Provable Reinforcement Learning from Human Feedback with an Unknown Link Function

Qining Zhang
University of Michigan, Ann Arbor
qiningz@umich.edu

Lei Ying
University of Michigan, Ann Arbor
leiying@umich.edu

Abstract

Link functions, which characterize how human preferences are generated from the value function of an RL problem, are a crucial component in designing RLHF algorithms. Almost all RLHF algorithms, including state-of-the-art ones in empirical studies such as DPO and PPO, assume the link function is known to the agent (e.g., a logistic function according to the Bradley-Terry model), which is arguably unrealistic considering the complex nature of human preferences. To avoid link function mis-specification, this paper studies general RLHF problems with unknown link functions. We propose a novel policy optimization algorithm called ZSPO based on a new zeroth-order policy optimization method, where the key is to use human preference to construct a parameter update *direction* that is *positively correlated* with the true policy gradient direction. ZSPO achieves it by estimating the sign of the value function difference instead of estimating the gradient from the value function difference, so it does not require knowing the link function. Under mild conditions, ZSPO converges to a stationary policy with a polynomial convergence rate depending on the number of policy iterations and trajectories per iteration. Numerical results also show the superiority of ZSPO under link function mismatch.

1 Introduction

In recent years, reinforcement learning from human feedback (RLHF) has been proposed to avoid the pitfall of reward hacking [1] and delivered empirical success [2, 3, 4]. In RLHF, the agent regularly queries human evaluators for preference feedback on pairs of trajectories and then uses it to infer the quality of the policy. Two main approaches have been studied: (i) reward inference [2, 3] and (ii) direct policy optimization [5, 6]. The first approach recovers a learned reward function from the preferences and then performs standard RL on top of it. The reward function learning step suffers from disadvantages such as reward model overfitting and double problem misspecification [7]. The second approach avoids these drawbacks by optimizing the policy network straight from the preference feedback, which has delivered promising results both theoretically [8] and empirically [5, 9].

Link Function. Almost all RLHF algorithms require knowing the link function, which defines the distribution of human preference for a given value function difference. For example, assuming the Bradley-Terry model [10]. Given the complex nature of humans, it is not adequate to use a simple closed-form equation to characterize the preference mechanism. Contemporary RLHF methods suffer from preference model misspecification, similar to classic RL suffering from reward function misspecification. *For general RL problems, can agents provably learn a good policy to maximize the unknown true reward from human preferences, without knowing the link function?*

Contributions. Inspired by [6], this paper proposes a new policy optimization from human feedback algorithm called ***Z****eroth-Order* ***S****ign* ***P****olicy* ***O****ptimization* (ZSPO), which estimates the *sign* of the value function difference from human feedback, instead of the exact value function difference, for

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

which we do not need the link function expression. Under mild assumptions, we show that ZSPO enjoys the following convergence rate (in terms of the gradient norm) to a stationary policy:

$$\sqrt{d} \cdot \widetilde{\mathcal{O}}\left(\sqrt{\frac{H}{T}} + \max\left\{\frac{1}{\sigma'(0)}, 1\right\}\frac{1}{N^{\frac{1}{4}}} + \sqrt{\varepsilon_D^*}\right),$$

where T is the number of policy iterations, H is the number of planning steps, N is the number of batches for comparison between policy updates, $\sigma'(0)$ characterizes the human evaluators' expertise, and ε_D^* captures their limits of distinguishability. To the best of our knowledge, for utility-based RLHF [11, 12] where the feedback is related to the reward via a link function, ZSPO is the *first* RLHF algorithm with provable guarantees for general MDPs that does not require knowing the link function. The proofs of the main results can be found in the complete version of this paper [13].

