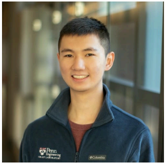
Partial Observability and Asymmetric Observability

BeNeRL Workshop - July 4th, 2025

Gaspard Lambrechts

Outline

Adrien Bolland
ULiège



Edward Hu
UPenn

Daniel Ebi
KIT

Aditya Mahajan
McGill

Damien Ernst
ULiège

Partial Observability

A matter of perception

Intelligence is usually understood as the ability to make **decisions**, *based on **perception***, in order to achieve an **objective** (McCarthy, 1998).

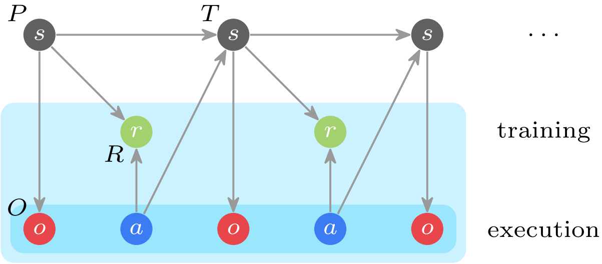
⇒ Intelligence is about (i) perceiving and abstracting past information about the world for (ii) acting on its future execution to achieve an objective.

In RL, we model these aspects as:

- **Perception**: past observations.
- **Decision**: current action.
- **Objective**: future rewards.

Unfortunately, RL overlooked (i) to focus on (ii), by assuming full observability.

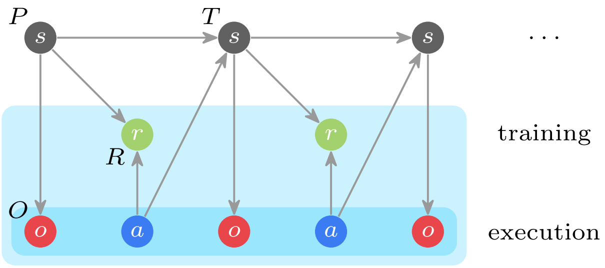
Partially observable Markov decision process



A **POMDP** is described by a model $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, O, P, \gamma)$.

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

History-dependent policies

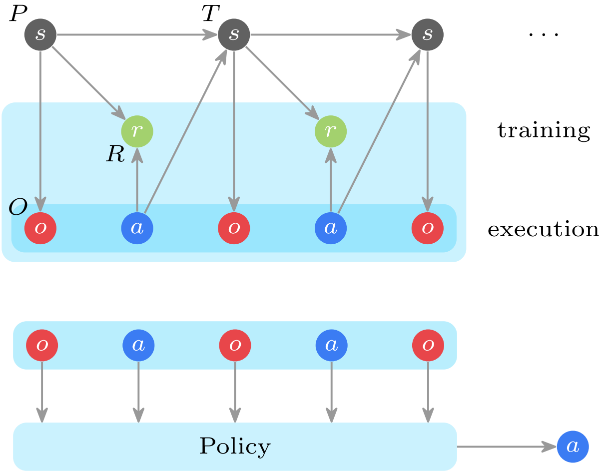


The **history** at time t is $h_t = (o_0, a_0, ..., o_t) \in \mathcal{H}$.

Definition 1: History-dependent policy.

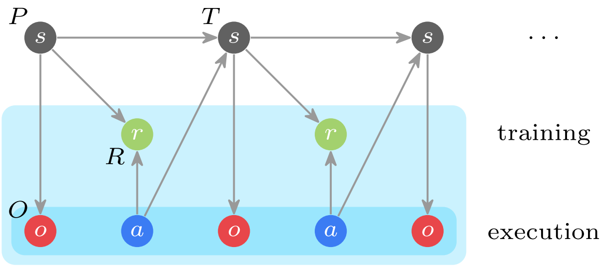
A history-dependent policy $\eta \in \mathrm{H} = \mathcal{H} \to \Delta(\mathcal{A})$ is a mapping from histories to distributions over the actions, with density $\eta(a|h)$.

