The following information should be so legible that a machine can read it

Roll No : 1234567890 **Dept.**:
e.g., 190040001 e.g., CSE

IIT Bombay
CS 217/337: AIML
Midsem Exam, 2023-24-II
Date: March 2, 2024

CS 217/337: Artificial Intelligence and Machine Learning

Total: **10 × 4 = 40 points,** *Duration:* **2 hours, ATTEMPT ALL QUESTIONS**

Instructions:

1. This question-and-answersheet booklet contains a total of **8 sheets** of paper **(16 pages, pages 2 and 16 are blank)**. Please verify.
2. Write your roll number and department on **every side of every sheet** (except the blank sheet) of this booklet. Use only **black/blue ball-point pen**. The first 5 minutes of additional time is given exclusively for this activity.
3. Write final answers neatly with a pen **only in the given boxes**.
4. Use the rough sheets for scratch works / attempts to solution. **Write only the final solution (which may be a sequence of logical arguments) in a precise and succinct manner in the boxes provided**. Do not provide unnecessarily elaborate steps. The space within the boxes is sufficient for the correct and precise answers.
5. Submit your answerscripts to the teaching staff when you leave the exam hall or the time runs out (whichever is earlier). **Your exam will not be graded if you fail to return the paper**.
6. **This is a closed book, notes, internet exam. No communication device, e.g., cellphones, iPad, etc., is allowed**. Keep it switched off in your bag and keep the bag away from you. If anyone is found in possession of such devices during the exam, that answerscript may be disqualified for evaluation and DADAC may be invoked.
7. One A4 assistance sheet (text **on both sides**) and a scientific calculator are allowed for the exam.

Problem 1 (10 points). Suppose we want to estimate a discrete random variable X with a distribution of $p(x) := P(X = x)$. However, we only observe a corrupted version of X that is represented by another random variable Y, which is related to X via the conditional distribution $p(y|x) := P(Y = y|X = x)$. From Y, we calculate a function $g(Y) = \hat{X}$, which is an estimate of X. In this question, we want to have a tight bound on the probability of making an error, i.e., $\hat{X} \neq X$. The dependency of these three random variables are as follows: $X \to Y \to \hat{X}$. Define the **probability of error** as $P_e := P(\hat{X} \neq X)$.

Recall the definitions of entropy and conditional entropy of a random variable.

$$H(X) := -\sum_{x\in\mathcal{X}} p(x)\ln p(x);\;\; H(X|Y) := -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\ln p(x|y) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\ln p(x|y).$$

Note that, here we have defined the entropies using the natural logarithm of the distributions. The sets $\mathcal{X}$ and $\mathcal{Y}$ are the the state-spaces of X and Y respectively and are both finite.

(a) Define a binary random variable called error E that captures how $\hat{X}$ makes an error estimating X.

1 point.

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \text{ , i.e. error is made.} \\ 0 & \text{ow} \end{cases}$$

(b) Prove that conditioning does not increase entropy, i.e., $H(X) \geqslant H(X|Y)$. You may use Jensen's inequality: for a concave function f, $\mathbb{E}f(X) \leqslant f(\mathbb{E}X)$.

2 points.

Consider the mutual information

$$I(X;Y) = H(X) - H(X|Y) = -\sum_x p(x)\ln p(x) + \sum_y\sum_x p(x,y)\ln p(x|y)$$

$$= -\underbrace{\sum_y\sum_x p(x,y)}_{\text{comes back to the same marginal } p(x)}\ln p(x) + \sum_y\sum_x p(x,y)\ln p(x|y)$$

$$= -\sum_y\sum_x p(x,y)\ln\frac{p(x)p(y)}{p(x,y)} \quad \text{---} \textcircled{1}$$

Since ln is a concave function, Jensen's ineq can be applied.

Note that $\sum_y \sum_x p(x,y) \ln \frac{p(x)\,p(y)}{p(x,y)} = \mathbb{E}_{x,y} \ln \frac{p(x)\,p(y)}{p(x,y)}$

$$\underset{\text{Jensen's}}{\leqslant} \ln\left(\mathbb{E}_{x,y} \frac{p(x)\,p(y)}{p(x,y)}\right) = \ln\left(\sum_y \sum_x p(x,y) \cdot \frac{p(x)\,p(y)}{p(x,y)}\right)$$

$$= \ln\left(\sum_y \sum_x p(x)\,p(y)\right) = \ln\left[\left(\sum_x p(x)\right)\left(\sum_y p(y)\right)\right]$$

$$= \ln 1 = 0$$

Hence, we get from ① :

$H(X) - H(X|Y) \geqslant 0$ as desired.

