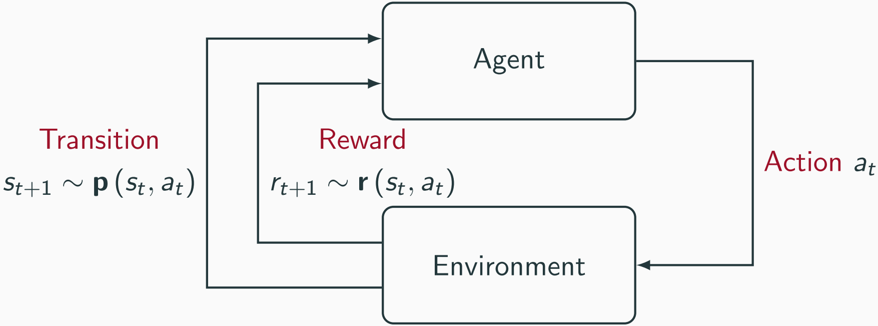
`IMED-RL`: Regret optimal learning of ergodic Markov decision processes

Fabien Pesquerel, Odalric-Ambrym Maillard

Outline - IMED-RL



We consider **Reinforcement Learning** in a discrete, **undiscounted**, infinite-horizon **Markov Decision Problem** (MDP) under the **average reward criterion**, and focus on the **maximization** of this criterion, when the learner does not know the rewards nor the transitions of the MDP.

Objective - Average reward criterion

The **cumulative reward** at time T, of policy π in MDP $\mathbf{M}$ is

$$V_{\pi,\mathbf{M}}(T) = \mathbb{E}_{\pi,\mathbf{M}}\left[\sum_{t=1}^{T} r_t\right].$$

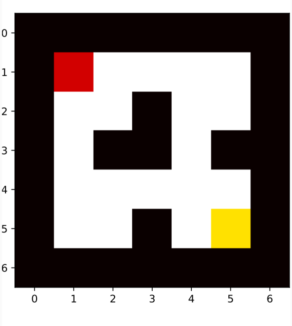
The **regret of policy** π at time T in MDP $\mathbf{M}$ is defined as

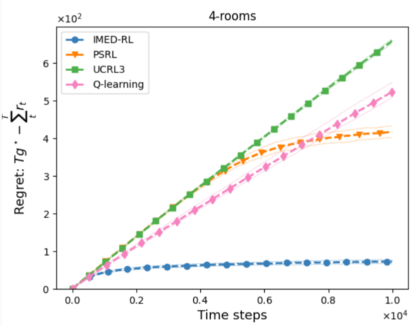
$$\mathcal{R}_\pi\left(\mathbf{M}, T\right) = \max_\eta\left(V_{\eta,\mathbf{M}}(T)\right) - V_{\pi,\mathbf{M}}(T)\,.$$

Objective: minimizing the regret in the long run, and thus maximizing the average-reward,

$$\lim_{T\to\infty} \frac{1}{T} V_{\pi,\mathbf{M}}(T).$$

We propose a **policy**, IMED-RL, that we prove to be **optimal** and show its impressive numerical performances.





Average **regret** *in the 4-rooms environment*

Instance dependent objective

In this work, one is interested in being **optimal** with respect to each **specific instance**. One must therefore assess the speed at which one can learn on each specific MDP. Hypothesis are therefore necessary to state the complexity of each instance.

Light-tail rewards and **Semi-bounded rewards** (support of the reward distribution is bounded from above)

Ergodicity The MDP is ergodic, $\forall s, s', \forall \pi, \exists t \in \mathbb{N} : \mathbf{p}_\pi^t(s'|s) > 0$.

Assess optimality

Thanks to the **ergodic assumption**, interesting quantities can be defined **locally**.

The **sub-optimality gap** $\Delta_{s,a}(\mathbf{M})$ in $\mathbf{M}$ is a measure of the **local regret** incurred by a policy that would play action a in state s.

The potential $\gamma_s(\mathbf{M})$ is a number used to assess optimality of actions in state s of MDP $\mathbf{M}$.

The **sub-optimality cost**, $\underline{\mathbf{K}}_{s,a}(\mathbf{M}) = \underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$, is a measure of the **local complexity** of distinguishing the sub-optimal action a from an optimal one in state s.

Assess optimality

Thanks to the **ergodic assumption**, interesting quantities can be defined **locally**.

The **sub-optimality gap** $\Delta_{s,a}(\mathbf{M})$ in $\mathbf{M}$ is a measure of the **local regret** incurred by a policy that would play action a in state s.

The potential $\gamma_s(\mathbf{M})$ is a number used to assess optimality of actions in state s of MDP $\mathbf{M}$.

