Exploration-Driven Policy Optimization in RLHF: Theoretical Insights on Efficient Data Utilization

Yihan Du
UIUC

Anna Winnick
UIUC

Gal Dalal
Nvidia

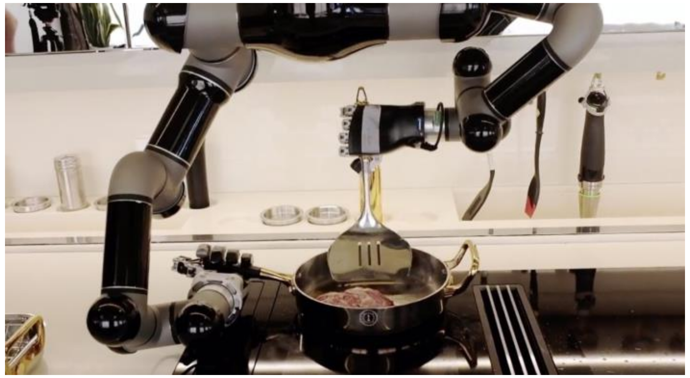
Shie Mannor
Technion/Nvidia

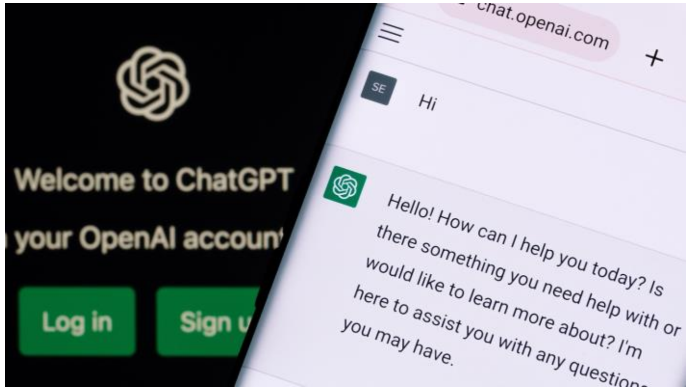
R. Srikant
UIUC

Speaker: Yihan Du
ICML 2024

Motivation

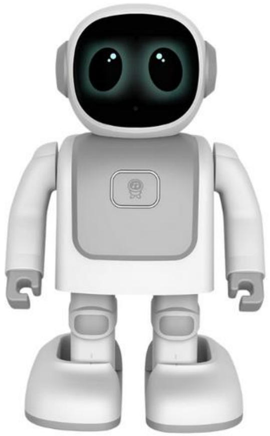
- Reinforcement Learning from Human Feedback (RLHF):
  - Goals are complex and hard to specify
    - E.g., let a robot cook
  - Misalignment with human's objective
    - E.g., ChatGPT
- Empirical success of RLHF:
  - Data efficiency: "Using feedback on <1% of the agent's interactions with the environment" [Christiano et al., 2017]





Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, Dario Amodei. Deep reinforcement learning from human preferences. NeurIPS, 2017

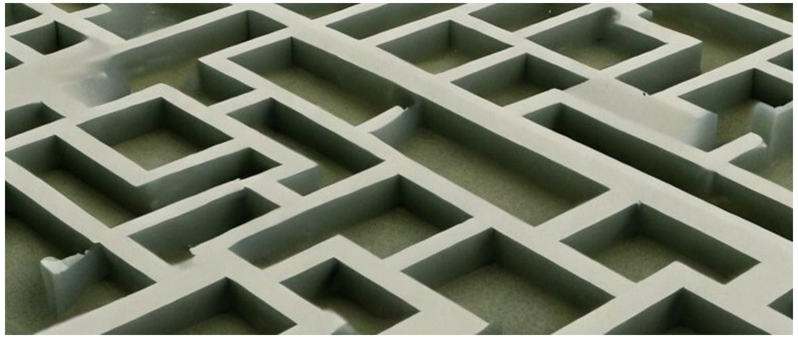
Formulation



Agent

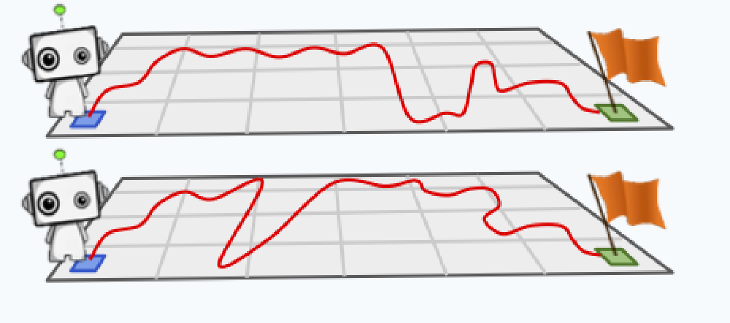
Action

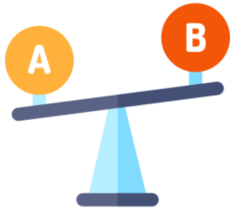
Observation
(State and ~~Reward~~)



