NATIONAL UNIVERSITY OF SINGAPORE

Mathematics PhD Qualifying Exam Paper 4

Stochastic Processes and Machine Learning

(August 2022)

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 **5** 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 **FIVE (5)** 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. Only non-programmable and non-graphing calculators are allowed.
7. You are not allowed to use any other electronic device (such as tablet, laptop or phone). 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.
9. Please start each part of a question (i.e., (a), (b), etc.) on a new page.

Q 1 (15 points) Let $L \in \mathbb{N}$ and $\mathbb{Z}/2L\mathbb{Z}$ denote the integers $\{0, 1, \ldots, 2L - 1\}$ (equipped with addition modulo $2L$). Consider the set $V = (\mathbb{Z}/2L\mathbb{Z})^d$, that is, each $v \in V$ is of the form $v = (v_1, \ldots, v_d)$, with each $v_i \in \mathbb{Z}/2L\mathbb{Z}$. For any element $v \in V$, let $N_{\text{even}}(v)$ denote the number of co-ordinates of v that are even numbers.

A particle starts moving on the set V according to the following rules. Let the particle be at $X_n \in V$ after n steps. Then, for the $n+1$-th step, we pick a co-ordinate i uniformly at random from the set $\{1, \ldots, d\}$ for possible updating. Then, with probability $1/2$, the particle stays at its current location (that is, we set $X_{n+1} = X_n$), whereas with probability $1/2$ the i-th co-ordinate $(X_n)_i$ is updated to an independently and uniformly chosen element $\in Z/2L\mathbb{Z}$ (i.e., $(X_{n+1})_i = U$, where U is uniformly chosen from $Z/2L\mathbb{Z}$ and independent of everything else, and $(X_{n+1})_j = (X_n)_j$ for $j \neq i$).

If the particle starts from $(0, \ldots, 0) \in V$, then calculate the limit

$$\lim_{n \to \infty} \mathbb{E}\left[N_{\text{even}}(X_n)\right].$$

Q 2 (15 points) Let $f : [0, 1] \mapsto \mathbb{R}$ be a continuous function. Let $\{U_i\}_{i=0}^{\infty}$ be independent and identically distributed (i.i.d.) random variables uniformly distributed on the interval $[0, 1]$. For each $N \geq 1$, define the random variable Λ_N as

$$\Lambda_N := \frac{1}{N} \sum_{j=0}^{N-1} f\left(\frac{j}{N} + \frac{U_j}{N}\right).$$

If $\mu_1 := \int_0^1 f(x)\mathrm{d}x$ and $\mu_2 := \int_0^1 f(x)^2\mathrm{d}x$, then :

- (a) (6 points) Calculate $\mathbb{E}\left[\Lambda_N\right]$ in terms of μ_1 and μ_2.
- (b) (9 points) Calculate $\left(\lim_{N\to\infty} \mathrm{Var}\left[\Lambda_N\right]\right)$ in terms of μ_1 and μ_2.

Q 3 (20 points) Answer each of the following questions.

- (a) (5 points) Let $X \sim N(0, \sigma^2)$ be a normal random variable on $\mathbb{R}$ with mean zero and variance σ^2. For $t > 0$, calculate $\mathbb{E}[\exp(-tX^2)]$ in terms of t and σ.
- (b) (15 points) Let $\mathbf{X} \sim N_d(\mathbf{0}, \Sigma)$ denote a d-dimensional normal random variable with mean $\mathbf{0} \in \mathbb{R}^d$ and covariance matrix Σ, and $\|\cdot\|_2$ denote the standard ℓ^2 norm on $\mathbb{R}^d$. For $t > 0$, calculate $\mathbb{E}\left[\exp(-t\|\mathbf{X}\|_2^2)\right]$ in terms of t and the eigenvalues of Σ.

