

Discussion: Underfitting of Plackett-Luce Models in Cyclic Preferences

Problem Statement

The Plackett-Luce (PL) model can handle lists of ranked items rather than just pairs. But it relies on the exact same fundamental assumption: Transitivity. The PL model assumes that every response y has a scalar reward score, $r(x, y)$. The probability of one item being chosen over another is determined by which one has the higher score. Because real numbers are transitive, the PL model enforces transitivity on the preferences.

1 Circumstances

The specific circumstance that leads to optimization failure is the presence of **Cyclic Preferences** in the dataset.

Assume, three responses y_A, y_B, y_C for a context x . In the dataset, we observe the following rankings:

- **Sample 1:** $y_A \succ y_B$
- **Sample 2:** $y_B \succ y_C$
- **Sample 3:** $y_C \succ y_A$

2 Effects on the Optimization Objective

The optimization objective for the Reward Model seeks to maximize the likelihood of the observed rankings. To minimize this loss, the model attempts to assign scalar reward scores such that the “winner” has a strictly higher value than the “loser.” Based on the cyclic samples above, the objective function effectively imposes the following contradictory constraints on the scalar rewards:

1. From Sample 1: The model pushes $RM(y_A) > RM(y_B)$.
2. From Sample 2: The model pushes $RM(y_B) > RM(y_C)$.
3. From Sample 3: The model pushes $RM(y_C) > RM(y_A)$.

This implies:

$$RM(y_A) > RM(y_B) > RM(y_C) > RM(y_A)$$

This is an impossibility for scalar values, as a number cannot be greater than itself.

3 Instability and Underfitting

Because the objective function attempts to satisfy impossible constraints, the training process suffers from two issues:

A. Unstable Training Updates

The gradients will fail to converge. When the model optimizes for a batch containing $y_A \succ y_B$, it increases the weight for A and decreases it for B . In a subsequent batch containing $y_C \succ y_A$, it decrease the weight for A . The model creates a loop where it endlessly chases the most recent data point.

B. Underfitting

Since no hierarchy satisfies the cycle $A \succ B \succ C \succ A$, the optimization tries to minimize the average error across all samples. The model minimizes the penalty by pushing all rewards to be approximately equal:

$$RM(y_A) \approx RM(y_B) \approx RM(y_C)$$

This represents underfitting. Although the data clearly shows strong preferences (e.g., A beats B), the model follows random guessing.