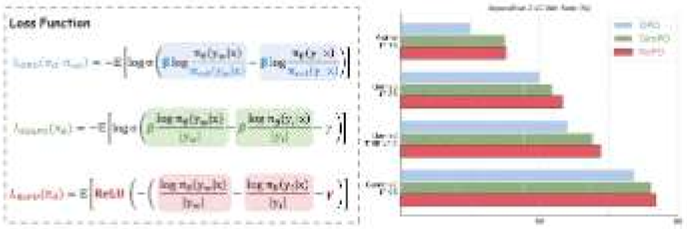
RePO: Understanding Preference Learning Through ReLU-Based Optimization

Junkang Wu, Kexin Huang, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, Xiang Wang

University of Science and Technology of China, Alibaba Group, Institute of Dataspace, Hefei Comprehensive National Science Center

Introduction: Rethinking Preference Optimization

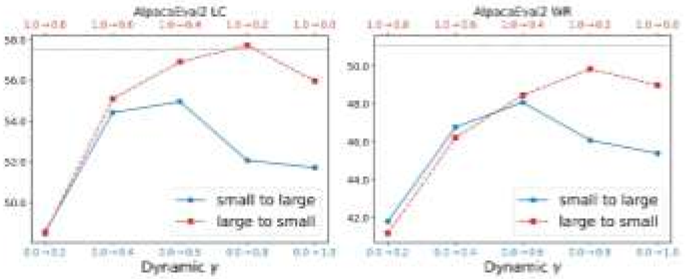
Aligning Large Language Models (LLMs) with human preferences is commonly achieved through methods like Direct Preference Optimization (DPO) and Simple Preference Optimization (SimPO). These approaches prevent over-optimization by using complex mechanisms like sigmoid-based gradient weighting or explicit regularization terms. This paper challenges this conventional wisdom by investigating a surprising finding: a simple ReLU activation can effectively align models without these components, suggesting that the fundamental principles of preference learning may be simpler than previously thought.



Dynamic Scheduling and Conclusion

Further experiments reveal that dynamically scheduling the margin γ can enhance performance. A 'large to small' strategy, where γ is gradually decreased during training (e.g., from 1.0 to 0.2), consistently outperforms a fixed γ or a 'small to large' schedule. This creates an effective curriculum: aggressive updates are permitted early on, while learning becomes more focused on challenging examples later, naturally preventing over-optimization.

Conclusion: Our work demonstrates that a simple binary thresholding mechanism, implemented via a ReLU function, is a powerful and fundamentally sound approach for preference learning. The key insight is that **selecting *which* examples to learn from is more critical than determining *how much* to learn from each one**. RePO provides a simple, effective baseline that encourages a re-evaluation of the core principles behind LLM alignment.

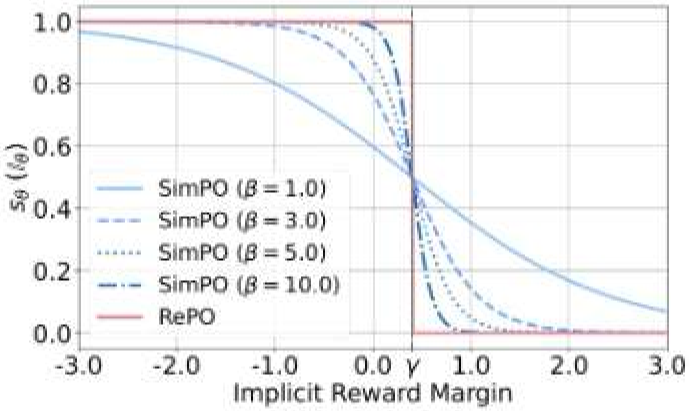


The RePO Method: Simple ReLU-based Filtering

We introduce ReLU-based Preference Optimization (RePO), which simplifies preference learning by removing the log-sigmoid function and SFT regularization. The method uses a straightforward loss function based on a ReLU activation and a single hyperparameter, the margin `γ`:

```
L_RePO(π_θ) = E[ReLU(−(M_θ − γ))]
```

where `M_θ` is the length-normalized implicit reward margin. This creates a binary filtering mechanism: gradient updates are applied only to sample pairs with an insufficient margin (`M_θ < γ`), while well-separated pairs are ignored. As shown in the figure below, SimPO's smooth, sigmoid-like weighting function becomes steeper as its `β` hyperparameter increases, converging to the hard-margin step function of RePO. This establishes RePO as the theoretical limit of SimPO, but without the complexity of tuning `β`.



Theoretical Foundation: Why RePO Works

RePO's empirical success is grounded in a strong theoretical connection to binary classification. The ideal, but non-differentiable, goal of preference learning is to minimize the 0-1 loss. Our analysis reveals a remarkable result:

Theorem 4.2: The RePO loss function, `ReLU(-x)`, is precisely the **convex envelope** (the tightest possible convex approximation) of the 0-1 loss `I(x < 0)`.

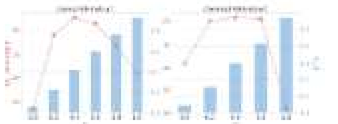
This finding explains why such a simple mechanism is so effective. By optimizing the ReLU surrogate, we are using gradient-based methods to solve for the same optimal solutions as the intractable 0-1 loss, a property not shared by the logistic loss used in DPO and SimPO.

Empirical Findings and Analysis

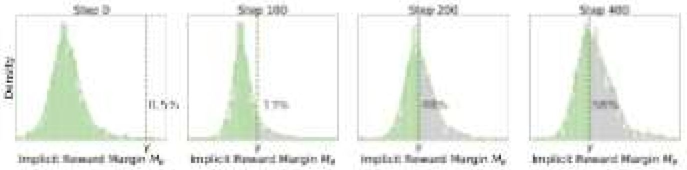
Experiments demonstrate that RePO's simple mechanism is highly effective.

1. Superior Performance: Across multiple models (Llama3, Gemma2), RePO consistently achieves competitive or superior results compared to established methods like DPO, IPO, and SimPO on benchmarks like AlpacaEval 2 and Arena-Hard (Table 1). As `β` increases, SimPO's performance improves but is still matched or slightly exceeded by the simpler RePO.

2. Over-optimization Control: The margin `γ` creates a natural trade-off. Performance follows an inverted U-shape as `γ` increases, peaking at an optimal value before declining. This shows `γ` is critical for preventing over-optimization by filtering out excessively easy examples.



3. Emergent Curriculum: During training, the distribution of reward margins progressively shifts to the right. This means the model naturally improves at discriminating between responses, and as a result, an increasing fraction of the data is filtered out (from 0.5% to 58%). The model effectively creates its own curriculum, focusing on more challenging examples in later stages.


Step 0
Step 100
Step 200
Step 400
Density
γ
Implicit Reward Margin M_θ