Environment

Humans provide comparison feedback between trajectories





Exploration-Driven RLHF Algorithm: PG-RLHF

π^{n+1} [Agarwal et al., 2020]

Reward Learning

Learn the reward model $\hat{r}^n$ from preference data $\left\{\tau_i^{(1)}, \tau_i^{(2)}\right\}_{i=1}^{M_{HF}}$ using policy $\pi_{cov}^n = avg(\pi^0, \dots, \pi^n)$

Add exploration bonus

$\hat{r}^n + b^n$

π_{cov}^n

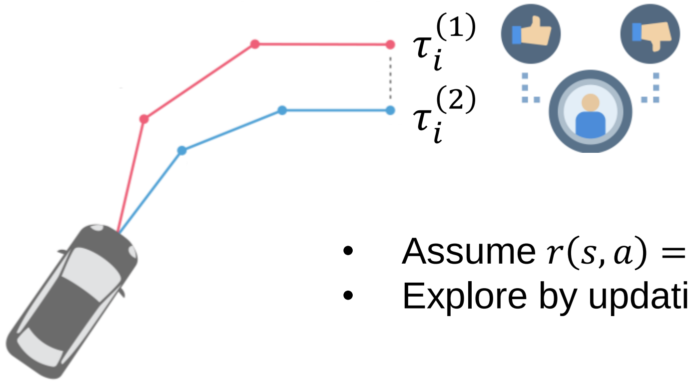
to generate initial state

Policy Improvement

$\pi(a|s) \propto \pi(a|s)\exp(\eta Q(s,a))$

Policy Evaluation

Collect samples from the trajectories generated by π, and estimate $Q \approx \phi(s,a)^\top \theta$



- Assume $r(s,a) = \phi(s,a)^\top \mu$
- Explore by updating the policy π_{cov}^n

Alekh Agarwal, Mikael Henaff, Sham Kakade, Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. NeurIPS, 2020.

Result of Algorithm PG-RLHF

Theorem 1. With probability $1-\delta$, the output policy π^{out} of algorithm PG-RLHF satisfies

$$V^{\pi^*}(s_0) - V^{\pi^{out}}(s_0) \leq \tilde{O}\left(\sqrt{\varepsilon_{bias}} + \frac{1}{\sqrt{T}} + \frac{\sqrt{N}}{M_{SGD}^{\frac{1}{4}}} + \frac{\sqrt{N}}{M_{HF}^{\frac{1}{4}}} + \frac{1}{N}\right)$$

- ε_{bias}: Q-value function approximation error
- T: # iterations of policy optimization
- M_{SGD}: # iterations in SGD for policy evaluation
- M_{HF}: # human trajectory comparisons for reward learning
- N: # outer loop iterations

- When T, M_{SGD}, M_{HF}, N increase, $V^{\pi^*}(s_0) - V^{\pi^{out}}(s_0)$ decreases to zero up to ε_{bias}
- M_{SGD} and M_{HF} have the same convergence rate

Comparison between PG-RLHF and PC-PG

	PG-RLHF (ours)	PC-PG [Agarwal et al., 2020] for Standard RL
# Samples	$\tilde{O}(NK + NTM_{SGD} + NM_{HF})$	$\tilde{O}(NK + NTM_{SGD})$
# True rewards	0	$\tilde{O}(NK + NTM_{SGD})$
# Queries	$O(NM_{HF})$	0

Remarks:

- $\tilde{O}(M_{SGD}) \approx \tilde{O}(M_{HF})$ due to the same convergence rate
- The ratio of query complexity over the overall sample complexity is about $\frac{NM_{HF}}{NTM_{SGD}} \approx \frac{1}{T}$

