CS7643: Deep Learning
Spring 2021
Problem Set 0

Instructor: Zsolt Kira
TAs: Manas Sahni, Yihao Chen, Mandy Xie, Hrishikesh Kale,
Michael Piseno, Zhuoran Yu , Rahul Duggal
Discussions: https://piazza.com/class/kjsselshfiz18c

Due: Wednesday, Jan 20, 11:59pm ET

Instructions

1. We will be using Gradescope to collect your assignments. Please read the following instructions for submitting to Gradescope carefully! Failure to follow these instructions may result in parts of your assignment not being graded. We will not entertain regrading requests for failure to follow instructions.

   - For Section 1: Multiple Choice Questions, it is mandatory to use the LaTeX template provided on the class webpage (https://www.cc.gatech.edu/classes/AY2021/cs7643_spring/assets/ps0.zip). For every question, there is only one correct answer. To mark the correct answer, change `\choice` to `\CorrectChoice`
   - For Section 2: Proofs - This section has 7 total problems/sub-problems (Q7, Q8a - Q8c and Q9a - Q9c) and your answer to each sub-problem should start on a new page. Please be sure to mark the corresponding pages to the correct question numbers marked on the PS0 outline while submitting on Gradescope.
   - For Section 2, LaTeX'd solutions are strongly encouraged (solution template available at https://www.cc.gatech.edu/classes/AY2021/cs7643_spring/assets/ps0.zip), but scanned handwritten copies are acceptable. If you scan handwritten copies, please make sure to append them to the PDF generated by LaTeX for Section 1 and please mark the pages correctly as mentioned above.

2. Hard copies are **not** accepted.

3. We generally encourage you to collaborate with other students. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and *not* as a group activity. Please list the students you collaborated with.
   Exception: PS0 is meant to serve as a background preparation test. You must NOT collaborate on PS0.

1 Multiple Choice Questions

1. (1 point) We are machine learners with a slight gambling problem (very different from gamblers with a machine learning problem!). Our friend, Bob, is proposing the following payout on the roll of a dice:

$$\text{payout} = \begin{cases} +\$2 & x = 1 \\ -\$1/4 & x \neq 1 \end{cases} \tag{1}$$

where $x \in \{1, 2, 3, 4, 5, 6\}$ is the outcome of the roll, $(+)$ means payout to us and $(-)$ means payout to Bob. Is this a good bet i.e. are we expected to make money?

● **True** ○ False

2. (1 point) X is a continuous random variable with probability density function:

$$p(x) = \begin{cases} 3x^2/32 & 0 \leq x \leq 2 \\ 3(6-x)/32 & 2 \leq x \leq 6 \end{cases} \tag{2}$$

Which of the following statements are true about the equation for the corresponding cumulative density function (CDF) $C(x)$?
[*Hint:* Recall that CDF is defined as $C(x) = Pr(X \leq x)$.]

● $C(x) = x^3/32$ **for** $0 \leq x \leq 2$
○ $C(x) = 9x/16 - 3x^2/64 - 15/16$ for $2 \leq x \leq 6$
○ All of the above
○ None of the above

3. (2 point) A random variable x in standard normal distribution has the following probability density:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3}$$

Evaluate the following integral:

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx \tag{4}$$

[*Hint:* We are not sadistic (okay, we're a little sadistic, but not for this question). This is not a calculus question.]

○ a + b + c ○ c ● **a + c** ○ b + c

4. (2 points) Consider the following function of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$:

$$f(\mathbf{x}) = \sigma\left(\log\left(5\left(\max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6)\right)\right) + \frac{1}{2}\right) \tag{5}$$

where σ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

Compute the gradient $\nabla_{\mathbf{x}} f(\cdot)$ and evaluate it at at $\hat{\mathbf{x}} = (-1, 3, 4, 5, -5, 7)$.

○ $\begin{bmatrix} 0 \\ 0.031 \\ 0.026 \\ -0.013 \\ -0.062 \\ -0.062 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.157 \\ 0.131 \\ -0.065 \\ -0.314 \\ -0.314 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.357 \\ 0.268 \\ -0.214 \\ -0.894 \\ -0.894 \end{bmatrix}$ ● $\begin{bmatrix} 0 \\ 0.357 \\ 0.268 \\ -0.214 \\ -0.447 \\ -0.447 \end{bmatrix}$

5. (2 points) Which of the following functions are convex?

