CS 4650 Problem Set 0 (Fall 2025)

Deadline: Tuesday, September 2, 2025, 11:59PM

IMPORTANT please read the following paragraphs carefully:

Problem Set 0 is a brief review of mathematical concepts necessary to succeed in this class. CS 4650 will cover deep learning and other machine learning methods for natural language processing. These will be discussed in a level of mathematical detail that is commonly understood by modern NLP engineers and researchers. We will do our best to make this material easy to follow, but there is a certain level of mathematical background required, that is not possible to fully cover in this course. To succeed in the course, it is important for students to be familiar with the concepts and notation from probability, linear algebra and calculus. If you see any concepts in this assignment you don't recognize, this is a sign that you need another math course before taking CS 4650 - please feel free to reach out to the course staff to discuss.

For legibility purposes, you are encouraged to use a typesetting software like Latex (see Overleaf for a quick and easy setup). However, you can also scan and upload your handwritten solutions to Gradescope as a PDF. Please write your answers clearly, as **we won't be able to award credit for answers that we deem illegible**. Please show your workings for every question unless specifically mentioned.

REMINDER: It is your responsibility to abide by academic integrity policies of this course. All incidents of suspected dishonesty, plagiarism, or violations of the Georgia Tech Honor Code will be subject to the institute's Academic Integrity procedures.

This may unfortunately lead to severe consequences, e.g. academic probation or dismissal, grade penalties, a 0 grade for assignments concerned, and prohibition from withdrawing from the class.

Please submit your solutions on Gradescope.

1 Joint and Marginal Probabilities

On a social media platform developed by Company Q, a new A.I. system was integrated with human moderators to help identify spam messages. With the assistance of this AI, moderators can work more efficiently, blending the best of technology with human judgment.

For this study:

- $\boldsymbol{S}$ indicates whether a message is spam, as judged by expert reviewers; this should be treated as ground truth.
- $\boldsymbol{A}$ represents the prediction of the A.I. system on whether a message is spam.
- $\boldsymbol{H}$ reflects the human moderators' decision. While they aim for accuracy, they might sometimes get it wrong, either missing spam or incorrectly flagging a legitimate message.

By examining a small portion of existing data, we can construct a complete joint probability distribution table for the three random variables: $\boldsymbol{S}$, $\boldsymbol{A}$ and $\boldsymbol{H}$.

	$\boldsymbol{A = 0}$		$\boldsymbol{A = 1}$	
	$\boldsymbol{H = 0}$	$\boldsymbol{H = 1}$	$\boldsymbol{H = 0}$	$\boldsymbol{H = 1}$
$\boldsymbol{S = 0}$	0.40	0.04	0.08	0.04
$\boldsymbol{S = 1}$	0.08	0.2	0.06	0.10

Table 1: Joint Probability Distribution of $\boldsymbol{S}$, $\boldsymbol{A}$, and $\boldsymbol{H}$

(a) **(1 point)** What is the marginal probability of $\boldsymbol{A = 0}$ i.e. A.I. system classification is "Not Spam"?

Sum of all cells with $A = 0$

$\Rightarrow 0.40 + 0.04 + 0.08 + 0.2 = 0.72$

(b) **(1 point)** What is the value of $\boldsymbol{P(A = 1 | S = 1, H = 0)}$ i.e. the A.I. system classification is "Spam" given the ground truth is "Spam" and human moderator classification is "Not Spam"?

$$P{:}(A = 1 | S = 1, H = 0) = \frac{P(A = 1, S = 1, H = 0)}{P(S = 1, H = 0)} = \frac{0.06}{0.08 + 0.06} = \frac{3}{7} \approx 0.429$$

(c) **(1 point)** What is the value of $\boldsymbol{P(A = 0 | H = 1)}$ i.e. the A.I. system classification is "Not Spam" given the human classification is "Spam"?

$$P(A = 0 | H = 1) = \frac{P(A = 0, H = 1)}{P(H = 1)} = \frac{0.04 + 0.2}{0.04 + 0.2 + 0.04 + 0.1} = \frac{12}{19} \approx 0.6316$$

2 Independence

(2 points) Select 3 statements that are **False**. (No need to show your working).

