Policy-labeled Preference Learning: Is Preference Enough for RLHF?

Taehyun Cho[1,†] Seokhun Ju[1,†] Seungyub Han[1]
Dohyeong Kim[1] Kyungjae Lee[2] Jungwoo Lee[1]

[1]Seoul National University, [2]Korea University
[†]Equal Contribution

ICML (spotlight)
June 20, 2025

Introduction

- Reinforcement Learning from Human Feedback (RLHF) enables agents to align with human goals using human preferences.

Introduction

- Reinforcement Learning from Human Feedback (RLHF) enables agents to align with human goals using human preferences.
- Existing RLHF methods often assume as if trajectories are generated by optimal policies π^*.

Introduction

- Reinforcement Learning from Human Feedback (RLHF) enables agents to align with human goals using human preferences.
- Existing RLHF methods often assume as if trajectories are generated by optimal policies π^*.
- This leads to **likelihood mismatch** in offline settings due to environmental stochasticity and diverse behavior policies.

Introduction

- Reinforcement Learning from Human Feedback (RLHF) enables agents to align with human goals using human preferences.
- Existing RLHF methods often assume as if trajectories are generated by optimal policies π^*.
- This leads to **likelihood mismatch** in offline settings due to environmental stochasticity and diverse behavior policies.
- Direct Preference Optimization (DPO) removes the need for explicit rewards, but fails to address this mismatch.

Our Contributions

1. We propose **Policy-labeled Preference Learning (PPL)**, a regret-based framework for RLHF, which explicitly models the **behavior policy** associated with preference data.
2. We introduce **contrastive KL regularization** to correct for likelihood mismatch.
3. PPL shows superior performance on MetaWorld offline tasks and is competitive in online RLHF.

Score-based Preference Model

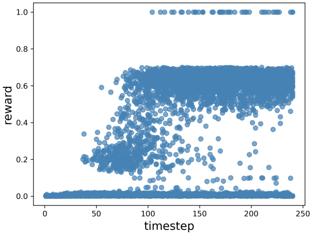
Table 1: Comparison for different preference models under PbRL framework.

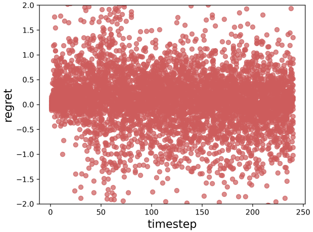
Algorithm	Score Function	Direct Preference Optimization	Likelihood Matching
PEBBLE (Lee et al., 2021)	$r_\psi(s_t, a_t)$	✗	✗
DPO (Rafailov et al., 2024b)	$\log \pi_\psi(y\|s)/\pi_{\text{ref}}(y\|s)$	✓	✗
DPPO (An et al., 2023)	$-\mathbb{E}_{a\sim\pi_\psi(\cdot\|s_t)}[\|\|a - a_t\|\|_2]$	✓	✗
CPL (Hejna et al., 2023)	$Q^{\pi_\psi}(s_t, a_t) - V^{\pi_\psi}(s_t)$	✓	✗
PPL [Ours]	$-(V^{\pi_\psi}(s_t) - Q^{\pi}(s_t, a_t))$	✓	✓

Model prediction: $P_{S_\psi}[\zeta^+ \succ \zeta^-] = \sigma\left(\sum_{t\geq 0} S_\psi(s_t^+, a_t^+) - S_\psi(s_t^-, a_t^-)\right),$

Loss function: $\mathcal{L}(S_\psi; \mathcal{D}) = -\mathbb{E}_{(\zeta^+,\zeta^-)\sim\mathcal{D}}\left[\log P_{S_\psi}[\zeta^+ \succ \zeta^-]\right]$

