STATS 700, Fall 2024
LLMs and Transformers
HW 1

Ambuj Tewari

August 2024

At first, solve all problems **without** using an LLM. Write down your answers. Then consult your favorite LLM (chatGPT, Gemini, Claude, etc.) and solve all the problems again. Submit both solutions (without and with LLM help) as a single typeset PDF file (no handwritten scans please) on Canvas. In the LLM version, please include an appropriate description of your prompts. Please also comment specifically on the strengths and weaknesses of the LLM's responses. Did it help? Did it make the task harder? Did it hallucinate? Did it surprise you? Feel free to include any other personal reflections you find worth sharing with others.

1 Convexity

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be convex iff

$$\forall x, y \in \mathbb{R}, \lambda \in [0,1],\ f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Suppose a function g is twice differentiable everywhere on $\mathbb{R}$ and satisfies $g''(x) \geq 0$ for all $x \in \mathbb{R}$. Prove that g is convex.

2 Optimization

Suppose $n > 1$ and you are given a probability distribution $(p_1, \ldots, p_n)$ such that $p_i > 0$ for all $i \in \{1, \ldots, n\}$. Find the n values of L_i that minimize

$$\sum_{i=1}^{n} p_i L_i$$

subject to the constraint:

$$\sum_{i=1}^{n} e^{-L_i} \leq 1.$$

Also compute the value of the minimum.

3 Generating Random Variables

Assume that the only source of randomness that you have access to is a fair coin. How would you simulate a fair 6-sided die given access to these random (fair) coin flips? How many coin flips does your method take, on average, to simulate one roll of the die? Do you think your method is optimal (in the sense of requiring the fewest coin flips per roll of dice, on average)?

4 Linear Representability

A Boolean function $f : \{0,1\} \times \{0,1\} \to \{0,1\}$ is said to be linearly representable if there are real-valued weights w_1, w_2 and a real-valued bias b such that

$$w_1 x_1 + w_2 x_2 < b \text{ if } f(x_1, x_2) = 0$$

$$w_1 x_1 + w_2 x_2 > b \text{ if } f(x_1, x_2) = 1$$

for all bits $x_1, x_2 \in \{0,1\}$. Prove that the XOR of two bits is not linearly representable. Recall that XOR of two bits is one iff exactly one of the bits is one.

5 Character Frequencies in English

The most common letter in the English language is "e". Write a paragraph on any topic that does not use the letter "e" at all. Your paragraph must be at least 100 words long.

6 Calculus

A two layer neural network is a function $f : \mathbb{R}^d \to \mathbb{R}$ parameterized using weights $w_i \in \mathbb{R}^d$ and $a_i \in \mathbb{R}$ as follows:

$$f(x) = \sum_{i=1}^{n} a_i \sigma(w_i^\top x) \ .$$

For a fixed input x, compute the derivatives of the network output $f(x)$ with respect to the parameters w_i and a_i. Assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is fixed and is differentiable everywhere.

7 Entropy

Using the approximation $x! \approx (x/e)^x$, show that

$$\binom{n}{m} \approx 2^{nH_2(m/n)}$$

where $H_2(p) := p \log_2 p + (1-p) \log_2 (1-p)$.

8 Rademacher Expectations

A ± 1-valued random variable with equal probability of being ± 1 is called a Rademacher random variable. Let ϵ_i be iid Rademacher random variables and let X_i be arbitrary iid random variables (and also independent of the ϵ_i's) taking values in some set $\mathcal{X}$. Show that, for any fixed real valued function $f : \mathcal{X} \to \mathbb{R}$,

$$E\left[\sum_i \epsilon_i f(X_i)\right] = 0 \ .$$

Also show that for any class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, we have

$$E\left[\sup_{f \in \mathcal{F}} \sum_i \epsilon_i f(X_i)\right] \geq 0 \ .$$

9 Lipschitz Constants and Matrix Norms

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be Lipschitz with constant L (or L-Lipschitz) if for all $x, y \in \mathbb{R}$:

$$|f(x) - f(y)| \leq L|x - y| \ .$$

Prove that, if f is differentiable, and $\forall x, |f'(x)| \leq L$ then f is L-Lipschitz.

Let us now pursue a generalization of this result for a vector-to-vector function $F : \mathbb{R}^d \to \mathbb{R}^d$. The notion of Lipschitz-ness now depends on which norms we use on the left and right hand sides of the inequality above. Recall the ℓ_1-norm

$$\|x\|_1 := \sum_i |x_i| \ ,$$

and the ℓ_∞-norm:

$$\|x\|_\infty := \max_i |x_i| \ .$$

For this problem, let us say that F is Lipschitz with constant L if for all $x, y \in \mathbb{R}^d$:

$$\|f(x) - f(y)\|_1 \leq L\|x - y\|_\infty \ .$$

Prove that if the Jacobian matrix (the $d \times d$ matrix of all partial derivatives) of F satisfies:

$$\forall x, \|\nabla F(x)\|_{1,1} \leq L$$

then F is L-Lipschitz. Here, the matrix norm $\|\cdot\|_{1,1}$ is defined simply as:

$$\|M\|_{1,1} := \sum_{i,j} |M_{i,j}| \ .$$

10 Softmax is Lipschitz

A key function appearing in the attention mechanism underlying transformers is the softmax function $F : \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$[F(x)]_j = \frac{e^{x_j}}{\sum_i e^{x_i}} \ .$$

Prove that softmax is 2-Lipschitz in the sense defined in the previous problem.