STATS 700, Fall 2024
LLMs and Transformers
HW 2

Ambuj Tewari

Due September 10

You are allowed to use LLMs provided you declare precisely how used them (but please be aware that they can hallucinate). You cannot search online or ask anyone. Discussing problems with other students in the course is allowed but all solutions should be your own work. If two students submit verbatim copies of any solution, it will be treated as a violation of academic integrity. Submit your solutions as a typeset PDF file (no handwritten scans please) on Canvas.

Corrections:

- Problem 4: corrected "entropy" to "negative entropy"

1 Bregman Divergence

Let $f : S \to \mathbb{R}$ be a twice differentiable convex function defined on some convex set $S \subseteq \mathbb{R}^d$. Its Bregman divergence is defined as:

$$D_f(x\|y) = f(x) - f(y) - \nabla f(y)^\top (x-y) \ .$$

Show that Bregman divergence is always non-negative (using the fact that f is convex and differentiable).

2 Bregman Divergence and Strong Convexity

If f is strongly convex, i.e., $\nabla^2 f \succeq \mu I$ (we say $A \succeq B$ for symmetric matrices if $A - B$ is positive semidefinite), then show that $D_f(x\|y) \geq \frac{\mu}{2}\|x-y\|_2^2$ where $\|x\|_2 = \sqrt{x^\top x}$.

3 Squared Euclidean Norm is a Bregman Divergence

Show that when $f(x) = \|x\|_2^2$, the Bregman divergence D_f is $\|x-y\|_2^2$.

4 Entropy is a Bregman Divergence

Show that when f is the negative entropy function, the Bregman divergence D_f is relative entropy.

5 Entropy of Powers of a Random Variable

Let k be a positive integer and X be a real-valued random variable taking values in a finite set. What is the best general relationship that you can establish between the entropy of X and the entropy of X^k?

6 Entropy of Random Graphs I

The classic Erdos-Renyi random (undirected) graph model on n vertices comes in two varieties. In the first model, we choose each possible edge independently with probability p. This is called the $G(n,p)$ model. Compute the entropy of a random graph drawn according to the $G(n,p)$ model.

7 Entropy of Random Graphs II

In the second kind of Erdos-Renyi model, we choose m edges uniformly at random from the total possible $\binom{n}{2}$ edges. This is called the $G(n,m)$ model. Compute the entropy of an Erdos-Renyi random graph under this model. Show that this entropy is close to the one in the previous problem when $m = \binom{n}{2}p$ and n is large.

8 Entropy of a Discrete Dynamical System I

Let X_0 be the outcome of the roll of a fair die recorded modulo 6 (so 1 through 5 are recorded as is, 6 is recorded as 0). For $n \geq 0$, define $X_{n+1} = f(X_n)$ where f is defined as $f(i) = 5i \bmod 6$. Compute the entropy H_n of X_n for all $n \geq 0$. What is $\lim_{n\to\infty} H_n$?

9 Entropy of a Discrete Dynamical System II

Let X_0 be the outcome of the roll of a fair die recorded modulo 6 (so 1 through 5 are recorded as is, 6 is recorded as 0). For $n \geq 0$, define $X_{n+1} = f(X_n)$ but, different from the previous problem, f is now defined as $f(i) = 3i \bmod 6$. Compute the entropy H_n of X_n for all $n \geq 0$. What is $\lim_{n\to\infty} H_n$? Are your answers same or different compared to the previous problem? Explain why.

10 Entropy of the Zipf-Mandelbrot Distribution

George Kingsley Zipf was an American linguist and Harvard professor who pioneered a statistical approach to linguistics. Zipf discovered an empirical law for the word frequencies in a language. He discovered that if you rank words in decreasing order of their frequency then the frequency of word with rank r roughly scales as

$$f(r) \approx \frac{1}{r^s}$$

for a value of s very close to 1. Benoit Mandelbrot was a French-American mathematician who spent most of his career at IBM but moved to Yale towards the end of his career. Wikipedia says that he was the oldest professor in Yale's history to receive tenure! Mandelbrot slightly modified Zipf's law and today the distribution with the following probability mass function, parameterized by $\theta = (N, q, s)$, is called a Zipf-Mandelbrot distribution:

$$f(r) = \frac{1}{Z_\theta} \frac{1}{(r+q)^s}, \; r \in \{1, \ldots, N\} \; .$$

Here N is a positive integer, q, s are real numbers with $q \geq 0$ and $s > 0$, and Z_θ is a normalizing constant. Show that the entropy of this distribution is

$$H(\theta) = \frac{s}{Z_\theta} \sum_{r=1}^{N} \frac{\log(r+q)}{(r+q)^s} + \log(Z_\theta) \; .$$