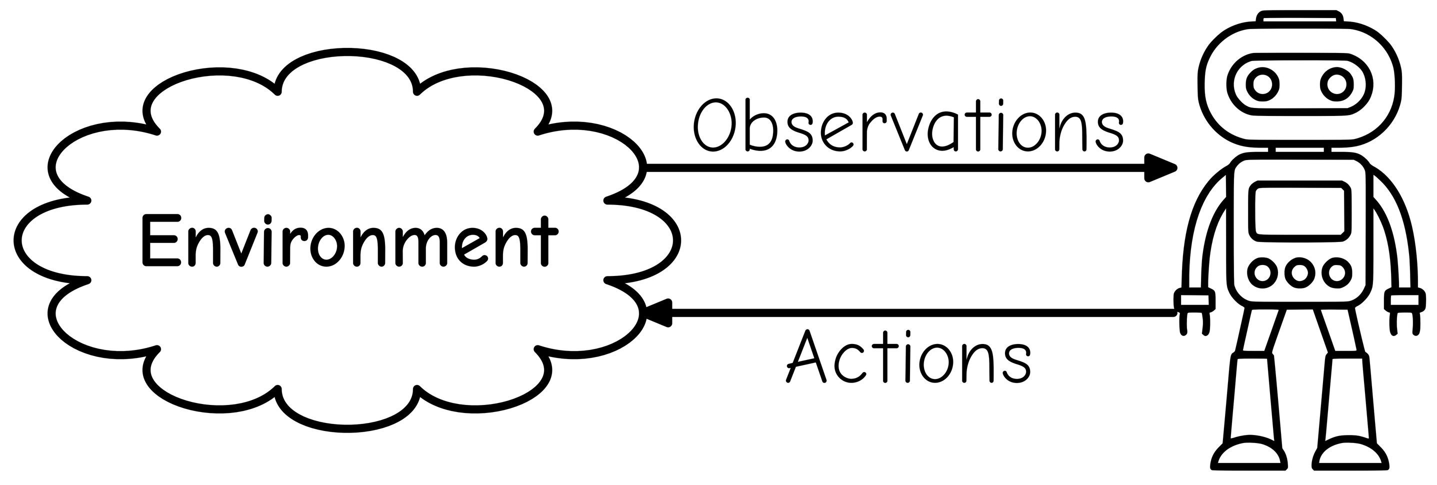
Q-learning for POMDPs

Does it converge? If so, to what?

Erfan Sayedsalehi, Nima Akbarzadeh, Amit Sinha, Aditya Mahajan

McGill University



Env state : $\mathbb{P}(S_{t+1} \mid S_t, A_t)$

Agent state : $Z_{t+1} = f(Z_t, Y_{t+1}, A_t)$

Recurrent Q-learning

$$\hat{Q}_{t+1}(z_t, a_t) = \hat{Q}_t(z_t, a_t) + \alpha_t(z_t, a_t)\left[R_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_t(z_{t+1}, a') - \hat{Q}_t(z_t, a_t)\right]$$

Conceptual Difficulty?

No DP corresponding to QL recursion because:

- $\mathbb{E}[R_t \mid Z_t, A_t]$ is not time-homogeneous.
- The controlled process $\{Z_t\}_{t \geq 1}$ is not Markov.

So **cannot** seek fixed point of "standard" Bellman op.

Q1: Does it converge?

Assumptions

(A1) Restrict to tabular setting

(A2) The exploration policy $\pi_{\text{exp}} \colon \mathcal{Z} \to \Delta(\mathcal{A})$ is such that the Markov chain $\{(S_t, Y_t, Z_t, A_t)\}_{t \geq 1}$ has a unique stationary distribution ξ. Plus, $\xi(s, y, z, a) > 0$.

(A3) $\alpha_t(z, a) = \mathbb{1}_{\{Z_t = z, A_t = a\}} / \sum_{\tau=1}^{t} \mathbb{1}_{\{Z_\tau = z, A_\tau = a\}}$

Convergence Result

Under (A1)–(A3), $\hat{Q}_t \to Q^*_\xi$ a.s., where

$$Q^*_\xi(z, a) = \sum_{s \in \mathcal{S}} \xi(s \mid z, a) \Big[r(s, a) + \gamma \sum_{(s', y')} P(s' \mid s, a)\, O(y' \mid s')\, V^*_\xi(f(z, y', a)) \Big]$$

Q2: How good is the converged solution?

