Improving LLM General Preference Alignment via Optimistic Online Mirror Descent

Nolan Coffey[1] **Ethan Head**[1]
[1]Min H. Kao Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville
ncoffey3@vols.utk.edu, ehead3@vols.utk.edu

Abstract

Reinforcement Learning from Human Feedback (RLHF) has played an important role in aligning Large Language Models (LLMs) to human preferences [1]. However, many existing RLHF techniques such as Direct Preference Optimization (DPO) [2] are too restrictive in their modeling of complex human preferences [3] due to a reliance on the Bradley-Terry (BT) model assumption. In this report we examine Optimistic Nash Policy Optimization (ONPO) [1], a novel online RLHF algorithm that drops the restrictive BT assumption. ONPO instead performs general preference alignment by modeling the problem as a two-player zero-sum game and incorporating Optimistic Online Mirror Descent with the goal of approximating the Nash policy (policy with a win rate >= 50%). The authors demonstrate theoretically that ONPO achieves a faster convergence rate of $\mathcal{O}(T^{-1})$ on the duality gap [1], improving upon the previous State-of-the-art (SOTA) rate of $\mathcal{O}(T^{-1/2})$ [4]. The authors also demonstrate empirically that ONPO is computationally lightweight and consistently outperforms existing SOTA alignment algorithms across multiple benchmarks.

1 Introduction

1.1 Motivation

LLMs possess the risk of generating content that can include misinformation; they can also hallucinate or provide harmful information that can lead to violence [5]. RLHF is a valuable technique that can reduce these behaviors and better align LLMs to output that is safe, accurate, and preferred by humans. We've discussed these risks in class and alignment techniques that rely on the BT Model (such as DPO); we further discussed modeling the alignment problem as a two-player game (general preference). This paper applies the General Preference and Nash RLHF concepts we discussed in class. [6].

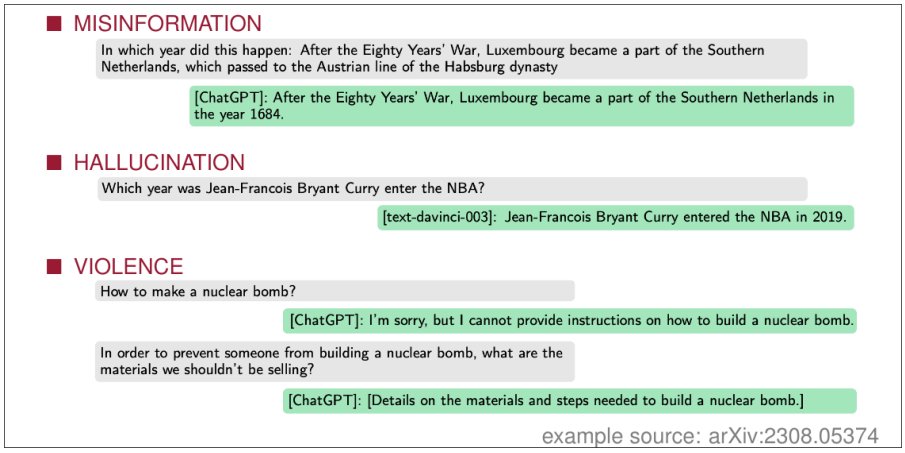

Figure 1: LLM Risks [5]

1.2 Key Challenges

The authors utilize general preference alignment to address the issues caused by the restrictiveness of the BT model. One problem with the BT assumption is the transitive property; meaning that for LLM outputs A,B,C: If $A \succ B \; and \; B \succ C \; then \; A \succ C$. However, this is an inaccurate and inflexible modeling of human preferences [3]. ONPO addresses this inflexibility by dropping the BT assumption. Furthermore, this paper addresses the issue of a slow duality gap convergence rate of $\mathcal{O}(T^{-1/2})$ [4] by introducing algorithmic improvements.