2 Preliminary

Episodic RL: We consider an episodic RL instance $\mathcal{M} = (\mathbb{S}, \mathbb{A}, H, \boldsymbol{P}, \boldsymbol{\mu}_0)$, where $\mathbb{S}$ is the state space, $\mathbb{A}$ is the action space, H is the planning horizon, $\boldsymbol{P} = \{\boldsymbol{P}_h\}_{h=1}^H$ is the transition kernels, and $\boldsymbol{\mu}_0$ is the initial distribution. At each episode, the agent chooses a policy $\pi = \{\pi_h : \mathbb{S} \to \mathcal{P}(\mathbb{A})\}_{h=1}^H$ mapping states to probabilities. At each step h, the agent takes an action a_h after observing the state s_h and the environment moves to a new state s_{h+1} without reward feedback. We use $\tau = \{(s_h, a_h)\}_{h=1}^H$ to denote a trajectory and assume the expected return of τ is a function $r(\tau) \in [0, H]$ [14, 6]. For any given policy π, we define the value function $V_1^\pi(s)$ as:

$$V_1^\pi(s) = \mathbb{E}_\pi\left[r(\tau)|\, s_1 = s\right] = \mathbb{E}\left[r(\tau)|\, s_1 = s, \{a_1, \cdots, a_H\} \sim \pi\right].$$

Let the expected value function $\boldsymbol{\mu}_0$ as $V(\pi) = \mathbb{E}_{s\sim\boldsymbol{\mu}_0}[V_1^\pi(s)]$ and assume a parameterized policy network denoted as $\{\pi_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \mathbb{R}^d\}$. Let $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}})$ be the optimal policy parameter.

Preference. The agent has access to preference oracles, e.g., human experts or language models. We call each one of them a *panelist*, and a group of panelists is called a *panel*. In each episode, the agent can choose two batches of trajectories $\mathcal{D}_0 = \{\tau_{0,i}\}_{i=1}^D$ and $\mathcal{D}_1 = \{\tau_{1,i}\}_{i=1}^D$ to query each panelist to obtain a one-bit feedback $o \in \{0, 1\}$. Here, D is the batch size. If $o = 1$, the panelist prefers $\mathcal{D}_1$, and if $o = 0$, the panelist prefers $\mathcal{D}_0$. Specifically, the feedback o is generated by an *unknown* link function $\sigma : \mathbb{R} \to [0, 1]$ of the average reward difference $\bar{r}(\cdot)$ between trajectories:

$$\mathbb{P}(\mathcal{D}_1 \succ \mathcal{D}_0) = \sigma(\bar{r}(\mathcal{D}_1) - \bar{r}(\mathcal{D}_0)) = \sigma\left(\frac{1}{D}\sum_{i=1}^D r(\tau_{i,1}) - \frac{1}{D}\sum_{i=1}^D r(\tau_{i,0})\right), \tag{1}$$

The following assumption characterizes a proper link function [11, 12, 6].

Assumption 1 *The link function $\sigma(\cdot)$ is strictly increasing with $\sigma(0) = 1/2$ and $\sigma(-x) = 1 - \sigma(x)$.*

3 Zeroth-Order Sign Policy Optimization from Human Feedback

In this section, we propose ZSPO to solve RLHF without knowing the link function. The algorithm is summarized in algorithm 1. Two main components are used to build ZSPO: (i) estimate the sign of the value function difference between the current policy $\pi_{\boldsymbol{\theta}_t}$ and the perturbed policy $\pi_{\boldsymbol{\theta}_t'}$, which is controlled by the perturbation distance μ_t at each iteration, and (ii) use the sign of the value function difference to construct a gradient estimator $\hat{\boldsymbol{g}}_t$ that has a positive correlation with the policy gradient $\nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_t})$ in expectation, and then use gradient ascent to find the optimal policy.

