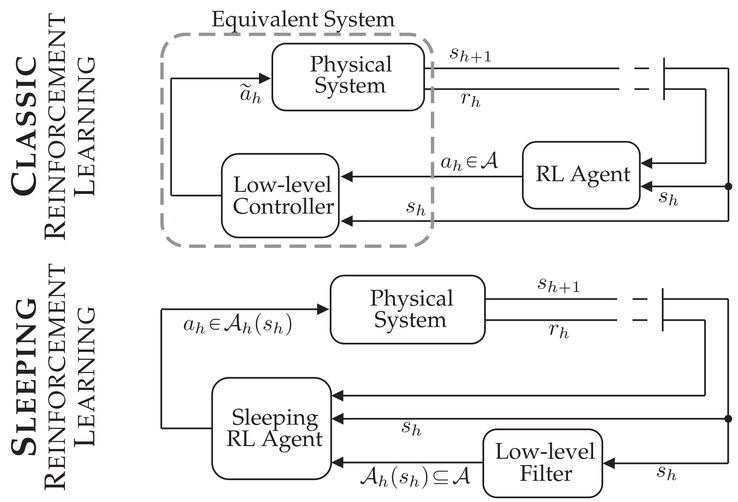
Sleeping Reinforcement Learning

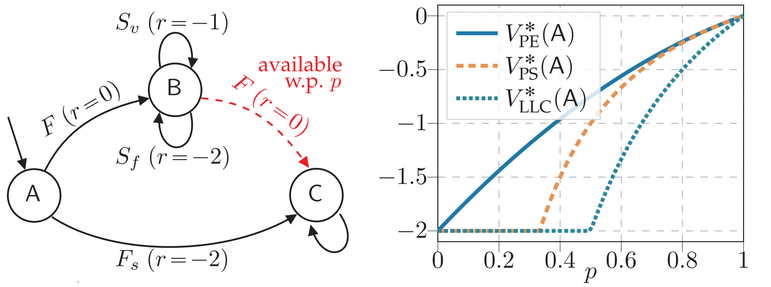
Simone Drago*, Marco Mussi*, and Alberto Maria Metelli
{simone.drago, marco.mussi, albertomaria.metelli}@polimi.it

Politecnico Milano 1863

Motivation



Example



Setting – Sleeping MDPs

$$\mathcal{M} := \big(\underbrace{\mathcal{S}, \mathcal{A}, P, R, H, \overline{s}}_{\text{Standard Finite States and Actions MDP}} , \underbrace{C}_{\text{Stochastic Action Availability Function}} \big)$$

We consider an **episodic** scenario and we call K the number of episodes

Available action set $\mathcal{A}_{k,h}(s) \subseteq \mathcal{A}$ is selected according to function C

Action Disclosure Models

- **Per-episode disclosure**: at the beginning of the current episode
- **Per-stage disclosure**: when the agent is asked to choose an action

References

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

O. D. Domingues, P. Ménard, E. Kaufmann, and M. Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, 2021.

R. D. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. In *Conference on Learning Theory*, 2008.

Per-Episode Disclosure

Per-Episode Regret

Given algorithm $\mathfrak{A}$ running $(\pi_k)_{k=1}^K$, we evaluate its performances in terms of **per-episode (PE) regret** w.r.t. the **optimal policy** π_k^* **for each episode**:

$$R_{\mathsf{PE}}(\mathfrak{A}, K) := \sum_{k \in [\![K]\!]} \left(V_{k,1}^{\pi_k^*}(\overline{s}) - V_1^{\pi_k}(\overline{s}) \right)$$

Lower Bound

Same as for MDPs ($|\mathcal{S}| = S,\ |\mathcal{A}| = A$):

$$\mathbb{E}\left[R_{\mathsf{PE}}(\mathfrak{A}, K)\right] \geqslant \Omega\left(H^{3/2}\sqrt{SAK}\right)$$

Regardless of C.

