NATIONAL UNIVERSITY OF SINGAPORE

Mathematics PhD Qualifying Exam Paper 4

Stochastic Processes and Machine Learning

(August 2024)

Time allowed: 3 hours

INSTRUCTIONS TO CANDIDATES

1. Please write your matriculation/student number only. Do not write your name.
2. Including this page, the examination paper comprises **4** printed pages.
3. At the top right corner of every page of your answer script, write the question and page numbers(eg. Q1 P1, Q1 P2, Q2 P1,. . .).
4. This examination contains **EIGHT (8)** questions. Answer all of them. **Properly justify** your answers.
5. There is a total of **ONE HUNDRED (100)** points. The points for each question are indicated at the beginning of the question.
6. This is an OPEN BOOK exam. No electronic device (such as calculator, tablet, laptop or phone) is allowed. You need to have your reference materials in hard copy with you.
7. A list containing information on the probability density / mass function, mean, variance and moment generating functions of some common distributions has been provided on the other side for possible consultation.
8. Please start each question on a new page.

---

1. **(10 points)** Let $\{X_i\}_{i=1}^n$ be i.i.d. exponential random variables with parameter $\lambda > 0$. Let $Y = \max\{\frac{1}{X_1}, \ldots, \frac{1}{X_n}\}$. Find the probability density function of Y.

2. **(12 points)** Let $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ be a d-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix Σ. Consider deterministic real numbers $a_1, a_2, \ldots, a_d$. Let $k \geq 1$ be an integer. Compute the following quantities in terms of k, $(a_i)_{i=1}^d$ and $(\Sigma_{ij})_{1\leq i,j\leq d}$. Set $a(\mathbf{X}) = a_1X_1 + a_2X_2 + \cdots + a_dX_d$. Compute the following quantities:
   - **(a) (6 points)** The moment generating function of $a(\mathbf{X})$, that is, $\mathbb{E}[e^{\theta\cdot a(\mathbf{X})}]$ for $\theta \in \mathbb{R}$
   - **(b) (6 points)** $\mathbb{E}\left[a(\mathbf{X})^k\right]$

3. **(18 points)** Let $\{\mathbf{u}_i\}_{i=1}^\infty$ be i.i.d. random vectors that are uniformly distributed on the d-dimensional unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. Define the matrices $\{\mathcal{P}_i\}_{i=1}^\infty$ as $\mathcal{P}_1 = \mathbf{u}_1\mathbf{u}_1^\top$ and $\mathcal{P}_i = \mathcal{P}_{i-1} + \mathbf{u}_i\mathbf{u}_i^\top$ for $i \geq 2$. For a non-zero vector $\mathbf{v} \in \mathbb{R}^d$, define the real valued random variables $Q_i(\mathbf{v}) = \mathbf{v}^\top\mathcal{P}_i\mathbf{v}$.
   - **(a) (8 points)** Show that the sequence of random variables $\{Q_n(\mathbf{v}) - \frac{n}{d}\cdot\|\mathbf{v}\|_2^2\}_{n=1}^\infty$ is a martingale (with respect to its own filtration).
   - **(b) (10 points)** Find a positive real sequence $\{b_n\}_{n=1}^\infty$ such that $b_n^{-1}\left(Q_n(\mathbf{v}) - \frac{n}{d}\cdot\|\mathbf{v}\|_2^2\right)$ converges to a non-degenerate random variable as $n \to \infty$, and identify the distribution of this limiting random variable including its mean and variance (in terms of $\|\mathbf{v}\|_2$ and universal constants). Properly justify your answer.

4. **(10 points)** Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a bounded continuous function, and $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be i.i.d. real valued random variables with the same distribution. Prove that, for any $\theta > 0$

$$\mathbb{P}\left[\left|\left(\frac{1}{n}\sum_{i=1}^n f(X_i)\right) - \left(\frac{1}{n}\sum_{i=1}^n f(Y_i)\right)\right| \geq \theta\right] \leq 2\exp\left(-\frac{n\theta^2}{8\|f\|_\infty^2}\right).$$

5. **[Linear PCA] (10 points)** In linear PCA, the covariance matrix of the data $\mathbf{C} = \mathbf{X}^\top\mathbf{X}$ is decomposed into weighted sums of its eigenvalues (λ) and eigenvectors $\mathbf{p}$:

$$\mathbf{C} = \sum_i \lambda\mathbf{p}_i\mathbf{p}_i^\top. \tag{1}$$

   Prove mathematically that the first eigenvalue λ_1 is identical to the variance obtained by projecting data into the first principal component $\mathbf{p}_1$.

