NATIONAL UNIVERSITY OF SINGAPORE

Mathematics PhD Qualifying Exam Paper 4

Stochastic Processes and Machine Learning

(August 2023)

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. This examination contains **FIVE (5)** questions. Answer all of them. **Properly justify** your answers.
4. There is a total of **ONE HUNDRED (100)** points. The points for each question are indicated at the beginning of the question.
5. 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, . . .).
6. Please start each part of a question (i.e., (a), (b), etc.) on a new page. Answer all parts of a question together.
7. 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.
8. 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.

Q1 (20 points) Let $[N] = \{1, \dots, N\}$ and $\mathcal{S}_k$ be the collection of all subsets of $[N]$ of size k, where $1 \le k \le N$.

We define a random dynamics on $\mathcal{S}_k$ as follows. We start with $\mathbb{X}_0 := \{1, \dots, k\} \in \mathcal{S}_k$. For $n \ge 1$, we obtain $\mathbb{X}_n$ from $\mathbb{X}_{n-1}$ in the following manner. With probability $1/2$, we leave $\mathbb{X}_{n-1}$ unchanged (i.e., set $\mathbb{X}_n = \mathbb{X}_{n-1}$), and with probability $1/2$ we exchange one element of $\mathbb{X}_{n-1}$ (chosen uniformly at random) with one element of $[N] \setminus \mathbb{X}_{n-1}$ (also chosen uniformly at random).

For any subset $A \subseteq [N]$, we define the *mean* $\mathcal{M}(A) = \frac{1}{|A|} \sum_{x \in A} x$, and the *correlation function* $\rho_n(A) := \mathbb{P}[A \subseteq \mathbb{X}_n]$.

- **(a) (6 points)** Justify that the limits in parts (b) and (c) below exist.
- **(b) (7 points)** Calculate the limiting mean $\lim_{n\to\infty} \mathbb{E}[\mathcal{M}(\mathbb{X}_n)]$.
- **(b) (7 points)** For any fixed subset $A \subseteq [N]$, calculate the limiting correlation function $\lim_{n\to\infty} \rho_n(A)$.

Q2 (20 points) Let $\chi_1, \dots, \chi_n$ be sets, and $f : \chi_1 \times \dots \times \chi_n \to \mathbb{R}$ be a function such that $\forall 1 \le i \le n$ there is a $\Delta_i > 0$ such that

$$\sup_{x,y \in \chi_i} |f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n)| \le \Delta_i$$

for all possible choices $x_j \in \chi_j; j \ne i$.

Let $(X_i)_{i=1}^n$ be random variables, with each X_i taking value in the set χ_i. For $1 \le i \le n$, consider the random variables $Y_i := \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_i]$.

- **(a) (6 points)** Show that $|Y_i - Y_{i-1}| \le \Delta_i$ a.s., $\forall 1 \le i \le n$.
- **(b) (4 points)** Show that, for $t \ge 0$, we have

$$\mathbb{P}[|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \ge t] \le 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^n \Delta_i^2}\right).$$

- **(b) (10 points)** Consider the standard square lattice $\mathbb{Z}^2$ endowed with edges connecting neighbouring points, i.e. the point $(x, y) \in \mathbb{Z}^2$ is connected by an edge each to the points $(x \pm 1, y \pm 1)$. Suppose each edge e in this graph is endowed with a random weight $w(e)$ that is a uniform random variable in the interval $[0, 1]$; the random weights are i.i.d. across the edges. An *upright path* in this graph is a directed path that starts from the origin $(0, 0)$ and moves to a neighbouring lattice site either upwards (ie due north) or to the right (ie due east). For an upright path $\mathcal{P}$ of finite length, we define the weight $w(\mathcal{P}) := \sum_{e \in \text{Edges}(\mathcal{P})} w(e)$. For $n \in \mathbb{Z}$, let the random variable W_n denote the maximum weight of an upright path from $(0, 0)$ to (n, n). Show that, for $t \ge 0$ we have

$$\mathbb{P}[|W_n - \mathbb{E}[W_n]| \ge t] \le 2 \exp(-\frac{t^2}{2n}).$$

Q3 (10 points) Let $\mathbb{D}$ be the closed unit disk in $\mathbb{R}^2$, with centre 0 and radius 1. Let $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ be random points that are distributed uniformly and independently in $\mathbb{D}$. Consider the random set $\mathbb{A}_n \subset \mathbb{D}$ consisting of all points $z \in \frac{1}{2} \cdot \mathbb{D}$ that are closer to 0 than to any of the points $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$. Calculate $\mathbb{E}[\text{Area}(\mathbb{A}_n)]$.