(a) If $P(A|B) = P(A)$, then $P(B|A) = P(B)$

(b) If $P(A, B|C) = P(C|A, B)$, then $P(A|C) = P(B|C)$ -> False

(c) If $P(B|C) = P(B)$, then $P(A|B, C) = \dfrac{P(A|B) \times P(A|C)}{P(A)}$ -> False

(d) If $P(A, B|C) = P(A|C)P(B|C)$, then $P(A|B) = P(A)$ -> False

3 Bayes Rule

Earthquakes and burglaries are independent events. The probability of either of them happening are shown in Table 2 and 3. Either of these events can cause a burglary alarm to go off. The value of 1 represents the occurrence of an event, and the value of 0 represents the non-occurrence.

$\boldsymbol{E}$	$\boldsymbol{P(E)}$
1	0.01
0	0.99

Table 2: The probability of an earthquake occurring

$\boldsymbol{B}$	$\boldsymbol{P(B)}$
1	0.02
0	0.98

Table 3: The probability of a burglary occurring

The conditional probability of $\boldsymbol{P(A|B, E)}$ is provided in Table 4.

For example, $\boldsymbol{P(A = 1|B = 0, E = 1)} = 1.0000$ denotes the probability that if there is an earthquake and no burglary, the alarm will go off.

(a) **(2 points)** Given all the information above, complete the joint probability distribution table provided in Table 5.

B	E	A	$P(A\|B,E)$
0	0	0	1.0000
0	0	1	0.0000
0	1	0	0.0000
0	1	1	1.0000
1	0	0	0.0000
1	0	1	1.0000
1	1	0	0.0000
1	1	1	1.0000

Table 4: Conditional probability table

B	E	A	$P(A,B,E)$
0	0	0	0.9702
0	0	1	0.0000
0	1	0	0.0000
0	1	1	0.0098
1	0	0	0.0000
1	0	1	0.0198
1	1	0	0.0000
1	1	1	0.0002

Table 5: Joint Probability Table

(b) **(1 point)** Given the alarm has sounded, what is the probability of a burglary having occurred, without any knowledge about an earthquake i.e., $P(B = 1|A = 1)$.

$$P(B=1|A=1) = \frac{P(B=1, A=1)}{P(A=1)} = \frac{P(B=1|A=1)}{1-P(B=0,E=0)} = \frac{0.0198 + 0.0002}{1-0.9702} \approx 0.6711$$

4 Entropy

An NLP researcher is trying to understand the relation between sentiment of movie reviews (Y) and their length in number of words (X). Given the two random variables X, and Y, the researcher has collected a dataset of 10 movie reviews and their corresponding sentiment (positive or negative) to estimate their distribution. The collected data is shown below.

Review ID	Length (X)	Sentiment (Y)
1	Short	Positive
2	Short	Negative
3	Medium	Positive
4	Short	Negative
5	Long	Positive
6	Medium	Negative
7	Medium	Positive
8	Long	Negative
9	Short	Positive
10	Medium	Positive

Consider that the random variable X takes the values 0, 1, and 2 for short, medium and long respectively, and the random variable Y takes the values 0 and 1 for Negative and Positive respectively. Use the above data to answer the following questions.

(a) (**1 point**) Write the marginal probability distribution tables, $p(X)$ and $p(Y)$ for the random variables X and Y.

(b) (**1 point**) Calculate the entropy of X and Y, i.e., $H(X)$ and $H(Y)$ in bits.

(c) (**1 point**) Write the conditional probability distribution table for the variable Y given X i.e. $p(Y \mid X)$.

