CS7643: Deep Learning
Fall 2020
Problem Set 0 Solutions

Instructor: Dhruv Batra

TAs: Prabhav Chawla, Yihao Chen, Sameer Dharur, Hrishikesh Kale, Michael Piseno, Joanne Truong, Tianyu Zhan

Discussions: https://piazza.com/gatech/fall2020/cs48037643

Due: Thursday, August 20, 11:59pm ET

1 Multiple Choice Questions

1. (1 point) true/false 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

Explanation: Assuming a fair die, there is a 1/6 chance of landing on any number,

$$p(1) = \frac{1}{6}; \qquad p(\text{not } 1) = \frac{5}{6} \tag{2}$$

The expected outcome for a turn is

$$\$2\left(\frac{1}{6}\right) - \$\frac{1}{4}\left(\frac{5}{6}\right) = \$\frac{3}{24} \tag{3}$$

So, we will gain money. Thus, it is a good deal.

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

$$p(x) = \begin{cases} 8x & 0 \leq x \leq 1/2 \\ -2x + 1 & 1/2 \leq x \leq 1 \end{cases} \tag{4}$$

Which of the following statements are true about equation for the corresponding cumulative density function (CDF) $C(x)$?

[*Hint:* Recall that CDF is defined as $C(x) = Pr(X \leq x)$.]

- ● **$C(x) = 4x^2$ for $0 \le x \le 1/2$**
- ○ $C(x) = -x^2 + x - 1/4$ for $1/2 \le x \le 1$
- ○ All of the above
- ○ None of the above

Explanation:

$$
\begin{aligned}
C(x) &= \int_0^x p(z)dz && (5)\\
&= \begin{cases} \int_0^x 8z & 0 \le x \le 1/2 \\ \int_0^{1/2} 8zdz + \int_{1/2}^x (-2z+1)dz & 1/2 \le x \le 1 \end{cases} && (6)\\
&= \begin{cases} 4x^2 & 0 \le x \le 1/2 \\ 1 + -x^2 + x + 1/4 - 1/2 \le x \le 1 \end{cases} && (7)\\
&= \begin{cases} 4x^2 & 0 \le x \le 1/2 \\ -x^2 + x + 3/4 & 1/2 \le x \le 1 \end{cases} && (8)
\end{aligned}
$$

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

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (9)$$

Evaluate following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx \quad (10)$$

[*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

Explanation: For standard normal distribution, we have,

$$\int_{-\infty}^{\infty} p(x)dx = 1 \quad (11a)$$

$$\int_{-\infty}^{\infty} p(x)xdx = E(X) = 0 \quad (11b)$$

$$\int_{-\infty}^{\infty} p(x)x^2dx = E(X^2) = VAR(X) + [E(X)]^2 = 1 + 0 = 1 \quad (11c)$$

Hence,

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx = a + c \quad (12)$$

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

$$f(\mathbf{x}) = \sigma\left(\log\left(5\left(\max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6)\right)\right) + \frac{1}{2}\right) \tag{13}$$

where σ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{14}$$

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

○ $\begin{bmatrix} 0 \\ 0.031 \\ 0.026 \\ -0.013 \\ -0.062 \\ -0.062 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.157 \\ 0.131 \\ -0.065 \\ -0.314 \\ -0.314 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.358 \\ 0.269 \\ -0.215 \\ -0.846 \\ -0.846 \end{bmatrix}$ ● $\begin{bmatrix} 0 \\ 0.358 \\ 0.269 \\ -0.215 \\ -0.448 \\ -0.448 \end{bmatrix}$

Explanation: Let

$$z_1 = 5\max\{x_1, x_2\}\frac{x_3}{x_4} - 5(x_5 + x_6) \tag{15}$$

$$z_2 = \log(z_1) + \frac{1}{2} \tag{16}$$

$$z_3 = \sigma(z_2) \qquad (= f(x)) \tag{17}$$

$$\tag{18}$$

Then

$$\nabla_x f^T = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \frac{\partial f}{\partial x_4} \\ \frac{\partial f}{\partial x_5} \\ \frac{\partial f}{\partial x_6} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_3} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_4} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_5} \\ \frac{\partial z_3}{\partial z_2}\frac{\partial z_2}{\partial z_1}\frac{\partial z_1}{\partial x_6} \end{bmatrix} \tag{19}$$

Now compute the partials listed above:

$$\begin{bmatrix} \frac{\partial z_1}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_1}{\partial x_3} \\ \frac{\partial z_1}{\partial x_4} \\ \frac{\partial z_1}{\partial x_5} \\ \frac{\partial z_1}{\partial x_6} \end{bmatrix} = \begin{bmatrix} 5\frac{x_3}{x_4}[[x_1 > x_2]] \\ 5\frac{x_3}{x_4}[[x_2 > x_1]] \\ 5\frac{\max\{x_1,x_2\}}{x_4} \\ -5\frac{\max\{x_1,x_2\}x_3}{x_4^2} \\ -5 \\ -5 \end{bmatrix} \tag{20}$$

$$\frac{\partial z_2}{\partial z_1} = \frac{1}{z_1} \tag{21}$$

$$\frac{\partial z_3}{\partial z_2} = \frac{e^{-z_2}}{(1+e^{-z_2})^2} = \sigma(z_2)(1-\sigma(z_2)) = z_3(1-z_3) \tag{22}$$

