Optimal regret minimization strategies in Markov Decision Processes.

Odalric-Ambrym Maillard

June 07, 2023

Inria Scool

Centre de Recherche en Informatique, Signal et Automatique de Lille

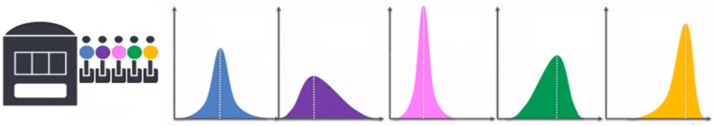
References

▷ Pesquerel, F. and Maillard, O.A., 2022, November. *IMED-RL: Regret optimal learning of ergodic Markov decision processes*. In NeurIPS 2022-Thirty-sixth Conference on Neural Information Processing Systems.

▷ Honda, J. and Takemura, A., 2015. *Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards*. J. Mach. Learn. Res., 16, pp.3721-3756.

▷ Agrawal, R., 1990, December. *Adaptive control of Markov chains under the weak accessibility*. In 29th IEEE Conference on Decision and Control (pp. 1426-1431). IEEE.

Model $\mathbf{M} = (\mathcal{A}, \mathbf{r})$ where $\mathcal{A} = \{1, \dots, 5\}$, $\mathbf{r} : \mathcal{A} \to \mathcal{P}([0,1])$.



Consider a score, e.g. the mean, $\mathbf{m} : \mathcal{A} \to [0,1]$

$$\text{Let } \mathbf{m}_\star = \max_{a \in \mathcal{A}} \mathbf{m}(a) \text{ and } \mathcal{O}(\mathbf{M}) = \operatorname*{Argmax}_{a \in \mathcal{A}} \mathbf{m}(a).$$

At each decision time t, choose $A_t \in \mathcal{A}$, receive $X_t \sim \mathbf{r}(A_t)$.

After a while

Form the trajectory $\tau_t = (A_1, X_1, \dots, A_t, X_t)$, and counts $N_t(a) = \sum_{t'=1}^{t} \mathbb{I}\{A_{t'} = a\}$.
Try to guess e.g. the best distribution.

Warming-up

Log-likelihood of model $\mathbf{M} = (\mathcal{A}, \mathbf{r})$ versus $\tilde{\mathbf{M}} = (\mathcal{A}, \tilde{\mathbf{r}})$ on τ_T:

$$\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} = \sum_{t=1}^{T} \log \frac{\mathbf{r}(A_t)(X_t)}{\tilde{\mathbf{r}}(A_t)(X_t)},$$

$$L_T(\mathbf{M}, \tilde{\mathbf{M}}) = \mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right] = \sum_{a} \mathbb{E}_{\mathbf{M}}\Big[N_T(a)\Big] \mathtt{KL}(\mathbf{r}(a), \tilde{\mathbf{r}}(a))\,.$$

Log-likelihood of model $\mathbf{M} = (\mathcal{A}, \mathbf{r})$ versus $\tilde{\mathbf{M}} = (\mathcal{A}, \tilde{\mathbf{r}})$ on τ_T:

$$\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} = \sum_{t=1}^{T} \log \frac{\mathbf{r}(A_t)(X_t)}{\tilde{\mathbf{r}}(A_t)(X_t)},$$

$$L_T(\mathbf{M}, \tilde{\mathbf{M}}) = \mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right] = \sum_{a} \mathbb{E}_{\mathbf{M}}\Big[N_T(a)\Big]\mathtt{KL}(\mathbf{r}(a), \tilde{\mathbf{r}}(a)).$$

Game

$\mathcal{B}(a, \mathbf{M})$: For some $a \notin \mathcal{O}(\mathbf{M})$, find $\tilde{\mathbf{M}} = \tilde{\mathbf{M}}_a$ such that

- $\forall a \in \mathcal{O}(\mathbf{M})$, $\mathbf{r}(a)$ and $\tilde{\mathbf{r}}(a)$ undistinguishable
- $a \in \mathcal{O}(\tilde{\mathbf{M}})$

Say $\mathcal{O}(\mathbf{M}) = \{5\}$, then $\tilde{\mathbf{r}}(5) = \mathbf{r}(5)$, $\tilde{\mathbf{r}}(a)$ has mean $\tilde{\mathbf{m}}(a) > \mathbf{m}(5)$.

Closest?

Warming-up

Log-likelihood of model $\mathbf{M} = (\mathcal{A}, \mathbf{r})$ versus $\tilde{\mathbf{M}} = (\mathcal{A}, \tilde{\mathbf{r}})$ on τ_T:

$$\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} = \sum_{t=1}^{T} \log \frac{\mathbf{r}(A_t)(X_t)}{\tilde{\mathbf{r}}(A_t)(X_t)},$$

$$L_T(\mathbf{M}, \tilde{\mathbf{M}}) = \mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right] = \sum_{a} \mathbb{E}_{\mathbf{M}}\Big[N_T(a)\Big] \mathtt{KL}(\mathbf{r}(a), \tilde{\mathbf{r}}(a))\,.$$

Game

$\mathcal{B}(a, \mathbf{M})$: For some $a \notin \mathcal{O}(\mathbf{M})$, find $\tilde{\mathbf{M}} = \tilde{\mathbf{M}}_a$ such that

- $\forall a \in \mathcal{O}(\mathbf{M})$, $\mathbf{r}(a)$ and $\tilde{\mathbf{r}}(a)$ undistinguishable
- $a \in \mathcal{O}(\tilde{\mathbf{M}})$

Say $\mathcal{O}(\mathbf{M}) = \{5\}$, then $\tilde{\mathbf{r}}(5) = \mathbf{r}(5)$, $\tilde{\mathbf{r}}(a)$ has mean $\tilde{\mathbf{m}}(a) > \mathbf{m}(5)$.

