STA355: Final Exam

Instructor: Wenlong Mou

Dec 19th, 2025

Student name: _____

Student ID: _____

Student signature: _____

This exam contains 12 pages.

Total marks: 100 pts

Time Allowed: 180 minutes

Question 1. [24 points, 3 points each] Mark each statement with T (true) or F (false). No justification required.

(1) Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. Let $g(x) := |x|$. Then $\sqrt{n}(g(\frac{1}{n}\sum_{i=1}^n X_i) - g(0)) \xrightarrow{d} \mathcal{N}(0,1)$. **Answer:** F

(2) If $\widehat{\theta}_n^{(1)}$ and $\widehat{\theta}_n^{(2)}$ are both consistent estimators of a scalar parameter θ, then $\max(\widehat{\theta}_n^{(1)}, \widehat{\theta}_n^{(2)})$ is also a consistent estimator of θ. **Answer:** T

(3) If $\widehat{\theta}_n^{(1)}$ and $\widehat{\theta}_n^{(2)}$ are both unbiased estimators of a scalar parameter θ, then $\max(\widehat{\theta}_n^{(1)}, \widehat{\theta}_n^{(2)})$ is also a unbiased estimator of θ. **Answer:** F

(4) Consider $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, where σ^2 is unknown. Then the statistic $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i^2$ is sufficient for σ^2. **Answer:** T

(5) Consider a collection of hypothesis testing problems $H_{0,i}$ vs. $H_{1,i}$, $i = 1, 2, \ldots, m$. Suppose that for each i, ϕ_i is a level-α test for $H_{0,i}$ vs. $H_{1,i}$. Then the test $\phi = \max_{1 \le i \le m} \phi_i$ is a level-α test for the global null hypothesis $H_0 : \bigcap_{i=1}^m H_{0,i}$ vs. the global alternative hypothesis $H_1 : \bigcup_{i=1}^m H_{1,i}$. **Answer:** F

(6) Consider a hypothesis testing problem H_0 vs. H_1. Let p be the p-value of a test computed based on the observed data. If we reject H_0 whenever $p \le \alpha$, then the resulting test is a level-α test. **Answer:** T

(7) If $\widehat{\theta}$ is a minimax optimal estimator for a parameter θ under squared error loss, then $\widehat{\theta}$ is admissible. **Answer:** F

(8) K-fold cross validation provides an exactly unbiased estimate of the prediction loss of a regression model trained on the entire dataset. **Answer:** F

Question 2. [14 points] Empirical estimator and bootstrap

Part (a). [6 points] Given samples $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$ for some $p \in (0,1)$, let $\widehat{p}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function, satisfying $g'(p) \neq 0$. Find the asymptotic distribution of $\sqrt{n}(g(\widehat{p}_n) - g(p))$.

Answer: By the Delta method, since g is continuously differentiable with $g'(p) \neq 0$, and $\sqrt{n}(\widehat{p}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$, we have:

$$\sqrt{n}(g(\widehat{p}_n) - g(p)) \xrightarrow{d} \mathcal{N}\big(0, g'(p)^2 p(1-p)\big).$$

Rubrics: 2 points for deriving the asymptotic distribution of $\widehat{p}_n - p$.

3 points for correctly invoking Delta method

1 point for the final calculation

(if all the intermediate steps are correct but the final answer is wrong, give 5 points)

Part (b). [8 points] Continuing the discussion in part (a), now that we want to use the bootstrap method to approximate the distribution of $\sqrt{n}(g(\widehat{p}_n) - g(p))$. Conditionally on the observed data $X_1, X_2, \ldots, X_n$, find the asymptotic distribution of the bootstrap quantity $\sqrt{n}(g(\widehat{p}_n^*) - g(\widehat{p}_n))$, where $\widehat{p}_n^* := \frac{1}{n}\sum_{i=1}^n X_i^*$. Does the bootstrap provide a valid approximation of the distribution of $\sqrt{n}(g(\widehat{p}_n) - g(p))$? Justify your answer.

