Zeroth-Order Policy Gradient for Reinforcement Learning from Human Feedback without Reward Inference

Qining Zhang

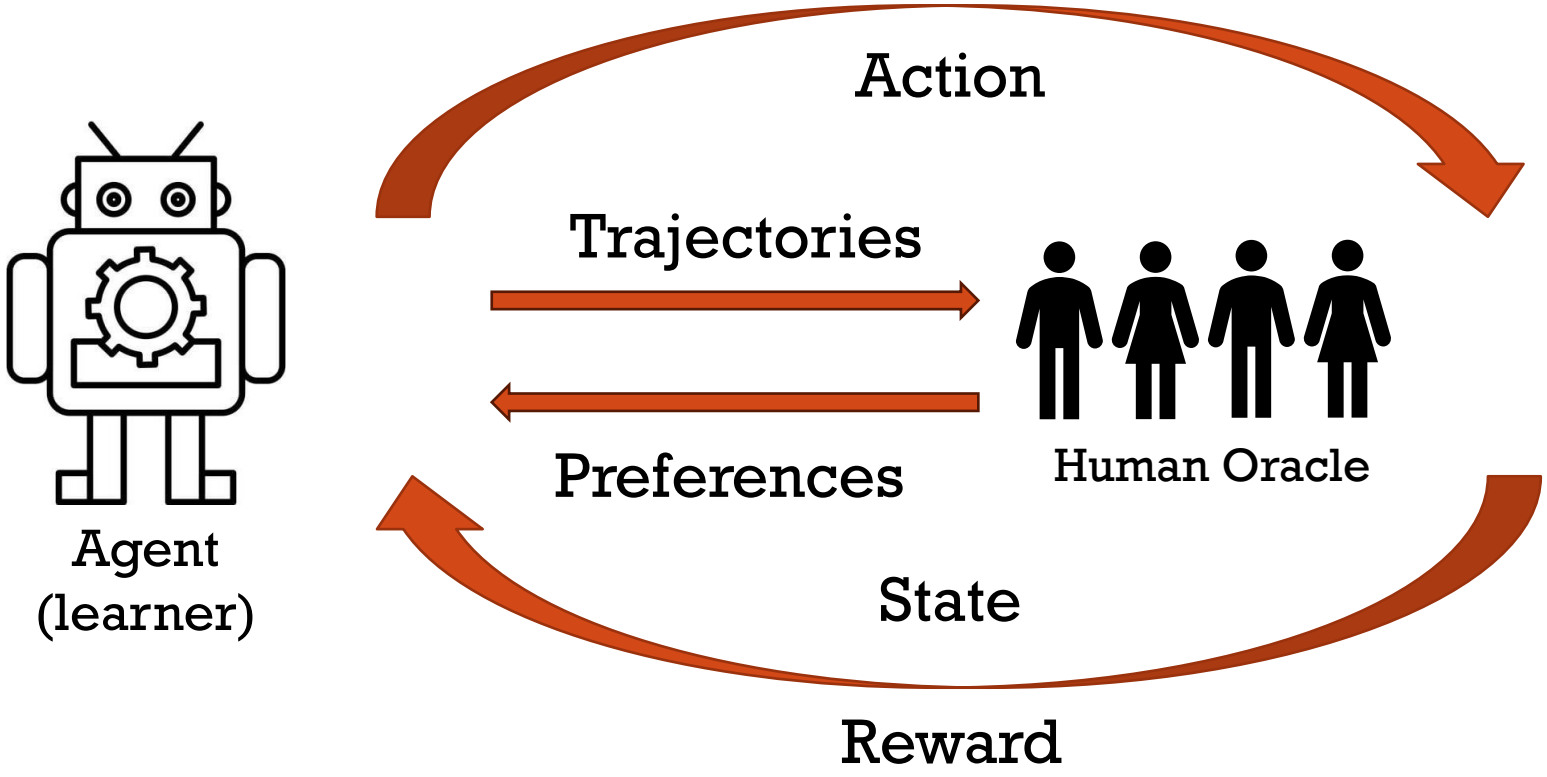
EECS, University of Michigan, Ann Arbor

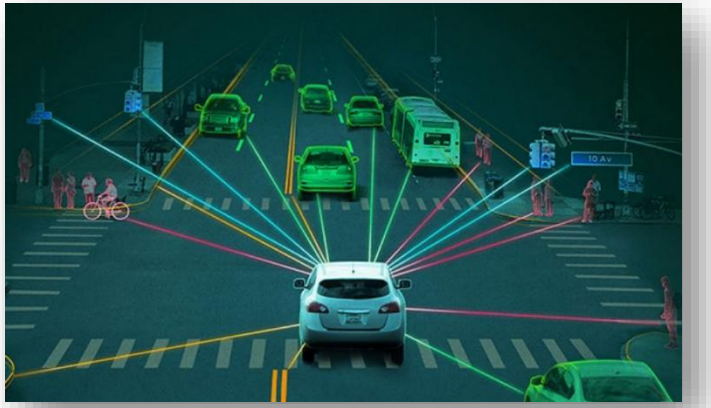
Joint work with Lei Ying(UMich)

RL from Human Feedback

- Learn from Reward v.s. Learn from Preference

- ❑ Inverse RL (Russell 1998, Ng&Russell 2000, …)
- ❑ Dueling bandits (Yue&Joachims 2000, Ailon, Karnin&Joachims, 2014, …)
- ❑ Preference-based RL (Akrour et al. 2011, Cheng et al. 2011, …)





Environment

Applications

Large Language Models

Preferences of Prompt Responses

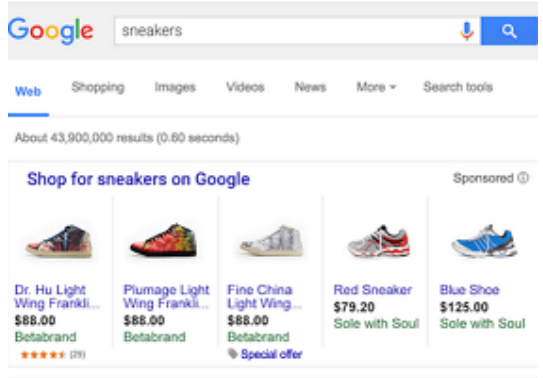
Video Game Bot

Preferences of Gaming Strategies

Robotics

