^2}\sum_{j=1}^m \widehat{\theta}_j^2 \int_0^1 \mathbb{1}(x \in B_j)dx - \frac{2}{h}\sum_{j=1}^m \widehat{\theta}_j^2 \\
&= \frac{1}{h}\sum_{j=1}^m \widehat{\theta}_j^2 - \frac{2}{h}\sum_{j=1}^m \widehat{\theta}_j^2 \\
&= -\frac{1}{h}\sum_{j=1}^m \widehat{\theta}_j^2 \\
&= -\frac{1}{hn^2}\sum_{j=1}^m Z_j^2
\end{aligned}$$

Considering the last quantity , we have that:

$$\sum_{j=1}^m Z_j^2 \geq \sum_{j=1}^m Z_j = n \quad , \quad \sum_{j=1}^m Z_j^2 = \sum_{j=1}^m Z_jZ_j \leq \sum_{j=1}^m Z_jn = n^2$$
$$\implies -\frac{1}{h} \leq \widehat{L}(h) \leq -\frac{1}{nh}$$

So $\widehat{L}(h) \to -\infty$ as $h \to 0$. Therefore, this loss is minimized by taking $h = 0$.

(b) **(8 pts.)**

From part (a) we have

$$\int \widehat{p}^2(x)dx = \frac{1}{h}\sum_{j=1}^m \widehat{\theta}_j^2. \tag{1}$$

And the second term in the leave-one-out loss is

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^n \widehat{p}_{(-i)}(X_i) &= \frac{2}{n(n-1)h}\sum_{j=1}^m\sum_{i=1}^n \mathbb{1}(X_i \in B_j)\sum_{k\neq i}\mathbb{1}(X_k \in B_j) \\
&= \frac{2}{n(n-1)h}\sum_{j=1}^m\sum_{i=1}^n \mathbb{1}(X_i \in B_j)(n\widehat{\theta}_j - \mathbb{1}(X_i \in B_j)) \\
&= \frac{2}{n(n-1)h}\sum_{j=1}^m (n^2\widehat{\theta}_j^2 - n\widehat{\theta}_j). \qquad (2)
\end{aligned}$$

Taking the difference of (1) and (2), we get

$$\begin{aligned}\widehat{L}(h) &= \frac{1}{h}\sum_{j=1}^{m}\widehat{\theta}_j^2 - \frac{2}{n(n-1)h}\sum_{j=1}^{m}(n^2\widehat{\theta}_j^2 - n\widehat{\theta}_j) \\ &= \frac{2}{(n-1)h}\underbrace{\sum_{j=1}^{m}\widehat{\theta}_j}_{=1} + \sum_{j=1}^{m}\widehat{\theta}_j^2\left(\frac{1}{h} - \frac{2n}{(n-1)h}\right) \\ &= \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\widehat{\theta}_j^2 \\ &= \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_{j=1}^{m}Z_j^2.\end{aligned}$$

Problem 2 [15 pts.]

Let $\widehat{p}_h$ be the kernel density estimator (in one dimension) with bandwidth $h = h_n$. Let $s_n^2(x) = \text{Var}(\widehat{p}_h(x))$.

(a) Show that

$$\frac{\widehat{p}_h(x) - p(x)}{s_n(x)} \rightsquigarrow N(0,1)$$

where $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$.

Hint: Recall that the Lyapunov central limit theorem says the following: Suppose that $Y_1, Y_2, \ldots$ are independent. Let $\mu_i = \mathbb{E}[Y_i]$ and $\sigma_i^2 = \text{Var}(Y_i)$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If

$$\lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}[|Y_i - \mu_i|^{2+\delta}] = 0$$

for some $\delta > 0$. Then $s_n^{-1}\sum_i(Y_i - \mu_u) \rightsquigarrow N(0,1)$.

(b) Assume that the smoothness is $\beta = 2$. Suppose that the bandwidth h_n is chosen optimally. Show that

$$\frac{\widehat{p}_h(x) - p(x)}{s_n(x)} \rightsquigarrow N(b(x),1)$$

for some constant $b(x)$ which is, in general, not 0.

Solution.

