Unbiased Asymmetric Reinforcement Learning under Partial Observability

AAMAS 2022, Auckland, New Zealand

Andrea Baisero **Christopher Amato**
`{baisero.a,c.amato}@northeastern.edu`
Northeastern University, Boston, USA

Overview

Setting

- Single-agent
- Model-free
- Partially observable (significant amounts)
- Reinforcement learning
- Offline training / online execution

Background

Offline Training / Online Execution (OTOE)

- Safety during training
- Faster training, e.g., via parallelization
- Access to privileged information

Privileged Information

- Multi-agent RL: Joint history $\bar{h}$
- Single-agent RL: Latent state s
- How to exploit it?
  - \+ Great potential
  - – Lack of theoretical justification
  - – Misuse $\implies$ grave issues

Northeastern University

Background

(Symmetric) Advantage Actor-Critic (A2C)

- Actor and critic models $\pi(h)$ and $\hat{V}(h)$, trained using

$$\nabla J \propto \mathbb{E}\left[\sum_t \gamma^t \delta_t \nabla \log \pi(a_t; h_t)\right] \tag{1}$$

$$\delta_t = r_t + \gamma \hat{V}(h_{t+1}) - \hat{V}(h_t) \tag{2}$$

Asymmetric Advantage Actor-Critic (Asym-A2C)

- Actor and critic models $\pi(h)$ and $\hat{V}(s)$, trained using

$$\delta_t = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t) \tag{3}$$

- True state $\implies$ more informative critic
- More informative critic $\implies$ improved policy gradient

Contributions

In Our Paper

- Theory of asymmetric A2C and $V^\pi(s)$
  - Expose conceptual and formal issues
  - $V^\pi(s)$ ill-defined and/or biased
- Unbiased Asymmetric A2C
  - Uses history-state values $V^\pi(h, s)$
  - $V^\pi(h, s)$ well-defined and unbiased!
- Interpretation of state as stochastic features of history
- Empirical evaluation on partially observable environments
  - Requires information gathering + memorization

Theory of State-Based Value Functions

Formal Methodology

- Policy gradient $\nabla J \propto \mathbb{E}\left[\sum_t \gamma^t Q^\pi(h_t, a_t) \nabla \log \pi(a_t; h_t)\right]$
- $Q^\pi(h, a)$ is the correct theoretical quantity
- V^π instead of Q^π (same implications)
- $V^\pi(s)$ as estimator of $V^\pi(h)$
  $\implies V^\pi(s)$ unbiased iff $V^\pi(h) = \mathbb{E}_{s|h}\left[V^\pi(s)\right]$

Theory of State-Based Value Functions

An Informal Argument Against State Values

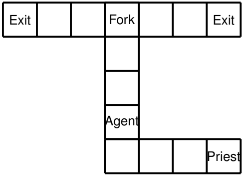


Figure: HeavenHell-3. The optimal agent will visit the priest to learn which exit leads to heaven, and which to hell.

History Aliasing

- s not a sufficient statistic of h
  $\implies$ s unable to determine agent behavior
  $\implies$ $V^\pi(s)$ unable to represent expected rewards
- Ideally, $V^\pi(s = \text{Fork})$ high if priest visited
  low if priest not visited
- Actually, $V^\pi(s = \text{Fork})$ unable to differentiate histories

Theory of State-Based Value Functions

Cases

- General policy under partial observability
  $\implies V_t^\pi(s)$ well-defined
  $\implies V^\pi(s)$ ill-defined (issue w/ time-invariant history RV)
- Reactive policy under partial observability
  $\implies V^\pi(s)$ well-defined but biased
- Reactive policy under virtually "full" observability
  $\implies V^\pi(s)$ well-defined and virtually unbiased

Takeaway

- $V^\pi(s)$ not suitable for partial observability

Northeastern University

Theory of State-Based Value Functions

Unbiased Asymmetric A2C

History-State Value Function $V^\pi(h, s)$

$$V^\pi(h, s) = \sum_a \pi(a; h) \left(R(s, a) + \gamma \, \mathbb{E}_{s', o \mid s, a} \left[V^\pi(hao, s')\right]\right)$$

- $V^\pi(h, s)$ as estimator of $V^\pi(h)$
  - Well-defined
  - Unbiased, $V^\pi(h) = \mathbb{E}_{s \mid h} \left[V^\pi(h, s)\right]$
  - Low state uncertainty $\implies$ low variance

Asymmetric Policy Gradient Theorem

$$\nabla J \propto \mathbb{E}\left[\sum_t \gamma^t Q^\pi(h_t, s_t, a_t) \nabla \log \pi(a_t, h_t)\right]$$

Northeastern University

Evaluation

Environments

- 8 environments with significant partial observability
  - Information gathering strategies
  - Mid-long term memorization

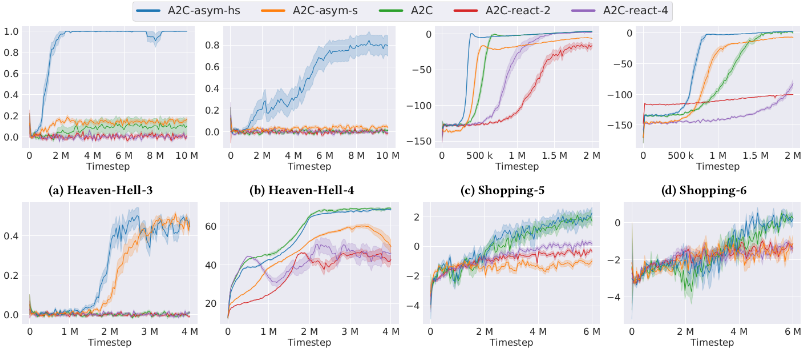
Algorithms

- A2C-react-$\{2,4\}$: history critic $\hat{V}(h)$ (short-term memory)
  - Short-term memory
  - Included to show partial observability
- A2C: history critic $\hat{V}(h)$
- A2C-asym-s: state critic $\hat{V}(s)$
- A2C-asym-hs: history-state critic $\hat{V}(h, s)$

Northeastern University

Evaluation

Results



(a) Heaven-Hell-3 (b) Heaven-Hell-4 (c) Shopping-5 (d) Shopping-6

(e) Car-Flag (f) Cleaner (g) Memory-Four-Rooms-7x7 (h) Memory-Four-Rooms-9x9

Northeastern University

Conclusions

Contributions

- Theory of asymmetric A2C and $V^\pi(s)$
    - Expose conceptual and formal issues
    - $V^\pi(s)$ ill-defined and/or biased
- Unbiased Asymmetric A2C
    - Uses history-state values $V^\pi(h, s)$
    - $V^\pi(h, s)$ well-defined and unbiased!
- Interpretation of state as stochastic features of history
- Empirical evaluation on partially observable environments
    - Requires information gathering + memorization

Northeastern University