Answer: Conditionally on the observed data $X_1, \ldots, X_n$, the bootstrap samples $X_1^*, \ldots, X_n^*$ are i.i.d. samples from $\text{Ber}(\widehat{p}_n)$, where $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Therefore, we have

$$n\widehat{p}_n^* = \sum_{i=1}^n X_i^* \sim \text{Binom}(n, \widehat{p}_n), \qquad \text{conditionally.}$$

By the Central Limit Theorem for Binomial distribution, we have

$$\sqrt{n}(\widehat{p}_n^* - \widehat{p}_n) \mid X_1, \ldots, X_n \xrightarrow{d} \mathcal{N}(0, \widehat{p}_n(1-\widehat{p}_n))$$

By the Delta method:

$$\sqrt{n}(g(\widehat{p}_n^*) - g(\widehat{p}_n)) \mid X_1, \ldots, X_n \xrightarrow{d} \mathcal{N}(0, g'(\widehat{p}_n)^2\widehat{p}_n(1-\widehat{p}_n))$$

Since $\widehat{p}_n \xrightarrow{p} p$ and g' is continuous, we have $g'(\widehat{p}_n)^2\widehat{p}_n(1-\widehat{p}_n) \xrightarrow{p} g'(p)^2p(1-p)$.

Therefore, the bootstrap does provide a valid approximation: the conditional distribution of $\sqrt{n}(g(\widehat{p}_n^*) - g(\widehat{p}_n))$ given the data converges in probability to the same limiting distribution $\mathcal{N}(0, g'(p)^2p(1-p))$ as the distribution of $\sqrt{n}(g(\widehat{p}_n) - g(p))$ from part (a).

Rubrics: 4 points for correctly identifying the conditionally distribution of $\widehat{p}_n^*$.

2 points for applying the Delta method to get the asymptotic distribution of $\sqrt{n}(g(\widehat{p}_n^*) - g(\widehat{p}_n))$.

2 points for justifying the validity of the bootstrap approximation.

Any other correct solution using different approaches can also get full marks.

If you only mention bootstrap is valid without justification (or just mentioning the terminology Hadamard differentiability, without verifying it), give 2 points.

Question 3. [14 poitns] MLE and parametric estimation.

Part (a). [6 points] Consider the class of Laplace distributions with density functions

$$p_\theta(x) = \frac{1}{2} e^{-|x-\theta|}, \quad x \in \mathbb{R}.$$

Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta^*}$ for some unknown parameter $\theta^* \in [-1, 1]$. Let $\widehat{\theta}_n$ be the maximum likelihood estimator (MLE) of θ^*. Show that $\widehat{\theta}_n$ is consistent.

Answer:

Solution I Define the sample-level and population-level log-likelihood functions as

$$\ell(\theta, X_i) := \log p_\theta(X_i) = \log \frac{1}{2} - |X_i - \theta|,$$
$$L(\theta) := \mathbb{E}_{\theta^*}[\ell(\theta, X)] = \int_{-\infty}^{\infty} p_{\theta^*}(x) \ell(\theta, x) \mathrm{d}x.$$

Since ℓ is continuous in θ and the model is identifiable (i.e., $p_\theta \neq p_{\theta'}$ for $\theta \neq \theta'$), the function L is uniquely maximized at $\theta = \theta^*$, and L is continuous in θ. Furthermore, we note that

$$\mathbb{E}\Big[\sup_{\theta \in [-1,1]} |\ell(\theta, X)|\Big] \leq \log 2 + \mathbb{E}_{\theta^*}[|X| + 1] < \infty.$$

So by the uniform law of large numbers, we have

$$\sup_{\theta \in [-1,1]} \Big| \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i) - L(\theta) \Big| \xrightarrow{p} 0.$$

Combining the above results, we obtain the consistency of the MLE $\widehat{\theta}_n$ for θ^*.

Rubrics: 2 points for correctly defining the log-likelihood functions.

4 points for verifying the conditions needed for applying the uniform law of large numbers and concluding consistency of MLE.

Another method using the property of sample median can also get full marks. If you use that method, give 3 points for showing MLE is the sample median, and 3 points for showing the sample median is consistent.

Part (b). [8 points] Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables that take positive values, and let each Y_i follow

$$Y_i \mid X_i \sim \text{Poisson}(\theta^* X_i),$$

where $\theta^* > 0$ is an unknown parameter. The probability mass function of Poisson(λ) distribution is given by

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

Find the form of the log-likelihood function $\ell(\theta; X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ based on the observed data (X_i, Y_i), $i = 1, 2, \ldots, n$. Compute the asymptotic distribution of the MLE $\widehat{\theta}_n$ for θ^*.

