COS 514: Fundamentals of Deep Learning

Fall 2025

Instructor: Prof. Sanjeev Arora
TA: Gon Buzaglo

Assignment 4

Instructions:

- Submission deadline is November 24.
- You may collaborate in groups of up to **3** students.
- If you collaborate on a problem, you must clearly state the names of your collaborators at the beginning of the solution to that problem.
- All group members must declare that they contributed equally to the solutions.
- You must write up your own solutions independently in LaTeX. **Handwritten or scanned solutions will not be accepted.**
- Cite any resources (papers, textbooks, websites) that you use.
- Submit your assignment as a single PDF on gradescope.

Problem 1: The Self-Attention Mechanism

The core of the Transformer architecture is the scaled dot-product attention mechanism. Given a set of input token embeddings, we project them into Query (Q), Key (K), and Value (V) vectors. The output is a weighted sum of the Value vectors, where the weights are determined by the similarity between Query and Key vectors.

The attention output is calculated as: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$. Assume the dimension of the key vectors, d_k, is 2.

Computational Cost of Self-Attention

Transformer processes tokens in parallel, overcoming the sequential bottleneck of RNNs. However, this comes at a computational cost.

(i) Consider the matrix multiplication QK^T in the self-attention formula for a sequence of length L and a model dimension of d. What is the computational complexity (in terms of Big-O notation) of this single operation with respect to the sequence length L?

(ii) This quadratic scaling in sequence length is a primary reason why the context window of early Transformers was limited. How does this compare to the computational complexity per step of a simple Recurrent Neural Network (RNN)? Explain why Transformers are still generally faster to train despite this.

(iii) Suppose the autoregressive transformer (eg, LLM) has an input of n tokens and it produces an output with m tokens. How many attention computations per layer must happen while producing the output?

Problem 2: Post Training

We posit an unobserved "true" reward function $r^*(x, y)$ that represents the latent human preference for response y to prompt x. Human choices are modeled by the Bradley–Terry (logistic) rule:

$$\Pr\big(y_w \succ y_l \mid x\big) \;=\; \sigma\big(r^*(x, y_w) - r^*(x, y_l)\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

That is, the probability that a human prefers response y_w over y_l increases with the difference in their latent rewards.

We now introduce a parametric model $r_\phi(x, y)$ intended to approximate $r^*(x, y)$, and we are given an i.i.d. dataset of pairwise preferences

$$\mathcal{D} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^n.$$

(a) Learning a Reward Model from Pairwise Preferences

Define the *maximum-likelihood estimator* (MLE) of r_ϕ under the model above as the parameter value

$$\hat{\phi}_{\mathrm{MLE}} \;=\; \arg\max_{\phi} \prod_{(x,y_w,y_l)\in\mathcal{D}} \sigma\big(r_\phi(x,y_w) - r_\phi(x,y_l)\big),$$

which maximizes the likelihood of the observed human preferences according to the Bradley–Terry model.

Show that finding $\hat{\phi}_{\mathrm{MLE}}$ is equivalent to minimizing the negative log-likelihood loss

$$\mathcal{L}_{\mathrm{RM}}(\phi) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \log\sigma\Big(r_\phi(x_i,y_{w,i}) - r_\phi(x_i,y_{l,i})\Big) \;=\; -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\log\sigma\big(r_\phi(x,y_w) - r_\phi(x,y_l)\big)\big].$$

(b) The Direct Path: Direct Preference Optimization (DPO)

The traditional RLHF pipeline uses the learned reward model r_ϕ to fine-tune a policy π_θ by maximizing a KL-regularized objective:

$$\mathcal{J}(\pi_\theta) \;=\; \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r_\phi(x,y)] \;-\; \beta\, D_{\mathrm{KL}}\big(\pi_\theta(\cdot|x)\,\|\,\pi_{\mathrm{ref}}(\cdot|x)\big),$$

where π_{ref} is the fixed supervised (SFT) policy and $\beta > 0$ controls the KL strength.

By Lemma 20.3.1 from the course book, the optimal policy π^* that maximizes $\mathcal{J}(\pi_\theta)$ satisfies

$$\pi^*(y|x) \;=\; \frac{1}{Z(x)}\,\pi_{\mathrm{ref}}(y|x)\,\exp\Big(\tfrac{1}{\beta}\,r_\phi(x,y)\Big), \qquad Z(x) = \sum_{y'} \pi_{\mathrm{ref}}(y'|x)\exp\Big(\tfrac{1}{\beta} r_\phi(x,y')\Big). \tag{1}$$

1. Using Eq. (1), express the reward in terms of the policy:

$$r_\phi(x,y) \;=\; \beta\log\frac{\pi^*(y|x)}{\pi_{\mathrm{ref}}(y|x)} \;+\; \beta\log Z(x).$$

2. For two responses (y_w, y_l), show that the constant $\log Z(x)$ cancels out:

$$r_\phi(x,y_w) - r_\phi(x,y_l) \;=\; \beta\left(\log\frac{\pi^*(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \log\frac{\pi^*(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right).$$

3. Replace π^* by a learnable policy π_θ and substitute the expression from step 2 into the loss $\mathcal{L}_{\text{RM}}(\phi)$ from part (a) to obtain the Direct Preference Optimization objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) \;=\; -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]. \tag{2}$$

(c) Additional Questions

(2) **Shift invariance.** Explain briefly why the reward $r_\phi(x, y)$ is identifiable only up to an additive constant for each prompt $c(x)$.

(4) **Gradient of DPO.** Derive the gradient of $\mathcal{L}_{\text{DPO}}$ with respect to θ in terms of $\nabla_\theta \log \pi_\theta(y|x)$.

