CS534 — Written Homework Assignment 0 —*Solution*

Overview and Objectives. In this homework, you are going to practice some of the skills we'll be using in class. If you can solve most of these problems (even if you have to Google some identities or brush up on your knowledge) then you should be very well-equipped for the course. Otherwise, you'll need to spend extra time revising these topics. **This assignment is to help you (and us) gauge your familiarity with these concepts and might be a bit challenging depending on your background.**

How to Do This Assignment. We prefer solutions typeset in LaTeX but will accept scanned written work if it is legible. If a TA can't read your work, they can't give you credit. Submit your solutions to Canvas as a PDF.

Advice. Start early. Start early. Start early. You may be rusty on some of this material. Some of it might be new to you depending on your background – seek out resources to refresh yourself if so. Some helpful references:

- **Probability Refresher:** CS229 Probability Review from Stanford
- **Linear Algebra Refresher:** Zico Koltur's Linear Algebra Review
- **Differential Calculus Refresher:** Jackie Nicholas' booklet from University of Sydney
- **Integration Refresher:** Mary Barnes' booklet from University of Sydney

Linear algebra

(a) **Transpose and Associative Property, Positive Semi-definite matrices [2pt]** Define a matrix $B = bb^T$, where $b \in \mathbf{R}^{d\times 1}$ is a column vector that is not all-zero. Show that B is a positive semi-definite matrix.

[*Hint*: To show that B is positive semi-definite, we need to show that B is symmetric, and for any vector $x \in \mathbf{R}^{d\times 1}$, $x^TBx \geq 0$. For the latter, try to get x^TBx to look like the product of two identical scalars. Note that $b^Tx = (x^Tb)^T$, that $a^T = a$ for scalar value a, and that matrix multiplication is associative.]
Symmetry of B: $B^T = (bb^T)^T = (b^T)^Tb^T = bb^T = B$

Note that as x and b are both column vectors (`d x 1` dimension), x^Tb results in a scalar value (product of matrices of dimensions `1 x d` and `d x 1` yields `1 x 1`). Let's call this scalar value $a = x^Tb$

$$\begin{aligned} x^TBx &= x^Tbb^Tx && \text{[Definition of } B\text{]} \\ &= x^Tb(x^Tb)^T && \text{[Property of Transpose]} \\ &= a(a)^T = a^2 && \text{[Our definition of } a\text{. And the transpose of a scalar is a scalar]} \\ a^2 &\geq 0 && \text{[}a^2\text{ is non-negative for real-valued } a\text{]} \end{aligned}$$

(b) **Solving systems of linear equations with matrix inverse. [2pt]** Consider the following set of linear equations:

$$\begin{cases} x_1 + x_2 - x_3 - x_4 = 1 \\ 2x_1 + 5x_2 - 7x_3 - 5x_4 = -2 \\ 2x_1 - x_2 + x_3 + 3x_4 = 4 \\ 5x_1 + 2x_2 - 4x_3 - 2x_4 = 6 \end{cases}$$

(a) (1 pt) Please express the system of equations as $A\mathbf{x} = \mathbf{b}$ by specifying the matrix A and vector $\mathbf{b}$

$$A = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 2 & 5 & -7 & -5 \\ 2 & -1 & 1 & 3 \\ 5 & 2 & -4 & -2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ -2 \\ 4 \\ 6 \end{bmatrix}$$

(b) (1 pt) Solve for $A\mathbf{x} = \mathbf{b}$ by using the matrix inverse of A (you can use software to compute the inverse).

$$A^{-1} = \begin{bmatrix} 1/2 & -1/6 & 0 & 1/6 \\ 2 & 1/6 & 1/2 & -2/3 \\ 7/4 & -1/4 & 0 & -1/4 \\ -1/4 & 1/4 & 1/2 & -1/4 \end{bmatrix}$$

$$\mathbf{x} = \mathbf{A^{-1}b} = \begin{bmatrix} 11/6 \\ -1/3 \\ 3/4 \\ -1/4 \end{bmatrix}$$

