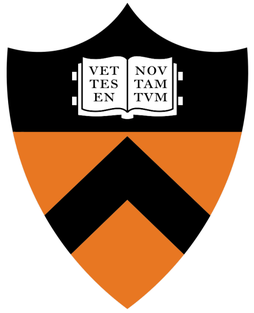
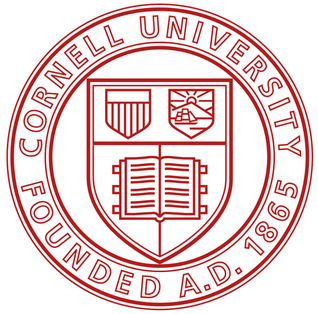
Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach

Xuezhou Zhang*, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun



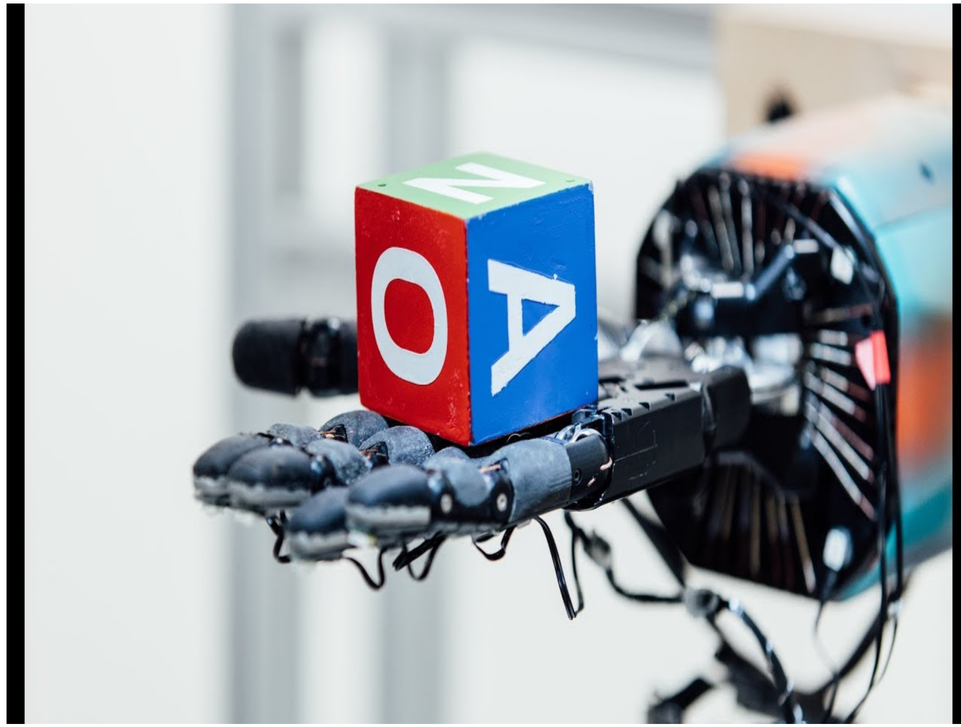


ICML 2022

Empirical RL for large-scale problems

[AlphaGo, Silver et.al, 15]

[OpenAI Five, 18]

