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

SFU SIMON FRASER UNIVERSITY

Max Qiushi Lin[1] Reza Asad[1] Kevin Tan[2] Haque Ishfaq[3,4] Csaba Szepesvári [5,6] Sharan Vaswani[1]

[1] Simon Fraser University [2] University of Pennsylvania [3] Mila [4] McGill University [5] Google DeepMind [6] University of Alberta

Motivation

- ✓ Actor-Critic methods are empirically successful in both on-policy and off-policy settings
- Consequently, there have been many attempts to provide a theoretical understanding of actor-critic methods, especially under function approximation
- × However, prior work has some limitations:
  - does not consider strategic exploration in a systematic manner
  - only analyzes complicated and impractical variants of the algorithm
- × In particular, many theoretical papers (e.g. [1]) study actor-critic methods that use the natural policy gradient (NPG) update *without an explicit policy parameterization*

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

Contributions

- Propose a general optimistic actor-critic framework that uses parameterized policies
- Instantiate the actor with Projected NPG that enables us to directly control the error between the explicitly parametrized policy and the easy-to-analyze implicit policy
- Instantiate the critic with Langevin Monte Carlo (LMC), a more practical and easy-to-implement approach that offers similar guarantees as UCB bonuses
- Analyze the proposed actor-critic framework in both the on-policy and off-policy settings and recover the state-of-the-art sample complexity results without bespoke tricks

Problem Setup

- Episodic Linear MDP: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$
  - $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s,a), \psi(s') \rangle$ and $r_h(s,a) = \langle \phi(s,a), \upsilon_h \rangle$
  - $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is the given feature map ; $\psi_h : \mathcal{S} \to \mathbb{R}^d$ and $\upsilon_h : \mathbb{R}^d$ are unknown
- Learning Objective: We aim to minimize the optimality gap after T iterations

$$\mathrm{OG(T)} \coloneqq \mathbb{E}\left[\mathrm{V}_1^\star(\mathrm{s}_1) - \mathrm{V}_1^{\overline{\pi}^{\mathrm{T}}}(\mathrm{s}_1)\right] = \frac{\mathrm{Reg(T)}}{\mathrm{T}} \text{ where } \overline{\pi}^T \text{ is the mixture policy}$$

- Log-Linear Policy: We consider the parametric policy of the following form: $\forall s, a, h$,

$$\pi_h(a \mid s, \theta) = \frac{\exp(z_h(s, a \mid \theta_h))}{\sum_{a' \in \mathcal{A}} \exp(z_h(s, a' \mid \theta_h))},$$

where $z_h(s, a \mid \theta_h) = \langle \varphi(s,a), \theta_h \rangle$ represents the logits parameterized by θ_h, and $\varphi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ are policy features available to the learner

Optimistic Actor-Critic Framework

Optimistic Actor-Critic with Parametric Policies

Input: number of update steps T, data collection batch size N (only for on-policy)
set $\mathcal{D}^0 \leftarrow \emptyset$, $w_h^1 \leftarrow \mathbf{0}$, $\pi_h^1(\cdot \mid s) \leftarrow \mathcal{U}(\mathcal{A}) \quad \forall (h, s)$
for $t = 1, \ldots, T-1$ **do**
  Collect data: $\mathcal{D}^t \leftarrow \begin{cases} \text{On-Policy:} & \left\{ N \text{ fresh traj.} \overset{\text{i.i.d.}}{\sim} \pi^t \right\} \\ \text{Off-Policy:} & \mathcal{D}^{t-1} \cup \{1 \text{ traj.} \sim \pi^t\} \end{cases}$
  Update the critic: $w^{t+1} \leftarrow \mathrm{Critic}(\mathcal{D}^t, \pi^t, w^t)$
  Update the actor: $\theta^{t+1} \leftarrow \mathrm{Actor}(w^{t+1}, \theta^t)$
  Instantiate the parametric policy: $\pi^{t+1} = \pi(\theta^{t+1})$
Return: mixture policy $\overline{\pi}^T$

Instantiating the Actor: Projected Natural Policy Gradient

- Standard NPG update: $\pi_h^{t+1}(\cdot|s) \propto \pi_h^1(\cdot|s) \exp(\eta \sum_{i=1}^t \widehat{Q}_h^i(s, \cdot))$
  - × require storing *all the clipped* Q functions and requires memory linear in $|\mathcal{S}|$ or T
  - ✓ Easy to analyze using standard OMD regret bound
- Projected NPG:
  $\pi_h^{t+1}(\cdot \mid s) = \mathrm{Proj}_\Pi\left[\pi_h^{t+1/2}(\cdot \mid s)\right]; \quad \pi_h^{t+1/2}(\cdot|s) \propto \pi_h^t(\cdot|s) \exp(\eta\, \widehat{Q}_h^t(s, \cdot))$,
  where Π is the class of realizable policies. The above update results in policies that
  - ✓ Can be realized by the explicit actor parameterization
  - ✓ Can provably approximate the policy induced by the standard NPG update.
- It can be implemented by optimizing the following loss:

$$\widetilde{\ell}_h^t(\theta) = \frac{1}{2} \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \left[\left\langle \varphi(s,a), \theta - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,a) \right],$$

where $\mathcal{D}_{\text{exp}} \subseteq \mathcal{S} \times \mathcal{A}$ and $\rho_{\text{exp}} \in \Delta(\mathcal{D}_{\text{exp}})$ ensure exploration of the state-action space
- ✓ Can be constructed via experimental design, guaranteeing that $|\mathcal{D}_{\text{exp}}| = \widetilde{\mathcal{O}}(d)$
- ✓ Makes the actor loss calculation computationally efficient
- ✓ Can also be constructed via reward-free exploration (e.g. CoverTraj) or regret minimization algorithms (e.g., OptCov) before the learning procedure

Actor: Projected Natural Policy Gradient

Input: critic parameters w^t, policy optimization learning rate η, $\mathcal{D}_{\text{exp}}$ and ρ_{exp}
for $h = 1, 2, \ldots, H$ **do**
  $\widehat{Q}_h^t(\cdot, \cdot) = \mathrm{clip}_{[0, \mathrm{H-h+1}]} \left\{ \max_{\mathrm{m} \in [\mathrm{M}]} \left\langle \phi(\cdot, \cdot), \mathrm{w}_{\mathrm{h}}^{\mathrm{t,m,J_t}} \right\rangle \right\}$ ▷ clip to ensure boundedness
  $\widetilde{\ell}_h^t(\theta) = \frac{1}{2} \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \left[\left\langle \varphi(s,a), \theta - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,a) \right]$
  for $k = 1, \ldots, K_t$ **do**
    $\theta_h^{t,k} \leftarrow \theta_h^{t,k-1} - \alpha_a^t \nabla_\theta \widetilde{\ell}_h^t(\theta_h^{t,k-1})$ ▷ off-policy updates
Return: actor parameters for the policy θ^t

Theoretical Guarantees for Projected Natural Policy Gradient

Extended OMD Regret Bound

Let $\epsilon^t \coloneqq \mathrm{KL}(\pi^\star \parallel \pi^{t+1}) - \mathrm{KL}(\pi^\star \parallel \pi^{t+1/2})$ be the projection error. Then, for any comparator $\pi^\star \in \Delta(\mathcal{A})$, it holds that

$$\sum_{t=1}^T \left\langle \pi^\star - \pi^t, \widehat{Q}^t \right\rangle \le \frac{\log|\mathcal{A}| + \sum_{t=1}^T \epsilon^t}{\eta} + \frac{\eta\, H^2\, T}{2}.$$

- If $\epsilon^t = 0$ for all $t \in [T]$, it recovers the standard OMD regret bound
- Controlling the projection error: $\overline{\epsilon} \coloneqq \sum_{t=1}^T \epsilon^t$ can be controlled by constructing $\mathcal{D}_{\text{exp}}$, ρ_{exp} via experimental design, and optimizing $\widetilde{\ell}_h^t(\theta)$ appropriately

Instantiating the Critic: Langevin Monte Carlo

- Prior works in linear MDPs usually utilize UCB bonuses for exploration
  - × Difficult to implement beyond linear MDPs
- `LMC` iteratively adds Gaussian noise to the gradient descent updates, and produces approximate samples of the critic parameter from its posterior distribution
  - ✓ Easy to implement even beyond the linear function approximation setting
  - ✓ Provides similar theoretical guarantees as UCB bonuses

Critic: Langevin Monte-Carlo (LMC)

Input: data $\mathcal{D}^t$, policy π^{t-1}, inverse temperature ζ, number of critic samples M
$\widehat{V}_{H+1}^t(\cdot) \leftarrow 0$
for $h = H, H-1, \ldots, 1$ **do**
  $\mathcal{L}_h^t(w) = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}^t|} \left[r_h(s_h^i, a_h^i) + \widehat{V}_{h+1}^t(s_{h+1}^i) - \left\langle \phi(s_h^i, a_h^i), w \right\rangle \right]^2 + \frac{\lambda}{2} \|w\|^2$
  $w_h^{t,m,0} \leftarrow w_h^{t-1,m,J_{t-1}} \quad \forall m \in [M]$
  for $j = 1, \ldots, J_t$ **do**
    $\nu_h^{t,m,j} \leftarrow \mathbf{N}(0, I) \quad \forall m \in [M]$ ▷ sample from the standard Gaussian distribution
    $w_h^{t,m,j} \leftarrow w_h^{t,m,j-1} - \alpha_c^{h,t} \nabla_w \mathcal{L}_h^t(w_h^{t,m,j-1}) + \sqrt{\alpha_c^{h,t}/\zeta}\, \nu_h^{t,m,j} \quad \forall m \in [M]$
    ▷ Langevin dynamic update
  $\widehat{Q}_h^t(\cdot, \cdot) = \mathrm{clip}_{[0, \mathrm{H-h+1}]} \left\{ \max_{\mathrm{m} \in [\mathrm{M}]} \left\langle \phi(\cdot, \cdot), \mathrm{w}_{\mathrm{h}}^{\mathrm{t,m,J_t}} \right\rangle \right\}$ ▷ clip to ensure boundedness
  $\widehat{V}_h^t(\cdot) = \mathbb{E}_{a \sim \pi^{t-1}(\cdot|s)} \widehat{Q}_h^t(\cdot, a)$
Return: critic parameters for the estimated Q-function $\{w_h^{t,m,J_t}\}_{(m,h) \in [M] \times [H]}$