(a) **[8 pts.]**

Caveat: The classical Central Limit Theorem cannot be applied here, as $h = h_n$ is a function of n and thus the $K\left(\frac{\|x-X_i\|}{h}\right)$ are not identically distributed. However, as the hint suggests, the Lyapunov CLT still holds for non-identically distributed random variables.

Claim. Let $p > 1$. Then

$$\mathbb{E}\left[\left|\frac{1}{h}K\left(\frac{\|x - X_i\|}{h}\right) - p_h(x)\right|^p\right] = \Theta\left(\frac{1}{h^{p-1}}\right).$$

Proof. See appendix

Now

$$\mathbb{E}\left[\left|\frac{1}{nh}K\left(\frac{\|x-X_i\|}{h}\right)-\frac{p_h(x)}{n}\right|^{2+\delta}\right]=\frac{1}{n^{2+\delta}}\mathbb{E}\left[\left|\frac{1}{h}K\left(\frac{\|x-X_i\|}{h}\right)-p_h(x)\right|^{2+\delta}\right]$$
$$=\Theta\left(\frac{1}{n^{2+\delta}h^{1+\delta}}\right),$$

and

$$s_n^2=\sum_{i=1}^n\mathbb{E}\left[\left|\frac{1}{nh}K\left(\frac{\|x-X_i\|}{h}\right)-\frac{p_h(x)}{n}\right|^2\right]$$
$$=\frac{1}{n}\mathbb{E}\left[\left|\frac{1}{h}K\left(\frac{\|x-X_i\|}{h}\right)-p_h(x)\right|^2\right]$$
$$=\Theta\left(\frac{1}{nh}\right).$$

Therefore,

$$\frac{1}{s_n^{2+\delta}}\sum_{i=1}^n\mathbb{E}\left[\left|\frac{1}{nh}K\left(\frac{\|x-X_i\|}{h}\right)-\frac{p_h(x)}{n}\right|^{2+\delta}\right]=\Theta\left((nh)^{1+\frac{\delta}{2}}\right)\cdot n\cdot\Theta\left(\frac{1}{n^{2+\delta}h^{1+\delta}}\right)$$
$$=\Theta\left((nh)^{-\frac{\delta}{2}}\right)$$
$$\to 0,$$

as $n\to\infty$ and $nh\to\infty$, for any $\delta>0$. So, by the Lyapunov CLT,

$$\frac{\widehat{p}_h(x)-p_h(x)}{s_n(x)}\rightsquigarrow N(0,1).$$

(b) **[7 pts.]** First note

$$\frac{\widehat{p}_h(x)-p(x)}{s_n(x)}=\frac{\widehat{p}_h(x)-p_h(x)}{s_n(x)}+\frac{p_h(x)-p(x)}{s_n(x)}$$
$$=\frac{\widehat{p}_h(x)-p_h(x)}{s_n(x)}+\frac{\text{Bias}(p_h(x))}{\sqrt{\text{Var}(\widehat{p}_h(x))}}.$$

From Theorem 5, the optimal bandwidth is $h_n=\Theta(n^{-1/5})$.

Now from part (a), we have

$$\text{Var}(\widehat{p}_h(x))=\Theta\left(\frac{1}{nh}\right)$$

and from Lemma 3,

$$\text{Bias}(p_h(x))=O(h^2).$$

Therefore,

$$\begin{aligned}
\frac{\widehat{p}_h(x) - p(x)}{s_n(x)} &= \frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{\text{Bias}(p_h(x))}{\sqrt{\text{Var}(\widehat{p}_h(x))}} \\
&= \frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{O(h^2)}{\Theta\left(\frac{1}{(nh)^{1/2}}\right)} \\
&= \frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{O(n^{-2/5})}{\Theta\left(n^{-2/5}\right)} \\
&= \underbrace{\frac{\widehat{p}_h(x) - p_h(x)}{s_n(x)}}_{\rightsquigarrow N(0,1)} + O(1) \\
&\rightsquigarrow N(b(x), 1).
\end{aligned}$$

Problem 3 [10 pts.]

Let $X_1, \dots, X_n \sim P$ where $X_i \in [0,1]$. Assume that P has density p which has bounded continuous derivative. Let $\widehat{p}_h(x)$ be the kernel density estimator. Show that, in general, the bias is of order $O(h)$ at the boundary. That is, show that $\mathbb{E}[\widehat{p}_h(0)] - p(0) = Ch$ for some $C > 0$.

Solution.

$$\begin{aligned}
\mathbb{E}\big[\widehat{p}(0)\big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{-X_i}{h}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{h}K\left(\frac{X_i}{h}\right)\right] \\
&= \frac{1}{h}\int_0^1 K\left(\frac{u}{h}\right)p(u)du \\
&= \int_0^{1/h} K(t)p(ht)dt \qquad \text{let } t = \frac{u}{h} \\
&= \int_0^{1/h} K(t)\left(p(0) + ht\cdot\partial_+ p(0) + \frac{h^2t^2}{2}\cdot\partial_+^2 p(0) + o(h^2)\right)dt \\
&= p(0)\int_0^{1/h} K(t)dt + O(h)\int_0^{1/h} tK(t)dt + O(h^2)\underbrace{\int_0^{1/h} t^2K(t)dt}_{\le \sigma_K^2/2<\infty} \\
&\le p(0) + O(h),
\end{aligned}$$

where we assumed $K(\cdot)$ is supported on $[-1,1]$, $h \le 1$, and $\int_0^{1/h} tK(t)dt$ is bounded.

Problem 4 [10 pts.]

Let p be a density on the real line. Assume that p is m-times continuously differentiable and that $\int |p^{(m)}|^2 < \infty$. Let K be a higher order kernel. This means that $\int K(y)dy = 1$, $\int y^j K(y)dy = 0$ for $1 \le j \le m-1$, $\int |y|^m K(y)dy < \infty$ and $\int K^2(y)dy < \infty$. Show that the kernel estimator with bandwidth h satisfies

$$\mathbb{E}\int (\widehat{p}(x) - p(x))^2 dx \le C\left(\frac{1}{nh} + h^{2m}\right)$$

for some $C > 0$. What is the optimal bandwidth and what is the corresponding rate of convergence (using this bandwidth)?

Solution.

We assume p has bounded m derivatives, and so $p \in \Sigma(m, L)$ for some constant $L > 0 \in \mathbb{R}$. Let's first analyze the bias $b(x)$:

$$\begin{aligned}
\mathbb{E}\left[\hat{p}(x)\right] - p(x) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)\right] - p(x) \\
&= \mathbb{E}\left[\frac{1}{h}K\left(\frac{x - x_1}{h}\right)\right] - p(x) \\
&= \int \frac{1}{h}K\left(\frac{x-u}{h}\right)p(u)du - p(x) \\
&= \int K(t)\, p(x - th)dt - p(x) && \text{where } t = \frac{x-u}{h} \\
&= \int K(t)\left[p(x) - thp'(x) + \frac{t^2h^2}{2}p''(x) + \ldots + \frac{(-th)^{m-1}}{(m-1)!}p^{(m-1)}(x)\right. \\
&\quad \left. + \frac{(-th)^m}{m!}p^{(m)}(w)\right]dt - p(x) && w \in (x - th, x)\text{, Taylor Exp.}
\end{aligned}$$

Given $\int K(y)dy = 1$, then $\int K(t)\, p(x)dt = p(x)$ and $\int y^j K(y)dy = 0$ for $1 \le j \le m-1$, so we are left with:

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{p}(x)\right] - p(x)\right| &= \left|\int K(t)\frac{(-th)^m}{(m)!}p^{(m)}(w)dt\right| \\
&\le \frac{Lh^m}{m!}\left|\int K(t)t^m dt\right| \\
&\le \frac{Lh^m}{m!}\int |K(t)|\,|t|^m\, dt = Ch^m && \text{for some } 0 < C < \infty
\end{aligned}$$

And so we have that $\int b(x)^2 dx \le C' h^{2m}$ for some $0 < C' < \infty$.
Analyzing now the variance we have that:

$$
\begin{aligned}
\mathbb{V}(\hat{p}(x)) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\right) \\
&= \frac{1}{nh^2}\mathbb{V}\left(K\left(\frac{x-x_1}{h}\right)\right) \\
&\leq \frac{1}{nh^2}\mathbb{E}\left(K\left(\frac{x-x_1}{h}\right)^2\right) \\
&= \frac{1}{nh^2}\int K\left(\frac{x-x_1}{h}\right)^2 p(x)dx \\
&= \frac{1}{nh}\int K(t)^2 p(x-th)dt && \text{where } t = \frac{x-u}{h} \\
&\leq \frac{\sup_x p(x)}{nh}\int K(t)^2\, dt \leq \frac{C''}{nh} && \text{for some } 0 < C'' < \infty
\end{aligned}
$$

Since densities in $\Sigma(m, L)$ are uniformly bounded.
The optimal bandwidth is therefore:

$$
\frac{\partial\left(\frac{1}{nh}+h^{2m}\right)}{\partial h} = 0 \implies -\frac{1}{nh^2} + 2mh^{2m-1} = 0 \implies h^* = (2mn)^{-\frac{1}{2m+1}} \asymp n^{-\frac{1}{2m+1}}
$$

And so the convergence rate is:

$$
\mathbb{E}\left[\int(\hat{p}(x)-p(x))^2dx\right] \leq n^{-\frac{2m}{2m+1}}
$$

Problem 5 [15 pts.]

Let $X_1, \dots, X_n \sim P$ where $X_i \in [0,1]$ and P has density p. Let $\phi_1, \phi_2, \dots$ be an orthonormal basis for $L_2[0,1]$. Hence $\int_0^1 \phi_j^2(x)dx = 1$ for all j and $\int_0^1 \phi_j(x)\phi_k(x)dx = 0$ for $j \neq k$. Assume that the basis is uniformly bounded, i.e. $\sup_j \sup_{0\leq x\leq 1}|\phi_j(x)| \leq C < \infty$. We may expand p as $p(x) = \sum_{j=1}^{\infty}\beta_j\phi_j(x)$ where $\beta_j = \int \phi_j(x)p(x)dx$. Define

$$
\widehat{p}(x) = \sum_{j=1}^{k}\widehat{\beta}_j\phi_j(x)
$$

where $\widehat{\beta}_j = (1/n)\sum_{i=1}^{n}\phi_j(X_i)$.

(a) Show that the risk is bounded by

$$
\frac{ck}{n} + \sum_{j=k+1}^{\infty}\beta_j^2
$$

for some constant $c > 0$.

(b) Define the Sobolev ellipsoid $E(m, L)$ of order m as the set of densities of the form $p(x) = \sum_{j=1}^{\infty}\beta_j\phi_j(x)$ where $\sum_{j=1}^{\infty}\beta_j^2 j^{2m} < L^2$. Show that the risk for any density in $E(m, L)$ is bounded by $c[(k/n) + (1/k)^{2m}]$. Using this bound, find the optimal value of k and find the corresponding risk.

Solution.

(a) **(10 pts.)**

First note,

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_j] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i)\right] \\ &= \mathbb{E}[\phi_j(x)] \\ &= \int_0^1 p(x)\phi_j(x)dx \\ &= \beta_j.\end{aligned}$$

So $\widehat{\beta}_j$ is unbiased. Now,

$$\begin{aligned}R(\widehat{p}(x)) &= \mathbb{E}\left[\int \left(\widehat{p}(x) - p(x)\right)^2 dx\right] \\ &= \mathbb{E}\left[\int \left(\sum_{j=1}^{k}\widehat{\beta}_j\phi_j(x) - \sum_{j=1}^{\infty}\beta_j\phi_j(x)\right)^2 dx\right] \\ &= \mathbb{E}\left[\int \left(\sum_{j=1}^{k}(\widehat{\beta}_j - \beta_j)\phi_j(x) - \sum_{j=k+1}^{\infty}\beta_j\phi_j(x)\right)^2 dx\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{k}(\widehat{\beta}_j - \beta_j)^2 + \sum_{j=k+1}^{\infty}\beta_j^2\right] && \text{since } \int \phi_i\phi_j = \delta_{ij} \\ &= \sum_{j=1}^{k}\mathrm{Var}(\widehat{\beta}_j) + \sum_{j=k+1}^{\infty}\beta_j^2 \\ &= \frac{k}{n}\mathrm{Var}\big(\phi_j(X_i)\big) + \sum_{j=k+1}^{\infty}\beta_j^2 \\ &\le \frac{C^2k}{n} + \sum_{j=k+1}^{\infty}\beta_j^2.\end{aligned}$$

(b) **(5 pts.)**

$$\begin{aligned}\sup_{p\in E(m,L)} R(\widehat{p}(x)) &\le \frac{C^2k}{n} + \sum_{j=k+1}^{\infty}\beta_j^2 && \text{from part (a)} \\ &= \frac{C^2k}{n} + \frac{k^{2m}\sum_{j=k+1}^{\infty}\beta_j^2}{k^{2m}} \\ &\le \frac{C^2k}{n} + \frac{\sum_{j=k+1}^{\infty}\beta_j^2 j^{2m}}{k^{2m}} \\ &\le \frac{C^2k}{n} + \frac{L^2}{k^{2m}} \\ &\le \max\{C^2, L^2\}\left(\frac{k}{n} + \frac{1}{k^{2m}}\right)\end{aligned}$$

Optimal k (up to some constant) can be found by, $\frac{k}{n} = \frac{1}{k^{2m}}$, which is, $k = O(n^{1/(2m+1)})$. And the corresponding risk is of the rate, $O(n^{-2m/(2m+1)})$.

Problem 6 [35 pts.]

Recall that the total variation distance between two distributions P and Q is $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$. In some sense, this would be the ideal loss function to use for density estimation. We only use L_2 because it is easier to deal with. Here you will explore some properties of TV.

(a) Suppose that P and Q have densities p and q. Show that

$$\mathrm{TV}(P,Q) = (1/2)\int |p(x) - q(x)|dx.$$

(b) Let T be any mapping. Let X and Y be random variables. Then

$$\sup_A |P(T(X) \in A) - P(T(Y) \in A)| \le \sup_A |P(X \in A) - P(Y \in A)|.$$

(c) Let K be a kernel. Recall that the convolution of a density p with K is $(p \star K)(x) = \int p(z)K(x-z)dz$. Show that

$$\int |p \star K - q \star K| \le \int |K| \int |p - q|.$$

Hence, smoothing reduces L_1 distance.

(d) Let p be a density on $\mathbb{R}$ and let p_n be a sequence of densities. Suppose that $\int (p - p_n)^2 \to 0$. Show that $\int |p - p_n| \to 0$.

(e) Let $\hat{p}$ be a histogram on $\mathbb{R}$ with binwidth h. Under some regularity conditions it can be shown that

$$\mathbb{E}\int |p - p_n| \approx \frac{\sqrt{2}}{\pi n h}\int \sqrt{p} + \frac{1}{4}h\int |p'|.$$

Hence, this risk can be unbounded if $\int \sqrt{p} = \infty$. A density is said to have a regularly varying tail of order r if $\lim_{x\to\infty} p(tx)/p(x) = t^r$ for all $t > 0$ and $\lim_{x\to-\infty} p(tx)/p(x) = t^r$ for all $t > 0$. Suppose that p has a regularly varying tail of order r with $r < -2$. Show that the risk bound above is bounded.

Solution.

(a) **(10 pts.)**

For any measurable $B \subseteq \mathbb{R}$,

$$\begin{aligned}
\frac{1}{2}\int |p - q| &= \frac{1}{2}\int |p(x) - q(x)|dx \\
&\ge \frac{1}{2}\int_B (p(x) - q(x))dx + \frac{1}{2}\int_{\mathbb{R}\setminus B}(q(x) - p(x))dx \\
&= \frac{1}{2}\int_B p(x)dx - \frac{1}{2}\int_B q(x)dx + \frac{1}{2}\int_{\mathbb{R}\setminus B} q(x)dx - \frac{1}{2}\int_{\mathbb{R}\setminus B} p(x)dx \\
&= \frac{1}{2}\int_B p(x)dx - \frac{1}{2}\int_B q(x)dx + \frac{1}{2}\Big(1 - \int_B q(x)dx\Big) - \frac{1}{2}\Big(1 - \int_B p(x)dx\Big) \\
&= \Big(\int_B p(x)dx - \int_B q(x)dx\Big) \\
&= P(B) - Q(B)
\end{aligned}$$

$$\implies \frac{1}{2}\int |p-q| \geq P(B) - Q(B) \quad \text{for any measurable } B \subseteq \mathbb{R}.$$

By noting,

$$\frac{1}{2}\int |p-q| = \frac{1}{2}\int |q-p|,$$

parallel reasoning shows

$$\frac{1}{2}\int |p-q| \geq Q(B) - P(B) \quad \text{for any measurable } B \subseteq \mathbb{R}.$$

So together we have,

$$\frac{1}{2}\int |p-q| \geq \big|P(B) - Q(B)\big|$$

and thus

$$\frac{1}{2}\int |p-q| \geq \sup_{B\subseteq\mathbb{R}} \big|P(B) - Q(B)\big|, \tag{3}$$

for any measurable $B \subseteq \mathbb{R}$.

