NATIONAL UNIVERSITY OF SINGAPORE

Mathematics PhD Qualifying Exam Paper 4

Stochastic Processes and Machine Learning

(January 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 **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 **SEVEN (7)** 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)** Balls are falling into two bins A and B according to the following randomized scheme. We begin with both bins empty, and the $(n+1)$-th ball is allocated to either bin with probability proportional to the existing number of balls in the other bin (after n steps). Let $\Lambda_n(t)$ be the moment generating function of the number of balls in bin A after n steps. Show that the following recursion holds (where f' denotes the first derivative of the function f) :

$$e^{-t}\Lambda_{n+1}(t) = \Lambda_n(t) - \frac{1-e^{-t}}{n} \cdot \Lambda_n'(t).$$

2. **(20 points)** Consider the following dynamics on the space of simple graphs on N vertices V (i.e., undirected graphs without any loops). Let the graph after $t \in \mathbb{N} \cup \{0\}$ steps of the dynamics be denoted by G_t. Then G_{t+1} is obtained as follows:
(1) we choose any pair of vertices from V, uniformly at random over all pairs (and independently of everything else), and then
(2) we connect that pair of vertices by an edge with probability $1/2$ (independently of everything else, and no matter if they were connected by an edge or not inside G_t).
   (a) (10 points) Show that the random graphs $\{G_t\}_{t\in\mathbb{N}}$ in the above dynamics converge to a limiting random graph G_∞ (no matter what the initial graph G_0 is), and give a complete description of the distribution of G_∞.
   (b) (10 points) If we start from $G_0 = K_N$, the complete graph on N vertices, then compute the expected minimum number of steps in the above dynamics to obtain the same graph K_N once again.

3. **(10 points)** Let $\{\xi_k\}_{k\geq 0}$ be i.i.d. random variables that are uniformly distributed in the interval $[-\sigma, \sigma]$, where $\sigma > 0$. Let θ be uniformly distributed on the interval $[0, 1]$ and independent of $\{\xi_k\}_{k\geq 1}$. For $n \geq 1$, consider the random variable

$$\mathbb{X}_n = \sum_{k=0}^{n} \xi_k \cos(2\pi k\theta).$$

Show that, $\forall n \geq 1, t \geq 0$, we have

$$\mathbb{P}\left[|\mathbb{X}_n - \mathbb{E}[\mathbb{X}_n]| \geq t\right] \leq 2\, \exp\left(-\frac{t^2}{2(n+1)\sigma^2}\right).$$

4. **(10 points)** Let X, Y be two i.i.d. random variables sampled uniformly from the (continuous) interval $[1, N]$, for a positive integer N. For any $t \in \mathbb{R}$, let $\lfloor t \rfloor$ denote the biggest integer $\leq t$, and let $\lceil t \rceil$ denote the smallest integer $\geq t$. Calculate

$$\mathbb{P}\left(\lfloor X \rfloor = \lceil Y \rceil\right).$$

5. **[Dropout Regulization] (15 points)** Having taken DSA5105 at NUS, Ethan is excited about regularization methods. He shares with his roommate, an AI engineer at MakeAICool, how useful those methods are. Much to Ethan's surprise, his roommate tells him that dropout regularization is used in training deep neural networks in addition to L2 and L1 regularizers covered in DSA5105. Believing that what he learns in DSA5105 can explain this dropout, Ethan is keen on exploring the effect of dropout regularization on a simple linear regression model trained using least squares. Given input vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top \in \mathbb{R}^D$ and output vector $\mathbf{y} = (y_1, y_2, \ldots, y_K) \in \mathbb{R}^K$, he considers a model of the form

$$y_k = \sum_{i=1}^{D} w_{ki} x_i$$

along with a sum-of-squares error function given by

$$L(\mathbf{W}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \left\{ y_{nk} - \sum_{i=1}^{D} w_{ki} R_{ni} x_{ni} \right\}^2$$

where w_{ki} are scalar learnable weights and the weight matrix $\mathbf{W}$ is given by $\mathbf{W}(k,i) = w_{ki}$. The elements $R_{ni} \in \{0,1\}$ of the dropout matrix are chosen randomly from a Bernoulli distribution with parameter ρ. Ethan now takes an expectation over the distribution of random dropout parameters. Since Ethan is busy attending the IMS Workshop on Mathematics of Data at NUS, he could not continue to work on the derivation this week. Thus, he asks Ph.D. students in our department to help him do the following.