Algorithm (AR-UCBVI)

We extend UCBVI to handle the action availabilities. AR-UCBVI is the **optimistic counterpart** of performing an **action-restricted value iteration**. This solution matches the LB (for every C):

$$\mathbb{E}\left[R_{\mathsf{PE}}(\texttt{AR-UCBVI}, K)\right] \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{SAK}\right)$$

Per-Stage Disclosure

Per-Stage Regret

Given algorithm $\mathfrak{A}$ running $(\pi_k)_{k=1}^K$, we evaluate its performances in terms of **per-stage (PS) regret** w.r.t. the **best constant policy** π^*:

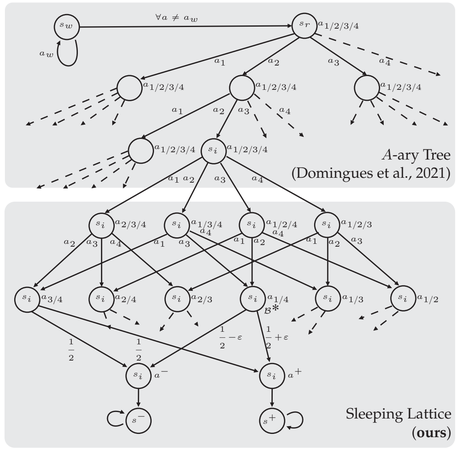
$$R_{\mathsf{PS}}(\mathfrak{A}, K) = K V_1^{\pi^*}(\overline{s}) - \sum_{k \in [\![K]\!]} V_1^{\pi_k}(\overline{s})$$

Markovian Availability ($C = C^{\mathsf{Markov}}$)

At every stage h, the **available actions** $\mathcal{A}_h(s_h)$ **depend** on:

- current state s_h
- previous state s_{h-1} and action a_{h-1}
- previous available actions $\mathcal{A}_{h-1}(s_{h-1})$

Lower Bound



The regret $\mathbb{E}\left[R_{\mathsf{PS}}(\mathfrak{A}, K)\right]$ that every algorithm suffer is:

$$\Omega\left(H^{3/2}\sqrt{SAK 2^{A/2}}\right)$$

There is a **statistical barrier** in the learning, as an **exponential dependence** on A is unavoidable.

How to Learn?

We can map the Sleeping MDP with an **Augmented MDP** in which we encode action availability sets in the state representation. This MDP will present an **augmented state space** $\widetilde{\mathcal{S}} := \mathcal{S} \times \mathcal{P}(\mathcal{A})$ with cardinality $S 2^A$.

Given this mapping, we can run UCBVI with Bernstein-Freedman bonus and get:

$$\mathbb{E}\left[R_{\mathsf{PS}}(\texttt{UCBVI}, K)\right] \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{SAK 2^A}\right)$$

Independent Availability ($C = C^{\mathsf{ind}}$)

At every stage h, the **available actions** $\mathcal{A}_h(s_h)$ **depend only** on **state** s_h

Lower Bound

$$\mathbb{E}\left[R_{\mathsf{PS}}(\mathfrak{A}, K)\right] \geqslant \Omega\left(H^{3/2}\sqrt{SAK}\right)$$

Algorithm (S-UCBVI)

S-UCBVI extends UCBVI to learn action availabilities $C_h^{\mathsf{ind}}(\cdot|s)$. It performs **optimism** on the estimates of both **state-action** $\widehat{Q}_{k,h}(s,a)$ and **state** $\widehat{V}_{k,h}(s)$ **value functions**:

$$\widehat{Q}_{k,h}(s,a) = R_h(s,a) + \sum_{s' \in \mathcal{S}} \widehat{P}_{k,h}(s'|s,a) \widehat{V}_{k,h+1}(s') + b_{k,h}^Q(s,a)$$

$$\widehat{V}_{k,h}(s) = \sum_{\mathcal{B} \in \mathcal{P}(\mathcal{A})} \widehat{C}_{k,h}^{\mathsf{ind}}(\mathcal{B}|s) \max_{a \in \mathcal{B}} \widehat{Q}_{k,h}(s,a) + b_{k,h}^V(s)$$

where $b_{k,h}^Q(s,a)$ and $b_{k,h}^V(s)$ are **Bernstein-Freedman** bonuses to account for the **uncertainty** on P and C. S-UCBVI enjoys a regret of:

$$\mathbb{E}\left[R_{\mathsf{PS}}(\texttt{S-UCBVI}, K)\right] \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{SAK}\right)$$

Experiments - Stochastic Frozen Lake

Frozen Lake with holes appearing during the episode with probability p