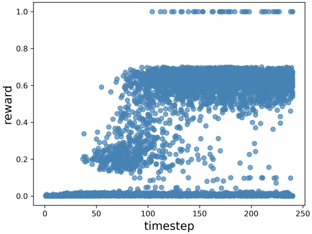
Reward vs Regret

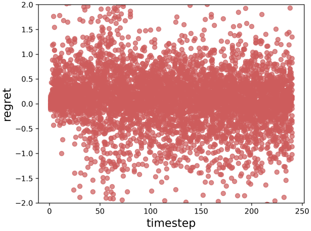




- **Reward :** Sparse and delayed feedback
- **Negative Regret :** Dense and stepwise feedback

Reward vs Regret





Negative Regret

- Performance difference between the behavior policy π and the optimal policy π^*
- $-\mathrm{Reg}^{\pi}_{\pi^*}(s, a) = Q^{\pi}(s, a) - V^{\pi^*}(s)$

Likelihood Mismatch

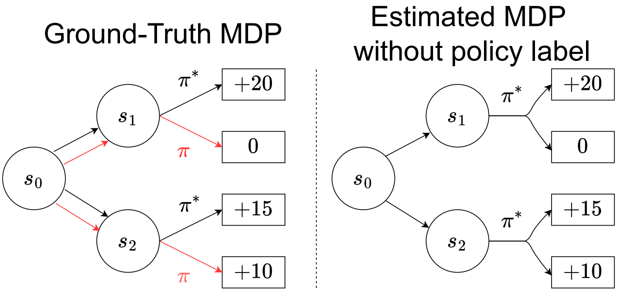
From a perspective of regret, existing RLHF/DPO disregards the source of the trajectories, implicitly treating all trajectories as if they were generated by the optimal policy.

Likelihood Mismatch

From a perspective of regret, existing RLHF/DPO disregards the source of the trajectories, implicitly treating all trajectories as if they were generated by the optimal policy.

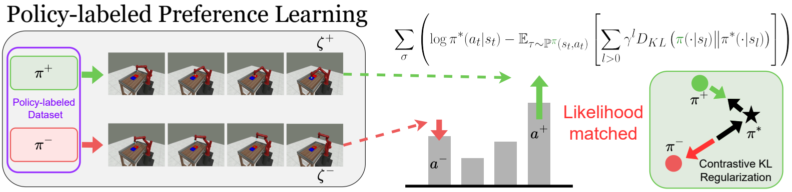
"What impact does this assumption – treating all behavior polices as optimal – have on the regret-based learning process?"

Likelihood Mismatch



- Offline data from π is misinterpreted as from π^*.
- Leads to erroneous preference modeling and degraded performance.

Policy-labeled Preference Learning



- Prior works like CPL use optimal advantage $Q^{\pi^*}(s,a) - V^{\pi^*}(s)$ as preference score.
- But this assumes all data comes from π^*, ignoring suboptimal behavior policies.
- PPL instead uses **negative regret**: $Q^{\pi}(s,a) - V^{\pi^*}(s)$, which incorporates behavior policy.

Sequential Forward KL Divergence

Theorem (Policy Deviation Theorem)

If a policy π^* is α-optimal, then for any policy π,

$$Q_*^{\pi^*}(s,a) - Q_*^{\pi}(s,a) = \alpha \bar{D}_{KL}(\pi || \pi^*; s, a)$$

*where the **sequential forward KL divergence** is defined as*

$$\bar{D}_{KL}(\pi || \pi'; s, a) := \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{s,a}} \left[\sum_{l>0} \gamma^l D_{KL}(\pi(\cdot|s_l) || \pi'(\cdot|s_l)) \right].$$

Here, $\mathbb{P}^{\pi}_{s,a}$ is the distribution of trajectories $\tau = (s_0, a_0, \cdots, s_l, a_l, \cdots)$ generated by policy π and the transition $\mathbb{P}$, starting at $(s_0, a_0) = (s, a)$.