(a) (5 points) Show that

$$\mathbb{E}[R_{ni}] = \rho$$
$$\mathbb{E}[R_{ni} R_{nj}] = \delta_{ij}\rho + (1 - \delta_{ij})\rho^2$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

(b) (5 points) Hence, show that the expected error function for this dropout model is given by

$$\mathbb{E}[L(\mathbf{W})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \left\{ y_{nk} - \rho \sum_{i=1}^{D} w_{ki} x_{ni} \right\}^2 + \rho(1-\rho) \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{i=1}^{D} w_{ki}^2 x_{ni}^2. \tag{1}$$

Thus, we see that the expected error function corresponds to a sum-of-squares error with a quadratic regularizer in which the regularization coefficient is scaled separately for each input variable according to the data values seen by that input.

(c) (5 points) Write down a closed-form solution for the weight matrix that minimizes the expected regularized error function in Eqn. (1).

6. **[Graph Laplacian] (20 points)** In DSA5105, Aanya studies graph-based methods. Given a graph $G = (V, E, \mathbf{W})$, V is the set of nodes, E is the set of edges, and the matrix $\mathbf{W}$ is the edge weights with $\mathbf{W}(i,j) = w_{ij}$. Recall that the degree matrix $\mathbf{D}$ of the graph G is a diagonal matrix with diagonal entries $\mathbf{D}(i,i) = d_i = \sum_{j=1}^{n} w_{ij}$, where n is the number of nodes in the graph. Aanya notices that there are two normalized versions of the graph Laplacian, a symmetric one and a non-symmetric one, given by

$$\mathbf{L}_S = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}},$$
$$\mathbf{L}_N = \mathbf{D}^{-1} \mathbf{L} = \mathbf{I} - \mathbf{D}^{-1} \mathbf{W},$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Aanya tries to understand their relationship and discovers some interesting results. However, since she is busy with her startup, she did not have time to prove them. Knowing that Ph.D. students in our department are very good, Aanya reaches out for help. Please help Aanya and prove the following results:

(a) (5 points) For every vector $\mathbf{f} \in \mathbb{R}^n$ there holds

$$\mathbf{f}^\top \mathbf{L}_S \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2.$$

(b) (5 points) λ is an eigenvalue of $\mathbf{L}_N$ with eigenvector $\mathbf{u}$ if and only if λ is an eigenvalue of $\mathbf{L}_S$ with eigenvector $\mathbf{w} = \mathbf{D}^{\frac{1}{2}} \mathbf{u}$.

(c) (5 points) λ is an eigenvalue of L_N with eigenvector $\mathbf{u}$ if and only if λ and $\mathbf{u}$ solve the generalized eigenvalue problem $\mathbf{L}\mathbf{u} = \lambda \mathbf{D}\mathbf{u}$.

(d) (5 points) 0 is an eigenvalue of $\mathbf{L}_N$ and the associated eigenvector is $\mathbf{1}$, where the vector $\mathbf{1} = (1, 1, \ldots, 1)^\top$. 0 is an eigenvalue of $\mathbf{L}_S$ and the associated eigenvector is $\mathbf{D}^{\frac{1}{2}} \mathbf{1}$.

