Informed Asymmetric Actor-Critic: Theoretical Insights and Open Questions

EWRL 2025 | Tübingen, Germany

Daniel Ebi (daniel.ebi@kit.edu) | Gaspard Lambrechts | Damien Ernst | Klemens Böhm

Asymmetric Observability

In many RL applications, agents must act under **partial observability**. However, assuming **identical observability** during training and execution is often **too restrictive**.

Asymmetric RL relaxes this assumption: It leverages additional state information at training time to accelerate learning.

Informed Partially Observable MDP

We consider an **informed POMDP** [1] $(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, I, O, R, P, \gamma)$:

- States $\boldsymbol{s}_t \in \mathcal{S}$,
- Actions $\boldsymbol{a}_t \in \mathcal{A}$,
- Information $\boldsymbol{i}_t \in \mathcal{I}$,
- Observations $\boldsymbol{o}_t \in \mathcal{O}$,
- Transition $\boldsymbol{s}_{t+1} \sim T(\cdot \mid \boldsymbol{s}_t, \boldsymbol{a}_t)$,
- Cognition $\boldsymbol{i}_t \sim I(\cdot \mid \boldsymbol{s}_t)$,
- Perception $\boldsymbol{o}_t \sim O(\cdot \mid \boldsymbol{i}_t)$,
- Rewards $r_t = R(\boldsymbol{s}_t, \boldsymbol{a}_t)$,
- Initial state $\boldsymbol{s}_0 \sim P(\cdot)$,
- Discount $\gamma \in [0, 1)$.

At execution, the agent observes $\boldsymbol{o}_t \sim \widetilde{O}(\cdot \mid \boldsymbol{s}_t) = \sum_{\boldsymbol{i} \in \mathcal{I}} O(\cdot \mid \boldsymbol{i})\ I(\boldsymbol{i} \mid \boldsymbol{s}_t)$.

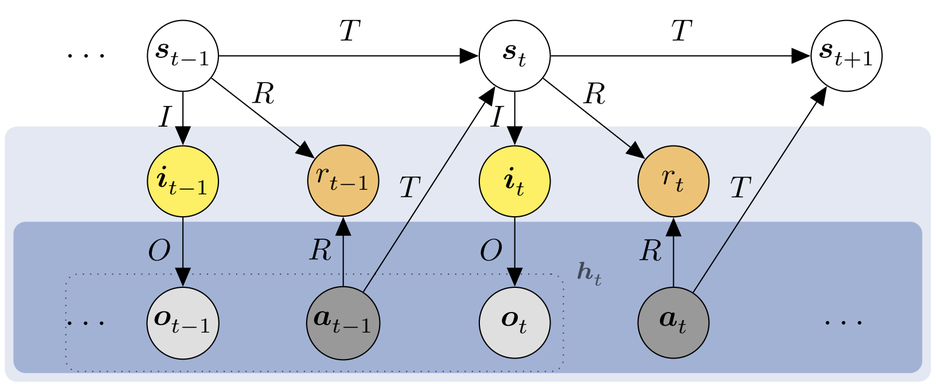

Figure 1 Bayesian network of an informed POMDP.

Objective: Given histories $\boldsymbol{h}_t = \begin{bmatrix}\boldsymbol{o}_0\ \boldsymbol{a}_0\ \dots\ \boldsymbol{o}_t\end{bmatrix}^\top \in \mathcal{H}$, we want the agent to learn a policy $\pi : \mathcal{H} \to \Delta(\mathcal{A})$ that maximizes $J(\pi) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$.

Unbiased Learning with Informed Critic

In **actor-critic** methods, the critic is not used during execution.
→ It can leverage additional information, typically the true state $\boldsymbol{s}_t$ [2].

Informed history-dependent value functions

We introduce $Q^\pi(\boldsymbol{h}, \boldsymbol{i}, \boldsymbol{a})$ and $V^\pi(\boldsymbol{h}, \boldsymbol{i})$ and prove that they are unbiased, i.e., $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}[Q^\pi(\boldsymbol{h}, \mathbf{i}, \boldsymbol{a})] = Q^\pi(\boldsymbol{h}, \boldsymbol{a})$ and $\mathbb{E}_{\mathbf{i}|\boldsymbol{h}}[V^\pi(\boldsymbol{h}, \mathbf{i})] = V^\pi(\boldsymbol{h})$.

Informed asymmetric policy gradient

We derive the informed asymmetric policy gradient $\nabla_\theta^{\text{IAAC}} J(\pi_\theta)$ and establish its equivalence to the symmetric policy gradient.

Theorem 1. Informed asymmetric policy gradient equivalence.

$$\nabla_\theta^{\text{IAAC}} J(\pi_\theta) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t\ Q^{\pi_\theta}(\boldsymbol{h}_t, \boldsymbol{i}_t, \boldsymbol{a}_t)\ \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t \mid \boldsymbol{h}_t)\right]$$

$$\overset{(\star)}{=} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t\ Q^{\pi_\theta}(\boldsymbol{h}_t, \boldsymbol{a}_t)\ \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t \mid \boldsymbol{h}_t)\right] =: \nabla_\theta\ J(\pi_\theta)$$

$(\star)$ Follows from expectation properties and our unbiasedness results.