○ $\|\mathbf{x}\|_{\frac{1}{2}}$

○ $\min_{i=1}^{k} \mathbf{a}_i^T \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$, and a finite set of arbitrary vectors: $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$

● $\log\left(1 + \exp(\mathbf{w}^T \mathbf{x})\right)$ **for** $\mathbf{w} \in \mathbb{R}^d$

○ All of the above

6. (2 points) Suppose you want to predict an unknown value $Y \in \mathbb{R}$, but you are only given a sequence of noisy observations $x_1, \ldots, x_n$ of Y with i.i.d. noise ($x_i = Y + \epsilon_i$). If we assume the noise is I.I.D. Gaussian ($\epsilon_i \sim N(0, \sigma^2)$), the maximum likelihood estimate ($\hat{y}$) for Y can be given by:

○ A: $\hat{y} = \text{argmin}_y \sum_{i=1}^{n} (y - x_i)^2$

○ B: $\hat{y} = \text{argmin}_y \sum_{i=1}^{n} |y - x_i|$

○ C: $\hat{y} = \frac{1}{n} \sum_{i=1}^{n} x_i$

● **Both A & C**

○ Both B & C

2 Proofs

7. (3 points) Prove that

$$\log_e x \leq x - 1, \qquad \forall x > 0 \tag{7}$$

with equality if and only if $x = 1$.

[*Hint:* Consider differentiation of $\log(x) - (x - 1)$ and think about concavity/convexity and second derivatives.]

Consider the function $f(x) = \log(x) - (x - 1)$. This function is differentiable when $x > 0$. Therefore, the maximum of this function must occur at the endpoints zero and infinity or where the derivative is zero. Let x approach 0. We have that $\log(x)$ approaches $-\infty$ and $x - 1 = -1$. Thus, the function approaches $-\infty$ as x approaches 0. Let x approach $+\infty$. We have $\log(x)$ approaches $+\infty$ and $x - 1$ approaches $+\infty$. However, $x - 1$ grows linearly while $\log(x)$ grows logarithmically. Thus, the function approaches $-\infty$ as x approaches $+\infty$.

Now we will look at the critical points. The derivative is $f'(x) = \frac{1}{x} - 1$. This vanishes when $x = 1$. The value of $f(1)$ is 0. Thus, because $f(x) \leq 0$ at every boundary point and critical point, $\log(x) \leq x - 1$ everywhere.

8. (6 points) Consider two discrete probability distributions p and q over k outcomes:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{8a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \ldots, k\} \tag{8b}$$

The Kullback-Leibler (KL) divergence (also known as the *relative entropy*) between these distributions is given by:

$$KL(p, q) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) \tag{9}$$

It is common to refer to $KL(p, q)$ as a measure of distance (even though it is not a proper metric). Many algorithms in machine learning are based on minimizing KL divergence between two probability distributions. In this question, we will show why this might be a sensible thing to do.

[*Hint:* This question doesn't require you to know anything more than the definition of $KL(p, q)$ and the identity in Q7]

(a) Using the results from Q7, show that $KL(p, q)$ is always non-negative.

$$\begin{aligned}
KL(p, q) &= \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) \\
-KL(p, q) &= \sum_{i=1}^{k} p_i \log\left(\frac{q_i}{p_i}\right) \\
-KL(p, q) &\leq \sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) \\
-KL(p, q) &\leq \sum_{i=1}^{k} (q_i - p_i) \\
-KL(p, q) &\leq 1 - 1 = 0 \\
KL(p, q) &\geq 0
\end{aligned}$$

(b) When is $KL(p, q) = 0$?

The KL divergence inequality reaches equality when the log inequality reaches equality. This occurs when $q_i/p_i = 1$ for all i. This occurs when $q_i = p_i$ for all i.

(c) Provide a counterexample to show that the KL divergence is not a symmetric function of its arguments: $KL(p, q) \neq KL(q, p)$

Consider two probability distributions $p_0 = 0.4$ and $p_1 = 0.6$. We also have $q_0 = 0.5$ and $p_1 = 0.5$. We have the KL-divergence one way as $0.4 \log(0.4/0.5) + 0.6 \log(0.6/0.5) = 0.00874$. The other way gets us $0.5 \log(0.5/0.4) + 0.5 \log(0.5/0.6) = 0.00886$. These are different.