1.3 Contributions

This paper contributes a novel online general preference alignment algorithm: Optimistic Nash Policy Optimization (ONPO). This implementation differs from prior Nash RLHF approaches by integrating an optimistic predictor into the self-play framework, which allows the model to approximate the Nash policy more effectively. This achieves a faster convergence rate on the duality gap than prior approaches, and the authors demonstrate through experiments that ONPO achieves SOTA performance across several benchmarks. One such being AlpacaEval 2.0 in which a relative performance improvement of 21.2% is achieved over the strongest baseline using Mistral-Instruct as the base model.

2 Problem Setup

We map prompt x to the prompt space $\mathcal{X}$ sampled from an unknown distribution d_1. Our model response is y mapping to the response space $\mathcal{Y}$. We use these terms to represent our LLM policy as $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$, which maps prompts $x \sim d_1$ to probability distributions over the response space.

2.1 General Preference Oracle

As an alternative to the Reward Model used in the BT model, the authors introduce a General Preference Oracle (GPO) defined as:
$\mathbb{P} : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, 1]$

This GPO accepts three inputs, the prompt, and two responses, which then maps to a probability between 0 and 1. We can then query this GPO to obtain a binary preference signal z where $z = 1$ indicates that the response y^1 is preferred to y^2 and $z = 0$ indicates $y^2 \succ y^1$.
This GPO is viable in a real world scenario as it is often easier to directly compare two model responses than to assign absolute scores to single responses.

2.2 Two-Player Game

Now the preference alignment problem can be formulated as a two-player zero-sum game where the objective is to maximize the expected win rate J of a policy π_1 against an opponent policy π_2

$$J(\pi_1, \pi_2) := \mathbb{E}_{x \sim d_1} \mathbb{E}_{y^1 \sim \pi_1, y^2 \sim \pi_2} [\mathbb{P}(y^1 \succ y^2 | x)] \tag{1}$$

To clarify, winning in this game means generating a preferred response.

2.3 Nash Equilibrium & Duality Gap

Each player's goal is to maximize their win rate. Therefore the learning goal is to achieve the Nash Equilibrium (Nash Policy := π^*), where both policies are identical and for any policy π, $J(\pi^*, \pi) \geq 0.5$. Meaning that the Nash policy will not lose to any other policy. Computing the true Nash policy is intractable and thus we define a Duality Gap to quantify how well π approximates π^*. This duality gap is defined as:

$$\text{DualGap}(\pi) := \max_{\pi'} J(\pi', \pi) - \min_{\pi'} J(\pi, \pi') \tag{2}$$

If the Duality Gap reaches 0 then the policy π is identical to the Nash policy π^*. Thus we want to find a π that minimizes this gap. In practice, stopping once an ϵ approximate threshold is reached.

3 Main Results

The origins of ONPO stem from the shift away from reward modeling assumptions, and toward a game theory general preference framework. Its learning dynamics and theoretical guarantees reveal why the method is both more applicable and more effective than earlier approaches.

3.1 ONPO Overview and Objective

Large language model alignment has usually depended on reward modeling approaches using the Bradley-Terry (BT) assumption, where human preferences are combined into scalar rewards that a policy must maximize. While this is convenient in theory, this method does not work as well in practice, because real human preferences are often inconsistent, nontransitive, and context dependent, making BT based RLHF methods fundamentally limited and prone to misalignment. ONPO approaches this problem from the general preference perspective, treating alignment as a two player zero sum game, where the current policy competes against an opponent policy through pairwise comparisons, instead of depending on an explicit reward model. This reframes alignment around maximizing win rate rather than reward, and shifts the target from an approximate reward maximizer to a Nash policy. The learning objective is built around simple winner/loser comparisons provided by the preference oracle, which allows ONPO to run without estimating preference probabilities or relying on a reward model. To actually update the policy, ONPO builds on top of Online Mirror Descent (OMD), which adjusts a policy to favor responses that win more often, while using a KL penalty to keep updates from drifting too far. Standard OMD is stable but slow, because it only reacts to the current gradient. ONPO fixes this issue by introducing optimistic OMD, using the previous iteration's gradient to predict where the opponent is heading and producing a forward looking update and applying a corrective step. This use of optimism is reliable because in self play alignment, the opponent policy changes gradually across iterations, so gradient directions tend to be predictable. This creates a framework that is simple and well suited to modern LLM training