All that's left is plugging in values:

$$\hat{z}_1 = 2 \tag{23}$$
$$\hat{z}_2 \approx 1.193 \tag{24}$$
$$f(\hat{x}) = \hat{z}_3 \approx 0.767 \tag{25}$$

And finally plug numbers into the gradient at $\hat{x}$. Start with the scalars

$$\frac{\partial z_2}{\partial z_1}|_{\hat{x}} = \frac{1}{2} = 0.5 \tag{26}$$

$$\frac{\partial z_3}{\partial z_2}|_{\hat{x}} = \hat{z}_3(1-\hat{z}_3) \approx 0.179 \tag{27}$$

$$\nabla_x f(x)^T|_{\hat{x}} \approx \begin{bmatrix} 0.179 \cdot 0.5 \frac{\partial z_1}{\partial x_1} \\ 0.179 \cdot 0.5 \frac{\partial z_1}{\partial x_2} \\ 0.179 \cdot 0.5 \frac{\partial z_1}{\partial x_3} \\ 0.179 \cdot 0.5 \frac{\partial z_1}{\partial x_4} \\ 0.179 \cdot 0.5 \frac{\partial z_1}{\partial x_5} \\ 0.179 \cdot 0.5 \frac{\partial z_1}{\partial x_6} \end{bmatrix} \approx \begin{bmatrix} 0.179 \cdot 0.5 \cdot 0 \\ 0.179 \cdot 0.5 \cdot 4 \\ 0.179 \cdot 0.5 \cdot 3 \\ 0.179 \cdot 0.5 \cdot -2.4 \\ 0.179 \cdot 0.5 \cdot -5 \\ 0.179 \cdot 0.5 \cdot -5 \end{bmatrix} \approx \begin{bmatrix} 0 \\ 0.358 \\ 0.269 \\ -0.215 \\ -0.448 \\ -0.448 \end{bmatrix} \tag{28}$$

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

   Explanation: The epigraph of A is not a convex set for $||x||_{\frac{1}{2}}$ or $||x||_{\frac{1}{2}}$, so the function is neither convex nor concave.

   B is concave. Affine functions are both concave and convex, and min of concave functions is concave.

   C is convex. For $f(z) = \log(1 + \exp z)$, Hessian $(f) = \frac{\exp z}{(1+\exp z)^2} > 0$ and $\mathbf{w}^T\mathbf{x}$ is linear in $\mathbf{w}$. Due to the property of composition with affine function, $f(\mathbf{w}^T\mathbf{x})$ is convex.

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

Explanation: Under I.I.D Gaussian noise, finding the maximum likelihood for Y is equivalent to finding the value $\hat{y}$ which minimizes the sum of least squares. That is to say: $\hat{y} = \text{argmin}_y \sum_{i=1}^{n}(y - x_i)^2$. The least squares solution also has a closed form solution: $\hat{y} = \frac{1}{n} \sum_{i=1}^{n} x_i$

2 Proofs

7. (3 points) Prove that

$$\log_e x \leq x - 1, \qquad \forall x > 0 \tag{29}$$

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

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

Solution: Define function $g(x)$ where:

$$g(x) = \log_e x - x + 1 \leq 0 \tag{30}$$

$g(x)$ is a strictly concave function ($g''(x) = -x^{-2} < 0$), therefore it is enough to show that the maximum is non-positive. At the maximum of $g(x)$ we must have $g'(x) = 0$. Therefore: $g'(x) = \frac{1}{x} - 1 = 0$. Solving this for x shows that the maximum of $g(x)$ is reached at $x = 1$. As the function value, there is $g(x = 1) = \log(1) - 1 + 1 = 0$. We know that $g(x) \leq 0$ for all $x \geq 0$.

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

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{31a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \ldots, k\} \tag{31b}$$

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

$$KL(p, q) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) \tag{32}$$

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.
Solution: Let $x = \frac{q_i}{p_i}$, we have,

$$\log\left(\frac{q_i}{p_i}\right) \leq \frac{q_i}{p_i} - 1 \tag{33}$$

$$
\begin{aligned}
KL(p,q) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) &= -\sum_{i=1}^{k} p_i \log\left(\frac{q_i}{p_i}\right) && \text{(34a)}\\
&\geq -\sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) && \text{(34b)}\\
&= -\sum_{i=1}^{k} (q_i - p_i) && \text{(34c)}\\
&= -\sum_{i=1}^{k} q_i + \sum_{i=1}^{k} p_i = 0 && \text{(34d)}
\end{aligned}
$$

(b) When is $KL(p,q) = 0$?
Solution: $KL(p,q) = 0$ if and only if $p_i = q_i \forall i$.

