Assignment 1: Foundations and efficient computation

42 points + 2 optional points

You may use a result requested in an earlier part even if you have not proved it.

Notation. Logarithms are natural and vectors are column vectors. For a matrix or vector A, $A^\top$ is its transpose and $\nabla_A L$ is the gradient of a real-valued function L, with entries $\partial L/\partial A_{ij}$ for a matrix and $\partial L/\partial A_i$ for a vector.

1 Warmup: FLOPs and gradients (14 points)

This problem reviews calculations used throughout the course.

(a) **(6 points)** Let $Q, K \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{n\times p}$, and let softmax act on each row: $[\mathrm{softmax}(S)]_{ij} = e^{S_{ij}} / \sum_{k=1}^{n} e^{S_{ik}}$ for $S \in \mathbb{R}^{n\times n}$. Count each scalar addition, multiplication, division, and exponentiation as one operation, and treat transposes as free. Count the operations needed to evaluate each expression below on its own, multiplying matrices by the row–column rule and, within softmax, computing each exponential and each row sum once:

$$\mathrm{softmax}(QK^\top)V, \qquad (QK^\top)V, \qquad Q(K^\top V).$$

State the dimensions of each output.

(b) **(2 points)** Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, and $Z = AB$. A differentiable loss L depends on A, B through Z. Given $G = \nabla_Z L$, compute $\nabla_A L$ and $\nabla_B L$ in terms of A, B, and G.

(c) For $x \in \mathbb{R}^n$, set $p_i = e^{x_i} / \sum_j e^{x_j}$. Let L be a differentiable function of p, with $g = \nabla_p L$.

(i) **(2 points)** Compute $\partial p_i/\partial x_j$ for every i, j.

(ii) **(2 points)** Compute $\nabla_x L$ in terms of p and g.

(iii) **(2 points)** Let $y_i \geq 0$, $\sum_i y_i = 1$, and $L = -\sum_i y_i \log p_i$. Use (ii) to compute $\nabla_x L$ in terms of p and y.

2 Scaling laws with inference costs (13 points)

A scaling law predicts a model's performance, for example its loss, from aggregate properties of the model and its training data, such as the number of parameters and the number of training tokens. Fitted on small runs, such laws help choose the model size and training duration that reach a target loss with the least training compute. However, a model served

repeatedly also spends compute on every inference token, and over its lifetime this can exceed the training cost. In this problem, we include inference compute and study how it changes the compute-optimal model size and training duration.

Suppose that a model with $N > 0$ parameters trained on $D > 0$ tokens has excess loss $L(N, D) = aN^{-\alpha} + bD^{-\beta}$ over a fixed baseline, where $a, b, \alpha, \beta > 0$ and N, D can be any positive reals. Training costs $6N$ operations per token and inference costs $2N$ per token. Given a target $\ell > 0$ and $T \geq 0$ expected inference tokens, we minimize total compute:

$$C_T(N, D) = 6ND + 2NT \qquad \text{subject to} \qquad L(N, D) \leq \ell.$$

(a) **(3 points)** Let $N_{\min} = (a/\ell)^{1/\alpha}$. For $N > N_{\min}$, find the least feasible D, denoted $D(N)$. Prove that the problem reduces to minimizing $C_T(N, D(N))$ over $N > N_{\min}$.

(b) **(4 points)** For $T = 0$, find the unique minimizer (N_0, D_0) and the ratio $aN_0^{-\alpha}/(bD_0^{-\beta})$.

(c) **(3 points)** For $T \geq 0$, let (N_T, D_T) be a minimizer (assume one exists). For $T_2 > T_1$, prove $N_{T_2} \leq N_{T_1}$ and $D_{T_2} \geq D_{T_1}$. (*Hint:* Compare the two pairs under each demand, then add the optimality inequalities.)

(d) **(3 points)** Let $N_s \in (N_{\min}, N_0)$ and $D_s = D(N_s)$. Find the demand T_{break} at which (N_s, D_s) and (N_0, D_0) cost the same. Prove $T_{\text{break}} > 0$. Which is cheaper on each side?