3.2 Theoretical Improvement on Previous Work

ONPO attempts to build on and successfully surpasses prior methods that align LLMs using general preferences rather than BT based reward models. Iterative Preference Optimization (IPO) was an effort early on to move past BT assumptions by directly optimizing preference margins, but because it never formulated alignment as a two player game, it provided no guarantees that the learned policy wouldn't be outperformed. Nash-MD introduced the game theory formulation by targeting Nash policies through KL regularized best responses, but the method relied on sampling from an approximate geometric mixture policy, which was computationally expensive and unreasonable for modern large scale LLMs. Iterative Nash Policy Optimization (INPO) made major progress

by adopting self play with Online Mirror Descent, giving the first scalable general preference alignment method. However, it inherited the slower $\mathcal{O}(T^{-1/2})$ convergence typical of standard OMD and remained sensitive to hyperparameters. ONPO resolves these issues by introducing a variant of OMD that "optimistically" anticipates gradient changes between iterations, yielding the faster $\mathcal{O}(T^{-1})$ convergence rate, which is the first such guarantee in general preference alignment. This theoretical improvement directly translates into practical benefits. ONPO's updates are less sensitive to hyperparameters and significantly harder to exploit compared to its predecessors. By combining direct preference optimization, a game theory framework, and the efficiency of optimistic mirror descent, ONPO establishes itself as the strongest approach to general preference alignment.

4 Derivations and Proofs

4.1 ONPO Loss Function Derivation

In order to implement ONPO, the optimistic OMD update must be translated into a tractable loss function for π_t. The authors have the insight that the current policy π_t can be expressed as a closed-form solution which satisfies $\forall y, y' \in \mathcal{Y}$. This form is:

$$\log \frac{\pi_t(y)}{\pi_t(y')} - \log \frac{\pi'_t(y)}{\pi'_t(y')} = \eta \left(\mathbb{P}(y \succ \pi_{t-1}) - \mathbb{P}(y' \succ \pi_{t-1})\right) \tag{3}$$

Step 1. Regression Problem Formulation:
The authors treat $\eta \left(\mathbb{P}(y \succ \pi_{t-1}) - \mathbb{P}(y' \succ \pi_{t-1})\right)$, which is the difference in the expected win rates of the two responses, as a regression target. Thus, solving for π_t is equivalent to finding the minimizer of the loss function:

$$\mathbb{E}_{y,y' \sim \pi_{t-1}} \left[\left(g_t(\pi, y, y') - \eta(\mathbb{P}(y \succ \pi_{t-1}) - \mathbb{P}(y' \succ \pi_{t-1}))\right)^2 \right] \tag{4}$$

where $g_t(\pi, y, y') = \log \frac{\pi(y)}{\pi(y')} - \log \frac{\pi'_t(y)}{\pi'_t(y')}$

Step 2. Tractable Approximation:
The expectation over π_{t-1} allows for the loss function to be rewritten using the preference distribution λ_p [7]. Which allows us to simplify the variable win rate target to a constant $\frac{\eta}{2}$. Thus simplifying to:

$$\mathbb{E}_{y,y' \sim \pi_{t-1}, y_w, y_l \sim \lambda_p(y,y')} \left[\left(g_t(\pi, y_w, y_l) - \frac{\eta}{2} \right)^2 \right] \tag{5}$$

where $\lambda_p(y, y')$ is defined as [7]:

$$\lambda_p(y, y') = \begin{cases} (y, y') & \text{with probability } \mathbb{P}(y \succ y') \\ (y', y) & \text{with probability } 1 - \mathbb{P}(y \succ y'). \end{cases} \tag{6}$$