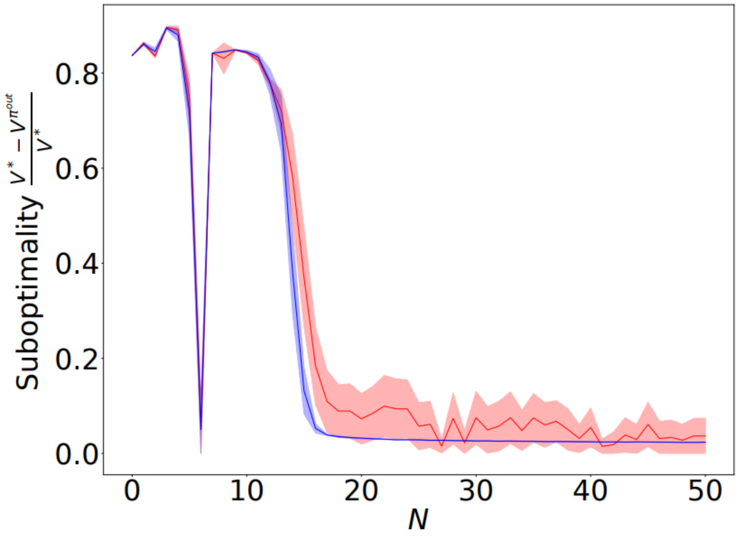
Alekh Agarwal, Mikael Henaff, Sham Kakade, Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. NeurIPS, 2020.

Experiments

- $N = 50,\ K = 2500,\ M_{SGD} = 2500,\ \ M_{HF} = 2500,$
  $S = 22,\ A = 5,\ \gamma = 0.9,\ \delta = 0.005$

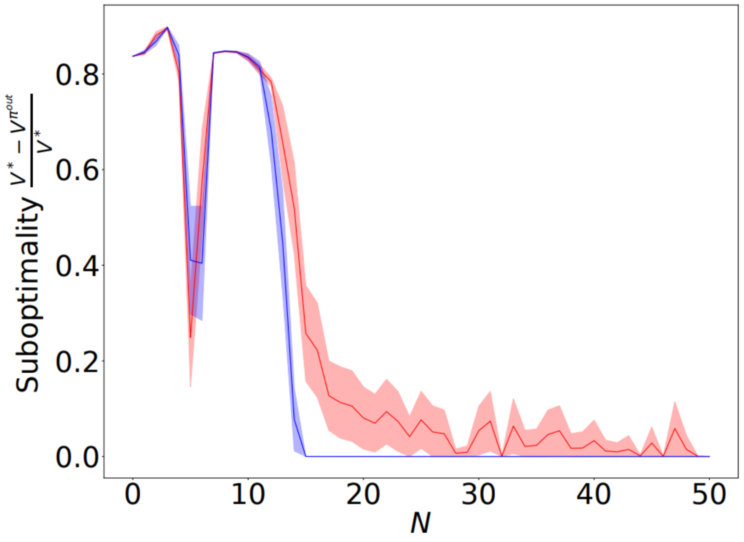
PG-RLHF (ours)

PC-PG [Agarwal et al., 2020]



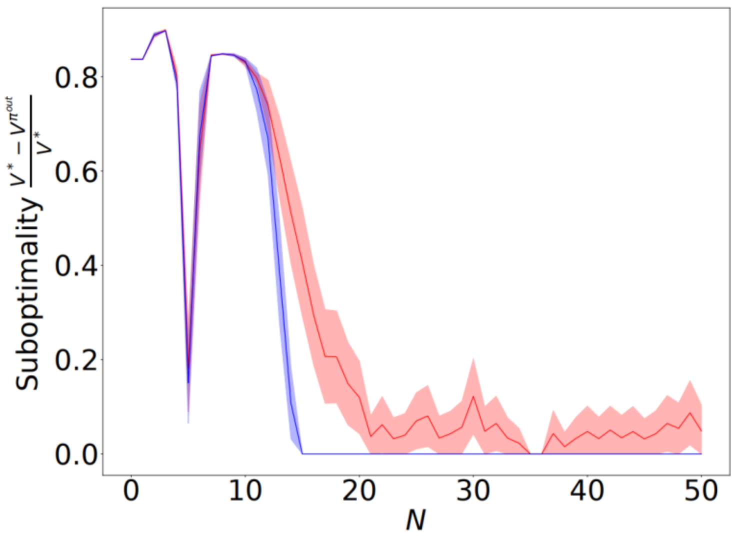
$T = 50$

$$\frac{\#\ Queries}{\#\ Samples} = 0.962\%$$



$T = 100$

$$\frac{\#\ Queries}{\#\ Samples} = 0.490\%$$



$T = 200$

$$\frac{\#\ Queries}{\#\ Samples} = 0.248\%$$

When T increases, $\frac{\#\ Queries}{\#\ Samples} \approx \frac{1}{T}$ decreases

Conclusion

A theoretical explanation for the data efficiency of RLHF:

- The reward model is first learned, and then fixed during policy optimization
- $\frac{\#\,Queries}{\#\,Total\ samples} \approx \frac{M_{HF}}{T M_{SGD}} \approx \frac{1}{T}$

Thank You