Preferences of Motions

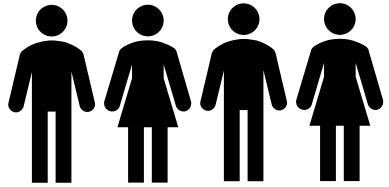
Recommendation Systems



Preferences of Merchandise

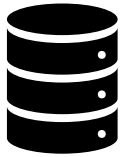
RLHF: The GPT-4 Approach

- Bradley-Terry (BT) Preference Model + Reward Inference + PPO (Achiam et al. 2023)



Human feedback $(\tau_1 \succ \tau_0)$

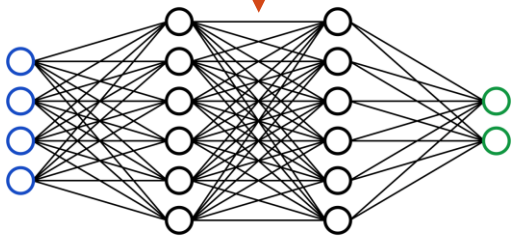
trajectory pair
(τ_0, τ_1)



logged dataset

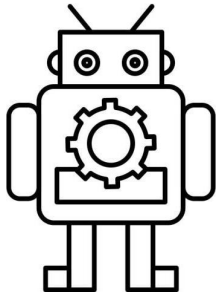
behavior policy π
trajectories
$(s_1, a_1, s_2, \cdots)$

BT: $\mathbb{P}(\tau_1 \succ \tau_0) = \left(1 + e^{r(\tau_0) - r(\tau_1)}\right)^{-1}$



Reward model: $r_h(s, a)$

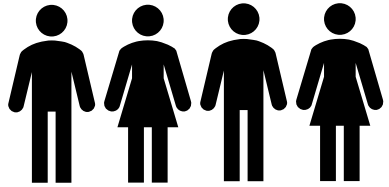
Reward inference



Proximal Policy Optimization (PPO)
(Schulman et al. 2017)

