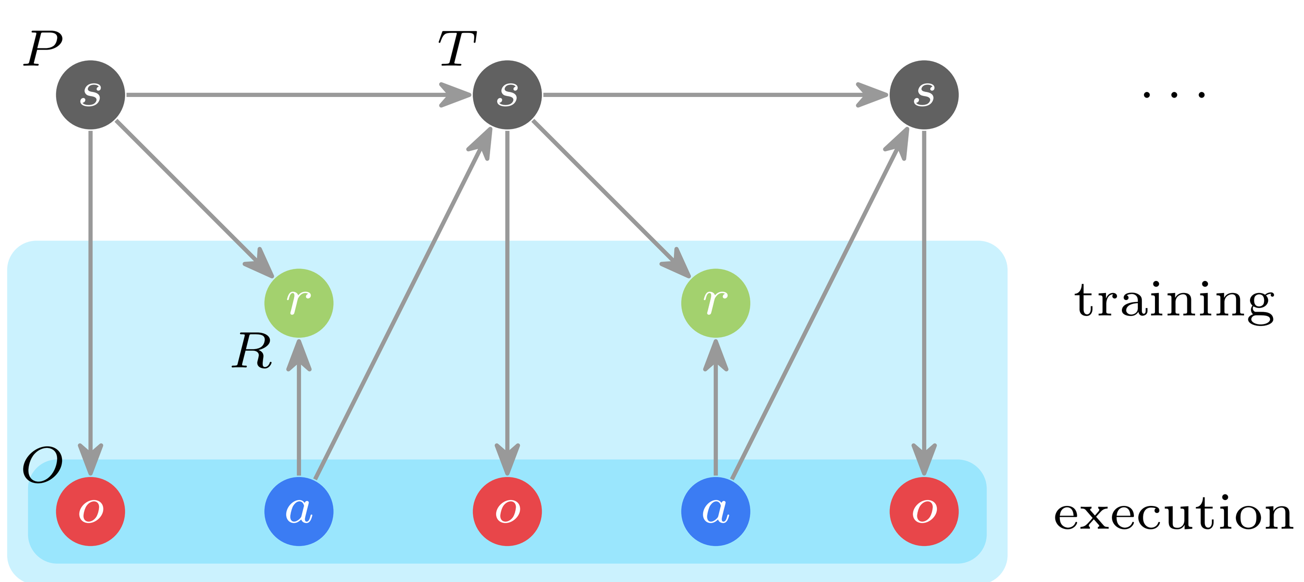
A Theoretical Justification for Asymmetric Actor-Critic Algorithms

Gaspard Lambrechts, Damien Ernst, Aditya Mahajan

Partial Observability

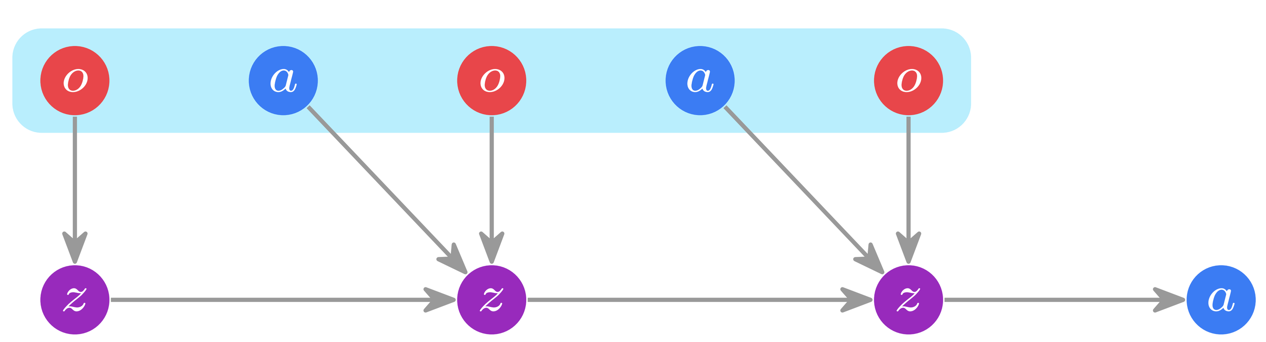
We consider a **POMDP** $(\mathcal{S}, \mathcal{A}, \mathcal{O}, P, O, T, R, \gamma)$:

- States $s_t \in \mathcal{S}$,
- Actions $a_t \in \mathcal{A}$,
- **Observations** $o_t \in \mathcal{O}$,
- Initialization $s_0 \sim P(\cdot)$,
- **Perception** $o_t \sim O(\cdot|s_t)$,
- Transition $s_{t+1} \sim T(\cdot|s_t, a_t)$,
- Reward $r_t \sim R(\cdot|s_t, a_t)$,
- Discount $\gamma \in [0, 1)$.



