CSCI 145 Problem 21: How Low Can Honest Next-Token Loss Go?

October 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Wednesday, November 25**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- A complete, good-faith attempt earns submission credit. The grader may select any recent assigned problem for a brief conversation, so keep your work and be prepared to explain each solution and answer a related question. Scheduling details will be announced in class.

Problem 21: How Low Can Honest Next-Token Loss Go?

A next-token model is tested on a deliberately unpredictable categorical target. The visible prefix is H, while the next token $Y \in \{1, \ldots, V\}$ for $V \geq 2$ is independent of H and satisfies

$$\Pr(Y = j) = p_j, \qquad p_j > 0, \qquad \sum_{j=1}^{V} p_j = 1.$$

A causal model outputs a distribution $\mathbf{q}(H) = (q_1(H), \ldots, q_V(H))$. During teacher-forced training, an unmasked model can instead expose the future one-hot target $\mathbf{e}_Y$ to the representation used for prediction.

1. Given the independence of Y and H, what is the smallest possible expected cross-entropy of a causal predictor, and which predictors attain it?

2. If a target-visible model maps $\mathbf{e}_Y$ to logits $\lambda \mathbf{e}_Y$ with $\lambda > 0$, what expected cross-entropy does it achieve, and what does its $\lambda \to \infty$ limit imply when compared with the honest limit?

Hint: You may use that a stack of causally masked transformer blocks keeps position i dependent only on positions $1, \ldots, i$, and that $\log x \leq x - 1$ for every $x > 0$, with equality only at $x = 1$.

Takeaway: *A model cannot predict an independent random next token better than its known probabilities. Letting it see the answer can drive training loss toward zero anyway. A low training loss only measures next-token prediction if the model cannot look ahead.*