(c) Prove the upper bound of entropy $H(X) \leqslant \ln|\mathcal{X}|$. You may use any of the previously proved or suggested inequalities. Note that this inequality also holds for the conditional entropy. However, you may be able to tighten the inequality, think about when and how (no point for this second part).

2 points.

$$H(X) = -\sum_{x\in\mathcal{X}} p(x) \ln p(x) = \sum_{x\in\mathcal{X}} p(x) \ln \frac{1}{p(x)}$$

$$= \mathbb{E}_X \ln\left(\frac{1}{p(X)}\right) \underset{\text{Jensen's}}{\leqslant} \ln\left[\mathbb{E}_X\left(\frac{1}{p(X)}\right)\right]$$

$$= \ln\left(\sum_{x\in\mathcal{X}} p(x) \cdot \frac{1}{p(x)}\right) = \ln|\mathcal{X}|$$

□

(d) Now prove the chain rule of entropy: $H(X,Y|Z) = H(X|Z) + H(Y|Z,X)$. Note that the earlier definition of conditional entropy can be extended to any number of variables, e.g., $H(X,Y|Z) = -\sum_{z\in\mathcal{Z}}\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y,z)\ln p(x,y|z)$. **2 points.**

From The chain rule of probability,

$$p(x,y|z) = p(x|z)\,p(y|z,x) \quad \text{- - - -} \; ①$$

$$H(X,Y|Z) = -\sum_z\sum_y\sum_x p(x,y,z) \ln p(x|z)\,p(y|z,x)$$

$$= \underbrace{-\sum_z\sum_x p(x,z)\ln p(x|z)}_{\text{marginalized over } y} - \sum_z\sum_y\sum_x p(x,y,z)\overset{\ln}{\wedge} p(y|z,x)$$

$$= H(X|Z) + H(Y|Z,X). \qquad \square$$

(e) Using the results obtained so far, expand $H(E,X|Y)$ in two different ways and obtain the missing terms of the following inequality. Note that the missing terms should not have any random variables. Show each step with justification. Write the final values of (i) and (ii) separately in the boxes provided at the end (the space below is only for the steps of the derivation). **2 + (0.5 × 2) points.**

$$\boxed{\text{(i)}} + P_e \cdot \boxed{\text{(ii)}} \geqslant H(X|Y).$$

Let us write $H(E,X|Y)$ in two different ways as follows.

$$H(E,X|Y) = H(E|Y) + H(X|Y,E) \quad \text{- - - -} \; ①$$

$$= H(X|Y) + H(E|Y,X) \quad \text{- - - -} \; ②$$

The following information should be so legible that a machine can read it

Roll No : e.g., 190040001

Dept.: e.g., CSE

IIT Bombay
CS 217/337: AIML
Midsem Exam, 2023-24-II
Date: March 2, 2024

From ①: $H(E|Y) \leq H(E)$ from part (b)

$$= -\left(P_e \ln P_e + (1-P_e)\ln(1-P_e)\right) \cdots\cdots ③$$

$$H(X|Y,E) = P(E=0)\underbrace{H(X|Y,E=0)}_{=0} + P(E=1)H(X|Y,E=1)$$

$=0$, if $E=0$, then $\hat{X}=X$ and there is no uncertainty of X.

also $H(X|Y,E=1) \leq \ln(|\mathcal{X}|-1)$, since $E=1$, then at least one option of X is ruled out which is $\hat{X}$, hence the uncertainty is over the rest $|\mathcal{X}|-1$

Hence, $H(X|Y,E) \leq P_e \ln(|\mathcal{X}|-1)$

From ②: Note that $H(E|Y,X)=0$

if X and Y are given, then there is no uncertainty about the error

Hence, collecting all terms together,

$$-\left(P_e \ln P_e + (1-P_e)\ln(1-P_e)\right) + P_e \ln(|\mathcal{X}|-1) \geq H(X|Y)$$

(i) = $-\left(P_e \ln P_e + (1-P_e)\ln(1-P_e)\right)$

(ii) = $\ln(|\mathcal{X}|-1)$

The following information should be so legible that a machine can read it

IIT Bombay
CS 217/337: AIML
Midsem Exam, 2023-24-II
Date: March 2, 2024

Roll No :
e.g., 190040001

Dept.:
e.g., CSE

Problem 2 (10 points). Consider the probability that a consumer buys a product is given by

$$p := \frac{\exp(\mathbf{a}^\top \mathbf{x} + b)}{1 + \exp(\mathbf{a}^\top \mathbf{x} + b)},$$

where $\mathbf{x} \in \mathbb{R}^n_{\geqslant 0}$ denotes the values of the features of the product that affect the probability of purchase, e.g., components of the product, advertising effort, wholesale price, retail price, packaging expenses, etc. The optimization variable $\mathbf{x}$ is constrained by the linear inequality $\mathbf{Fx} \leqslant \mathbf{g}$, where $\mathbf{F} \in \mathbb{R}^{m \times n}, \mathbf{g} \in \mathbb{R}^m$.

