Problem 1 [40 pts.]
Let \( X_1, \ldots, X_n \sim P \) where \( X_i \in [0,1] \) and \( P \) has density \( p \). Let \( \bar{p} \) be the histogram estimator using \( m \) bins. Let \( h = 1/m \). Recall that the \( L_2 \) error is 
\[
\int (\bar{p}(x) - p(x))^2 = \int \bar{p}^2(x) dx - 2 \int \bar{p}(x)p(x) dx + \int p^2(x) dx.
\]
As usual, we may ignore the last term so we define the loss to be 
\[
L(h) = \int \bar{p}^2(x) dx - 2 \int \bar{p}(x)p(x) dx.
\]

(a) Suppose we used the direct estimator of the loss, namely, we replace the integral with the average to get 
\[
\bar{L}(h) = \int \bar{p}^2(x) dx - \frac{2}{n} \sum_i \bar{p}(X_i).
\]
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 
\[
\tilde{L}(h) = \int \bar{p}^2(x) dx - \frac{2}{n} \sum_i \bar{p}_{-i}(X_i),
\]
Show that 
\[
\tilde{L}(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_j Z_j^2
\]
where \( Z_j \) is the number of observations in bin \( j \).

(c) Show that \( \bar{L}(h) - L(h) \xrightarrow{P} 0 \).

Solution.
Define 
\[
\bar{\theta}_j = \frac{1}{n} \sum_{i=1}^n 1(X_i \in B_j) \quad \text{and} \quad Z_j = n\bar{\theta}_j
\]
for \( j = 1, \ldots, m \).
(a) (15 pts.)

\[ \hat{L}(h) = \int \hat{\rho}^2(x)dx - \frac{2}{n} \sum_i \hat{\rho}(X_i) \]

\[ = \int \left( \frac{m}{h} \bar{\theta}_j \mathbb{1}(x \in B_j) \right)^2 dx - \frac{2}{n} \sum_i \frac{m}{h} \bar{\theta}_j \mathbb{1}(X_i \in B_j) \]

\[ = \frac{1}{h^2} \int \left( \sum_{j=1}^m \bar{\theta}_j \mathbb{1}(x \in B_j) \right)^2 dx - \frac{2}{h} \sum_{j=1}^m \bar{\theta}_j \cdot \frac{1}{n} \sum_i \mathbb{1}(X_i \in B_j) \]

\[ = \frac{1}{h^2} \sum_{j=1}^m \bar{\theta}_j^2 \int \mathbb{1}(x \in B_j)dx - \frac{2}{h} \sum_{j=1}^m \bar{\theta}_j^2 \]

\[ = \frac{1}{h} \sum_{j=1}^m \bar{\theta}_j^2 - \frac{2}{h} \sum_{j=1}^m \bar{\theta}_j \]

\[ = -\frac{1}{h} \sum_{j=1}^m \bar{\theta}_j^2 \]

\[ = -\frac{1}{hn^2} \sum_{j=1}^m Z_j^2 \]

\( \sum_{j=1}^m Z_j^2 \) is bounded below by \( n \), so \( \hat{L}(h) \to -\infty \) as \( h \to 0 \). Therefore, this loss is minimized by taking \( h = 0 \).

(b) (10 pts.)

From part (a) we have

\[ \int \hat{\rho}^2(x)dx = \frac{1}{h} \sum_{j=1}^m \bar{\theta}_j^2. \]  

(1)

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

\[ \frac{2}{n} \sum_{i=1}^n \hat{\rho}_{(-i)}(X_i) = \frac{2}{n(n-1)h} \sum_{j=1}^m \sum_{i=1}^n \mathbb{1}(X_i \in B_j) \sum_{k \neq i} \mathbb{1}(X_k \in B_j) \]

\[ = \frac{2}{n(n-1)h} \sum_{j=1}^m \sum_{i=1}^n \mathbb{1}(X_i \in B_j)(n\bar{\theta}_j - \mathbb{1}(X_i \in B_j)) \]

\[ = \frac{2}{n(n-1)h} \sum_{j=1}^m (n^2\bar{\theta}_j^2 - n\bar{\theta}_j). \]  

(2)
Taking the difference of (1) and (2), we get
\[
\hat{L}(h) = \frac{1}{h} \sum_{j=1}^{m} \theta_j^2 - \frac{2}{n(n-1)h} \sum_{j=1}^{m} (n^2 \theta_j^2 - n \theta_j)
\]
\[
= \frac{2}{(n-1)h} \sum_{j=1}^{m} \theta_j + \sum_{j=1}^{m} \theta_j^2 \left( \frac{1}{h} - \frac{2n}{(n-1)h} \right)
\]
\[
= \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h} \sum_{j=1}^{m} \theta_j^2
\]
\[
= \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_{j=1}^{m} Z_j^2.
\]

