Efficient Preference Alignment via Pareto Exploration (Student Abstract)

Pengfei Liu, Rui Kong, Zongzhang Zhang*

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
School of Artificial Intelligence, Nanjing University, Nanjing 210023, China
{liupf, kongr}@lamda.nju.edu.cn, zzzhang@nju.edu.cn

Abstract

Hand-craft reward engineering requires domain knowledge with numerous trials and errors, while Preference-based Reinforcement Learning (PbRL) avoids manual reward design but often suffers from limited interpretability and unstable training. To address these issues, we propose a novel preference alignment framework. Our approach leverages large language models to generate sub-reward functions informed by prior knowledge and further align human preferences by optimizing the weights combining these sub-rewards. For policy learning, we introduce Policy Optimization via Pareto Regularization (POPR) which regularizes updates along Pareto-optimal directions. Experiments show that our framework improves reward quality and policy stability, achieving superior performance to expert-designed rewards across most tasks.

Introduction

Reinforcement Learning (RL) seeks a policy that maximizes expected cumulative reward for a task. A core difficulty is designing a reward function that accurately captures the task objective and reliably guides learning. In locomotion and robotic manipulation, crafting such rewards is time-consuming and requires repeated hand-tuning of component weights by experts. This trial-and-error process is computationally expensive and limited by the designer's expertise.

PbRL circumvents the need for manual reward engineering by learning a parameterized reward model directly from human preferences, and has demonstrated effectiveness across a variety of tasks (Christiano et al. 2017; Zhang et al. 2024). However, implicit reward models are hard to interpret and the co-adaptation between the reward model and the policy is prone to brittle, unstable learning dynamics. We argue these issues stem in part from ignoring the inherent structure of most control tasks, which require balancing multiple objectives (e.g., efficiency, stability, energy). The optimal trade-offs among these objectives naturally form a Pareto front.

Instead of learning a black-box reward model, we propose to exploit this multi-objective structure. We introduce a novel preference alignment framework which models the overall reward as a composition of interpretable sub-rewards and leverages the geometric efficiency of the Pareto front to regularize policy optimization, leading to a more stable and direct search for the user-preferred policy. Concretely, the method first obtains initial optimal policy for uniformly weighted sub-rewards, then fits reward weights to human preference data, and finally employs a POPR step to efficiently locate the policy on the frontier that aligns with those preferences. Compared to PbRL, our approach is more interpretable and easier to optimize because it tunes intuitive sub-reward weights rather than opaque reward parameters. Compared to manual reward engineering, using POPR in place of repeated RL retraining reduces computation time and yields more stable, progressively optimized solutions.

*Corresponding author: Zongzhang Zhang.

Method

Our framework enables efficient preference alignment through two components: low-dimensional preference fitting and Pareto-regularized policy optimization.

Preference Alignment via Weight Optimization

We cast preference alignment as a low-dimensional optimization problem. Rather than learning a high-dimensional, monolithic reward model, we represent the overall reward as a convex combination of m interpretable sub-rewards: $r_{\boldsymbol{\omega}} = \sum_{i=1}^{m} \omega_i r_i$, where the sub-reward functions $\{r_1, \ldots, r_m\}$ are automatically generated using a Large Language Model (LLM) to incorporate domain knowledge. The weights $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_m]$ are constrained to the probability simplex (i.e., $\sum_i \omega_i = 1, \omega_i \geq 0$). This formulation reduces reward learning to the efficient estimation of a low-dimensional vector $\boldsymbol{\omega}$. We fit $\boldsymbol{\omega}$ to human preference data using standard PbRL techniques. Restricting the search to this simplex yields a compact, transparent representation. It facilitates faster convergence and makes human inspection easier compared to unconstrained, high-dimensional reward models.

Pareto-Regularized Policy Optimization

Once a new weight vector $\boldsymbol{\omega}$ is obtained, the agent must efficiently transfer its existing policy to one that is optimal for the new composite reward. We frame this transfer as a Multi-Objective Optimization (MOO) problem:

$$\max_{\boldsymbol{\theta}} \boldsymbol{F}(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), f_2(\boldsymbol{\theta}), \cdots, f_m(\boldsymbol{\theta})], \tag{1}$$

where $\boldsymbol{\theta}$ represents the parameters of policy $\pi_{\boldsymbol{\theta}}$ and $f_i(\boldsymbol{\theta})$ represents the expected return under the i-th sub-reward. The optimal policies for all possible weight combinations collectively form an efficient set in the parameter space known as the Pareto front. Naively retraining with a standard policy gradient can be inefficient: it might move the policy away from the Pareto front, leading to unstable convergence.

Our key contribution is the novel integration of Continuous Pareto Exploration (CPE) from the MOO field (Ma, Du, and Matusik 2020) into the PbRL framework, addressing the aforementioned issues of poor policy adaptation and training instability. CPE efficiently computes the tangent direction to the Pareto front at the current solution's location, representing a "safe" update path that avoids performance degradation on any sub-objective. Further details regarding this methodological migration are provided in the appendix[1]. Based on this, we propose POPR. The policy is optimized by first calculating a composite gradient g_{total} according to the following rule:

$$g_{\text{total}} = \nabla_{\boldsymbol{\theta}} J_\pi(\boldsymbol{\theta}) - \lambda_{\text{pareto}} \text{ParetoExplore}(\boldsymbol{\theta}), \qquad (2)$$