Answer: The log-likelihood function for each sample is

$$\ell(\theta; X_i, Y_i) = \log p_\theta(Y_i \mid X_i) = Y_i \log(\theta X_i) - \theta X_i - \log(Y_i!).$$

So the overall log-likelihood function is

$$\ell(\theta; X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \sum_{i=1}^n Y_i \log \theta + \sum_{i=1}^n Y_i \log X_i - \theta \sum_{i=1}^n X_i - \sum_{i=1}^n \log(Y_i!).$$

Here we provide two methods to compute the asymptotic distribution of the MLE.

Method I: using Fisher information The score function is:

$$\frac{\partial \ell(\theta; X_i, Y_i)}{\partial \theta} = \frac{Y_i}{\theta} - X_i.$$

So the Fisher information for one sample is:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ell(\theta; X_i, Y_i)}{\partial \theta}\right)^2\right] = \mathbb{E}\left[\text{Var}\left(\frac{Y_i}{\theta} \mid X_i\right)\right] = \frac{\mathbb{E}[X_i]}{\theta}.$$

By the asymptotic normality of MLE

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta^*}{\mathbb{E}[X_i]}\right).$$

Method II: using the explicit form of MLE To find the MLE, we take the derivative with respect to θ and set it to zero:

$$\frac{\partial \ell}{\partial \theta}(\theta; X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \frac{1}{\theta} \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i = 0.$$

Therefore, the MLE is:

$$\widehat{\theta}_n = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}.$$

For the asymptotic distribution, note that $\mathbb{E}_{\theta^*}[Y_i \mid X_i] = \theta^* X_i$, so $\mathbb{E}[Y_i] = \theta^* \mathbb{E}[X_i]$. By the law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}[X_i]$$

By central limit theorem, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \theta^* X_i) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{var}(Y_1 - \theta^* X_1)\right) = \mathcal{N}\left(0, \theta^* \mathbb{E}[X_i]\right).$$

Using Slutsky's theorem, we obtain

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \theta^* X_i)}{\frac{1}{n}\sum_{i=1}^n X_i} \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta^*}{\mathbb{E}[X_i]}\right).$$

Rubrics: 3 points for the expression of log-likelihood function.
4 points for using the correct methods to derive the asymptotic distribution.
1 points for calculating the final answer correctly.
Intermdiate steps will also be given partial credits.

Question 4. [14 points] Hypothesis testing

Part (a). [6 points] Consider two independent samples $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_X, 1)$ and $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_Y, 1)$, where $\theta_X, \theta_Y \in \mathbb{R}$ are unknown parameters. Consider the hypothesis testing problem $H_0 : \theta_X = \theta_Y$ vs. $H_1 : \theta_X \neq \theta_Y$. Construct an exact level-α test for this problem using what you have learned in this course.

[Note: the answer is not unique. You will get full marks as long as your test is valid and well justified.]

Answer: The solution is not unique. Here we provide two possible solutions.

Solution I: Wald's Test Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ be the sample means. Under $H_0 : \theta_X = \theta_Y$, we have $\overline{X} - \overline{Y} \sim \mathcal{N}(0, 2/n)$.

The test statistic is:

$$W_n = \frac{\overline{X} - \overline{Y}}{\sqrt{2/n}} \sim \mathcal{N}(0,1) \quad \text{under } H_0.$$

We reject H_0 if $|W_n| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. This is an exact level-α test.

Solution II: Permutation Test Pool the two samples to get $Z_1, Z_2, \ldots, Z_{2n}$ (first n are X_i's, next n are Y_i's). Compute the test statistic $T_{\text{obs}} = |\overline{X} - \overline{Y}|$.

For each permutation π of $\{1, 2, \ldots, 2n\}$, split the permuted data into two groups of size n and compute the test statistic T_π.

The p-value is:

$$\text{p-value} = \frac{1}{(2n)!}\sum_{\pi} \mathbf{1}_{T_\pi \geq T_{\text{obs}}}.$$

(Computationally, we can approximate this by randomly sampling permutations.)