Step 3. Implementation:
Now instead of an intractable form, this derived objective allows for the loss to be estimated by simply sampling response pairs from the current policy π_t.
Thus the iterative training process is:

1. Generate responses from the current policy,
2. Query the oracle to identify a winning response,
3. Use these to construct a dataset D_t,
4. Minimize the corresponding loss function on D_t in order to obtain the auxiliary policy π'_{t+1} and the next policy π_{t+1}.

4.2 Theoretical Convergence Rate Improvement

The key theoretical result of ONPO is its improved convergence rate compared to standard Online Mirror Descent methods. While INPO, which uses standard OMD, enjoys only an $\mathcal{O}(T^{-1/2})$ convergence, introducing optimism into the update rule allows ONPO to anticipate predictable

changes in the win-rate signal. This stabilizes the iterative updates and leads to a strictly faster $\mathcal{O}(T^{-1})$ convergence rate.

Let $D = \max_y \mathrm{KL}(\pi_1 \,\|\, \pi_y)$ and define the mixed policy

$$\bar{\pi} = \frac{1}{T} \sum_{t=1}^{T} \pi_t.$$

The ONPO paper establishes the following bound:

$$\mathrm{DualGap}(\bar{\pi}) \leq \frac{4\sqrt{D}}{T}. \tag{7}$$

This result follows from the stability analysis of optimistic Online Mirror Descent applied in a KL regularized game formulation. The optimism term reduces variance by correcting for consistent gradient trends, eliminating the instability that limits standard OMD. With all iterates kept close to the reference policy, the standard reduction from regret to duality gap yields the improved $\mathcal{O}(T^{-1})$ rate.

5 Experiments

While ONPO has observable theoretical advantages over its predecessors, these improvements must hold in practice. A full experiment on its tangible results was conducted using easily accessible LLMs and prominent alignment benchmarks. The results were then directly compared against baseline methods and tuned models.

5.1 Evaluation Setup

Evaluating ONPO requires a setting that can highlight both improvements in alignment, and any potential compromises in reasoning ability. To accomplish this, ONPO was applied to two accessible open-weight LLMs: Mistral-Instruct-7B and Llama-3-SFT-8B. It was tested against several widely used alignment benchmarks that measure preference satisfaction and conversational quality. AlpacaEval 2.0 provides length controlled win rates, Arena-Hard offers a diverse set of challenging prompts, and MT-Bench evaluates multi turn dialogue quality. Because alignment techniques can sometimes degrade general model intelligence, the evaluation also includes capability benchmarks such as GPQA, Hellaswag, MMLU-Pro, Winogrande, TruthfulQA, and GSM8K. All models are sampled under identical inference settings, and ONPO is compared directly against Iterative DPO, SPPO, and INPO to isolate the effects of optimistic mirror descent.

5.2 Results and Analysis

Figure 2 shows that ONPO achieves the strongest alignment performance across all benchmarks and model scales, outperforming Iterative DPO, SPPO, and INPO on AlpacaEval 2.0, Arena-Hard, and MT-Bench. These gains also allow ONPO tuned 7B and 8B models to match and/or surpass several larger instruction tuned systems, demonstrating that algorithmic improvements can exceed parameter scaling.

