1 Probability and Statistics

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:

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

(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? Are we expected to make money?

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

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

(2)

What is the equation for the corresponding cumulative density function (cdf) \(C(x)\)?

[\text{Hint: Recall that CDF is defined as } C(x) = Pr(X \leq x).]
3. (1 point) Recall that the variance of a random variable is defined as $\text{Var}[X] = E[(X - \mu)^2]$, where $\mu = E[X]$. Use the properties of expectation to show that we can rewrite the variance of a random variable $X$ as

$$\text{Var}[X] = E[X^2] - (E[X])^2 \tag{3}$$

4. (1 point) A random variable $x$ in standard normal distribution has following probability density

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{4}$$

Evaluate following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx \tag{5}$$

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

2 Proving Stuff

5. (2 points) Prove that

$$\log_e x \leq x - 1, \quad \forall x > 0 \tag{6}$$

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

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

6. (3 points) Consider two discrete probability distributions $p$ and $q$ over $k$ outcomes:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{7a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \ldots, k\} \tag{7b}$$

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

$$KL(p, q) = \sum_{i=1}^{k} p_i \log \left( \frac{p_i}{q_i} \right) \tag{8}$$

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.

(a) Using the results from Q5, show that $KL(p, q)$ is always positive.

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

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

[Hint: This question doesn’t require you to know anything more than the definition of $KL(p, q)$ and the identity in Q5]
3 Calculus

7. (3 points) Consider the following function of $x = (x_1, x_2, x_3, x_4, x_5, x_6)$:

$$f(x) = \sigma \left( \log \left( 5 \left( \max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6) \right) \right) + \frac{1}{2} \right)$$ (9)

where $\sigma$ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$ (10)

Evaluate $f(\cdot)$ at $\mathbf{x} = (5, -1, 6, 12, 7, -5)$. Then, compute the gradient $\nabla_x f(\cdot)$ and evaluate it at the same point.

4 Softmax Classifier

8. (5 points) Implement a Softmax classifier (from scratch, no ML libraries allowed), and train it (via SGD) on CIFAR-10: cc.gatech.edu/classes/AY2018/cs7643_fall/hw0-q8/.

9. (3 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 $(x, y)$. Simplifying the notation slightly from the implementation writeup, let

$$z = Wx + b,$$ (11)

$$p_j = \frac{e^{z_j}}{\sum_k e^{z_k}},$$ (12)

$$L(W) = -\log (p_y)$$ (13)

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.]