Vector Calculus

(a) **Derivatives.[2pt]**. Compute the derivative $f'(x)$ for

(a) (1 pts) the logistic (aka sigmoid) function $f(x) = \frac{1}{1+\exp(-x)}$

$$\begin{aligned} f'(x) &= \frac{-1}{(1+exp(-x))^2} \times \frac{d(1+exp(-x))}{dx} \\ &= \frac{-1}{(1+exp(-x))^2} \times \frac{d(exp(-x))}{dx} \\ &= \frac{-1}{(1+exp(-x))^2} \times (-exp(-x)) = \frac{exp(-x)}{(1+exp(-x))^2} \\ &= \frac{1}{(1+exp(-x))} \times \frac{exp(-x)}{(1+exp(-x))} = f(x) \times (1-f(x)) \end{aligned}$$

(b) (1 pts) $f(x) = \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$

$$\begin{aligned} f'(x) &= \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \times \frac{d(-\frac{1}{2\sigma^2}(x-\mu)^2)}{dx} \\ &= \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \times (-\frac{2}{2\sigma^2}(x-\mu)) \\ &= \frac{-(x-\mu)}{\sigma^2} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \end{aligned}$$

(b) **Gradients.** **[3pt]**Compute the gradient $\nabla_{\mathbf{x}} f$ of the following functions. Please clearly specify the dimension of the gradient.

(a) (1pt)

$$f(z) = \log(1+z),\ z = \mathbf{x}^T\mathbf{x},\ \mathbf{x} \in R^D$$

Use the property for derivative of a log first, then apply the chain rule and vector derivative properties:

$$\nabla_{\mathbf{x}} f = \frac{1}{1+z} \times \nabla_{\mathbf{x}} z, \text{ where } z = \mathbf{x}^{\mathbf{T}}\mathbf{x}$$

.

Now, we have:

$$\nabla_{\mathbf{x}} z = \nabla_{\mathbf{x}} \mathbf{x}^{\mathbf{T}}\mathbf{x} = 2\mathbf{x}$$

Finally, sub in $\mathbf{x}^{\mathbf{T}}\mathbf{x}$ *for* z*:*

$$\nabla_{\mathbf{x}} f = \frac{2\mathbf{x}}{1+\mathbf{x}^{\mathbf{T}}\mathbf{x}} \in R^D$$

(b) (2pt)

$$f(z) = \exp\left(-\frac{1}{2}z\right)$$
$$z = g(\mathbf{y}) = \mathbf{y}^T S^{-1}\mathbf{y}$$
$$\mathbf{y} = h(\mathbf{x}) = \mathbf{x} - \mu$$
where $\mathbf{x}, \mu \in R^D, S \in R^{D\times D}$ is a symmetric matrix.

Recall first the property (from Matrix cookbook) that:

$$\nabla_{\mathbf{x}} \mathbf{x}^{\mathbf{T}} M\mathbf{x} = (M + M^T)\mathbf{x}, \text{ and } 2 \times M \times \mathbf{x} \text{ if } M \text{ is symmetric}$$

Here we have $M = S^{-1}$ *and a symmetric matrix. The rest comes directly from derivatives of functions and the chain rule:*

$$\nabla_{\mathbf{x}} f = -\frac{1}{2}\exp\left(-\frac{1}{2}z\right)(\nabla_{\mathbf{x}} z) \quad \textit{[Derivative of } f(x) = e^x \textit{ is } f \textit{ itself, and Chain rule]}$$

$$\nabla_{\mathbf{x}} f = -\frac{1}{2}\exp\left(-\frac{1}{2}z\right) 2S^{-1}\mathbf{y}(\nabla_{\mathbf{x}}\mathbf{y}) \quad \textit{[From the statement above]}$$

$$\nabla_{\mathbf{x}} f = -\exp\left(-\frac{1}{2}z\right) S^{-1}\mathbf{y}\nabla_{\mathbf{x}}(\mathbf{x}-\mu) \quad \textit{[note } \nabla_{\mathbf{x}}(\mathbf{x}-\mu) = I\textit{]}$$

$$\nabla_{\mathbf{x}} f = -\exp\left(-\frac{1}{2}z\right) S^{-1}\mathbf{y} \quad \textit{[now substitute in } z = \mathbf{y}^T S^{-1}\mathbf{y} \textit{ and } \mathbf{y} = \mathbf{x} - \mu\textit{]}$$

$$\nabla_{\mathbf{x}} f = -\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T S^{-1}(\mathbf{x}-\mu)\right) S^{-1}(\mathbf{x}-\mu) \quad \textit{[Final solution} \in R^D\textit{]}$$

