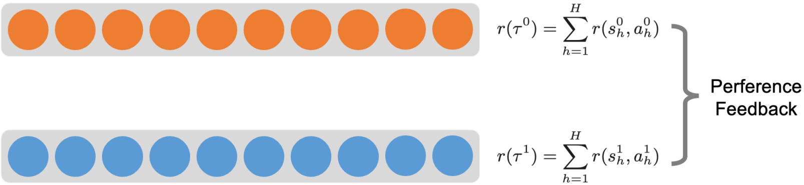
Adversarial Policy Optimization for Offline Preference-Based Reinforcement Learning

Hyungkyu Kang and **Min-hwan Oh**

Seoul National University

Offline Preference-Based RL



- RL needs well-defined reward functions, which are often hard to design
- Preference-based RL (PbRL) learns from human feedback via trajectory comparisons
- Collecting online preferences is expensive → offline PbRL

Theoretical Guarantee vs. Computational Efficiency

- Although several works have developed empirical PbRL algorithms, FREEHAND (Zhan et al., 2024) and Sim-OPRL (Pace et al., 2024) are only provably efficient offline PbRL algorithms with general function approximation
- However, they ensure conservatism using explicit confidence sets over reward and transition models
  - FREEHAND relies on confidence-set-constrained optimization
  - Sim-OPRL uses the width of the confidence sets as uncertainty penalties

  $\rightarrow$ Computationally intractable with complex function classes such as neural networks

Our goal: **Statistically & Computationally efficient offline PbRL**

Problem Setting

Episodic MDP $(\mathcal{S}, \mathcal{A}, H, \{P_h^\star\}_{h=1}^H, \{r_h^\star\}_{h=1}^H)$

- Rewards are unobservable to the agent, only trajectory-based preference feedback is available

Offline datasets: preference dataset $\mathcal{D}_{\text{pref}} = \{(\tau^{m,0}, \tau^{m,1}, y^m)\}_{m=1}^M$ and unlabeled trajectory dataset $\mathcal{D}_{\text{traj}} = \{(\tau^{0,n}, \tau^{1,n})\}_{n=1}^N$

- $\mathbb{P}(y = 1 \mid \tau^0, \tau^1) = \mathbb{P}(\tau^1 \text{ is preferred over } \tau^0) = \Phi(r^\star(\tau^1) - r^\star(\tau^0))$ where $r^\star(\tau) = \sum_{h=1}^H r_h^\star(s_h, a_h)$
- Assume $|r(\tau)| \leq R$ and $\kappa = 1/(\inf_{x \in [-R,R]} \Phi'(x))$ is finite

General function approximation: the function class of rewards $\mathcal{R}$ and the function class of transitions $\mathcal{P}$

- Maximum likelihood reward estimation $\hat{r} \in \arg\min_{r \in \mathcal{R}^H} \hat{\mathcal{L}}_R(r)$ where

$$\hat{\mathcal{L}}_R(r) = -\mathop{\mathbb{E}}_{(\tau^0, \tau^1, y) \sim \mathcal{D}_{\text{pref}}} \left[\mathbb{1}_{\{y=1\}} \log \Phi(r(\tau^1) - r(\tau^0)) + \mathbb{1}_{\{y=0\}} \log \Phi(r(\tau^0) - r(\tau^1)) \right]$$

- Similarly, $\hat{P}_h \in \arg\min_{P \in \mathcal{P}} \hat{\mathcal{L}}_T(P; h)$ where

$$\hat{\mathcal{L}}_T(P; h) = \mathbb{E}_{(s_h, a_h, s_{h+1}) \sim \mathcal{D}_{\text{traj}}} [\log P(s_{h+1} \mid s_h, a_h)]$$

Adversarial Optimization for PbRL

Zhan et al. (2024) proves that the following optimization problem yields a near-optimal policy $\hat{\pi}$, for a proper constant ζ:

$$\hat{\pi} \in \arg\max_{\pi} \min_{r \in \hat{\mathcal{R}}} \left(V_{1,r}^{\pi}(s_1) - V_{1,r}^{\mu}(s_1) \right) \text{ where } \hat{\mathcal{R}} = \left\{ r \in \mathcal{R}^H : \hat{\mathcal{L}}_R(r) \leq \hat{\mathcal{L}}_R(\hat{r}) + \zeta \right\}.$$

However, the constrained optimization is not computationally efficient.

Adversarial Optimization for PbRL

Zhan et al. (2024) proves that the following optimization problem yields a near-optimal policy $\hat{\pi}$, for a proper constant ζ:

$$\hat{\pi} \in \arg\max_{\pi} \min_{r \in \hat{\mathcal{R}}} \left(V^{\pi}_{1,r}(s_1) - V^{\mu}_{1,r}(s_1) \right) \text{ where } \hat{\mathcal{R}} = \{ r \in \mathcal{R}^H : \hat{\mathcal{L}}_R(r) \leq \hat{\mathcal{L}}_R(\hat{r}) + \zeta \}.$$

However, the constrained optimization is not computationally efficient.

