Informed Asymmetric Actor-Critic: Leveraging Privileged Signals Beyond Full-State Access

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

Asymmetric Observability

In many RL applications, agents must act under **partial observability**. Yet, **additional state information** may be available **during training**.

Asymmetric RL exploits this information to accelerate learning.

In **actor-critic** methods, the critic is not used during execution.
→ It can leverage additional information, typically the full state s_t [1].

Informed Partially Observable MDP

We consider an **informed POMDP** $(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, I, O, R, P, \gamma)$ [2].
It extends a POMDP with **training-time information** about the state.

- States $s_t \in \mathcal{S}$ with $s_0 \sim P(\cdot)$, actions $a_t \in \mathcal{A}$, and observations $o_t \in \mathcal{O}$.
- **Information** $i_t \in \mathcal{I}$, where $i_t \sim I(\cdot|s_t)$ and $o_t \sim O(\cdot|i_t)$.
- Dynamics $s_{t+1} \sim T(\cdot|s_t, a_t)$ with rewards $r_t = R(s_t, a_t)$, and $\gamma \in [0, 1)$.

At execution, the agent observes $o_t \sim \tilde{O}(\cdot|s_t) = \mathbb{E}_{i_t|s_t}\big[O(o_t|i_t)\big]$.

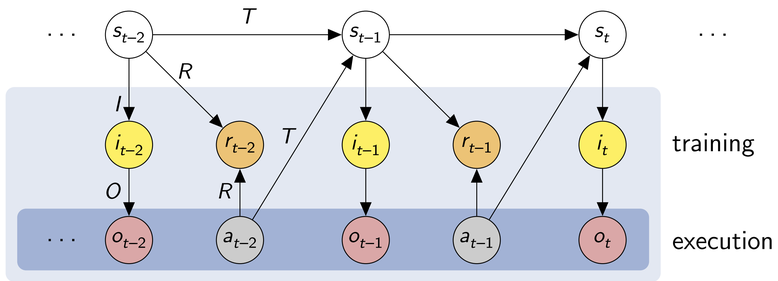

Figure 1 Bayesian network of an informed POMDP.

Objective: Given histories $h_t = \big[o_0\ a_0\ \dots\ o_t\big] \in \mathcal{H}$, we want to learn a policy $\pi : \mathcal{H} \to \Delta(\mathcal{A})$ that maximizes $J(\pi) = \mathbb{E}^{\pi}\big[\sum_{t=0}^{\infty} \gamma^t r_t\big]$.

Unbiased Learning with Informed Critic

Informed History-Dependent Value Functions

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

Informed Asymmetric Policy Gradient

We derive a **new expression for the policy gradient** $\nabla_\theta J(\pi_\theta)$ that uses $Q^{\pi}(h_t, i_t, a_t)$.

Hence, informing the critic with **any state-conditioned** privileged signal $i_t \sim I(\cdot|s_t)$ yields **unbiased policy gradients**, even when $i_t \neq s_t$.

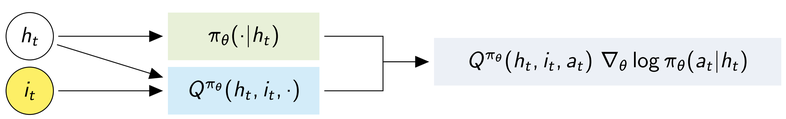

Figure 2 Informed asymmetric actor-critic architecture.

Results

Main Theoretical Result

Theorem 1. Policy gradient equivalence under informed critics.

$$\nabla_\theta^{\text{IAAC}} J(\pi_\theta) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t\ Q^{\pi_\theta}(h_t, i_t, a_t)\ \nabla_\theta \log \pi_\theta(a_t|h_t)\right]$$
$$= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t\ Q^{\pi_\theta}(h_t, a_t)\ \nabla_\theta \log \pi_\theta(a_t|h_t)\right] =: \nabla_\theta\ J(\pi_\theta)$$

Learning Performance on Navigation Tasks and POPGym Suite [3]

We evaluate the informed critic against symmetric and full-state critics.

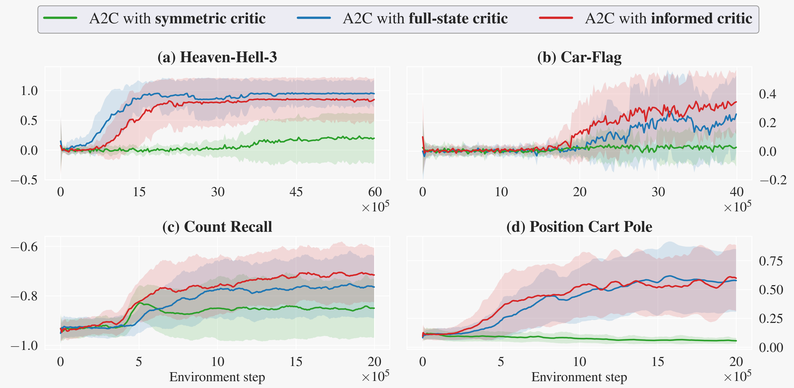

Figure 3 Episodic returns on four benchmark environments (mean and standard deviation across 20 runs).