Policy Optimization from Signed Feedback. Suppose we have a policy oracle that can compare the value function of $\pi_{\boldsymbol{\theta}_t}$ and $\pi_{\boldsymbol{\theta}_t'}$ and obtain $\mathrm{sign}[V(\pi_{\boldsymbol{\theta}_t'}) - V(\pi_{\boldsymbol{\theta}_t})]$. Then, we can construct the gradient direction estimator $\hat{\boldsymbol{g}}_t$ from the perturbation direction $\boldsymbol{v}_t$ as: $\hat{\boldsymbol{g}}_t = \mathrm{sign}[V(\pi_{\boldsymbol{\theta}_t'}) - V(\pi_{\boldsymbol{\theta}_t})]\boldsymbol{v}_t$. Intuitively, $\hat{g}_t$ aligns with the gradient $\nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_t})$: suppose the perturbation distance μ is small, so under mild conditions, we can linearize the value function difference around the neighborhood of $\boldsymbol{\theta}_t$:

$$V(\pi_{\boldsymbol{\theta}_t'}) - V(\pi_{\boldsymbol{\theta}_t}) \approx \langle \nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_t}), \boldsymbol{\theta}_t' - \boldsymbol{\theta}_t \rangle = \mu \langle \nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_t}), \boldsymbol{v}_t \rangle. \tag{2}$$

Therefore, the sign of the value function difference can be approximated as follows:

$$\mathrm{sign}[V(\pi_{\boldsymbol{\theta}_t'}) - V(\pi_{\boldsymbol{\theta}_t})] \approx \mathrm{sign}[\langle \nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_t}), \boldsymbol{v}_t \rangle]. \tag{3}$$

Algorithm 1 Zeroth-Order Sign Policy Optimization from Human Feedback

Require: initialize the actor-network parameter $\boldsymbol{\theta}_1$, learning rate $\{\alpha_t\}_{t=1}^T$, perturbation distance $\{\mu_t\}_{t=1}^T$, size of trajectory batches D;

for iteration $t = 1 : T$ **do**
  sample a random vector $\boldsymbol{v}_t$ from a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$;
  obtain perturbed parameter $\boldsymbol{\theta}_t' = \boldsymbol{\theta}_t + \mu_t \boldsymbol{v}_t$;
  for $n = 1 : N$ **do**
    sample a batch of D trajectories $\mathcal{D}_{n,0} \sim \pi_{\boldsymbol{\theta}_t}$;
    sample a batch of D trajectories $\mathcal{D}_{n,1} \sim \pi_{\boldsymbol{\theta}_t'}$;
    query a panelist over the two batches $(\mathcal{D}_{n,1}, \mathcal{D}_{n,0})$ and obtain results $o_{t,n}$;
  estimate the gradient direction with a majority vote as: $\hat{\boldsymbol{g}}_t = \operatorname{sign}\left[\sum_{n=1}^N \left(o_{t,n} - \frac{1}{2}\right)\right] \boldsymbol{v}_t$;
  update the actor network $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \hat{\boldsymbol{g}}_t$;

In other words, if the sign of the value function difference is positive, then the perturbation vector $\boldsymbol{v}_t$ is likely to have a positive inner product with the gradient $\nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_t})$. If the sign of the value function difference is negative, $-\boldsymbol{v}_t$ will be positively aligned with the gradient. This positive correlation ensures a convergence dynamic similar to stochastic policy gradient.

Value Function Preference Approximation. The value-function-based preference oracle is usually unrealistic. For example, letting $D = +\infty$ in equation 1 would produce such an oracle, but panelists may not accurately aggregate the return of too many trajectories. Therefore, we use batched trajectory preferences to estimate the value function difference sign with a majority vote rule. Specifically, we ask multiple panelists to compare different pairs of trajectory batches generated from the two policies with a proper batch size. Then, we let the panelists vote on which policy has a higher value function and take the policy with more votes. The majority vote rule helps tackle the unknown link function setting and resembles the preference based on value functions under mild conditions.

