

# Assignment 1: Foundations and efficient computation

42 points + 2 optional points [With solutions](#)

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:  $[\text{softmax}(S)]_{ij} = e^{S_{ij}} / \sum_{k=1}^n e^{S_{ik}}$  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:

$$\text{softmax}(QK^\top)V, \quad (QK^\top)V, \quad Q(K^\top V).$$

State the dimensions of each output.

**Solution.** All three outputs are  $n \times p$  matrices. The intermediate matrices  $QK^\top$  and  $\text{softmax}(QK^\top)$  are  $n \times n$ , and  $K^\top V$  is  $d \times p$ .

Multiplying an  $a \times b$  matrix by a  $b \times c$  matrix costs  $ac(2b - 1)$  operations. Each of the  $ac$  entries needs  $b$  multiplications and  $b - 1$  additions. Each row of the softmax needs  $n$  exponentials,  $n - 1$  additions, and  $n$  divisions, or  $3n - 1$  operations. The three expressions therefore cost

$$\begin{aligned} n^2(2d - 1) + n(3n - 1) + np(2n - 1), \\ n^2(2d - 1) + np(2n - 1), \\ dp(2n - 1) + np(2d - 1). \end{aligned}$$

Here  $\Theta(f)$  denotes a quantity bounded above and below by positive constant multiples of  $f$ . The first two costs are  $\Theta(n^2(d + p))$ , and the third is  $\Theta(ndp)$ .

- (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  $G$ .

**Solution.** Since  $Z_{ij} = \sum_k A_{ik} B_{kj}$ , the chain rule gives

$$\frac{\partial L}{\partial A_{ik}} = \sum_j G_{ij} B_{kj}, \quad \frac{\partial L}{\partial B_{kj}} = \sum_i A_{ik} G_{ij}.$$

Therefore  $\nabla_A L = GB^\top \in \mathbb{R}^{m \times n}$  and  $\nabla_B L = A^\top G \in \mathbb{R}^{n \times p}$ .

These matrix forms are much cleaner than the entrywise formulas and are worth remembering. The gradient for each factor multiplies  $G$  by the transpose of the other factor, on the side where that factor appears.

(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  $i, j$ .

**Solution.** Write  $\mathbb{1}_{i=j}$  for the indicator that equals one when  $i = j$  and zero otherwise. Let  $s = \sum_k e^{x_k}$ , so that  $p_i = e^{x_i} / s$ . The quotient rule gives

$$\frac{\partial p_i}{\partial x_j} = \frac{\mathbb{1}_{i=j} e^{x_i} s - e^{x_i} e^{x_j}}{s^2} = p_i (\mathbb{1}_{i=j} - p_j).$$

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

**Solution.** By part (i) and the chain rule,

$$\frac{\partial L}{\partial x_j} = \sum_i g_i p_i (\mathbb{1}_{i=j} - p_j) = p_j \left( g_j - \sum_i p_i g_i \right).$$

Let  $\mathbf{1}_n$  denote the vector of  $n$  ones and  $\odot$  denote entrywise multiplication, so  $(u \odot v)_i = u_i v_i$ . Also,  $p^\top g = \sum_i p_i g_i$ . In vector form,  $\nabla_x L = p \odot (g - (p^\top g) \mathbf{1}_n)$ .

(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  $y$ .

**Solution.** Here  $g_i = -y_i / p_i$ , so  $p^\top g = -\sum_i y_i = -1$ . Part (ii) then gives  $\nabla_x L = p \odot (g + \mathbf{1}_n) = p - y$ .

## 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:

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

- (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  $N > N_{\min}$ .

**Solution.** Since  $bD^{-\beta} > 0$ , every feasible pair has  $aN^{-\alpha} < \ell$ , that is,  $N > N_{\min}$ . For such  $N$ , the constraint holds exactly when  $bD^{-\beta} \leq \ell - aN^{-\alpha}$ . So the least feasible training count is

$$D(N) = \left( \frac{b}{\ell - aN^{-\alpha}} \right)^{1/\beta}.$$

At fixed  $N$ , the cost strictly increases in  $D$ . So an optimal pair uses  $D = D(N)$ , and the problem reduces to minimizing  $6ND(N) + 2NT$  over  $N > N_{\min}$ .

- (b) **(4 points)** For  $T = 0$ , find the unique minimizer  $(N_0, D_0)$  of  $C_0(N, D)$  subject to  $L(N, D) \leq \ell$ , and also compute the ratio  $aN_0^{-\alpha} / (bD_0^{-\beta})$ .