3 Entropy and KL divergence from axioms (15 points)

Entropy and KL divergence appear throughout machine learning. For instance, the next-token log loss is closely related to both (cf. part (d)). We derive entropy from natural axioms and study KL divergence through a similar characterization.

Write $\Delta_n = \{p \in \mathbb{R}^n : p_i \geq 0, \sum_i p_i = 1\}$ for the probability simplex, $u_n = (1/n, \ldots, 1/n)$ for the uniform distribution, and (x, y) for the concatenation of x and y.

Shannon entropy. We seek a measure $H_n(p)$ of the "uncertainty" of a distribution $p \in \Delta_n$ that satisfies the following axioms:

(E1) **Continuity and invariance.** For every n, H_n is continuous and permutation invariant.

(E2) **Monotonicity.** The uncertainty $H_n(u_n)$ is nondecreasing in n.

(E3) **Additivity.** For any positive integers m, n, distributions $p \in \Delta_m$, $r \in \Delta_n$, and $t \in [0, 1]$,

$$H_{m+n}(tp, (1-t)r) = H_2(t, 1-t) + tH_m(p) + (1-t)H_n(r)\,. \tag{1}$$

A term with zero weight contributes zero.

We fix the unit of measurement by setting $H_2(u_2) = \log 2$.

To see why these axioms are natural, think of $H_n(p)$ as the uncertainty resolved by observing a draw from p. It should change little when p changes little, and it should not depend on how outcomes are labeled. More equally likely outcomes should not make a draw less uncertain. Finally, we can observe a draw in two stages: first its group, chosen with probabilities t and $1 - t$, then the outcome within that group. Additivity says that the total uncertainty is that of the first stage plus the expected uncertainty of the second.

(a) Let $A(n) = H_n(u_n)$.

(i) **(3 points)** Prove $A(mn) = A(m) + A(n)$ for positive integers m, n. (*Hint:* Extend additivity to any finite number of groups by induction.)

(ii) **(1 point)** Deduce $A(2^k) = k \log 2$ for integers $k \geq 0$.

(iii) **(3 points)** Use monotonicity to prove $A(n) = \log n$ for every positive integer n. (*Hint:* Compare n^k with consecutive powers of 2, and let $k \to \infty$.)

(b) **(4 points)** Suppose $p_i = a_i / N$, where a_i are positive integers and $\sum_i a_i = N$. Derive $H_n(p)$ by splitting outcome i into a_i equally likely outcomes.

Distributions with positive rational coordinates are dense in Δ_n, so continuity extends part (b) to $H_n(p) = -\sum_i p_i \log p_i$ for all $p \in \Delta_n$, with $0 \log 0 = 0$. This is the unique function satisfying the axioms and the normalization.

Kullback–Leibler divergence. Analogous axioms characterize a measure of mismatch between a data distribution $p \in \Delta_n$ and a model distribution $q \in \Delta_n$ with positive coordinates: the Kullback–Leibler divergence $\mathrm{KL}(p\|q) = \sum_i p_i \log(p_i/q_i)$, where terms with $p_i = 0$ contribute zero.

(c) **(2 optional points)** Prove $\mathrm{KL}(p\|q) \geq 0$, with equality exactly when $p = q$. (*Hint:* Use $\log x \leq x - 1$, with equality exactly at $x = 1$.)

(d) **(4 points)** Let P and Q be distributions on sequences $X = (X_1, \ldots, X_T) \in \mathcal{V}^T$ over a finite vocabulary $\mathcal{V}$, with Q positive on every sequence. Let $p_t(\cdot \mid x_{<t})$ and $q_t(\cdot \mid x_{<t})$ be their next-token distributions, where p_t is arbitrary at prefixes that have probability zero under P. For $H(P) = -\sum_x P(x) \log P(x)$, prove

$$\mathbb{E}_{X\sim P}[-\log Q(X)] - H(P) = \mathrm{KL}(P\|Q) = \sum_{t=1}^{T} \mathbb{E}_{X_{<t}\sim P}\left[\mathrm{KL}\big(p_t(\cdot \mid X_{<t})\|q_t(\cdot \mid X_{<t})\big)\right].$$

Deduce when the expected log loss $\mathbb{E}_{X\sim P}[-\log Q(X)]$ equals $H(P)$.