IMPERIAL

MATH70081

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2024

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Nonparametric Statistics

Date: Thursday, May 9, 2024

Time: 10:00 – 11:30 (BST)

Time Allowed: 1.5 hours

This paper has 3 Questions.

Please Answer All Questions in 1 Answer Booklet

Candidates should start their solutions to each question on a new sheet of paper.

Supplementary books may only be used after the relevant main book(s) are full.

Any required additional material(s) will be provided.

Allow margins for marking.

Credit will be given for all questions attempted.

Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

The open-book material allowed during the examinations consists of any material provided by the lecturers and annotated by the students, i.e. annotated lecture notes, annotated slides, and annotated problem class sheets. Books and electronic devices are not allowed.

1. Consider nonparametric regression

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with *fixed* design points $X_i = i/n \in [0,1]$ for $i = 1, \ldots, n$ and where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with $E\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Write $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Recall that *local polynomial regression* of order ℓ involves finding $\beta_0(x), \beta_1(x), \ldots, \beta_\ell(x)$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{\ell} \beta_j(x)(X_i - x)^j \right)^2 K_h(X_i - x), \tag{1}$$

where $K_h(u) = \frac{1}{h} K(u/h)$ with K a symmetric kernel.

(i) Equation (1) can be written in the form

$$(\boldsymbol{Y} - X_x \boldsymbol{\beta}_x)^T W_x (\boldsymbol{Y} - X_x \boldsymbol{\beta}_x).$$

Define the terms $X_x, \boldsymbol{\beta}_x, W_x$.

(ii) Define

$$\hat{\boldsymbol{\beta}}_x = \mathrm{argmin}_{\boldsymbol{\beta}_x} (\boldsymbol{Y} - X_x \boldsymbol{\beta}_x)^T W_x (\boldsymbol{Y} - X_x \boldsymbol{\beta}_x).$$

Show that $\hat{\boldsymbol{\beta}}_x = (X_x^T W_x X_x)^{-1} X_x^T W_x \boldsymbol{Y}$.

What is the relationship between local polynomial regression and the Nadaraya-Watson estimator?

In what follows, let $K(u) = \frac{1}{2} 1_{(-1,1]}(u)$ denote the rectangular kernel. Denote the corresponding Nadaraya-Watson estimator by $\hat{m}_n$ and the *local linear regression* estimator (i.e. local polynomial regression with $\ell = 1$) by $\hat{m}_n^1$. You are given that $\hat{m}_n^1$ can be written in the form

$$\hat{m}_n^1(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{2}$$

where the weights satisfy

$$\sum_{i=1}^{n} W_{ni}(x) = 1, \qquad \sum_{i=1}^{n} |W_{ni}(x)| \le C, \qquad \sum_{i=1}^{n} (X_i - x) W_{ni}(x) = 0 \qquad \text{for all } x$$

and $W_{ni}(x) = 0$ if $|X_i - x| \ge h$.

(iii) Let $m : [0,1] \to \mathbb{R}$ be twice differentiable with $\sup_{u \in [0,1]} |m'(u)| < M$ and $\sup_{u \in [0,1]} |m''(u)| \le M$ for some $M < \infty$. You may assume that the bandwidth $h \to 0$ is large enough that $\sum_{i=1}^{n} K_h(x - X_i) > 0$ for all $x \in [0,1]$.

(a) Show that the bias of the Nadaraya-Watson estimator at a point $x \in (h, 1-h)$ satisfies

$$|E\hat{m}_n(x) - m(x)| \le Ch$$

where C can depend on M.

*[Recall that we are in the **fixed** design setting so X_i are non-random].*

[This question continues on the next page . . .]

(b) Show that the bias of the local linear estimator $\hat{m}_n^1$ given in (2) at a point $x \in (h, 1-h)$ satisfies

$$|E\hat{m}_n^1(x) - m(x)| \leq M \sum_{i=1}^{n} |W_{ni}(x)||X_i - x|^2 \leq Ch^2,$$

where C can depend on M.

(iv) Based on your answers to (iii)(a) and (iii)(b), discuss how the degree ℓ of the local polynomial estimator based on the same kernel K affects the size of the bias when the true regression function m has increasingly many derivatives.

How could you modify the Nadaraya-Watson estimator to take advantage of increasing smoothness to reduce the bias?

