UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 7 June, 2021 12:00 pm to 3:00 pm

PAPER 205

MODERN STATISTICAL METHODS

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

Attempt no more than ***FOUR*** *questions.*
There are ***SIX*** *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 $X \in \mathbb{R}^{n\times p}$ be a design matrix with rows $x_1, \ldots, x_n$, and $(y_1, \ldots, y_n) \in \{-1, 1\}^n$ a vector of responses. Let $\hat{\beta}$ be a solution to the following ℓ_1-penalised logistic regression problem

$$\underset{\beta\in\mathbb{R}^p}{\text{minimise}}\ \frac{1}{n}\sum_{i=1}^{n}[-y_i x_i^T\beta + \log(1+\exp(y_i x_i^T\beta))] + \lambda\|\beta\|_1.$$

(a) Derive the KKT conditions for the problem above. [You may cite any result from the course.]

(b) Prove that $X\hat{\beta}$ is unique.

(c) Define

$$\mathcal{E} = \left\{ j \in \{1, \ldots, p\} : \Big| \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1+\exp(y_i x_i^T\beta)} \Big| = n\lambda \right\}.$$

Prove that $\hat{\beta}$ is unique if $\operatorname{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$.

2 Let $X \in \mathbb{R}^{n\times p}$, where the columns of X have mean 0 and ℓ^2 norm $\sqrt{n}$, and consider a normal linear model with responses $Y = X\beta^0 + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$. Suppose that the predictors are partitioned into disjoint blocks $B_1, \ldots, B_K$ with $B_1 \cup \cdots \cup B_K = \{1, \ldots, p\}$, and let $b(j)$ denote the block to which the jth predictor belongs. Predictors in different blocks are nearly orthogonal with

$$\frac{1}{n}|X_j^T X_\ell| < \frac{\eta}{32p} \quad \text{if } b(j) \neq b(\ell),$$

for some constant $\eta > 0$. Furthermore, the predictors within a given block are linearly independent, and for any block B_k with $k = 1, \ldots, K$, the smallest eigenvalue of $n^{-1} X_{B_k}^T X_{B_k}$ is greater than η.

(a) Let $S \subseteq \{1, \ldots, p\}$ and $\hat{\Sigma} = n^{-1} X^T X$. Define the *compatibility constant* $\phi^2_{\hat{\Sigma}}(S)$.

(b) Prove that for any $S \subseteq \{1, \ldots, p\}$,

$$\phi^2_{\hat{\Sigma}}(S) \geqslant \frac{\eta}{2}.$$

(c) Let $\hat{\beta}$ be a Lasso estimator with parameter $\lambda = A\sigma\sqrt{\log p/n}$ for some $A > 0$. Show that with probability at least $1 - 2p^{-(A^2/8-1)}$, we have

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta} - \beta^0\|_1 \leqslant \frac{32A^2\sigma^2 \log p}{\eta} \frac{s}{n}$$

where s is the number of non-zero entries in β^0.

[*Throughout this question, you may use any result from the course without proof provided it is clearly stated.*]

3 (a) Let $A \in \mathbb{R}^{d\times p}$ have i.i.d. Uniform($\{-1, 1\}$) entries. Fixing $u \in \mathbb{R}^p$, prove that for $t \in (0, 1)$,

$$\mathbb{P}\left(\left|\frac{\|Au\|_2^2}{d\|u\|_2^2} - 1\right| \geqslant t\right) \leqslant 2e^{-dt^2/136}.$$

[You may cite any result from the lecture notes without proof.]

(b) Suppose we have data $u_1, \ldots, u_n \in \mathbb{R}^p$, with p large and $n \geqslant 2$. Show that for a given $t, \epsilon \in (0, 1)$ and $d > 272\log(n/\sqrt{\epsilon})/t^2$, each data point may be compressed down through $u_i \mapsto Au_i/\sqrt{d} := w_i$ whilst approximately preserving the distances between the points:

$$\mathbb{P}\left(1 - t \leqslant \frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} \leqslant 1 + t \text{ for all } i, j \in \{1, \ldots, n\},\ i \neq j\right) \geqslant 1 - \epsilon.$$

4 (a) Let $\mathcal{X}$ be a finite set. Let $(g(x))_{x\in\mathcal{X}}$ be a stochastic process with $\mathbb{E}g(x) = 0$ and $\mathbb{E}g^2(x) < \infty$ for all $x \in \mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined by

$$k(x, x') = \exp\left(-\frac{\mathrm{Var}(g(x) - g(x'))}{2\eta^2}\right),$$

for some constant $\eta^2 > 0$. Prove that k is a positive definite kernel. [You need not prove basic closure properties of kernels.]