Our approach: Frame PbRL as a two-player Stackelberg game

$$\hat{\pi} \in \arg\max_{\pi} \left(V^{\pi}_{1,r^\pi}(s_1) - V^{\mu}_{1,r^\pi}(s_1) \right)$$
$$\text{subject to } r^\pi \in \arg\min_{r \in \mathcal{R}^H} \left(V^{\pi}_{1,r}(s_1) - V^{\mu}_{1,r}(s_1) + \mathcal{E}(r; \hat{r}) \right)$$
$$\text{where } \mathcal{E}(r; \hat{r}) = \mathbb{E}_{\tau^0, \tau^1 \sim \mu} \left[\left| \{ r(\tau^0) - r(\tau^1) \} - \{ \hat{r}(\tau^0) - \hat{r}(\tau^1) \} \right| \right]$$

- Policy π: Maximizes return for r^π
- Reward model r: Minimizes advantage of π over behavior policy μ
- We use the trajectory-pair ℓ_1 loss instead of log-likelihood

Adversarial Optimization for PbRL

The optimization $r^\pi \in \arg\min_{r \in \mathcal{R}^H} \left(V^\pi_{1,r}(s_1) - V^\mu_{1,r}(s_1) + \mathcal{E}(r;\hat{r})\right)$ requires online trajecotries from π $\to$ Unavailable in offline PbRL

We use reparameterization of reward model to address this challenge

- Fix a policy π. For given reward model $r = \{r_h\}_{h=1}^H$, we have the value function $\{Q^\pi_{h,r}\}_{h=1}^H$ such that $Q^\pi_{h,r} = r_h + P^\star_h(Q^\pi_{h+1,r} \circ \pi_{h+1})$
- Conversely, for a value function $f = \{f_h\}_{h=1}^H$, we can construct a reward model $\{r_h\}_{h=1}^H$ satisfying $r_h = f_h - P^\star_h(f_{h+1} \circ \pi_{h+1})$

Using the reparameterization, we reduce the two-player game to a single unconstrained optimization problem

Proposed algorithm: APPO

Algorithm 2 Adversarial Preference-based Policy Optimization (APPO)

Input: KL regularization η, Initial policy $\pi_h^1 = \mathrm{Unif}(\mathcal{A})$ for all $h \in [H]$
Estimate $\hat{r} \in \arg\min_{r \in \mathcal{R}^H} \hat{\mathcal{L}}_R(r)$, $\hat{P}_h \in \arg\min_{P \in \mathcal{P}} \hat{\mathcal{L}}_T(P; h)$ for all $h \in [H]$
for $t = 1, \cdots, T$ **do**
    $f^t \in \underset{f \in \mathcal{F}^H}{\arg\min} \left(\sum_{h=1}^{H} \mathbb{E}_{(s_h, a_h) \sim \mathcal{D}_{\text{traj}}} \left[f_h \circ \pi_h^t(s_h) - f_h(s_h, a_h) \right] + \lambda \hat{\mathcal{E}}_{\mathcal{D}_{\text{traj}}}(f; \hat{P}, \hat{r}) \right)$
    Update policy $\pi_h^{t+1}(a \mid s) \propto \pi_h^t(a \mid s) \exp(\eta f_h^t(s, a))$ for $h \in [H]$
end for
Return $\hat{\pi} = \frac{1}{T} \sum_{t=1}^{T} \pi^t$

$$\hat{\mathcal{E}}_{\mathcal{D}_{\text{traj}}}(f; \hat{P}, \hat{r}) = \mathbb{E}_{(\tau_0, \tau_1) \sim \mu} \left[\left| \{ r^{\pi^t}_{\hat{P}, f}(\tau^0) - r^{\pi^t}_{\hat{P}, f}(\tau^1) \} - \{ \hat{r}(\tau^0) - \hat{r}(\tau^1) \} \right| \right]$$

$$r^{\pi}_{h, P^\star, r} = f_h - P_h^\star(f_{h+1} \circ \pi_h)$$

Theoretical Analysis

Assumption (Reward realizability)

We have $r_h^\star \in \mathcal{R}$ *for all* $h \in [H]$*. In addition, every* $r \in \mathcal{R}^H$ *satisfies* $0 \leq r(\tau) \leq R$ *for any trajectory* τ*.*

Assumption (Transition realizability)

We have $P_h^\star \in \mathcal{P}$ *for all* $h \in [H]$*.*

Assumption (Value function class)