[Total 25 marks]

2. Let $X_1, \ldots, X_n$ be i.i.d. random variables coming from some probability density function f on $\mathbb{R}$, and let E denote the corresponding expectation under the joint distribution of $X_1, \ldots, X_n$. For a kernel K and bandwidth $h > 0$, set $K_h(x) = h^{-1}K(x/h)$.

(i) Define the the kernel density estimator $\hat{f}_n = \hat{f}_{n,h}$ of f based on K.

For a continuously differentiable kernel K, consider estimating the *derivative* f' of f using the derivative of the kernel density estimator: $\hat{f}'_n(x) = \frac{d}{dx}\hat{f}_n(x)$. For $m = 0, 1, 2, \ldots$ consider the expected loss at a point $x \in \mathbb{R}$:

$$R_m(x) = R_m(\hat{f}_n, f, x) = E[(\hat{f}_n^{(m)}(x) - f^{(m)}(x))^2],$$

so that $R_1(x) = E[(\hat{f}'_n(x) - f'(x))^2]$ and $R_0(x) = E[(\hat{f}_n(x) - f(x))^2]$ is the mean-squared error.

(ii) Show that

$$E\hat{f}'_n(x) = [K_h * f'](x),$$

where $g * h(x) = \int_{\mathbb{R}} g(x-y)h(y)dy$ denotes the usual convolution.
Hint: use integration by parts.

(iii) Suppose now that f is bounded and twice differentiable with $\sup_{y\in\mathbb{R}} |f''(y)| \le M$ for some $0 < M < \infty$. Show that the bias satisfies

$$|E\hat{f}'_n(x) - f'(x)| \le CMh$$

for some constant $C = C(K) < \infty$ depending only on K.
Show that the variance satisfies $\mathrm{Var}(\hat{f}'_n(x)) \le \frac{C'}{nh^3}\sup_{y\in\mathbb{R}}|f(y)|$, where $C' = C'(K) < \infty$ is a constant depending only on K.

(iv) Show that as $n \to \infty$, the expected loss satisfies $R_1(x) \le cn^{-2/5}$ for a bandwidth choice $h = h_n = c'n^{-\alpha} \to 0$, where you should specify α.
[You do not need to evaluate c, c' exactly].

Recall that for a density f satisfying the conditions in (iii), the mean-squared error satisfies $R_0(x) = O(n^{-4/5})$. In view of this, briefly comment on the difficulty of estimating the derivative of a density $f'(x)$ versus its value $f(x)$ at a point x.

Let $m, \ell \in \mathbb{N}$ be integers with $0 \le m < \ell$. Consider now estimating the m^{th}-derivative $f^{(m)} = \frac{d^m}{dx^m}f(x)$ of a bounded density f that is ℓ-times differentiable with $\sup_{x\in\mathbb{R}}|f^{(\ell)}(x)| \le M < \infty$.

(v) You may assume without proof that the bias and variance of the estimator $\hat{f}_n^{(m)}$ satisfy

$$|E\hat{f}_n^{(m)}(x) - f^{(m)}(x)| \le CMh^{\ell-m}, \qquad \mathrm{Var}(\hat{f}_n^{(m)}(x)) \le \frac{1}{nh^{2m+1}}\sup_{y\in\mathbb{R}}|f(y)|\int_{\mathbb{R}} K^{(m)}(u)du.$$

Using this, find the best possible bound you can of the form $R_m(x) \le cn^{-\beta}$, where β depends on m and ℓ.

[This question continues on the next page . . .]

Note: only the dependence on the sample size n matters - you do not need to work out c precisely.

In view of the β you obtain, discuss how both m and ℓ affect the difficulty of the statistical estimation problem.

(vi) Consider selecting the bandwidth parameter h via cross-validation. Would you expect this approach to work if you are interested in estimating $f^{(m)}$ rather than f? Justify your answer.

[Total 25 marks]

3. Let $DP(\alpha, H)$ denote the Dirichlet process distribution with concentration parameter $\alpha > 0$ and base probability measure H on $\mathbb{R}$. Consider the Bayesian model with $X_1, \ldots, X_n | G \sim^{i.i.d} G$ and assign the prior $G \sim DP(\alpha, H)$ with $\alpha > 0$.

