Optimistic Actor-Critic with Parametric Policies for Linear Markov Decision Processes

AISTATS 2026

Max Qiushi Lin

SFU

Reza Asad

SFU

Kevin Tan

UPenn

Haque Ishfaq

Mila

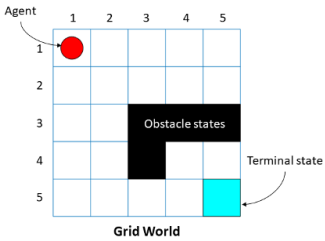
Csaba Szepesvári

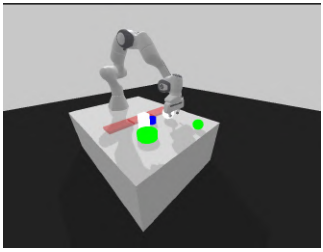
UAlberta

Sharan Vaswani

SFU

Reinforcement Learning with Function Approximation





- Tabular MDPs: methods have complexity that scales with $|\mathcal{S}|$
- Function Approximation:
  - make structural assumptions on the underlying MDPs
  - aim to find methods that have complexity independent of $|\mathcal{S}|$
- Linear MDP: r and P are linear in $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$

Motivations

- Most theoretical work on actor-critic methods in linear MDPs employs Natural Policy Gradient (NPG) [Kakade, 2001] as an actor

  $\pi_{t+1}(\cdot \mid s) \propto \pi_t(\cdot \mid s) \exp\left(\eta\, \widehat{Q}_t(s,\cdot)\right) \propto \exp\left(\eta \sum_{\tau=1}^{t} \widehat{Q}_\tau(s,\cdot)\right)$ "implicit" policy

- To facilitate exploration, Q-functions, provided by the critic, are often **non-linear** due to added bonuses and clipping

$$\widehat{Q}_{t,h}(s,a) = \mathrm{Clip}_{[0,H-h+1]}\left[\widehat{Q}_h^{\pi_t}(s,a) + b_t(s,a)\right]$$

- To use policy π_{t+1}, we need to store all the previous Q-functions $\widehat{Q}_1, \ldots, \widehat{Q}_t$
  - ✗ This results in a high runtime cost for policy inference
  - ✗ This deviates from practice, where people often use parametric policies as actors

Can we design provably efficient actor–critic algorithms with parametric policies?

Algorithm Design

- Log-Linear Policy Class: Given some policy feature $\mathbf{X} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$,

$$\pi_h(\cdot \mid s, \theta) \propto \exp(z(s, \cdot \mid \theta)) \quad \text{where } z(s, a \mid \theta) = \langle \mathbf{X}(s, a), \theta \rangle$$

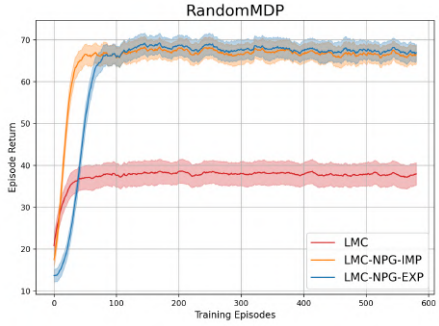
- Instantiating the Actor: Projected NPG
  - $\widehat{\ell}_{t,h}^{\text{actor}}(\theta) = \frac{1}{2} \sum_{(s,a) \in \mathcal{C}} \rho(s, a) \big[\langle \mathbf{X}(s, a), \theta \rangle - \underbrace{\Big(\langle \mathbf{X}(s, a), \theta_{t,h} \rangle + \eta \, \widehat{Q}_{t,h}(s, a) \Big)}_{z_{t+1/2}} \big]^2$
    - ✓ Project the implicit policies onto a parametric policy class
    - ✓ Control the projection error through experimental design
- Instantiating the Critic: Langevin Monte Carlo
  - $\ell_{t,h}^{\text{critic}}(w) = \frac{1}{2} \sum_{(s,a,s') \in \mathcal{D}_{t,h}} \big[\underbrace{r_h(s, a) + \widehat{V}_{t,h+1}(s')}_{\text{Bellman Target}} - \underbrace{\langle \phi(s, a), w \rangle}_{\text{Current Estimate}} \big]^2 + \frac{\lambda}{2} \|w\|^2$
  - $w_{t,j+1} = w_{t,j} - \alpha \nabla_w \ell_{t,h}^{\text{critic}}(w_{t,j}) + \sqrt{\alpha/\zeta} \, \nu_j$ where $\nu_j \leftarrow \mathbf{N}(0, I)$ is the Gaussian noise
    - ✓ Approximate Thompson sampling

Sample Complexity Analysis

$$\text{Sub-Opt}(\overline{\pi}_T) \leq \widetilde{\mathcal{O}}\Big(\frac{H^2\sqrt{d^4 \log |\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\bar{\epsilon}}\Big)$$