Given an **agent state** $z = f(h)$, recurrent s.t. $f(h') = u(f(h), a, o')$, we want an optimal **agent-state policy** $\pi^* \in \operatorname*{argmax}_{\pi \in \Pi} J(\pi)$ with $\Pi = \mathcal{Z} \to \Delta(\mathcal{A})$ and,

$$J(\pi) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R_t \right].$$



Asymmetric Observability

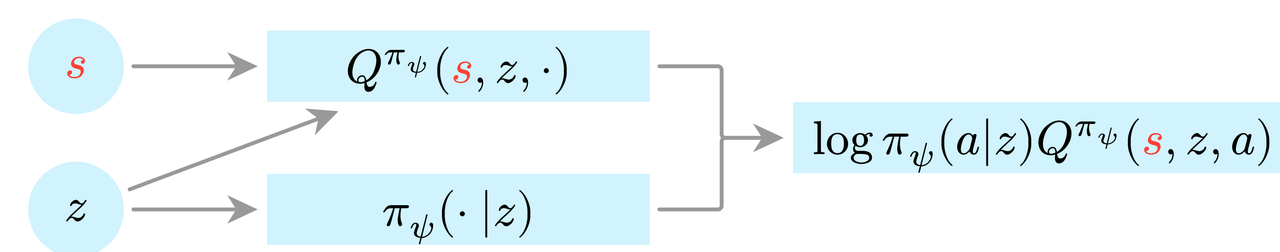
Partial observability is more realistic than **full observability**. But in some cases, the state may still be available during training.

Decision Process	Execution	Training
MDP	s	s
POMDP	o	o
Privileged POMDP	o	s + o

Asymmetric RL leverages the state at training time to learn faster.

Asymmetric Actor-Critic

In **actor-critic** methods, the critic is not needed at execution.
⇒ The critic can be **informed** with the state: $Q^\pi(z, a) \to Q^\pi(s, z, a)$.



While the asymmetric policy gradient is **unbiased** compared to the symmetric one [1], a **theoretical justification for its benefits is still missing**.

Proposed Analysis

We provide a **theoretical justification** by adapting a **finite-time bound** for symmetric actor-critic [2] to the asymmetric setting.

- **Linear finite-state critics:**
  - $\hat{\mathcal{Q}}^\pi_\beta(s, z, a) = \langle \beta, \varphi(s, z, a) \rangle$ and $\hat{Q}^\pi_\beta(z, a) = \langle \beta, \chi(z, a) \rangle$.
- **Log-linear finite-state policy:**
  - $\pi_\theta(a|z) \propto \exp(\langle \theta, \psi(z, a) \rangle)$.

Algorithm 1. (A)symmetric natural actor-critic.

1. Initialize policy parameters ψ_0.
2. For $t = 1 \ldots T$:
   1. Estimate $\hat{\mathcal{Q}}^{\pi_\psi}_\varphi \approx \mathcal{Q}^{\pi_\psi}$ or $\hat{Q}^\pi_\chi \approx Q^{\pi_\psi}$.
      - **TD learning** for K steps.
   2. Estimate $g_{t-1} \approx F^\dagger_{\pi_{\psi_{t-1}}} \nabla_\psi J(\pi_{\psi_{t-1}})$ with $\hat{\mathcal{Q}}^{\pi_\psi}_\varphi$ or $\hat{Q}^{\pi_\psi}_\chi$.
      - **NPG estimation** for N steps.
   3. Update policy $\psi_t = \psi_{t-1} + \eta g_{t-1}$.
3. Return π_{ψ_T}.

Because we use **TD learning with agent states**, we note that:

- The fixed point $\tilde{\mathcal{Q}}^\pi$ of the asymmetric Bellman operator is $\mathcal{Q}^\pi$,
- The fixed point $\tilde{Q}^\pi$ of the symmetric Bellman operator is not Q^π.

Using the belief $b(s|h) = \Pr(s|h)$ and approximate belief $\hat{b}(s|z) = \Pr(s|z)$, we introduce a **measure of aliasing** for the agent state.