Sequential Forward KL Divergence

Now we can derive the (negative) regret into policy expression,

$$
\begin{aligned}
-\mathrm{Reg}^{\pi}_{\pi^*}(s_t, a_t) &:= -\underbrace{V^{\pi^*}(s_t)}_{\text{expected return under } \pi^*} + \underbrace{Q^{\pi}(s_t, a_t)}_{\text{achieved return under } \pi} \\
&= \alpha\Big(\underbrace{\log \pi^*(a_t|s_t)}_{\text{increase likelihood}} - \underbrace{\bar{D}_{\mathrm{KL}}\big(\pi||\pi^*; s_t, a_t\big)}_{\text{decrease sequential forward KL}}\Big).
\end{aligned}
$$

Policy-labeled Preference Learning

$$\mathcal{L}_{\mathsf{PPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(-\sum_{t\geq 0} \mathrm{Reg}^{\pi^+}_{\pi_\psi}(s_t^+, a_t^+) - \mathrm{Reg}^{\pi^-}_{\pi_\psi}(s_t^-, a_t^-)\right)\right]$$

Policy-labeled Preference Learning

$$\mathcal{L}_{\mathsf{PPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(-\sum_{t\geq 0}\mathrm{Reg}^{\pi^+}_{\pi_\psi}(s^+_t, a^+_t) - \mathrm{Reg}^{\pi^-}_{\pi_\psi}(s^-_t, a^-_t)\right)\right]$$

Substitute $-\mathrm{Reg}^{\pi}_{\pi^*}(s_t, a_t) = \alpha\Big(\log\pi^*(a_t|s_t) - \bar{D}_{\mathsf{KL}}\big(\pi||\pi^*; s_t, a_t\big)\Big)$

Policy-labeled Preference Learning

$$\mathcal{L}_{\mathrm{PPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(-\sum_{t\geq 0} \mathrm{Reg}_{\pi_\psi}^{\pi^+}(s_t^+, a_t^+) - \mathrm{Reg}_{\pi_\psi}^{\pi^-}(s_t^-, a_t^-)\right)\right]$$

Substitute $-\mathrm{Reg}_{\pi^*}^{\pi}(s_t, a_t) = \alpha\left(\log \pi^*(a_t|s_t) - \bar{D}_{\mathrm{KL}}(\pi||\pi^*; s_t, a_t)\right)$

$$-\sum_{t\geq 0} \mathrm{Reg}_{\pi_\psi}^{\pi^+}(s_t^+, a_t^+) - \mathrm{Reg}_{\pi_\psi}^{\pi^-}(s_t^-, a_t^-)$$

$$= \alpha \sum_{t\geq 0}\left(\log \frac{\pi_\psi(a_t^+|s_t^+)}{\pi_\psi(a_t^-|s_t^-)} \underbrace{-\bar{D}_{KL}(\pi^+||\pi_\psi; s_t^+, a_t^+) + \bar{D}_{KL}(\pi^-||\pi_\psi; s_t^-, a_t^-)}_{\text{contrastive KL regularization } \mathcal{R}(\pi_\psi;\ \pi^+,\pi^-)}\right)$$

Contrastive KL Regularization

$$R(\pi_\psi; \pi^+, \pi^-) := -\bar{D}_{KL}(\pi^+ || \pi_\psi; s_t^+, a_t^+) + \bar{D}_{KL}(\pi^- || \pi_\psi; s_t^-, a_t^-)$$
$$\approx \frac{1}{L} \sum_{l=1}^{L} \left[\log \frac{\pi^+(a_{t+l}^+ | s_{t+l}^+)}{\pi_\psi(a_{t+l}^+ | s_{t+l}^+)} - \log \frac{\pi^-(a_{t+l}^- | s_{t+l}^-)}{\pi_\psi(a_{t+l}^- | s_{t+l}^-)} \right]$$