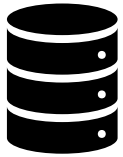
RLHF: The GPT-4 Approach

- Bradley-Terry (BT) Preference Model + Reward Inference + PPO (Achiam et al. 2023)



Human feedback $(\tau_1 \succ \tau_0)$

trajectory pair (τ_0, τ_1)

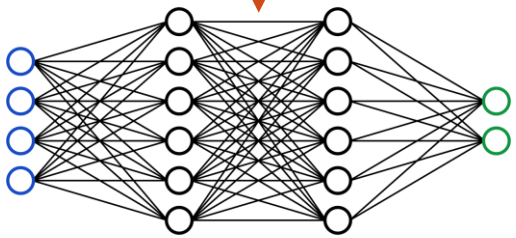


logged dataset

behavior policy π

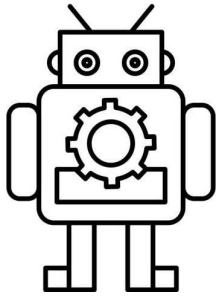
trajectories $(s_1, a_1, s_2, \cdots)$

BT: $\mathbb{P}(\tau_1 \succ \tau_0) = \left(1 + e^{r(\tau_0) - r(\tau_1)}\right)^{-1}$



Reward model: $r_h(s, a)$

Reward inference



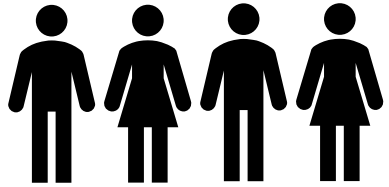
Proximal Policy Optimization (PPO)
(Schulman et al. 2017)

Theoretical analysis of RLHF

- ❑ Value-based (Wang, Liu&Jin 2023, …)
- ❑ Policy-optimization-based (Du et al. 2024, …)

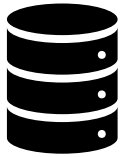
RLHF: The GPT-4 Approach

- Bradley-Terry (BT) Preference Model + Reward Inference + PPO (Achiam et al. 2023)



Human feedback $(\tau_1 \succ \tau_0)$

trajectory pair (τ_0, τ_1)

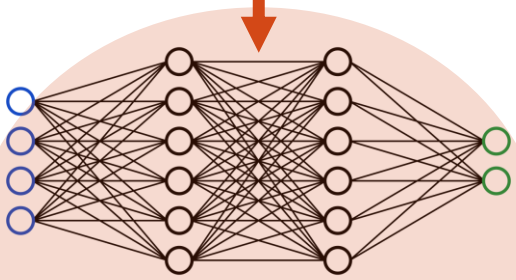


logged dataset

behavior policy π

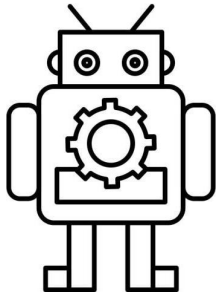
trajectories $(s_1, a_1, s_2, \cdots)$

BT: $\mathbb{P}(\tau_1 \succ \tau_0) = \left(1 + e^{r(\tau_0) - r(\tau_1)}\right)^{-1}$



Reward model: $r_h(s, a)$

Reward inference



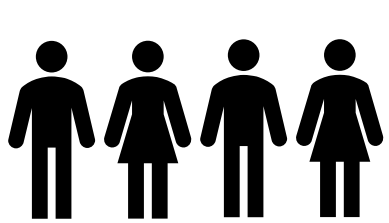
Proximal Policy Optimization (PPO)
(Schulman et al. 2017)

Issues of reward inference (Casper et al. 2023)

- ❑ reward model evaluation without ground truth
- ❑ distribution shift
- ❑ overfitting in joint reward model and policy training

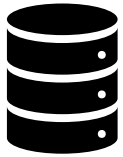
Direct Preference Optimization

- DPO (Rafailov et al. 2023)