For any $h \in [H]$*,* $r \in \mathcal{R}^H$*, and policy* π*, we have* $Q_{h,r}^\pi \in \mathcal{F}$*. In addition, every* $f \in \mathcal{F}$ *satisfies* $0 \leq f(s,a) \leq R$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*.*

Assumption (Trajectory concentrability)

There exists a finite constant C_{TR} *such that the behavior policy* μ *and the optimal policy* $\pi^\star$ *satisfy* $\sup_\tau \frac{d^{\pi^\star}(\tau)}{d^\mu(\tau)} \leq C_{\mathrm{TR}}$*.*

Theoretical Analysis

Theorem (Sub-optimality bound of APPO)

Suppose Assumptions 1,2,3, and 4 hold. With probability at least $1-\delta$*,*
APPO with $\lambda = \Theta(C_{\mathrm{TR}}), \lambda > C_{\mathrm{TR}}, \eta = \sqrt{\frac{2\log|\mathcal{A}|}{R^2 T}}$ *achieves*

$$V_{1,r^\star}^{\pi^\star} - V_{1,r^\star}^{\hat{\pi}}$$
$$\leq \mathcal{O}\left(C_{\mathrm{TR}}\sqrt{\frac{\kappa^2 H}{M}\log\frac{|\mathcal{R}|}{\delta}} + RH\sqrt{\frac{1}{N}\max\left\{HT\log\frac{H|\mathcal{F}|}{\delta}, \log\frac{H|\mathcal{P}|}{\delta}\right\}} + RH\sqrt{\frac{\log|\mathcal{A}|}{T}}\right).$$

Theoretical Analysis

Theorem (Sub-optimality bound of APPO)

Suppose Assumptions 1,2,3, and 4 hold. With probability at least $1-\delta$*, APPO with* $\lambda = \Theta(C_{\mathrm{TR}}), \lambda > C_{\mathrm{TR}}, \eta = \sqrt{\frac{2\log|\mathcal{A}|}{R^2 T}}$ *achieves*

$$V_{1,r^\star}^{\pi^\star} - V_{1,r^\star}^{\hat{\pi}}$$
$$\leq \mathcal{O}\left(C_{\mathrm{TR}}\sqrt{\frac{\kappa^2 H}{M}\log\frac{|\mathcal{R}|}{\delta}} + RH\sqrt{\frac{1}{N}\max\left\{HT\log\frac{H|\mathcal{F}|}{\delta}, \log\frac{H|\mathcal{P}|}{\delta}\right\}} + RH\sqrt{\frac{\log|\mathcal{A}|}{T}} \right).$$

- ϵ-optimal policy with $T = \Theta\left(\frac{R^2H^2\log|\mathcal{A}|}{\epsilon^2}\right)$, $M = \Theta\left(\frac{C_{\mathrm{TR}}^2\kappa^2 H\log(|\mathcal{R}|/\delta)}{\epsilon^2}\right)$, $N = \Theta\left(\max\left\{\frac{R^4H^5\log|\mathcal{A}|\log(H|\mathcal{F}|/\delta)}{\epsilon^4}, \frac{R^2H^2\log(H|\mathcal{P}|/\delta)}{\epsilon^2}\right\}\right)$.
- Same bound with Zhan et al. (2024) and Pace et al. (2024) for preference dataset (M)
- The bound for unlabeled dataset is looser than $\Theta\left(\frac{C_P^2R^2H^2\log(H|\mathcal{P}|/\delta)}{\epsilon^2}\right)$ bound of them (C_P is the concentrability for transition models)
- This highlights a trade-off: While FREEHAND and Sim-OPRL have tighter bounds for unlabeled data, they are computationally intractable.