(c) **(15 pts.)**

It suffices to show
\[
\frac{1}{n} \sum_i \hat{p}_{(-i)}(X_i) - \int \hat{p}(x) p(x) dx = P \rightarrow 0. \tag{3}
\]

Caveat: It is not a valid line of reasoning to directly say (3) holds by the weak law of large numbers since the \(\hat{p}_{(-i)}\) are functions of the sample size.

\[
\int \hat{p}(x) p(x) dx = \int \sum_{j=1}^{m} \frac{\hat{\theta}_j}{h} 1(x \in B_j) p(x) dx
\]
\[
= \int_{B_j} \sum_{j=1}^{m} \frac{\hat{\theta}_j}{h} p(x) dx
\]
\[
= \sum_{j=1}^{m} \frac{\hat{\theta}_j}{h} \int_{B_j} p(x) dx \quad \text{by Fubini/Tonelli theorem}
\]
\[
= \sum_{j=1}^{m} \frac{\hat{\theta}_j}{h} P(X \in B_j) = \hat{\theta}_j
\]
\[
= \frac{1}{h} \sum_{j=1}^{m} \hat{\theta}_j \theta_j
\]

And following from (2) we have
\[
\frac{1}{n} \sum_i \hat{p}_{(-i)}(X_i) = \frac{1}{n(n-1)h} \sum_{j=1}^{m} (n^2 \theta_j^2 - n \theta_j)
\]
\[
= \frac{n}{(n-1)h} \sum_{j=1}^{m} \theta_j^2 - \frac{1}{(n-1)h} \sum_{j=1}^{m} \theta_j
\]
\[
= \frac{1}{h} \sum_{j=1}^{m} \theta_j^2 + \frac{1}{(n-1)h} \sum_{j=1}^{m} \theta_j^2 - \frac{1}{(n-1)h}
\]
\[
= \frac{1}{h} \sum_{j=1}^{m} \theta_j^2 + O(n^{-1}).
\]
Therefore,

\[
\frac{1}{n} \sum_i \hat{p}_{(-i)}(X_i) - \int \hat{p}(x)p(x)dx = \frac{1}{h} \sum_{j=1}^{m} \hat{\theta}_j^2 - \frac{1}{h} \sum_{j=1}^{m} \hat{\theta}_j \theta_j + O(n^{-1})
\]

\[
= \frac{1}{h} \sum_{j=1}^{m} \hat{\theta}_j (\hat{\theta}_j - \theta_j) + O(n^{-1})
\]

\[
P \to 0
\]

since

\[
\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \in B_j) \to P(X \in B_j) = \theta_j
\]

by the weak law of large numbers.
Problem 2

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

(a) Show that
\[
\frac{\hat{p}_h(x) - p(x)}{s_n(x)} \sim N(0, 1)
\]
where \( p_h(x) = \mathbb{E}[\hat{p}_h(x)] \).

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 \( s_n^{-1} \sum_i (Y_i - \mu_i) \sim N(0, 1) \).

(b) Assume that the smoothness is \( \beta = 2 \). Suppose that the bandwidth \( h_n \) is chosen optimally. Show that
\[
\frac{\hat{p}_h(x) - p(x)}{s_n(x)} \sim N(b(x), 1)
\]
for some constant \( b(x) \) which is, in general, not 0.

Solution.

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

Proof. See appendix

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

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

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

\[
= \Theta \left( \frac{1}{nh} \right).
\]

Therefore,

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

\[
= \Theta \left( \left( \frac{1}{n^{\delta}} \right)^{2} \right)
\]

\[
\rightarrow 0,
\]

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

\[
\frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} \sim N(0, 1).
\]

(b) First note

\[
\frac{\hat{p}_n(x) - p(x)}{s_n(x)} = \frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} + \frac{p_h(x) - p(x)}{s_n(x)}
\]

\[
= \frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} + \frac{\text{Bias}(p_h(x))}{\sqrt{\text{Var}(\hat{p}_n(x))}}
\]

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

Now from part (a), we have

\[
\text{Var}(\hat{p}_n(x)) = \Theta \left( \frac{1}{nh} \right)
\]

and from Lemma 3,

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

Therefore,

\[
\frac{\hat{p}_n(x) - p(x)}{s_n(x)} = \frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} + \frac{\text{Bias}(p_h(x))}{\sqrt{\text{Var}(\hat{p}_n(x))}}
\]

\[
= \frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} + \frac{O(h^2)}{\Theta \left( \frac{1}{n h^{1/2}} \right)}
\]

\[
= \frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} + \frac{O(n^{-2/5})}{\Theta \left( n^{-2/5} \right)}
\]

\[
= \frac{\hat{p}_n(x) - p_h(x)}{s_n(x)} + O(1)
\]

\[
\sim N(0, 1)
\]

\[
\sim N(b(x), 1).
\]
Problem 3

Let $X_1, \ldots, X_n \sim P$ where $X_i \in [0, 1]$. Assume that $P$ has density $p$ which has bounded continuous derivative. Let $\hat{P}_h(x)$ be the kernel density estimator. Show that

$$ E[\hat{P}_h(0)] = \frac{p(0)}{2} + O(h). $$

Solution.