History-dependent policies (ii)



NB: POMDP $\approx$ MDP whose state is the history: the “history MDP”.

Optimal control under partial observability



The problem of **RL in POMDP** is to find an optimal history-dependent policy,

$$\eta^* \in \underset{\eta \in \mathrm{H}}{\operatorname{argmax}} \underbrace{\mathbb{E}^{\eta}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]}_{J(\eta)},$$

from samples $(o_0, a_0, r_0, ..., o_n)$.

Structure of the optimal policy

Definition 2: Belief of a history.

The belief $b = f(h)$ of a history $h \in \mathcal{H}$ is defined as,

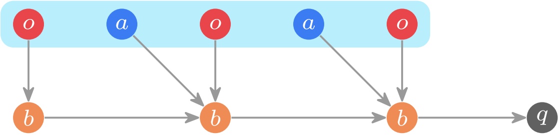
$$b(s) = \Pr(s|h).$$

Theorem 1: Belief recurrence.

$$f(h') = u(f(h), a, o').$$

Theorem 2: Belief sufficiency.

$$Q(h, a) = Q'(f(h), a).$$



⇒ If the belief is known, we can discard the history. But it is **usually unknown**.

History-dependent reinforcement learning

We use **function approximators** for the **policy** or **Q-function** estimation.

- Feedforward network, transformer, recurrent networks, etc.

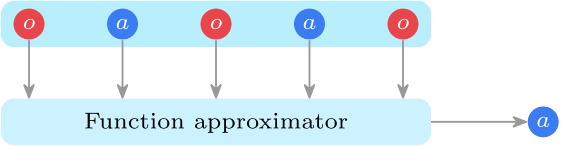


Fig. 1: Policy approximator.

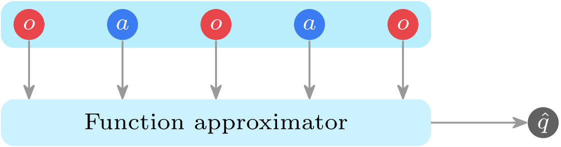


Fig. 2: Q-function approximator.

History-dependent reinforcement learning (ii)

We use **function approximators** for the **policy** or **Q-function** estimation.

- Feedforward network, transformer, recurrent networks, etc.

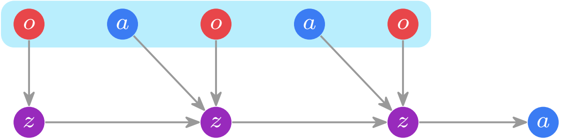


Fig. 1: Policy approximator.

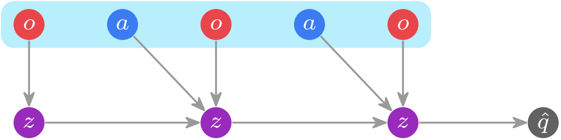


Fig. 2: Q-function approximator.

History-dependent reinforcement learning (iii)

We use **function approximators** for the **policy** or **Q-function** estimation.

- Feedforward network, transformer, recurrent networks, etc.

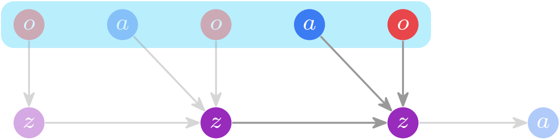


Fig. 1: Policy approximator.

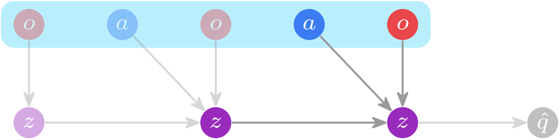


Fig. 2: Q-function approximator.

Agent-states

Learning from histories is infeasible, even with function approximators:

- **Feedforward networks** use a fixed window size (no extrapolation),
- **Transformers** use a fixed context size ($O(t)$ extrapolation),
- **Recurrent networks** use BPTT truncation ($O(1)$ extrapolation).

Will a history-dependent approximator generalize in extrapolation? Not sure.

