Paper 4, Section II

28K Principles of Statistics

Let $g : \mathbb{R} \to \mathbb{R}$ be an unknown function, twice continuously differentiable with $|g''(x)| \leqslant M$ for all $x \in \mathbb{R}$. For some $x_0 \in \mathbb{R}$, we know the value $g(x_0)$ and we wish to estimate its derivative $g'(x_0)$. To do so, we have access to a pseudo-random number generator that gives $U_1^*, \dots, U_N^*$ i.i.d. uniform over $[0,1]$, and a machine that takes input $x_1, \dots, x_N \in \mathbb{R}$ and returns $g(x_i) + \varepsilon_i$, where the ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$.

(a) Explain how this setup allows us to generate N independent $X_i = x_0 + hZ_i$, where the Z_i take value 1 or -1 with probability $1/2$, for any $h > 0$.

(b) We denote by Y_i the output $g(X_i) + \varepsilon_i$. Show that for some independent $\xi_i \in \mathbb{R}$

$$Y_i - g(x_0) = hZ_i\, g'(x_0) + \frac{h^2}{2} g''(\xi_i) + \varepsilon_i\,.$$

(c) Using the intuition given by the least-squares estimator, justify the use of the estimator $\hat{g}_N$ given by

$$\hat{g}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{Z_i (Y_i - g(x_0))}{h}\,.$$

(d) Show that

$$\mathbb{E}[|\hat{g}_N - g'(x_0)|^2] \leqslant \frac{h^2 M^2}{4} + \frac{\sigma^2}{Nh^2}\,.$$

Show that for some choice h_N of parameter h, this implies

$$\mathbb{E}[|\hat{g}_N - g'(x_0)|^2] \leqslant \frac{\sigma M}{\sqrt{N}}\,.$$

Paper 3, Section II

28K Principles of Statistics

In the model $\{\mathcal{N}(\theta, I_p), \theta \in \mathbb{R}^p\}$ of a Gaussian distribution in dimension p, with unknown mean θ and known identity covariance matrix I_p, we estimate θ based on a sample of i.i.d. observations $X_1, \ldots, X_n$ drawn from $\mathcal{N}(\theta_0, I_p)$.

(a) Define the *Fisher information* $I(\theta_0)$, and compute it in this model.

(b) We recall that the *observed Fisher information* $i_n(\theta)$ is given by

$$i_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log f(X_i, \theta) \, \nabla_\theta \log f(X_i, \theta)^\top .$$

Find the limit of $\hat{i}_n = i_n(\hat{\theta}_{MLE})$, where $\hat{\theta}_{MLE}$ is the maximum likelihood estimator of θ in this model.

(c) Define the *Wald statistic* $W_n(\theta)$ and compute it. Give the limiting distribution of $W_n(\theta_0)$ and explain how it can be used to design a confidence interval for θ_0.

[*You may use results from the course provided that you state them clearly.*]

Paper 2, Section II

28K Principles of Statistics

We consider the model $\{\mathcal{N}(\theta, I_p), \theta \in \mathbb{R}^p\}$ of a Gaussian distribution in dimension $p \geqslant 3$, with unknown mean θ and known identity covariance matrix I_p. We estimate θ based on one observation $X \sim \mathcal{N}(\theta, I_p)$, under the loss function

$$\ell(\theta, \delta) = \|\theta - \delta\|_2^2 .$$

(a) Define the *risk* of an estimator $\hat{\theta}$. Compute the maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ and its risk for any $\theta \in \mathbb{R}^p$.

(b) Define what an *admissible estimator* is. Is $\hat{\theta}_{MLE}$ admissible?

(c) For any $c > 0$, let $\pi_c(\theta)$ be the prior $\mathcal{N}(0, c^2 I_p)$. Find a Bayes optimal estimator $\hat{\theta}_c$ under this prior with the quadratic loss, and compute its Bayes risk.

(d) Show that $\hat{\theta}_{MLE}$ is minimax.

[*You may use results from the course provided that you state them clearly.*]

Paper 1, Section II

29K Principles of Statistics

A scientist wishes to estimate the proportion $\theta \in (0,1)$ of presence of a gene in a population of flies of size n. Every fly receives a chromosome from each of its two parents, each carrying the gene A with probability $(1-\theta)$ or the gene B with probability θ, independently. The scientist can observe if each fly has two copies of the gene A (denoted by AA), two copies of the gene B (denoted by BB) or one of each (denoted by AB). We let $n_{\mathrm{AA}}, n_{\mathrm{BB}}$, and n_{AB} denote the number of each observation among the n flies.

(a) Give the probability of each observation as a function of θ, denoted by $f(X,\theta)$, for all three values $X = \mathrm{AA}$, BB, or AB.

(b) For a vector $w = (w_{\mathrm{AA}}, w_{\mathrm{BB}}, w_{\mathrm{AB}})$, we let $\hat{\theta}_w$ denote the estimator defined by

$$\hat{\theta}_w = w_{\mathrm{AA}}\frac{n_{\mathrm{AA}}}{n} + w_{\mathrm{BB}}\frac{n_{\mathrm{BB}}}{n} + w_{\mathrm{AB}}\frac{n_{\mathrm{AB}}}{n}\,.$$

Find the unique vector w^* such that $\hat{\theta}_{w^*}$ is unbiased. Show that $\hat{\theta}_{w^*}$ is a consistent estimator of θ.

(c) Compute the maximum likelihood estimator of θ in this model, denoted by $\hat{\theta}_{MLE}$. Find the limiting distribution of $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$. [You may use results from the course, provided that you state them clearly.]