Can RLHF be More Efficient with Imperfect Reward Models? A Policy Coverage Perspective

Jiawei Huang [1] Bingcong Li [1] Christoph Dann[2] Niao He [1]
Contact Email: jiawei.huang@inf.ethz.ch
[1]ETH Zurich [2]Google Research

ICML International Conference On Machine Learning

ETH zürich

Google

Paper Link

Motivation

- Sample efficiency (# of human annotations) is crucial in online RLHF.
- Previous works focus on strategic exploration, while we study from a underexplored perspective—transfer learning.
- Rich scenarios for transfer learning:
  - Reward models (RMs) from relevant tasks
  - Easy-to-access metric other than human feedback
  - Guidance from advanced LLMs

Key Question: How to improve sample efficiency in online RLHF by leveraging those imperfect source RMs?

Setting and Assumptions

Standard Contextual Bandit Framework

- $\mathcal{S}$: prompt space; $\mathcal{A}$: response space,
- $\pi : \mathcal{S} \to \Delta(\mathcal{A})$: LLM as a policy. W.L.O.G., $\pi(\cdot|\cdot) > 0$ everywhere.
- r^*: unknown true (human intrinsic) RM,
- Learning objective:

$$\pi^*_{r^*} \leftarrow \arg\max_{\pi} J_\beta(\pi) := \mathbb{E}_{s\sim\rho,a\sim\pi}[r^*(s,a)] - \beta\mathbb{KL}(\pi\|\pi_{\texttt{ref}}), \tag{1}$$

with ρ as prompt distribution, $\pi_{\texttt{ref}}$ as the reference policy, and yields:

$$\pi^*_{r^*}(a|s) \propto \pi_{\texttt{ref}}(a|s)\exp(\frac{r^*(s,a)}{\beta}), \tag{2}$$

- Bradley-Terry preference model

$$\mathbb{P}_{r^*}(\mathbb{I}[a \succ \widetilde{a}]|s,a,\widetilde{a}) = \text{Sigmoid}(r(s,a) - r(s,\widetilde{a})).$$

Standard Assumptions

- **Bounded rewards**: $r^* \in [0, R]$,
- **Function approximation**: A policy class Π is available.
  (i) $\pi^*_{r^*} \in \Pi$. **(ii)** $\forall\pi \in \Pi$, $\|\log\frac{\pi}{\pi_{\texttt{ref}}}\|_\infty \le \frac{R}{\beta}$.

Online RLHF with Reward Transfer

Besides human feedback, we access to W source RMs $r^1, ..., r^W \in [0, R]$:

- no prior knowledge on their quality
- due to Eq. (2), any LLM policy can be converted as a RM.

Transfer RL has been studied for decades.
But the KL-regularization in Eq. (1) makes something different!

Blessing of Regularization: A Policy Coverage Perspective

Policy Coverage: The coverage coefficient of policy $\widetilde{\pi}$ by another policy π:

$$\text{Cov}^{\widetilde{\pi}|\pi} := \mathbb{E}_{s\sim\rho,a\sim\widetilde{\pi}}[\frac{\widetilde{\pi}(a|s)}{\pi(a|s)}].$$

Why Policy Coverage Perspective?

- Fundamental complexity measure in online [1] and offline [2] RLHF.
- Optimization/exploration on policy (LLM) space is more efficient in RLHF.

Key Lemma: special structure due to KL regularization ($\beta > 0$)

Lemma 3.1: For any $\pi \in \text{conv}(\Pi) \cup \{\pi^*_{r^w}\}_{w=1}^W$,

$$\text{Cov}^{\pi^*_{r^*}|\pi} = 1 + O(1)\cdot\frac{J_\beta(\pi^*_{r^*}) - J_\beta(\pi)}{\beta}.$$

Interpretation

- $\text{Cov}^{\pi^*_{r^*}|\pi}$ can be identified by π's value gap
  - vastly distinguished from pure reward maximization
- KL-reg "reconciles" exploration and exploitation
  - exploiting policies with high policy value coincides with exploration!