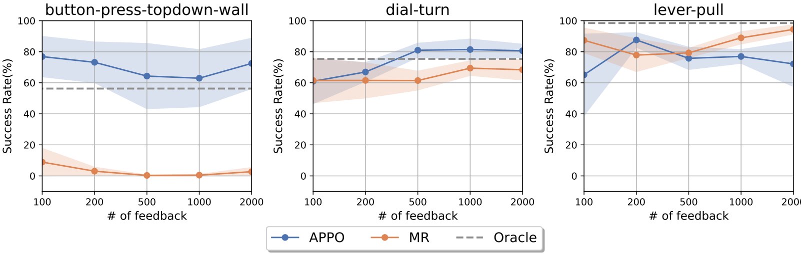
Experiments

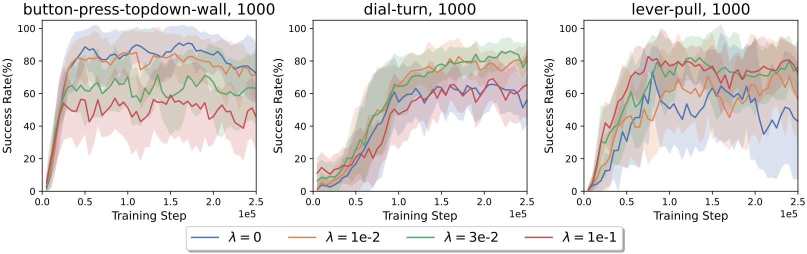
Dataset & # of feedback	Oracle	MR	PT	DPPO	IPL	APPO (ours)
BPT-500	88.33±4.76	10.08±7.57	22.87±9.06	3.93±4.34	34.73±13.9	**53.52**±13.9
box-close-500	93.40±3.10	**29.12**±13.2	0.33±1.16	10.20±11.5	5.93±5.81	**18.24**±15.6
dial-turn-500	75.40±5.47	61.44±6.08	**68.67**±12.4	26.67±22.2	31.53±12.5	**80.96**±4.49
sweep-500	98.33±1.87	**86.96**±6.93	43.07±24.6	10.47±15.8	27.20±23.8	26.80±5.32
BPT-wall-500	56.27±6.32	0.32±0.30	0.87±1.43	0.80±1.51	8.93±9.84	**64.32**±21.0
sweep-into-500	78.80±7.96	**28.40**±5.47	20.53±8.26	23.07±7.02	**32.20**±7.35	24.08±5.91
drawer-open-500	100.00±0.00	**98.00**±2.32	88.73±11.6	35.93±11.2	19.00±13.6	87.68±10.0
lever-pull-500	98.47±1.77	**79.28**±2.95	**82.40**±22.7	10.13±12.2	31.20±15.8	**75.76**±7.17
BPT-1000	88.33±4.76	8.48±5.80	18.27±10.6	3.20±3.04	36.67±17.4	**59.04**±19.0
box-close-1000	93.40±3.10	**27.04**±14.5	2.27±2.86	9.33±9.60	6.73±8.41	**34.24**±18.5
dial-turn-1000	75.40±5.47	69.44±4.70	68.80±5.50	36.40±21.9	43.93±13.4	**81.44**±6.73
sweep-1000	98.33±1.87	**87.52**±7.87	29.13±14.6	8.73±16.4	38.33±24.9	17.36±12.4
BPT-wall-1000	56.27±6.32	0.48±0.47	2.13±2.96	0.27±0.85	14.07±11.5	**62.96**±18.4
sweep-into-1000	78.80±7.96	**26.00**±5.53	20.27±7.84	**23.33**±7.80	**30.40**±7.74	18.16±11.1
drawer-open-1000	100.00±0.00	**98.40**±2.82	**95.40**±7.27	36.47±7.30	28.53±18.4	**98.56**±2.68
lever-pull-1000	98.47±1.77	**88.96**±3.94	72.93±10.2	8.53±9.96	40.40±17.4	76.96±4.40
Average Rank	-	2.316	3.125	4.375	3.063	**2.125**

Baselines: IQL with markovian reward model (MR), Preference transformer (PT) (Kim et al., 2023), DPPO (An et al., 2023), IRL (Hejna and Sadigh, 2024)

APPO outperforms or shows comparable performance with the baselines

Experiments





Summary and Contributions

APPO: Statistically and Computationally Efficient Offline PbRL

- Based on the two-player game formulation of PbRL and our reparameterization technique
- Unconstrained optimization which allows practical implementation

Theoretical Guarantee

- General function approximation, standard assumptions on function classes and trajectory concentrability
- The first computationally efficient offline PbRL algorithm providing a sample complexity bound

Empirical Performance

- Practical implementation leveraging deep learning techniques
- Performance comparable to existing state-of-the-art algorithms

Achieves both theoretical guarantee and practical efficiency!

References I

An, G., Lee, J., Zuo, X., Kosaka, N., Kim, K.-M., and Song, H. O. (2023). Direct preference-based policy optimization without reward modeling. Advances in Neural Information Processing Systems, 36:70247–70266.

Hejna, J. and Sadigh, D. (2024). Inverse preference learning: Preference-based rl without a reward function. Advances in Neural Information Processing Systems, 36.

Kim, C., Park, J., Shin, J., Lee, H., Abbeel, P., and Lee, K. (2023). Preference transformer: Modeling human preferences using transformers for RL. In The Eleventh International Conference on Learning Representations.

Pace, A., Schölkopf, B., Ratsch, G., and Ramponi, G. (2024). Preference elicitation for offline reinforcement learning. In ICML 2024 Workshop: Foundations of Reinforcement Learning and Control – Connections and Perspectives.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. (2024). Provable offline preference-based reinforcement learning. In The Twelfth International Conference on Learning Representations.