(d) (**1 point**) Calculate the conditional entropy values of the following in bits:

(i) Y given $X = 0$ i.e., $H(Y|X = 0)$

(ii) Y given $X = 1$ i.e., $H(Y|X = 1)$

(iii) Y given $X = 2$, i.e., $H(Y|X = 2)$

4/

a) $p(X=0) = 0.4$, $p(X=1) = 0.4$, $p(X=2) = 0.2$

$p(Y=0) = 0.4$, $p(Y=1) = 0.6$

b) Entropy in bits: $H(U) = -\sum p \log_2 p$

- $H(X) = H(0.4, 0.4, 0.2) \approx 1.5219$ bits
- $H(Y) = H(0.4, 0.6) \approx 0.9710$ bits

c) $p(Y|X)$:

- For $X=0$ (Short): 2 Pos, 2 Neg -> $p(Y=1|X=0) = 0.5$, $p(Y=0|X=0) = 0.5$
- For $X=1$ (Medium): 3 Pos, 1 Neg -> $p(Y=1|X=1) = 0.75$, $p(Y=0|X=1) = 0.25$
- For $X=2$ (Long): 1 Pos, 1 Neg -> $p(Y=1|X=2) = 0.5$, $p(Y=0|X=2) = 0.5$

d) i) $H(Y|X=0) = H(0.5, 0.5) = 1.0000$ bit

ii) $H(Y|X=1) = H(0.25, 0.75) \approx 0.8113$ bit

iii) $H(Y|X=2) = H(0.5, 0.5) = 1.0000$ bit

e) $H(Y|X) = \sum_x p(x) H(Y|X=x) = (0.4 \cdot 1) + (0.4 \cdot 0.8113) + (0.2 \cdot 1) \approx 0.9245$ bit

(e) (**1 point**) Calculate the conditional entropy of Y given X, i.e., $H(Y|X)$ in bits.

5 Probability

(a) A probability density function for three different independent random variables x_i, where $i = \{1, 2, 3\}$ is given by

$$f_i(x_i) = \begin{cases} \beta_i e^{-\alpha_i x_i} & \text{if } x_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

(i) (**1 point**) If $\alpha_3 > \alpha_2 > \alpha_1$, find the relation between $\beta_1, \beta_2, \beta_3$
Hint: The β symbol for each of the densities is the normalization constant of the respective densities.

(ii) (**2 points**) Compute the expected value and variance of x_1,i.e., $E[x_1]$ and $Var(x_1)$ if $\alpha_1 = 1$.

(b) (**1 point**) Imagine a potluck dinner with four guests, each of who bring a unique and distinct dish. If the dishes are collected at one place and randomly redistributed to the guests to eat, what is the probability that at least one guest ends up with their own dish?

6 Calculus Review

Consider the following function - KL divergence between two distributions p(X) and q(X) for a discrete random variable X over 3 possible outcomes. $X = 1$ indicates that an input word is a noun, $X = 2$ indicates it is a verb, and $X = 3$ indicates it is an adjective:

$$\mathrm{D}_{KL}(p||q) = \sum_{i=1}^{3} p_i \ln\left(\frac{p_i}{q_i}\right)$$

Here p_i, q_i are shorthand for $p(X = i), q(X = i)$ respectively.
Assume p = $[p_1, p_2, p_3]$ to be a distribution defining ground truth labels on a classification task.

Take q to the your guess weightage spread across a probability distribution such that $q_1 + q_2 + q_3 = 1$ and $0 \leq q_i \leq 1$. For example, if our input word is "happy", an adjective, so that p = $[0.05, 0.05, 0.9]$, but our q might be imperfect like $[0.3, 0.6, 0.1]$. Lets investigate the gradient of this function w.r.t. our guess distribution q !

5/

a)

i) For $f_i(x) = \beta_i e^{-\alpha_i x}$ on $x > 0$: normalization gives

$$1 = \int_0^\infty \beta_i e^{-\alpha_i x} dx = \frac{\beta_i}{\alpha_i} \Rightarrow \beta_i = \alpha_i$$

$$\boxed{\Rightarrow \alpha_3 > \alpha_2 > \alpha_1 \text{ , then } \beta_3 > \beta_2 > \beta_1}$$

ii) With $\alpha_1 = 1$, this is the exponential $Exp(\alpha_1)$ with

$$\boxed{E[x_1] = \frac{1}{\alpha_1} = 1 \text{ , } Var(x_1) = \frac{1}{\alpha_1^2} = 1}$$

b) "At least one guest gets their own dish" for 4 permutations = 1 - (no one gets their own)

$$\boxed{\Rightarrow 1 - \frac{!4}{4!} = 1 - \frac{9}{24} = \frac{15}{24} = \frac{5}{8} = 0.625}$$

