Role of State in Partially Observable RL

Andrea Baisero {baisero.a@northeastern.edu}

Khoury College of Computer Sciences, Northeastern University, Boston, USA

Problem Statement

- Many control problems are **partially observable**:
  Agent does not observe *state* s, must rely on *observable history* h.
- Privileged training frameworks use state **during training** to improve agent performance during evaluation.
- Empirically successful [1, 2, 3], but still **poorly understood**:
  Belief-MDPs (and history-MDPs) dictate *state should not matter*!

Research Question

Why does state help privileged training algorithms?

Background

Partially Observable Control

- Partially observable tasks require information gathering, memory.
- Agent relies on good representation of history $\phi(h)$, **hard** to learn.
- Good $\phi(h)$ extracts key events and filters the rest, but ...
  ... identifying key events is like finding a needle in a haystack ...
  ... while learning to recognize needles and haystacks ...
  ... without supervision.

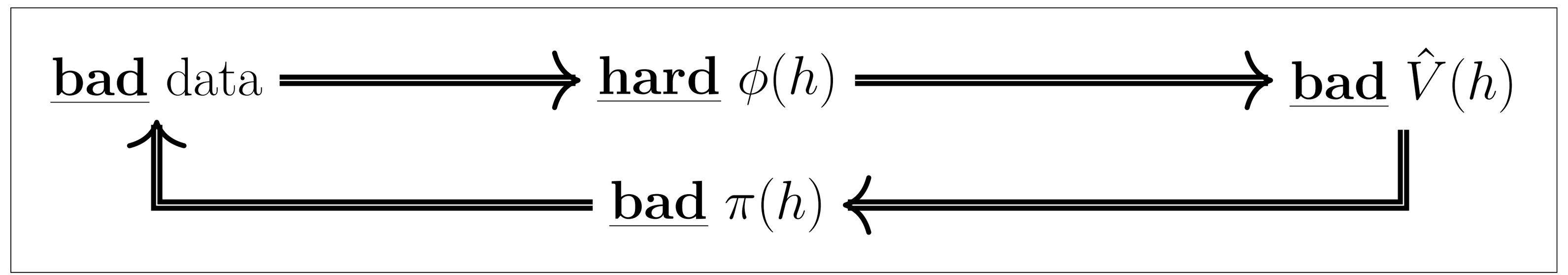


Figure: A vicious Actor-Critic cycle.

Privileged Training Frameworks

- Based on *history-state values* $V^\pi(h, s)$ and $Q^\pi(h, s, a)$, e.g.,

$$V^\pi(h, s) = \mathbb{E}_{a\sim\pi(h)}\Big[R(s, a) + \gamma\, \mathbb{E}_{s',o|s,a}\big[V^\pi(hao, s')\big]\Big] . \quad (1)$$

- (Unbiased) **Asymmetric A2C** [1]

$$\nabla J \approx \mathbb{E}\Big[\sum_t \gamma^t \hat{Q}(h_t, s_t, a_t)\nabla \log \pi(h_t, a_t)\Big] , \quad (2)$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\Big(r + \gamma\hat{V}(h_{t+1}, s_{t+1}) - \hat{V}(h_t, s_t)\Big)^2 . \quad (3)$$

- **Asymmetric DQN** [2]

$$\mathcal{L}_{\hat{U}} = \frac{1}{2}\Big(r + \gamma\hat{U}(hao, s', \operatorname*{argmax}_{a'} \hat{Q}(hao, a')) - \hat{U}(h, s, a)\Big)^2 , \quad (4)$$

$$\mathcal{L}_{\hat{Q}} = \frac{1}{2}\Big(r + \gamma\hat{U}(hao, s', \operatorname*{argmax}_{a'} \hat{Q}(hao, a')) - \hat{Q}(h, a)\Big)^2 . \quad (5)$$

Role of State Hypotheses

State as Information

- State provides information that is **extrinsic** to the history.
- Strongest when $\mathbb{H}[S \mid H = h] \gg 0$.

State as a Feature

- State provides information that is **intrinsic** to the history.
- Strongest when $\mathbb{H}[S \mid H = h] \approx 0$.

State as Exploration

- State injects **context-dependent** variance $\mathbb{V}_{s|h}[V^\pi(h, s)]$.

State as Bootstrapping

- State representation $\phi(s)$ is **easier** to learn than $\phi(h)$ ...
  ... which helps learn a better critic $\hat{V}(h, s)$ ...
  ... bootstrap a better $\phi(h)$ ...
  ... leading to a better policy $\pi(h)$.