Probability

(a) **Joint, Marginal, and Conditional Probabilities [2pt]** Consider two discrete random variables X and Y with the following joint distribution:

Please compute:

(a) (1 pt) The marginal distributions $p(x)$ and $p(y)$

$p(y)$:

y_1	y_2	y_3
0.26	0.47	0.27

$p(x)$:

x_1	x_2	x_3	x_4	x_5
0.16	0.17	0.11	0.22	0.34

	y_1	0.01	0.02	0.03	0.1	0.1
Y	y_2	0.05	0.1	0.05	0.07	0.2
	y_3	0.1	0.05	0.03	0.05	0.04
		x_1	x_2	x_3	x_4	x_5

X

(b) (1 pt) The Conditional distribution $p(x|Y = y_1)$ and $p(y|X = x_3)$

$p(x|Y = y_1)$:

x_1	x_2	x_3	x_4	x_5
1/26	*2/26*	*3/26*	*10/26*	*10/26*

$p(y|X = x_3)$:

y_1	y_2	y_3
3/11	*5/11*	*3/11*

(b) **Conditional probabilities, Marginalization and Bayes Rule [5pt]** Consider two coins, one is fair and the other one has a 1/10 probability for head. Now you randomly pick one of the coins, and toss it twice. Answer the following questions.

(a) (1pt) What is the probability that you picked the fair coin? What is the probability of the first toss being head?

The probability that you picked the fair coin is 0.5. Let x_1 denote the outcome of the first toss and let y denote the coin that is selected. We can write down the following probabilities.

$P(y = f) = P(y = u) = \frac{1}{2}$

$P(x_1 = h|y = f) = \frac{1}{2}$

$p(x_1 = h|y = u) = 1/10$

Now we can write out the probability of the first toss being head as:

$P(x_1 = h) = P(x_1 = h, y = f) + P(x_1 = h, y = u)$

$= P(y = f)P(x_1 = h|y = f) + P(y = u)P(x_1 = h|y = u) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{10}$

$= \frac{1}{2} \times \frac{6}{10} = \frac{3}{10}$

(b) (2pts) If both tosses are heads, what is the probability that you have chosen the fair coin (Hint: you should apply Bayes Rule for this)?

Let x_1, x_2 denote the outputs of the first two tosses. It is easy to see that

$P(x_1 = h, x_2 = h|y = f) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$P(x_1 = h, x_2 = h|y = u) = \frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$

Now we need to compute $P(y = f|x_1 = h, x_2 = h)$, to do so, we use Bayes Theorem:

$P(y = f|x_1 = h, x_2 = h) = \frac{P(x_1=h,x_2=h|y=f)P(y=f)}{P(x_1=h,x_2=h)}$

To compute the denominator, we use the same approach as used in (a):

$P(x_1 = h, x_2 = h) = P(x_1 = h, x_2 = h|y = f)P(y = f) + P(x_1 = h, x_2 = h|y = u)P(y = u)$

$= \frac{1}{4} \times \frac{1}{2} + \frac{1}{100} \times \frac{1}{2} = \frac{13}{100}$

Plug this into the Bayes Theorem, we have:

$P(y = f|x_1 = h, x_2 = h) = \frac{P(x_1=h,x_2=h|y=f)P(y=f)}{P(x_1=h,x_2=h)} = \frac{1/4 \times 1/2}{13/100} = \frac{25}{26}$