6/

$$D_{KL}(p\|q) = \sum_{i=1}^{3} p_i \ln \frac{p_i}{q_i} = \sum_i p_i \ln q_i - \sum_i p_i \ln q_i$$

a) $$\frac{\partial}{\partial q_i} D_{KL}(p\|q) = -\frac{p_i}{q_i}$$

Vector form: $\nabla_q D_{KL}(p\|q) = -\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}, \frac{p_3}{q_3}\right)$

b) For $p = [0.05, 0.05, 0.9]$, $q = [0.3, 0.6, 0.1]$

$$\boxed{\begin{aligned} &\nabla_q = -\left(\frac{0.05}{0.3}, \frac{0.05}{0.6}, \frac{0.9}{0.1}\right) = (-0.1667, -0.0833, -9) \\ &\ell_2 \text{ norm} = \sqrt{0.1667^2 + 0.0833^2 + 9^2} \approx 9.0019 \end{aligned}}$$

c) For $q = [0.1, 0.2, 0.7]$: $\nabla_q = -(0.5, 0.25, 1.285714...)$, $\|\nabla_q\|_2 \approx 1.4020$

The closer q is to p, the smaller the gradient magnitude. At the optimum $q = p$, the gradient is zero

(a) (**1 point**) Find $\frac{\partial}{\partial q_i}\mathrm{D}_{KL}$ symbolically and report the gradient $\frac{\partial}{\partial q}\mathrm{D}_{KL}$.

(b) (**1 point**) Plug in p = $[0.05, 0.05, 0.9]$, and q = $[0.3, 0.6, 0.1]$ in the expression for the gradient above. What is the L_2 norm of this gradient?

(c) (**2 points**) If **q** were instead $[0.1, 0.2, 0.7]$, what would the value for $\frac{\partial}{\partial \mathbf{q_1}}\mathrm{D}_{KL}$, $\frac{\partial}{\partial \mathbf{q_2}}\mathrm{D}_{KL}$, and $\frac{\partial}{\partial \mathbf{q_3}}\mathrm{D}_{KL}$ be as a result? What is the L_2 norm of this gradient? What is the relationship between the correctness of your guess and the magnitude of your gradient?

7 Multivariate Calculus

Here are the equations which define a GRU, an architecture in sequence modeling explicitly designed to keep gradient norms through long sequences (biases omitted in equations for brevity).

Gating functions:

$$\mathbf{z_t} = \sigma(\mathbf{W^{z_t}_{x_t} x_t} + \mathbf{W^{z_t}_{h_{t-1}} h_{t-1}})$$
$$\mathbf{r_t} = \sigma(\mathbf{W^{r_t}_{x_t} x_t} + \mathbf{W^{r_t}_{h_{t-1}} h_{t-1}})$$

Update and reset gates:

$$\mathbf{\tilde{h}_t} = \tanh\Big(\mathbf{W^{\tilde{h}_t}_{x_t} x_t} + \mathbf{W^{\tilde{h}_t}_{h_{t-1}}}(\mathbf{r_t} \odot \mathbf{h_{t-1}})\Big)$$

Hidden state update:

$$\mathbf{h_t} = (1 - \mathbf{z_t}) \odot \mathbf{h_{t-1}} + \mathbf{z_t} \odot \mathbf{\tilde{h}_t}$$

The details of the function symbols above, and some helpful identities:

$$\mathbf{z_t}, \mathbf{r_t}, \mathbf{h_t}, \mathbf{\tilde{h}_t}, \mathbf{h_{t-1}} \in \mathbb{R}^{d\times 1}$$
$$\mathbf{W^{z_t}_{x_t}}, \mathbf{W^{z_t}_{h_{t-1}}}, \mathbf{W^{r_t}_{x_t}}, \mathbf{W^{r_t}_{h_{t-1}}}, \mathbf{W^{\tilde{h}_t}_{x_t}}, \mathbf{W^{\tilde{h}_t}_{h_{t-1}}}, \in \mathbb{R}^{d\times d}$$

For scalars

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$
$$\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1-\sigma(x))$$
$$\frac{\partial}{\partial x}\tanh(x) = 1 - \tanh^2(x)$$