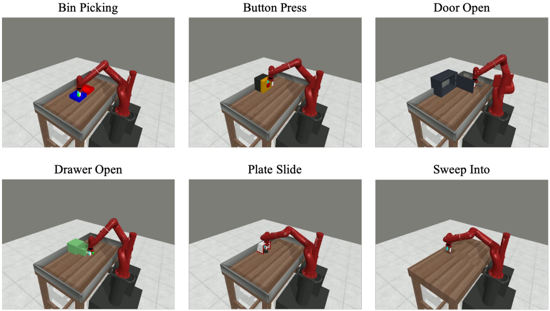
- Encourages π_ψ to align with preferred policy π^+ and diverge from less preferred π^-.
- Approximates into L-horizon undiscounted sum with sampled segments $\{s_t^+, a_t^+\} \sim \zeta^+$ and $\{s_t^-, a_t^-\} \sim \zeta^-$
- Mitigates likelihood mismatch over sequential rollouts.

Deterministic Pseudo Labeling

$$S_{\text{PPL-d}}(\pi_\psi; \zeta^+) - S_{\text{PPL-d}}(\pi_\psi; \zeta^-) =$$
$$\sum_{t \geq 0} \left[\log \frac{\pi_\psi(a_t^+ \mid s_t^+)}{\pi_\psi(a_t^- \mid s_t^-)} + \frac{1}{L} \sum_{l=1}^{L} \log \frac{\pi_\psi(a_{t+l}^+ \mid s_{t+l}^+)}{\pi_\psi(a_{t+l}^- \mid s_{t+l}^-)} \right]$$

- Behavior policy is typically unknown in offline setting.
- Assign a pseudo label as if each segment is generated by deterministic policy.

Experiments



- 1. Is PPL robust when learning from heterogeneous datasets that include suboptimal data?
- 2. Does incorporating policy labels lead to improved performance?
- 3. Can PPL be applied effectively in an online setting?

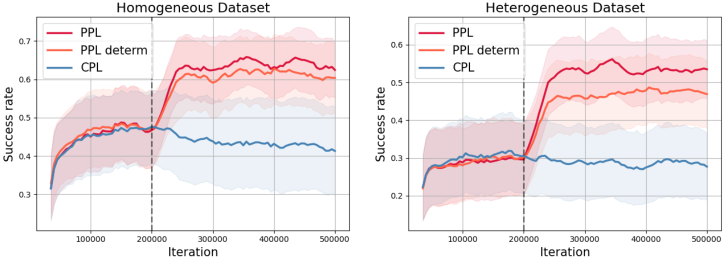
Offline MetaWorld Results

Table 2: Success rates of all methods across six tasks on the MetaWorld benchmark on different datasets. Each score is reported with the maximum average performance across four seeds over 200 episode evaluation window.