Definition 1. Aliasing measure.

$$\varepsilon_{\text{alias/inf}} \propto \mathbb{E}\left[\left\| b(\cdot|h) - \hat{b}(\cdot|z) \right\| \right]$$

Finite-Time Bounds

1 **Theorem 1.** For any $\pi \in \Pi$ and any $m \in \mathbb{N}$, these finite-time bounds hold for **TD learning** with $\alpha = \frac{1}{K}$.

$$\begin{aligned} \sqrt{\mathbb{E}\left[\left\| \mathcal{Q}^\pi - \overline{\mathcal{Q}}^\pi \right\|^2_{d^\pi} \right]} &\leq \varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}} \\ \sqrt{\mathbb{E}\left[\left\| Q^\pi - \overline{Q}^\pi \right\|^2_{d^\pi} \right]} &\leq \varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}} + \varepsilon_{\text{alias}} \end{aligned} \tag{1}$$

$$\varepsilon_{\text{td}} = \sqrt{\frac{4B^2 + \left(\frac{1}{1-\gamma} + 2B \right)^2}{2\sqrt{K}(1 - \gamma^m)}}$$

$$\varepsilon_{\text{app}} = \frac{1 + \gamma^m}{1 - \gamma^m} \min_{f \in \mathcal{F}^B_\varphi} \| f - Q^\pi \|_{d^\pi}$$

$$\varepsilon_{\text{shift}} = \left(B + \frac{1}{1-\gamma} \right) \sqrt{\frac{2\gamma^m}{1 - \gamma^m} \sqrt{\| d^\pi_m \otimes \pi - d^\pi \otimes \pi \|_{\text{TV}}}}$$

$$\varepsilon_{\text{alias}} = \frac{2}{1-\gamma} \left\| \mathbb{E}^\pi \left[\sum_{k=0}^{\infty} \gamma^{km} \left\| \hat{b}_{km} - b_{km} \right\|_{\text{TV}} \middle| Z_0 = \cdot, A_0 = \cdot \right] \right\|_{d^\pi}$$

2 **Theorem 2.** For any $f : \mathcal{H} \to \mathcal{Z}$, this finite-time bound holds for **Algorithm 1** with $\alpha = \frac{1}{K}$, $\zeta = \frac{B\sqrt{1-\gamma}}{\sqrt{2N}}$ and $\eta = \frac{1}{\sqrt{T}}$.

$$\begin{aligned} &(1 - \gamma) \min_{0 \leq t < T} \mathbb{E}[J(\pi^*) - J(\pi_t)] \\ &\leq \varepsilon_{\text{nac}} + \varepsilon_{\text{actor}} + \varepsilon_{\text{grad}} + \varepsilon_{\text{inf}} + \frac{1}{T} \sum_{t=0}^{T-1} \varepsilon^{\pi_t}_{\text{critic}} \end{aligned} \tag{2}$$

$$\varepsilon_{\text{nac}} = \frac{B^2 + 2\log|A|}{2\sqrt{T}} \qquad \varepsilon_{\text{actor}} = \overline{C}_\infty \sqrt{\frac{(2-\gamma)B}{(1-\gamma)\sqrt{N}}}$$

$$\varepsilon^{\text{asym}}_{\text{grad}} = 2\overline{C}_\infty \sup_{0 \leq t < T} \sqrt{\min_w \mathcal{L}_t(w)} \qquad \varepsilon^{\text{sym}}_{\text{grad}} = 2\overline{C}_\infty \sup_{0 \leq t < T} \sqrt{\min_w L_t(w)}$$

$$\varepsilon^{\text{asym}}_{\text{inf}} = 0 \qquad \varepsilon^{\text{sym}}_{\text{inf}} = 2\mathbb{E}^{\pi^*} \left[\sum_{k=0}^{\infty} \gamma^k \left\| \hat{b}_k - b_k \right\|_{\text{TV}} \right]$$

$$\varepsilon^{\pi_t}_{\text{critic}} = 2\overline{C}_\infty \sqrt{6} (\text{RHS of } (1))$$

Conclusion

Asymmetric learning is less sensitive to aliasing in the agent state.

Future works:

- Consider learnable agent states or nonlinear approximators,
- Relax some assumptions (iid sampling and concentrability) [3],
- Generalize to non Markovian additional information.

[1] A. Baisero and C. Amato, "Unbiased Asymmetric Reinforcement Learning under Partial Observability," *AAMAS*, 2022.

[2] S. Cayci, N. He, and R. Srikant, "Finite-Time Analysis of Natural Actor-Critic for POMDPs," *SIMODS*, 2024.

[3] Y. Cai, X. Liu, A. Oikonomou, and K. Zhang, "Provable Partially Observable Reinforcement Learning with Privileged Information," *NeurIPS*, 2024.

arXiv:2501.19116