Recall the above notation means if $Z \sim H$ is random variable with distribution H, then $H(A) = P(Z \in A)$ for any measurable set A. In what follows, you may use any results from the lecture notes without proof provided these are clearly stated. Note also that if $W \sim Beta(r, s)$ then

$$EW = \frac{r}{r+s}, \qquad Var(W) = \frac{rs}{(r+s)^2(r+s+1)}.$$

(i) State the posterior distribution of $G|X_1, \ldots, X_n$. Explain how you can sample from the posterior distribution using the stick-breaking construction.

Given an interpretation for the prior concentration parameter α.

Let now H be the exponential distribution having density function $h(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ with parameter $\lambda > 0$.

(ii) Set $A = (a, \infty)$ for some $a > 0$. What is the *prior* distribution of $G(A)$? State its expectation and variance.

What is the *posterior* distribution of $G(A)$?

(iii) Consider two cases for the posterior of $G|X_1, \ldots, X_n$: (I) letting $\alpha \to 0$ or (II) letting $\lambda \to 0$. In each case, explain the interpretation behind taking these limits. Which choice seems more reasonable? Give a brief justification for your answer.

Consider estimating the *expectation* of the X_i's using the mean of G under the posterior:

$$m(G) = \int x \, dG(x),$$

i.e. we draw a random G from the *posterior* and compute $m(G)$.

Recall that for a discrete measure of the form $Q = \sum_k w_k \delta_{z_k}$, we have $\int \psi(y) dQ(y) = \sum_k w_k \psi(z_k)$.

(iv) Find the posterior mean $E[m(G)|X_1, \ldots, X_n]$ of $m(G)$ given $X_1, \ldots, X_n$.

Hint: use the explicit form of a draw from a Dirichlet process.

Being unsure about what value of λ to pick in our prior, we set λ to be the maximum likelihood estimator in the i.i.d. exponential distribution model: $\lambda = \hat{\lambda}_n = 1/\bar{X}_n$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Briefly comment on how this choice affects $E[m(G)|X_1, \ldots, X_n]$.

[Total 25 marks]

End of examination paper

Imperial College London

Module: MATH70081
Setter: K. Ray
Checker: Ernst
Editor: Varty
External: Woods
Date: March 4, 2024

MSc EXAMINATIONS (STATISTICS)

MATH70081 Nonparametric Statistics

Time: 1 hour 30 minutes

Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

Questions on next page

1. Consider nonparametric regression

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with *fixed* design points $X_i = i/n \in [0,1]$ for $i = 1, \ldots, n$ and where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with $E\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Write $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Recall that *local polynomial regression* of order ℓ involves finding $\beta_0(x), \beta_1(x), \ldots, \beta_\ell(x)$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{\ell} \beta_j(x)(X_i - x)^j \right)^2 K_h(X_i - x), \tag{1}$$

where $K_h(u) = \frac{1}{h}K(u/h)$ with K a symmetric kernel.

(i) Equation (1) can be written in the form

$$(\boldsymbol{Y} - X_x\boldsymbol{\beta}_x)^T W_x (\boldsymbol{Y} - X_x\boldsymbol{\beta}_x).$$

Define the terms $X_x, \boldsymbol{\beta}_x, W_x$.

ANSWER: (SEEN) Let $W_x = \mathrm{diag}(K_h(X_1 - x), K_h(X_2 - x), ..., K_h(X_n - x))$, $\boldsymbol{\beta}_x = (\beta_0(x), ..., \beta_\ell(x))^T$ and

$$X_x = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^\ell \\ 1 & (X_2 - x) & \cdots & (X_2 - x)^\ell \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^\ell \end{pmatrix}.$$

[4 marks]

(ii) Define

$$\hat{\boldsymbol{\beta}}_x = \mathrm{argmin}_{\boldsymbol{\beta}_x} (\boldsymbol{Y} - X_x\boldsymbol{\beta}_x)^T W_x (\boldsymbol{Y} - X_x\boldsymbol{\beta}_x).$$

Show that $\hat{\boldsymbol{\beta}}_x = (X_x^T W_x X_x)^{-1} X_x^T W_x \boldsymbol{Y}$.
What is the relationship between local polynomial regression and the Nadaraya-Watson estimator?