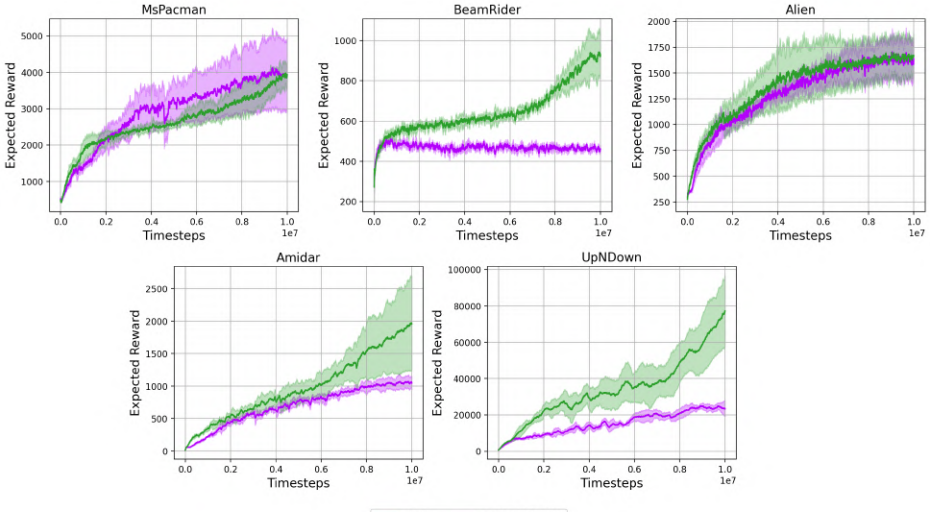
- $\bar{\epsilon}$ is the bound for projection error, which depends on the expressivity of **X**
- On-Policy: Collect N fresh trajectories per episode (throw away old trajectories)
  - ▶ Requires $\widetilde{\mathcal{O}}(1/\epsilon^4)$ to obtain a $(\epsilon + H^2\sqrt{\bar{\epsilon}})$-optimal mixture policy
  - ✓ Recover the results from Liu et al. [2023]
- Off-Policy: Collect 1 trajectory per episode and aggregate with old trajectories
  - ▶ Requires $\widetilde{\mathcal{O}}(1/\epsilon^2)$ to obtain a $(\epsilon + H^2\sqrt{\bar{\epsilon}})$-optimal mixture policy
  - ✓ Recover the results from Sherman et al. [2023], Cassel and Rosenberg [2024] w/o any bespoke tricks

Experiments in Linear MDP

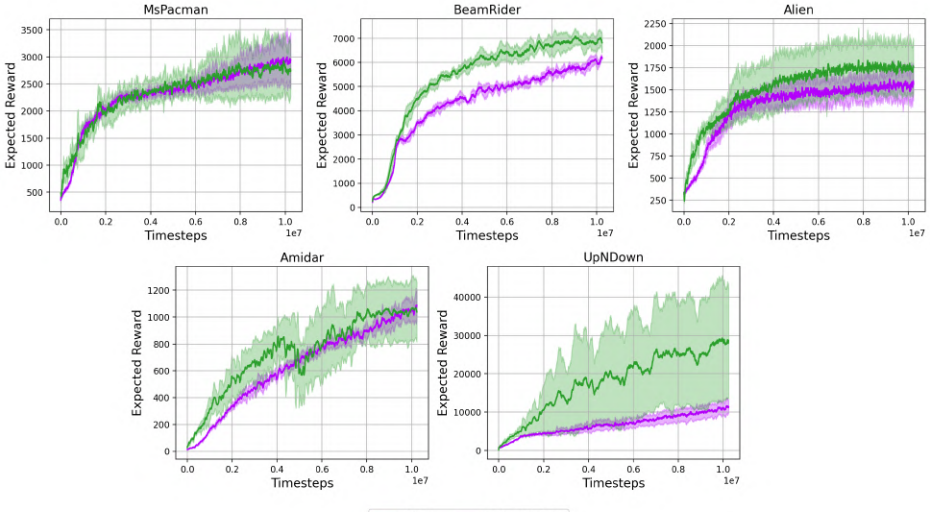


- We use the proposed optimistic actor-critic framework with Projected NPG (explicit policy) as the actor and LMC as the critic (`LMC-NPG-EXP`)
- We compare our method against the same framework with implicit policies (`LMC-NPG-IMP`) and the value-based baseline (LMC) [Ishfaq et al., 2024]

Experiments in Atari: Ours vs. PPO for On-Policy Learning



Experiments in Atari: Ours vs. DQN for Off-Policy Learning



Questions?

Contact: maxqslin@gmail.com

link to arXiv

References i

Asaf Cassel and Aviv Rosenberg. Warm-up free policy optimization: Improved regret in linear Markov decision processes. *Advances in Neural Information Processing Systems*, 37:3275–3303, 2024.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024.

Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvári. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *Advances in Neural Information Processing Systems*, 36:3560–3577, 2023.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.