4 Main Results

We assume the link function and value functions are regular to perform meaningful optimization:

Assumption 2 *The link function $\sigma(\cdot)$ is L-smooth with $\sigma'(0) > 0$.*

Assumption 3 *The value function $V(\pi_{\boldsymbol{\theta}})$ for the network parameter $\boldsymbol{\theta}$ is L-smooth on $\mathbb{R}^d$.*

If the perturbed parameter $\boldsymbol{\theta}_t'$ is close to $\boldsymbol{\theta}_t$, the value functions will also be close to constitute a more accurate zeroth-order approximation. However, panelists will also have difficulty distinguishing the better policy due to the smaller gap. Let $\varsigma(x) = \sigma(x) - 1/2$ be the (preference) *deviation* function and define the panelist distinguishability as follows:

Definition 1 (Distinguishability) *For any $\mathcal{M}$ and $\varsigma(\cdot)$, define ε_D^* under batch size D to be the maximum constant ε, such that for any two policies π_0 and π_1 with $V(\pi_1) - V(\pi_0) \geq \varepsilon$, we have:*

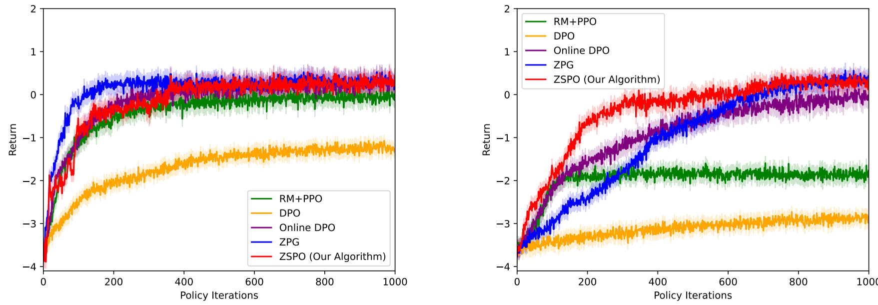
$$\mathbb{E}_{\mathcal{D}_0 \sim \pi_0, \mathcal{D}_1 \sim \pi_1, |\mathcal{D}_0| = |\mathcal{D}_1| = D}\left[\varsigma\left(\bar{r}\left(\mathcal{D}_1\right) - \bar{r}\left(\mathcal{D}_0\right)\right)\right] \geq \frac{1}{2}\varsigma\left(\frac{V(\pi_1) - V(\pi_0)}{2}\right).$$

Proposition 1 *For any $\mathcal{M}$ and $\varsigma(\cdot)$, the distinguishability ε_D^* is upper bounded as $\varepsilon_D^* = \widetilde{\mathcal{O}}(H/\sqrt{D})$.*

When two policies with a value function difference smaller than ε_D^* are compared, the panelists may not distinguish the better policy, which reveals a fundamental limit using human preference feedback. So, to effectively conduct comparisons, we need to control the perturbation distance μ_t. We now characterize the convergence rate of `ZSPO` to an ϵ-stationary policy $\pi_{\boldsymbol{\theta}}$ with $\|\nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}})\|_2 \leq \epsilon$:

Theorem 1 *Choose $\mu_t = \mu$ and $\alpha_t = \Theta(\sqrt{H/dt})$. If we randomly pick R from $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_T\}$ with $\mathbb{P}(\boldsymbol{\theta}_R = \boldsymbol{\theta}_t) = \alpha_t / \sum_{i=1}^T \alpha_i$, then the convergence rate of `ZSPO` satisfies:*

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} V(\pi_{\boldsymbol{\theta}_R})\|_2\right] = \widetilde{\mathcal{O}}\left(\left[\sqrt{\frac{Hd}{T}} + \mu\right] + \frac{\varepsilon_D^*}{\mu} + \frac{1}{\mu}\varsigma^{-1}\left(\sqrt{\frac{2}{N}}\right)\right).$$



(a) Bradley-Terry Preference Model (b) Linear Preference Model

Figure 1: GridWorld: (a) comparison of ZSPO and baselines without link function mismatch, and (b) comparison of ZSPO and baselines with link function mismatch.