Evaluation Methodology

Latent Observations (observations available during training)
Estimate policy gradient using *latent observations* (designed or learned):

- Latent space $\mathcal{Z}$, function $Z\colon \mathcal{S} \to \Delta\mathcal{Z}$, values $V^\pi(h, z), Q^\pi(h, z, a)$,

$$\nabla J \approx \mathbb{E}\Big[\sum_t \gamma^t \hat{Q}(h_t, z_t, a_t)\nabla \log \pi(h_t, a_t)\Big] . \quad (6)$$

Counterfactual History-State Values
Estimate policy gradient using *counterfactual* states $V^\pi(h, \tilde{s})$:

$$\mathbb{E}_{\tilde{s}|h}[V^\pi(h, \tilde{s})] = \mathbb{E}_{s|h}[V^\pi(h, s)] = V^\pi(h) , \quad (7)$$

$$\mathbb{V}_{\tilde{s}|h}[V^\pi(h, \tilde{s})] = \mathbb{V}_{s|h}[V^\pi(h, s)] . \quad (8)$$

Noisy History Values
Estimate policy gradient using *noisy* values $V^\pi(h, \omega) = V^\pi(h) + \omega$:

- Inject noise $\omega \sim \mathrm{Normal}\big(0, \sigma^2(h)\big)$ where $\sigma^2(h) = \mathbb{V}_{s|h}\big[V^\pi(h, s)\big]$,

$$\mathbb{V}_{\omega|h}[V^\pi(h, \omega)] = \mathbb{V}_{s|h}[V^\pi(h, s)] . \quad (9)$$

Feature Importance Analysis
Compare relative importance of history/state features during training:

- Permutation feature importance [4].
- SHapley Additive exPlanations (SHAP) [5].

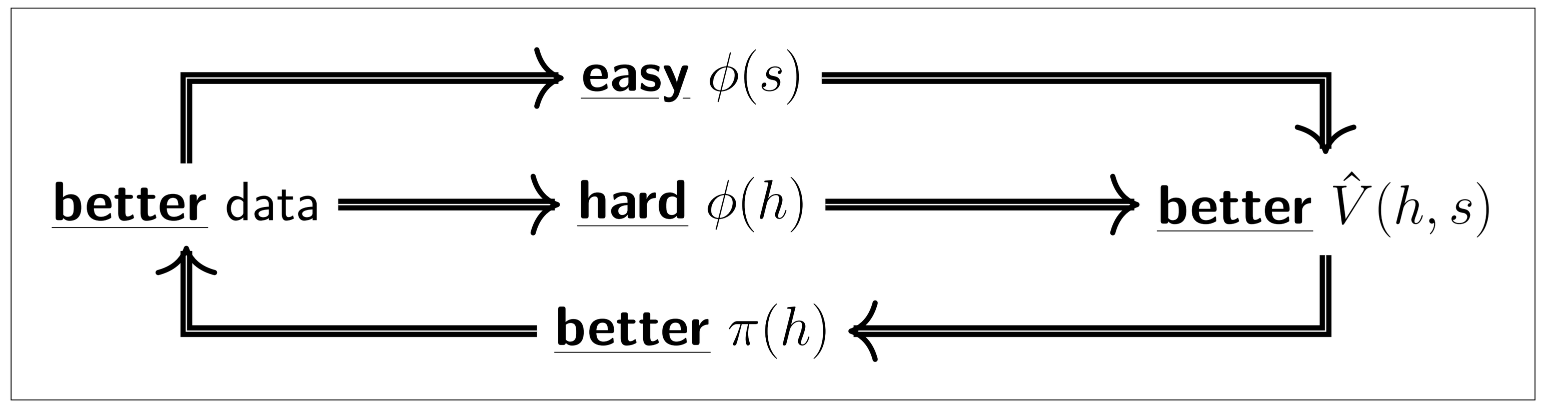
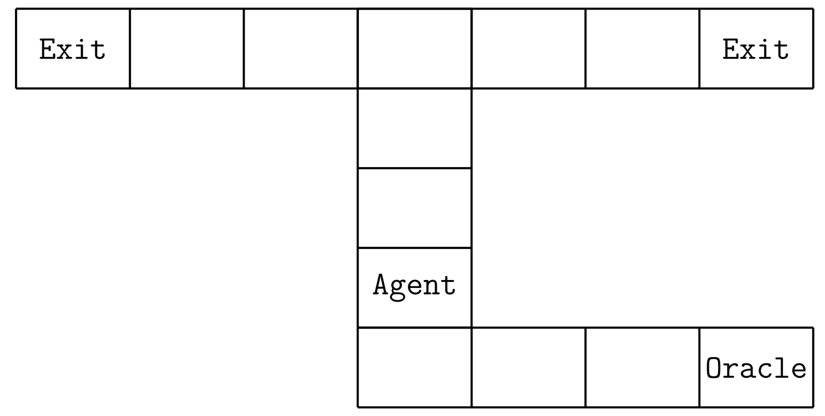
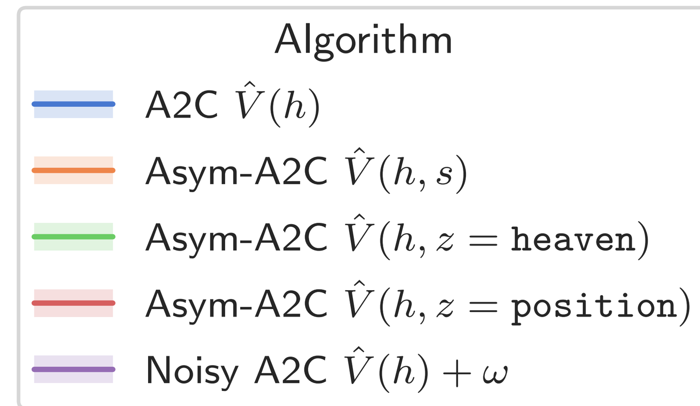