ANSWER: (SEEN SIMILAR) Expanding out the objective function,

$$\varphi(\boldsymbol{\beta}_x) = \boldsymbol{Y}^T W_x \boldsymbol{Y} - 2\boldsymbol{Y}^T W_x X_x \boldsymbol{\beta}_x + \boldsymbol{\beta}_x^T X_x^T W_x X_x \boldsymbol{\beta}_x.$$

Since this is a quadratic in $\boldsymbol{\beta}_x$, we know that it is minimized at its stationary point. Taking gradients,

$$\begin{aligned} \nabla\varphi(\boldsymbol{\beta}_x) = -2\boldsymbol{Y}^T W_x X_x + 2\boldsymbol{\beta}_x^T X_x^T W_x X_x &= 0 \\ \Rightarrow \boldsymbol{\beta}_x^T X_x^T W_x X_x &= \boldsymbol{Y}^T W_x X_x \\ \Rightarrow (X_x^T W_x X_x)^T \boldsymbol{\beta}_x &= X_x^T W_x \boldsymbol{Y} \\ \Rightarrow \boldsymbol{\beta}_x &= (X_x^T W_x X_x)^{-1} X_x^T W_x \boldsymbol{Y}. \end{aligned}$$

[This question continues on the next page . . .]

The Nadaraya-Watson estimator is the local polynomial regression estimator of order $\ell = 0$. [5 marks]

In what follows, let $K(u) = \frac{1}{2}1_{(-1,1]}(u)$ denote the rectangular kernel. Denote the corresponding Nadaraya-Watson estimator by $\hat{m}_n$ and the *local linear regression* estimator (i.e. local polynomial regression with $\ell = 1$) by $\hat{m}_n^1$. You are given that $\hat{m}_n^1$ can be written in the form

$$\hat{m}_n^1(x) = \sum_{i=1}^n W_{ni}(x) Y_i, \tag{2}$$

where the weights satisfy

$$\sum_{i=1}^n W_{ni}(x) = 1, \qquad \sum_{i=1}^n |W_{ni}(x)| \leq C, \qquad \sum_{i=1}^n (X_i - x) W_{ni}(x) = 0 \qquad \text{for all } x$$

and $W_{ni}(x) = 0$ if $|X_i - x| \geq h$.

(iii) Let $m : [0,1] \to \mathbb{R}$ be twice differentiable with $\sup_{u\in[0,1]} |m'(u)| < M$ and $\sup_{u\in[0,1]} |m''(u)| \leq M$ for some $M < \infty$. You may assume that the bandwidth $h \to 0$ is large enough that $\sum_{i=1}^n K_h(x - X_i) > 0$ for all $x \in [0,1]$.

(a) Show that the bias of the Nadaraya-Watson estimator at a point $x \in (h, 1-h)$ satisfies

$$|E\hat{m}_n(x) - m(x)| \leq Ch$$

where C can depend on M.
*[Recall that we are in the **fixed** design setting so X_i are non-random].*

ANSWER: (SEEN SIMILAR) Consider the Nadaraya-Watson estimator

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{i=1}^n K_h(x - X_i)}.$$

For the rectangular kernel, we can rewrite the kernel terms as

$$K_h(x - X_i) = \frac{1}{2h} 1\{-h < x - X_i \leq h\}$$

so that it is non-zero only if $|x - X_i| \leq h$. Taking expectations of the NW estimator and Taylor expanding each $m(X_i)$ about $m(x)$:

$$\begin{aligned} E\hat{m}_n(x) &= \frac{\sum_i K_h(x - X_i) m(X_i)}{\sum_i K_h(x - X_i)} \\ &= \frac{\sum_i K_h(x - X_i)[m(x) + m'(x)(X_i - x) + O(M(X_i - x)^2))]}{\sum_i K_h(x - X_i)}. \end{aligned}$$

[This question continues on the next page . . .]

But the kernel terms are all non-negative for the rectangular kernel and hence

$$\begin{aligned} |E\hat{m}_n(x) - m(x)| &\leq \frac{|m'(x)| \sum_i K_h(x - X_i)|X_i - x|}{\sum_i K_h(x - X_i)} + O(Mh^2) \\ &\leq Mh + O(M^2h^2). \end{aligned}$$

[6 marks]

(b) Show that the bias of the local linear estimator $\hat{m}_n^1$ given in (2) at a point $x \in (h, 1-h)$ satisfies

$$|E\hat{m}_n^1(x) - m(x)| \leq M \sum_{i=1}^n |W_{ni}(x)||X_i - x|^2 \leq Ch^2,$$

where C can depend on M.