The manufacturer of the product is considering optimizing two different objectives given as follows.

(a) *Objective 1: Maximizing the buying probability.* The aim here is to maximize p by choosing the right $\mathbf{x}$. Write down the optimization problem that the manufacturer needs to solve to find the optimal $\mathbf{x}$.

2 points.

$$\max \quad \frac{\exp(a^T x + b)}{1 + \exp(a^T x + b)}$$

$$\text{s.t.} \quad F x \leqslant g$$
$$x \geqslant 0$$

(b) Is the above optimization problem a linear program? If yes, explain how it is obtained. If not, reduce it to an LP with appropriate justification. [Note: if the earlier optimization problem itself is incorrect, then this part may not be checked]

3 points.

No. However, we observe that $\frac{e^u}{1+e^u}$ is a monotone increasing function.

The following information should be so legible that a machine can read it

Roll No : e.g., 190040001 **Dept.**: e.g., CSE

IIT Bombay
CS 217/337: AIML
Midsem Exam, 2023-24-II
Date: March 2, 2024

Hence, the previous optimization problem can be simplified to:

$$\max \quad a^T x + b$$
$$\text{s.t.} \quad Fx \leq g$$
$$x \geq 0$$

which is an LP.

(c) *Objective 2: Maximizing the expected profit.* Let $(\mathbf{c}^\top \mathbf{x} + d)$ be the profit derived when the product is sold, and is assumed to be positive for all feasible values of $\mathbf{x}$. Write down the optimization problem that the manufacturer needs to solve to find the optimal $\mathbf{x}$ for maximizing expected profit.

2 points.

Here the maximization objective is $p \cdot (c^T x + d)$ which is equivalent to maximize its log.

$$\max \quad a^T x + b - \log\left(1 + \exp(a^T x + b)\right) + \log(c^T x + d)$$
$$\text{s.t.} \quad Fx \leq g \quad , \quad x \geq 0$$

This is the desired optimization problem

[keeping the problem in a non-log form is also okay and will be given same credit if it is correct. This forme helps in proving the next part of this question]

(d) Is the above optimization problem a convex optimization program? If yes, explain mathematically using only the properties of convex functions. If not, provide a counterexample. [Note: if the earlier optimization problem itself is incorrect, then this part may not be checked] **3 points.**

Yes. To prove it we need to show that the objective function is concave in x (since this is a maximization problem), as the constraints are linear.

We show the concavity of the objective function by showing that its second derivative (called hessian) matrix is negative semidefinite.

let $f(x) = a^Tx+b - \log(1+\exp(a^Tx+b)) + \log(c^Tx+d)$

$$\nabla_x f = a - \frac{\exp(a^Tx+b)}{1+\exp(a^Tx+b)}\cdot a + \frac{1}{c^Tx+d}\cdot c$$

$$\nabla_x^2 f = -\frac{(1+\exp(a^Tx+b))\exp(a^Tx+b) - \exp^2(a^Tx+b)}{(1+\exp(a^Tx+b))^2}\cdot aa^T - \frac{1}{(c^Tx+d)^2}\cdot cc^T$$

This is negative semidefinite, because $\forall\, x \neq 0 \in \mathbb{R}^n$

$$x^T(\nabla_x^2 f)x = -\frac{\exp(a^Tx+b)}{(1+\exp(a^Tx+b))^2}(a^Tx)^2 - \frac{1}{(c^Tx+d)^2}(c^Tx)^2 \leq 0$$

The following information should be so legible that a machine can read it

Roll No : □ e.g., 190040001 **Dept.**: □ e.g., CSE

IIT Bombay
CS 217/337: AIML
Midsem Exam, 2023-24-II
Date: March 2, 2024

Problem 3 (10 points). In this question, we will learn the Gaussian approximation of any random variable using truncated Taylor series. Recall that the Taylor series for vector functions is given by

$$f(\mathbf{x}) := f(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^\top \nabla f(\mathbf{x}_0) + \frac{1}{2!}(\mathbf{x}-\mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \dots,$$

where $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. The above Taylor series is expanded around the stationary point $\mathbf{x}_0$. Using a second order approximation (i.e., terms above the second degree of $\mathbf{x}$ are ignored), obtain the reduced expression of $f(\mathbf{x})$ and answer the following sub-questions.