Human feedback $(\tau_1 \succ \tau_0)$

trajectory pair (τ_1, τ_2)



logged dataset

behavior Policy π

trajectories $(s_1, a_1, s_2, \cdots)$

human preference

reward

optimal policy

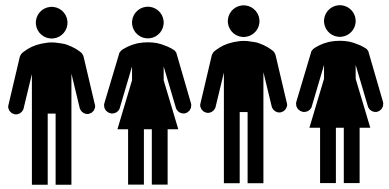
BT model

$$\mathbb{P}(\tau_1 \succ \tau_0) = \left(1 + e^{r(\tau_0) - r(\tau_1)}\right)^{-1}$$

nonparametric; bandit setting

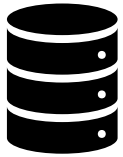
$$\pi^*(a|x) \propto \pi_{\text{ref}}(a|x) \exp\left(\frac{1}{\beta} r(x, a)\right)$$

DPO



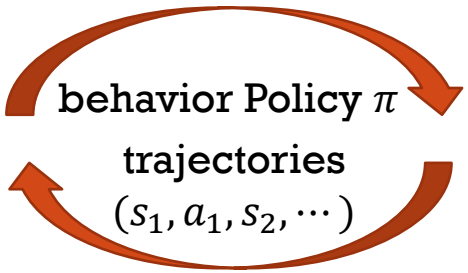
Human feedback

trajectory pair
(τ_1, τ_2)

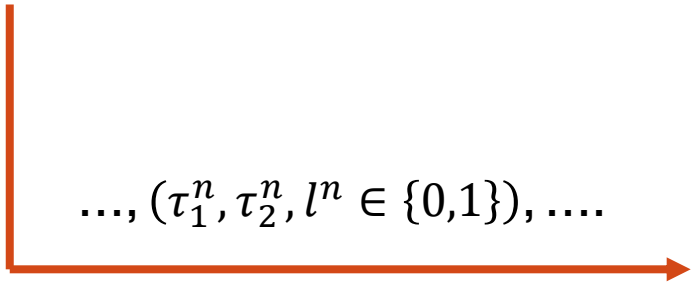


logged dataset

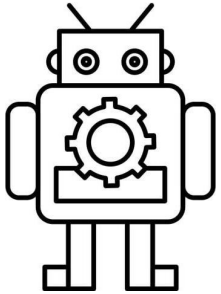
behavior Policy π
trajectories
$(s_1, a_1, s_2, \cdots)$



$\ldots, (\tau_1^n, \tau_2^n, l^n \in \{0,1\}), \ldots.$



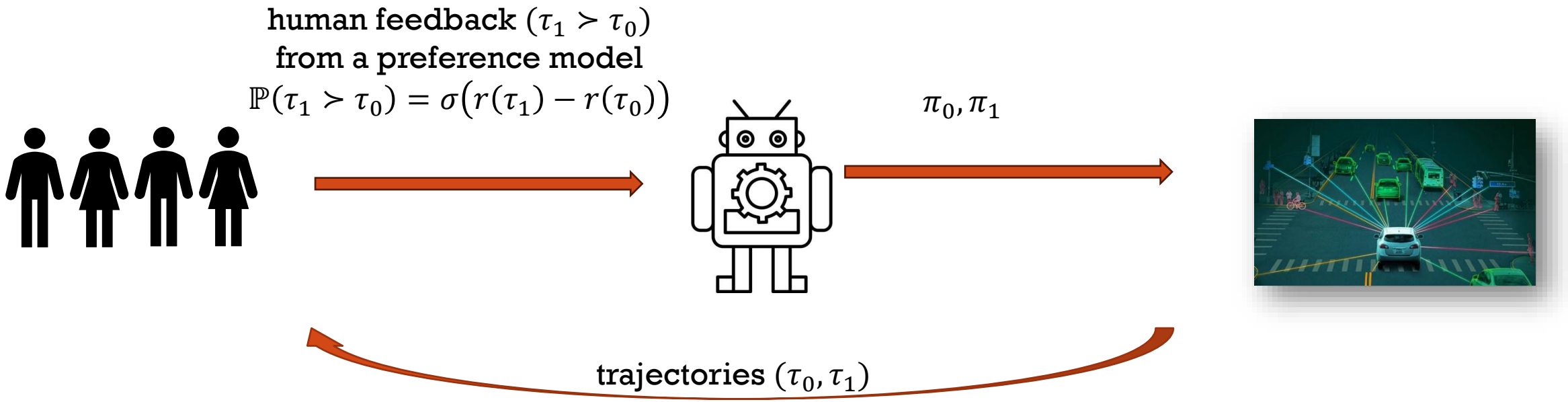
$\mathbb{P}(\tau_1 \succ \tau_2) =$ function of π^*