Closest?

$$\inf_{\tilde{M} \in \mathcal{B}(a,\mathbf{M})} L_T(\mathbf{M}, \tilde{\mathbf{M}}) = \mathbb{E}_{\mathbf{M}}\Big[N_T(a)\Big] \inf_{\tilde{r}} \{\mathtt{KL}(\mathbf{r}(a), \tilde{r}) : \tilde{r} \text{ has mean} > \mathbf{m}_\star\}$$

$$= \mathbb{E}_{\mathbf{M}}[N_T(a)]\, \underline{\mathbf{K}}(\mathbf{r}(a), \mathbf{m}_\star)$$

Warming-up: Consequences

How much does it cost to fool you and deviate from optimality?

Fooling cost The larger $\inf_{\tilde{M}\in\mathcal{B}(a,\mathbf{M})} L(\mathbf{M},\tilde{\mathbf{M}})$ the most difficult to make arm a look optimal while it isn't.

Lower bound We consider algorithms that are uniformly "good" on all $\mathbf{M}\in\mathfrak{M}$. In particular on $\mathbf{M}$, and $\tilde{\mathbf{M}}_a \in \mathcal{B}(a,\mathbf{M})$ for each a.

Control achievable "performances".

Algorithm

$$\text{Pick MLE } \widehat{\mathbf{M}}_t \text{, choose to sample } A_t \in \mathcal{O}(\widehat{\mathbf{M}}_t)$$

vs

$$\text{Pick MLE } \widehat{\mathbf{M}}_t \text{, choose to ``track'' } \inf_{\tilde{M} \in \mathcal{B}(a, \widehat{\mathbf{M}}_t)} L(\widehat{\mathbf{M}}_t, \tilde{\mathbf{M}}).$$

Information Minimal Empirical Distribution

Algorithm

$$\text{Pick MLE } \widehat{\mathbf{M}}_t, \text{ choose to sample } A_t \in \mathcal{O}(\widehat{\mathbf{M}}_t)$$

vs

$$\text{Pick MLE } \widehat{\mathbf{M}}_t, \text{ choose to ``track'' } \inf_{\tilde{M}\in\mathcal{B}(a,\widehat{\mathbf{M}}_t)} L(\widehat{\mathbf{M}}_t, \tilde{\mathbf{M}}).$$

Ensure $\log(\frac{N_t(a)}{t}) \simeq - \inf_{\tilde{M}\in\mathcal{B}(a,\widehat{\mathbf{M}}_t)} L(\widehat{\mathbf{M}}_t, \tilde{\mathbf{M}})$ by choosing:

$$\begin{aligned} A_t &\in \mathop{\mathbf{argmin}}_{a\in\mathcal{A}} \inf_{\tilde{M}\in\mathcal{B}(a,\widehat{\mathbf{M}}_t)} L(\widehat{\mathbf{M}}_t, \tilde{\mathbf{M}}) + \log(\frac{N_t(a)}{t})\,. \\ &= \mathop{\mathbf{argmin}}_{a\in\mathcal{A}} N_T(a)\underline{\mathbf{K}}(\widehat{\mathbf{r}}_t(a), \widehat{\mathbf{m}}_{\star,t}) + \log(\frac{N_t(a)}{t})\,. \end{aligned}$$

▷ Provably optimal strategy for "regret minimization"

Table of contents

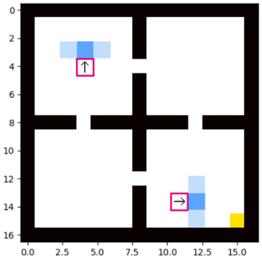
Markov Decision Processes

$$\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}_1, \mathbf{p}, \mathbf{r})$$

$\mathcal{S}$: set of states,
$\mathcal{A} = (\mathcal{A}_s)_{s\in\mathcal{S}}$: set of actions available in each state.
$\mathcal{X} = \{(s, a) : s \in \mathcal{A}, a \in \mathcal{A}_s\}$: set of pairs.
$\mathbf{p}_1 \in \mathcal{P}(\mathcal{S})$: initial state distribution,
$\mathbf{p} : \mathcal{X} \to \mathcal{P}(\mathcal{S})$: transition distribution function.
$\mathbf{r} : \mathcal{X} \to \mathcal{P}(\mathbb{R})$: reward distribution function, with mean reward function $\mathbf{m} : \mathcal{X} \to \mathbb{R}$

In RL: $\mathbf{p}, \mathbf{r}$ are **unknown**.