(a) Suppose the *probability density function (PDF)* of a random variable X is given by $q(\mathbf{x}) = \frac{1}{F} f(\mathbf{x})$, where $F = \int f(\mathbf{x}) d\mathbf{x}$ is the normalization factor. Using the *mode* of X, say $\mathbf{x}_m$, and the second order approximation of the Taylor series, find the Gaussian approximation of $f(\mathbf{x})$. A Gaussian approximation of a function is where the functional form is same as a multivariate Gaussian distribution (see below) with possibly different parameters and multipliers. Show each step of your derivation with explanation. [Hint: expand the `log` of $f(\mathbf{x})$] **3 points.**

Recall that

- *mode* is the point where the probability density of the random variable is maximum, and
- the PDF of a multivariate Gaussian distribution is

$$g(\mathbf{z}) := \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \right\}, \text{ where } \mathbf{z} \in \mathbb{R}^n.$$

We expand $\ln f(x)$ using the second order approximation of Taylor series around the stationary point x_m

$$\ln f(x) \approx \ln f(x_m) + (x-x_m)^T \nabla_x \ln f(x_m) + \frac{1}{2}(x-x_m)^T \nabla_x^2 \ln f(x_m)(x-x_m)$$

Note that, since x_m is the mode of X, i.e., it maximizes $q(x)$, and therefore $f(x)$ at $x = x_m$, it also maximizes $\ln f(x)$ at $x = x_m$.

Hence, $\nabla_x \ln f(x_m) = 0$.

The following information should be so legible that a machine can read it

Roll No : e.g., 190040001 **Dept.**: e.g., CSE

The approximation can be reduced to

$$\ln f(x) = \ln f(x_m) - \frac{1}{2}(x - x_m)^T A (x - x_m)$$

where $A = -\nabla_x^2 \ln f(x_m)$

$$\Rightarrow \quad f(x) = f(x_m) \exp\left\{-\frac{1}{2}(x - x_m)^T A (x - x_m)\right\}$$

(b) What is a necessary condition (e.g., on $\nabla^2 \ln f(\mathbf{x}_m)$) so that the above Gaussian approximation is well defined?

2 points.

A necessary condition for the Gausian approximation to be well defined is that $A = -\nabla_x^2 \ln f(x_m)$ should be positive definite. This is because x_m has to be a local maxima of this distribution (and not a minima or a saddle point).

(c) Now consider a dataset D with the parameter vector given by $\boldsymbol{\theta}$. Recall that $p(D|\boldsymbol{\theta})$ is the *likelihood* and $p(\boldsymbol{\theta})$ is the *prior* of the parameter vector. How is the *maximum aposteriori probability (MAP)* estimate of $\boldsymbol{\theta}$, given by $\boldsymbol{\theta}_{\text{MAP}}$, related to the joint distribution $p(D, \boldsymbol{\theta})$? **1 point.**

$$\theta_{MAP} \in \underset{\theta}{\operatorname{argmax}}\ p(\theta|D) = \underset{\theta}{\operatorname{argmax}}\ p(D, \theta)$$

(d) Using the results derived so far and using the Gaussian approximation of $p(D, \boldsymbol{\theta}) =: f(\boldsymbol{\theta})$ and identifying $p(D) =: F$ (the normalization factor), fill in the blank below with an expression involving only $\boldsymbol{\theta}_{\text{MAP}}$, $\nabla^2_{\boldsymbol{\theta}} \ln p(D, \boldsymbol{\theta}_{\text{MAP}})$, and other constants. **4 points.**

$$\ln p(D) \approx \ln p(D|\boldsymbol{\theta}_{\text{MAP}}) + \boxed{\quad ? \quad}.$$

Show every important step of your derivation.