$$ E[\hat{P}(0)] = E \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left( \frac{-X_i}{h} \right) \right] $$

$$ = E \left[ \frac{1}{h} K \left( \frac{X_i}{h} \right) \right] $$

$$ = \frac{1}{h} \int_{0}^{1} K \left( \frac{u}{h} \right) p(u) du $$

$$ = \int_{0}^{1/h} K(t) p(ht) dt $$

$$ = \int_{0}^{1/h} K(t) \left( p(0) + ht \cdot \partial_x p(0) + \frac{h^2 t^2}{2} \partial^2_x p(0) + o(h^2) \right) dt $$

$$ = p(0) \int_{0}^{1/h} K(t) dt + O(h) \int_{0}^{1/h} t K(t) dt + O(h^2) \int_{0}^{1/h} t^2 K(t) dt $$

$$ \leq \frac{\sigma^2_K}{2} $$

$$ \leq \sigma^2_K/2 < \infty $$

$$ = \frac{p(0)}{2} + O(h), $$

where we assumed $K(\cdot)$ is supported on $[-1, 1]$, $h \leq 1$, and $\int_{0}^{1/h} t K(t) dt$ is bounded.
Problem 4

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 $1 \leq j \leq m-1$, $\int y^m K(y)dy < \infty$ and $\int K^2(y)dy < \infty$. Show that the kernel estimator with bandwidth $h$ satisfies

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

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

Solution.

See Chapter 6 of [1].
Problem 5 [20 pts.]

Let \( X_1, \ldots, X_n \sim P \) where \( X_i \in [0, 1] \) and \( P \) has density \( p \). Let \( \phi_1, \phi_2, \ldots \) 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

\[
\hat{p}(x) = \sum_{j=1}^k \hat{\beta}_j \phi_j(x)
\]

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

(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,

\[
\mathbb{E}[\hat{\beta}_j] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \phi_j(X_i) \right] = \mathbb{E} \left[ \phi_j(x) \right] = \int_0^1 p(x)\phi_j(x)dx = \beta_j.
\]
So $\hat{\beta}_j$ is unbiased. Now,

$$R(\hat{p}(x)) = \mathbb{E} \left[ \int (\hat{p}(x) - p(x))^2 dx \right]$$

$$= \mathbb{E} \left[ \int \left( \sum_{j=1}^{k} \hat{\beta}_j \phi_j(x) - \sum_{j=1}^{\infty} \beta_j \phi_j(x) \right)^2 dx \right]$$

$$= \mathbb{E} \left[ \int \left( \sum_{j=1}^{k} (\hat{\beta}_j - \beta_j) \phi_j(x) - \sum_{j=k+1}^{\infty} \beta_j \phi_j(x) \right)^2 dx \right]$$

$$= \mathbb{E} \left[ \sum_{j=1}^{k} (\hat{\beta}_j - \beta_j)^2 + \sum_{j=k+1}^{\infty} \beta_j^2 \right]$$

$$= \sum_{j=1}^{k} \text{Var}(\hat{\beta}_j) + \sum_{j=k+1}^{\infty} \beta_j^2$$

$$= \frac{k}{n} \text{Var}(\phi_j(X_i)) + \sum_{j=k+1}^{\infty} \beta_j^2$$

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

(b) (10 pts.)

$$\sup_{p \in \mathcal{E}(m, L)} R(\hat{p}(x)) \leq \frac{C^2 k}{n} + \sum_{j=k+1}^{\infty} \beta_j^2$$

from part (a)

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

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

$$\leq \frac{C^2 k}{n} + \frac{L^2}{k^{2m}}$$

$$\leq \max\{C^2, L^2\} \left( \frac{k}{n} + \frac{1}{k^{2m}} \right)$$
Problem 6 [40 pts.]

Recall that the total variation distance between two distributions $P$ and $Q$ is $\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. 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

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

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

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

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

$$\int |p \ast K - q \ast K| \leq \int |K| \int |p - q|.$$ 

Hence, smoothing reduces $L_1$ distance.

Solution.

(a) (15 pts.)