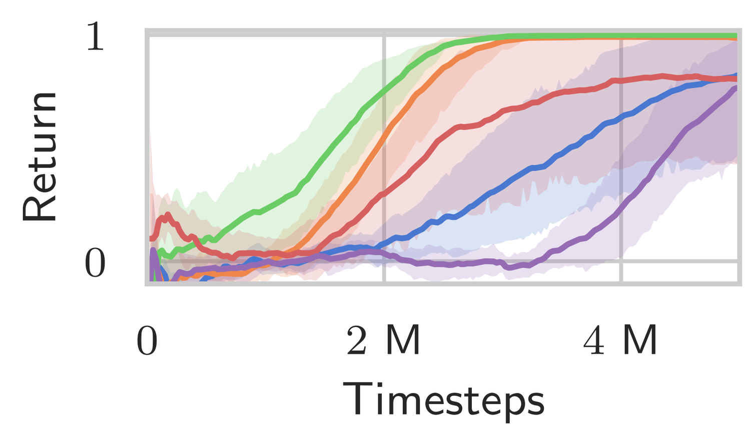
Figure: A better Asymmetric Actor-Critic cycle.

Preliminary Results

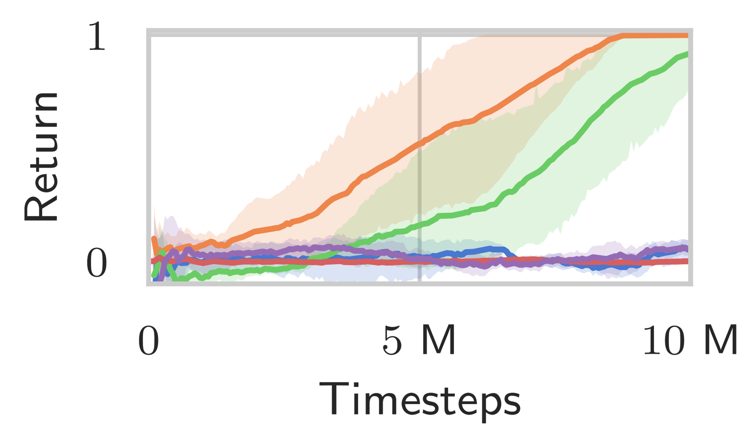


(a) HeavenHell





(b) HeavenHell-3



(c) HeavenHell-4

Figure: Mean returns over 5 seeds, bootstrapped 95% CI.

References

[1] A. Baisero and C. Amato, "Unbiased Asymmetric Reinforcement Learning under Partial Observability," in *Proceedings of the Conference on Autonomous Agents and Multiagent Systems*, 2022.

[2] A. Baisero, B. Daley, and C. Amato, "Asymmetric DQN for Partially Observable Reinforcement Learning," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2022.

[3] E. Marchesini, A. Baisero, R. Bhati, and C. Amato, "On Stateful Value Factorization in Multi-Agent Reinforcement Learning," in *Proceedings of the Conference on Autonomous Agents and Multiagent Systems*, 2025.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30.

[5] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," vol. 20, no. 177, pp. 1–81.