ANSWER: (UNSEEN) Using the form (2) and again using a Taylor expansion,

$$\begin{aligned} E\hat{m}_n^1(x) &= \sum_{i=1}^n W_{ni}(x)m(X_i) \\ &= \sum_{i=1}^n W_{ni}(x)[m(x) + m'(x)(X_i - x) + \frac{1}{2}m''(\xi_i)(X_i - x)^2] \\ &= m(x)\sum W_{ni}(x) + m'(x)\sum W_{ni}(x)(X_i - x) + \frac{1}{2}\sum W_{ni}(x)m''(\xi_i)(X_i - x)^2 \end{aligned}$$

where ξ_i lie between X_i and x. Using the weight properties, the second term is zero and we thus deduce that

$$|E\hat{m}_n^1(x) - m(x)| \leq M \sum_{i=1}^n |W_{ni}(x)||X_i - x|^2.$$

We see that the only non-zero weights occur when $|X_i - x| \leq h$ and so we can bound the sum by $Ch^2 \sum_i |W_{ni}(x)| \leq C'h^2$ as required. [6 marks]

(iv) Based on your answers to (iii)(a) and (iii)(b), discuss how the degree ℓ of the local polynomial estimator based on the same kernel K affects the size of the bias when the true regression function m has increasingly many derivatives.

How could you modify the Nadaraya-Watson estimator to take advantage of increasing smoothness to reduce the bias?

ANSWER: (UNSEEN) We see that increasing the degree of the local polynomial allows to exploit increasing smoothness of m to reduce the bias. This can also be seen from the idea in the lecture notes that such estimators are heuristically

[This question continues on the next page . . .]

like fitting a Taylor expansion of order ℓ, which takes advantage of increasing smoothness of m.

One can use a higher order kernels to reduce the bias since the NW is a kernel based estimator. [4 marks]

[Total 25 marks]

2. Let $X_1, \ldots, X_n$ be i.i.d. random variables coming from some probability density function f on $\mathbb{R}$, and let E denote the corresponding expectation under the joint distribution of $X_1, \ldots, X_n$. For a kernel K and bandwidth $h > 0$, set $K_h(x) = h^{-1}K(x/h)$.

(i) Define the the kernel density estimator $\hat{f}_n = \hat{f}_{n,h}$ of f based on K.

ANSWER: (SEEN) A kernel is an integrable function $K : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{\mathbb{R}} K(x)dx = 1$. For a bandwidth $h > 0$, the kernel density estimator is

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

[2 marks]

For a continuously differentiable kernel K, consider estimating the *derivative* f' of f using the derivative of the kernel density estimator: $\hat{f}'_n(x) = \frac{d}{dx}\hat{f}_n(x)$. For $m = 0, 1, 2, \ldots$ consider the expected loss at a point $x \in \mathbb{R}$:

$$R_m(x) = R_m(\hat{f}_n, f, x) = E[(\hat{f}_n^{(m)}(x) - f^{(m)}(x))^2],$$

so that $R_1(x) = E[(\hat{f}'_n(x) - f'(x))^2]$ and $R_0(x) = E[(\hat{f}_n(x) - f(x))^2]$ is the mean-squared error.

(ii) Show that

$$E\hat{f}'_n(x) = [K_h * f'](x),$$

where $g * h(x) = \int_{\mathbb{R}} g(x - y)h(y)dy$ denotes the usual convolution.
Hint: use integration by parts.

ANSWER: (SEEN SIMILAR) Differentiating, we get $\hat{f}'_n(x) = h^{-2}\sum_{i=1}^{n} K'((x - X_i)/h)$. Taking the expectation and integrating by parts,

$$\begin{aligned} E\hat{f}'_n(x) &= \frac{1}{h^2}\int_{\mathbb{R}} K'\left(\frac{x-y}{h}\right) f(y)dy \\ &= \left[-\frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)\right]_{-\infty}^{\infty} + \int_{\mathbb{R}} K_h(x-y)f'(y)dy \\ &= K_h * f'(x) \end{aligned}$$

where we used that $K(\pm\infty) = 0$ since the kernel is integrable [or one can also use $f(\pm\infty) = 0$]. [3 marks]

(iii) Suppose now that f is bounded and twice differentiable with $\sup_{y\in\mathbb{R}} |f''(y)| \leq M$ for some $0 < M < \infty$. Show that the bias satisfies

$$|E\hat{f}'_n(x) - f'(x)| \leq CMh$$

[This question continues on the next page . . .]

for some constant $C = C(K) < \infty$ depending only on K.
Show that the variance satisfies $\mathrm{Var}(\hat{f}_n'(x)) \leq \frac{C'}{nh^3} \sup_{y\in\mathbb{R}} |f(y)|$, where $C' = C'(K) < \infty$ is a constant depending only on K.

