Randomized Exploration for Reinforcement Learning with General Value Function Approximation

Haque Ishfaq
McGill University, Mila

Joint work with Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang,

Doina Precup and Lin F. Yang

Recent Progress on Provably Efficient RL

- Provably efficient RL with linear function approximation and general value function approximation. [Wang and Yang 2019, Jin et al. 2020, Du et al. 2020, Zanette et al. 2020, Wang et al, 2020]

- Based on Upper Confidence Bonus (UCB) function
    - UCB in general is not efficient
    - UCB based bonus can be extremely optimistic
    - Naïve bonus function has high complexity

Randomized Value Function Approximation

Repeat in each episode:

1. Randomly perturb the training data

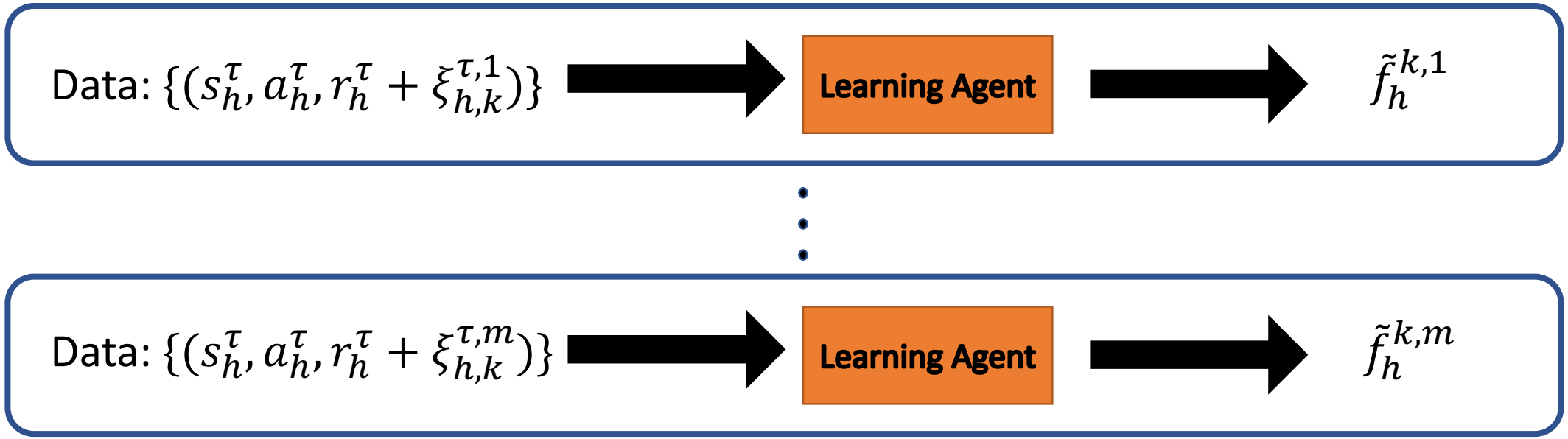
2. Apply an existing method for value function estimation

3. Follow the greedy policy under the estimated value function

Can be thought as approximate Thompson sampling

Can we perform provably efficient randomized exploration for RL with general value function approximation

Optimistic Reward Sampling



$$Q_h^{k,m}(\cdot,\cdot) \leftarrow \tilde{f}_h^{k,m}(\cdot,\cdot)$$

$$Q_h^k(\cdot,\cdot) \leftarrow \min\{\max_{m\in[M]}\{\mathrm{Q}_\mathrm{h}^{k,m}(\cdot,\cdot)\}, H - h + 1\}$$

Our Result

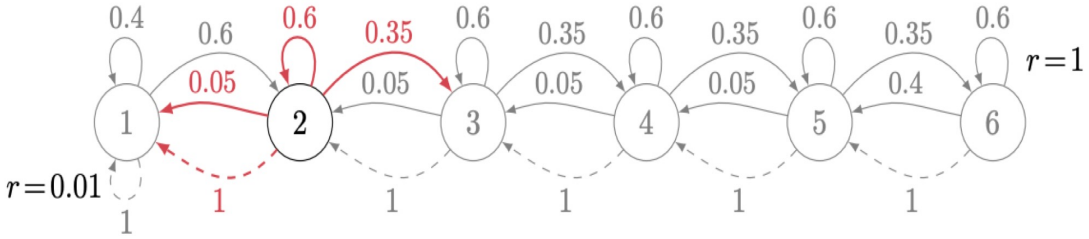
- A new algorithm $\mathcal{F}$-LSVI-PHE that works provably efficiently for a general function class $\mathcal{F}$.

