NATIONAL UNIVERSITY OF SINGAPORE

Mathematics PhD Qualifying Exam Paper 4

Stochastic Processes and Machine Learning

(January 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 **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 **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 $G(n,p)$ denote the random graph on n vertices, where every pair of vertices is connected independently by an edge with probability p, independently across edges. Let N_4 denote the (random) number of copies K_4-s (i.e., complete graph on 4 vertices) that are inside $G(n,p)$.

- (a) (5 points) Calculate $\mathbb{E}[N_4]$.
- (b) (10 points) Calculate $\mathrm{Var}[N_4]$.

Q 2 (15 points) We consider opinion polling in a population of size N. Each individual in the population has one of two possible opinions - either YES or NO. We say that the population has reached *consensus* if all individuals have the same opinion – i.e., under consensus, everyone in the population is either YES or everyone is NO. Till consensus is reached, we update the opinions as follows : we pick an individual uniformly at random from the population, and update his/her opinion to YES with probability p, or to NO with probability $1-p$. Calculate the probability that, when consensus is reached, everyone has opinion YES.

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

- (a) (10 points) Let $\mathcal{X}$ be a random subset of $[N] = \{1,\dots,N\}$, formed by including each element $i \in [N]$ inside $\mathcal{X}$ with probability p, independently across $i \in [N]$. Let $\mathbb{X}$ and $\mathbb{Y}$ be two independent copies of the random set $\mathcal{X}$. Calculate the probability $\mathbb{P}[\mathbb{X} \cap \mathbb{Y} = \phi]$.
- (b) (10 points) Let U and V be two independent random variables, each distributed uniformly on the interval $[0,1]$. Show that the random variables $|U-V|$ and $\min\{U,V\}$ have the same distribution.

Q 4 (20 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 $R_{emp}(\hat{h}_n) = \frac{1}{n}\sum_{i=1}^n 1_{\{y_i\neq\hat{h}_n(x_i)\}} = 0$. 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 $\mathcal{H} = \{h_1, h_2, \dots, h_m\}$.

- (a) (5 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.

Q 5 (30 points)

Consider a 1 layer 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 $\sigma(z) = \max(z,0)$ is the ReLU 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 $W = (w_i)_{1\leq i\leq n}$'s and $V = (v_i)_{1\leq i\leq n}$'s.

- (a) (7 points) Let $x \in \mathbb{R}^d$. What is $\mathbb{E}_{W,V}[f(x)]$ and $\mathrm{Var}_{W,V}[f(x)]$? (give $\mathrm{Var}_{W,V}[f(x)]$ as a function of $\|x\|$)
- (b) (6 points)Is $f(x)$ Gaussian? what is the distribution of $f(x)$ in the limit $n \to \infty$?
- (c) (11 points) Show that

$$\mathbb{E}_{Z,Z'}[\sigma(Z)\sigma(cZ + \sqrt{1-c^2}Z')] = \frac{1}{2\pi}c \times \arcsin(c) + \frac{1}{2\pi}\sqrt{1-c^2} + \frac{1}{4}c$$

  for all $c \in [-1,1]$, where Z, Z' are iid $\mathcal{N}(0,1)$.

  (Hint: You can also assume without proof that the derivative of the function $h(c) = \mathbb{E}_{Z,Z'}[1_{Z>0}1_{cZ+\sqrt{1-c^2}Z'>0}]$ is given by $h'(c) = \frac{1}{2\pi\sqrt{1-c^2}}$.)

- (d) (6 points) Let $x, x' \in \mathbb{R}^d$. We define the *Neural Network Kernel* by $k(x,x') = \mathbb{E}_{W,V}[f(x)f(x')]$.

  Show that

$$k(x,x') = \frac{x^T x'}{2\pi}\left(\arcsin\left(\frac{x^T x'}{\|x\|\|x'\|}\right) + \frac{\pi}{2}\right) + \frac{1}{2\pi}\sqrt{\|x\|^2\|x'\|^2 - (x^T x')^2}.$$

— **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 $t<\lambda.$