For vectors

$$\sigma(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}} = \frac{e^{\mathbf{x}}}{e^{\mathbf{x}}+1} \in \mathbb{R}^{d\times 1}$$

$$\frac{\partial}{\partial \mathbf{x}}\sigma(\mathbf{x}) = \text{diag}(\sigma(\mathbf{x})(1-\sigma(\mathbf{x}))) \in \mathbb{R}^{d\times d}$$

$$\frac{\partial}{\partial \mathbf{x}}\tanh(\mathbf{x}) = \text{diag}(1-\tanh^2(\mathbf{x})) \in \mathbb{R}^{d\times d}$$

Note: $\sigma(\mathbf{x}), e^{\mathbf{x}}, \tanh(\mathbf{x})$ (and $\tanh^2(x)$) are applied element wise

$\odot$ is element wise multiplication

the ones used with vectors are also applied element wise

A useful operator for your answers diag(x): The operator diag(x) transforms a vector $\mathbf{x}$ into a square matrix by placing the elements of x on the diagonal of the matrix, and filling the off-diagonal elements with zeros.

$$\text{For a vector } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{the matrix diag}(\mathbf{x}) \text{ is given by:}$$

$$\begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_n \end{pmatrix}$$

You may find Matrix Cookbook page 8 useful for some of its identities (also uploaded on canvas).

(a) (**1 point**) For this portion, consider d = 1, what would the dimension of $\frac{\partial}{\partial h_{t-1}} z_t$ be and find $\frac{\partial}{\partial h_{t-1}} z_t$ symbolically. What is the dimension of $\frac{\partial}{\partial \tilde{h}_t} h_t$. Find the symbolic expression for $\frac{\partial}{\partial \tilde{h}_t} h_t$.

(b) (**2 points**) Repeat part (a) considering now d as larger than 1. Explicitly: what would the dimension of $\frac{\partial}{\partial \mathbf{h_{t-1}}} \mathbf{z_t}$ be and find $\frac{\partial}{\partial \mathbf{h_{t-1}}} \mathbf{z_t}$ symbolically. What is the dimension of $\frac{\partial}{\partial \mathbf{\tilde{h}_t}} \mathbf{h_t}$. Find the symbolic expression for $\frac{\partial}{\partial \mathbf{\tilde{h}_t}} \mathbf{h_t}$.

7/

Given:

$$z_t = \sigma(W_x^{(z)} x_t + W_h^{(z)} h_{t-1})$$

$$r_t = \sigma(W_x^{(r)} x_t + W_h^{(r)} h_{t-1})$$

$$\tilde{h}_t = \tanh(W_x^{(\tilde{h})} x_t + W_h^{(\tilde{h})} (r_t \odot h_{t-1}))$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

a) Scolar case $d = 1$

- Dimension of $\dfrac{\partial z_t}{\partial h_{t-1}}$: 1×1 (a scalar)

Let $a^{(z)} = W_x^{(z)} x_t + W_h^{(z)} h_{t-1}$

$$\Rightarrow \frac{\partial z_t}{\partial h_{t-1}} = \sigma'(a^{(z)}) W_h^{(z)} = z_t (1 - z_t) W_h^{(z)}$$

- Dimension of $\dfrac{\partial h_t}{\partial \tilde{h}_t}$: 1×1 (a scalar)

Holding z_t fixed in this partial

$$\Rightarrow \frac{\partial h_t}{\partial \tilde{h}_t} = z_t$$

b) Vector case $d > 1$

- Dimension of $\dfrac{\partial z_t}{\partial h_{t-1}}$: $d \times d$ (Jacobian)

With $a^{(z)} = W_x^{(z)} x_t + W_h^{(z)} h_{t-1}$

$$\frac{\partial z_t}{\partial h_{t-1}} = \operatorname{diag}(z_t \odot (1 - z_t)) W_h^{(z)} \qquad (\in \mathbb{R}^{d \times d})$$

- Diamension of $\dfrac{\partial h_t}{\partial \tilde{h}_t}$: $d \times d$

From $h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$

$$\frac{\partial h_t}{\partial \tilde{h}_t} = \operatorname{diag}(z_t) \qquad (\in \mathbb{R}^{d \times d})$$