(c) Provide a counterexample to show that the KL divergence is not a symmetric function of its arguments: $KL(p,q) \neq KL(q,p)$
Solution: Let $p = [1/2, 1/2]$ and $q = [1/4, 3/4]$. Then

$$
KL(p,q) = \frac{1}{2}\log(\frac{1/2}{1/4}) + \frac{1}{2}\log(\frac{1/2}{3/4}) \approx 0.144 \tag{35}
$$

$$
KL(q,p) = \frac{1}{4}\log(\frac{1/4}{1/2}) + \frac{3}{4}\log(\frac{3/4}{1/2}) \approx 0.131 \tag{36}
$$

$$
\tag{37}
$$

So $KL(p,q) \neq KL(q,p)$.

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:

$$f(\mathbf{x}) = \log\left(\sum_{i=1}^{n} e^{x_i}\right) \tag{38}$$

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

Solution: To show that $f(\mathbf{x})$ is differentiable on $\mathbb{R}^n$, we must show that its gradient, $\nabla_{\mathbf{x}} f(\mathbf{x})$, always exists and is continuous at each $\mathbf{x} \in \mathbb{R}^n$.

Let $s_i = e^{x_i}$. The gradient of f is

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\mathbf{s}}{\mathbf{1}^T\mathbf{s}} \tag{39}$$

Notice that $\nabla_{\mathbf{x}} f(\mathbf{x})$ is the softmax function! Softmax is always defined and differentiable (thus continuous) in $\mathbb{R}^n$. See: https://en.wikipedia.org/wiki/Softmax_function.

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

Solution: Building off of $\nabla_{\mathbf{x}} f(\mathbf{x})$ computed above, we calculate the Hessian:

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{\mathbf{1}^T\mathbf{s}} \operatorname{diag}(\mathbf{s}) - \frac{1}{(\mathbf{1}^T\mathbf{s})^2}\mathbf{s}\mathbf{s}^T \tag{40}$$

Consider an arbitrary vector $\mathbf{y}$ with the same dimensionality as $\mathbf{x}$. We want to show:

$$\mathbf{y}^T\nabla^2_{\mathbf{x}} f(\mathbf{x})\mathbf{y} \geq 0 \tag{41}$$

Now, consider the vectors $\mathbf{t}$ and $\mathbf{u}$, where $t_i = \sqrt{s_i}$ and $u_i = y_i\sqrt{s_i}$. Then,

$$\begin{aligned}
\mathbf{y}^T\nabla^2_{\mathbf{x}} f(\mathbf{x})\mathbf{y} &= \mathbf{y}^T\left(\frac{1}{\mathbf{1}^T\mathbf{s}} \operatorname{diag}(\mathbf{s}) - \frac{1}{(\mathbf{1}^T\mathbf{s})^2}\mathbf{s}\mathbf{s}^T\right)\mathbf{y} && (42)\\
&= \frac{1}{\mathbf{t}^T\mathbf{t}}\mathbf{u}^T\mathbf{u} - \frac{1}{(\mathbf{t}^T\mathbf{t})^2}(\mathbf{t}^T\mathbf{u})^2 && (43)\\
& && (44)
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
(\mathbf{t}^T\mathbf{t})(\mathbf{u}^T\mathbf{u}) \geq (\mathbf{t}^T\mathbf{u})^2 && (45)\\
(\mathbf{t}^T\mathbf{t})(\mathbf{u}^T\mathbf{u}) - (\mathbf{t}^T\mathbf{u})^2 \geq 0 && (46)\\
\frac{1}{\mathbf{t}^T\mathbf{t}}\mathbf{u}^T\mathbf{u} - \frac{1}{(\mathbf{t}^T\mathbf{t})^2}(\mathbf{t}^T\mathbf{u})^2 \geq 0 && (47)
\end{aligned}$$

This means the Hessian is positive semi-definite, so $f(\mathbf{x})$ is convex.

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

$$\max\{x_1, \ldots, x_n\} \leq f(\mathbf{x}) \leq \max\{x_1, \ldots, x_n\} + \log(n) \tag{48}$$

Solution: For any $\mathbf{x} \in \mathbb{R}^n$, we know that $\sum_{i=1}^{n} x_i \leq n \cdot \max\{x_1, \ldots, x_n\}$ since a sum of n numbers is at most n times its maximum term. Additionally, $\sum_{i=1}^{n} e^{x_i} \geq e^{x_i}\ \forall x_i$ since $\{e^{x_1}, \ldots, e^{x_n}\}$ is a set of positive numbers.

Putting these two inequalities together, and the fact that $log(\cdot)$ is monotonically increasing in $\mathbb{R}^{+}$:

$$\begin{aligned}
\max\{x_1, \ldots, x_n\} &= \log(e^{\max\{x_1,\ldots,x_n\}}) \\
&\leq \log(\sum_{i=1}^{n} e^{x_i}) \\
&\leq \log(n \cdot e^{\max\{x_1,\ldots,x_n\}}) \\
&= \max\{x_1, \ldots, x_n\} + \log(n)
\end{aligned}$$

Hence proved.