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:

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

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 the probability density function:

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

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 $C(x) = Pr(X \leq x)$.]

● $C(x) = 2x^2$ for $0 \leq x \leq 1/2$
○ $C(x) = -2x^2 + 4x - 3/2$ for $1/2 \leq x \leq 1$
○ All of the above
○ None of the above

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

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

Evaluate 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 $\hat{\mathbf{x}} = (5, -1, 6, 12, 7, -5)$.

○ $\begin{bmatrix} 0.157 \\ 0.0 \\ 0.131 \\ -0.065 \\ -0.846 \\ -0.846 \end{bmatrix}$ ● $\begin{bmatrix} 0.157 \\ 0 \\ 0.131 \\ -0.065 \\ -0.314 \\ -0.314 \end{bmatrix}$ ○ $\begin{bmatrix} 0.031 \\ 0 \\ 0.026 \\ -0.013 \\ -0.062 \\ -0.062 \end{bmatrix}$ ○ $\begin{bmatrix} -0.468 \\ 0 \\ -0.390 \\ 0.195 \\ 0.937 \\ 0.937 \end{bmatrix}$

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

○ $||\mathbf{x}||_{\frac{1}{2}}$

○ $\min_i \mathbf{a}_i^T \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$

○ $\log(1 + \exp(\mathbf{w}^T \mathbf{x}_i))$ 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 $x_1 ... x_n$ 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

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

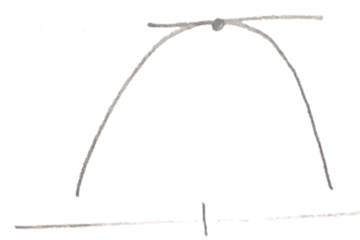
with equality if and only if $x = 1$.

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

Let $f(x) = \ln x - (x-1)$

then $f'(x) = \frac{1}{x} - 1$

and $f''(x) = (-1)x^{-2} = \frac{-1}{x^2}$



Note: $f''(x) < 0 \quad \forall x > 0$

so, $f(x)$ is concave down.

Set $f'(x) = 0, \implies x = 1$

@ $x = 1$, f has maximum value

i.e. $f(1) = \ln 1 - (1-1) = 0$, is the max value.

So $f(x) \leq 0$ i.e. $\ln x - (x-1) \leq 0$

$\implies \ln x \leq x - 1$

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.

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

$= \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right)^{-1} = \sum_{i=1}^{k} p_i \log\left(\frac{q_i}{p_i}\right)$

$\leq \sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right)$ (★★) (Q7 identity)

$\leq \sum_{i=1}^{k} q_i - p_i = \sum_{i=1}^{k} q_i - \sum_{i=1}^{k} p_i = 0$ (8a)

i.e. $-KL(p,q) \leq 0$ or $KL(p,q) \geq 0$, or $\underline{KL(p,q) \text{ is nonnegative}}$

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

Assuming $f(p,q) = -KL(p,q)$,

from step (**) of proof in Q8a, we had

$$f(p,q) \leq \sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right)$$

since $KL(p,q) \geq 0$, then $f(p,q) \leq 0$.

and $f(p,q)$ is at a maximum, when $KL(p,q) = 0$.

i.e. $KL(p,q) = 0 \rightarrow f(p,q) = \text{max value} = \sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) = 0$.

this equality, $\sum_{i=0}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) = 0$ holds only when:

$p_i = q_i \quad \forall i$

So, $KL(p,q) = 0$ when $p_i = q_i \ \forall i$

i.e. when $p = q$.