Hence, informing the critic with state-conditioned information yields unbiased policy gradients beyond the special case $\boldsymbol{i}_t = \boldsymbol{s}_t$.

Informed Asymmetric Rec-NAC

We propose an **informed asymmetric variant** of the recurrent natural actor-critic (Rec-NAC) algorithm [3], where the **informed critic** receives additional input $\boldsymbol{i}_t$, breaking symmetry with the actor.

Algorithm 1. Recurrent Natural Actor-Critic (Rec-NAC)

1. Initialize actor parameters θ_0 and critic parameters ϑ_0.
2. For $t = 0$ to $T-1$:
   a. Estimate $\widehat{Q}_{\bar{\vartheta}}^{\pi_{\theta_t}}$ with $\bar{\vartheta} = \frac{1}{K}\sum_{k=0}^{K-1} \vartheta_k$. (via K steps of **Rec-TD**)
   b. Estimate gradient $\hat{\boldsymbol{g}}_{t,n}$. (via N steps of **Rec-NPG**)
   c. Update policy parameters: $\theta_{t+1} = \theta_t + \eta_{\text{npg}} \cdot \frac{1}{N}\sum_{n=0}^{N-1} \hat{\boldsymbol{g}}_{t,n}$.
3. Return final policy π_{θ_T}.

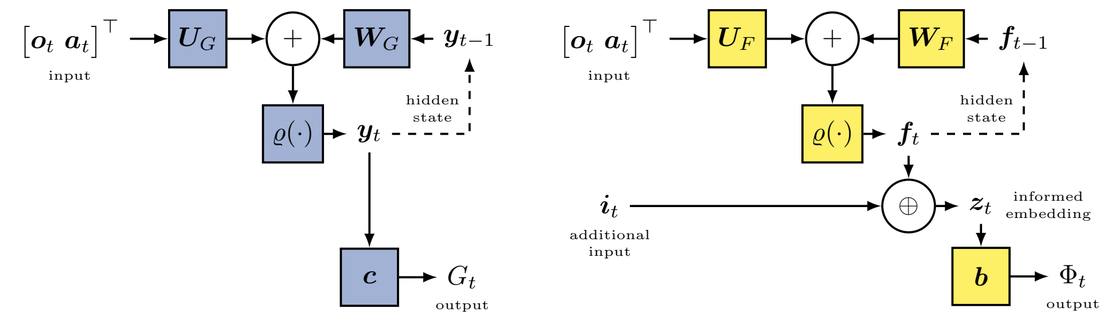

Figure 2 Architecture of the actor (left) and the informed critic (right).

Finite-Time and Finite-Width Analysis

We study the effect of an informed critic on Rec-TD convergence.

Theorem 2. Error bound of informed asymmetric Rec-TD.

Under standard smoothness and regularization assumptions, for a critic of width $m_F \in \mathbb{N}$ and $K \in \mathbb{N}$ gradient steps:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \mathcal{R}_T^\pi(\vartheta_k)\right] \leq \frac{\|\boldsymbol{\nu}\|_2^2}{\eta_{td}\ (1-\gamma)\sqrt{K}} + \frac{\eta_{td}\ C_T^{(1)}}{(1-\gamma)^3\sqrt{K}} + \frac{C_T^{(2)}}{(1-\gamma)^2\sqrt{m_F}} + \frac{\gamma^T}{(1-\gamma)K}\sum_{k=0}^{K-1} \omega_{T,k}^2.$$

Trade-off: increased complexity vs. approximation gain

Rec-TD with informed critic achieves the **same asymptotic rate** as in the symmetric case [3], but the error bound reveals a key **trade-off**:

↑ **Complexity**
- Generally larger Lipschitz and smoothness constants.
- Increased dimensionality of parameter-related quantities.

↓ **Approximation error**
- Richer function class from additional information.
- Informative $\boldsymbol{i}_t$ may mitigate partial observability.

Conclusion

Additional inputs improve learning only if their informativeness outweighs the added complexity.

Central question for future work:
How can we formalize this trade-off to guide auxiliary input design?

References
1. G. Lambrechts, A. Bolland, and D. Ernst. Informed POMDP: Leveraging additional information in model-based RL. Reinforcement Learning Journal, 2024.
2. A. Baisero and C. Amato. Unbiased asymmetric reinforcement learning under partial observability. 21st Intl. Conference on Autonomous Agents and Multiagent Systems, 2022.
3. S. Cayci and A. Eryilmaz. Recurrent natural policy gradient for POMDPs. 17th European Workshop of Reinforcement Learning (EWRL), 2024.

Paper