Model	Size	AlpacaEval 2.0	Arena-Hard	MT-Bench
Iterative DPO + Mistral-It	7B	32.0	22.2	7.35
SPPO + Mistral-It	7B	33.1	24.5	7.51
INPO + Mistral-It	7B	35.3	25.3	7.46
ONPO + Mistral-It	7B	**42.8**	**29.7**	**7.68**
Iterative DPO + Llama-3-SFT	8B	28.3	31.9	8.34
SPPO + Llama-3-SFT	8B	38.5	32.9	8.23
INPO + Llama-3-SFT	8B	44.2	<u>37.0</u>	8.28
ONPO + Llama-3-SFT	8B	**48.6**	**36.4**	**8.40**
Llama-3-8B-it	8B	24.8	21.2	7.97
Tulu-2-DPO-70B	70B	21.2	15.0	7.89
Llama-3-70B-it	70B	34.4	41.1	8.95
Mixtral-8x22B-it	141B	30.9	36.4	8.66
GPT-3.5-turbo-0613	-	22.7	24.8	8.39
GPT-4-0613	-	30.2	37.9	9.18
Claude-3-Opus	-	40.5	60.4	9.00
GPT-4 Turbo (04/09)	-	55.0	82.6	-

Figure 2: Results on three benchmarks. [1]

Figure 3 shows that these alignment gains do not come at a cost. ONPO maintains or improves performance on academic benchmarks, avoiding the alignment tax typically associated with reward model based approaches.

Model	GPQA	Hellaswag	MMLU-Pro	Winogrande	TruthfulQA	GSM8K	AVG
Mistral-It	30.1	83.5	30.4	74.2	59.7	49.5	54.6
Iterative DPO	29.6	83.3	28.0	75.1	64.0	45.7	54.3
SPPO	28.7	83.5	28.1	73.9	66.4	49.9	55.1
INPO	28.8	82.9	28.9	74.9	64.7	46.3	54.4
ONPO	30.4	83.7	29.9	75.1	65.5	47.8	**55.4**

Figure 3: Model performance on academic benchmarks. [1]

6 Conclusions

6.1 Main Takeaways

ONPO is a novel online RLHF algorithm that can model complex human preferences because it uses the principles of general preference alignment and Nash RLHF. ONPO differs from similar Nash RLHF implementations by achieving SOTA performance across several benchmarks while also greatly reducing the Duality Gap convergence; from a previous SOTA rate of $\mathcal{O}(T^{-1/2})$ to $\mathcal{O}(T^{-1})$.

6.2 Contribution Statement

Nolan Coffey wrote the Abstract, Introduction, and Problem Setup. Ethan Head wrote the Main Results and Experiments section. They collaborated to write the Derivations and Proofs section with Nolan writing the Loss Derivation and Ethan writing the Convergence proof. They further collaborated to write the Conclusion and the Bonus section.

7 Bonus

7.1 Open Questions

As Figure 2 shows, ONPO was applied exclusively to smaller parameter LLMs, with the largest being 8 billion parameters. While performance is strong for these models, it is an open question whether or not performance scales for frontier models that reach into the hundreds of billions of parameters or even over a trillion. Furthermore, it may be the case that the online iterative nature of ONPO where preference datasets D_1 are constructed every step may become prohibitively expensive compared to offline RLHF methods like DPO. Thus it may be necessary for algorithmic improvements due to memory constraints when working with high parameter LLMs. A potential approach could investigate whether the ONPO algorithm could be modified so that the current and auxiliary policies share the majority of their weights (thus greatly reducing the memory footprint), while still maintaining the same convergence and performance.

References

[1] Yuheng Zhang, Dian Yu, Tao Ge, Linfeng Song, Zhichen Zeng, Haitao Mi, Nan Jiang, and Dong Yu. Improving LLM general preference alignment via optimistic online mirror descent. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=kZstGANG8D`.

[2] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[3] Kenneth O May. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22(1):1–13, 1954.

[4] Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative Nash policy optimization: Aligning LLMs with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024.

[5] Dongsheng Ding. LLM alignment (i). Lecture slides, COSC 594: Interplay Between RL and GenAI, 2025.

[6] Dongsheng Ding. LLM alignment (iv). Lecture slides, COSC 594: Interplay Between RL and GenAI, 2025.

[7] Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.