Q 4 (20 points) Consider a binary classification problem with a training sample $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}, i = 1, \ldots, n\}$ and a predictor $\hat{h}_n$ obtained as the output of the 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 which depends on $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{H}$. Assume that the predictor

$\hat{h}_n$ is consistent, that is $R_{emp}(\hat{h}_n) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i \neq \hat{h}_n(x_i)\}} = 0$. In PAC learning 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 $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$.

- (a) (6 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) (6 points) We say that a set $C = \{c_1, c_2, \ldots, c_k\} \subset \mathbb{R}^d$ is shattered by $\mathcal{H}$ if for any $\{b_1, b_2, \ldots, b_k\} \in \{0, 1\}^d$, there exists a function $h \in \mathcal{H}$ such that $h(c_i) = b_i$ for all $i \in \{1, 2, \ldots, k\}$. The VC dimension of $\mathcal{H}$ is defined by

$$\text{VCdim}(\mathcal{H}) = \max_{\mathcal{C}}\{|\mathcal{C}|, \text{ s.t. } \mathcal{C} \text{ is shattered by } \mathcal{H}\},$$

  where the max is taken over all subsets $C \subset \mathbb{R}^d$, and $|C|$ refers to the cardinality (number of elements) of C. If for any $k \geq 1$, there exists a set C that is shattered by $\mathcal{H}$, we set $\text{VCdim}(\mathcal{H}) = \infty$.

  In the case of finite hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$, show that $VCdim(\mathcal{H}) \geq \log_2(|\mathcal{H}|)$, and give an example where the equality holds.

- (c) (8 points) Consider the hypothesis space $\mathcal{H} = \{f_a \,, a \in \mathbb{R}\}$, where $f_a(x) = \sin(ax)$ for all $x \in \mathbb{R}$. What is $\text{VCdim}(\mathcal{H})$?

Q 5 (30 points)

Consider a fully connected neural network given by

$$f(x) = \sum_{i=1}^{n} v_i \sigma(w_i^T x),$$

where $x, w_i \in \mathbb{R}^d$, $v_i \in \mathbb{R}$, and $\sigma(z) = \sin(z)$ is the sine activation function. We assume that the weights w_i's are iid multivariate Gaussian random variables with identity covariance matrix, i.e. $w_i \sim \mathcal{N}(0, I)$. Similarly, we assume that v_i's are iid Gaussian random variables with variance $1/n$, i.e. $v_i \sim \mathcal{N}(0, 1/n)$. Hereafter, the expectation $\mathbb{E}$ will always be taken with respect to random variables w_i's and v_i's.

- (a) (7 points) Let $x \in \mathbb{R}^d$. What is $\mathbb{E}[f(x)]$ and $\text{Var}[f(x)]$? (Express $\text{Var}[f(x)]$ in terms of $\|x\|$, Sine, and some expectation over a one-dimensional standard Gaussian variable Z.)

- (b) (10 points) Is $f(x)$ Gaussian? What is the distribution of $f(x)$ in the limit $n \to \infty$?

- (c) (13 points) We restrict our analysis to the case where $x \in \mathbb{S}^d := \{x \in \mathbb{R}^d, \text{s.t. } \|x\| = 1\}$. We define the Neural Kernel by $k(x, x') = \alpha\mathbb{E}[f(x)f(x')]$, where $\alpha = (\mathbb{E}[\sigma(Z)^2])^{-1}$ (where $Z \sim \mathcal{N}(0, 1)$).

Show that the kernel k can be expressed in the form

$$k(x, x') = a \exp(x^T x') + b \exp(-x^T x'), \forall x, x' \in \mathbb{S}^d,$$

for some constants $a, b \in \mathbb{R}$ (compute a, b explicitly).

Hint: use the Gaussian property $\mathbb{E}[ZG(Z)] = \mathbb{E}[G'(Z)]$ satisfied by any function G such that $\mathbb{E}[|G'(Z)|] < \infty$, where $Z \sim \mathcal{N}(0, 1)$.

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

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

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

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

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

- 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 $t<\lambda$.