Transfer Policy Optimization (TPO): Provably Efficient Online Transfer Learning

Main Idea in TPO Algorithm Design: transfer from π with the lowest $\text{Cov}^{\pi^*_{r^*}|\pi}$

- **Principle 1**: Transfer from the source policy with the highest policy value.
- **Principle 2**: "Self-Transfer Learning": treat the offline policy distilled from collected data as a transfer candidate (see paper for details).

Theorem 4.4: In sharp constrast to $\widetilde{O}(\sqrt{\text{Complx}(\Pi)\cdot T})$ regret in previous works, our TPO achieves:

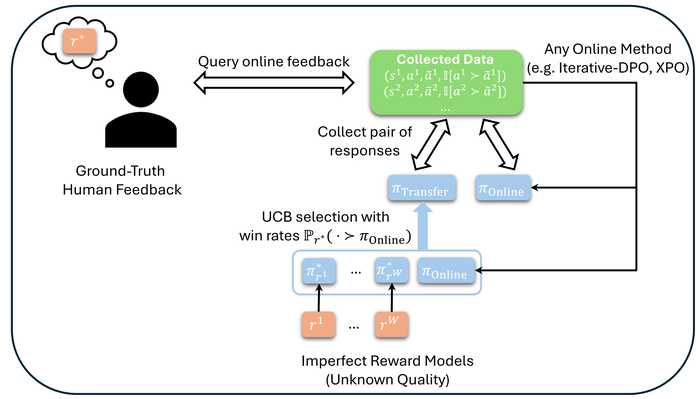
- $\widetilde{O}(\sqrt{WT})$, when T is small
  — Reduce dependence on complexity of Π to W
- $\widetilde{O}(\sqrt{T})$, when T is large enough
  — No dependence on complexity of Π

Empirical TPO: From Theory to Practice

- Estimating policy value can be computationally expensive.
- Is there a more accessible indicator for $\text{Cov}^{\pi^*_{r^*}|\cdot}$? Yes, the win rates!
- Lemma 5.1 [Informal] A lower bound for $\text{Cov}^{\pi^*_{r^*}|\pi}$:

$$\text{Cov}^{\pi^*_{r^*}|\pi} \ge \max_{\gamma>0}\frac{1}{\sqrt{\gamma + 2\cdot\mathbb{P}_{r^*}(\pi \succ \pi^*_{r^*}))\log\frac{1+\gamma}{\gamma}}} \tag{3}$$

- Empirical algorithm design



Main Idea: *"Transfer from the expert until beat it"*
- $\pi^*_{r^*}$ is unknown, but the online learning policy π_{Online} converges to it.
- Transfer from $\arg\max_\pi \mathbb{P}_{r^*}(\pi \succ \pi_{\text{Online}})$.
- Formulate as a multi-armed bandit problem since win rates are unknown.
- Scalable in practice!

Experiments in Summarization Tasks with T5

- Fine-tuning T5-small (60M) on XSum dataset.
- 4 source RMs:
  - 2 metrics of similarity with human summary: (i) ROUGE (ii) BERTScore
  - 2 advanced LLMs: (iii) T5-Base (250M) (iv) T5-Large (770M)
- Llama3-8B to simulate human feedback.

	Without Transfer	Purely Exploit ROUGE	Purely Exploit T5-Large
Iter 1	52.1 ± 1.2	53.1 ± 1.1	49.5 ± 0.9
Iter 2	53.3 ± 1.6	54.5 ± 1.3	49.1 ± 0.4
Iter 3	54.0 ± 1.2	53.3 ± 1.5	50.6 ± 0.3

Table 1. Win rates (%) of the policies trained by empirical TPO competed with 3 baselines.

References

[1] Xie et al., Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf.
[2] Liu et al., Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer.