Partially Observable Control & Stateful Partially Observable RL (PORL)

CS 4180/5180 Guest Lecture

Andrea Baisero

baisero.a@northeastern.edu

Northeastern University, Boston, USA

Northeastern University

Overview

Partial Observability:

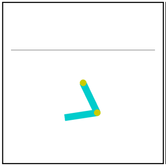
- Single-agent RL perspective
- Mainly theory and difficulties of PO control
- Practical methods and extensions

Asymmetric RL for Partial Observability (aka Stateful PORL):

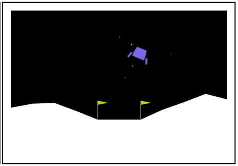
- **Note:** Not related to equivariance-type symmetry as presented by Dian
- Main subject of my research & thesis dissertation
- Good results, ongoing work, many open questions

Examples of Full and Partial Observability

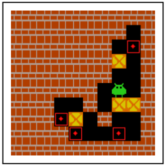
Informal definition: Some state information is hidden.



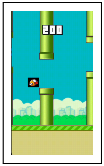
(a) Acrobot.



(b) Lunar Lander.

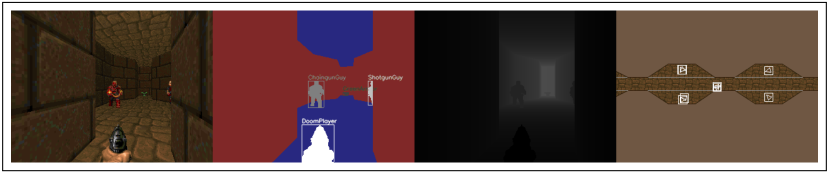


(c) Sokoban.



(d) Flappybird.

(e) Breakout.



(f) ViZDoom.

Implications of Partial Observability

What is partial observability?

- Environment state is in some way hidden from agent
  e.g., behind the agent, behind the corner, contents of a box, etc.
- Agent sees indirect observations, not full state
  - Sometimes filtered state
  - Sometimes separate observation space altogether
- Non-Markovian $\Pr(o_t \mid o_{t-1}, \ldots, o_1) \neq \Pr(o_t \mid o_{t-1})$

Is that such a big deal?

- No, just use the same methods with available observation

Reactive policies $\pi \colon \mathcal{O} \to \Delta\mathcal{A}$

Northeastern University

Implications of Partial Observability

What is partial observability?

- Environment state is in some way hidden from agent
  e.g., behind the agent, behind the corner, contents of a box, etc.
- Agent sees indirect observations, not full state
  - Sometimes filtered state
  - Sometimes separate observation space altogether
- Non-Markovian $\Pr(o_t \mid o_{t-1}, \ldots, o_1) \neq \Pr(o_t \mid o_{t-1})$

Is that such a big deal?

- ~~No, just use the same methods with available observation~~
- YES! Significant theoretical & practical consequences!

Reactive policies $\pi \colon \mathcal{O} \to \Delta\mathcal{A}$ are far from ideal!

A Positive Example: Sutton's Gridworld

Observation: 3x3 region around agent

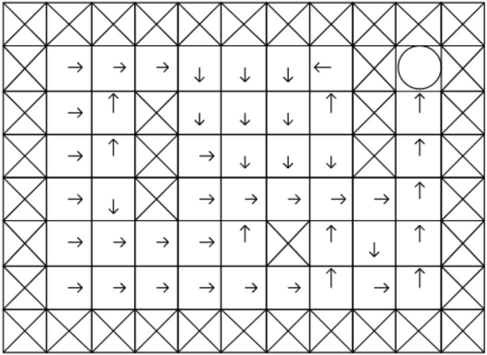


Figure: Reactive solution to Sutton's gridworld (Littman, 1994).

Keypoints: State aliasing
Suboptimality, reactive control as constrained control
Still, we can guarantee goal

Northeastern University

A Negative Example: McCallum's Maze

Observation: 3x3 region around agent

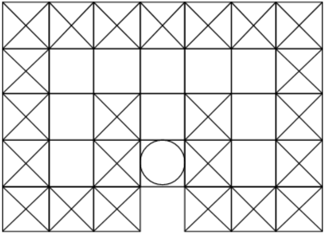


Figure: McCallum's Maze (Littman, 1994).

Question: Can we guarantee goal?

A Negative Example: McCallum's Maze

Observation: 3x3 region around agent

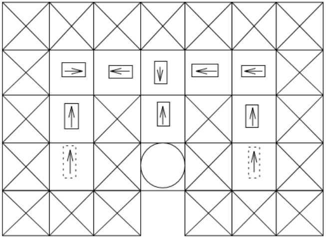


Figure: McCallum's Maze (Littman, 1994).

Question: Can we guarantee goal?

Keypoints: Deterministic π fails from one direction (constrained policy)
Stochastic π succeeds from both directions (still suboptimal)
Optimal deterministic policy not guaranteed?

A Negative Example: McCallum's Maze

Observation: 3x3 region around agent

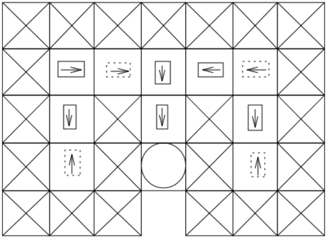


Figure: McCallum's Maze (Littman, 1994).

Question: Can we guarantee goal?

Keypoints: Deterministic π fails from one direction (constrained policy)
Stochastic π succeeds from both directions (still suboptimal)
Optimal deterministic policy not guaranteed?

Resolution: Deterministic π w/ memory, guarantee goal from both directions!

Information Gathering & Memorization

Goal: Reach good exit, avoid bad exit

Observability:

- Full observability of position
- Partial observability of exits

FO Optimality: Just exit

PO Optimality: Backtrack, memorize, exit
Easy to execute
Deviously hard to discover

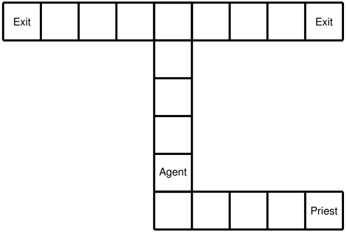


Figure: Heaven/Hell Problem.

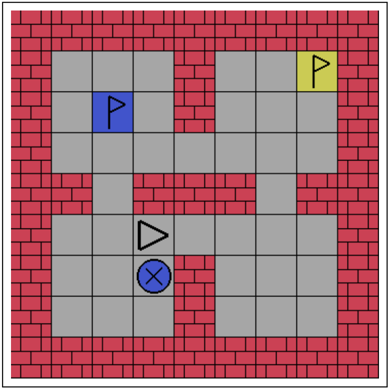
Information Gathering:

- Broader than state estimation, about reaching desireable belief-states
- Problem-dependent, e.g., info gathering actions/states/experiences, etc.
- Commonly harmful upfront cost, hard to overcome

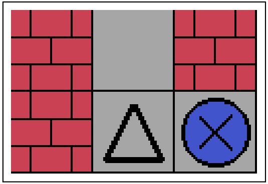
Memorization: History contains all the info, memorize what?

- Memory as information extraction
- Needle in haystack problem, key to extract and integrate right info
- Remember too little, remember too much, remember wrong things

Information Gathering & Memorization



(a) State.



(b) Observation.

Figure: **Memory-Four-Rooms-9x9**, a procedurally generated navigation task which requires information-gathering and memorization. The agent must avoid the *bad* exit and reach the *good* exit, which is identifiable by the color of the *beacon*.

Full vs Partial Observability

Fully Observable Control

- Access to full system state
  $\pi : \mathcal{S} \to \Delta\mathcal{A}$
- Solid theory, comparatively easy!
- Strong assumption, not always possible!

Partially Observable Control

- Access to partial/indirect observations derived from state
  $\pi : ? \to \Delta\mathcal{A}$
- Better match to reality, common in real world problems
- Extremely wide range of difficulties (technical vs significant PO)
  - At best, about as difficult as full observability
  - At worst, orders of magnitude more difficult, virtually impossible

Northeastern University

Markov Decision Processes (MDPs) Refresher

Definition: MDP tuple $M = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Transition function $T \colon \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$
- Reward function $R \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Discount factor $\gamma \in [0, 1)$

Policy: $\pi \colon \mathcal{S} \to \Delta\mathcal{A}$

Goal: $\max_\pi J \doteq \mathbb{E}\left[\sum_t \gamma^t R(s_t, a_t)\right]$

State Values via Bellman equations

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(s)}\left[R(s,a) + \gamma\,\mathbb{E}_{s'|s,a}\left[V^\pi(s')\right]\right] \tag{1}$$
$$Q^\pi(s,a) = R(s,a) + \gamma\,\mathbb{E}_{s'|s,a}\left[\mathbb{E}_{a'\sim\pi(s')}\left[Q^\pi(s',a')\right]\right] \tag{2}$$

Partially Observable Markov Decision Processes (POMDPs)

Definition: POMDP tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, p, T, O, R, \gamma \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Observation space $\mathcal{O}$
- Starting state distribution $p \in \Delta\mathcal{S}$
- Transition function $T\colon \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$
- Observation function $O\colon \mathcal{A} \times \mathcal{S} \to \Delta\mathcal{O}$
  $O\colon \mathcal{S} \to \Delta\mathcal{O}$ (optional starting observation)
- Reward function $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Discount factor $\gamma \in [0, 1)$

Policy: $\pi\colon ? \to \Delta\mathcal{A}$
(spoiler: histories or beliefs)

Goal: $\max_\pi J \doteq \mathbb{E}\left[\sum_t \gamma^t R(s_t, a_t)\right]$

Northeastern University

(PO)MDP Graphical Models

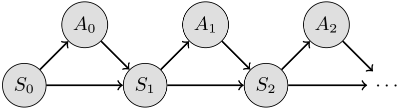


Figure: MDP graphical model.

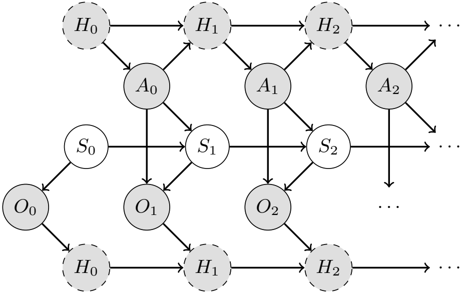


Figure: POMDP graphical model.

Histories

Notation:

- History Space $\mathcal{H} \doteq \mathcal{O} \times (\mathcal{A} \times \mathcal{O})^*$
  e.g., $h = (o_0, a_0, o_1, \dots a_{l-1}, o_{l-1})$ history with length $|h| = l$
- Concatenation notation $hao = (*h, a, o)$

History-based Control:

- History policy $\pi \colon \mathcal{H} \to \Delta\mathcal{A}$
- History values via Bellman equations

$$V^\pi(h) = \mathbb{E}_{a \sim \pi(h)} \left[R(h, a) + \mathbb{E}_{o \mid h,a} \left[V^\pi(hao) \right] \right] \tag{3}$$

$$Q^\pi(h, a) = R(h, a) + \mathbb{E}_{o \mid h,a} \left[\mathbb{E}_{a' \sim \pi(hao)} \left[Q^\pi(hao, a') \right] \right] \tag{4}$$

where

$$R(h, a) \doteq \mathbb{E}_{s \mid h} \left[R(s, a) \right] \tag{5}$$

$$\Pr(o \mid h, a) = \sum_{s, s'} \Pr(s \mid h) T(s, a, s') O(a, s', o) \tag{6}$$

Northeastern University

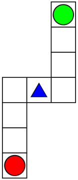
Histories - Why Actions?

Question: Agent chooses actions, they are not observations of state
Do we actually need them?

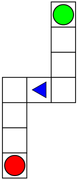
Histories - Why Actions?

Question: Agent chooses actions, they are not observations of state
Do we actually need them?

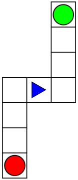
Answer: Yes, actions may influence unobserved part of state



(a) Starting state.



(b) Left turn.



(c) Right turn.

Figure: Spiral problem where action-memory is necessary.

In Practice:

- Highly dependent on problem and action semantics
  - Generic case: Strictly necessary
  - Special case: Plausibly unnecesaary and ignorable

Agent POV Example

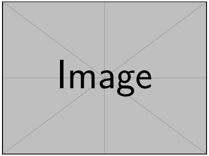


Figure: ????? problem.

Table: Agent POV.

Time	History	Action	Reward	Observation
0	( )	A3	-100	O1
1	(A3, O1)	A1	-1	O2
2	(A3, O1, A1, O2)	A1	-1	O2
3	(A3, O1, A1, O2, A1, O2)	A2	10	O1
4	(A3, O1, A1, O2, A1, O2, A2, O1)	A2	-100	O1

Keypoints:

- Unlikely seeing same histories enough times to learn (despite tiny size)
- Hard to interpret long sequences despite semantics

Agent POV Example

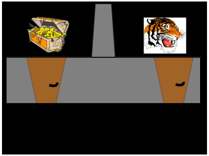


Figure: Tiger problem[1].

Table: Tiger Problem POV.

Time	History	Action	Reward	Observation
0	( )	OpenRight	-100	HearLeft
1	(OR, HL)	Listen	-1	HearRight
2	(OR, HL, L, HR)	Listen	-1	HearRight
3	(OR, HL, L, HR, L, HR)	OpenLeft	10	HearLeft
4	(OR, HL, L, HR, L, HR, OL, HL)	OpenLeft	-100	HearLeft

Keypoints:

- Unlikely seeing same histories enough times to learn (despite tiny size)
- Hard to interpret long sequences despite semantics

[1] (Possibly recursive) image credit to Chris Amato's PORL slides.

Beliefs

Belief: Distribution over states given observed history

- Belief Space $\mathcal{B} \doteq \Delta\mathcal{S}$
- Concatenation notation $bao = (*b, a, o)$
- POMDP initial belief $p \in \mathcal{B}$, prior to any observation
- **Notation overload:** $b\colon \mathcal{H} \to \Delta\mathcal{S}$, likelihood of s given observed history h
  $$b(h, s) \doteq \Pr(s \mid h)$$

Belief-based Control:

- Belief policy $\pi\colon \mathcal{B} \to \Delta\mathcal{A}$
- Belief values via Bellman equations

$$V^\pi(b) = \mathbb{E}_{a\sim\pi(b)}\left[R(b,a) + \gamma\,\mathbb{E}_{o\mid b,a}\left[V^\pi(bao)\right]\right] \tag{7}$$

$$Q^\pi(b,a) = R(b,a) + \gamma\,\mathbb{E}_{o\mid b,a}\left[\mathbb{E}_{a'\sim\pi(bao)}\left[Q^\pi(bao,a')\right]\right] \tag{8}$$

where

$$R(b,a) \doteq \mathbb{E}_{s\sim b}\left[R(s,a)\right] \tag{9}$$

$$\Pr(o \mid b,a) = \sum_{s,s'} b(s)T(s,a,s')O(s',a,o) \tag{10}$$

Belief Update

Rules: Just Bayes' rule

$$po(s) \propto p(s)O(s,o) \tag{11}$$

$$bao(s') \propto \left(\sum_s b(s)T(s,a,s')\right) O(a,s',o) \tag{12}$$

In practice:

- Discrete state space $\mathcal{S} \equiv \{s_i\}_{i=1}^n$
  $\implies$ Tensor operations, exact recursive update
- Continuous state space, linear dynamic system, e.g.

$$s_{t+1} = F_t s_t + B_t a_t + w_t \tag{13}$$

$$o_t = H_t s_t + v_t \tag{14}$$

  $\implies$ Kalman filter, exact recursive update on $\mu = \mathbb{E}\left[s \mid h\right], \Sigma = \mathbb{C}\left[s \mid h\right]$
- General system
  $\implies$ Particle filter $\{s_k\}_{k=1}^K$, approximate sample-based update
  Particles $\{s_k\}_{k=1}^K$ + rejection sampling $\rightarrow \{s'_k\}_{k=1}^K$
  Particles $\{s_k, w_k\}_{k=1}^K$ + importance sampling $\rightarrow \{s'_k, w'_k\}_{k=1}^K$

Belief Update Example: Gridworld

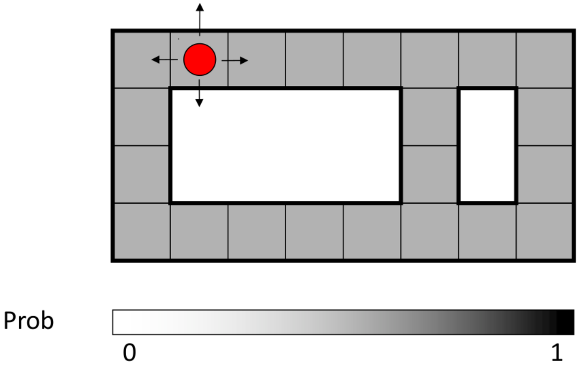


Figure: Belief estimation over time[2].

[2](Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Gridworld

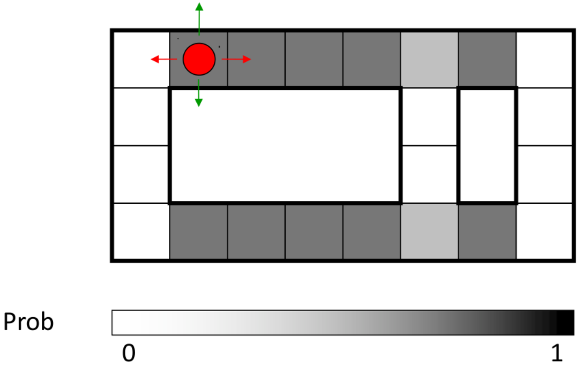


Figure: Belief estimation over time[2].

[2](Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Gridworld

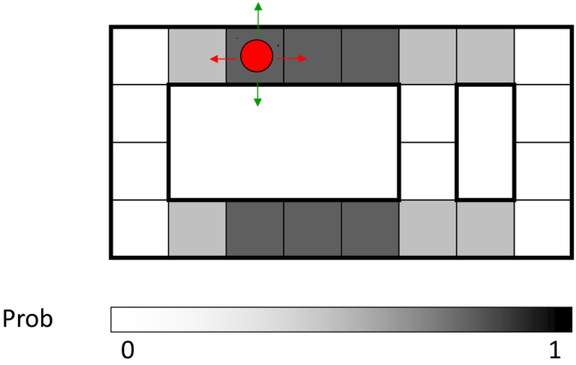


Figure: Belief estimation over time[2].

[2](Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Gridworld

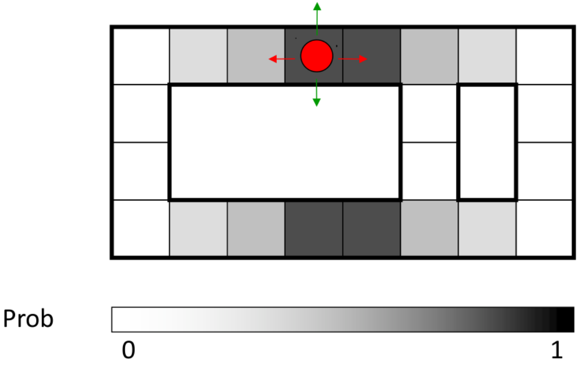


Figure: Belief estimation over time[2].

[2](Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Gridworld

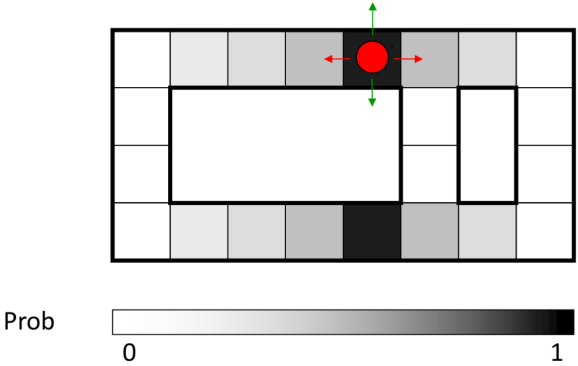


Figure: Belief estimation over time[2].

[2](Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Gridworld

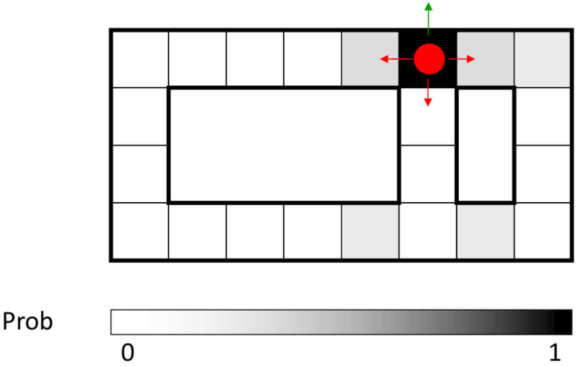


Figure: Belief estimation over time[2].

Idea: Maybe PO control is all about state estimation?

[2](Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Gridworld

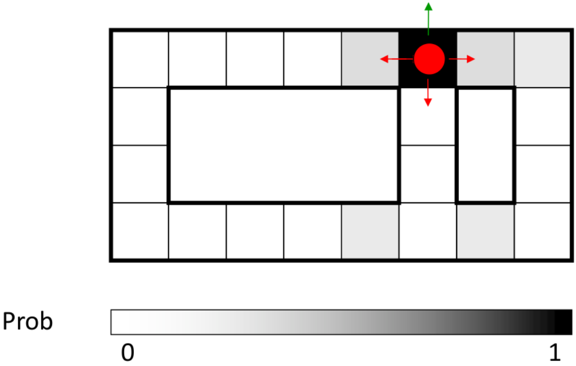
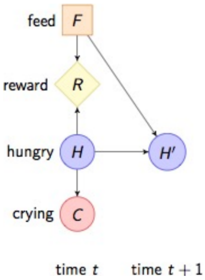


Figure: Belief estimation over time[2].

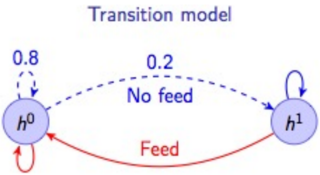
Idea: Maybe PO control is all about state estimation? NO

[2] (Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Crying Baby



(a) Model[3]



(b) Transition[3]

Details:

- States: not hungry (c=0), hungry (h=1)
- Actions: no feed (f=0), feed (f=1)
- Observations: no cry (c=0), cry (c=1)
- Costs: 5 to feed, 10 if hungry

[3] (Possibly recursive) image credit to Chris Amato's PORL slides.

Belief Update Example: Crying Baby

(a) Transitions.

Next State	State, Action			
	$h=0$, $f=0$	$h=0$, $f=1$	$h=1$, $f=0$	$h=1$, $f=1$
$h=0$	0.8	1.0	0.0	1.0
$h=1$	0.2	0.0	1.0	0.0

(b) Observations.

Observation	State	
	$h=0$	$h=1$
$c=0$	0.9	0.2
$c=1$	0.1	0.8

Table: Beliefs over time.

Time	Action	Observation	Belief	
			$h=0$	$h=1$
0			0.5	0.5
1	$f=0$	$c=1$	0.0928	0.9072
2	$f=1$	$c=0$	1.0	0.0
3	$f=0$	$c=0$	0.9759	0.0241
4	$f=0$	$c=0$	0.9701	0.0299
5	$f=0$	$c=1$	0.4624	0.5376

Note: Never really sure if baby is hungry!
Keypoint: Not always about state estimation!

Northeastern University

POMDPs as History-MDPs

Given: POMDP tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, p, T, O, R, \gamma \rangle$

Definition: History-MDP tuple $M_h \doteq \langle \mathcal{S}_h, \mathcal{A}_h, T_h, R_h, \gamma \rangle$

- State space $\mathcal{S}_h \doteq \mathcal{H}$
- Action space $\mathcal{A}_h \doteq \mathcal{A}$
- Transition function $T_h(h, a, h') \doteq \mathbb{E}_{s|h} \left[\sum_{s', o} T(s, a, s') O(a, s', o) \, \mathbb{I} \left[h' = hao \right] \right]$
- Reward function $R(h, a) \doteq \mathbb{E}_{s|h} \left[R(s, a) \right]$

Rejoice! MDP theory fully applies to POMDPs via history-MDPs!

- No need to rederive all of MDP theory for history-MDPs
- POMDPs just as easy as MDPs!

Northeastern University

POMDPs as History-MDPs

Given: POMDP tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, p, T, O, R, \gamma \rangle$

Definition: History-MDP tuple $M_h \doteq \langle \mathcal{S}_h, \mathcal{A}_h, T_h, R_h, \gamma \rangle$

- State space $\mathcal{S}_h \doteq \mathcal{H}$
- Action space $\mathcal{A}_h \doteq \mathcal{A}$
- Transition function
  $T_h(h, a, h') \doteq \mathbb{E}_{s|h} \left[\sum_{s', o} T(s, a, s') O(a, s', o) \, \mathbb{I}\left[h' = hao \right] \right]$
- Reward function $R(h, a) \doteq \mathbb{E}_{s|h} \left[R(s, a) \right]$

Rejoice! MDP theory fully applies to POMDPs via history-MDPs!

- No need to rederive all of MDP theory for history-MDPs
- ~~POMDPs just as easy as MDPs!~~
- POMDPs just as “easy” as VERY HARD MDPs!

Practical Difficulties:

- Exponential space size
- Never encounter same history twice
- Hard to generalize well
  Do similar histories imply similar optimal actions?
- Needle in a haystack problem

Northeastern University

POMDPs as Belief-MDPs

Given: POMDP tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, p, T, O, R, \gamma \rangle$

Definition: Belief-MDP tuple $M_b \doteq \langle \mathcal{S}_b, \mathcal{A}_b, T_b, R_b, \gamma \rangle$

- State space $S_b \doteq \mathcal{B} = \Delta\mathcal{S}$
- Action space $A_b \doteq \mathcal{A}$
- Transition function $T_b(b, a, b') \doteq \mathbb{E}_{s \sim b}\left[\sum_{s', o} T(s, a, s') O(a, s', o) \, \mathbb{I}\left[b' = bao\right]\right]$
- Reward function $R(b, a) \doteq \mathbb{E}_{s \sim b}\left[R(s, a)\right]$

Rejoice! MDP theory fully applies to POMDPs via belief-MDPs!

- No need to rederive all of MDP theory for belief-MDPs
- POMDPs just as easy as MDPs!

POMDPs as Belief-MDPs

Given: POMDP tuple $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, p, T, O, R, \gamma \rangle$

Definition: Belief-MDP tuple $M_b \doteq \langle \mathcal{S}_b, \mathcal{A}_b, T_b, R_b, \gamma \rangle$

- State space $S_b \doteq \mathcal{B} = \Delta\mathcal{S}$
- Action space $A_b \doteq \mathcal{A}$
- Transition function $T_b(b, a, b') \doteq \mathbb{E}_{s\sim b}\left[\sum_{s',o} T(s,a,s')O(a,s',o)\,\mathbb{I}\left[b' = bao\right]\right]$
- Reward function $R(b, a) \doteq \mathbb{E}_{s\sim b}\left[R(s,a)\right]$

Rejoice! MDP theory fully applies to POMDPs via belief-MDPs!

- No need to rederive all of MDP theory for belief-MDPs
- ~~POMDPs just as easy as MDPs!~~
- Well... no... but we're getting closer... with some big caveats!
  - Belief as sufficient statistic of history for control, good generalization
  - Similar beliefs usually imply similar optimal actions

Practical Difficulties:

- Continuous MDP even for discrete POMDPs
- Requires model of environment
- Estimating beliefs and belief updates is hard

Northeastern University

Deep Q-Networks (DQN)

"Human-level control through deep reinforcement learning" (Mnih et al., 2015)

Control Problem: Atari 2600, high-dimensional highly structured data

$$\mathcal{L}_{\hat{Q}} = \frac{1}{2}\left(r + \gamma \max_{a'} \hat{Q}(s', a') - \hat{Q}(s, a)\right)^2 \tag{15}$$

Frame Stacking: Approximate $\hat{s}_t \approx (o_t, o_{t-1}, o_{t-2}, o_{t-3})$

- $\hat{s}_t$ approximates a Markov state much more than o_t,
  e.g., movements and velocities

Limitations:

- Requires domain knowledge
  How many observations are enough?
- Works fine in many Atari games, but not all
- Not generalizable
- Not really about partial observability
  Main success about deep RL on high-dimensional highly structured data

Deep Recurrent Q-Networks (DRQN) - Recurrent Layers for Memory

"Deep recurrent q-learning for partially observable mdps" (Hausknecht and Stone, 2015)

Control Problem: Flickering Atari, frame obscured with $p = 50\%$

Approach: Ditch frame stacking, employ RNNs!

- $\hat{h}_t, y_t = F(\hat{h}_{t-1}, x_t)$!
- Theoretical "infinite" memory!
- Vanishing gradients problem
  - LSTMs (Hochreiter and Schmidhuber, 1997)
  - GRUs (Cho et al., 2014)

Keypoints:

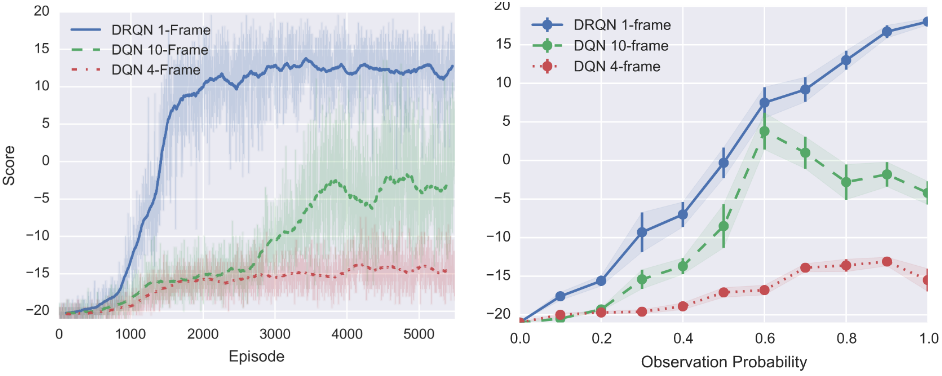
- Use RNN to combine all observation information over time, $\phi(h)$
- Combine RNN with any deep RL method, PO "solved"!

Limitations:

- Training RNNs is slow and requires A LOT of data
- Vanishing gradient still a problem for LSTMs and GRUs
- RNNs may not easily learn good history representations just via RL

Deep Recurrent Q-Networks (DRQN) - Recurrent Layers for Memory

"Deep recurrent q-learning for partially observable mdps" (Hausknecht and Stone, 2015)



(a) Flickering Pong (b) Policy Generalization

Figure: Left: Training performance. Right: Resistance to varying levels of flickering probability[4].

[4] Image and caption credit to Hausknecht and Stone, 2015.

Summary of Partial Observability

Partial Observability:

- Inevitable property of our world (read: our senses and sensors)
- PO as default case, FO as special case

Great divide between theory and practice:

- Theory: POMDPs interpretable as MDPs
- Practice: POMDPs VERY hard MDPs

History representation learning as sequence modeling

- Big recent advances, especially NLP, e.g., attention, transformers, LLMs
- However, NLP benefits from BIG DATA
- While RL almost fundamentally about small data, sample efficiency

Personal hunch: Complex PO not solved by RL alone

- model-based approaches to internal state representation
- planning + model-free approaches to control

Next: Stateful PORL

Northeastern University

Stateful PORL

Groundbreaking new idea:

Stateful PORL

Groundbreaking new idea: Let's use state in PORL!

Stateful PORL

Groundbreaking new idea: Let's use state in PORL!

... ... Wait, what?

Stateful PORL

Groundbreaking new idea: Let's use state in PORL!

... ... Wait, what?

State available at training time:

- Agent learning algorithm uses state to improve agent policy
- Agent policy does not use state to select actions

Stateful PORL - Offline RL

Training with static dataset:

- FO dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_i^N$
- PO dataset $\mathcal{D} = \{(h_i, a_i, r_i, o_i, r_i)\}_i^N$
- PO+state dataset $\mathcal{D} = \{(h_i, s_i, a_i, r_i, s_i', o_i, r_i)\}_i^N$

Training with dynamic simulator:

- Exploit simulated state representation
- Other practical advantages: Physical safety, training efficiency
- Other practical concerns (sim-to-real):
  - State completeness (e.g., missing variables)
  - Observation realism (e.g., out-of-distribution issues)
  - Dynamics accuracy (e.g., impossible transitions, collisions, clipping)
- Multi-agent: Centralized Training for Decentralized Execution (CTDE)
- Single-agent: Offline Training for Online Execution (OTOE)

Question: How to use state (correctly) to improve history-policy?

Northeastern University

Stateful PORL - Asymmetric Actor-Critic

"Asymmetric Actor Critic for Image-Based Robot Learning" (Pinto et al., 2018)

Goal: Robot control from images

Topics:

- Robotic manipulation: Pick-n-place, push, move, etc.
- Vision based control: Third POV of workspace
- Goal based control (with explicit goal representation)
- Sim-to-real: Transfer from simulated training to real environment
- Asymmetric training: Exploit (compact) simulator state for training

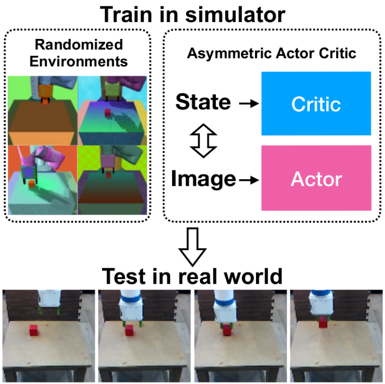
Asymmetric Training:

- History policy constrained by problem definition
  $\pi\colon \mathcal{H} \to \Delta\mathcal{A}$, where $\mathcal{H} \equiv$ history of images
- State critic as training construct, not used during execution
  $\hat{V}\colon \mathcal{S} \to \mathbb{R}$, where $\mathcal{S} \equiv$ simulator internal state

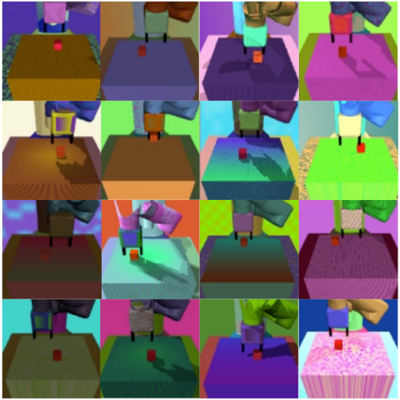
Northeastern University

Stateful PORL - Asymmetric Actor-Critic

"Asymmetric Actor Critic for Image-Based Robot Learning" (Pinto et al., 2018)



(a) Asymmetric training framework[5].



(b) Randomized simulated environments[5].

Northeastern University

[5] (Pinto et al., 2018)

Stateful PORL - Asymmetric Actor-Critic

"Asymmetric Actor Critic for Image-Based Robot Learning" (Pinto et al., 2018)

(Symmetric) Actor-Critic:

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

$$Q^\pi(h_t, a_t) \approx r_t + \gamma \hat{V}(h_t a_t o_t)$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\left(r_t + \gamma \hat{V}(h_t a_t o_t) - \hat{V}(h_t)\right)^2$$

Asymmetric Actor-Critic:

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

$$Q^\pi(s_t, a_t) \approx r_t + \gamma \hat{V}(s_{t+1})$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\left(r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)\right)^2$$

Question: Does the asymmetric $\nabla J \stackrel{?}{=}$ equality hold?

Answer: Not in general =(

- *Can* work in practice (e.g., for reactive control, as in paper!)
- *But* concerning theoretical issues re: partial observability

Stateful PORL - Asymmetric Actor-Critic

"Asymmetric Actor Critic for Image-Based Robot Learning" (Pinto et al., 2018)

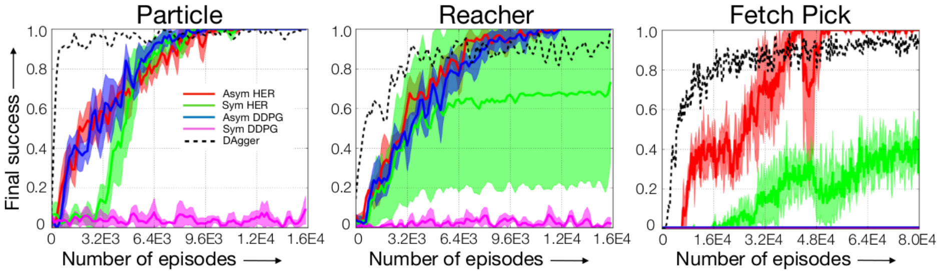


Figure: We show that asymmetric inputs for training outperforms symmetric inputs by significant margins. The shaded region corresponds to ± 1 standard deviation across 5 random seeds. Although the behaviour cloning (BC) by expert imitation baseline (dashed lines) learn faster initially, it saturates to a sub optimal value compared to asymmetric HER. Also note that the BC baseline doesn't include the iterations the expert poicy was trained on[6].

[6] Image and caption credit to Pinto et al., 2018.

Stateful PORL - Theory of Stateful Values

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

Questions of interest:

- Is $V^\pi(s)$ well-defined in PORL as in FORL?

$$V^\pi(s) \stackrel{?}{=} \mathbb{E}_{a|s}\left[R(s,a) + \gamma\, \mathbb{E}_{s'|s,a}\left[V^\pi(s')\right]\right] \tag{16}$$

- Is $V^\pi(s)$ an unbiased estimate of $V^\pi(h)$, i.e.,

$$V^\pi(h) \stackrel{?}{=} \mathbb{E}_{s|h}\left[V^\pi(s)\right] \tag{17}$$

(sufficient condition for unbiased gradient)

Stateful PORL - Theory of Stateful Values

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

Definition of $V^\pi(s)$: Unique solution to the stateful Bellman equality

$$V^\pi(s) = \mathbb{E}_{a|s}\left[R(s,a) + \gamma\,\mathbb{E}_{s'|s,a}\left[V^\pi(s')\right]\right] \tag{18}$$

Red flag:

- $V^\pi(s)$ measures expected performance
- Performance depends on:
  - Extrinsic context, i.e., environment behavior
  - Intrinsic context, i.e., agent behavior

Issue: What is the nature of $\Pr(a \mid s)$?