- $\varepsilon = \sup_{t \geq 1} \max_{h_t, a_t} \Big| \mathbb{E}[r(S_t, a_t) \mid h_t, a_t] - \sum_{s \in \mathcal{S}} r(s, a_t)\, \xi(s \mid \sigma_t(h_t), a_t) \Big|$
- $\delta_{\mathfrak{F}} = \sup_{t \geq 1} \max_{h_t, a_t} d_{\mathfrak{F}}\Big(\mathbb{P}(Z_{t+1} = \cdot \mid h_t, a_t), P_\xi(\cdot \mid \sigma_t(h_t), a_t) \Big)$

IPM (Integral probability metric)

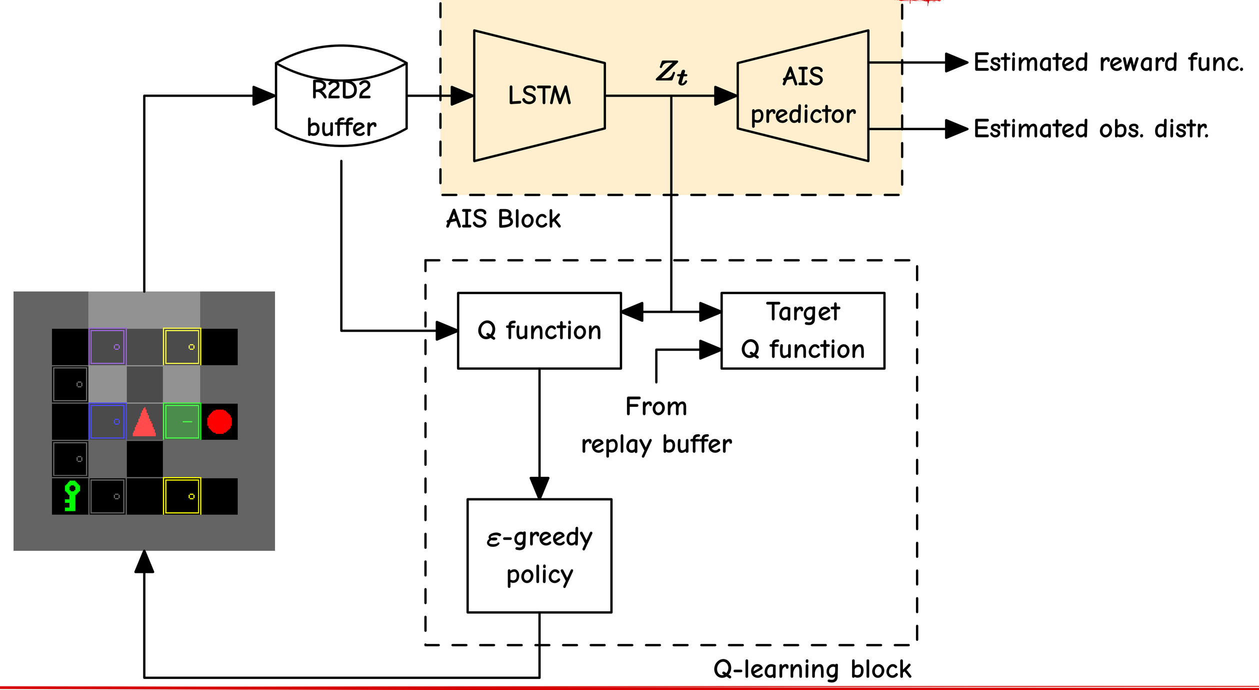
- $d_{\mathfrak{F}}$: IPM, e.g., TV, Wasserstein, MMD, etc.
- $\rho_{\mathfrak{F}}$: Depends on IPM, e.g., $\|\cdot\|_\infty$, $\mathrm{Lip}(\cdot)$, $\|\cdot\|_{\mathcal{H}}$, etc.

Approximation Result

For any history h_t:

$$\big| V^*_t(h_t) - V_t^{\pi^*_\xi \circ \sigma_t}(h_t) \big| \leq (1-\gamma)^{-1} \big[\varepsilon + \gamma \delta_{\mathfrak{F}} \rho_{\mathfrak{F}}(V^*_\xi) \big]$$

Q3: Can this help in RL?



Environment	RQL-AIS	ND-R2D2
SimpleCrossingS9N2	0.944 ± 0.007	0.757 ± 0.423
LavaCrossingS9N2	0.926 ± 0.014	0.934 ± 0.034
RedBlueDoors-8x8	0.977 ± 0.009	0.962 ± 0.018
MultiRoom-N2-S4	0.790 ± 0.049	0.839 ± 0.010
DoorKey-8x8	0.942 ± 0.038	0.371 ± 0.508
ObstructedMaze-1Dl	0.916 ± 0.020	0.000 ± 0.000
KeyCorridorS3R2	0.885 ± 0.038	0.000 ± 0.000
UnlockPickup	0.517 ± 0.474	0.000 ± 0.000

AIS loss vs. performance