UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 6 June, 2022 1:30 pm to 3:30 pm

PAPER 210

TOPICS IN STATISTICAL THEORY

Before you begin please read these instructions carefully

Candidates have TWO HOURS to complete the written examination.

Attempt no more than **THREE** questions.
There are **FOUR** questions in total.
The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury tag
Script paper
Rough paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Let $X_1, \ldots, X_n$ be independent random variables taking values in $[0,1]$, with $\mathbb{E}(X_i) = \mu_i \in (0,1)$ for $i = 1, \ldots, n$. Writing $\bar{X} := n^{-1}\sum_{i=1}^n X_i$ and $\bar{\mu} := n^{-1}\sum_{i=1}^n \mu_i$, prove that

$$\mathbb{P}(\bar{X} - \bar{\mu} \geqslant x) \leqslant e^{-n\mathrm{kl}(x+\bar{\mu},\bar{\mu})}$$

for every $x \in [0, 1-\bar{\mu}]$, where $\mathrm{kl}(a,b) := a\log\bigl(\frac{a}{b}\bigr) + (1-a)\log\bigl(\frac{1-a}{1-b}\bigr)$ for $a \in [0,1]$ and $b \in (0,1)$, and where $0 \log 0 := 0$.

Hence or otherwise, show that for every $\delta \geqslant 0$,

$$\mathbb{P}\bigl(\bar{X} \geqslant (1+\delta)\bar{\mu}\bigr) \leqslant \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{n\bar{\mu}} \leqslant e^{-n\delta^2\bar{\mu}/(2+\delta)}.$$

Prove further that for every $\delta \in [0,1]$,

$$\mathbb{P}\bigl(\bar{X} \leqslant (1-\delta)\bar{\mu}\bigr) \leqslant \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{n\bar{\mu}} \leqslant e^{-n\delta^2\bar{\mu}/2}.$$

[You may use the inequalities $\log(1+z) \leqslant z$ for all $z > -1$, $2z/(2+z) \leqslant \log(1+z)$ for all $z \geqslant 0$ and $z + (1-z)\log(1-z) \geqslant z^2/2$ for all $z \in [0,1)$ without proof.]

2 In the context of kernel density estimation, define what is meant by a *kernel*. If $X_1,\ldots,X_n$ are independent with density f on the real line, define a *kernel density estimator* $\hat{f}_n \equiv \hat{f}_{n,h,K}(\cdot)$ of f with bandwidth $h>0$ and kernel K.

What does it mean to say that a kernel is *of order* $\ell \in \mathbb{N}$?

Given $\delta \in (0,1]$ and $x \in \mathbb{R}$, describe Lepski's method for choosing a bandwidth $\hat{h}_\delta$ designed to ensure that $\hat{f}_{n,\hat{h}_\delta,K}(x)$ is close to $f(x)$ with probability at least $1-\delta$.

[*You may assume the existence of a bounded kernel of order ℓ that vanishes outside* $[-1,1]$, *and may find the following definitions helpful:* $\Gamma := \frac{\|K\|_\infty^2}{9R(K)}$, *where* $R(K) := \int_{-\infty}^{\infty} K^2(u)\,du$,

$$\mathcal{H}_{n,\delta} := \biggl\{\frac{2^j\Gamma}{n}\log\biggl(\frac{2\lfloor\frac{2\ell}{2\ell+1}\log_2(4n)\rfloor}{\delta}\biggr) : j=1,\ldots,\biggl\lfloor\frac{2\ell}{2\ell+1}\log_2(4n)\biggr\rfloor\biggr\},$$

and

$$\hat{\sigma}_{n,h,\delta} := \biggl(\frac{32\hat{f}_\infty R(K)\log(2|\mathcal{H}_{n,\delta}|/\delta)}{nh}\biggr)^{1/2},$$

for a suitable $\hat{f}_\infty$ that you should define.]

With $\beta \in (0,\ell]$ and $L>0$, assume that the density f belongs to the Hölder class $\mathcal{F}(\beta,L)$ and let

$$h_{\mathrm{opt}} := \biggl(D^2_{\beta,L,K}\frac{\log(2|\mathcal{H}_{n,\delta}|/\delta)}{n}\biggr)^{1/(2\beta+1)},$$

where $D_{\beta,L,K} := (\lceil\beta\rceil-1)!\sqrt{8R(K)}/\{L\mu_\beta(K)\}$ and $\mu_\beta(K) := \int_{-\infty}^{\infty}|u|^\beta|K(u)|\,du$. Explain why there exists $n_1 \equiv n_1(\beta,L,K,\delta) \in \mathbb{N}$ such that for $n \geqslant n_1$, we have $\tilde{h}_{\mathrm{opt}} := \max\bigl(\mathcal{H}_{n,\delta}\cap[0,h_{\mathrm{opt}}]\bigr) \geqslant h_{\mathrm{opt}}/2$.