Q4 (25 points) Consider a binary classification problem with a training sample $\mathcal{D} = \{(x_i, y_i) : i = 1, \dots, n\}$ and a predictor $\hat{h}_n$ obtained as the output of some learning algorithm, i.e. $\hat{h}_n = \mathcal{A}(\mathcal{D}, \mathcal{H})$, where $\mathcal{A}$ is the algorithm (e.g. SGD for neural networks) and $\mathcal{H}$ is the hypothesis class.

Given the training data and the hypothesis space, the generalization risk is given by $R(\hat{h}_n) = \mathbb{E}_{(X,Y)\sim\mu}\left[1_{Y \neq \hat{h}_n(X)}\right]$, where μ is the underlying probability distribution from which $\mathcal{D}$ is sampled. The risk $R(\hat{h}_n)$ is a random variable that depends on $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{H}$. Assume that the predictor $\hat{h}_n$ is consistent with data $\mathcal{D}$, that is $R_{emp}(\hat{h}_n) = \frac{1}{n}\sum_{i=1}^n 1_{\{y_i \neq \hat{h}_n(x_i)\}} = 0$. We are interested in its tail distribution, i.e. finding a bound which holds with large probability:

$$\mathbb{P}(R(\hat{h}_n) \geq \epsilon) \leq \delta.$$

The basic idea is to set the probability of being misled to δ and find a suitable ϵ to the satisfy the inequality above.

Consider the case of finite hypothesis space $\mathcal{H} = \{h_1, h_2, \dots, h_m\}$ for some $m \geq 1$.

- (a) (10 points) Show that for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$R(\hat{h}_n) \leq \frac{\log(m) + \log(\frac{1}{\delta})}{n}.$$

- (b) (15 points) Now we want to assign a weight $w_h \in (0, 1)$ to each of the predictors $h \in \mathcal{H}$ such that $\sum_{h\in\mathcal{H}} w_h = 1$. By carefully choosing ϵ such that $\mathbb{P}(R(\hat{h}_n) \geq \epsilon) \leq w_h\delta$ for all $h \in \mathcal{H}$, show that for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$R(\hat{h}_n) \leq \frac{\log\left(\frac{1}{\min_{h\in\mathcal{H}} w_h}\right) + \log(\frac{1}{\delta})}{n}.$$

  Compare this bound with the one in question (a). Give an interpretation to the result.

Q5 (25 points)

Consider a discrete-time Markov decision process with finite state space $\mathcal{S}$ and action space $\mathcal{A}$. We use the usual notation of $\{S_t, A_t, R_t\}$ to denote the state, action and reward at time t respectively. Let the transition probability kernel be $p(s', r \mid s, a) = P[S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a]$.

- (a) (2 points) Define the value function v_π with respect to a policy π. You may assume that we consider a discount rate of $0 < \gamma < 1$ when computing the returns.
- (b) (3 points) Write down the Bellman's optimality equation that the value function corresponding to an optimal policy should satisfy.
- (c) (10 points) Show that there exists a unique solution to the Bellman's optimality equation.
- (d) (10 points) Is an optimal policy always unique ? If so, prove this statement. If not, give a counterexample.

—— 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 $t \neq 0$.

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

- Uniform on the disk in $\mathbb{R}^2$ with centre z_0 and radius r :
  $f(z)=\begin{cases} \frac{1}{\pi r^2} \text{ if } |z-z_0| \le r \\ 0 \text{ otherwise .} \end{cases}$

- 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 $t<\lambda$.