Now consider the set

$$B' = \{x \in \mathbb{R} : \; p(x) > q(x)\}.$$

B' is measurable and

$$\begin{aligned}
\frac{1}{2}\int |p-q| &= \frac{1}{2}\int |p(x)-q(x)|dx \\
&= \frac{1}{2}\int_{B'} (p(x)-q(x))dx + \frac{1}{2}\int_{\mathbb{R}\setminus B'} (q(x)-p(x))dx \\
&= \frac{1}{2}\int_{B'} p(x)dx - \frac{1}{2}\int_{B'} q(x)dx + \frac{1}{2}\int_{\mathbb{R}\setminus B'} q(x)dx - \frac{1}{2}\int_{\mathbb{R}\setminus B'} p(x)dx \\
&= \frac{1}{2}\int_{B'} p(x)dx - \frac{1}{2}\int_{B'} q(x)dx + \frac{1}{2}\Big(1 - \int_{B'} q(x)dx\Big) - \frac{1}{2}\Big(1 - \int_{B'} p(x)dx\Big) \\
&= \Big(\int_{B'} p(x)dx - \int_{B'} q(x)dx\Big) \\
&= P(B') - Q(B'). \\
&= \big|P(B') - Q(B')\big|.
\end{aligned}$$

We have found a set $B' \subseteq \mathbb{R}$ such that

$$\frac{1}{2}\int |p-q| = \big|P(B') - Q(B')\big|,$$

therefore,

$$\frac{1}{2}\int |p-q| \leq \sup_{B\subseteq\mathbb{R}} \big|P(B) - Q(B)\big|. \tag{4}$$

Combining (3) and (4), we have

$$TV(P,Q) = \frac{1}{2}\int |p-q|.$$

(b) **(5 pts.)**

Let $\mathcal{F}$ be the σ-field generated by the sets A on the sample space Ω, and

$$\mathcal{C} = T(\mathcal{F}) = \{T(A) : \; A \in \mathcal{F}\}.$$

Define $T^{-1}(C) = \{\omega \in \Omega : T(\omega) \in C\}$, i.e. the pre-image mapping. By definition,

$$T^{-1}(\mathcal{C}) = \{T^{-1}(C) : C \in \mathcal{C}\} \subseteq \mathcal{F}.$$

Then,

$$\begin{aligned}\sup_{C\in\mathcal{C}} \big|P(T(X)\in C) - P(T(Y)\in C)\big| &= \sup_{A\in T^{-1}(\mathcal{C})} \big|P(X\in A) - P(Y\in A)\big| \\ &\le \sup_{A\in\mathcal{F}} \big|P(X\in A) - P(Y\in A)\big|.\end{aligned}$$

(c) **(5 pts.)**

$$\begin{aligned}\int |p\star K - q\star K| &= \int \left| \int p(z)K(x-z)dz - \int q(z)K(x-z)dz \right| dx \\ &= \int \left| \int (p(z)-q(z))K(x-z)dz \right| dx \\ &\le \int\int |p(z)-q(z)||K(x-z)|dzdx \\ &\le \int\int |p(z)-q(z)||K(x-z)|dxdz && \text{Fubini's theorem} \\ &= \int \left(|p(z)-q(z)| \int |K(x-z)|dx\right)dz \\ &= \int \left(|p(z)-q(z)| \int |K(x)|dx\right)dz && \text{invariant to translation} \\ &= \int |K(x)|dx \int |p(z)-q(z)|dz \\ &= \int |K| \int |p-q|\end{aligned}$$

(d) **(10 pts.)** Here we can further assume that the density has bounded support, see appendix for a proof without this assumption. By Cauchy inequality,

$$\left(\int |p-p_n|\right)^2 \le \int (p-p_n)^2 \int 1^2 \to 0,$$

where $\int 1^2$ is finite because density has bounded support.

