Version MAG/2

EGT3
ENGINEERING TRIPOS PART IIB

Thursday 30 April 2026 14.00 to 15.40

Module 4M24

COMPUTATIONAL STATISTICS AND MACHINE LEARNING

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Write on single-sided paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 Consider a random vector $\mathbf{x} \in \mathbb{R}^d$ having Lebesgue probability density which is multivariate Gaussian with zero-mean and identity covariance such that $p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. An estimate of the expectation $\mathbb{E}_p\{f(\mathbf{x})\} = \int_{\mathbb{R}^d} f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ can be obtained using an Importance Sampling estimator where the Importance Density is $q(\mathbf{x}) = \mathcal{N}(\mathbf{1}_d, \mathbf{I}_d)$, with $\mathbf{1}_d$ a d-dimensional vector whose elements all have value one, and the Importance Weights are defined as $w(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}$.

(a) Show that the Importance Weights for samples $\mathbf{x} \sim q(\mathbf{x})$ takes the form $\exp(-\alpha d)$ when taken at the mean of the importance distribution and find the value of the positive non-zero constant α. [30%]

(b) For N samples from the Importance Distribution the Effective Sample Size (ESS) is a way to measure the efficiency of a Monte Carlo estimator and is given by the expression $N\mathbb{E}_q\{w(\mathbf{x})\}^2/\mathbb{E}_q\{w(\mathbf{x})^2\}$. Show that ESS $\approx N\exp(-d)$. [40%]

(c) Discuss the implications on the efficiency of the Importance Sampling Estimator from the expressions derived in part (a) and part (b) above when $d \to \infty$. [30%]

2 A Langevin diffusion on $\mathbb{R}$ has an invariant distribution $\pi(x)$ with respect to Lebesgue measure and is defined such that each $x_t \in \mathbb{R}$ evolves as

$$dx_t = b(x_t, \alpha)dt + \sqrt{2}dW_t = \alpha \tanh(\alpha x_t)dt - \alpha x_t dt + \sqrt{2}dW_t$$

where $\alpha \in \mathbb{R}$ is a constant scalar coefficient, and W_t is standard Brownian motion.

(a) From the Fokker-Planck equation for a Langevin diffusion on $\mathbb{R}$ defined as

$$\frac{\partial}{\partial t}\pi_t(x) = -\frac{\partial}{\partial x}(b(x, \alpha)\pi_t(x)) + \frac{\partial^2}{\partial x^2}\pi_t(x)$$

derive the invariant distribution $\pi(x) = \pi_{t=\infty}(x)$ and represent this as a mixture of Gaussian distributions. [40%]

(b) Define the conditions which the coefficient α must satisfy for the Langevin diffusion to have an exponential rate of convergence to the invariant distribution $\pi(x)$. [30%]

(c) Discuss and explain the effect of the coefficient α not satisfying the conditions required in both part (a) and part (b). How would this affect the practical performance of a Metropolis Adjusted Langevin Algorithm (MALA) ? [30%]

3 For a random vector $\mathbf{x} \in \mathbb{R}^n$ with a parametric probability density $p_\theta(\mathbf{x})$, the score function is defined as $s_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$. Given an unknown probability density $p(\mathbf{x})$ which we would like to approximate, *score matching* minimises the divergence between the unknown and approximate parametric score functions, based on samples from the unknown true distribution.

In this divergence appears the term:

$$\operatorname{tr}\left(\nabla_{\mathbf{x}} s_\theta(\mathbf{x})\right),$$

where $\operatorname{tr}(\cdot)$ is the trace operator. The computation of the trace of the Jacobian, denoted $J = \nabla_{\mathbf{x}} s_\theta(\mathbf{x})$ can be very expensive for large n. An estimator for this term can be generated using random projections where each $\mathbf{v} \in \mathbb{R}^n$ such that

$$\mu = \mathbb{E}_{\mathbf{v} \sim p(\mathbf{v})}\left[\mathbf{v}^\top J \mathbf{v}\right],$$

(a) Given the projections are sampled from a standard n-dimensional Gaussian, i.e. $p(\mathbf{v}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, show that $\mu = \operatorname{tr}(J)$. [20%]

(b) Consider the Monte Carlo estimator $\hat{\mu}$ of μ:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}^{(i)\top} J \mathbf{v}^{(i)}, \quad \mathbf{v}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$$

(i) Show the estimator $\hat{\mu}$ is unbiased. [10%]

(ii) Show that the variance of the estimator is given by

$$\operatorname{Var}(\hat{\mu}) = \frac{2}{N} \operatorname{tr}\left(J^2\right)$$

Hint: Isserlis's theorem states that for zero-mean Gaussian variables v_i, v_j, v_k, v_l:

$$\mathbb{E}\left[v_i v_j v_k v_l\right] = \mathbb{E}\left[v_i v_j\right]\mathbb{E}\left[v_k v_l\right] + \mathbb{E}\left[v_i v_k\right]\mathbb{E}\left[v_j v_l\right] + \mathbb{E}\left[v_i v_l\right]\mathbb{E}\left[v_j v_k\right]$$

[40%]

(c) Suppose the parametric density $p_\theta(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \theta^2 \mathbf{I}_n)$. From the Central Limit Theorem for Monte Carlo estimators, determine the distribution of the estimator $\hat{\mu}$ in terms of θ, the number of projections N, and dimension n. [30%]

4 Bayesian Machine Learning often needs Monte Carlo estimates with respect to distributions of the form:

$$p(\theta, \phi|x)$$

where x is the data, and θ, ϕ are the model variables.

(a) Provide a detailed description of the Gibbs sampler for such a distribution and derive the acceptance probability of the sampler. [40%]

(b) Suppose $x_1, \ldots, x_N$ are iid random variables and each x_i are Gaussian distributed with mean μ and precision (inverse variance) τ:

$$x_i \sim \mathcal{N}(\mu, \tau^{-1})$$

The Bayesian model has priors on the mean and precision given by:

$$\mu \sim \mathcal{N}(0, w^{-1})$$
$$\tau \sim \text{Gamma}(\alpha, \beta)$$

Where α, β, w are fixed. The form of the Gamma distribution is given by:

$$\text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}$$

Denoting $x_1, \ldots, x_N$ as $\mathbf{x}$, define the Gibbs sampler for $p(\mu, \tau|\mathbf{x})$, giving both relevant full conditional distributions. [60%]

END OF PAPER

THIS PAGE IS BLANK