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

$$\frac{1}{2} \int |p - q| = \frac{1}{2} \int |p(x) - q(x)|dx$$

$$\geq \frac{1}{2} \int_B (p(x) - q(x))dx + \frac{1}{2} \int_{\mathbb{R}\setminus B} (q(x) - p(x))dx$$

$$= \frac{1}{2} \int_B p(x)dx - \frac{1}{2} \int_B q(x)dx + \frac{1}{2} \int_{\mathbb{R}\setminus B} q(x)dx - \frac{1}{2} \int_{\mathbb{R}\setminus B} p(x)dx$$

$$= \frac{1}{2} \int_B p(x)dx - \frac{1}{2} \int_B q(x)dx + \frac{1}{2} (1 - \int_B q(x)dx) - \frac{1}{2} (1 - \int_B p(x)dx)$$

$$= (\int_B p(x)dx - \int_B q(x)dx)$$

$$= P(B) - Q(B)$$

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

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,

$$\frac{1}{2} \int |p - q| \geq |P(B) - Q(B)|.$$
and thus
\[
\frac{1}{2} \int |p - q| \geq \sup_{B \subseteq \mathbb{R}} |P(B) - Q(B)|, \quad (4)
\]
for any measurable \( B \subseteq \mathbb{R} \).

Now consider the set
\[
B' = \{ x \in \mathbb{R} : p(x) > q(x) \}.
\]

\( B' \) is measurable and
\[
\frac{1}{2} \int |p - q| = \frac{1}{2} \int |p(x) - q(x)|dx
= \frac{1}{2} \int_{B'} (p(x) - q(x))dx + \frac{1}{2} \int_{\mathbb{R} \setminus B'} (q(x) - p(x))dx
= \frac{1}{2} \int_{B'} p(x)dx - \frac{1}{2} \int_{B'} q(x)dx + \frac{1}{2} \int_{\mathbb{R} \setminus B'} q(x)dx - \frac{1}{2} \int_{\mathbb{R} \setminus B'} p(x)dx
= \frac{1}{2} \int_{B'} p(x)dx - \frac{1}{2} \int_{B'} q(x)dx + \frac{1}{2} (1 - \int_{B'} q(x)dx) - \frac{1}{2} (1 - \int_{B'} p(x)dx)
= (\int_{B'} p(x)dx - \int_{B'} q(x)dx)
= P(B') - Q(B')
= |P(B') - Q(B')|.
\]

We have found a set \( B' \subseteq \mathbb{R} \) such that
\[
\frac{1}{2} \int |p - q| = |P(B') - Q(B')|,
\]
therefore,
\[
\frac{1}{2} \int |p - q| \leq \sup_{B \subseteq \mathbb{R}} |P(B) - Q(B)|, \quad (5)
\]
Combining (4) and (5), we have
\[
TV(P, Q) = \frac{1}{2} \int |p - q|.
\]

(b) (10 pts.)

Let \( \mathcal{F} \) be the \( \sigma \)-field generated by the sets \( A \) on the sample space \( \Omega \), and
\[
\mathcal{C} = \mathcal{T}(\mathcal{F}) = \{ T(A) : A \in \mathcal{F} \}.
\]

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,
\[
\sup_{C \in \mathcal{C}} |P(T(X) \in C) - P(T(Y) \in C)| = \sup_{A \in T^{-1}(\mathcal{C})} \sup_{A' \in \mathcal{F}} |P(X \in A) - P(Y \in A)|
\leq \sup_{A' \in \mathcal{F}} |P(X \in A) - P(Y \in A)|.
\]
\( \int |p \ast K - q \ast K| = \int \left| \int p(z) K(x - z) \, dz - \int q(z) K(x - z) \, dz \right| \, dx \)

\[ = \int \left| \int (p(z) - q(z)) K(x - z) \, dz \right| \, dx \]

\[ \leq \int \int |p(z) - q(z)| |K(x - z)| \, dz \, dx \]

\[ \leq \int \int |p(z) - q(z)| |K(x - z)| \, dx \, dz \quad \text{Fubini’s theorem} \]

\[ = \int \left( |p(z) - q(z)| \int |K(x - z)| \, dx \right) \, dz \]

\[ = \int \left( |p(z) - q(z)| \int |K(x)| \, dx \right) \, dz \quad \text{invariant to translation} \]

\[ = \int |K(x)| \, dx \int |p(z) - q(z)| \, dz \]

\[ = \int |K| \int |p - q| \]
References


Appendix

Proof of Claim.

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

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

Then,

$$E[|Z_i|^p] = \frac{1}{h^p} \int |K|^p\left(\frac{\|x - u\|}{h}\right) p(u) du$$

$$= \frac{1}{h^{p-1}} \int |K|^p(\|v\|) p(x + hv) dv.$$  

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,

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

Then

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

Hence

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

which implies

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