The theory only works for

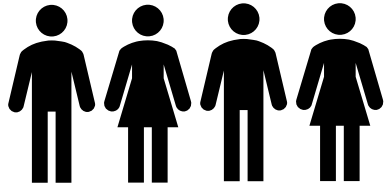
- ❑ non-parametric settings
- ❑ bandits (Rafailov et al. 2023) or deterministic MDPs (Rafailov et al. 2024)

RLHF without Reward Inference for General RL Problems



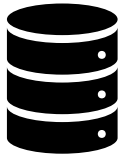
General RL, Parameterized Policy, Uncountable (s,a)

- DPO (Rafailov et al. 2023)



Human feedback $(\tau_1 \succ \tau_0)$

trajectory pair (τ_1, τ_2)



logged dataset

behavior Policy π

trajectories $(s_1, a_1, s_2, \cdots)$

human preference

reward

optimal policy

BT model

$$\mathbb{P}(\tau_1 \succ \tau_0) = \left(1 + e^{r(\tau_0) - r(\tau_1)}\right)^{-1}$$

nonparametric; bandit setting

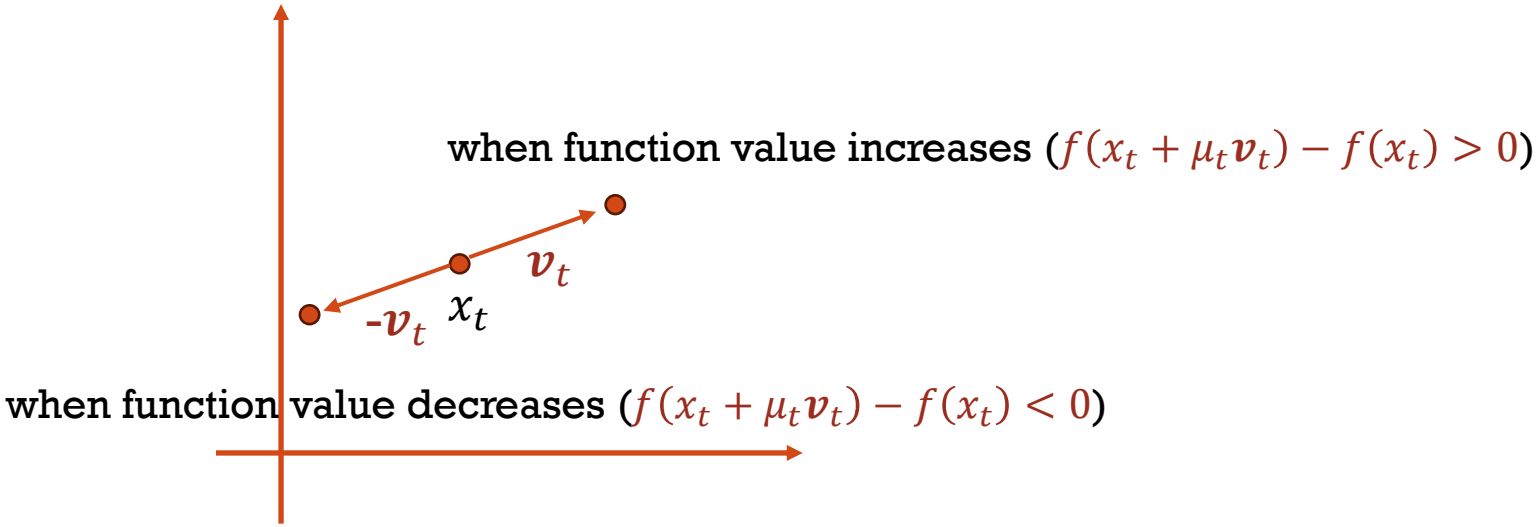
$$\pi^*(a|x) \propto \pi_{\text{ref}}(a|x) \exp\left(\frac{1}{\beta} r(x,a)\right)$$