6. **[Kernel Method] (15 points)** Consider a parametric model governed by the parameter vector $\mathbf{w}$ together with a dataset of input values $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and a nonlinear feature mapping $\phi(\mathbf{x})$. Suppose that the dependence of the error function on $\mathbf{w}$ takes the form

$$J(\mathbf{w}) = f(\mathbf{w}^\top\phi(\mathbf{x}_1), \ldots, \mathbf{w}^\top\phi(\mathbf{x}_N)) + g(\mathbf{w}^\top\mathbf{w}) \tag{2}$$

where $g(\cdot)$ is a monotonically increasing function. By writing $\mathbf{w}$ in the form

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n \phi(\mathbf{x}_n) + \mathbf{w}_\perp \tag{3}$$

where $\mathbf{w}_\perp^\top \phi(\mathbf{x}_n) = 0$ for all n, show that the value of $\mathbf{w}$ that minimizes $J(\mathbf{w})$ takes the form of a linear combination of the basis function $\phi(\mathbf{x}_n)$ for $n = 1, \ldots, N$.

7. **[Two-layer Network] (10 points)** Consider a two-layer network function of the form

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{4}$$

where the set of all weight and bias parameters have been grouped together into a vector $\mathbf{w}$ and σ is the logistic sigmoid function given by:

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{5}$$

Show that there exists an equivalent network, which computes exactly the same function, but with hidden unit activation functions given by $tanh(a)$ where the $tanh$ function is defined by:

$$tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}. \tag{6}$$

8. **[Markov Decision Process] (15 points)** In the game micro-blackjack, you repeatedly draw a number (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw (a number) or Stop if the sum of the numbers you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your reward equals the sum if the sum is less than 6, or equals 0 if the sum is 6 or higher. When you Draw, you receive no reward for that action and continue to the next round. The game ends only when you Stop. There is no discount ($\gamma = 1$).
   (a) (2 points) What is the state space for this MDP?
   (b) (3 points) What is the reward function for this MDP?
   (c) (5 points) Apply one iteration of Value Iteration on this MDP. Let the initial estimate $V^0(x)$ be set to $V^0(x) = \max_a R(x, a)$, where a is an action and $R(x, a)$ is the reward function. Clearly state your answer for $V^1(x)$.
   (d) (5 points) What is the optimal policy for this MDP?

End of Paper

- Bernoulli (p) :
$\mathbb{P}(X=i) = \begin{cases} p \text{ if } i=1 \\ 1-p \text{ if } i=0. \end{cases}$
$\mathbb{E}[X]=p, \quad \mathrm{Var}[X]=p(1-p), \quad \mathbb{E}[e^{tX}]=(1-p)+pe^t.$

- Binomial (n,p):
$\mathbb{P}(X=i)=\binom{n}{i}p^i(1-p)^{n-i}; 0\le i\le n.$
$\mathbb{E}[X]=np, \quad \mathrm{Var}[X]=np(1-p), \quad \mathbb{E}[e^{tX}]=[(1-p)+pe^t]^n.$

- Geometric (p) :
$\mathbb{P}(X=i)=(1-p)^{i-1}p; i\ge 1.$
$\mathbb{E}[X]=\frac{1}{p}, \quad \mathrm{Var}[X]=\frac{1-p}{p^2}, \quad \mathbb{E}[e^{tX}]=\frac{pe^t}{1-(1-p)e^t}$ for $t<-\log(1-p)$.

- Poisson (λ):
$\mathbb{P}(X=i)=e^{-\lambda}\frac{\lambda^i}{i!}; i\ge 1.$
$\mathbb{E}[X]=\lambda, \quad \mathrm{Var}[X]=\lambda, \quad \mathbb{E}[e^{tX}]=\exp(\lambda(e^t-1)).$

- Uniform (a,b) :
$f(x) = \begin{cases} \frac{1}{b-a} \text{ if } a\le x\le b \\ 0 \text{ otherwise }. \end{cases}$
$\mathbb{E}[X]=(a+b)/2, \quad \mathrm{Var}[X]=\frac{(b-a)^2}{12}, \quad \mathbb{E}[e^{tX}]=\frac{e^{tb}-e^{ta}}{t(b-a)}$ if $t\ne 0$.

- Uniform on the square $(a,b)\times(c,d)$:
$f(x,y) = \begin{cases} \frac{1}{(b-a)(d-c)} \text{ if } a\le x\le b, c\le y\le d \\ 0 \text{ otherwise }. \end{cases}$

- Normal / Gaussian ($N(\mu,\sigma^2)$):
$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}).$
$\mathbb{E}[X]=\mu, \quad \mathrm{Var}[X]=\sigma^2, \quad \mathbb{E}[e^{tX}]=\exp(\mu t+\frac{1}{2}\sigma^2t^2).$

- Exponential (λ):
$f(x) = \begin{cases} \lambda\exp(-\lambda x) \text{ if } x>0 \\ 0 \text{ otherwise.} \end{cases}$
$\mathbb{E}[X]=1/\lambda, \quad \mathrm{Var}[X]=1/\lambda^2, \quad \mathbb{E}[e^{tX}]=\frac{\lambda}{\lambda-t}$ for $t<\lambda$.