Feedback-Augmented Loss: Mathematical Derivation

1. Standard Loss

For a model $p_\theta(y \mid x)$, the standard cross-entropy loss is:

$$L_{\text{CE}} = -\log p_\theta(y \mid x) \tag{1}$$

For sequence models, this is typically averaged over all tokens in the sequence.

2. Feedback Mechanism

Agent_b provides, for each sample:

- **solution_score**: $S_{\text{sol}} \in [0, 1]$
- **reasoning_score**: $S_{\text{reas}} \in [0, 1]$
- **is_correct**: $C \in \{0, 1\}$

We define the combined feedback score:

$$S = \frac{S_{\text{sol}} + S_{\text{reas}}}{2} \tag{2}$$

3. Feedback-Augmented Loss

The feedback-augmented loss for each sample is:

$$L_{\text{aug}} = L_{\text{CE}} - \lambda_1 S \cdot \overline{\log p_\theta(y \mid x)} + \lambda_2 (1 - S) \cdot \overline{\max_j p_\theta(j \mid x)} \cdot (1 - C) \tag{3}$$

Where:

- L_{CE} is the mean cross-entropy over tokens in the sample.
- $\overline{\log p_\theta(y \mid x)}$ is the average log-probability assigned to the correct tokens.
- $\overline{\max_j p_\theta(j \mid x)}$ is the average maximum softmax probability (model confidence) per token.
- λ_1, λ_2 are scalar hyperparameters controlling feedback influence.
- $(1 - C)$ applies the penalty term only if the prediction is incorrect.

4. Batch Loss

For a batch of N samples, the loss is:

$$L_{\text{batch}} = \frac{1}{N} \sum_{i=1}^{N} L_{\text{aug}}^{(i)} \tag{4}$$

5. Derivation and Interpretation

- **Cross-Entropy:** Regular maximum likelihood objective, encourages correct predictions.
- **Reward Term:** $-\lambda_1 S \cdot \overline{\log p_\theta(y \mid x)}$; high feedback ($S \approx 1$) increases probability for correct outputs.
- **Penalty Term:** $\lambda_2(1 - S) \cdot \overline{\max_j p_\theta(j \mid x)} \cdot (1 - C)$; low feedback ($S \approx 0$) and incorrect ($C = 0$) penalizes overconfident mistakes.
- **Differentiability:** All terms are differentiable, enabling gradient-based optimization.

6. Final Formula

For each sample i:

$$L_{\text{aug}}^{(i)} = L_{\text{CE}}^{(i)} - \lambda_1 S^{(i)} \cdot \overline{\log p_\theta(y \mid x)} + \lambda_2(1 - S^{(i)}) \cdot \overline{\max_j p_\theta(j \mid x)} \cdot (1 - C^{(i)}) \tag{5}$$

Batch loss:

$$L_{\text{batch}} = \frac{1}{N} \sum_{i=1}^{N} L_{\text{aug}}^{(i)} \tag{6}$$

7. Summary

- **High feedback:** Model is rewarded for high confidence on correct outputs.
- **Low feedback & incorrect:** Penalizes overconfident wrong answers.
- **Fully differentiable:** All terms influence learning directly via gradients.