UNIVERSITY OF CAMBRIDGE

MAMA/210, NST3AS/210, MAAS/210

MAT3

MATHEMATICAL TRIPOS Part III

Monday 3 June 2024 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 Define what it means for a random variable X with $\mathbb{E}(X) = 0$ to be *sub-Gamma in the right tail with variance factor* $\sigma^2 > 0$ *and scale parameter* $c > 0$. Prove that if X is such a random variable, then

$$\mathbb{P}(X \geqslant x) \leqslant e^{-\frac{x^2}{2(\sigma^2 + cx)}} \leqslant \max\bigl\{e^{-x^2/(4\sigma^2)}, e^{-x/(4c)}\bigr\} \tag{1}$$

for all $x \geqslant 0$.

State and prove Bernstein's inequality.

[You may use the facts that $e^u - 1 - u \leqslant u^2/2$ for all $u \leqslant 0$, and $\log u \leqslant u - 1$ for $u > 0$.]

Let $X_1, \ldots, X_n$ be independent, real-valued random variables with distribution function F. Define the *empirical distribution function* $\mathbb{F}_n$. By using a version of the second bound in (1), or otherwise, prove that for every $x \in \mathbb{R}$ and $\delta \in (0, 1]$, we have

$$\mathbb{P}\left(\mathbb{F}_n(x) - F(x) \geqslant 2\sqrt{\frac{F(x)\bigl(1 - F(x)\bigr)\log(1/\delta)}{n}} + \frac{4\bigl(1 - F(x)\bigr)}{3n}\log(1/\delta)\right) \leqslant \delta.$$

2 Let $X_1, \ldots, X_n$ be independent, real-valued random variables with density f. Define what is meant by a *kernel* K and by a *kernel density estimator* $\hat{f}_n \equiv \hat{f}_{n,h,K}$ of f with bandwidth $h > 0$ and kernel K. What is meant by saying that a kernel is of *order* $\ell \in \mathbb{N}$?

Let $p \in [2, \infty)$, and suppose that we wish to estimate f with loss function

$$L(\hat{f}, f) := \int_{-\infty}^{\infty} |\hat{f}(x) - f(x)|^p \, dx.$$

For $\beta, L > 0$ and $m := \lceil \beta \rceil - 1$, let $\mathcal{N}_p(\beta, L)$ denote the set of $(m-1)$-times differentiable $g : \mathbb{R} \to \mathbb{R}$, for which $g^{(m-1)}$ is locally absolutely continuous, with weak derivative $g^{(m)}$ satisfying

$$\left\{ \int_{-\infty}^{\infty} \left| g^{(m)}(x+t) - g^{(m)}(x) \right|^p dx \right\}^{1/p} \leqslant L|t|^{\beta - m}$$

for all $t \in \mathbb{R}$. Further, let $\mathcal{F}_{\mathcal{N},p}(\beta, L)$ denote the set of densities that are in $\mathcal{N}_p(\beta, L)$. Prove that if $f \in \mathcal{F}_{\mathcal{N},p}(\beta, L)$, and if K is of order $\lceil \beta \rceil$, then

$$\mathbb{E} L(\hat{f}_{n,h,K}, f) \leqslant \frac{2^{2p-1} C_p R_p(K)}{(nh)^{p-1}} + \frac{2^{p-1} C_p R(K)^{p/2} \|f\|_{p/2}^{p/2}}{(nh)^{p/2}} + \frac{2^{p-1} L^p}{(m!)^p} \mu_\beta^p(K) h^{p\beta},$$

where $R_p(K) := \int_{-\infty}^{\infty} |K(u)|^p \, du$, $R(K) := \int_{-\infty}^{\infty} K(u)^2 \, du$, $\mu_\beta(K) := \int_{-\infty}^{\infty} |u|^\beta |K(u)| \, du$, $\|f\|_q := \left(\int_{-\infty}^{\infty} f(x)^q \, dx \right)^{1/q}$ for $q \in [1, \infty)$, and where C_p appears below.

[You may use the fact that if $W_1, \ldots, W_n$ are independent and identically distributed with $\mathbb{E}(W_1) = 0$, and if $p \in [2, \infty)$, then

$$\mathbb{E}\left(\left| \frac{1}{n} \sum_{i=1}^{n} W_i \right|^p \right) \leqslant C_p \left\{ \frac{\mathbb{E}(|W_1|^p)}{n^{p-1}} + \frac{\left(\mathbb{E}(W_1^2) \right)^{p/2}}{n^{p/2}} \right\},$$

where $C_p > 0$ depends only on p. You may also use the fact that if $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ are Borel measurable with $\int_{-\infty}^{\infty} |g_1(u)| \, du < \infty$ and $\int_{-\infty}^{\infty} |g_2(u)|^q \, du < \infty$ for some $q \in [1, \infty)$, then their convolution $g_1 * g_2$ is Borel measurable and satisfies

$$\left(\int_{-\infty}^{\infty} |(g_1 * g_2)(u)|^q \, du \right)^{1/q} \leqslant \left(\int_{-\infty}^{\infty} |g_1(u)| \, du \right) \left(\int_{-\infty}^{\infty} |g_2(u)|^q \, du \right)^{1/q}.$$

The inequality $(a+b)^r \leqslant 2^{r-1}(a^r + b^r)$ for $a, b \geqslant 0$ and $r \geqslant 1$ may be used without proof.]

3 Let $n \geqslant 3$ and let $a \leqslant x_1 < \ldots < x_n \leqslant b$. What is a *cubic spline* with knots at $x_1, \ldots, x_n$? What does it mean to say that such a cubic spline is *natural*?