(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 the following counterexample

$$p_0 = \frac{1}{3}, \quad p_1 = \frac{2}{3} \quad \text{and} \quad q_0 = 1/2, \quad q_1 = 1/2$$

$$KL(p,q) = \sum_{i=0}^{1} p_i \log \frac{p_i}{q_i} = \frac{1}{3} \log \left(\frac{\frac{1}{3}}{1/2}\right) + \frac{2}{3} \log \left(\frac{\frac{2}{3}}{\frac{1}{2}}\right)$$

$$= \frac{1}{3} \log \frac{2}{3} + \frac{2}{3} \log \frac{4}{3}$$

$$KL(q,p) = \sum_{i=0}^{1} q_i \log \left(\frac{q_i}{p_i}\right) = \frac{1}{2} \log \left(\frac{\frac{1}{2}}{\frac{1}{3}}\right) + \frac{1}{2} \log \left(\frac{\frac{1}{2}}{\frac{2}{3}}\right)$$

$$= \frac{1}{2} \log \frac{3}{2} + \frac{1}{2} \log \frac{3}{4}$$

$$= -\frac{1}{2} \log \frac{2}{3} - \frac{1}{2} \log \frac{4}{3}$$

So, $\boxed{KL(p,q) \neq KL(q,p)}$

9. (6 points) In this question, you will prove that cross-entropy loss for a softmax classifier is convex in the model parameters, thus gradient descent is guaranteed to find the optimal parameters. Formally, consider a single training example $(\mathbf{x}, y)$. Simplifying the notation slightly from the implementation writeup, let

$$\mathbf{z} = W\mathbf{x} + \mathbf{b}, \tag{10}$$

$$p_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{11}$$

$$L(W) = -\log(p_y) \tag{12}$$

Prove that $L(\cdot)$ is convex in W.

[*Hint:* One way of solving this problem is "brute force" with first principles and Hessians. There are more elegant solutions.]

$$L(w) = -\log_e\left(\frac{e^{Wy+b}}{\sum_k e^{Wk+b}}\right) = \log_e\left(\frac{\sum_k e^{Wk+b}}{e^{Wy+b}}\right)$$

$$L'(w) = \frac{e^{Wy+b}}{\sum_k e^{Wk+b}}\left(\frac{(e^{Wy+b})\sum_k ke^{Wk+b} - (\sum_k e^{Wk+b})\,y\,e^{Wy+b}}{(e^{Wy+b})^2}\right)$$

$$= \frac{\cancel{e^{Wy+b}}}{\sum_k e^{Wk+b}} \circ \frac{1}{\cancel{(e^{Wy+b})^2}} \cdot \cancel{e^{Wy+b}}\left(\sum_k ke^{Wk+b} - y\sum_k e^{Wk+b}\right)$$

$$= \frac{\sum_k ke^{Wk+b}}{\sum_k e^{Wk+b}} - y\,\cancel{\frac{\sum_k e^{Wk+b}}{\sum_k e^{Wk+b}}} = \frac{\sum_k ke^{Wk+b}}{\sum_k e^{Wk+b}}$$

Sign of $L''(w)$ determined by num, since den is positive.

$$L''(w) = \frac{\sum_k e^{Wk+b}\left(\sum_k k^2 e^{Wk+b}\right) - \sum_k ke^{Wk+b}\left(\sum_k ke^{Wk+b}\right)}{\left(\sum_k e^{Wk+b}\right)^2} = \frac{\text{num}}{\text{den}}$$

$$\text{num} = \sum_{k_1=0}^{n}\sum_{k_2=0}^{k_1} e^{Wk_1+b}(k_1-k_2)^2\left(e^{W(k_1-k_2)+b}\right) - \sum_{k_1=0}^{n}\sum_{k_2=0}^{k_1} k_2 e^{Wk_1+b}(k_1-k_2)\,e^{W(k_1-k_2)+b}$$

$$= \sum\sum \left(e^{Wk_1+b}\right)\left(e^{W(k_1-k_2)+b}\right)\left((k_1-k_2)^2 - (k_1-k_2)\right)$$

positive positive

$$= (k_1-k_2)(k_1-k_2-1), \text{ which is non-negative.}$$

∴ the numerator, thus $L''(w)$ is positive, which proves that $L(\cdot)$ is convex in W.