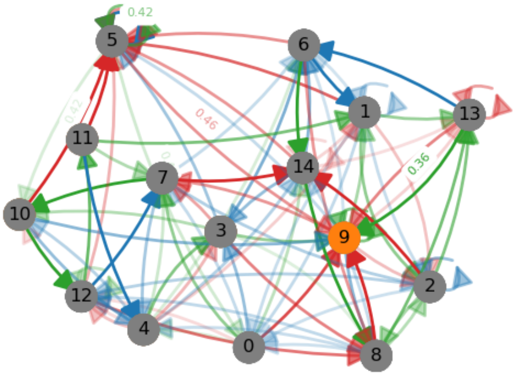
Example: Grid-world



Actions: {←, ↓, →, ↑}

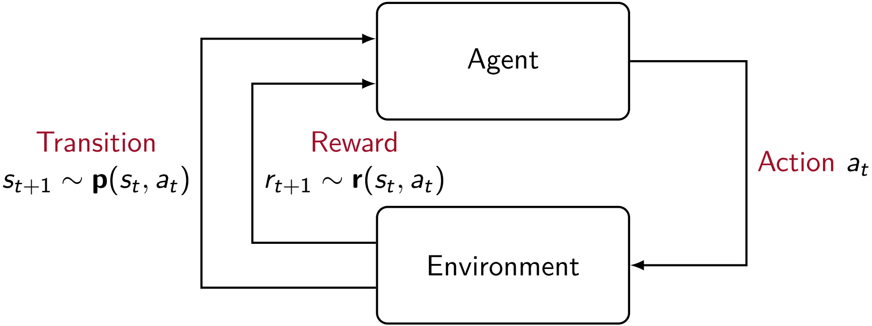
Rewards: 1 when in goal state, 0 otherwise.

Transitions: "Frozen-lake". Reset to random initial state after reaching goal (yellow)



Actions: Colors
Transitions: Shaded arrows
Rewards: sparse.

Reinforcement Learning



Discrete time (no autonomous dynamic)

Goal: Maximize "cumulative return".

Policies

Each stationary policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ acting in $\mathbf{M}$ induces a Markov chain on $\mathcal{S}$.

$$\begin{aligned}
&\text{(Policy mean)} & \mathbf{m}_\pi(s) &= \sum_{a \in \mathcal{A}_s} \mathbf{m}(s,a)\pi(a|s) \\
&\text{(Policy transition)} & \mathbf{p}_\pi(s'|s) &= \sum_{a \in \mathcal{A}_s} \mathbf{p}(s'|s,a)\pi(a|s)
\end{aligned}$$

For a sequence of policies $\boldsymbol{\pi} = (\pi_t)_t$, its value up to time T is

$$\mathbf{V}_{\boldsymbol{\pi},T}(s_1) = \mathbb{E}\Big[\sum_{t=1}^T r_t\Big] = \sum_{t=1}^T \Big(\prod_{t'=1}^{t-1} \mathbf{p}_{\pi_{t'}} \mathbf{m}_{\pi_t}\Big)(s_1)\,.$$

Optimality gap and Regret

Learning **with episodes** (of length H):

- Learner wants to find a policy with **maximal value** $\mathbf{V}^\star_H(s) = \max_\pi \mathbf{V}_{\pi,H}(s)$.
- At episode k, agent builds policy π_k. We measure the **gap to optimality**:

$$\Delta(\pi_k) = \mathbf{V}^\star_H(s_1) - \mathbf{V}_{\pi_k,H}(s_1)$$

- As $k \to \infty$, we expect the gap to vanish, that is $\Delta(\pi_k) \to 0$.

vs

Optimality gap and Regret

Learning **with episodes** (of length H):

- Learner wants to find a policy with **maximal value** $\mathbf{V}^\star_H(s) = \max_\pi \mathbf{V}_{\pi,H}(s)$.
- At episode k, agent builds policy π_k. We measure the **gap to optimality**:

$$\Delta(\pi_k) = \mathbf{V}^\star_H(s_1) - \mathbf{V}_{\pi_k,H}(s_1)$$

- As $k \to \infty$, we expect the gap to vanish, that is $\Delta(\pi_k) \to 0$.

vs

Learning **within one episode**: (this talk!)

- Learner wants to get **maximal reward** and maximize its **own** value.
- At decision step T, we measure the **regret** of the agent choosing $\boldsymbol{\pi} = (\pi_t)_t$, accumulated **while learning**, compared to an optimal policy $\star$ with

$$\mathfrak{R}_T(\boldsymbol{\pi}, \mathbf{M}) = \mathbf{V}_{\star,T}(s_1) - \mathbf{V}_{\pi,T}(s_1)$$

- As $T \to \infty$, the regret cannot decreases, we still expect $\mathfrak{R}_T(\boldsymbol{\pi}, \mathbf{M})/T \to 0$.