Zeroth-Order Optimization

- Two-point, zeroth-order optimization (Nesterov&Spokoiny, 2017)

$$\max f(x) \qquad x_{t+1} = x_t + \frac{h_t}{\mu_t}\left(f(x_t + \mu_t \boldsymbol{v}_t) - f(x_t)\right)\boldsymbol{v}_t$$

  - $\boldsymbol{v}_t$: sampled uniformly from a unit sphere



Zeroth-Order Optimization

- Two-point, zeroth-order optimization
  (Nesterov&Spokoiny, 2017)

$$\max f(x)$$

$$x_{t+1} = x_t + \frac{h_t}{\mu_t}\left(f(x_t + \mu_t \boldsymbol{v}_t) - f(x_t)\right)\boldsymbol{v}_t$$

  - $\boldsymbol{v}_t$: sampled uniformly from a unit sphere

Policy optimization in RL

$$\max_{\theta} V(\pi_\theta)$$

Value function $V(\pi_{\theta_t})$ under policy π_{θ_t}

Parameter θ_t that defines policy π_{θ_t}

Zeroth-Order Optimization

- Two-point, zeroth-order optimization
  (Nesterov&Spokoiny, 2017)

$$\max f(x)$$

$$x_{t+1} = x_t + \frac{h_t}{\mu_t}\left(f(x_t + \mu_t \boldsymbol{v}_t) - f(x_t)\right)\boldsymbol{v}_t$$

- $\boldsymbol{v}_t$: sampled uniformly from a unit sphere

Policy optimization in RL

$$\max_{\theta} V(\pi_\theta)$$

Value function $V(\pi_{\theta_t})$ under policy π_{θ_t}

Parameter θ_t that defines policy π_{θ_t}

How to obtain the value function difference?

Human feedback!

Zeroth-Order Policy Gradient from Human Feedback

- ZPG: $\theta_{t+1} = \theta_t + \frac{h_t}{\mu_t}\left(V\left(\pi_{\theta_t+\mu_t v_t}\right) - V\left(\pi_{\theta_t}\right)\right)\boldsymbol{v}_t$
- ZPG in HF: $\theta_{t+1} = \theta_t + \frac{h_t}{\mu_t}\hat{g}_t\boldsymbol{v}_t$

Q. Zhang and L. Ying. *Reinforcement Learning from Direct Human Feedback via Zeroth-Order Policy Gradient.* ICLR, 2025.

Zeroth-Order Policy Gradient from Human Feedback

Current policy π_{θ_t} → N trajectories $\{\tau_{n,0}\}$

Perturbed policy $\pi_{\theta_t+\mu_t v_t}$ → N trajectories $\{\tau_{n,1}\}$

M human evaluations $\{o_{n,1}, \cdots, o_{n,M}\}$ for each pair of trajectories $\{\tau_{n,1}, \tau_{n,0}\}$

Estimate the preference probability from human feedback

$$p_{t,n} = \left(\sum_m \frac{o_{n,m}}{M}\right)_{\Delta}^{1-\Delta} \approx \mathbb{P}(\tau_{n,1} \succ \tau_{n,0})$$

Estimate value function difference

$$\hat{g}_t = \frac{\sum_n \sigma^{-1}(p_{t,n})}{N} \approx \frac{\sum_n \left(r(\tau_{n,1}) - r(\tau_{n,0})\right)}{N} \approx V(\pi_{\theta_t+\mu_t v_t}) - V(\pi_{\theta_t})$$