Stateful PORL - Theory of Stateful Values

“Unbiased Asymmetric Reinforcement Learning under Partial Observability” (Baisero and Amato, 2022)

Issue: What is the nature of $\Pr(a \mid s)$?

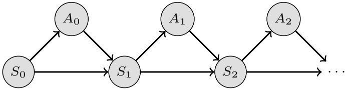


Figure: MDP graphical model.

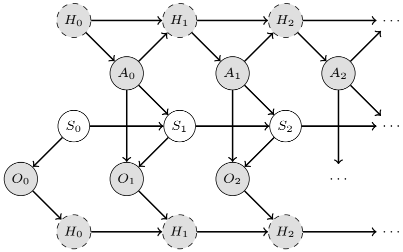


Figure: POMDP graphical model.

- **FORL:** Direct (causal) relationship $\Pr(a \mid s) = \pi(a; s)$
- **PORL:** Unclear relationship $\Pr(a \mid s) =$ ¯_(ツ)_/¯
  - Policy acts based on histories, not state
  - $\Pr(A_t = a \mid S_t = s)$ dependent on time

$\implies$ State insufficient predictor of behavior, consequently of performance!

Conclusion: In general case, $\Pr(a \mid s)$ is ill-defined, $V^\pi(s)$ is ill-defined

Stateful PORL - Theory of Stateful Values

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

Theorem (State Value Bias)

Even if/when well-defined, $V^\pi(s)$ is biased. Assume well-defined $V^\pi(s)$, then

$$V^\pi(h) = \mathbb{E}_{s|h}\left[V^\pi(s)\right] \tag{19}$$

is not guaranteed to hold.

Proof by contradiction.

Assume Equation (19) holds. Take $h_A \neq h_B$ s.t. $b(h_A) = b(h_B)$; a common occurrence in POMDPs. Different histories $h_A \neq h_B$ imply different behaviors and $V^\pi(h_A) \neq V^\pi(h_B)$. Same beliefs $b(h_A) = b(h_B)$ imply

$$V^\pi(h_A) = \mathbb{E}_{s|h_A}\left[V^\pi(s)\right] = \mathbb{E}_{s|h_B}\left[V^\pi(s)\right] = V^\pi(h_B)\,. \tag{20}$$

A contradiction is found, therefore Equation (19) does not hold. ∎

Stateful PORL - Theory of Stateful Values

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

Reactive Policy: State-observation value function

$$V^\pi(s, o) = \mathbb{E}_{a \sim \pi(o)} \left[R(s, a) + \gamma \, \mathbb{E}_{s', o' \mid s, a} \left[V^\pi(s', o') \right] \right] \tag{21}$$

Theorem (State-Observation Value Bias, or Lack Thereof)

$$V^\pi(h) = \mathbb{E}_{s \mid h} \left[V^\pi(s, o_h) \right] \tag{22}$$

Stateful PORL - Theory of Stateful Values

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

Reactive Policy: State-observation value function

$$V^\pi(s, o) = \mathbb{E}_{a\sim\pi(o)}\left[R(s, a) + \gamma\, \mathbb{E}_{s', o' \mid s, a}\left[V^\pi(s', o')\right]\right] \tag{21}$$

Theorem (State-Observation Value Bias, or Lack Thereof)

$$V^\pi(h) = \mathbb{E}_{s\mid h}\left[V^\pi(s, o_h)\right] \tag{22}$$

History Policy: History-state value function

$$V^\pi(h, s) = \mathbb{E}_{a\sim\pi(h)}\left[R(s, a) + \gamma\, \mathbb{E}_{s', o \mid s, a}\left[V^\pi(hao, s')\right]\right] \tag{23}$$

Theorem (History-State Value Bias, or Lack Thereof)

$$V^\pi(h) = \mathbb{E}_{s\mid h}\left[V^\pi(h, s)\right] \tag{24}$$

Stateful PORL - Theory of Stateful Values

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

Reactive Policy: State-observation value function

$$V^\pi(s,o) = \mathbb{E}_{a\sim\pi(o)}\left[R(s,a) + \gamma\,\mathbb{E}_{s',o'|s,a}\left[V^\pi(s',o')\right]\right] \tag{21}$$

Theorem (State-Observation Value Bias, or Lack Thereof)

$$V^\pi(h) = \mathbb{E}_{s|h}\left[V^\pi(s,o_h)\right] \tag{22}$$

History Policy: History-state value function

$$V^\pi(h,s) = \mathbb{E}_{a\sim\pi(h)}\left[R(s,a) + \gamma\,\mathbb{E}_{s',o|s,a}\left[V^\pi(hao,s')\right]\right] \tag{23}$$

Theorem (History-State Value Bias, or Lack Thereof)

$$V^\pi(h) = \mathbb{E}_{s|h}\left[V^\pi(h,s)\right] \tag{24}$$

Keypoints:

- Decision making context w/ state always well-defined
- Decision making context w/o state not always well-defined, e.g., $V^\pi(o)$

Stateful PORL - Theory of Stateful Values

Table: Theoretical properties of value functions.

Observations	Policy	Value	Well Defined	Unbiased
Generic $o \sim O(a, s)$	History $\pi(h)$	$V^\pi(h)$	✓	✓
		$V^\pi(s)$		
		$V^\pi(s, o)$		
		$V^\pi(h, s)$	✓	✓
		$V^\pi(h, z)$	✓	✓
Generic $o \sim O(a, s)$	Reactive $\pi(o)$	$V^\pi(h)$	✓	✓
		$V^\pi(s)$		
		$V^\pi(s, o)$	✓	✓
		$V^\pi(h, s)$	✓	✓
		$V^\pi(h, z)$	✓	✓
Reactive $o \sim O(s)$	Reactive $\pi(o)$	$V^\pi(h)$	✓	✓
		$V^\pi(s)$	✓	
		$V^\pi(s, o)$	✓	✓
		$V^\pi(h, s)$	✓	✓
		$V^\pi(h, z)$	✓	✓
Reactive $o \sim O(s)$, w/o aliasing	Reactive $\pi(o)$	$V^\pi(h)$	✓	✓
		$V^\pi(s)$	✓	✓
		$V^\pi(s, o)$	✓	✓
		$V^\pi(h, s)$	✓	✓
		$V^\pi(h, z)$	✓	✓

Stateful PORL - Asymmetric Actor-Critic (Revisited)

"Asymmetric Actor Critic for Image-Based Robot Learning" (Pinto et al., 2018)

Question: Why does it work if there are issues?

[7] Image credit to Pinto et al., 2018

Stateful PORL - Asymmetric Actor-Critic (Revisited)

"Asymmetric Actor Critic for Image-Based Robot Learning" (Pinto et al., 2018)

Question: Why does it work if there are issues?
Answer: Reactive tasks, almost fully observable
Observation information approximates state information

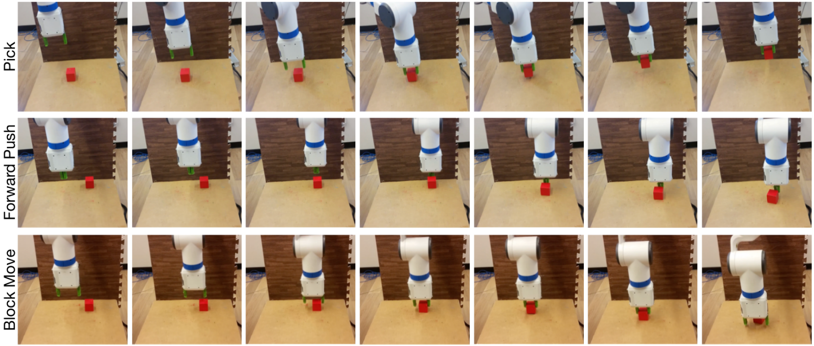


Figure: Task observations[7] are clean, occlusionless, almost fully observable.