Policies

We are interested in policies that maximize rewards in the long run,

$$\lim_{T} \frac{1}{T} \mathbf{V}_{\pi,T}(s_1)$$

For stationary policy π, this is $\left(\lim_{T} \frac{1}{T} \sum_{t=1}^{T} \mathbf{p}_\pi^{t-1} \mathbf{m}_\pi\right)(s_1)$, hence we define

$$\begin{aligned}
&\text{(Average transition)} \quad \overline{\mathbf{p}}_\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{p}_\pi^{t-1}\,,\\
&\text{(Average frequency)} \quad \mathbf{f}_{\pi,s_1}(s,a) = \overline{\mathbf{p}}_\pi(s|s_1)\pi(a|s)
\end{aligned}$$

The **average value** (aka gain) of a policy π starting in state s_1:

$$\mathbf{g}_\pi(s_1) = (\overline{\mathbf{p}}_\pi \mathbf{m}_\pi)(s_1) = \langle \mathbf{f}_{\pi,s_1}, \mathbf{m}_\pi \rangle.$$

The **average value** (aka gain) of a policy π starting in state s_0:

$$\mathbf{g}_\pi(s_1) = (\overline{\mathbf{p}}_\pi \mathbf{m}_\pi)(s_1) = \langle \mathbf{f}_{\pi,s_1}, \mathbf{m}_\pi \rangle .$$

The **bias function** is given by $\mathbf{b}_\pi(s) = \Big(\sum_{t=1}^{\infty} (\mathbf{p}_\pi^{t-1} - \overline{\mathbf{p}}_\pi)\mathbf{m}_\pi\Big)(s)$.

Fixed point property:

$$\textit{(Poisson equation)} \qquad \mathbf{b}_\pi(s) \;=\; \mathbf{m}_\pi(s) - \mathbf{g}_\pi(s) + (\mathbf{p}_\pi \mathbf{b}_\pi)(s)\,.$$

(Span $\mathbb{S}(f) = \max_x f(x) - \min_x f(x)$ of function $f : \mathcal{X} \to \mathbb{R}$. The span operator induces a semi-norm on functions.)

MDP classification

An MDP M is **communicating** if every pair of state is communicating under some policy, that is

$$\forall s, s' \in \mathcal{S}, \exists \pi, \exists t < \infty : \mathbf{p}_\pi^t(s'|s) > 0.$$

An MDP M is **irreducible** if every policy π on M induces an irreducible Markov chain, that is

$$\forall \pi, \forall s, s' \in \mathcal{S}, \exists t < \infty : \mathbf{p}_\pi^t(s'|s) > 0.$$

An MDP M is **ergodic**[1] if every policy π on M induces an ergodic Markov chain, that is $\forall \pi$,

$$\begin{aligned}
&\text{(irreducibility)} && \forall s, s' \in \mathcal{S}, \exists t < \infty,\ \mathbf{p}_\pi^t(s'|s) > 0,\\
&\text{(aperiodicity)} && \forall s \in \mathcal{S},\ \mathrm{GCD}(\{t : \mathbf{p}_\pi^t(s|s) > 0\}) = 1,\\
&\text{(positive recurrence)} && \forall s \in \mathcal{S},\ \mathbb{E}(\min\{t > 1 : s_t = s\} | s_1 = s]) < \infty\,.
\end{aligned}$$

[1] In the MDP learning literature it is common to use "ergodic" terminology in lieu of "irreducible".

Optimal policies and regret

- Gain optimal policies:

$$\overline{\mathcal{O}}(\mathbf{M}, s_1) \quad = \quad \left\{\pi : \mathbf{g}_\pi(s_1) \geqslant \mathbf{g}_{\pi'}(s_1) \forall \pi'\right\},$$

- In a communicating MDP with finite $\mathcal{S}$ and $\mathcal{A}$, $\bigcap\limits_{s_1 \in \mathcal{S}} \overline{\mathcal{O}}(\mathbf{M}, s_1) \neq \emptyset$ and contains a **unichain** policy.
- The **cumulative regret** up to time $T \in \mathbb{N}$ of an algorithm executing policy $\boldsymbol{\pi} = (\pi_t)_t$ (that is, it plays policy π_t at time $t \in \{1, \dots, T\}$) in MDP $\mathbf{M}$ starting from initial state s_1, against optimal policies for the different optimality criteria are given by

$$\overline{\mathfrak{R}}_{T,s_1}(\boldsymbol{\pi}; \mathbf{M}) \quad = \quad \mathbf{V}_{\star,T}(s_1) - \mathbf{V}_{\pi,T}(s_1) \text{ where } \star \in \overline{\mathcal{O}}(\mathbf{M}, s_1)$$

Regret decomposition

The **cumulative regret** up to time T for the average gain and discounted criterion satisfy the following decomposition provided that an optimal policy $\star$ is **unichain**,

$$\overline{\mathfrak{R}}_{T,s_1}(\pi;\mathbf{M}) \quad = \quad \sum_{x\in\mathcal{X}} \mathbb{E}_{\pi,s_1}[N_T(x)]\Delta(x) + \underbrace{\Big(\Big[\prod_{t'=1}^{T}\mathbf{p}_{\pi_{t'}} - \mathbf{p}_\star^T\Big]\mathbf{b}_\star\Big)(s_1)}_{\leqslant \mathbb{S}(\mathbf{b}_\star)},$$

where we introduced the **sub-optimality gap**

$$\Delta(s,a) = \mathbf{m}_\star(s) - \mathbf{m}(s,a) + ((\mathbf{p}_\star - \mathbf{p}_a)\mathbf{b}_\star)(s).$$

Minimizing $\Delta(s,a)$ is equivalent to maximizing $\varphi_{\mathbf{M}}(s,a) = \mathbf{m}(s,a) + (\mathbf{p}_a\mathbf{b}_\star)(s)$.