Update the policy

$$\theta_{t+1} = \theta_t + \frac{h_t}{\mu_t}\hat{g}_t \boldsymbol{v}_t$$

Challenge: a biased estimator of the value function difference

Q. Zhang and L. Ying. *Reinforcement Learning from Direct Human Feedback via Zeroth-Order Policy Gradient.* ICLR, 2025.

Rate of Convergence

Theorem: ZPG has the following rate of convergence to a stationary point:

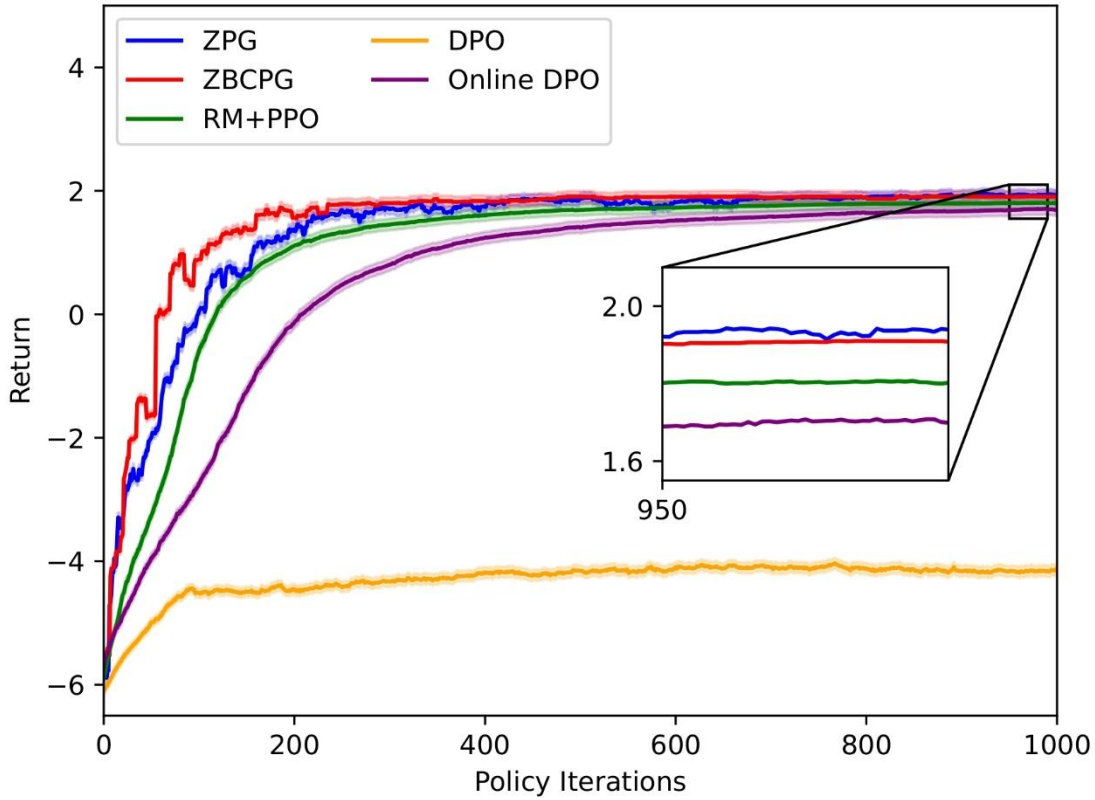
$$\widetilde{O}\left(\frac{Hd}{T}+\frac{d^2\sqrt{\log M}}{\sqrt{M}}+\frac{Hd\sqrt{d}}{\sqrt{N}}\right)$$

- d: policy dimension
- T: number of policy gradient steps
- N: number of trajectory pairs per pg step
- M: number of human preferences per trajectory pair

Q. Zhang and L. Ying. *Reinforcement Learning from Direct Human Feedback via Zeroth-Order Policy Gradient.* ICLR, 2025.

Experiments

- Stochastic Gridworld
  - Softmax policy parameterization
  - Bradley-Terry human feedback
  - Weibull human feedback



Q. Zhang and L. Ying. *Reinforcement Learning from Direct Human Feedback via Zeroth-Order Policy Gradient.* ICLR, 2025.

Thanks!
Questions?