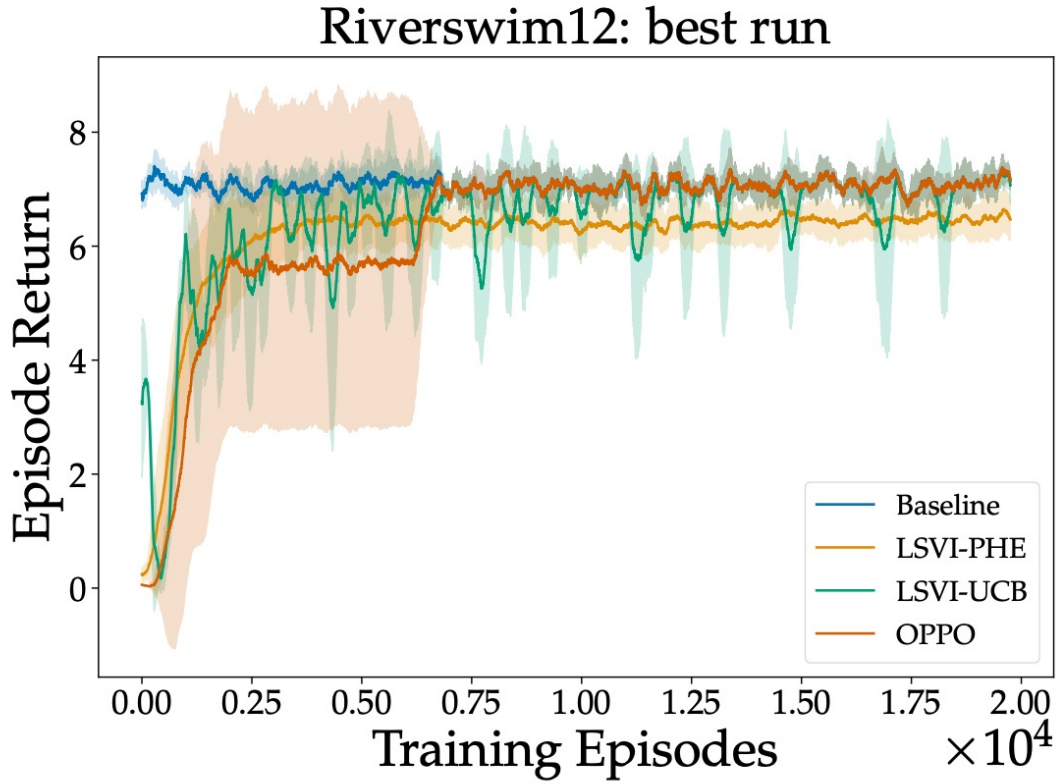
Theorem

$\mathcal{F}$-LSVI-PHE achieves a regret bound of **poly$(H, d).T^{1/2}$**

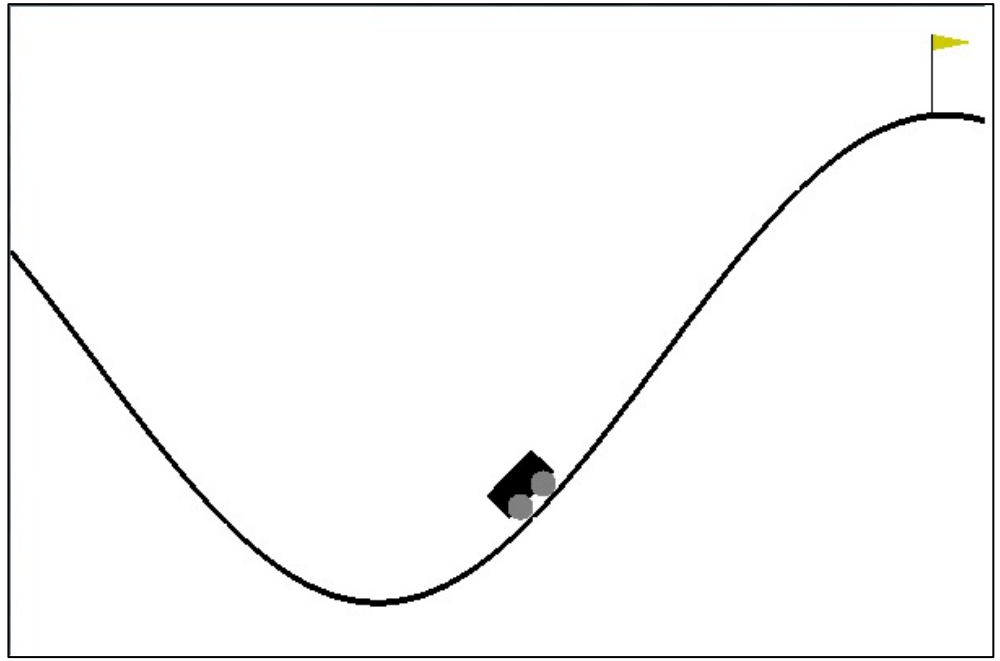
- d characterizes the complexity of the function class
- d depends on eluder dimension [Russo and Van Roy 2013] and log-covering numbers of the function class.