Regret decomposition

The **cumulative regret** up to time T for the average gain and discounted criterion satisfy the following decomposition provided that an optimal policy $\star$ is **unichain**,

$$\overline{\mathfrak{R}}_{T,s_1}(\pi;\mathbf{M}) \quad = \quad \sum_{x\in\mathcal{X}} \mathbb{E}_{\pi,s_1}[N_T(x)]\Delta(x) + \underbrace{\Big(\Big[\prod_{t'=1}^{T}\mathbf{p}_{\pi_{t'}} - \mathbf{p}_\star^T\Big]\mathbf{b}_\star\Big)(s_1)}_{\leqslant \mathbb{S}(\mathbf{b}_\star)}\,,$$

where we introduced the **sub-optimality gap**

$$\Delta(s,a) = \mathbf{m}_\star(s) - \mathbf{m}(s,a) + ((\mathbf{p}_\star - \mathbf{p}_a)\mathbf{b}_\star)(s).$$

Minimizing $\Delta(s,a)$ is equivalent to maximizing $\varphi_{\mathbf{M}}(s,a) = \mathbf{m}(s,a) + (\mathbf{p}_a\mathbf{b}_\star)(s)$.

- In Bandits $|\mathcal{S}| = 1$, $\Delta(a) = \mathbf{m}_\star - \mathbf{m}(a)$ and

$$\overline{\mathfrak{R}}_T(\pi,\mathbf{M}) \quad = \quad \sum_{a\in\mathcal{A}} \mathbb{E}_\pi[N_T(a)]\Delta(a)\,.$$

Table of contents

Assumptions

Regular model $\mathfrak{M}$

For each $x \in \mathcal{X}$, $\mathbf{r}(x) \in \mathcal{D}_x \subset \mathcal{P}(\mathbb{R})$ where $\mathcal{D}_x$ is known to the learner, $\mathcal{D} = \otimes_x \mathcal{D}_x$.

Light-tail rewards For all $x \in \mathcal{X}$, the moment generating function of reward $\mathbf{r}(x)$ exists in a neighborhood of 0.

Semi-bounded rewards There exists a known quantity $m_{\max}(x) \in \mathbb{R}$ such that $\mathrm{Supp}(\mathbf{r}(x)) \subset (-\infty, m_{\max}(x)]$ and $\mathbf{m}(x) < m_{\max}(x)$.

Good learner

Definition (Uniformly Good strategies)

A agent is **uniformly-good** on $\mathcal{D}$ if

$$\forall \mathbf{M} = (\mathbf{r}_a)_{a \in \mathcal{A}} \in \mathcal{D}, \forall a \notin \mathcal{O}(\mathbf{M}), \quad \mathbb{E}[N_T(a)] = o(T^\alpha) \quad \text{for all } \alpha \in (0,1].$$

Regret lower bounds for bandits

Theorem ("Price for being uniformly-good")

Any uniformly good strategy on $\mathcal{D} = \mathrm{Bern}^{\mathcal{A}}$ must satisfy (Lai & Robbins, 85)

$$\forall a \notin \mathcal{O}(\mathbf{M}) \quad \liminf_{T\to\infty} \frac{\mathbb{E}_{\mathbf{M}}[N_T(a)]}{\ln(T)} \geqslant \frac{1}{\mathtt{kl}(\mathbf{m}_a, \mathbf{m}_\star)} .$$

More generally (Burnetas & Katehakis, 96) for $\mathbf{r} \in \otimes_{a\in\mathcal{A}}\mathcal{D}_a$

$$\forall a \notin \mathcal{O}(\mathbf{M}) \quad \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \geqslant \frac{1}{\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star)},$$

$$\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star) = \inf\{\mathtt{KL}(\mathbf{r}_a, \tilde{r}) : \tilde{r} \in \mathcal{D}_a, \mathbb{E}_{\tilde{r}}[X] > \mathbf{m}_\star\}$$

Bandit complexity

For any **uniformly-good** strategy $\boldsymbol{\pi}$ on $\mathfrak{M}$ and $\mathbf{M} \in \mathfrak{M}$,

$$\liminf_{T\to\infty} \frac{\overline{\mathfrak{R}}_T(\boldsymbol{\pi}, \mathbf{M})}{\ln T} \geqslant \sum_{a\in\mathcal{A}} \frac{\Delta_a}{\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star)}$$

where

$$\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star) = \inf\{\mathtt{KL}(\mathbf{r}_a, \tilde{r}) : \tilde{r} \in \mathcal{D}_a, \mathbb{E}_{\tilde{r}}[X] > \mathbf{m}_\star\}$$

Bandit complexity

For any **uniformly-good** strategy $\boldsymbol{\pi}$ on $\mathfrak{M}$ and $\mathbf{M} \in \mathfrak{M}$,

$$\liminf_{T\to\infty} \frac{\overline{\mathfrak{R}}_T(\boldsymbol{\pi}, \mathbf{M})}{\ln T} \geqslant \sum_{a\in\mathcal{A}} \frac{\Delta_a}{\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star)}$$

where

$$\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star) = \inf\{\mathtt{KL}(\mathbf{r}_a, \tilde{r}) : \tilde{r} \in \mathcal{D}_a, \mathbb{E}_{\tilde{r}}[X] > \mathbf{m}_\star\}$$