7. **[Sparse SVM] (15 points)** Waking up in the morning after the New Year celebration, Professor Nguyen realizes that there are two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. He then wonders suppose that instead of maximizing the margin, he chooses to maximize sparsity by minimizing the L_p norm of the vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ for some $p \geq 1$, where m is the number of data points in the training set and $\mu_1, \mu_2, \ldots, \mu_m$ are Lagrange multipliers, a.k.a., dual variables in the SVM problem. Professor Nguyen first considers the case $p = 2$. This gives the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\mu}, b, \boldsymbol{\xi}} \frac{1}{2} \sum_{i=1}^{m} \mu_i^2 + C \sum_{i=1}^{m} \xi_i \\ &\text{subject to } y_i \left(\sum_{j=1}^{m} \mu_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b \right) \geq 1 - \xi_i, i = 1, 2, \ldots, m \\ &\xi_i, \mu_i \geq 0, i = 1, 2, \ldots, m, \end{aligned} \tag{2}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)^\top$. Believing that Ph.D. students in our department could help him further develop the Sparse SVM given in (2), Professor Nguyen decides to put it in the QE and asks the students to do the following.

(a) (5 points) Show that the problem in (2) coincides with an instance of the primal optimization problem of SVM with the additional non-negativity constraint on $\boldsymbol{\mu}$.

(b) (5 points) Derive the dual optimization of the Sparse SVM in (2).

(c) (5 points) Setting $p = 1$ will induce a more sparse $\boldsymbol{\mu}$. The Sparse SVM in (2) is now become

$$\begin{aligned} &\min_{\boldsymbol{\mu}, b, \boldsymbol{\xi}} \sum_{i=1}^{m} |\mu_i| + C \sum_{i=1}^{m} \xi_i \\ &\text{subject to } y_i \left(\sum_{j=1}^{m} \mu_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b \right) \geq 1 - \xi_i, i = 1, 2, \ldots, m \\ &\xi_i, \mu_i \geq 0, i = 1, 2, \ldots, m. \end{aligned} \tag{3}$$

Derive the dual optimization in this case.

End of Paper

- Bernoulli (p) :
  $\mathbb{P}(X = i) = \begin{cases} p \text{ if } i = 1 \\ 1 - p \text{ if } i = 0. \end{cases}$
  $\mathbb{E}[X] = p, \quad \text{Var}[X] = p(1-p), \quad \mathbb{E}[e^{tX}] = (1-p) + pe^t.$
- Binomial (n,p):
  $\mathbb{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}; 0 \leq i \leq n.$
  $\mathbb{E}[X] = np, \quad \text{Var}[X] = np(1-p), \quad \mathbb{E}[e^{tX}] = [(1-p) + pe^t]^n.$
- Geometric (p) :
  $\mathbb{P}(X = i) = (1-p)^{i-1} p; i \geq 1.$
  $\mathbb{E}[X] = \frac{1}{p}, \quad \text{Var}[X] = \frac{1-p}{p^2}, \quad \mathbb{E}[e^{tX}] = \frac{pe^t}{1-(1-p)e^t}$ for $t < -\log(1-p)$.
- Poisson (λ):
  $\mathbb{P}(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}; i \geq 1.$
  $\mathbb{E}[X] = \lambda, \quad \text{Var}[X] = \lambda, \quad \mathbb{E}[e^{tX}] = \exp(\lambda(e^t - 1)).$
- Uniform (a,b) :
  $f(x) = \begin{cases} \frac{1}{b-a} \text{ if } a \leq x \leq b \\ 0 \text{ otherwise .} \end{cases}$
  $\mathbb{E}[X] = (a+b)/2, \quad \text{Var}[X] = \frac{(b-a)^2}{12}, \quad \mathbb{E}[e^{tX}] = \frac{e^{tb} - e^{ta}}{t(b-a)}$ 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 ($N(\mu, \sigma^2)$):
  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.
  $\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2, \quad \mathbb{E}[e^{tX}] = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.
- Exponential (λ):
  $f(x) = \begin{cases} \lambda \exp(-\lambda x) \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$
  $\mathbb{E}[X] = 1/\lambda, \quad \text{Var}[X] = 1/\lambda^2, \quad \mathbb{E}[e^{tX}] = \frac{\lambda}{\lambda - t}$ for $t < \lambda$.