The **sub-optimality cost**, $\underline{\mathbf{K}}_{s,a}(\mathbf{M}) = \underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$, is a measure of the **local complexity** of distinguishing the sub-optimal action a from an optimal one in state s.

Assess optimality

Thanks to the **ergodic assumption**, interesting quantities can be defined **locally**.

The **sub-optimality gap** $\Delta_{s,a}(\mathbf{M})$ in $\mathbf{M}$ is a measure of the **local regret** incurred by a policy that would play action a in state s.

The potential $\gamma_s(\mathbf{M})$ is a number used to assess optimality of actions in state s of MDP $\mathbf{M}$.

The **sub-optimality cost**, $\underline{\mathbf{K}}_{s,a}(\mathbf{M}) = \underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$, is a measure of the **local complexity** of distinguishing the sub-optimal action a from an optimal one in state s.

Assess optimality

Thanks to the **ergodic assumption**, interesting quantities can be defined **locally**.

The **sub-optimality gap** $\Delta_{s,a}(\mathbf{M})$ in $\mathbf{M}$ is a measure of the **local regret** incurred by a policy that would play action a in state s.

The potential $\gamma_s(\mathbf{M})$ is a number used to assess optimality of actions in state s of MDP $\mathbf{M}$.

The **sub-optimality cost**, $\underline{\mathbf{K}}_{s,a}(\mathbf{M}) = \underline{\mathbf{K}}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$, is a measure of the **local complexity** of distinguishing the sub-optimal action a from an optimal one in state s.

Regret decomposition

Under the ergodic assumption, the regret of any policy π can be decomposed as

$$\mathcal{R}_\pi(\mathbf{M}, T) = \sum_{s,a} \mathbb{E}_\pi\left[N_{s,a}(T)\right]\Delta_{s,a}(\mathbf{M}) + C,$$

where $N_{s,a}(T) = \sum_{t=1}^{T} 1\{s_t = s, a_t = a\}$ **counts** the number of time the state-action pair (s, a) has been sampled.

Regret bounds and optimality of `IMED-RL`

Theorem (Regret lower bound)

Let $\mathbf{M}$ *be an MDP satisfying hypothesis. For all policy* π*, the* **regret lower bound** *is*

$$\liminf_{T\to\infty} \frac{\mathcal{R}_\pi(\mathbf{M}, T)}{\log T} \geqslant \sum_{(s,a)\in\mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})}.$$

Theorem (Regret upper bound - Asymptotic Optimality)

`IMED-RL` is asymptotically optimal, that is,

$$\lim_{T\to+\infty} \frac{\mathcal{R}_{\texttt{IMED-RL}}(\mathbf{M}, T)}{\log T} \leqslant \sum_{(s,a)\in\mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})}.$$

IMED-RL - **Intuition**

IMED-RL is a *model-based* algorithm that keeps empirical estimates of the transitions **p** and rewards **r**.

While **policy iteration** constructs a **sequence of policies** that are increasingly better, IMED-RL constructs a **sequence of sub-MDPs** of the original MDP that are increasingly better with high probability.

A sub-MDP is better than another if its optimal gain is better. Sub-MDP are built by restricting the action space of the original MDP: the **skeleton** (sub-MDP) at time t, is defined by

$$\mathcal{A}_s(t) = \left\{ a \in \mathcal{A}_s \;:\; N_{s,a}(t) \geqslant \log^2 \left(\max_{a' \in \mathcal{A}_s} N_{s,a'}(t) \right) \right\}.$$

Indexed Minimum Empirical Divergence for RL

Algorithm 1: `IMED-RL`

Require State-Action space of an MDP with hypothesis

Initialisation State s_1

for $t \geqslant 1$ **do**

  Sample $a_t \in \arg\min_{a \in \mathcal{A}_{s_t}} \mathbf{H}_{s,a}(t)$

where $\mathbf{H}_{s,a}(t) = N_{s,a}(t)\underline{\mathbf{K}}_{s,a}\left(\widehat{\mathbf{M}}_t(\mathcal{A}(t)), \widehat{\gamma}_s(t)\right) + \log N_{s,a}(t)$.

Take-home message

IMED-RL is a **provably optimal** RL algorithm in the average-reward setting under the **ergodic** dynamic hypothesis.

Nonetheless, IMED-RL has **impressive numerical performances** beyond the ergodic case, in the communicating one.

This raises the question on how to adapt IMED-RL to handle the theoretically more **challenging framework of communicating MDPs**.

Thank you

Talk with us at poster 52874

Code available on github at **fabienpesquerel/IMED-RL**

Reach us at **fabien.pesquerel@inria.fr**

More research at

fabienpesquerel.github.io

odalricambrymmaillard.neowordpress.fr