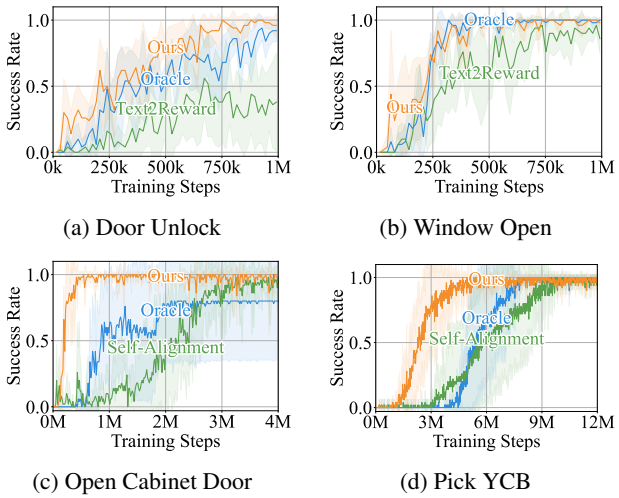
where the first term $\nabla_{\boldsymbol{\theta}} J_\pi(\boldsymbol{\theta})$ is the policy loss gradient from a standard policy optimization algorithm (e.g., SAC or PPO), guiding the policy toward the new preference target. The second term $\text{ParetoExplore}(\boldsymbol{\theta})$ is the CPE-provided tangent direction to regularize the update remains aligned with the Pareto front. The coefficient λ_{pareto} controls the strength of this regularization. By applying gradient descent with g_{total}, POPR ensures the policy transitions smoothly along the Pareto front, achieving rapid and stable adaptation.

Our complete framework operates by iterating between these two components. First, we fit the weights $\boldsymbol{\omega}$ to reflect the current human preferences. Then, we use POPR to efficiently update the policy to be optimal for these new weights. This iterative cycle is particularly well-suited for real-world scenarios where human preferences may evolve or require refinement throughout the training process.

Experiments

We evaluate our framework in two challenging robotic manipulation environments: Meta-World (Yu et al. 2020) and ManiSkill2 (Gu et al. 2023). Our method is compared against three baselines: a policy trained with expert-designed oracle rewards, Text2Reward (Xie et al. 2024), and a fully LLM-based PbRL approach, Self-Alignment (Zeng, Mu, and Shao 2024). We use Success Rate as the primary evaluation metric because it directly reflects the policy's learning quality at different training stages. Detailed experimental settings are provided in the appendix.

Meta-World On the two Meta-World tasks Door Unlock and Window Open, our method converges substantially faster and with greater stability than Text2Reward, while closely matching the oracle reward's final success rates. As shown in Figs. 1a and 1b, the learning curves show a steep early rise for our approach, a result of high reward quality that guides the policy effectively from the start. Our method also demonstrates excellent policy stability by stably converging to the oracle's final performance. We attribute this improved sample efficiency and stability to our use of low-dimensional preference fitting and policy updates via POPR.



(a) Door Unlock (b) Window Open

(c) Open Cabinet Door (d) Pick YCB

Figure 1: Learning curves on tasks from Meta-World and ManiSkill2. All results are obtained from five runs with different random seeds.

ManiSkill2 In the more complex ManiSkill2 environment, our approach produces smoother, more consistent training over long horizons and achieves higher final performance than the baselines. As shown in Figs. 1c and 1d, Self-Alignment shows slow, noisy progress and frequent plateaus, whereas our method yields steady, monotonic improvement and even surpasses the oracle reward on the Open Cabinet Door task. These results highlight our method's effectiveness in complex, high-dimensional tasks, where constraining policy updates to Pareto-optimal directions improves convergence reliability and final performance.

Conclusion

This paper introduced a novel framework for efficient preference alignment that addresses the common challenges of unstable training and poor interpretability in PbRL. By decomposing the reward function into a weighted combination of interpretable sub-rewards generated by LLM, our method simplifies the preference alignment process to a low-dimensional weight optimization problem. We also proposed POPR, which leverages CPE to guide policy updates along the Pareto front, ensuring stable and efficient adaptation to new preferences. Our experiments in complex robotic manipulation tasks confirm that this approach improves reward quality and policy stability, achieving performance that surpasses other methods.

Acknowledgments

This work is supported by the National Science Foundation of China (No. 62276126).

[1]https://www.lamda.nju.edu.cn/liupf/files/POPRAppendix.pdf

References

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *NIPS*, 4302–4310.

Gu, J.; Xiang, F.; Li, X.; Ling, Z.; Liu, X.; Mu, T.; Tang, Y.; Tao, S.; Wei, X.; Yao, Y.; et al. 2023. ManiSkill2: A unified benchmark for generalizable manipulation skills. In *ICLR*.

Ma, P.; Du, T.; and Matusik, W. 2020. Efficient continuous pareto exploration in multi-task learning. In *ICML*, 6522–6531.

Xie, T.; Zhao, S.; Wu, C. H.; Liu, Y.; Luo, Q.; Zhong, V.; Yang, Y.; and Yu, T. 2024. Text2Reward: Reward shaping with language models for reinforcement learning. In *ICLR*.

Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; and Levine, S. 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 1094–1100.

Zeng, Y.; Mu, Y.; and Shao, L. 2024. Learning reward for robot skills using large language models via self-alignment. In *ICML*, 58366–58386.

Zhang, Z.; Sun, Y.; Ye, J.; Liu, T.-S.; Zhang, J.; and Yu, Y. 2024. Flow to better: Offline preference-based reinforcement learning via preferred trajectory generation. In *ICLR*.