$\Rightarrow$ **Agent state** that is fixed $z = f(h)$ and recurrent $f(h') = u(f(h), a, o')$.

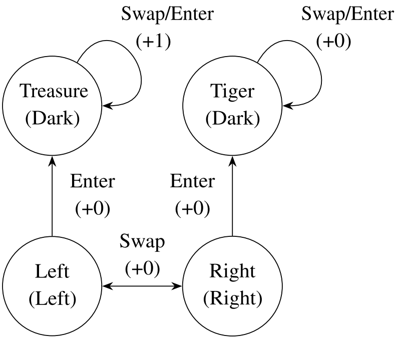
- Sliding window, last observation, Bayes/Kalman filter, etc.

Now, we thus focus on learning an **agent-state policy** $\pi(a|z)$, which form the **history-dependent policy** $\eta(a|h) = \pi(a|f(h))$.

NB: We can learn a Transformer or RNN on top of an agent state.

Aliasing

Let us consider the "Aliased Tiger" POMDP, with $z = o$. Let us look at $V(z)$[1] versus $\tilde{V}(z)$ the fixed point of the aliased Bellman equations.



- We have $V(z = \text{Left}) = \frac{\gamma}{1-\gamma}$ and $V(z = \text{Right}) = \frac{\gamma^2}{1-\gamma}$.
- But we have $\tilde{V}(z = \text{Left}) = \tilde{V}(z = \text{Right}) = \frac{\gamma}{2(1-\gamma)}$.

[1]**NB:** The aliased value functions $V(z)$ should be more carefully defined (timed).

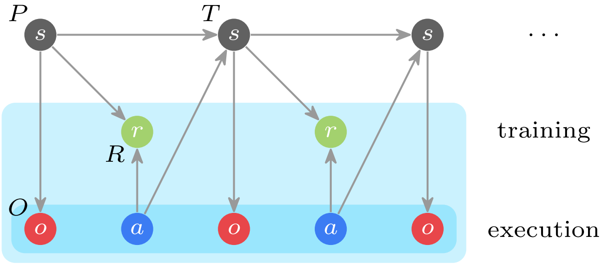
Asymmetric Observability

Asymmetric observability

Decision process	Execution	Training	Generality
MDP	s	s	**Too optimistic.**
POMDP	o	o	**Too pessimistic.**
Privileged POMDP	o	s	**Too optimistic.**
Informed POMDP	o	i	**Just right?**

Examples: simulator state, trajectory in hindsight, additional sensors, additional viewpoints, observations of other agents, etc.

Optimal control under partial observability

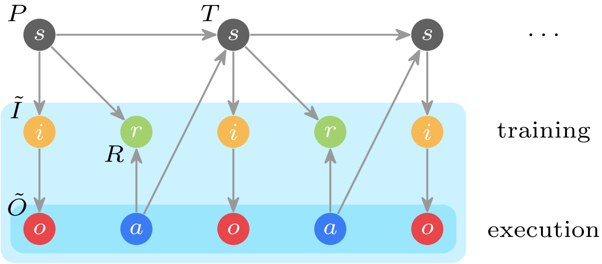


The problem of **RL in POMDP** is to find an optimal history-dependent policy,

$$\eta^* \in \underset{\eta \in \mathrm{H}}{\operatorname{argmax}} \underbrace{\mathbb{E}^{\eta}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]}_{J(\eta)},$$

from samples $(o_0, a_0, r_0, \ldots, o_n)$.

Optimal control under asymmetric observability



The problem of **RL in POMDP** is to find an optimal history-dependent policy,

$$\eta^* \in \underset{\eta \in \mathrm{H}}{\operatorname{argmax}} \underbrace{\mathbb{E}^{\eta}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]}_{J(\eta)},$$

from samples $(i_0, o_0, a_0, r_0, \ldots, i_n, o_n)$.

Asymmetric reinforcement learning

Asymmetric RL leverages i (usually s) to learn a policy $\pi(a|f(h))$ faster.

1. **Imitation learning** approaches:
   - Learn a fully observable policy $\pi(a|s)$, imitate the policy $\pi(a|z) \approx \pi(a|s)$.
2. **Asymmetric actor-critic** approaches:
   - Use additional state information as input to the critic $Q(s, z, a)$.
3. **Model-based** approaches:
   - Learn to predict the next state $q(r, s'|z, a)$.
4. **Representation learning** approaches:
   - Learn to predict the belief $q(s|z)$ as an auxiliary loss.

While initial methods were **heuristic**, a recent line of work has proposed **theoretically grounded** asymmetric learning objectives (Baisero & Amato, 2022; Lambrechts et al., 2024; Wang et al., 2023; Warrington et al., 2021).

Asymmetric actor-critic is successful

- Magnetic Control of Tokamak Plasma through Deep RL (Degrave et al., 2022).
- Champion-Level Drone Racing using Deep RL (Kaufmann et al., 2023).
- A Super-Human Vision-Based RL Agent in Gran Turismo (Vasco et al., 2024).

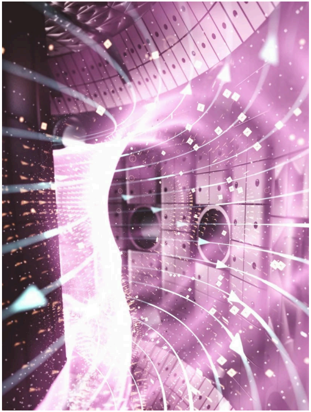


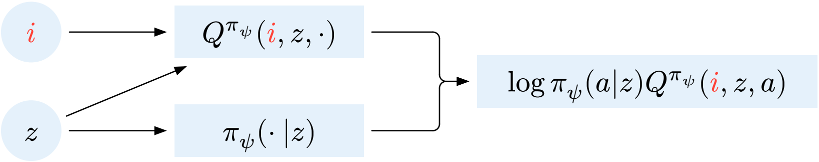
Image credits: first, second, third.

Theoretical Justifications

Asymmetric actor-critic algorithm

Actor-critic algorithms are policy-gradient methods with a critic $Q_\varphi^{\pi_\psi} \approx Q^{\pi_\psi}$.

- The critic is **only used** for estimating the policy-gradient.
- It can be **informed** with additional information: $Q^{\pi_\psi}(z, a) \to Q^{\pi_\psi}(i, z, a)$.



This policy gradient was proved valid for any $I(\cdot \mid s)$ for history-dependent policies (Ebi et al., 2025).

$\Rightarrow$ Effective, but **no theoretical justification for its benefits** until recently.

Aliasing and asymmetric observability

In addition to aliased z ($\tilde{V}(z) \neq V(z)$), we also have aliased s ($\tilde{V}(s) \neq V(s)$).

Indeed, while the state is sufficient for the environment execution, it is not for the agent execution: $a \sim \eta(\cdot\,|z)$ is not conditionally independent on z given s.

Instead, the **environment-agent state** (s, z) is sufficient for the execution of the environment and agent.

$\Rightarrow$ POMDP (with z) $\approx$ MDP whose state is (s, z): the "environment-agent MDP".

The agent simply ignores s. As in any MDP, $\tilde{V}(s, z) = V(s, z)$.

Agent-state asymmetric actor-critic algorithm

We provide a theoretical justification by comparing the finite-time bound for an asymmetric actor-critic algorithm (Lambrechts et al., 2025) and for its symmetric counterpart (Cayci et al., 2024).

- **State-informed**:
  - We study the case where $i = s$.
- **Fixed agent state:**
  - Fixed update $z' \sim U(\cdot\,|z, a, o')$, and policy $a \sim \pi(\cdot\,|z)$.
- **Finite state Q-functions:**
  - Asymmetric $\mathcal{Q}^\pi(s, z, a)$ and symmetric $Q^\pi(z, a)$.
- **Linear approximations:**
  - $\hat{\mathcal{Q}}^\pi_\beta(s, z, a) = \langle \beta, \varphi(s, z, a) \rangle$ and $\hat{Q}^\pi_\beta(z, a) = \langle \beta, \chi(z, a) \rangle$.
  - $\pi_\theta(a|z) \propto \exp(\langle \theta, \psi(z, a) \rangle)$.

Actor-critic algorithm

Algorithm 1: Asymmetric and symmetric actor-critic.

1. Initialize policy parameters ψ_0.
2. For $t = 1 \ldots T$
   1. Estimate $\hat{\mathcal{Q}}^{\pi}_{\varphi} \approx \mathcal{Q}^{\pi_\psi}$ or $\hat{Q}^{\pi}_{\chi} \approx Q^{\pi_\psi}$ (**TD learning**).
   2. Estimate $g_{t-1} \approx \nabla_\psi J\left(\pi_{\psi_{t-1}}\right)$ using $\mathcal{Q}_\varphi$ or Q_χ (**NPG estimation**).
   3. Update policy $\psi_t = \psi_{t-1} + \eta g_{t-1}$.
3. Return π_{ψ_T}

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

Aliasing measure.

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

Finite-time bound for the critics

Theorem 3: Finite-time bound for asymmetric and symmetric Q-functions.

For any $\pi \in \Pi_{\mathcal{M}}$, and any $m \in \mathbb{N}$, we have for TD learning with $\alpha = \frac{1}{K}$,

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

$$\sqrt{\mathbb{E}\left[\left\|Q^\pi - \overline{Q}^\pi\right\|^2_{d^\pi}\right]} \leq \varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}} + \varepsilon_{\text{alias}}.$$

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

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

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

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

Finite-time bound for the actors

Theorem 4: Finite-time bound for asymmetric and symmetric NAC.

For any $(\mathcal{Z}, U)$, we have for NAC with $\alpha = \frac{1}{K}$, $\zeta = \frac{B\sqrt{1-\gamma}}{\sqrt{2N}}$, $\eta = \frac{1}{\sqrt{T}}$,

$$(1-\gamma) \min_{0 \leq t < T} \mathbb{E}[J(\pi^*) - J(\pi_t)] \leq \varepsilon_{\text{nac}} + \varepsilon_{\text{actor}} + \varepsilon_{\text{inf}} + \varepsilon_{\text{grad}} + \frac{1}{T} \sum_{t=0}^{T-1} \varepsilon_{\text{critic}}^{\pi_t},$$

$$\varepsilon_{\text{nac}} = \frac{B^2 + 2\log|A|}{2\sqrt{T}}$$

$$\varepsilon_{\text{actor}} = \overline{C}_\infty \sqrt{\frac{(2-\gamma)B}{(1-\gamma)\sqrt{N}}}$$

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

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

$$\varepsilon_{\text{critic,asym}}^{\pi_t} = 2\overline{C}_\infty \sqrt{6}\left(\varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}}\right) \qquad \varepsilon_{\text{critic,sym}}^{\pi_t} = 2\overline{C}_\infty \sqrt{6}\left(\varepsilon_{\text{td}} + \varepsilon_{\text{app}} + \varepsilon_{\text{shift}} + \varepsilon_{\text{alias}}\right)$$

Some insights

When using an asymmetric actor-critic algorithm:

- The **critic error** has a **smaller upper bound**.
  - Because the asymmetric critic is the solution of a Bellman equation.
- The **actor suboptimality** has a **smaller upper bound**.
  - This benefit mainly comes from the smaller upper bound on the critic error.

Some limitations:

- The analysis assumes a **fixed agent state process**.
  - Shed light on the effect of an aliased agent state (e.g., RNN at initialization).
  - It can easily be extended to learnable agent state processes: $\mathcal{A}^+ = \mathcal{A} \times \mathcal{Z}$.
- Requires **samples from the discounted visitation measure**.
  - But it is still feasible without assumption on the mixing time.
  - Reveal an interesting term $\varepsilon_{\text{shift}}$ when not assuming stationary distribution.

Future Directions

Other algorithms with an asymmetric critic

Asymmetric and symmetric actor-critic algorithms were both valid,[2]

$$\begin{aligned}\nabla J(\pi) &= \mathbb{E}^{d^\pi}[\log \pi(a|z) Q^\pi(i, z, a)] \\ &= \mathbb{E}^{d^\pi}[\log \pi(a|z) Q^\pi(z, a)]\end{aligned}$$

because $\nabla J(\pi)$ is linear in Q^π and $\mathbb{E}[Q^\pi(I, z, a)] = Q^\pi(z, a)$.

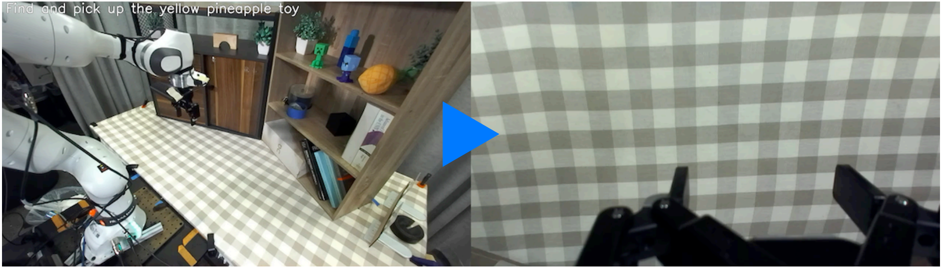
Other RL objectives, such as the advantage weighted regression (AWR), do not present such properties (Hu et al., 2025).

$$\begin{aligned}\mathcal{L}_{\text{AWR}}(\pi) &= J(\pi) - J(\mu) - \beta \ \text{KL}(\mu \parallel \pi) \\ &= \mathbb{E}^{d^\mu}[\log \pi(a|z) \exp(A^\pi(s, z, a)/\beta)] \\ &\neq \mathbb{E}^{d^\mu}[\log \pi(a|z) \exp(A^\pi(z, a)/\beta)].\end{aligned}$$

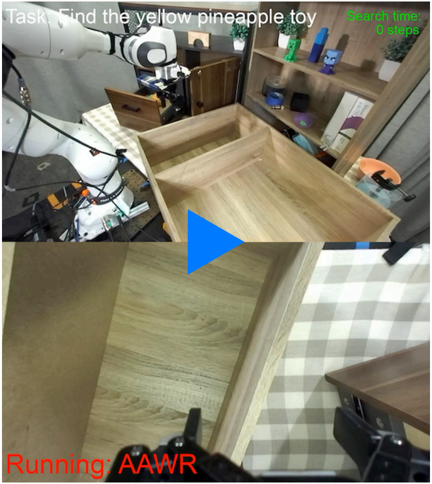
This further motivate the usage of asymmetric critics for nonlinear objectives.

[2] **NB:** The aliased value functions $Q^\pi(z, a)$ should be more carefully defined (timed).

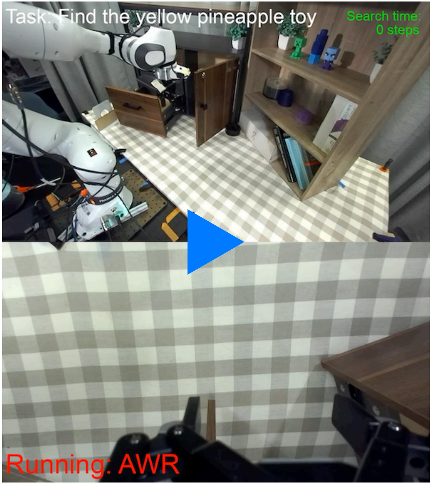
Foundations policies on partially observable tasks

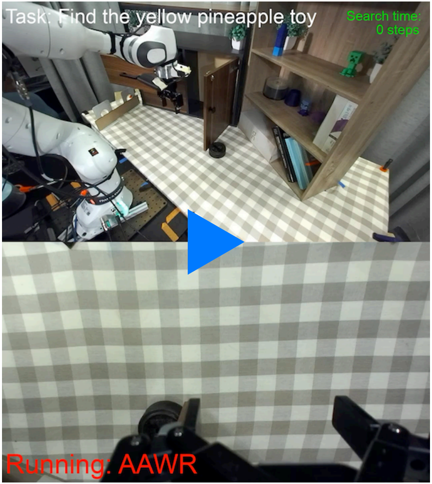


Example results

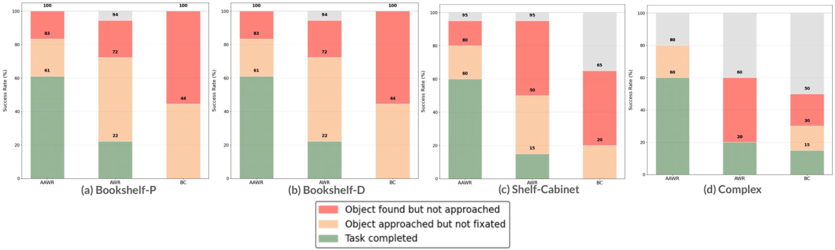


Example results (ii)





Asymmetric advantage weighted regression



Conclusion

Take-home message

Don't make the problem harder than it is.

Consider all available information at training.

Further readings:

- Lambrechts, G., Bolland, A., & Ernst, D. (2024). Informed POMDP: Leveraging Additional Information in Model-Based RL. *Reinforcement Learning Journal.*
- Lambrechts, G., Ernst, D., & Mahajan, A. (2025). A Theoretical Justification for Asymmetric Actor-Critic Algorithms. *International Conference on Machine Learning*.
- Others coming soon (Ebi et al., 2025; Hu et al., 2025).

Follow me on Bluesky @gsprd.be for updates.

References

Baisero, A., & Amato, C. (2022). Unbiased Asymmetric Reinforcement Learning under Partial Observability. *International Conference on Autonomous Agents and Multiagent Systems*.

Cayci, S., He, N., & Srikant, R. (2024). Finite-Time Analysis of Natural Actor-Critic for POMDPs. *SIAM Journal on Mathematics of Data Science*.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B. D., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., Las Casas, D. de, Donner, C., Fritz, L., Galperti, C., Huber, A., Keeling, J., Tsimpoukelli, M., Kay, J., Merle, A., Moret, J.-M., … Riedmiller, M. A. (2022). Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning. *Nature*.

Ebi, D., Lambrechts, G., Ernst, D., & Böhm, K. (2025). [title kept confidential during double-blind reviews]. *Under Review*.

References (ii)

Hu, E. S., Wang, J., Yuan, X., Luo, F., Li, M., Lambrechts, G., Rybkin, O., & Jayaraman, D. (2025). [title kept confidential during double-blind reviews]. *Under Review*.

Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Champion-Level Drone Racing using Deep Reinforcement Learning. *Nature*.

Lambrechts, G., Bolland, A., & Ernst, D. (2024). Informed POMDP: Leveraging Additional Information in Model-Based RL. *Reinforcement Learning Journal*.

Lambrechts, G., Ernst, D., & Mahajan, A. (2025). A Theoretical Justification for Asymmetric Actor-Critic Algorithms. *International Conference on Machine Learning*.

McCarthy, J. (1998). *What is Artificial Intelligence*.

References (iii)

Vasco, M., Seno, T., Kawamoto, K., Subramanian, K., Wurman, P. R., & Stone, P. (2024). A Super-Human Vision-Based Reinforcement Learning Agent for Autonomous Racing in Gran Turismo. *Reinforcement Learning Journal*.

Wang, A., Li, A. C., Klassen, T. Q., Icarte, R. T., & McIlraith, S. A. (2023). Learning Belief Representations for Partially Observable Deep RL. *International Conference on Machine Learning*.

Warrington, A., Lavington, J. W., Scibior, A., Schmidt, M., & Wood, F. (2021). Robust Asymmetric Learning in POMDPs. *International Conference on Machine Learning*.