		Bin Picking	Button Press	Door Open	Drawer Open	Plate Slide	Sweep Into
Homogeneous Dense	SFT	39.7 ± 19.2	71.5 ± 3.3	**48.0 ± 15.6**	56.2 ± 1.8	64.8 ± 0.8	70.0 ± 6.5
	P-IQL	62.0 ± 4.4	72.3 ± 1.0	47.7 ± 5.1	58.0 ± 5.7	**70.5 ± 6.1**	65.8 ± 1.3
	CPL	22.7 ± 5.5	64.3 ± 1.4	29.0 ± 4.3	54.0 ± 4.3	65.5 ± 3.1	69.8 ± 3.3
	PPL	**83.5 ± 4.4**	**79.8 ± 4.8**	39.3 ± 2.0	**69.2 ± 5.5**	64.7 ± 2.0	**72.8 ± 4.8**
Homogenous Sparse	SFT	33.5 ± 5.4	67.4 ± 1.5	31.3 ± 2.1	54.9 ± 2.7	67.1 ± 3.7	78.3 ± 2.5
	P-IQL	72.4 ± 6.6	74.5 ± 0.0	**58.5 ± 1.4**	51.4 ± 4.6	**76.3 ± 1.6**	**79.0 ± 2.6**
	CPL	26.5 ± 1.0	63.7 ± 1.3	28.5 ± 5.8	50.1 ± 4.5	65.1 ± 2.8	72.9 ± 6.1
	PPL	**87.2 ± 3.5**	**87.3 ± 2.8**	49.3 ± 6.5	**68.5 ± 5.3**	64.0 ± 6.4	73.9 ± 3.5
Heterogeneous Dense	SFT	18.5 ± 23.8	63.7 ± 12.2	26.0 ± 12.5	32.0 ± 5.7	62.8 ± 1.6	53.0 ± 9.1
	P-IQL	51.2 ± 5.3	62.5 ± 4.9	**32.0 ± 3.5**	41.8 ± 3.8	67.0 ± 3.0	**59.3 ± 3.7**
	CPL	1.2 ± 0.8	49.7 ± 3.0	17.3 ± 2.5	26.0 ± 2.2	59.2 ± 7.7	51.2 ± 3.0
	PPL	**59.7 ± 18.6**	**73.8 ± 3.3**	25.8 ± 2.0	**58.5 ± 3.8**	**69.8 ± 2.3**	57.3 ± 8.6
Heterogeneous Sparse	SFT	12.2 ± 1.0	63.7 ± 4.7	17.8 ± 0.8	38.7 ± 3.0	70.7 ± 3.8	60.7 ± 2.5
	P-IQL	48.0 ± 5.6	71.0 ± 6.6	**44.1 ± 3.2**	47.5 ± 3.0	**72.0 ± 4.0**	**64.3 ± 1.0**
	CPL	18.0 ± 6.1	50.8 ± 0.8	18.5 ± 3.0	32.1 ± 1.6	67.3 ± 5.5	55.5 ± 3.3
	PPL	**83.8 ± 3.8**	**83.5 ± 1.8**	34.3 ± 7.6	**60.8 ± 7.3**	71.2 ± 1.9	63.3 ± 4.2

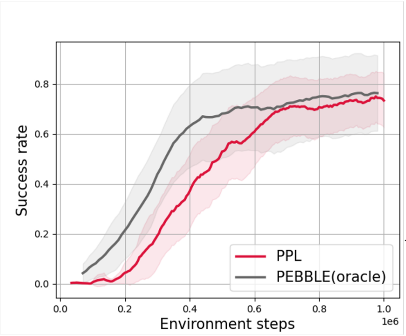
- PPL outperforms CPL and P-IQL especially in sparse and heterogeneous settings.
- Robust across 6 MetaWorld tasks.

Ablation on Policy Labels



- Deterministic pseudo-labels perform worse in heterogeneous data.
- Shows benefit of incorporating true or approximated behavior policies.

Online RLHF Setting



Method	Params
PPL	76k
PEBBLE	859k

- PPL can be directly applied to online setting.
- Achieves competitive performance with 1/10 of PEBBLE's parameters.

Conclusion

- PPL resolves likelihood mismatch by modeling regret w.r.t. behavior policies.
- Theoretical foundations show regret minimization $\Leftrightarrow$ forward KL.
- Contrastive KL regularization provides robustness across offline and online RLHF.
- PPL is sample-efficient and scalable for real-world RLHF tasks.

Thank you!

Project Page

References

Gaon An (2023)
Direct preference-based policy optimization without reward modeling

Paul F Christiano (2017)
Deep reinforcement learning from human preferences

Joey Hejna (2024)
Contrastive preference learning: Learning from human feedback without reinforcement learning

W. Bradley Knox (2024)
Learning optimal advantage from preferences and mistaking it for reward.

Kimin Lee (2021)
Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training.

References

Rafael Rafailov (2024)
From r to q^*: Your language model is secretly a q-function.

Rafael Rafailov (2024)
Direct preference optimization: Your language model is secretly a reward model.

Omar Shaikh (2024)
Show, Don't Tell: Aligning Language Models with Demonstrated Feedback.

Yongcheng Zeng (2024)
Token-level Direct Preference Optimization

Brian D. Ziebart (2010)
Modeling purposeful adaptive behavior with the principle of maximum causal entropy.