**Solution.** Differentiating  $aN^{-\alpha} + bD(N)^{-\beta} = \ell$  gives

$$D'(N) = -\frac{\alpha a}{\beta b} N^{-\alpha-1} D(N)^{\beta+1}.$$

Hence the derivative  $D(N) + ND'(N)$  of  $ND(N)$  has the same sign as

$$1 - \frac{\alpha a N^{-\alpha}}{\beta b D(N)^{-\beta}} = 1 - \frac{\alpha a N^{-\alpha}}{\beta(\ell - aN^{-\alpha})}.$$

This expression increases strictly from  $-\infty$  to 1 on  $(N_{\min}, \infty)$ . Setting it to zero gives

$$N_0 = \left( \frac{a(\alpha + \beta)}{\beta \ell} \right)^{1/\alpha}, \quad D_0 = \left( \frac{b(\alpha + \beta)}{\alpha \ell} \right)^{1/\beta}.$$

The derivative is negative for  $N < N_0$  and positive for  $N > N_0$ , so  $(N_0, D_0)$  is the unique minimizer. At this optimum, the ratio of the two loss terms is

$$\frac{aN_0^{-\alpha}}{bD_0^{-\beta}} = \frac{\beta}{\alpha}.$$

- (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.)

**Solution.** Write  $(N_i, D_i) = (N_{T_i}, D_{T_i})$  for  $i = 1, 2$ . Optimality at the two demands gives

$$\begin{aligned} 6N_1D_1 + 2T_1N_1 &\leq 6N_2D_2 + 2T_1N_2, \\ 6N_2D_2 + 2T_2N_2 &\leq 6N_1D_1 + 2T_2N_1. \end{aligned}$$

Adding and canceling yields  $2(T_2 - T_1)(N_2 - N_1) \leq 0$ , so  $N_2 \leq N_1$ . By part (a),  $D_i = D(N_i)$ . The formula for  $D(N)$  is decreasing in  $N$ , hence  $D_2 \geq D_1$ .

- (d) **(3 points)** Let  $N_s \in (N_{\min}, N_0)$  and  $D_s = D(N_s)$ . Find the demand  $T_{\text{break}}$  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?

**Solution.** Since  $(N_0, D_0)$  is the unique minimizer of training compute,  $N_s D_s > N_0 D_0$ . The cost difference is

$$C_T(N_s, D_s) - C_T(N_0, D_0) = 6(N_s D_s - N_0 D_0) - 2T(N_0 - N_s).$$

It is positive at  $T = 0$  and strictly decreasing in  $T$  because  $N_s < N_0$ . It equals zero at

$$T_{\text{break}} = \frac{3(N_s D_s - N_0 D_0)}{N_0 - N_s} > 0.$$

Below  $T_{\text{break}}$ , the training-only optimum  $(N_0, D_0)$  is cheaper. Above it, the smaller model  $(N_s, D_s)$  is cheaper.

### 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,  $u_n = (1/n, \dots, 1/n)$  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  $t \in [0, 1]$ ,

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

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.)

**Solution.** Additivity with  $m = n = 1$  gives  $H_2(t, 1 - t) = H_2(t, 1 - t) + H_1(1)$ , so  $H_1(1) = 0$ . We first extend additivity to  $k$  groups: for any  $w \in \Delta_k$  and  $r^{(i)} \in \Delta_{m_i}$ ,

$$H_{m_1+\dots+m_k}(w_1r^{(1)}, \dots, w_kr^{(k)}) = H_k(w) + \sum_{i=1}^k w_i H_{m_i}(r^{(i)}).$$

The case  $k = 1$  follows from  $H_1(1) = 0$ , and the case  $k = 2$  is the axiom. For the induction step, first suppose every  $w_i > 0$ . Put  $t = 1 - w_k$ ,  $\hat{w} = (w_1/t, \dots, w_{k-1}/t)$ , and  $R = (\hat{w}_1r^{(1)}, \dots, \hat{w}_{k-1}r^{(k-1)})$ . Apply binary additivity to  $(tR, (1 - t)r^{(k)})$  and use the induction hypothesis to expand the entropy of  $R$ . The terms involving only group choices combine to  $H_k(w)$ , because additivity also gives

$$H_k(w) = H_2(t, 1 - t) + tH_{k-1}(\hat{w}).$$

The remaining terms are the required weighted sum. Continuity in  $w$  extends the identity to zero group probabilities, with the conditional distributions held fixed. Now split  $mn$  equally likely outcomes into  $m$  groups of  $n$ . Taking  $w = u_m$  and  $r^{(i)} = u_n$  gives  $A(mn) = A(m) + A(n)$ .