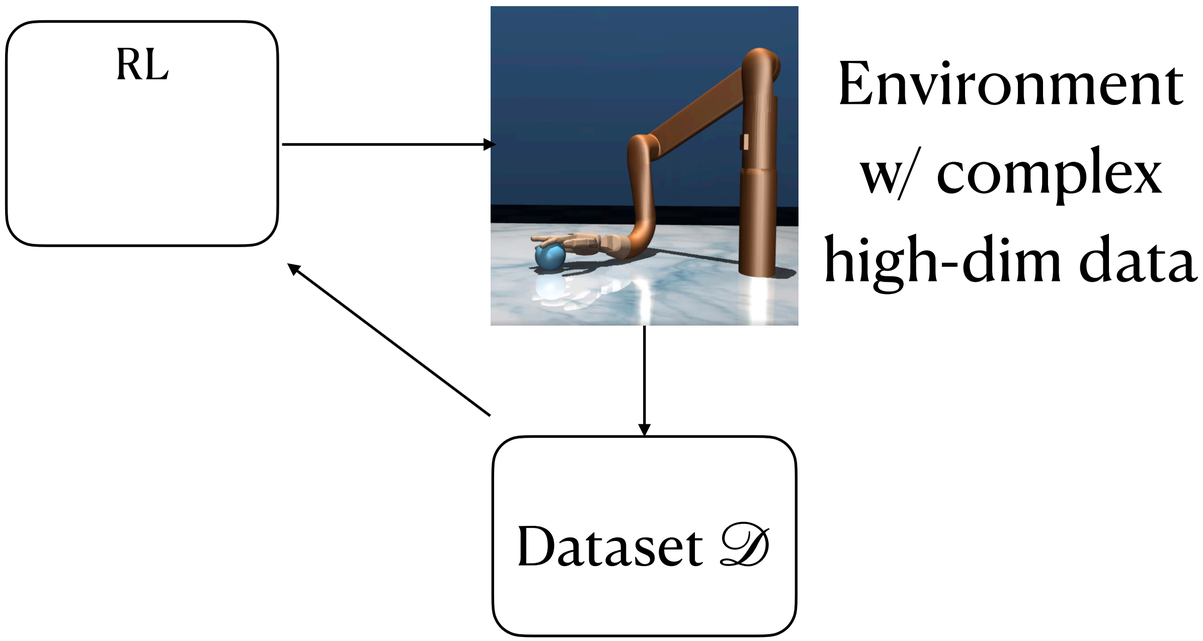


[OpenAI,19]

Rich (nonlinear) function approximation + RL can work well w/ enough samples

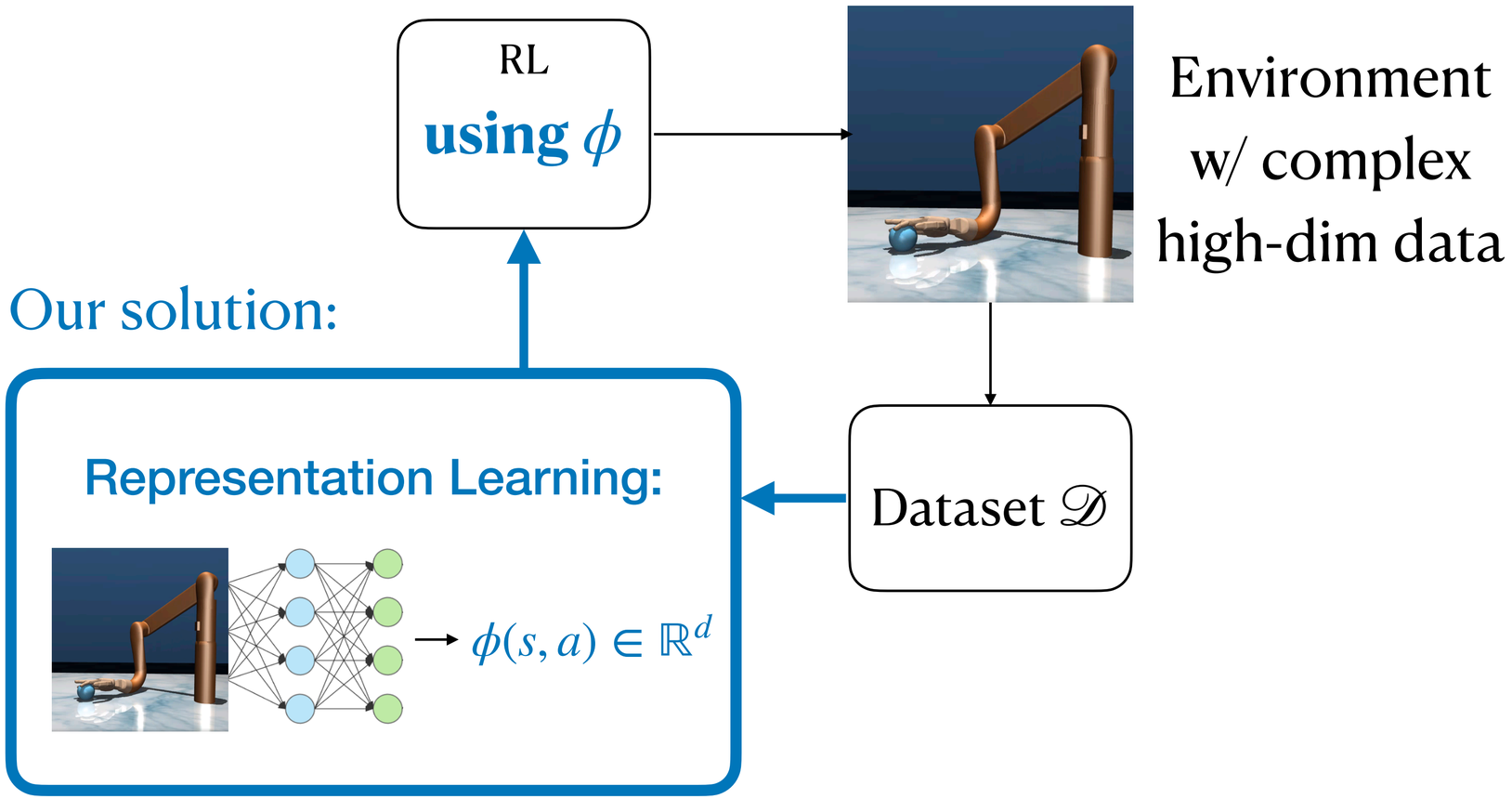
Can we design provably efficient algorithms for