▷ IMED strategy $\boldsymbol{\pi}_I$ is optimal in the sense that for each $\mathbf{M} \in \mathfrak{M}$ (with regular $\mathfrak{M}$)

$$\limsup_{T\to\infty} \frac{\overline{\mathfrak{R}}_T(\boldsymbol{\pi}_I, \mathbf{M})}{\ln T} \leqslant \sum_{a\in\mathcal{A}} \frac{\Delta_a}{\underline{\mathbf{K}}_a(\mathbf{r}_a, \mathbf{m}_\star)}$$

- For any **test function** h bounded in $[0, 1]$,

$$\mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right] \geqslant \mathtt{kl}\left(\mathbb{E}_{\mathbf{M}}[h(\tau_T)], \mathbb{E}_{\tilde{\mathbf{M}}}[h(\tau_T)]\right),$$

where $\mathtt{kl}(x, y) = x \log(x/y) + (1-x)\log((1-x)/(1-y))$.

Fundamental inequality

▷ For any **test function** h bounded in $[0,1]$,

$$\mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right] \geqslant \mathtt{kl}\left(\mathbb{E}_{\mathbf{M}}[h(\tau_T)], \mathbb{E}_{\tilde{\mathbf{M}}}[h(\tau_T)]\right),$$

where $\mathtt{kl}(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$.

▷ For $h(\tau_T) = 1 - N_T(a)/T$, using **uniformly-good** assumption, we can show

$$\mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}_a(\tau_T)}\right] \quad \geqslant \quad (1-\alpha)\log(T) + o(\log(T)) \ \text{ for all } \alpha,$$

from which we get for each $a \notin \mathcal{O}(\mathbf{M})$

$$\liminf_T \frac{\mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}_a(\tau_T)}\right]}{\log(T)} \geqslant 1 \text{ i.e. } \liminf_T \frac{\mathbb{E}_{\mathbf{M}}[N_T(a)]}{\log(T)} \geqslant \frac{1}{\underline{\mathbf{K}}(\mathbf{r}(a), \mathbf{m}_\star)}.$$

MDP Likelihood ratios

For a given trajectory $\tau_T = (s_1, a_1, r_1, s_2, \dots, s_{T+1})$ and two MDPs $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}_1, \mathbf{p}, \mathbf{r}), \tilde{\mathbf{M}} = (\mathcal{S}, \mathcal{A}, \tilde{\mathbf{p}}_1, \tilde{\mathbf{p}}, \tilde{\mathbf{r}})$:

$$\begin{aligned}
\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} &= \log \frac{p_1(s_1)}{\tilde{p}_1(s_1)} + \sum_{t=1}^{T} \log \frac{p(s_t, a_t)(s_{t+1})}{\tilde{p}(s_t, a_t)(s_{t+1})} + \log \frac{r(s_t, a_t)(r_t)}{\tilde{r}(s_t, a_t)(r_t)} \\
\mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right] &= \mathtt{KL}(\mathbf{p}_1, \tilde{\mathbf{p}}_1) + \sum_x \mathbb{E}_{\mathbf{M}}\Big[N_T(x)\Big]\Big[\mathtt{KL}(\mathbf{p}(x), \tilde{\mathbf{p}}(x)) + \mathtt{KL}(\mathbf{r}(x), \tilde{\mathbf{r}}(x))\Big]
\end{aligned}$$

▷ For a policy π generating τ_T from starting state s_1 (i.e. $\mathbf{p}_1 = \tilde{\mathbf{p}}_1 = \delta_{s_1}$), we let

$$L_T(\mathbf{M}, \tilde{\mathbf{M}}) = \sum_x \mathbb{E}_{\mathbf{M},\pi,s_1}\Big[N_T(x)\Big]\mathtt{KL}_x(\mathbf{M}, \tilde{\mathbf{M}}) .$$

where $\mathtt{KL}_x(\mathbf{M}, \tilde{\mathbf{M}}) = \mathtt{KL}(\mathbf{p}(x), \tilde{\mathbf{p}}(x)) + \mathtt{KL}(\mathbf{r}(x), \tilde{\mathbf{r}}(x)) .$

▷ **Regret decomposition**

$$\overline{\mathfrak{R}}_{T,s_1}(\boldsymbol{\pi};\mathbf{M}) = \sum_{x\in\mathcal{X}} \mathbb{E}_{\boldsymbol{\pi}}[N_T(x)]\Delta(x) + \text{constant}$$

with $\Delta(s,a) = \varphi_{\mathbf{M}}(s,\star(s)) - \varphi_{\mathbf{M}}(s,a)$.

