CS 771A: Intro to Machine Learning, IIT Kanpur			**Quiz II** (20 Mar 2024)
Name	MELBO		**20 marks**
Roll No	240007	Dept. AWSM	Page **1** of **2**

Instructions:
1. This question paper contains 1 page (2 sides of paper). Please verify.
2. Write your name, roll number, department above in **block letters neatly with ink**.
3. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
4. Don't overwrite/scratch answers especially in MCQ – such cases will get straight 0 marks.
5. Do not rush to fill in answers. You have enough time to solve this quiz.

(Noise to Regularize) The underlying principle behind the deep learning technique *dropout* is that adding noise to data can prevent models from overfitting. Let us derive this fact formally.

Q1. Let $\boldsymbol{\epsilon} \in \{-1,+1\}^D$ be a D-dim Rademacher vector with coordinates chosen i.i.d. $\epsilon_j = 1$ or -1 uniformly randomly. Find the following (no derivation) Note: $j, k \in [D], j \neq k$. **(6 x 1 = 6 marks)**

$\mathbb{E}[\epsilon_j + \epsilon_k] = 0$	$\mathrm{Var}[\epsilon_j + \epsilon_k] = 2$
$\mathbb{E}[\epsilon_j \epsilon_k] = 0$	$\mathrm{Var}[\epsilon_j \epsilon_k] = 1$
$\mathbb{E}[\epsilon_j / \epsilon_k] = 0$	$\mathrm{Var}[\epsilon_j / \epsilon_k] = 1$

Q2. Let $y, \lambda \in \mathbb{R}$ and $\mathbf{w}, \mathbf{x} \in \mathbb{R}^D$ be constants and $\boldsymbol{\epsilon} \in \{-1,+1\}^D$ be a Rademacher vector sampled independently of $y, \lambda, \mathbf{w}, \mathbf{x}$. Obtain a simplified expression (expectation is over the choice of $\boldsymbol{\epsilon}$ only). Give brief derivation. Your expression should not contain any ϵ_j terms. **(2 + 4 = 6 marks)**

Write final expression in the box

$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left(y - \mathbf{w}^\top(\mathbf{x} + \lambda \cdot \boldsymbol{\epsilon})\right)^2\right] =$ $(y - \mathbf{w}^\top \mathbf{x})^2 + \lambda^2 \cdot \|\mathbf{w}\|_2^2$ or else $(y - \mathbf{w}^\top \mathbf{x})^2 + \lambda^2 \cdot \mathbf{w}^\top \mathbf{w}$

Brief derivation for simplification

Expanding the expression gives us $\mathbb{E}_{\boldsymbol{\epsilon}}\left[y^2 + (\mathbf{w}^\top \mathbf{x} + \lambda \cdot \mathbf{w}^\top \boldsymbol{\epsilon})^2 - 2y\left(\mathbf{w}^\top(\mathbf{x} + \lambda \cdot \boldsymbol{\epsilon})\right)\right]$

Using linearity of expectation yields $y^2 + (\mathbf{w}^\top \mathbf{x})^2 - 2y \cdot \mathbf{w}^\top \mathbf{x} + \mathbb{E}_{\boldsymbol{\epsilon}}[(\lambda \cdot \mathbf{w}^\top \boldsymbol{\epsilon})^2 - 2y\lambda \cdot \mathbf{w}^\top \boldsymbol{\epsilon}]$

Using the fact that $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}] = \mathbf{0}$ gives us $y^2 + (\mathbf{w}^\top \mathbf{x})^2 - 2y \cdot \mathbf{w}^\top \mathbf{x} + \lambda^2 \cdot \mathbb{E}_{\boldsymbol{\epsilon}}[(\mathbf{w}^\top \boldsymbol{\epsilon})^2]$

Expanding the last term gives us $\mathbb{E}_{\boldsymbol{\epsilon}}[(\mathbf{w}^\top \boldsymbol{\epsilon})^2] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sum_{d \in [D]} w_d^2 \epsilon_d^2 + \sum_{\substack{d \neq d' \\ d,d' \in [D]}} w_d w_{d'} \epsilon_d \epsilon_{d'}\right]$

Using results from Q1 simplifies this to $\mathbb{E}_{\boldsymbol{\epsilon}}[(\mathbf{w}^\top \boldsymbol{\epsilon})^2] = \sum_{d \in [D]} w_d^2 = \mathbf{w}^\top \mathbf{w} = \|\mathbf{w}\|_2^2$

Completing the squares then yields the final expression.