Note that θ_{MAP} is the mode of the distribution $p(\theta|D)$, hence is a maximizer of $p(D,\theta)$ as well. Using the result of part (a), we get

$$p(D,\theta) = f(\theta) = f(\theta_{MAP}) \exp\left\{-\frac{1}{2}(\theta-\theta_{MAP})^T A (\theta-\theta_{MAP})\right\}$$ --- ①

where $\boxed{A = -\nabla_\theta^2 \ln f(\theta_{MAP})}$ --- ②

Also, $$p(D) = F = \int f(\theta)\, d\theta = f(\theta_{MAP}) \cdot \frac{(2\pi)^{n/2}}{|A|^{1/2}}$$ --- ③

Since, the Gaussian distribution ensures

$$1 = \int \frac{1}{(2\pi)^{n/2}|A|^{-1/2}} \exp\left\{-\frac{1}{2}(\theta-\theta_{MAP})^T A (\theta-\theta_{MAP})\right\} d\theta$$

So, from ② we get

$$\ln p(D) = \ln f(\theta_{MAP}) + \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|A|$$

next page →

$$= \ln p(D, \theta_{MAP}) + \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |A|$$

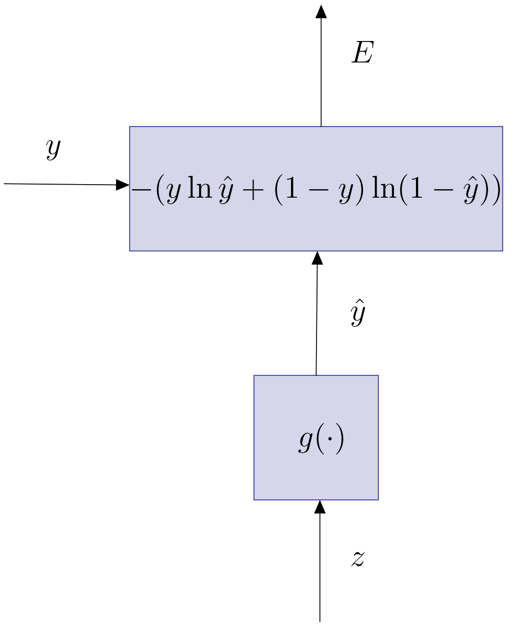
$$= \ln p(D | \theta_{MAP}) + \boxed{\ln p(\theta_{MAP}) + \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |A|}$$

Hence, the expression for the blank is

$$\ln p(\theta_{MAP}) + \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |A|$$

where $A = -\nabla_\theta^2 \ln p(D, \theta_{MAP})$

Problem 4 (10 points). Consider the following computation graph for the final layer of a feedforward neural network.



In this question, we will consider two different activation functions and find the derivative of the error function expressed in terms of the true label y and the estimate $\hat{y}$.

(a) Express $\partial E/\partial z$ in terms of the true label y and the estimate $\hat{y}$ when $g \equiv \sigma$, the sigmoid function. Show the steps of derivation. Recall:

$$\sigma(x) = \frac{1}{1+\exp(-x)}.$$

4 points.

Using the chain rule of differentiation,

$$\frac{\partial E}{\partial z} = \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z}$$

① $$\frac{\partial E}{\partial \hat{y}} = -\frac{\partial}{\partial \hat{y}}\left(y \ln \hat{y} + (1-y)\ln(1-\hat{y})\right)$$

$$= -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

② $$\frac{\partial \hat{y}}{\partial z} = \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1+e^{-z}-1}{(1+e^{-z})^2}$$

$$= \left(\frac{1}{1+e^{-z}}\right)\left(1-\frac{1}{1+e^{-z}}\right) = \hat{y}(1-\hat{y})$$

$$\Rightarrow \frac{\partial E}{\partial z} = \hat{y} - y.$$

The following information should be so legible that a machine can read it

Roll No : e.g., 190040001 ______ **Dept.**: e.g., CSE ______

IIT Bombay
CS 217/337: AIML
Midsem Exam, 2023-24-II
Date: March 2, 2024

(b) Express $\partial E/\partial z$ in terms of the true label y and the estimate $\hat{y}$ when $g \equiv \tanh$, the hyperbolic tangent function. Show the steps of derivation. Recall:

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}.$$

6 points.

We know that tanh is a scaled and translated version of the sigmoid function.

$$\hat{y} = \tanh(z) = 2\sigma(2z) - 1$$

hence, $\frac{\partial \hat{y}}{\partial z} = 2\sigma'(2z) \cdot 2$

$$= 4\left[\sigma(2z)(1-\sigma(2z))\right]$$

$$= 4\left[\frac{1}{2}(1+\hat{y}) \cdot \frac{1}{2}(1-\hat{y})\right] = (1+\hat{y})(1-\hat{y})$$

$\frac{\partial E}{\partial \hat{y}}$ remains as derived in part (a).

Hence, $\frac{\partial E}{\partial z} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \cdot (1+\hat{y})(1-\hat{y})$

$$= \left(1 - \frac{y}{\hat{y}}\right)(1+\hat{y}) \qquad \square$$

END OF QUESTION PAPER. GOOD LUCK!