We reject H_0 if p-value $\leq \alpha$. This test is exact and does not require normality assumptions.

Rubrics: Any valid test construction will get full marks.

3 points for correctly defining the test statistic.

3 points for justifying the level-α property of the test.

Part (b). [8 points] Given i.i.d. samples from a one-dimensional Gaussian mixture model

$$X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(-\theta^*, 1) + \frac{1}{2}\mathcal{N}(\theta^*, 1),$$

where $\theta^* \in \mathbb{R}$ is an unknown parameter. Consider the hypothesis testing problem $H_0 : \theta^* = 0$ vs. $H_1 : |\theta^*| \geq \varepsilon$. Consider the test statistic

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Show that there exists a constant $c > 0$, such that when $\varepsilon \geq cn^{-1/4}$, there exists a test based on T_n that makes the sum of type-I and type-II errors less than $1/4$.

[Hint: for $Z \sim \mathcal{N}(0, 1)$, we have $\mathbb{E}[Z^4] = 3$. You may use this fact to compute the mean and variance of T_n under both null and alternative hypotheses.]
[Note: you do not need to find the optimal constant c.]

Answer: Under $H_0 : \theta^* = 0$, we have

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}[X_1^2] = 1,\\ \operatorname{var}(T_n) &= \frac{1}{n}\operatorname{var}(X_1^2) = \frac{1}{n}(\mathbb{E}[X_1^4] - (\mathbb{E}[X_1^2])^2) = \frac{1}{n}(3-1) = \frac{2}{n}.\end{aligned}$$

Under $\theta^* \neq 0$, we have

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}[X_1^2] = 1 + (\theta^*)^2,\\ \operatorname{var}(T_n) &= \frac{1}{n}\operatorname{var}(X_1^2) = \frac{1}{n}(\mathbb{E}[X_1^4] - (\mathbb{E}[X_1^2])^2) = \frac{1}{n}(2 + 4(\theta^*)^2).\end{aligned}$$

We construct the test that rejects H_0 if $T_n > 1 + \frac{\varepsilon^2}{2}$. By Chebyshev's inequality, under H_0, we have

$$\mathbb{P}_{H_0}\left(T_n > 1 + \frac{\varepsilon^2}{2}\right) \leq \mathbb{P}_{H_0}\left(|T_n - 1| > \frac{\varepsilon^2}{2}\right) \leq \frac{4\operatorname{var}(T_n)}{\varepsilon^4} = \frac{8}{n\varepsilon^4}.$$

Under H_1, we have

$$\mathbb{P}_{\theta^*}\left(T_n \leq 1 + \frac{\varepsilon^2}{2}\right) \leq \mathbb{P}_{\theta^*}\left(|T_n - (1 + (\theta^*)^2)| \geq \frac{(\theta^*)^2}{2}\right) \leq \frac{4\operatorname{var}(T_n)}{(\theta^*)^4} \leq \frac{8 + 16\varepsilon^2}{n\varepsilon^4}.$$

Choosing $\varepsilon \geq 16n^{-1/4}$, we can ensure that both type-I and type-II errors are less than $1/8$. Therefore, the sum of type-I and type-II errors is less than $1/4$.

Rubrics: 4 points for correctly computing upper bound on type-I error.
4 points for correctly computing upper bound on type-II error.
For each case, computing mean/variance correctly gets 2 points.

Question 5. [10 points] Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}\big([\theta - 1/2, \theta + 1/2]\big)$, where $\theta \in \mathbb{R}$ is an unknown parameter. Consider the improper prior density function $\pi(\theta) = 1$ for all $\theta \in \mathbb{R}$. Find the posterior distribution of θ given the observed data $X_1, X_2, \ldots, X_n$. Find the Bayes estimator of θ under squared error loss.

Answer: The likelihood function is:

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \mathbf{1}_{[\theta-1/2, \theta+1/2]}(X_i) = \mathbf{1}_{[\max_i X_i - 1/2, \min_i X_i + 1/2]}(\theta).$$

The posterior distribution is proportional to the product of the prior and the likelihood:

$$p(\theta \mid X_1, \ldots, X_n) \propto \pi(\theta) L(\theta; X_1, \ldots, X_n) = \mathbf{1}_{[\max_i X_i - 1/2, \min_i X_i + 1/2]}(\theta).$$

Therefore, the posterior distribution of θ given the data is a uniform distribution on the interval $[\max_i X_i - 1/2, \min_i X_i + 1/2]$.