ANSWER: (SEEN SIMILAR) Using the usual change of variable $u = (x-y)/h$ and that $\int_{\mathbb{R}} K(u)du = 1$,

$$\begin{aligned} |E\hat{f}_n'(x) - f'(x)| &= \left| \int_{\mathbb{R}} K(u)f'(x-uh)du - f'(x) \right| \\ &= \left| \int_{\mathbb{R}} K(u)[f'(x-uh) - f'(x)]\, du \right| \end{aligned}$$

Using the Taylor expansion $f'(x-uh) = f'(x) - f''(\xi)uh$ for some ξ between x and uh, the above equals

$$\left| \int_{\mathbb{R}} K(u)f''(\xi)uhdu \right| \leq h \sup_{y\in\mathbb{R}} |f''(y)| \int_{\mathbb{R}} |u||K(u)|du \leq C_K Mh.$$

Turning to the variance, using the i.i.d. structure and that $\mathrm{Var}(Y) \leq EY^2$,

$$\begin{aligned} \mathrm{Var}(\hat{f}_n'(x)) = \frac{1}{nh^4}\mathrm{Var}(K'((x-X_i)/h)) &\leq \frac{1}{nh^4} EK'((x-X_i)/h)^2 \\ &= \frac{1}{nh^3}\int_{\mathbb{R}} K'(u)^2 f(x-uh)du \\ &\leq \frac{1}{nh^3}\|f\|_\infty \int_{\mathbb{R}} K'(u)^2 du. \end{aligned}$$

[7 marks]

(iv) Show that as $n \to \infty$, the expected loss satisfies $R_1(x) \leq cn^{-2/5}$ for a bandwidth choice $h = h_n = c'n^{-\alpha} \to 0$, where you should specify α.
[You do not need to evaluate c, c' exactly].

Recall that for a density f satisfying the conditions in (iii), the mean-squared error satisfies $R_0(x) = O(n^{-4/5})$. In view of this, briefly comment on the difficulty of estimating the derivative of a density $f'(x)$ versus its value $f(x)$ at a point x.

ANSWER: (SEEN SIMILAR) Using the bias-variance decomposition and the answers derived above, $R_1(x) \leq CMh^2 + \frac{C'\|f\|_\infty}{nh^3}$. Differentiating the right-hand side with respect to h, we get

$$CMh - \frac{3C'\|f\|_\infty}{nh^4} = 0,$$

thereby giving $h^5 = \frac{CM}{3C'\|f\|_\infty} n^{-1}$. Thus we get $h_n \simeq n^{-1/5}$ or $\alpha = 1/5$. Plugging this back into the upper bound gives $R_1(x) \leq c'n^{-2/5}$.
We see that it is more difficult to estimate the derivative of a density than its value, since our estimation rate is slower. [5 marks]

[This question continues on the next page . . .]

Let $m, \ell \in \mathbb{N}$ be integers with $0 \leq m < \ell$. Consider now estimating the m^{th}-derivative $f^{(m)} = \frac{d^m}{dx^m} f(x)$ of a bounded density f that is ℓ-times differentiable with $\sup_{x\in\mathbb{R}} |f^{(\ell)}(x)| \leq M < \infty$.

(v) You may assume without proof that the bias and variance of the estimator $\hat{f}_n^{(m)}$ satisfy

$$|E\hat{f}_n^{(m)}(x) - f^{(m)}(x)| \leq CMh^{\ell-m}, \qquad \mathrm{Var}(\hat{f}_n^{(m)}(x)) \leq \frac{1}{nh^{2m+1}} \sup_{y\in\mathbb{R}} |f(y)| \int_{\mathbb{R}} K^{(m)}(u)du.$$

Using this, find the best possible bound you can of the form $R_m(x) \leq cn^{-\beta}$, where β depends on m and ℓ.