Now assume further that there exist an event $\Omega_0 \equiv \Omega_0(n,\delta)$ with $\mathbb{P}(\Omega_0^c) \leqslant \delta$, as well as $A \equiv A(\beta,L)>0$ and a positive integer $N \equiv N(\beta,L,K,\delta) \geqslant n_1$ such that for $n \geqslant N$, we have on Ω_0 that $\hat{f}_\infty \leqslant A$ and

$$\bigl|\hat{f}_{n,h,K}(x)-f(x)\bigr| \leqslant \hat{\sigma}_{n,h,\delta} \quad \text{for all } h \in \mathcal{H}_{n,\delta}\cap[0,\tilde{h}_{\mathrm{opt}}].$$

Prove that for $n \geqslant N$, we have on Ω_0 that

$$\bigl|\hat{f}_{n,\hat{h}_\delta,K}(x)-f(x)\bigr| \leqslant C(\beta,L,K)\biggl(\frac{\log\bigl(\log(4n)/\delta\bigr)}{n}\biggr)^{\beta/(2\beta+1)},$$

for an appropriate $C(\beta,L,K)>0$.

Finally, assume in addition that $f(x) \leqslant A$ and that, when $\delta \geqslant n^{-3}$, we may take N to be independent of δ. Prove that there exists a data-driven bandwidth $\hat{h}$ and $C' \equiv C'(\beta,L,K)>0$ such that for every $n \in \mathbb{N}$ and $\beta \in (0,\ell]$, we have

$$\mathbb{E}\bigl[\bigl\{\hat{f}_{n,\hat{h},K}(x)-f(x)\bigr\}^2\bigr] \leqslant C'\biggl(\frac{\log(en)}{n}\biggr)^{2\beta/(2\beta+1)}.$$

3 Consider the nonparametric regression model

$$Y_i = m(x_i) + \epsilon_i,$$

for $i = 1, \ldots, n$, where $x_1 < \ldots < x_n$ are fixed and $\epsilon_1, \ldots, \epsilon_n$ are independent with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = 1$ for $i = 1, \ldots, n$. For $x \in \mathbb{R}$, write down the optimisation problem solved by the local polynomial estimator $\hat{m}_n(x) \equiv \hat{m}_n(x; p, h, K)$ of degree $p \in \mathbb{N}_0$ with bandwidth $h > 0$ and kernel K. Explain what is meant by a *linear estimator* of $m(x)$, and under a positive definiteness condition that you should specify and then assume throughout, prove that $\hat{m}_n(x)$ is such a linear estimator.

Define what is meant by the *effective kernel* $\{w_{p,i}(x) \equiv w_{p,i}(x; x_1, \ldots, x_n) : i = 1, \ldots, n\}$ of $\hat{m}_n(x)$. If R is a polynomial of degree at most p, prove that

$$\frac{1}{n} \sum_{i=1}^{n} w_{p,i}(x) R(x_i) = R(x).$$

For $\beta, L > 0$, write down the definition of the Hölder class $\mathcal{H}(\beta, L)$ of functions on $[0,1]$. Let $\tilde{\mathcal{H}}(\beta, L)$ denote the subclass of $\mathcal{H}(\beta, L)$ consisting of functions f that satisfy

$$\max_{j=0,1,\ldots,\beta_0-1} |f^{(j)}(x) - f^{(j)}(x')| \leqslant L|x - x'|$$

for all $x, x' \in [0,1]$, where $\beta_0 := \lceil \beta \rceil - 1$. Assume that $m \in \tilde{\mathcal{H}}(\beta, L)$, that $x_i = i/n$ for $i = 1, \ldots, n$ and that K is bounded and vanishes outside $[-1, 1]$. By first developing an appropriate bound on the effective kernel, or otherwise, prove that if $p < \beta_0$ and $h \geqslant 1/(2n)$, then

$$|\mathrm{Bias}\, \hat{m}_n(x; p, h, K)| \leqslant C(\lambda_0, p, L, K) h^{p+\gamma},$$

where λ_0 is the smallest eigenvalue of a matrix that you should specify, and where both $C(\lambda_0, p, L, K) > 0$ and the universal constant $\gamma > 0$ should be specified.

4 Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and let $(\mathcal{Y}, \mathcal{B})$ be a measurable space. If $g : \mathcal{X} \to \mathcal{Y}$ is measurable, define the *pushforward* measure of μ under g.

State the Lebesgue decomposition theorem. What is meant by an *f-divergence* between two probability measures P and Q on $(\mathcal{X}, \mathcal{A})$?

State and prove the data processing inequality. [You may assume the Radon–Nikodym theorem.]

Define the *χ^2-divergence* from Q to P, denoted $\chi^2(P, Q)$. Prove that if $P_1, \ldots, P_M$ are probability measures on $\mathcal{X}$ and $A_1, \ldots, A_M \in \mathcal{A}$ form a partition of $\mathcal{X}$, then

$$\frac{1}{M} \sum_{j=1}^{M} P_j(A_j) \leqslant \frac{1}{M} + \sqrt{\frac{1}{M}\left(1 - \frac{1}{M}\right)} \sqrt{\frac{1}{M} \inf_{Q \in \mathcal{Q}} \sum_{j=1}^{M} \chi^2(P_j, Q)},$$

where $\mathcal{Q}$ denotes the set of all probability distributions on $\mathcal{X}$. [You may assume that f-divergences are jointly convex.]

END OF PAPER