Theoretical Guarantees of Langevin Monte Carlo

Optimism and Error Bound

Let $\iota_h^t(s,a) \coloneqq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)$ be the model projection error and $\Lambda_h^t \coloneqq \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a)\, \phi(s,a)^\top + \lambda\, I$. Using `LMC` in our proposed framework can guarantee that

$$-\Gamma_{\texttt{LMC}} \times \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \le \iota_h^t(s,a) \le 0\,,$$

where $\Gamma_{\texttt{LMC}} \in \widetilde{\mathcal{O}}(Hd)$ is a problem-dependent constant

- $\iota_h^t(s,a) \le 0$ implies optimism (i.e., $\widehat{Q}_h^t(s,a) \le r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)$)
- $\iota_h^t(s,a) \ge -\Gamma_{\texttt{LMC}} \times \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$ holds for both on-policy and off-policy settings
  - On-policy: $\mathcal{D}^t$ is sampled only by $\pi^t \implies$ Use Self-Normalized Concentration
  - Off-policy: $\mathcal{D}^t$ is sampled by $\{\pi^1, \ldots \pi^t\} \implies$ Use Value-Aware Uniform Concentration
    - ✓ log-covering number $\log(\mathcal{N}_\Delta(\mathcal{V}))$ is bounded since we are projecting onto a fixed policy class

Sample Complexity in On-Policy and Off-Policy Settings

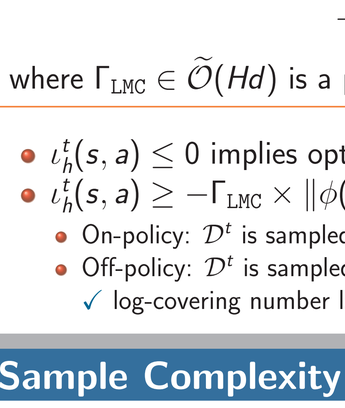
Sample Complexity

$$\mathrm{OG(T)} \lesssim \underbrace{\frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^\star}\left[\left\langle \pi_h^\star - \pi_h^t, \widehat{Q}_h^t \right\rangle \right]}_{\substack{\text{policy optimization (actor) error} \\ \text{bounded by the extended OMD regret}}} + \underbrace{\frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H (\mathbb{E}_{\pi^\star}[\iota_h^t] - \mathbb{E}_{\pi^t}[\iota_h^t])}_{\substack{\text{policy evaluation (critic) error} \\ \text{bounded by optimism and error bound}}}$$

$$\lesssim \widetilde{\mathcal{O}}\left(\frac{H^2 \sqrt{\log|\mathcal{A}|}}{\sqrt{T}} + H^2 \sqrt{\overline{\epsilon}} \right) + \begin{cases} \widetilde{\mathcal{O}}(\sqrt{d^3\, H^4\, \log^2(N/\delta)/N}) & \text{on-policy} \\ \widetilde{\mathcal{O}}(\sqrt{d^3\, H^4/T} & \text{off-policy} \end{cases}$$

- On-policy: Setting $N = d^3\, T/(H^2 \log|\mathcal{A}|) \implies \widetilde{\mathcal{O}}(1/\epsilon^4)$ sample complexity
- Off-policy: $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity

Experiments



- `LMC-NPG-EXP` uses our proposed actor-critic framework that is instantiated with Projected NPG and LMC
- `LMC-NPG-IMP` uses the standard NPG without parametric policies and requires storing all the past Q functions (hence memory intensive)
- `LMC` [2] is the value-based method (i.e., $\pi_h(s) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_h(s,a)$)
- `LMC-NPG-EXP` matches `LMC-NPG-IMP` while outperforming `LMC`

References

[1] Liu, Q., Weisz, G., György, A., Jin, C., Szepesvári, C. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. NeurIPS 2023

[2] Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., Azizzadenesheli, K. Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo. ICLR 2024