Note: only the dependence on the sample size n matters - you do not need to work out c precisely.

In view of the β you obtain, discuss how both m and ℓ affect the difficulty of the statistical estimation problem.

ANSWER: (UNSEEN) By the bias-variance tradeoff, we need to balance $Ch^{2(\ell-m)} + Cn^{-1}h^{2m+1}$, which is minimized by $h = h_n \simeq n^{-\frac{1}{2\ell+1}}$. Substituting this back into the expression gives the bound $n^{-\frac{2(\ell-m)}{2\ell+1}}$, i.e. $\beta = \frac{2(\ell-m)}{2\ell+1}$.

We see that the larger m the slower the rate, while the larger the ℓ the faster the rate. Moreover, we see that the problem is qualitatively different from estimating an $\ell - m$ smooth function at a point x, as one might expect, with a faster rate possible. Any reasonable discussion is fine. [5 marks]

(vi) Consider selecting the bandwidth parameter h via cross-validation. Would you expect this approach to work if you are interested in estimating $f^{(m)}$ rather than f? Justify your answer.

ANSWER: (UNSEEN) Yes. We see that the optimal bandwidth choice for estimating the derivative $f^{(m)}$ is the same as for estimating f, so a method which picks good bandwidth for estimating f should do so for estimating $f^{(m)}$.

[3 marks]

[Total 25 marks]

3. Let $DP(\alpha, H)$ denote the Dirichlet process distribution with concentration parameter $\alpha > 0$ and base probability measure H on $\mathbb{R}$. Consider the Bayesian model with $X_1, \ldots, X_n | G \sim^{i.i.d} G$ and assign the prior $G \sim DP(\alpha, H)$ with $\alpha > 0$.

Recall the above notation means if $Z \sim H$ is random variable with distribution H, then $H(A) = P(Z \in A)$ for any measurable set A. In what follows, you may use any results from the lecture notes without proof provided these are clearly stated. Note also that if $W \sim Beta(r, s)$ then

$$EW = \frac{r}{r+s}, \qquad Var(W) = \frac{rs}{(r+s)^2(r+s+1)}.$$

(i) State the posterior distribution of $G|X_1, \ldots, X_n$. Explain how you can sample from the posterior distribution using the stick-breaking construction.

Given an interpretation for the prior concentration parameter α.

ANSWER: (SEEN) The posterior has distribution

$$G|X_1, \ldots, X_n \sim DP\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{X_i}}{\alpha + n}\right).$$

Regarding the stick-breaking construction, we sample $z_1, z_2, \ldots \sim^{iid} \frac{\alpha H + \sum_{i=1}^n \delta_{X_i}}{\alpha+n}$, that is with probability $\frac{\alpha}{\alpha+n}$ we draw $z_i \sim H$ and with probability $\frac{n}{\alpha+n}$ we draw z_k equal to one of the observed values $X_1, \ldots, X_n$, each with equal probability. We then draw $V_1, V_2, \ldots \sim^{iid} \text{Beta}(1, \alpha)$ and set $W_k = V_k \prod_{l=1}^{k-1}(1 - V_l)$. We then finally take our draw as

$$G = \sum_{k=1}^{\infty} W_k \delta_{z_k}.$$

From the form of the posterior for G, we see that α represents the 'count weight' or number of observations assigned to the prior distribution. [7 marks]

Let now H be the exponential distribution having density function $h(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ with parameter $\lambda > 0$.

(ii) Set $A = (a, \infty)$ for some $a > 0$. What is the *prior* distribution of $G(A)$? State its expectation and variance.

What is the *posterior* distribution of $G(A)$?

ANSWER: (SEEN SIMILAR) For the base measure, we have $H((a, \infty)) = \int_a^\infty \lambda e^{-\lambda x} dx = e^{-\lambda a}$. Under the DP prior, we have

$$G(A) \sim \text{Beta}(\alpha e^{-\lambda a}, \alpha(1 - e^{-\lambda a})).$$

[This question continues on the next page . . .]