[7] Image credit to Pinto et al., 2018

Stateful PORL - Unbiased Asymmetric Actor-Critic

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

(Symmetric) Actor-Critic:

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

$$Q^\pi(h_t, a_t) \approx r_t + \gamma \hat{V}(h_t a_t o_t)$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\left(r_t + \gamma \hat{V}(h_t a_t o_t) - \hat{V}(h_t)\right)^2$$

(Unbiased) Asymmetric Actor-Critic:

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

$$Q^\pi(h_t, s_t, a_t) \approx r_t + \gamma \hat{V}(h_{t+1}, s_{t+1})$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\left(r_t + \gamma \hat{V}(h_{t+1}, s_{t+1}) - \hat{V}(h_t, s_t)\right)^2$$

Question: Does the asymmetric $\nabla J \stackrel{?}{=}$ equality hold?

Northeastern University

Stateful PORL - Unbiased Asymmetric Actor-Critic

"Unbiased Asymmetric Reinforcement Learning under Partial Observability" (Baisero and Amato, 2022)

(Symmetric) Actor-Critic:

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

$$Q^\pi(h_t, a_t) \approx r_t + \gamma \hat{V}(h_t a_t o_t)$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\left(r_t + \gamma \hat{V}(h_t a_t o_t) - \hat{V}(h_t)\right)^2$$

(Unbiased) Asymmetric Actor-Critic:

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

$$Q^\pi(h_t, s_t, a_t) \approx r_t + \gamma \hat{V}(h_{t+1}, s_{t+1})$$

$$\mathcal{L}_{\hat{V}} = \frac{1}{2}\left(r_t + \gamma \hat{V}(h_{t+1}, s_{t+1}) - \hat{V}(h_t, s_t)\right)^2$$

Question: Does the asymmetric $\nabla J \stackrel{?}{=}$ equality hold?

Theorem (Stateful Policy Gradient)

$$\nabla J = \mathbb{E}\left[\sum_t \gamma^t Q^\pi(h_t, s_t, a_t) \nabla \log \pi(a_t; h_t)\right] \tag{25}$$

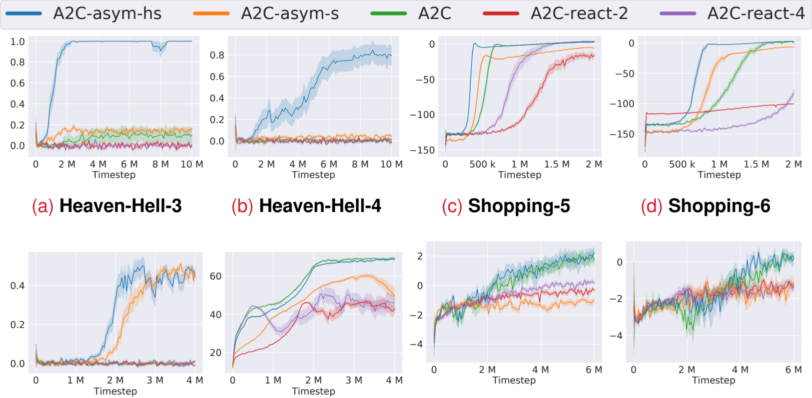
Keypoints:

- Applicable to generic non-reactive problems and policies
- Small implementation change, huge empirical advantages

Northeastern University

Stateful PORL - Unbiased Asymmetric Actor-Critic

“Unbiased Asymmetric Reinforcement Learning under Partial Observability” (Baisero and Amato, 2022)



(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**

Figure: Learning performance curves of episodic returns averaged over the last 100 episodes, with statistics computed over 20 independent runs. Shaded areas are centered around the empirical mean and show one standard error of the mean.

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

Question: How to apply asymmetry to value-based methods?

- Actor-critic has two learned models $\pi, \hat{V}$
- Value-based methods only has one $\hat{Q}(h, a)$ (behavior *and* evaluator)
  $\Longrightarrow$ Introduce training construct $\hat{U}(h, s, a)$

Bottom-Up Approach: Develop asymmetric versions of basic algorithms

- Asymmetric Policy Iteration (API)
- Asymmetric Action-Value Iteration (AAVI)
- Asymmetric Q-learning (AQL)
- Asymmetric DQN (ADQN)

Mutual Consistency: $\hat{Q}$ and $\hat{U}$ are mutually consistent if

$$\hat{Q}(h, a) = \mathbb{E}_{s|h}\left[\hat{U}(h, s, a)\right] \tag{26}$$

Northeastern University

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

Algorithm Asymmetric Policy Iteration (API)

Require: U_0, Q_0, π_0 arbitrarily initialized tabular models.
Ensure: $\lim_{k\to\infty}\{U_k, Q_k, \pi_k\} = \{U^*, Q^*, \pi^*\}$.

```
1: for k ← 0, 1, 2, 3, ... do
2:     U_{k+1} ← U_k
3:     repeat
4:         U_{k+1} ← B_{π_k} U_{k+1}
5:     until convergence
6:     Q_{k+1} ← E U_{k+1}
7:     π_{k+1} ← g(Q_{k+1})
8: end for
```

Theorem (API Optimality)

The sequences U_k, Q_k, and π_k generated by API converge to U^*, Q^*, and π^*.

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

Algorithm Asymmetric Action-Value Iteration (AAVI)

Require: U_0, Q_0 arbitrarily initialized tabular models.
Ensure: $\lim_{k\to\infty}\{U_k, Q_k\} = \{U^*, Q^*\}$.
1: **for** $k \leftarrow 0, 1, 2, 3, \ldots$ **do**
2: $\quad U_{k+1} \leftarrow B_{g(Q_k)} U_k$
3: $\quad Q_{k+1} \leftarrow E U_{k+1}$
4: **end for**

Theorem (AAVI Optimality)

The sequences U_k and Q_k generated by AAVI converge to U^* and Q^*.

Northeastern University

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

Algorithm Asymmetric Q-Learning (AQL)

Require: U, Q mutually consistent tabular models.
Ensure: $\{U, Q\} \to \{U^*, Q^*\}$.

```
while True do
    Initialize history and state (h, s)
    while s is not terminal do
        Choose action a from ε-greedy policy on Q
        Take action a, observe r, s', o
        y ← r + γU(hao, s', argmax_a' Q(hao, a'))
        U(h, s, a) ← (1 − α)U(h, s, a) + αy
        Q(h, a) ← (1 − α)Q(h, a) + αy
        (s, h) ← (s', hao)
    end while
end while
```

Northeastern University

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

Theorem (AQL Optimality)

Assume stepsizes α_k satisfying the following asymptotic conditions,

$$\sum_{k=0}^{\infty} \alpha_k = \infty \,, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty \,. \tag{27}$$

If Q_0, U_0 are mutually consistent ($Q_0 = EU_0$), then the sequences Q_k and U_k generated by AQL converge to Q^* and U^* with probability 1.