Q3. We have N datapoints $(\mathbf{x}^n, y^n) \in \mathbb{R}^D \times \mathbb{R}, n \in [N], \lambda \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^D$ all of which can be treated as constants. We also sample N Rademacher vectors $\boldsymbol{\epsilon}^n \in \{-1, +1\}^D, n \in [N]$ i.i.d. of each other as well as independent of the datapoints and $\lambda, \mathbf{w}$. Expectation is over the choice of $\{\boldsymbol{\epsilon}^n, n \in [N]\}$ only. Write down a simplified expression for the following (no derivation needed). **(2 marks)**

$$\mathbb{E}_{\{\boldsymbol{\epsilon}^n\}}\left[\sum_{n\in[N]} \left(y^n - \mathbf{w}^\top(\mathbf{x}^n + \lambda \cdot \boldsymbol{\epsilon}^n)\right)^2\right] = \left(\sum_{n\in[N]} (y^n - \mathbf{w}^\top \mathbf{x}^n)^2\right) + N\lambda^2 \cdot \|\mathbf{w}\|_2^2$$

or else $\sum_{n\in[N]}((y^n - \mathbf{w}^\top\mathbf{x}^n)^2 + \lambda^2 \cdot \|\mathbf{w}\|_2^2)$.

Note: this is exactly the objective function that ridge regression tries to minimize 😊

Q4 (IPL Intrigue). Melbo is a big IPL fan and is trying to analyse the performance of MI vs CSK on various kinds of pitches. Let M be the event that MI won a MI-vs-CSK match and C be the event that CSK won a MI-vs-CSK match. There are 3 kinds of pitches F = flat, G = green, D = dusty. A total of 24 matches were played between MI and CSK, 1/4$^{\text{th}}$ of which were on green pitches and 1/3$^{\text{rd}}$ on flat pitches. MI won 6 of the matches played on flat pitches. Both MI and CSK won equal number of matches played on green pitches i.e., $\mathbb{P}[M \mid G] = \mathbb{P}[C \mid G]$. Also, both flat and dusty pitches have been equally favourable for MI in that $\mathbb{P}[F \mid M] = \mathbb{P}[D \mid M]$. Find out the following quantities as fractions or decimals (no derivations needed). *Hint: either use Bayes rule or fill-up a 2 x 3 matrix showing which team won how many matches on what kind of pitch*. **(6 x 1 = 6 marks)**

$\mathbb{P}[F \mid M] = \frac{2}{5}$ $\qquad$ $\mathbb{P}[F \mid C] = \frac{2}{9}$

$\mathbb{P}[G \mid M] = \frac{1}{5}$ $\qquad$ $\mathbb{P}[G \mid C] = \frac{1}{3}$

$\mathbb{P}[D \mid M] = \frac{2}{5}$ $\qquad$ $\mathbb{P}[D \mid C] = \frac{4}{9}$

$\mathbb{P}[F] = \frac{1}{3}, \mathbb{P}[G] = \frac{1}{4}, \mathbb{P}[D] = \frac{5}{12}$ which gives us $\mathbb{P}[M \mid F] = \frac{\mathbb{P}[M \cap F]}{\mathbb{P}[F]} = \frac{\frac{6}{24}}{\frac{1}{3}} = \frac{3}{4}$ and $\mathbb{P}[C \mid F] = \frac{1}{4}$.

Since $\mathbb{P}[M \mid G] + \mathbb{P}[C \mid G] = 1$, we get $\mathbb{P}[M \mid G] = \mathbb{P}[C \mid G] = \frac{1}{2}$. Since $\mathbb{P}[F \mid M] = \mathbb{P}[D \mid M]$, we get $\mathbb{P}[M \mid D] = \frac{\mathbb{P}[D \mid M] \cdot \mathbb{P}[M]}{\mathbb{P}[D]} = \frac{\mathbb{P}[F \mid M] \cdot \mathbb{P}[M]}{\mathbb{P}[D]} = \mathbb{P}[M \mid F] \cdot \frac{\mathbb{P}[F]}{\mathbb{P}[D]} = \frac{3}{5}$ and $\mathbb{P}[C \mid D] = \frac{2}{5}$.

The law of total probability then allows us to calculate the following:

$\mathbb{P}[M] = \mathbb{P}[M \mid F] \cdot \mathbb{P}[F] + \mathbb{P}[M \mid G] \cdot \mathbb{P}[G] + \mathbb{P}[M \mid D] \cdot \mathbb{P}[D] = \frac{5}{8}$ and $\mathbb{P}[C] = 1 - \mathbb{P}[M] = \frac{3}{8}$.

Applying the Bayes rule then tells us that $\mathbb{P}[F \mid M] = \frac{\mathbb{P}[M \mid F] \cdot \mathbb{P}[F]}{\cdot \mathbb{P}[M]} = \frac{2}{5} = \mathbb{P}[D \mid M]$ as promised and $\mathbb{P}[G \mid M] = 1 - \mathbb{P}[F \mid M] - \mathbb{P}[D \mid M] = \frac{1}{5}$.

Applying the Bayes rule again tells us $\mathbb{P}[F \mid C] = \frac{\mathbb{P}[C \mid F] \cdot \mathbb{P}[F]}{\cdot \mathbb{P}[C]} = \frac{2}{9}$, $\mathbb{P}[G \mid C] = \frac{\mathbb{P}[C \mid G] \cdot \mathbb{P}[G]}{\cdot \mathbb{P}[C]} = \frac{1}{3}$ and $\mathbb{P}[D \mid C] = 1 - \mathbb{P}[F \mid C] - \mathbb{P}[G \mid C] = \frac{4}{9}$.