Let $\mathcal{S}_2[a,b]$ denote the set of real-valued functions on $[a,b]$ that have an absolutely continuous first derivative. Prove that for any $(g_1, \ldots, g_n)^\top \in \mathbb{R}^n$, the natural cubic spline interpolant to $g_1, \ldots, g_n$ at $x_1, \ldots, x_n$ is the unique minimiser of $R(\tilde{g}'') := \int_a^b \tilde{g}''(x)^2 \, dx$ over all $\tilde{g} \in \mathcal{S}_2[a,b]$ that interpolate $g_1, \ldots, g_n$ at $x_1, \ldots, x_n$.

[You may assume the existence and uniqueness of such a natural cubic spline interpolant.]

Consider the nonparametric regression model

$$Y_i = g(x_i) + \sigma\epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = 1$ for $i \in [n]$. Prove that for each $\lambda > 0$, there exists a unique minimiser $\hat{g}_\lambda$ of $S_\lambda(\tilde{g}) := \sum_{i=1}^n \{Y_i - \tilde{g}(x_i)\}^2 + \lambda \int_a^b \tilde{g}''(x)^2 \, dx$ over $\tilde{g} \in \mathcal{S}_2[a,b]$, and find a closed-form expression for $\big(\hat{g}_\lambda(x_1), \ldots, \hat{g}_\lambda(x_n)\big)^\top$.

[You may assume that given $\mathbf{g} = (g_1, \ldots, g_n)^\top \in \mathbb{R}^n$, there exists a non-negative definite matrix $K \in \mathbb{R}^{n \times n}$ such that the natural cubic spline interpolant g to $g_1, \ldots, g_n$ at $x_1, \ldots, x_n$ satisfies

$$\int_a^b g''(x)^2 \, dx = \mathbf{g}^\top K \mathbf{g}.$$

]

For $\lambda > 0$, define the *cross validation score* $\mathrm{CV}(\lambda)$ in terms of the data and solutions $\hat{g}_{-i,\lambda}$, for $i \in [n]$, to optimisation problems that you should also define. For $i \in [n]$, write down a vector $\tilde{\mathbf{Y}}^{(i)} \in \mathbb{R}^n$ such that $\big(\hat{g}_{-i,\lambda}(x_1), \ldots, \hat{g}_{-i,\lambda}(x_n)\big)^\top = (I + \lambda K)^{-1} \tilde{\mathbf{Y}}^{(i)}$. Deduce that, for a given $\lambda > 0$, we can compute $\mathrm{CV}(\lambda)$ via a single natural cubic spline fit.

4 State Assouad's lemma.

Fix $n \in \mathbb{N}$ with $n \geqslant 2$, and let $x_i := i/n$ for $i \in \{0\} \cup [n]$. Let $\theta_0 := 0$ and let

$$\mathcal{C}_n := \left\{ \theta = (\theta_1, \ldots, \theta_n)^\top \in \mathbb{R}^n : \frac{\theta_{i+1} - \theta_i}{x_{i+1} - x_i} \geqslant \frac{\theta_i - \theta_{i-1}}{x_i - x_{i-1}} \text{ for } i \in [n-1] \right\}$$

denote the *convex cone* in $\mathbb{R}^n$. Consider the convex regression model

$$Y_i = \theta_i + \epsilon_i$$

for $i \in [n]$, where $\theta = (\theta_1, \ldots, \theta_n)^\top \in \mathcal{C}_n \cap [0,1]^n$ and $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0,1)$. Let $\hat{\Theta}$ denote the set of Borel measurable functions from $\mathbb{R}^n$ to $\mathbb{R}^n$, and define $\theta^* = (\theta_1^*, \ldots, \theta_n^*)^\top \in \mathcal{C}_n \cap [0,1]^n$ by $\theta_i^* := (i/n)^2$ for $i \in [n]$. By considering piecewise linear perturbations of θ^*, or otherwise, prove that there exist universal constants $c > 0$ and $n_0 \in \mathbb{N}$ such that

$$\inf_{\hat{\theta} \in \hat{\Theta}} \sup_{\theta \in \mathcal{C}_n \cap [0,1]^n} \frac{1}{n} \mathbb{E}_\theta \big(\|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|^2 \big) \geqslant c \cdot n^{-4/5}$$

for all $n \geqslant n_0$.

[You may use the fact that for each $k \in [n]$ and for $m := \lfloor n/k \rfloor$, the squared Euclidean distance between such perturbations that differ on only one segment of the form $\{i \in [n] : (j-1)k + 1 \leqslant i \leqslant jk\}$ does not depend on $j \in [m]$.]

Writing $\mathcal{M}_n$ for the monotone cone in $\mathbb{R}^n$, could there exist an estimator $\hat{\theta} \in \hat{\Theta}$ and universal constants $C > 0$, $\gamma > 4/5$ such that

$$\sup_{\theta \in \mathcal{C}_n \cap \mathcal{M}_n \cap [0,1]^n} \frac{1}{n} \mathbb{E}_\theta \big(\|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\|^2 \big) \leqslant C \cdot n^{-\gamma}$$

for all $n \in \mathbb{N}$? Justify your answer briefly.

[Pinsker's inequality may be used without proof.]

[*Hint: For $k \in \mathbb{N}$,*

$$\sum_{\ell=0}^{k-1} \ell^2 (k - 1 - \ell)^2 = \frac{1}{30} k(k-1)(k-2)(k^2 - 2k + 2).$$

]

END OF PAPER