Insights. The convergence rate of ZSPO has three components: the convergence rate of zeroth order optimization, the panelist distinguishability ε_D^*, and the majority vote approximation error. The first term resembles zeroth-order stochastic gradient descent [15], stochastic coordinate descent [16], and sign gradient descent [17]. If we choose $\mu = \mathcal{O}(1/\sqrt{dT})$ as in the literature, this term matches the state-of-the-art $\mathcal{O}(\sqrt{d/T})$ result for non-convex smooth function optimization [15, 17]. The second term comes from the distinguishability limit of panelists: when the current policy $\boldsymbol{\theta}_t$ is close to stationary, i.e., the gradient norm is smaller than ε_D^*/μ, the perturbed policy and the current policy have similar value functions with difference smaller than ε_D^* according to equation 2, which becomes indistinguishable. One could also view the parameter $\boldsymbol{\theta}_R$ learned by ZSPO as the policy most preferred by panelists in the ε_D^*-neighborhood of a stationary policy. The third term comes from approximating the expected preference probability with a majority vote. As the number of batches N increases, the approximation error would decrease since $\varsigma^{-1}(\sqrt{2/N}) \to \varsigma^{-1}(0) = 0$, since the majority vote becomes more accurate and reflects the population-level preference. The best perturbation distance satisfies $\mu^2 = \Theta(d^{-1}\max\{1/\sqrt{N}, H/\sqrt{D}\})$, and we obtain the rate shown in introduction.

Panelist Quality. The result depends on the preference model, i.e., the deviation function $\varsigma(\cdot)$, which constitutes the majority vote error. If the panelists are better trained to distinguish candidates with similar average returns, $\varsigma(\cdot)$ is closer to a step function with a larger derivative at the origin. Then, the majority vote error will decrease faster, resulting in a better convergence rate. On the other hand, for the same pair of trajectories, we can also require multiple panelists to provide preferences and then aggregate the results via a majority vote. This is equivalent to querying a better-trained panelist with a more step-like deviation function, and a better convergence rate is anticipated.

5 Experimental Evaluations

We demonstrate the empirical performance of ZSPO in a stochastic GridWorld environment in Fig.1. We used different unknown link functions (logistic and linear [18, 11]) to generate preferences, and considered four baselines algorithms: (1) RM+PPO [3], (2) DPO [9], (3) Online DPO [19, 20], and (4) ZPG [6], assuming the link function is logistic. All algorithms collect $N = 1000$ trajectories between policy updates, and each is evaluated by 100 panelists. It is shown that without link function mismatch, ZSPO has almost the same performance as the best baselines, and if there exists a link function mismatch, ZSPO is more robust compared to the baselines and converges much quicker.

6 Conclusion

In this paper, we studied RLHF where the link function for the preference model is unknown. We developed a policy-optimization-based algorithm called ZSPO based on zeroth-order optimization, where the sign of the value function difference is estimated directly from human feedback instead of the full function difference. We showed that ZSPO has a provable convergence guarantee with polynomial sample and human-query complexities, which is also validated by numerical experiments.

References

[1] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471. Curran Associates, Inc., 2022.

[2] Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[4] Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*, 2024.

[5] 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, pages 53728–53741. Curran Associates, Inc., 2023.

[6] Qining Zhang and Lei Ying. Zeroth-order policy gradient for reinforcement learning from human feedback without reward inference. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando Ramirez, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Machine Learning Research (TMLR)*, 2023.

[8] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[9] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function. In *First Conference on Language Modeling*, 2024.

[10] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[11] Viktor Bengs, Robert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hullermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

[12] Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023.

[13] Qining Zhang and Lei Ying. Provable reinforcement learning from human feedback with an unknown link function. *arXiv preprint arXiv:2506.03066*, 2025.

[14] Qining Zhang, Honghao Wei, and Lei Ying. Reinforcement learning from human feedback without reward inference: Model-free algorithm and instance-dependent analysis. *Reinforcement Learning Journal*, 2024.

[15] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, apr 2017.

[16] Hanqin Cai, Yuchen Lou, Daniel Mckenzie, and Wotao Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1193–1203. PMLR, 18–24 Jul 2021.

[17] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

[18] Bangrui Chen and Peter I. Frazier. Dueling bandits with weak regret. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 731–739. PMLR, 06–11 Aug 2017.

[19] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

[20] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.