Rich Function Approx + RL ?

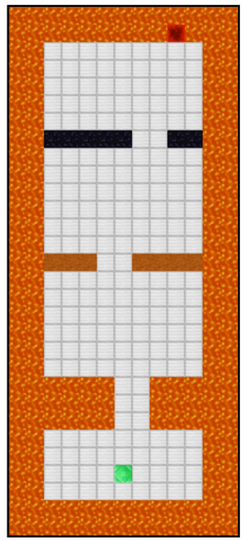


Can we design provably efficient algorithms for

Rich Function Approx + RL ?



Block MDP



Block MDP

Decoder: for any $s \in \mathcal{S}, z = \psi^\star(s)$.

Latent Transition: $z' \sim T^\star(\cdot \mid z, a)$

Emission: $s' \sim o^\star(\cdot \mid z')$

Block-structured Representation learning with Interleaved Explore Exploit

Block-structured Representation learning with Interleaved Explore Exploit

1. Data collection using the current policy: $\pi := \{\pi_1, \ldots, \pi_H\}$, for all h:

Block-structured Representation learning with Interleaved Explore Exploit

1. Data collection using the current policy: $\pi := \{\pi_1, \ldots, \pi_H\}$, for all h:

$$s \sim d_h^{\pi}, a \sim U(A), s' \sim P_h^{\star}(s,a), D_h = D_h \cup \{s, a, s'\}$$

$$s \sim d_{h-1}^{\pi}, a \sim U(A), s' \sim P_{h-1}^{\star}(s,a), a' \sim U(A), s'' \sim P_h^{\star}(s,a), \quad D_h' = D_h' \cup \{s', a', s''\}$$

Block-structured Representation learning with Interleaved Explore Exploit

1. Data collection using the current policy: $\pi := \{\pi_1, \ldots, \pi_H\}$, for all h:

$$s \sim d_h^{\pi}, a \sim U(A), s' \sim P_h^{\star}(s,a), D_h = D_h \cup \{s, a, s'\}$$

$$s \sim d_{h-1}^{\pi}, a \sim U(A), s' \sim P_{h-1}^{\star}(s,a), a' \sim U(A), s'' \sim P_h^{\star}(s,a), \quad D'_h = D'_h \cup \{s', a', s''\}$$

2. Run Representation Learning subprotocol with the $\mathscr{D}_h$ and $\mathscr{D}'_h$

$$\hat{\phi}_h = \text{RepLearn}(\mathscr{D}_h \cup \mathscr{D}'_h, \Phi)$$

Block-structured Representation learning with Interleaved Explore Exploit

1. Data collection using the current policy: $\pi := \{\pi_1, \ldots, \pi_H\}$, for all h:

$$s \sim d_h^{\pi}, a \sim U(A), s' \sim P_h^{\star}(s,a), D_h = D_h \cup \{s, a, s'\}$$

$$s \sim d_{h-1}^{\pi}, a \sim U(A), s' \sim P_{h-1}^{\star}(s,a), a' \sim U(A), s'' \sim P_h^{\star}(s,a), \quad D_h' = D_h' \cup \{s', a', s''\}$$