Signal Evaluation on Synthetic Informed POMDPs

We construct i_t by selecting feature subsets from the state vector s_t.

Privileged signal	Residual Inform.		Prediction Inform.		AUC (↑)
	ρ_{obs} (↑)	p-value (↓)	L_τ (↑)	p-value (↓)	(sum of returns)
$i_t = [s_t^1, s_t^2]$	3.0e-05	0.438	-1.4e-02	0.801	1.07e+05
$i_t = [s_t^1, s_t^2, s_t^3]$	5.5e-05	0.170	0.007	0.397	1.08e+05
$i_t = [s_t^1, s_t^2, s_t^4]$	7.6e-05	0.119	0.035	0.043	1.08e+05
$i_t = [s_t^1, s_t^2, s_t^5]$	5.0e-05	0.191	0.018	0.232	1.10e+05
$i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$	7.4e-05	0.068	0.046	0.009	1.07e+05
$i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$	5.8e-05	0.106	0.026	0.120	1.11e+05
$i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$	7.6e-05	0.035	0.064	0.003	1.23e+05
$i_t = [s_t^1, s_t^2, s_t^3, s_t^4, s_t^5]$	7.0e-05	0.038	0.056	0.072	1.19e+05

Table 1 Effect of signal choice on informativeness and return (means across 10 runs).

- **Return-relevant signals** improve value estimation and policy learning.
- **Irrelevant/redundant signals** are less helpful for return prediction.
- **Full-state critics** are not always the best choice.

Informativeness of Privileged Signals

If any $i_t \sim I(\cdot|s_t)$ can be used, which signals help learning?
→ We propose **two informativeness criteria**.

Residual Informativeness (Pre-Training Selection)

Idea: Test whether i_t contains additional information about future returns $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ beyond (h_t, a_t).

Algorithm 1. Test for α-residual informativeness.

1. Estimate conditional means $\mathbb{E}[G_t|h_t, a_t]$ and $\mathbb{E}[i_t|h_t, a_t]$.
2. Compute residuals $\hat{e}_{G_t} := G_t - \widehat{\mathbb{E}}[G_t|h_t, a_t]$ and $\hat{e}_{i_t} := i_t - \widehat{\mathbb{E}}[i_t|h_t, a_t]$.
3. Test whether $\hat{e}_{G_t} \perp\!\!\!\perp \hat{e}_{i_t}$ at significance level $\alpha > 0$.

Interpretation: Rejecting residual independence indicates that i_t is informative about G_t beyond (h_t, a_t).

Informativeness via Return Prediction Gain (Post-Hoc Evaluation)

Idea: Test whether i_t improves the critic's prediction of future returns.

Algorithm 2. Test for (ϵ, δ)-prediction informativeness.

1. Train critics without and with access to i_t: $\widehat{Q}(h_t, a_t)$ and $\widehat{Q}(h_t, i_t, a_t)$.
2. For each episode τ_n with $n = 0, \dots, N-1$:
   a. Compute the episode-level mean prediction gain
$$L^{\tau_n} := \frac{1}{T_n} \sum_{t=0}^{T_n - 1} \Big(\big(\widehat{Q}(h_t, a_t) - G_t\big)^2 - \big(\widehat{Q}(h_t, i_t, a_t) - G_t\big)^2 \Big).$$
3. Test whether $\mathbb{E}[L^{\tau}] > \epsilon$ for $\epsilon = 0$ at significance level $\delta > 0$.

Interpretation: A positive gain indicates improved return prediction when i_t is used as additional critic input.

Selecting Helpful Privileged Signals

We **evaluate all candidate signals** using the informativeness criteria, retain statistically significant signals, and **prioritize higher effect sizes**.

Conclusion

Any state-conditioned privileged signal preserves unbiased policy updates, but selecting informative signals is crucial for learning.

Future work:

- Extend criteria to account for signal complexity or direct policy effects.
- Integrate signal selection end-to-end within the learning loop.
- Explore critics that condition on histories of privileged signals.

References

1. A. Baisero and C. Amato. Unbiased asymmetric reinforcement learning under partial observability. 21st International Conference on Autonomous Agents and Multiagent Systems, 2022.
2. G. Lambrechts, A. Bolland, and D. Ernst. Informed POMDP: Leveraging additional information in model-based RL. Reinforcement Learning Journal, 2024.
3. S. Morad, R. Kortvelesy, M. Bettini, S. Liwicki, and A. Prorok. POPGym: Benchmarking partially observable reinforcement learning. 11th International Conference on Learning Representation, 2023.