The Bayes estimator under squared error loss is the posterior mean:

$$\widehat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta \mid X_1, \ldots, X_n] = \frac{\max_i X_i - 1/2 + \min_i X_i + 1/2}{2} = \frac{\max_i X_i + \min_i X_i}{2}.$$

Rubrics: 2 points for deriving the likelihood function correctly.

3 points for deriving the posterior distribution correctly using likelihood and prior.

5 points for computing the Bayes estimator correctly.

Intermdiate steps will also be given partial credits.

Question 6. [14 points] Linear regression

Part (a). [8 points] Suppose we observe i.i.d. data (X_i, Y_i), for $i = 1, 2, \ldots, n$. Assume that

$$Y_i = X_i^\top \beta^* + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \mid X_i \sim \mathcal{N}(0, \sigma^2(X_i)),$$

where $\sigma^2(x)$ is a known function of x. Given a weight function $w : \mathbb{R}^d \to \mathbb{R}_+$, consider the weighted least-squares estimator

$$\widehat{\beta}_n^{(w)} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^n w(X_i)(Y_i - X_i^\top \beta)^2$$

Find the asymptotic distribution of $\sqrt{n}(\widehat{\beta}_n^{(w)} - \beta^*)$. Express the asymptotic variance in terms of $w(\cdot)$, $\sigma^2(\cdot)$, and the distribution of X_i.
[If you solve the one-dimensional case $d = 1$, you will get 75 % of the marks.]

Answer: The weighted least-squares estimator can be expressed as:

$$\widehat{\beta}_n^{(w)} = \left(\sum_{i=1}^n w(X_i) X_i X_i^\top \right)^{-1} \sum_{i=1}^n w(X_i) X_i Y_i.$$

Substituting $Y_i = X_i^\top \beta^* + \varepsilon_i$, we have:

$$\widehat{\beta}_n^{(w)} - \beta^* = \left(\sum_{i=1}^n w(X_i) X_i X_i^\top \right)^{-1} \sum_{i=1}^n w(X_i) X_i \varepsilon_i.$$

By the law of large numbers, we have:

$$\frac{1}{n} \sum_{i=1}^n w(X_i) X_i X_i^\top \xrightarrow{p} \mathbb{E}[w(X) X X^\top] =: A_w,$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n w(X_i) X_i \varepsilon_i \xrightarrow{d} \mathcal{N}(0, B_w),$$

where

$$B_w = \mathbb{E}[w(X)^2 X X^\top \sigma^2(X)].$$

Therefore, by Slutsky's theorem, we have:

$$\sqrt{n}(\widehat{\beta}_n^{(w)} - \beta^*) \xrightarrow{d} \mathcal{N}(0, A_w^{-1} B_w A_w^{-1}).$$

Rubrics: 4 points for deriving the expression of $\widehat{\beta}_n^{(w)} - \beta^*$ correctly.

4 points for applying the law of large numbers and central limit theorem correctly to derive the asymptotic distribution.

If you only solve the one-dimensional case, give 3 points for deriving the expression and 3 points for applying LLN and CLT.

Part (b). [6 points] Find the weight function $w(\cdot)$ that minimizes the asymptotic variance of $\widehat{\beta}_n^{(w)}$.
[Hint: try to make an educated guess based on the maximal likelihood estimator, and then verify the guess using Cauchy–Schwarz inequality.]