2. Run Representation Learning subprotocol with the $\mathscr{D}_h$ and $\mathscr{D}_h'$

$$\hat{\phi}_h = \text{RepLearn}(\mathscr{D}_h \cup \mathscr{D}_h', \Phi)$$

3.(Linear bandit style) bonus under $\hat{\phi}$:

$$b_h(s,a) = c\sqrt{\hat{\phi}_h(s,a)\Sigma_h^{-1}\hat{\phi}_h(s,a)}, \quad \Sigma_h = \sum_{s,a \in \mathscr{D}_h} \hat{\phi}_h(s,a)\hat{\phi}_h(s,a)^{\top} + \lambda I$$

Block-structured Representation learning with Interleaved Explore Exploit

1. Data collection using the current policy: $\pi := \{\pi_1, \ldots, \pi_H\}$, for all h:

$$s \sim d_h^{\pi}, a \sim U(A), s' \sim P_h^{\star}(s,a), D_h = D_h \cup \{s, a, s'\}$$

$$s \sim d_{h-1}^{\pi}, a \sim U(A), s' \sim P_{h-1}^{\star}(s,a), a' \sim U(A), s'' \sim P_h^{\star}(s,a), \quad D_h' = D_h' \cup \{s', a', s''\}$$

2. Run Representation Learning subprotocol with the $\mathscr{D}_h$ and $\mathscr{D}_h'$

$$\hat{\phi}_h = \text{RepLearn}(\mathscr{D}_h \cup \mathscr{D}_h', \Phi)$$

3.(Linear bandit style) bonus under $\hat{\phi}$:

$$b_h(s,a) = c\sqrt{\hat{\phi}_h(s,a)\Sigma_h^{-1}\hat{\phi}_h(s,a)}, \quad \Sigma_h = \sum_{s,a \in \mathscr{D}_h} \hat{\phi}_h(s,a)\hat{\phi}_h(s,a)^{\top} + \lambda I$$