(b) Let $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ for $i = 1, \ldots, n$. Suppose that $y_i = f^0(x_i) + \varepsilon_i$ for each $i = 1, \ldots, n$ where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$, and f^0 is an arbitrary function. Define the *kernel ridge regression estimator* $\hat{f}_\lambda$.

Suppose that the kernel matrix K with $K_{i,j} = k(x_i, x_j)$ has eigenvalues $d_1 > d_2 > \cdots > d_n > 0$. Prove that

$$\mathbb{E}\left\{\sum_{i=1}^{n}(f^0(x_i) - \hat{f}_\lambda(x_i))^2\right\} \leqslant \sigma^2 \sum_{i=1}^{n} \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda}{4}\mathbf{f}^T K^{-1}\mathbf{f}$$

where $\mathbf{f} = (f^0(x_1), \ldots, f^0(x_n))^T$. Describe the value of $\mathbf{f}$ which maximises the upper bound in this inequality over all vectors with unit Euclidean norm.

5 Let $x_1, \ldots, x_n$ be i.i.d. random vectors with a $N_d(0, \Sigma)$ distribution, where Σ has eigenvalues $1, 1/2, 1/3, \ldots, 1/d$.

Given fixed vectors $a_1, \ldots, a_L$ in the unit sphere S^{d-1}, suppose we are interested in estimating $v_\ell = \mathrm{Var}(a_\ell^T x_1)$ for each $\ell = 1, \ldots, L$. Find estimators $\hat{v}_1, \ldots, \hat{v}_\ell$ such that there is a constant C, such that whenever $\log d + 1 + \delta \leqslant n$ and $\delta > 0$,

$$\mathbb{P}\left((v_\ell - \hat{v}_\ell)^2 \leqslant C\sqrt{\frac{\log d + 1 + \delta}{n}} \text{ for all } \ell \in \{1, \ldots, L\}\right) \geqslant 1 - e^{-\delta}.$$

Prove this inequality, stating carefully any necessary result from the lecture notes.

6 (a) A positive semi-definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ is said to be η-invertible if there is an approximate inverse matrix Θ such that

$$\max_{j,k} |(\Sigma\Theta - I)_{j,k}| \leqslant \eta. \tag{1}$$

Show that any matrix Σ is 1-invertible. Show that finding the smallest value of η such that Σ is η-invertible is a convex optimisation problem, i.e. minimising a convex function over a convex set.

(b) Let $Y = X\beta^0 + \varepsilon$ where X is a design matrix in $\mathbb{R}^{n\times p}$, and $\varepsilon \sim N_p(0, I)$. Define the *Lasso estimator* $\hat{\beta}$ with regularisation parameter λ.

Suppose that using a convex optimisation algorithm, we establish that $\hat{\Sigma} = X^T X/n$ is $\sqrt{\log p/n}$-invertible, with the approximate inverse $\hat{\Theta}$. Let

$$\hat{b} = \hat{\beta} + \hat{\Theta}^T X^T (Y - X\hat{\beta})/n.$$

Show that it is possible to write

$$\sqrt{n}(\hat{b} - \beta^0) = W + \Delta$$

where W has a normal distribution which you must specify, and $\|\Delta\|_\infty \leqslant \rho(n,p)\|\hat{\beta} - \beta^0\|_1$ for some function $\rho(n,p)$ which you must specify.

(c) Take $\lambda = A\sqrt{\log p/n}$ for some $A > 0$. Consider a sequence of models with increasing dimensions n and p, and deterministic design matrices; state assumptions which guarantee that $\mathbb{P}(\|\Delta\|_\infty > cs\log p/\sqrt{n}) \to 0$ as $n \to \infty$, where s is the number of nonzero entries in β^0 and c is a constant. [You may cite any result from the lecture notes.]

END OF PAPER