Periodic agent-state based Q-learning (PASQL) for POMDPs

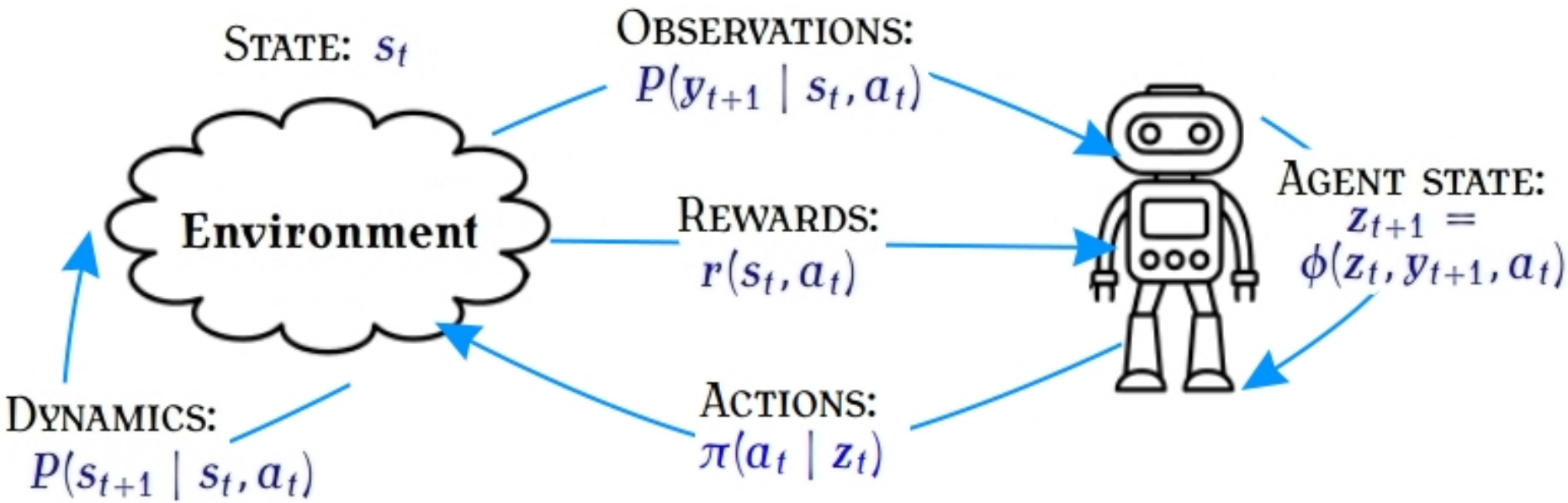
Amit Sinha[1] Matthieu Geist[2] Aditya Mahajan[1]

[1]McGill University, Mila

[2]Cohere

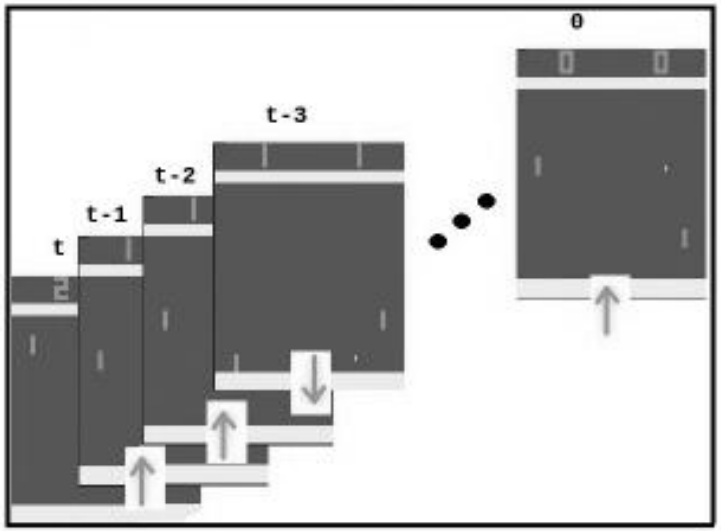
November 13, 2024

Partially observable MDP (POMDP) with agent states



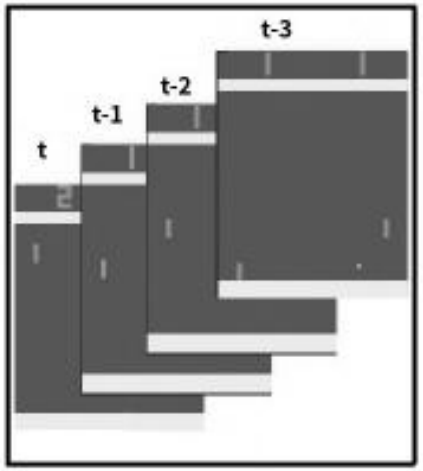
Example: Agent states in Atari-pong

Observation:  **Action:** ↑/↓



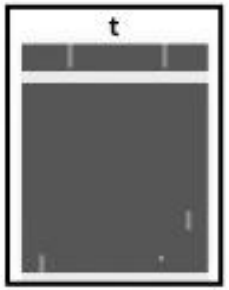
History

$$z_t = y_{0:t}, a_{0,t-1}$$



Finite Window (Size 4)

$$z_t = y_{t-4:t}$$



Abstract Features

$$z_t =$$

(x,y) = (0.1, 0.1)
(x,y) = (0.9, 0.2)
(x,y) = (0.8, 0.1)

Main problem

1. Agent state Z_t may be non-Markovian (e.g. due to function approx. in neural networks)

Main problem

1. Agent state Z_t may be non-Markovian (e.g. due to function approx. in neural networks)
2. **Non-stationary** policies will do better than **stationary** ones

Why periodic policies?

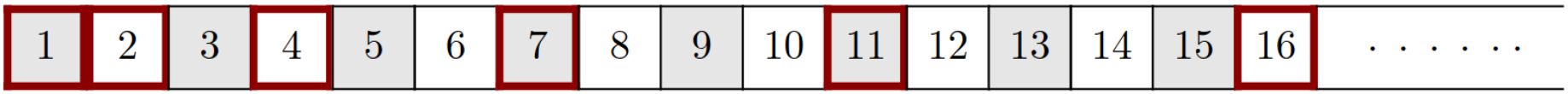
Definition (period L): $\pi = (\pi_1, \cdots, \pi_L, \pi_1, \cdots, \pi_L, \cdots)$

Why periodic policies?

Definition (period L): $\pi = (\pi_1, \cdots, \pi_L, \pi_1, \cdots, \pi_L, \cdots)$

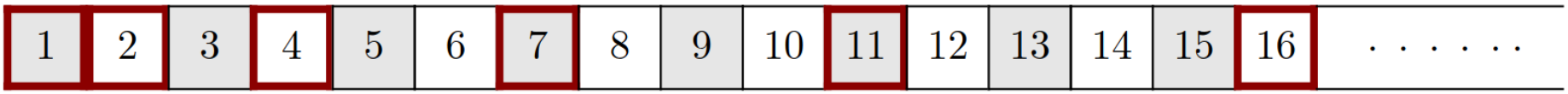
Reason : Non-stationary policies are not realizable in practice for infinite horizon problems.

Periodic performance



Period (L)	1	2	3	4	5	6	7	8	**9**	10
$J_L^\star$	4.022	4.022	7.479	6.184	8.810	7.479	9.340	8.488	**9.607**	8.810

Periodic performance



Period (L)	1	2	3	4	5	6	7	8	**9**	10
$J_L^\star$	4.022	4.022	7.479	6.184	8.810	7.479	9.340	8.488	**9.607**	8.810

Periodic policies perform better than **stationary policies** !

Periodic agent-state based Q-learning (PASQL) algorithm

Regular Q-learning:

$$Q_{t+1}(z, a) = Q_t(z, a) + \alpha_t(z, a)\Big[R_t + \gamma \max_{a' \in \mathsf{A}} Q_t(Z', a') - Q_t(z, a)\Big].$$

Periodic agent-state based Q-learning (PASQL) algorithm

Regular Q-learning:

$$Q_{t+1}(z,a) = Q_t(z,a) + \alpha_t(z,a)\Big[R_t + \gamma \max_{a' \in \mathsf{A}} Q_t(Z',a') - Q_t(z,a)\Big].$$

PASQL is just **Q-learning** for **periodic policies**!

$$Q^{\ell}_{t+1}(z,a) = Q^{\ell}_t(z,a) + \alpha^{\ell}_t(z,a)\Big[R_t + \gamma \max_{a' \in \mathsf{A}} Q^{[\![\ell+1]\!]}_t(z_{t+1},a') - Q^{\ell}_t(z,a)\Big]$$

$$\forall \ell \in \mathsf{L}, \quad [\![\ell]\!] = \ell \pmod{L}.$$

(PASQL)

Periodic agent-state based Q-learning (PASQL) algorithm

Regular Q-learning:

$$Q_{t+1}(z,a) = Q_t(z,a) + \alpha_t(z,a)\Big[R_t + \gamma \max_{a' \in \mathsf{A}} Q_t(Z',a') - Q_t(z,a)\Big].$$

PASQL is just **Q-learning** for **periodic policies**!

$$Q^{\ell}_{t+1}(z,a) = Q^{\ell}_t(z,a) + \alpha^{\ell}_t(z,a)\Big[R_t + \gamma \max_{a' \in \mathsf{A}} Q^{[\![\ell+1]\!]}_t(z_{t+1},a') - Q^{\ell}_t(z,a)\Big]$$

$$\forall \ell \in \mathsf{L}, \quad [\![\ell]\!] = \ell \pmod{L}.$$

(PASQL)

Problem:

Standard Q-learning convergence cannot apply (since Z_t is non-Markovian)

Main result: Convergence theorem

$$\{(Q_t^0, \ldots, Q_t^{L-1})\}_{t \geq 1} \to (Q_\mu^0, \ldots, Q_\mu^{L-1}) \quad \text{a.s.}$$

Main result: Convergence theorem

$$\{(Q_t^0, \ldots, Q_t^{L-1})\}_{t \geq 1} \to (Q_\mu^0, \ldots, Q_\mu^{L-1}) \quad \text{a.s.}$$

$$Q_\mu^\ell(z, a) = r_\mu^\ell(z, a) + \gamma \sum_{z' \in \mathsf{Z}} P_\mu^\ell(z' | z, a) \max_{a' \in \mathsf{A}} Q_\mu^{[\![\ell+1]\!]}(z', a')$$

$$\forall \ell \in \mathsf{L}, \forall (z, a) \in \mathsf{Z} \times \mathsf{A}$$

Main result: Convergence theorem

$$\{(Q_t^0, \dots, Q_t^{L-1})\}_{t \geq 1} \to (Q_\mu^0, \dots, Q_\mu^{L-1}) \quad \text{a.s.}$$

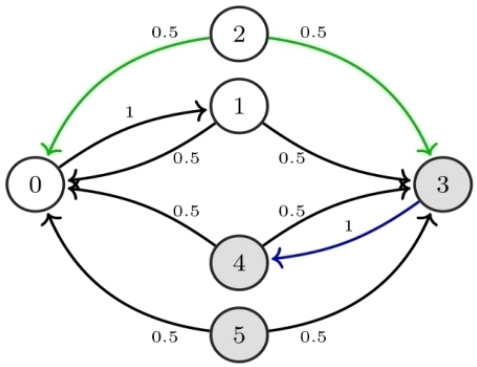
$$Q_\mu^\ell(z, a) = r_\mu^\ell(z, a) + \gamma \sum_{z' \in \mathsf{Z}} P_\mu^\ell(z' | z, a) \max_{a' \in \mathsf{A}} Q_\mu^{[\![\ell+1]\!]}(z', a')$$

$$\forall \ell \in \mathsf{L}, \forall (z, a) \in \mathsf{Z} \times \mathsf{A}$$

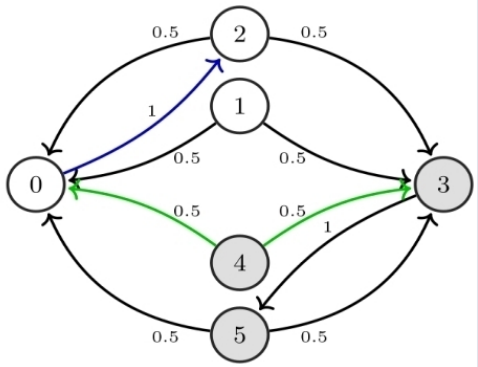
$$r_\mu^\ell(z, a) := \sum_{s \in \mathsf{S}} r(s, a) \zeta_\mu^\ell(s \mid z)$$

$$P_\mu^\ell(z' | z, a) := \sum_{(s, y') \in \mathsf{S} \times \mathsf{Y}} \mathbb{1}_{\{z' = \phi(z, y', a)\}} P(y' | s, a) \zeta_\mu^\ell(s | z)$$

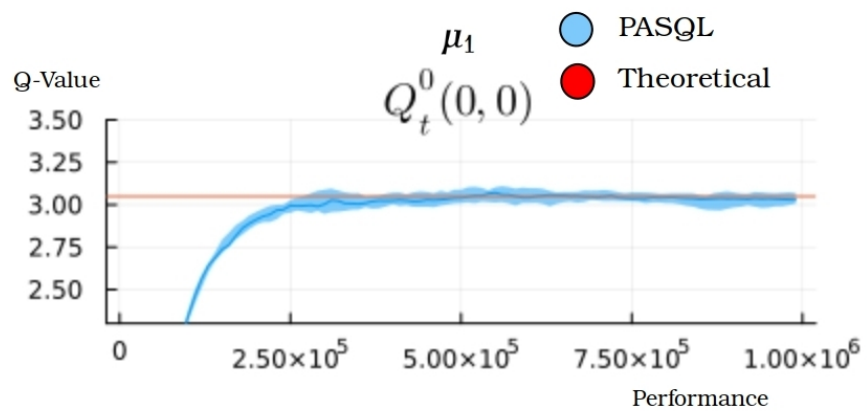
PASQL Numerical



(a) Dynamics under action 0.



(b) Dynamics under action 1.



$$\mu_1 := \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}, \mu_2 := \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \mu_3 := \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

Performance of converged *periodic* (L=2) policies.

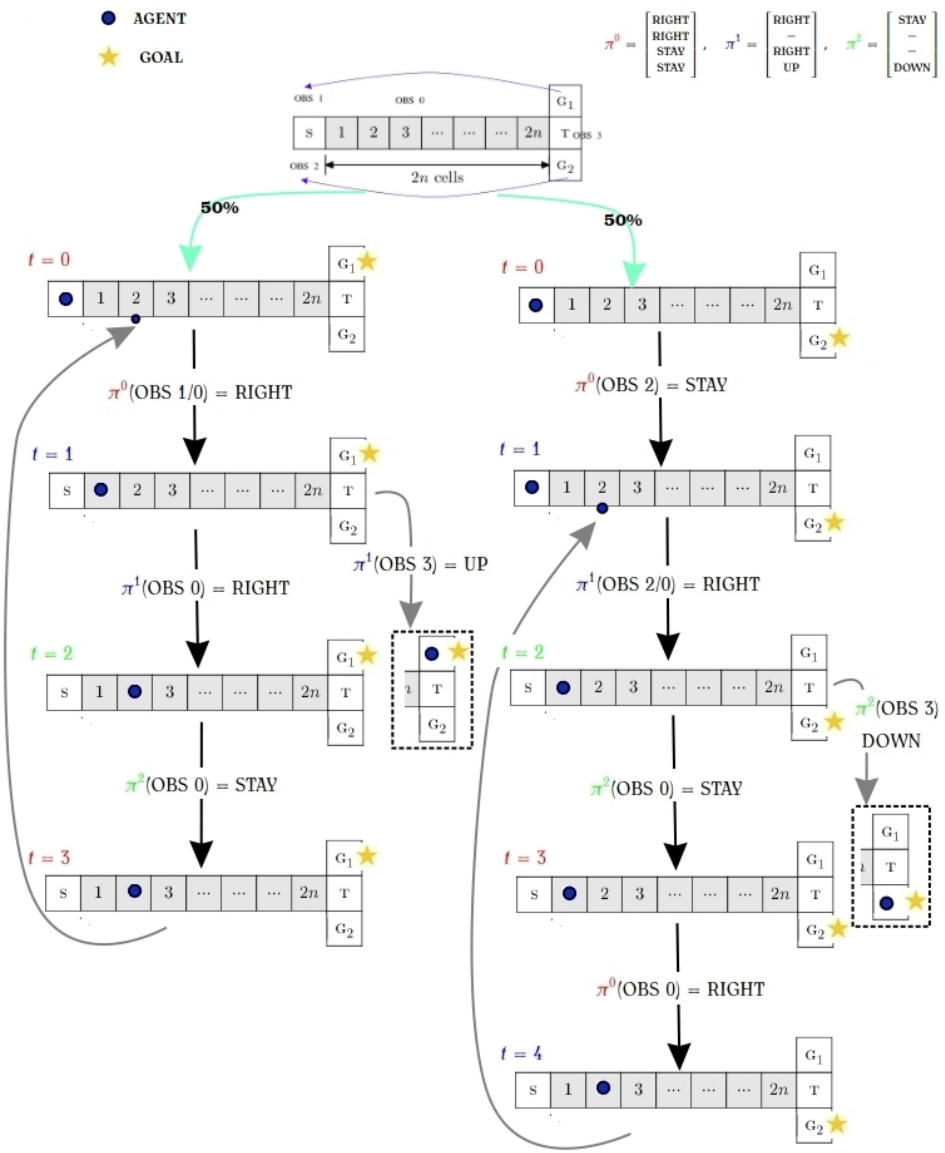
$J_2^\star$	$J^{\pi_{\mu_1}}$	$J^{\pi_{\mu_2}}$	$J^{\pi_{\mu_3}}$
6.793	6.793	1.064	0.532

$$\bar{\mu}_1 := [0.2; 0.8], \bar{\mu}_2 := [0.5; 0.5], \bar{\mu}_3 := [0.8; 0.2]$$

Performance of converged *stationary* policies.

$J_1^\star$	$J^{\bar{\pi}_{\mu_1}}$	$J^{\bar{\pi}_{\mu_2}}$	$J^{\bar{\pi}_{\mu_3}}$
2.633	0.0	1.064	2.633

T-maze example



Conclusions

1. We show that considering periodic policies over stationary policies can be beneficial

Conclusions

1. We show that considering periodic policies over stationary policies can be beneficial
2. We provide an algorithm PASQL with several useful theoretical properties

The End