Northeastern University

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

(Symmetric) DQN:

$$\mathcal{L}_{\hat{Q}} = \frac{1}{2}\left(y - \hat{Q}(h,a)\right)^2$$
$$y = r + \gamma \max_{a'} \hat{Q}(hao, a')$$

Asymmetric DQN:

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

Note: Ignoring implementation details, e.g., target networks

Keypoints:

- Same y for $\mathcal{L}_{\hat{Q}}$ and $\mathcal{L}_{\hat{U}}$
- Difference between

$$\max_{a'} \hat{Q}(hao, a') \tag{28}$$
$$\max_{a'} \hat{U}(hao, s', a') \tag{29}$$
$$\hat{U}(hao, s', \operatorname*{argmax}_{a'} \hat{Q}(hao, a')) \tag{30}$$

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)

Algorithm Asymmetric DQN (ADQN)

Require: $\hat{U}, \hat{Q}$ deep models parameterized by θ.

```
Initialize parameters θ
Initialize and prepopulate episode buffer
while True do
    From simulated environment, sample and append episodes to episode buffer
    From episode buffer, sample batch of transitions {(h_i, s_i, a_i, r_i, s'_i, o_i)}_{i=1}^N
    L_U ← (1/N) Σ_{i=1}^N 𝓛_Û(h_i, s_i, a_i, r_i, s'_i, o_i).
    L_Q ← (1/N) Σ_{i=1}^N 𝓛_Q̂(h_i, s_i, a_i, r_i, s'_i, o_i).
    Perform gradient step on θ using ∇_θ(L_U + L_Q)
end while
```

Where:

$$\{(h_i, s_i, a_i, r_i, s'_i, o_i)\}_{i=1}^N$$

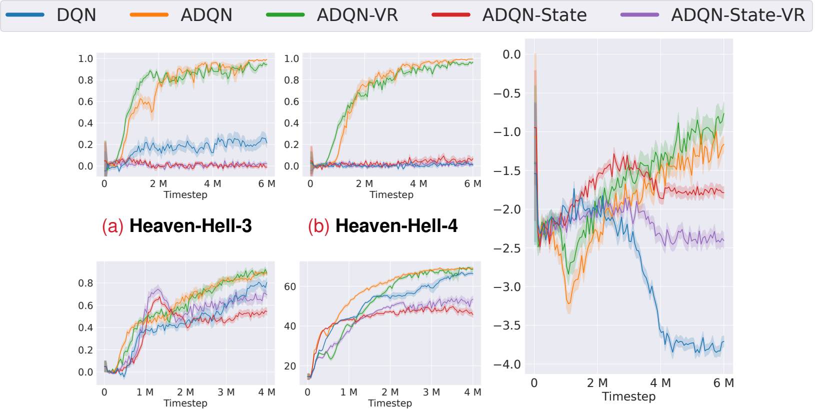
$$L_U \leftarrow \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\hat{U}}(h_i, s_i, a_i, r_i, s'_i, o_i).$$

$$L_Q \leftarrow \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\hat{Q}}(h_i, s_i, a_i, r_i, s'_i, o_i).$$

$$\nabla_\theta(L_U + L_Q)$$

Stateful PORL - Asymmetric DQN

"Asymmetric DQN for partially observable reinforcement learning" (Baisero, Daley, and Amato, 2022)



(a) **Heaven-Hell-3** (b) **Heaven-Hell-4**

(c) **Car-Flag** (d) **Cleaner**

(e) **Memory-Four-Rooms-7x7**

Figure: Performance curves showing episodic returns averaged over the last 100 completed episodes, with statistics computed over 20 independent runs. The shaded areas represent one standard error around the mean.

Stateful PORL - Learning Belief Representations for PORL

"Learning belief representations for partially observable deep RL" (Wang et al., 2023)

Objective: Train and use a sepresentation of belief $b(h)$ for PORL

Compact Representation of State $\phi(s)$

- Train state and observation representations $\hat{u}_s = \phi(s), \hat{u}_o = \phi(o)$, bisimulation $\hat{r}, \hat{u}_{s'}, \hat{u}_{o'} = g(\hat{u}_s, \hat{u}_o, \hat{a})$
- Avoid redundancy via information bottleneck $\implies$ compact $\phi(s)$
- Throw everything away except $\phi(s)$

Generative Belief Modeling with VAEs (Kingma, Welling, et al., 2019)

- Train generative model $p(\hat{u} \mid h, z)$, discriminative model $q(z \mid \hat{u}, h)$
- Generate sample-based belief representation $b(h) = \{\hat{u}_i\}_{i=1}^N$
- Generate belief representation $\hat{b} = W_{agg}\left(\frac{1}{n}\sum_{i=1}^n W_{enc}\left(\hat{u}_i\right)\right)$

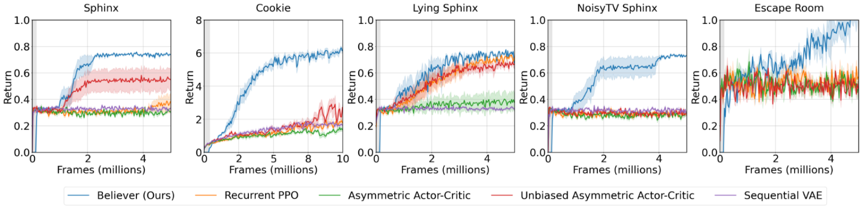
Policy Training

- Train belief-observation policy $\pi(\hat{b}, o)$ with standard FORL

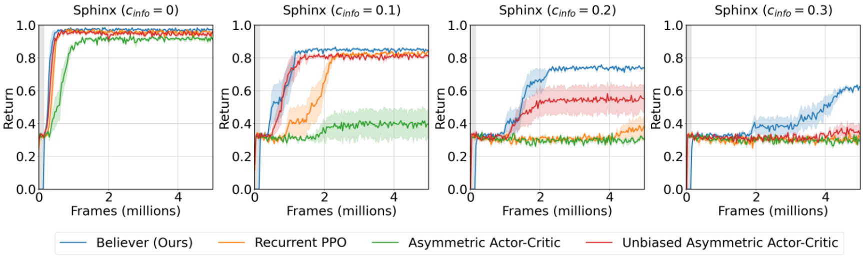
Northeastern University

Stateful PORL - Learning Belief Representations for PORL

"Learning belief representations for partially observable deep RL" (Wang et al., 2023)



(a) Performance of Believer and Asym-A2C variants on PO tasks[8].



(b) Resiliance to cost of information gathering[8].

[8] Image credit to Wang et al., 2023.

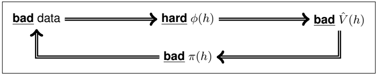
Stateful PORL - Role of State

Question: Why does stateful RL work so well?

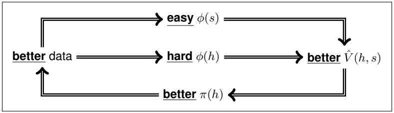
- State provides valuable information
- State boosts exploration
- State helps ground values before history

My hypothesis: (WIP)

- Not about information
  Nothing special about state
- Tiny bit exploration
- Primarily representation learning
  - $\phi(s)$ easier than $\phi(h)$
  - $\implies$ $\hat{V}(h, s)$ faster than $\hat{V}(h)$
  - $\implies$ better behavior
  - $\implies$ better data
  - $\implies$ better $\phi(s)$ and $\phi(h)$
  - $\implies$ $\hat{V}(h, s)$ faster than $\hat{V}(h)$
  - $\implies$ better behavior
  - $\implies$ better data
  - $\implies$ . . .



(a) A vicious Actor-Critic cycle.



(b) A better Actor-Critic cycle.

Northeastern University

References I

Baisero, Andrea and Christopher Amato (2022). "Unbiased Asymmetric Reinforcement Learning under Partial Observability". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*.

Baisero, Andrea, Brett Daley, and Christopher Amato (2022). "Asymmetric DQN for partially observable reinforcement learning". In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*.

Cho, Kyunghyun et al. (2014). "On the properties of neural machine translation: Encoder-decoder approaches". In: arXiv: 1409.1259 [cs.CL].

Hausknecht, Matthew and Peter Stone (2015). "Deep recurrent q-learning for partially observable mdps". In: *2015 aaai fall symposium series*.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

References II

Kingma, Diederik P, Max Welling, et al. (2019). "An introduction to variational autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4, pp. 307–392.

Littman, Michael L (1994). "Memoryless policies: Theoretical limitations and practical results". In.

Mnih, Volodymyr et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533.

Pinto, Lerrel et al. (2018). "Asymmetric Actor Critic for Image-Based Robot Learning". In: *14th Robotics: Science and Systems, RSS 2018*. MIT Press Journals.

Wang, Andrew et al. (2023). "Learning belief representations for partially observable deep RL". In: *International Conference on Machine Learning*.