(e) **(5 pts.)** We need to show that the integral is finite, $\int \sqrt{p} < +\infty$.

First, the regularly varying tail condition can be translated (not rigorously) as an expression for large value x,

$$p(tx) = t^r p(x), \forall |x| > B,$$

where $B > 0$ is a constant. Then we decompose the integral into three parts,

$$\int_x \sqrt{p(x)} = \int_{|x|\le B} \sqrt{p(x)} + \int_{x\ge B} \sqrt{p(x)} + \int_{x\le B} \sqrt{p(x)},$$

where the first term, integrating on bounded region, is finite. In the following, we argue that the second term $\int_{x\ge B} \sqrt{p(x)}$ is finite, and the third term is also finite using similar argument. By substituting variable $x = Bt$, and using regularly varying tail condition, the second term is,

$$\int_{x\ge B} \sqrt{p(x)}dx = B\int_{t\ge 1} \sqrt{p(tB)}dt = B\int_{t\ge 1} \sqrt{p(B)}t^{r/2}dt.$$

Since $r < -2$, the integral, $\int_{t\ge 1} t^{r/2}dt$, is finite.

Appendix

Proof of Claim in Problem 2.

From $\frac{1}{2^p}|a|^p - |b|^p \leq |a-b|^p \leq 2^p|a|^p + 2^p|b|^p$, we have

$$2^{-p}\mathbb{E}[|Z_i|^p] - p_h(x)^p \leq \mathbb{E}[|Z_i - p_h(x)|^p] \leq 2^p\mathbb{E}[|Z_i|^p] + 2^p p_h(x)^p.$$

Then,

$$\begin{aligned}\mathbb{E}[|Z_i|^p] &= \frac{1}{h^p}\int |K|^p\left(\frac{\|x-u\|}{h}\right)p(u)du \\ &= \frac{1}{h^{p-1}}\int |K|^p(\|v\|)p(x+hv)dv.\end{aligned}$$

So as $h \to 0$, choose any $[a,b]$ such that $|K|^p(\|v\|) > 0$ for some $v \in [a,b]$, then $\int \|K\|^p(\|v\|)p(x+hv)dv \geq \int_a^b |K|^p(\|v\|)p(x+hv)dv \to \int_a^b |K|^p(\|v\|)p(x)dv > 0$ by the Bounded Convergence Theorem. Also, $\int |K|^p(\|v\|)p(x+hv)dv \leq \int |K|^p(\|v\|)\sup_x p(x)dv < \infty$, hence $\int |K|^p(\|v\|)p(x+hv)dv = \Theta(1)$, and accordingly,

$$\mathbb{E}[|Z_i|^p] = \Theta\left(\frac{1}{h^{p-1}}\right).$$

Then

$$|p_h(x)| = |\mathbb{E}[Z_i]| \leq \mathbb{E}[|Z_i|] = O(1).$$

Hence

$$\Theta\left(\frac{1}{h^{p-1}}\right) = 2^{-p}\mathbb{E}[|Z_i|^p] - p_h(x)^p \leq \mathbb{E}[|Z_i - p_h(x)|^p] \leq 2^p\mathbb{E}[|Z_i|^p] + 2^p p_h(x)^p = \Theta\left(\frac{1}{h^{p-1}}\right)$$

which implies

$$\mathbb{E}[|Z_i - p_h(x)|^p] = \Theta\left(\frac{1}{h^{p-1}}\right).$$

Proof for Problem 6 (d).

First by $\int (p-p_n)^2 \to 0$, we claim $p_n \to p, a.s.$

It's because by contradiction, if there exist set A with $\int 1_A > 0$ such that $p_n(x) \nrightarrow p(x), \forall x \in A$, then $\int (p-p_n)^2 \geq \int_A (p-p_n)^2 > 0$.

Then note that $\int |p-p_n|$ is bounded,

$$\int 0 \leq \int |p-p_n| \leq \int p + p_n = 2.$$

Thus by Dominated convergence theorem,

$$\int |p-p_n| \to \int (\lim |p-p_n|) = 0.$$