▷ **Optimal policies**

$$\overline{\mathcal{O}}(\mathbf{M},s_1) = \left\{\pi : \mathbf{g}_\pi(s_1) \geqslant \mathbf{g}_{\pi'}(s_1)\forall\pi'\right\},$$

▷ For each π deviating from $\overline{\mathcal{O}}(\mathbf{M},s_1)$, find **confusing** $\tilde{\mathbf{M}}_\pi$ that minimizes

$$L_T(\mathbf{M},\tilde{\mathbf{M}}) = \sum_{x\in\mathcal{X}} \mathbb{E}_{\mathbf{M}}\Big[N_T(x)\Big]\mathrm{KL}_x(\mathbf{M},\tilde{\mathbf{M}}).$$

Good strategies and beneficial MDPs

Uniformly good strategies An algorithm is said uniformly η-good on $\mathfrak{M}$, where $\eta : \mathbb{N} \to \mathbb{R}^+$ that satisfies $\lim_{t\to\infty} \eta(t) = 0$ specifies a rate function, if for all T, for all $\mathbf{M} \in \mathfrak{M}$, for all $x \in \mathcal{X}$ such that $\Delta_{\mathbf{M}}(x) > 0$, then

$$\mathbb{E}_{\mathbf{M},\pi,s_1}[N_T(x)/T] \leqslant \eta(T)\,.$$

(e.g. $\eta(T) = O(T^{\alpha-1})$)

Good strategies and beneficial MDPs

Uniformly good strategies An algorithm is said uniformly η-good on $\mathfrak{M}$, where $\eta : \mathbb{N} \to \mathbb{R}^+$ that satisfies $\lim_{t\to\infty} \eta(t) = 0$ specifies a rate function, if for all T, for all $\mathbf{M} \in \mathfrak{M}$, for all $x \in \mathcal{X}$ such that $\Delta_{\mathbf{M}}(x) > 0$, then

$$\mathbb{E}_{\mathbf{M},\pi,s_1}[N_T(x)/T] \leqslant \eta(T)] \,.$$

(e.g. $\eta(T) = O(T^{\alpha-1})$)

Beneficial MDPs for policy π, are $\tilde{\mathbf{M}} \in \overline{\mathcal{B}}_{s_1}(\pi, \mathbf{M}; \mathfrak{M})$ such that

- $\forall \star \in \mathcal{O}(\mathbf{M})$, $(\mathbf{r}_\star, \mathbf{p}_\star)$ and $(\tilde{\mathbf{r}}_\star, \tilde{\mathbf{p}}_\star)$ are undistinguishable.
- $\forall \star \in \mathcal{O}(\mathbf{M}), \tilde{\mathbf{g}}_\pi(s_1) > \tilde{\mathbf{g}}_\star(s_1)$

Note: So $\tilde{\mathbf{g}}_\star = \mathbf{g}_\star$ and $\varphi_{\tilde{\mathbf{M}}}(s, \star(s)) = \varphi_{\mathbf{M}}(s, \star(s)) = \varphi^\star_{\mathbf{M}}(s)$

Confusing instance for Ergodic MDPs

$$\inf_{\tilde{M}\in\overline{\mathcal{B}}_{s_1}(\pi,\mathbf{M})} \sum_{x\in\mathcal{X}} \mathbb{E}_{\mathbf{M}}\Big[N_T(x)\Big]\mathtt{KL}_x(\mathbf{M},\tilde{\mathbf{M}})$$

▷ **Perturbation of optimal policy** For $\star \in \mathcal{O}(\mathbf{M})$, consider $\pi = \star_{s,a}$ that plays action a in state s and otherwise $\star$.

▷ In ergodic MDPs, all states are recurrent under **each policy**.

In particular $\mathcal{X}_\star \Delta \mathcal{X}_{\star_{s,a}} = \{(s,a)\}$, where $\mathcal{X}_\pi = \{x : \mathbf{f}_{\pi,s_1}(x) > 0\}$.

Confusing instance for Ergodic MDPs

$$\inf_{\tilde{M}\in\overline{\mathcal{B}}_{s_1}(\pi,\mathbf{M})} \sum_{x\in\mathcal{X}} \mathbb{E}_{\mathbf{M}}\Big[N_T(x)\Big]\mathtt{KL}_x(\mathbf{M},\tilde{\mathbf{M}})$$

▷ **Perturbation of optimal policy** For $\star \in \mathcal{O}(\mathbf{M})$, consider $\pi = \star_{s,a}$ that plays action a in state s and otherwise $\star$.

▷ In ergodic MDPs, all states are recurrent under **each policy**.

In particular $\mathcal{X}_\star \Delta \mathcal{X}_{\star_{s,a}} = \{(s,a)\}$, where $\mathcal{X}_\pi = \{x : \mathbf{f}_{\pi,s_1}(x) > 0\}$.

▷ Also, since $\star$ and π only differ in (s,a), $\tilde{\mathbf{g}}_{\star_{s,a}}(s_1) > \tilde{\mathbf{g}}_\star(s_1)$ iff $\varphi_{\tilde{\mathbf{M}}}(s,a) > \varphi_{\tilde{\mathbf{M}}}(s,\star(s)) = \varphi^\star_{\mathbf{M}}(s)$, that is $\tilde{\mathbf{m}}(s,a) + (\tilde{\mathbf{p}}_a\mathbf{b}_\star)(s) > \varphi^\star_{\mathbf{M}}(s)$.