4. Run Least-square VI with $\hat{\phi}_h, \mathscr{D}_h \cup \mathscr{D}_h', r + b$

Oracle Efficient Algorithms

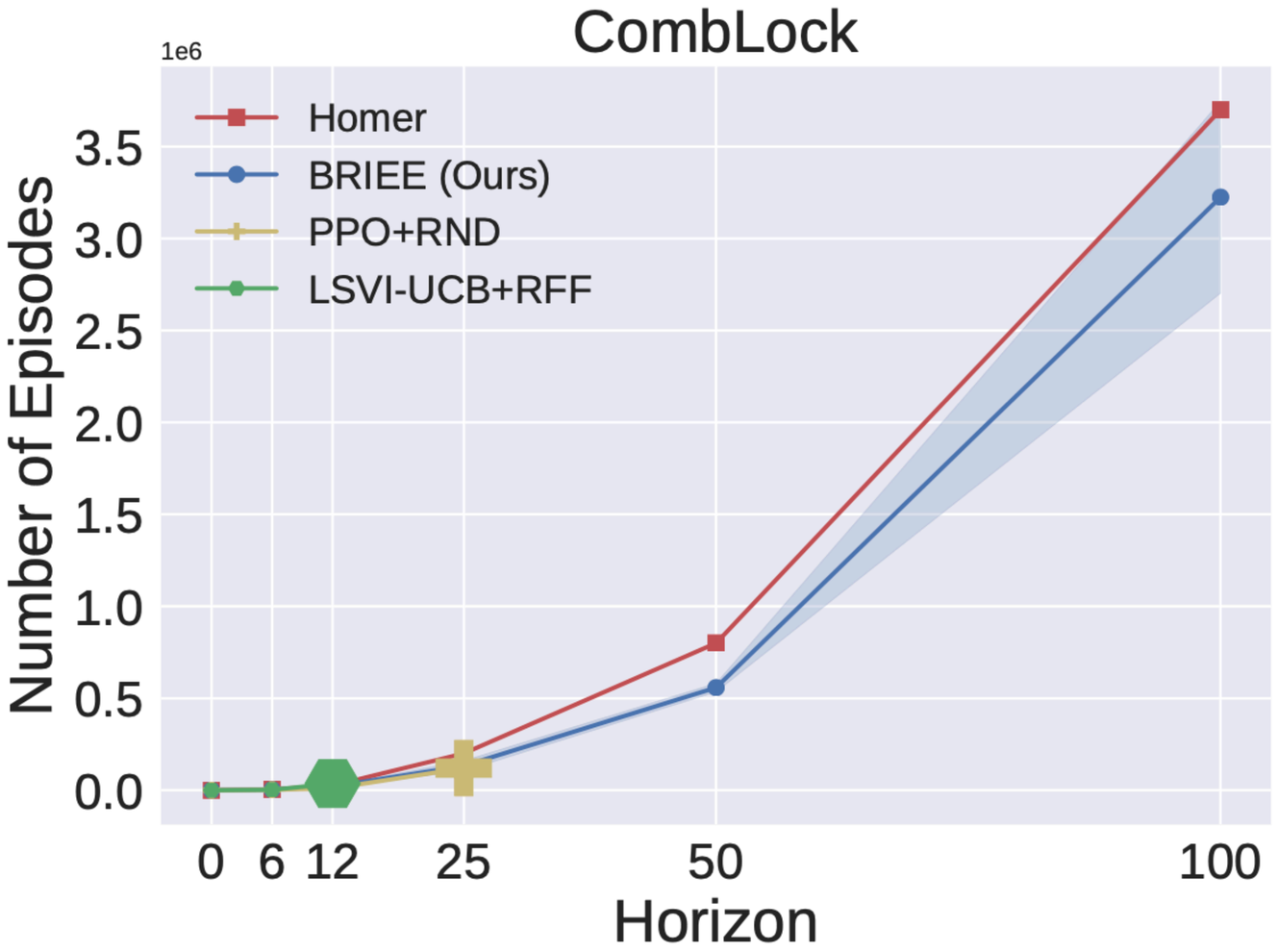
	Sample Complexity	Model-based ?	Reward?
FLAMBE [Agarwal et al., 2020]	$H^{22}d^7A^9\epsilon^{-10}$	Model-based	Reward-free
MOFFLE [Modi et al., 2021]	$H^8d^7A^{13}\epsilon^{-2}\eta_{\min}^{-1}$	Model-free	Reward-free
HOMER [Misra et al., 2019]	$H\lvert\mathcal{Z}\rvert^8A^4\epsilon^{-2}\eta_{\min}^{-3}$	Model-free	Reward-free
REP-UCB [Uehara, 2021]	$H^5d^4A^2\epsilon^{-2}$	Model-based	Reward-driven
BRIEE [this work]	$H^9\lvert\mathcal{Z}\rvert^8A^{14}\epsilon^{-4}$	Model-free	Reward-driven