(c) (2pts) If both tosses are heads, what is the probability that the third coin toss will be head? (you should build on results of c)

We will use x_3 to denote the outcome of the third coin toss.

$$
\begin{aligned}
&P(x_3 = h | x_1 = h, x_2 = h) \\
&= P(x_3 = h, y = f | x_1 = h, x_2 = h) + P(x_3 = h, y = u | x_1 = h, x_2 = h) \\
&= P(x_3 = 1 | y = f) P(y = f | x_1, x_2) + P(x_3 = 1 | y = u) P(y = u | x_1, x_2) \\
&= \frac{1}{2} \times \frac{25}{26} + \frac{1}{10} \times \frac{1}{26} \\
&= \frac{63}{130}
\end{aligned}
$$

(c) **Linearity of Expectation [2 pt]** A random variable x distributed according to a standard normal distribution (mean zero and unit variance) has the following probability density function (pdf):

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Using the properties of expectations, evaluate the following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx$$

[*Hint:* This is *not* a calculus question. The simple solution relies on linearity of expectation and the provided mean/variance of p(x).]

Note that equation (3) is an expectation of $ax^2 + bx + c$ with respect to the normal random variable x.

$$E[ax^2 + bx + c] = \int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx$$

Applying linearity of expectation breaks this down to:

$$E[ax^2 + bx + c] = aE[x^2] + bE[x] + c$$

For standard normal distribution (0 mean, variance 1), we have,

$$E[x] = \int_{-\infty}^{\infty} p(x)x \; dx = 0 \qquad \text{[This is just the mean.]}$$

$$VAR[x] = E[x^2] - E[x]^2 = \left(\int_{-\infty}^{\infty} p(x)x^2 \; dx \right) - 0 = 1 \qquad \text{[This is just the variance.]}$$

$$\rightarrow E[x^2] = 1$$

Hence,

$$aE[x^2] + bE[x] + cE[1] = a + c$$

(d) **Cumulative Density Functions / Calculus [2 pt]** X is a continuous random variable over the interval [0,1], show that the following function p is a valid probability density function (PDF) and derive the corresponding cumulative density function (CDF).

$$p(x) = \begin{cases} 4x & 0 \leq x \leq 1/2 \\ -4x + 4 & 1/2 \leq x \leq 1 \end{cases}$$

[*Hint:* Recall that a function is a valid PDF function if it integrates to 1: $\int_{-\infty}^{\infty} p(x)\, dx = 1$. And the cumulative density function (CDF) is defined as $C(x) = P(X \leq x)$ or the probability that a sample from p is less than x – which can be computed as $C(x) = \int_{-\infty}^{x} p(x)\, dx$. This *is* a calculus question. But the PDf is a piece-wise linear function, hence it is straightforward.]

This question comes down to integrating the piece-wise linear function $p(x)$ with respect to x.

$$\begin{aligned}\int_{-\infty}^{\infty} p(x)\, dx &= \int_{0}^{0.5} 4x\, dx + \int_{0.5}^{1} (-4x+4)\, dx \\ &= 2x^2\Big|_{0}^{0.5} + \left(-2x^2+4x\right)\Big|_{0.5}^{1} \\ &= [2 \cdot 0.25 - 0] + [(4-2) - (-0.5+2)] \\ &= 0.5 + [2 - 1.5] = 0.5 + 0.5 = 1\end{aligned}$$

Remember that $C(x)$ is supposed to equal $\int_{-\infty}^{x} p(x)\, dx$ so we'll need to accumulate the $0 \leq x \leq 1/2$ piece of the integral when writing the expression for the $1/2 \leq x \leq 1$ portion. For similar reasons, we'll also need to define $C(x)$ for the regions outside the interval $[0, 1]$

$$C(x) = \begin{cases} 0 & x \leq 0 \\ \int_{0}^{x} 4x\ dx = 2x^2 & 0 \leq x \leq 1/2 \\ (\int_{0}^{1/2} 4x\ dx) + (\int_{1/2}^{x} -4x+4\ dx) = -2x^2 + 4x - 1 & 1/2 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$