9. (6 points) In this question, we will get familiar with a fairly popular and useful function, called the log-sum-exp function. For $\mathbf{x} \in \mathbb{R}^n$, the log-sum-exp function is defined (quite literally) as:

$$f(\mathbf{x}) = \log\left(\sum_i e^{x_i}\right) \tag{10}$$

(a) Prove that $f(\mathbf{x})$ is differentiable everywhere in $\mathbb{R}^n$.

[*Hint:* Multivariable functions are differentiable if the partial derivatives exist and are continuous.]

We will find the partial derivative with respect to x_n.

$$\frac{\partial f}{\partial x_n} = \frac{e^{x_n}}{\sum_i e^{x_i}}$$

The partial derivative is defined everywhere. Therefore, the function f is everywhere differentiable.

(b) Prove that $f(\mathbf{x})$ is convex on $\mathbb{R}^n$.

[*Hint:* One approach is to use the second-order condition for convexity.]

We will compute the Hessian. We will first compute the partial derivative $\frac{\partial^2 f}{\partial x_n \partial x_m}$. First consider when $n \neq m$.

$$\begin{aligned}
\frac{\partial f}{\partial x_n} &= \frac{e^{x_n}}{\sum_i e^{x_i}} \\
\frac{\partial^2 f}{\partial x_n \partial x_m} &= \frac{\left(\sum_i e^{x_i}\right) 0 - e^{x_n} e^{x_m}}{\left(\sum_i e^{x_i}\right)^2} \\
\frac{\partial^2 f}{\partial x_n \partial x_m} &= -\frac{e^{x_n} e^{x_m}}{\left(\sum_i e^{x_i}\right)^2}
\end{aligned}$$

Now we will consider what happens when $n = m$.

$$\begin{aligned}
\frac{\partial f}{\partial x_n} &= \frac{e^{x_n}}{\sum_i e^{x_i}} \\
\frac{\partial^2 f}{\partial x_n \partial x_m} &= \frac{\left(\sum_i e^{x_i}\right) e^{x_n} - e^{2x_n}}{\left(\sum_i e^{x_i}\right)^2}
\end{aligned}$$

We will now show that the Hessian matrix of partial derivatives is diagonally dominant. Consider the magnitude of the ith diagonal element. That is just the absolute value of the partial derivative when $n = m$ i.e. $\frac{\left(\sum_i e^{x_i}\right) e^{x_n} - e^{2x_n}}{\left(\sum_i e^{x_i}\right)^2}$. Note that this is positive because all the exponentials are positive and e^{2x_n} is included in the summation. We just subtract it out later.

The sum of the non-diagonal elements is the sum of the absolute values of each $n \neq m$ term. The $n \neq m$ term is $-\frac{e^{x_n} e^{x_m}}{\left(\sum_i e^{x_i}\right)^2}$. Note that is is negative, so we can remove the negative sign. When we sum across all $n \neq m$, we get $\frac{\left(\sum_i e^{x_i}\right) e^{x_n} - e^{2x_n}}{\left(\sum_i e^{x_i}\right)^2}$.

Because the off-diagonal sum equals the diagonal sum, we get that the Hessian is diagonally dominant. Therwfore, it is positive semi-definite. Therefore, the log-sum-exp is convex.

(c) Show that $f(\mathbf{x})$ can be viewed as an approximation of the max function, bounded as follows:

$$\max\{x_1, \ldots, x_n\} \leq f(\mathbf{x}) \leq \max\{x_1, \ldots, x_n\} + \log(n) \tag{11}$$

$$f(\mathbf{x}) = \log\left(\sum_i e^{x_i}\right)$$

$$\log\left(\max_i e^{x_i}\right) \leq f(\mathbf{x}) \leq \log\left(n \cdot \max_i e^{x_i}\right)$$

$$\max\{x_1, \ldots, x_n\} \leq f(\mathbf{x}) \leq \max\{x_1, \ldots, x_n\} + \log(n)$$