Answer: The MLE under the heteroscedastic model is given by:

$$\widehat{\beta}_n^{(\mathrm{MLE})} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \frac{(Y_i - X_i^\top \beta)^2}{\sigma^2(X_i)}.$$

This corresponds to choosing the weight function:

$$w(X) = \frac{1}{\sigma^2(X)}.$$

To verify that this choice minimizes the asymptotic variance, we note that for any weight function $w(X)$, the asymptotic variance is given by:

$$A_w^{-1} B_w A_w^{-1} = \Big(\mathbb{E}[w(X) X X^\top]\Big)^{-1} \mathbb{E}[w(X)^2 X X^\top \sigma^2(X)] \Big(\mathbb{E}[w(X) X X^\top]\Big)^{-1}.$$

By the Cauchy–Schwarz inequality, for any unit vector $u \in \mathbb{S}^{d-1}$, we have:

$$u^\top \mathbb{E}[w(X)^2 X X^\top \sigma^2(X)] u \geq \frac{\big(u^\top \mathbb{E}[w(X) X X^\top] u\big)^2}{\mathbb{E}[\sigma^2(X)]}.$$

Equality holds when $w(X) = c/\sigma^2(X)$ for some constant $c > 0$. Therefore, the choice $w(X) = 1/\sigma^2(X)$ minimizes the asymptotic variance of $\widehat{\beta}_n^{(w)}$.

Rubrics: 3 points for identifying the MLE weight function.
3 points for verifying the optimality using Cauchy–Schwarz inequality.
You will still get full credit if you only derive in the one-dimensional case.

Question 7. [10 points] Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$, where p is a density function supported on $[0,1]$. Suppose that $\sup_x p(x) \le p_{\max}$ for some $p_{\max} > 0$, and that p satisfies the Hölder condition

$$|p(x) - p(y)| \le L\sqrt{|x-y|}, \qquad \text{for all } x, y \in [0,1],$$

for some constants $L > 0$. Write down the kernel density estimator $\widehat{p}_n$ with a bandwidth $h > 0$ and a kernel function $K : \mathbb{R} \to \mathbb{R}$. Now we want to estimate $p(x_0)$ for some fixed $x_0 \in (0,1)$. With your choice of the kernel function K, find the optimal bandwidth h to achieve optimal rate for mean squared error $\mathbb{E}\big[|\widehat{p}_n(x_0) - p(x_0)|^2\big]$.

[Note: you can see $(L, p_{\max})$ as a fixed constant. You only need to find the rate in terms of n, and you do not need to track the constant factors.]

Answer: The kernel density estimator is defined as:

$$\widehat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

To estimate $p(x_0)$, we analyze the mean squared error (MSE):

$$\mathbb{E}\big[|\widehat{p}_n(x_0) - p(x_0)|^2\big] = \operatorname{var}(\widehat{p}_n(x_0)) + \left(\mathbb{E}[\widehat{p}_n(x_0)] - p(x_0)\right)^2.$$

The variance term is:

$$\operatorname{var}(\widehat{p}_n(x_0)) = \frac{1}{nh^2}\operatorname{var}\left(K\left(\frac{x_0 - X_1}{h}\right)\right) \le \frac{p_{\max}}{nh}\int K(u)^2 du.$$

The bias term is:

$$\mathbb{E}[\widehat{p}_n(x_0)] - p(x_0) = \int K(u)p(x_0 - hu)du - p(x_0) = \int K(u)(p(x_0 - hu) - p(x_0))du.$$

Using the Hölder condition, we have:

$$|p(x_0 - hu) - p(x_0)| \le L\sqrt{|hu|} = Lh^{1/2}|u|^{1/2}.$$

Therefore, the bias is bounded by:

$$|\mathbb{E}[\widehat{p}_n(x_0)] - p(x_0)| \le Lh^{1/2}\int |K(u)||u|^{1/2}du.$$

Choosing the kernel function $K(u) := \frac{1}{2}\mathbf{1}_{[-1,1]}(u)$, combining the variance and bias terms, we have:

$$\mathbb{E}\big[|\widehat{p}_n(x_0) - p(x_0)|^2\big] \le \frac{C_1}{nh} + C_2 h,$$

for some constants $C_1, C_2 > 0$ depending on $p_{\max}, L$. To minimize the MSE, we choose optimal bandwidth is $h_n = c_0 n^{-1/2}$, for some constant $c_0 > 0$. This gives the optimal rate for MSE as:

$$\mathbb{E}\big[|\widehat{p}_n(x_0) - p(x_0)|^2\big] \leq C n^{-1/2},$$

for some constant $C > 0$.

Rubrics: 4 points for deriving the variance term correctly.

4 points for deriving the bias term correctly.

2 points for finding the optimal bandwidth and the corresponding MSE rate.

You do not need to track constant factors, so minor mistakes in constants will not affect the score.