Problem 3: Adversarial Training as DRO

This exercise connects standard adversarial training with Distributionally Robust Optimization (DRO). Consider the standard adversarial objective with an ℓ_2 adversary:

$$\text{(LHS)} \quad \min_\theta \mathbb{E}_{x\sim P}\left[\max_{\|\delta\|_2\le\epsilon} \mathcal{L}(\theta, x+\delta)\right]$$

And the following DRO objective, which seeks robustness to a worst-case distribution Q that is close to the empirical data distribution P in Wasserstein-1 distance:

$$\text{(RHS)} \quad \min_\theta \sup_{Q: W_1(Q,P)\le\rho} \mathbb{E}_{x\sim Q}[\mathcal{L}(\theta, x)]$$

Assume the loss function $\mathcal{L}(\theta, x)$ is K-Lipschitz with respect to its input x. You will use the Kantorovich-Rubinstein duality for the W_1 distance:

$$W_1(Q, P) = \sup_{f:\|f\|_L\le 1} \left(\mathbb{E}_{x\sim Q}[f(x)] - \mathbb{E}_{x\sim P}[f(x)]\right)$$

where the supremum is over all 1-Lipschitz functions f.

(a) Using the KR duality, show that the increase in expected loss for any valid distribution Q is bounded:

$$\mathbb{E}_Q[\mathcal{L}] - \mathbb{E}_P[\mathcal{L}] \le K \cdot W_1(Q, P)$$

(b) The worst-case distribution Q^* that achieves the supremum in the DRO objective can be constructed by deterministically shifting each data point x from the distribution P to a new point $x + \delta(x)$. For such a Q^*, the Wasserstein distance simplifies to $W_1(Q^*, P) = \mathbb{E}_{x\sim P}[\|\delta(x)\|_2]$. To maximize the loss under the constraint $\mathbb{E}_{x\sim P}[\|\delta(x)\|_2] \le \rho$, a simple and effective strategy is to choose a uniform shift length $\|\delta(x)\|_2 = \rho$ for all x.

Using a first-order Taylor approximation for the loss, $\mathcal{L}(x+\delta) \approx \mathcal{L}(x) + \nabla_x \mathcal{L}(x)^\top \delta$, what is the optimal direction for the shift $\delta(x)$ to maximize the loss?

(c) Substitute this optimal perturbation into the DRO objective (RHS). What is the resulting optimization problem?

(d) Now consider the adversarial training objective (LHS). The inner maximization is often approximated by taking a single gradient ascent step (the FGSM method). Solve this inner maximization, $\max_{\|\delta\|_2 \le \epsilon} \mathcal{L}(\theta, x+\delta)$, using this approximation.

(e) Compare your results from (c) and (d). What is the relationship between the adversarial budget ϵ and the distributional budget ρ that makes the two formulations equivalent?

Problem 4: Mode Collapse with Linear Discriminators

This problem makes the theory of mode collapse concrete for a simple discriminator class by explicitly working out the sample complexity required for a generator to fool it.

Let the data be in $\mathbb{R}^d$. Consider a class of linear discriminators $\mathcal{D} = \{D(x) = \sigma(w^T x) \mid \|w\|_2 \le L\}$, where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function.

(a) Generalization bounds for a function class often depend on its Lipschitz constant. Show that for any $D \in \mathcal{D}$, the function $f_D(x) = \log(1 - D(x))$ is L-Lipschitz. [1] This implies the loss function in the GAN objective is Lipschitz with a constant C proportional to L.

[1] Hint: Recall that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, which is maximized at $z = 0$.

Based on the result from (a) and the generalization bounds presented in Chapter 5 (e.g., Theorem 5.2.7), the number of samples required to guarantee that the empirical loss is within ϵ of the true loss for all discriminators in $\mathcal{D}$ is $M = \Omega\left(\frac{L^2 d}{\epsilon^2}\right)$. Let us fix M to be this sample complexity.

Now, consider a true data distribution p_{data} that is a uniform mixture of K well-separated modes, where $K \gg M$. For instance, the mixture of K gaussians $p_{\text{data}} = \frac{1}{k}\sum_{i=1}^{K} \mathcal{N}(c \cdot e_i, \sigma^2 I)$ for large c and small σ, where $\{e_i\}$ are standard basis vectors. Let the generator's distribution, p_g, be the uniform distribution over a set S of just M samples drawn i.i.d. from p_{data}.

(b) **(Mode Collapse!)** In this setting, recall that the GAN value function is

$$V(D, G) \;=\; \mathbb{E}_{x\sim p_{\text{data}}}\big[\log D(x)\big] \;+\; \mathbb{E}_{x\sim p_g}\big[\log(1 - D(x))\big].$$

The generator G induces the empirical distribution $p_g = \frac{1}{M}\sum_{j=1}^{M} \delta_{x_j}$, where the x_j are i.i.d. samples from p_{data}.

(i) Using the law of large numbers (or Hoeffding's inequality), show that for any fixed w with $\|w\|_2 \le L$,

$$\Big|\mathbb{E}_{x\sim p_{\text{data}}}[w^\top x] \;-\; \mathbb{E}_{x\sim p_g}[w^\top x]\Big| \;\xrightarrow[M\to\infty]{}\; 0.$$

Intuitively, this means that every linear projection $w^\top x$ has nearly the same average under p_{data} and under p_g.

(ii) Show that no linear discriminator $D(x) = \sigma(w^\top x)$ can reliably distinguish between p_{data} and p_g: their inputs $w^\top x$ have nearly the same distributions.

(iii) Conclude that for sufficiently large M (as defined in part (a)), the value of the GAN objective,

$$\max_{D\in\mathcal{D}} V(D, G),$$

will be very close to its baseline value when the two distributions are identical,

$$V(D, G) \approx -2\log 2,$$

even though the generator has collapsed from K well-separated modes down to only M sample points.