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

**Solution.** Taking  $m = n = 1$  in part (i) gives  $A(1) = 2A(1)$ , so  $A(1) = 0$ . Applying  $A(2n) = A(2) + A(n)$  repeatedly with  $A(2) = \log 2$  gives  $A(2^k) = k \log 2$  for every integer  $k \geq 0$ .

- (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  $k \rightarrow \infty$ .)

**Solution.** Part (i) gives  $A(n^k) = kA(n)$ . Let  $r_k = \lfloor k \log_2 n \rfloor$ , where  $\log_2 n = \log n / \log 2$  and  $\lfloor x \rfloor$  is the greatest integer at most  $x$ . Since  $2^{r_k} \leq n^k < 2^{r_k+1}$ , monotonicity and part (ii) give

$$r_k \log 2 \leq kA(n) \leq (r_k + 1) \log 2.$$

Divide by  $k$  and let  $k \rightarrow \infty$ . Since  $r_k/k \rightarrow \log_2 n$ , we get  $A(n) = \log n$ .

- (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.

**Solution.** Partition  $N$  equally likely outcomes into groups of sizes  $a_1, \dots, a_n$ . Group  $i$  has probability  $p_i = a_i/N$ , and its conditional distribution is uniform on  $a_i$  outcomes. By the  $k$ -group form of additivity from part (a)(i),

$$H_N(u_N) = H_n(p) + \sum_i p_i H_{a_i}(u_{a_i}).$$

Part (a)(iii) gives  $H_N(u_N) = \log N$  and  $H_{a_i}(u_{a_i}) = \log a_i$ . Substituting and using  $\sum_i p_i = 1$  yields

$$H_n(p) = \log N - \sum_i p_i \log a_i = - \sum_i p_i \log p_i.$$

Distributions with positive rational coordinates are dense in  $\Delta_n$ , 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  $\text{KL}(p\|q) = \sum_i p_i \log(p_i/q_i)$ , where terms with  $p_i = 0$  contribute zero.

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

**Solution.** Let  $S = \{i : p_i > 0\}$ . Applying  $\log x \leq x - 1$  to  $x = q_i/p_i$  gives

$$\text{KL}(p\|q) \geq \sum_{i \in S} (p_i - q_i) = 1 - \sum_{i \in S} q_i \geq 0.$$

Equality requires  $q_i = p_i$  for every  $i \in S$  and  $\sum_{i \in S} q_i = 1$ . Together these give  $p = q$ . Conversely,  $\text{KL}(p\|p) = 0$ .

- (d) **(4 points)** Let  $P$  and  $Q$  be distributions on sequences  $X = (X_1, \dots, X_T) \in \mathcal{V}^T$  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

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

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

**Solution.** Write  $x = (x_1, \dots, x_T)$  and  $x_{<t} = (x_1, \dots, x_{t-1})$ . We use  $\mathbb{E}_P$  for expectation when  $X$  has distribution  $P$ . The first equality follows from the two formulas:

$$\mathbb{E}_P[-\log Q(X)] - H(P) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = \text{KL}(P\|Q).$$

For the second, write  $P(x) = \prod_t p_t(x_t \mid x_{<t})$ , and similarly for  $Q$ . On sequences with  $P(x) > 0$ ,

$$\log \frac{P(x)}{Q(x)} = \sum_{t=1}^T \log \frac{p_t(x_t \mid x_{<t})}{q_t(x_t \mid x_{<t})}.$$

Take the expectation under  $P$ . For each  $t$ , averaging over  $X_t$  given  $X_{<t}$  gives the KL divergence between the next-token conditionals. Averaging this over  $X_{<t} \sim P$  gives the  $t$ -th term.

Hence the expected log loss equals  $H(P)$  exactly when  $\text{KL}(P\|Q) = 0$ . By part (c), this holds exactly when  $P = Q$ . Equivalently, the next-token conditionals of  $P$  and  $Q$  agree at every prefix of positive  $P$ -probability.

## Sources

Problem 1 reviews standard matrix and softmax calculations. Problem 2 follows the model of Sardana et al. (2024), Section 2, which adds inference to the compute-optimal analysis of Hoffmann et al. (2022). Problem 3 follows the characterizations of entropy by Shannon (1948), Theorem 2 and of KL divergence by Leinster (2019).