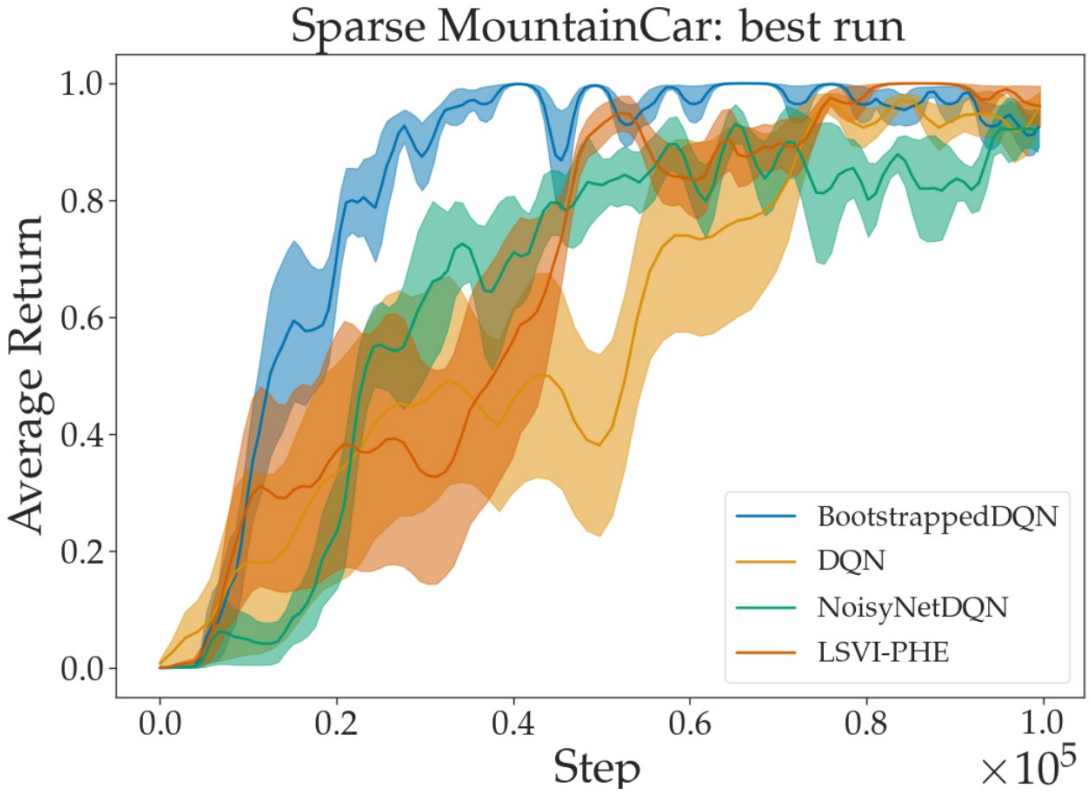
Experiment: Riverswim





Experiment: Sparse Mountain Car





Conclusion

- Provably efficient RL algorithm that works for general function classes.
- Unifying Optimism and approximate Thompson sampling
- Regret bound depends on the complexity of the function class