Using the expectation and variance of a Beta random variable (or directly expressions for the Dirichlet process from the lecture notes),

$$EG(A) = e^{-\lambda a}, \qquad \mathrm{Var}(G(A)) = \frac{e^{-\lambda a}(1 - e^{-\lambda a})}{1 + \alpha}.$$

Given the form of the posterior above, we just need to update the parameters. Let N_A denote the number of observations $X_1, \ldots, X_n$ falling in $A = (a, \infty)$. Then

$$G(A)|X_1, \ldots, X_n \sim \mathrm{Beta}(\alpha e^{-\lambda a} + N_a, \alpha(1 - e^{-\lambda a}) + (n - N_a)).$$

[7 marks]

(iii) Consider two cases for the posterior of $G|X_1, \ldots, X_n$: (I) letting $\alpha \to 0$ or (II) letting $\lambda \to 0$. In each case, explain the interpretation behind taking these limits. Which choice seems more reasonable? Give a brief justification for your answer.

ANSWER: (UNSEEN) (I) Letting $\alpha \to 0$ means we assign zero weights or counts to the prior, so draws from the posterior DP put all their mass on the observations $X_1, \cdots, X_n$. (II) Letting $\lambda \to 0$ means we put a heavy tailed prior measure as can be seen from the mean tending to infinity, "adding α" of these observations via the prior.

Either can be considered reasonable as long as this is properly justified.

[4 marks]

Consider estimating the *expectation* of the X_i's using the mean of G under the posterior:

$$m(G) = \int x \, dG(x),$$

i.e. we draw a random G from the *posterior* and compute $m(G)$.

Recall that for a discrete measure of the form $Q = \sum_k w_k \delta_{z_k}$, we have $\int \psi(y) dQ(y) = \sum_k w_k \psi(z_k)$.

(iv) Find the posterior mean $E[m(G)|X_1, \ldots, X_n]$ of $m(G)$ given $X_1, \ldots, X_n$.
Hint: use the explicit form of a draw from a Dirichlet process.

Being unsure about what value of λ to pick in our prior, we set λ to be the maximum likelihood estimator in the i.i.d. exponential distribution model: $\lambda = \hat{\lambda}_n = 1/\bar{X}_n$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Briefly comment on how this choice affects $E[m(G)|X_1, \ldots, X_n]$.

ANSWER: (UNSEEN) Using the explicit form of the Dirichlet process draws coming from the stick-breaking representation,

$$m(G) = \int x \, d\left(\sum_k W_k \delta_{z_k}\right) = \sum_k W_k z_k.$$

[This question continues on the next page . . .]

But since the $z_1, z_2, \ldots$ are i.i.d. and independent of the $W_1, W_2, \ldots$, under the Dirichlet process (including the posterior),

$$E[m(G)|X_1, \ldots, X_n] = E[z_1|X_1, \ldots, X_n]E\left[\sum_k W_k \middle| X_1, \ldots, X_n\right].$$

But $\sum_k W_k = 1$ (with probability one) and by (i)

$$E[z_1|X_1, \ldots, X_n] = \frac{\alpha}{\alpha + n}\frac{1}{\lambda} + \frac{n}{\alpha + n}\bar{X}_n,$$

which gives the posterior mean.

Taking $1/\lambda = \bar{X}_n$ gives the posterior mean to be $\bar{X}_n$. The prior is also 'centered' at the data in some sense, so the posterior is fully data-driven [7 marks]

[Total 25 marks]

MATH70081 Nonparametric Statistics

Question	Marker's comment
1	The first parts of this question were well-answered. Computing the bias of the estimators caused more issues - most candidates understood some kind of Taylor expansion was required, but sometimes struggled with the details. Several candidates also missed that that the covariates were not random, which was highlighted in the question.
2	The first half of this question was generally well-answered, with many candidates getting close to full marks on (i)-(iii). However, many struggled (or didn't attempt) the later discussion on interpreting the role of smoothness and the derivative being estimated on the difficulty of the statistical estimation problem. Few candidates correctly answered the last question on using cross-validation (one possible solution being that the optimal bandwidths derived earlier in the question did not depend on the derivative being estimated, and hence a method that works well for one problem should also work for the others. Any other reasonable discussion was also acceptable).
3	This question caused the most difficulty. The definitions/explanations were generally fine, but many candidates did not evaluate the specific prior base measure in (ii). The intrepretation parts were better answered here than in other questions (e.g. (iii)). Part (iv) caused a lot of trouble, with many candidates struggling to write the first definitions. Those that did, however, did usually make some progress.