Oracle Efficient Algorithms

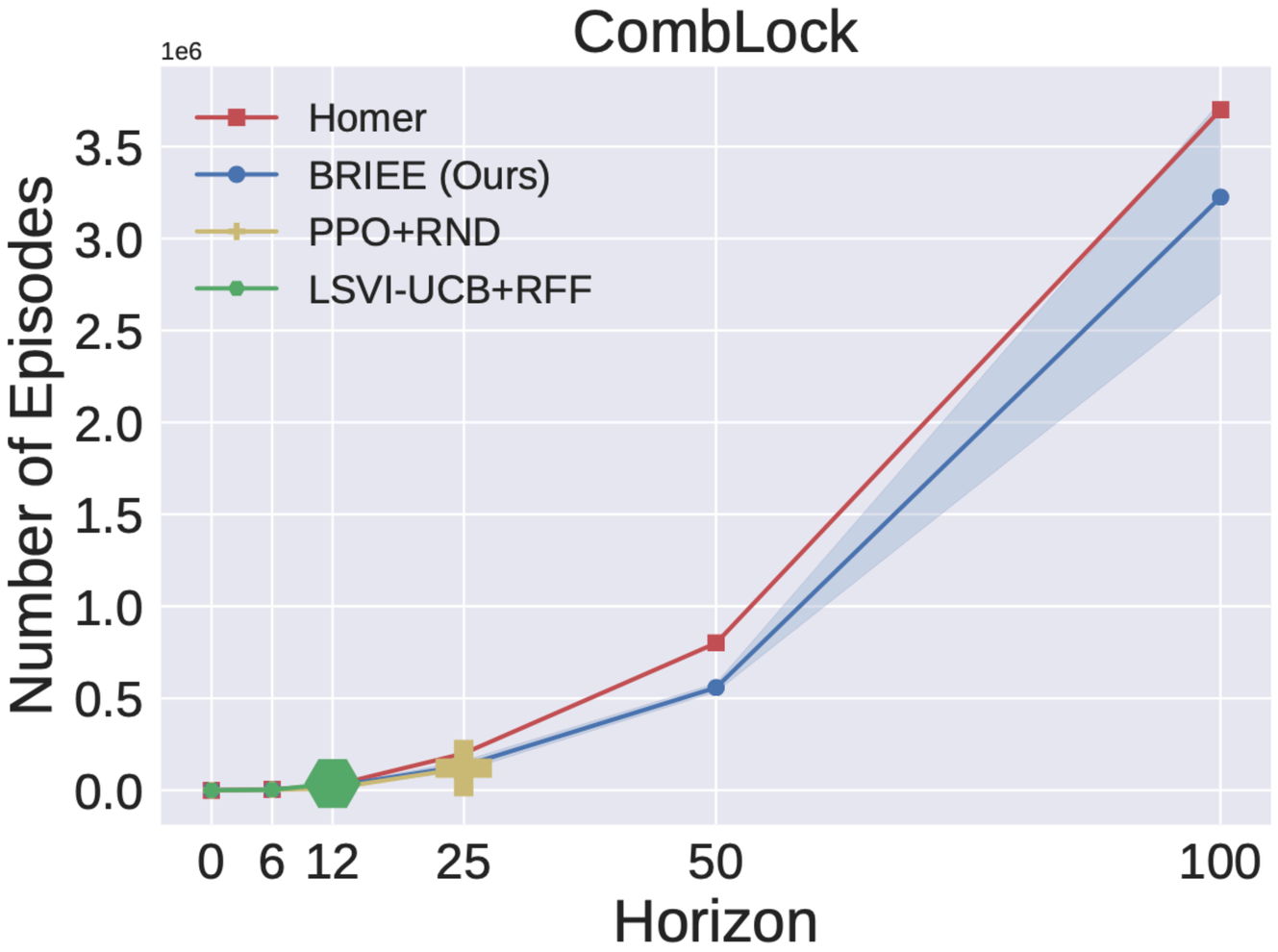
	Sample Complexity	Model-based ?	Reward?
FLAMBE [Agarwal et al., 2020]	$H^{22}d^7A^9\epsilon^{-10}$	Model-based	Reward-free
MOFFLE [Modi et al., 2021]	$H^8d^7A^{13}\epsilon^{-2}\eta_{\min}^{-1}$	Model-free	Reward-free
HOMER [Misra et al., 2019]	$H\lvert\mathcal{Z}\rvert^8A^4\epsilon^{-2}\eta_{\min}^{-3}$	Model-free	Reward-free
REP-UCB [Uehara, 2021]	$H^5d^4A^2\epsilon^{-2}$	Model-based	Reward-driven
BRIEE [this work]	$H^9\lvert\mathcal{Z}\rvert^8A^{14}\epsilon^{-4}$	Model-free	Reward-driven

Reachability

Efficient DeepRL Implementation

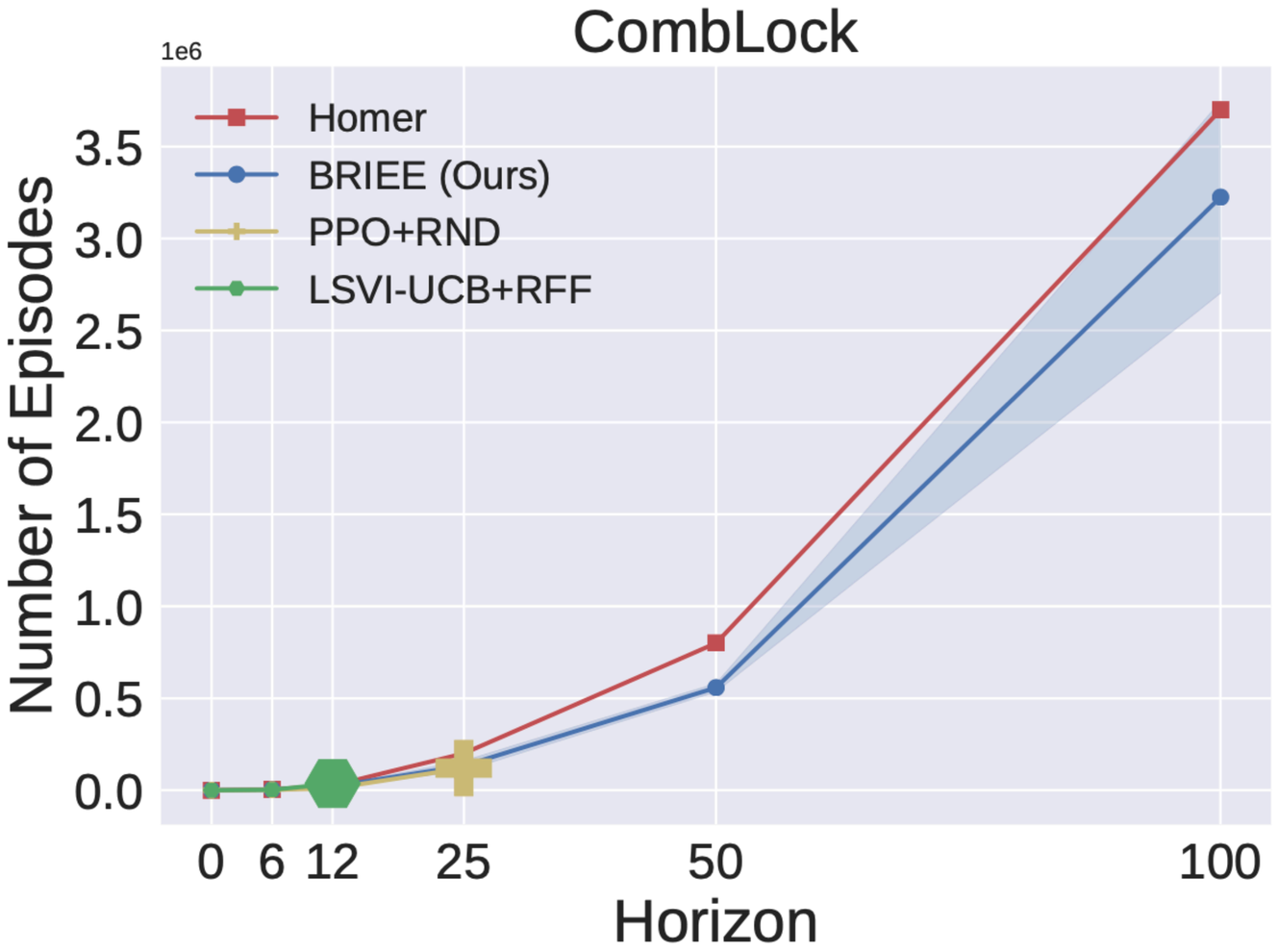


Efficient DeepRL Implementation



1. Lifting Linear MDP to kernel is not enough (i.e., linear RL theory fails here..)

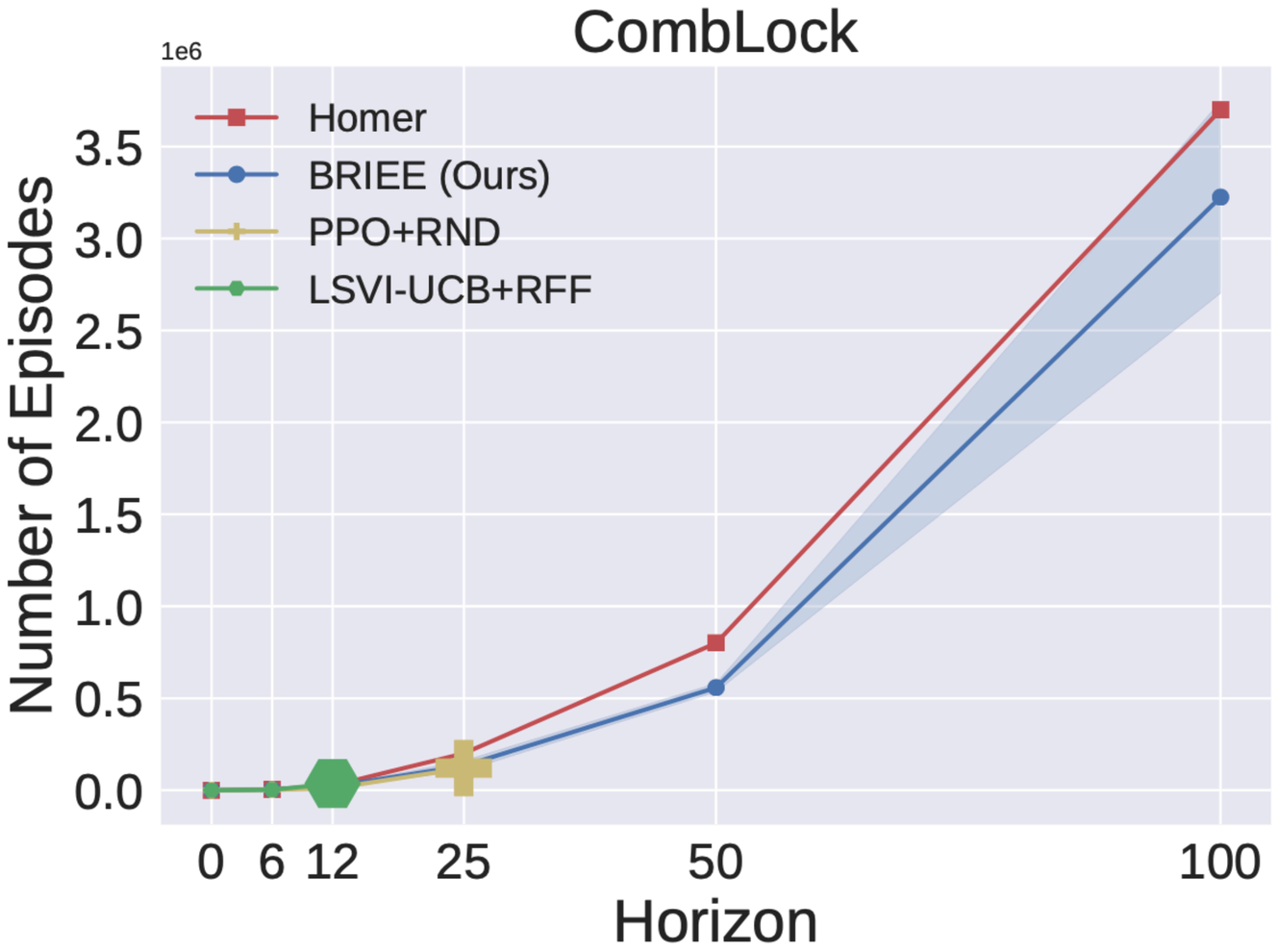
Efficient DeepRL Implementation



1. Lifting Linear MDP to kernel is not enough (i.e., linear RL theory fails here..)

2. Implicit Rep learning via deep RL (PPO) is not enough (i.e., deep rl fails here..)

Efficient DeepRL Implementation



1. Lifting Linear MDP to kernel is not enough (i.e., linear RL theory fails here..)

2. Implicit Rep learning via deep RL (PPO) is not enough (i.e., deep rl fails here..)

3. Heuristic deep exploration approach (RND) fails..

References

- BRIEE paper: https://arxiv.org/pdf/2202.00063.pdf
- BRIEE code: code: https://github.com/yudasong/briee