Confusing instance for Ergodic MDPs

$$\inf_{\tilde{M}\in\overline{\mathcal{B}}_{s_1}(\pi,\mathbf{M})} \sum_{x\in\mathcal{X}} \mathbb{E}_{\mathbf{M}}\Big[N_T(x)\Big]\mathtt{KL}_x(\mathbf{M},\tilde{\mathbf{M}})$$

▷ **Perturbation of optimal policy** For $\star\in\mathcal{O}(\mathbf{M})$, consider $\pi=\star_{s,a}$ that plays action a in state s and otherwise $\star$.

▷ In ergodic MDPs, all states are recurrent under **each policy**.

In particular $\mathcal{X}_\star\Delta\mathcal{X}_{\star_{s,a}}=\{(s,a)\}$, where $\mathcal{X}_\pi=\{x:\mathbf{f}_{\pi,s_1}(x)>0\}$.

▷ Also, since $\star$ and π only differ in (s,a), $\tilde{\mathbf{g}}_{\star_{s,a}}(s_1)>\tilde{\mathbf{g}}_\star(s_1)$ iff $\varphi_{\tilde{\mathbf{M}}}(s,a)>\varphi_{\tilde{\mathbf{M}}}(s,\star(s))=\varphi^\star_{\mathbf{M}}(s)$, that is $\tilde{\mathbf{m}}(s,a)+(\tilde{\mathbf{p}}_a\mathbf{b}_\star)(s)>\varphi^\star_{\mathbf{M}}(s)$.

Ergodic simplification

For each (s,a), the confusing cost becomes

$$\mathbb{E}_{\mathbf{M}}\Big[N_T(s,a)\Big]\underline{\mathbf{K}}_{s,a}(\mathbf{M},\varphi^\star_{\mathbf{M}}(s))$$

$$\begin{aligned}\text{where } \underline{\mathbf{K}}_x(\mathbf{M},\varphi) &= \inf\left\{\mathtt{KL}(\mathbf{r}(x),\tilde{\mathbf{r}}(x))+\mathtt{KL}(\mathbf{p}(x),\tilde{\mathbf{p}}(x)) : \varphi_{\tilde{\mathbf{M}}}(x)>\varphi\right\}\\ &= \inf\left\{\mathtt{KL}(\mathbf{r}(x)\otimes\mathbf{p}(x),\tilde{r}\otimes\tilde{p}) : \tilde{m}+\tilde{q}\mathbf{b}_\star>\varphi\right\}\end{aligned}$$

For all η-uniformly-good strategy with $\eta(T) = O(T^{\alpha-1})$ for all $\alpha \in (0,1)$,

$$\begin{aligned}\liminf_{T} \frac{\overline{\mathfrak{R}}_{T,s_1}(\pi,\mathbf{M})}{\ln(T)} &\geqslant \inf_{\kappa\in\mathbb{R}_+^{|\mathcal{X}|}} \left\{ \sum_{x\in\mathcal{X}} \kappa_x \Delta(x) : \forall(s,a), \kappa_{s,a}\underline{\mathbf{K}}_{s,a}(\mathbf{M},\varphi^\star_{\mathbf{M}}(s)) \geqslant 1 \right\}, \\ &\geqslant \sum_{(s,a)\in\mathcal{X}} \frac{\Delta(s,a)}{\underline{\mathbf{K}}_{s,a}(\mathbf{M},\varphi^\star_{\mathbf{M}}(s))}\end{aligned}$$

Table of contents

Hypothesis - Ergodic dynamic & light-tail reward

Light-tail rewards For all $x \in \mathcal{X}$, the moment generating function of reward $\mathbf{r}(x)$ exists in a neighborhood of 0.

Semi-bounded rewards For all $x \in \mathcal{X}$, $\mathbf{r}(x)$ belongs to a subset $\mathcal{F}_x \subset \mathcal{P}(\mathbb{R})$ known to the learner.
There exists a known quantity $m_{\max}(x) \in \mathbb{R}$ such that $\mathrm{Supp}(\mathbf{r}(x)) \subset (-\infty, m_{\max}(x))$ and its mean satisfies $\mathbf{m}(x) < m_{\max}(x)$.

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

Intuition

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

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.

Empirical MDP $\widehat{\mathbf{r}}(s,a)$ and $\widehat{\mathbf{p}}_t(s,a)$ the empirical reward and transition distributions after t time steps, *i.e.* using $N_{s,a}(t)$ samples from the distributions. $\widehat{\mathbf{M}}_t$ is the empirical MDP built from those estimations.

Skeleton The skeleton at time t, is defined by

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

It is **homogeneous** in the sense that it does not depends explicitly on t nor $N_s(t)$, only $N_{s,a}(t)$ and $\max_{a'} N_{s,a'}(t)$.

Restricted MDP $\widehat{\mathbf{M}}_{|t} = \widehat{\mathbf{M}}_t(\mathcal{A}(t))$ is the empirical MDP whose action space is restricted to the skeleton.

A value iteration scheme (or Policy iteration scheme) is applied to $\widehat{\mathbf{M}}_{|t}$, to find a policy $\widehat{\star}_t$.

▷ This enables to build **estimate** $\varphi_{\widehat{\mathbf{M}}_{|t}}$ of: $\varphi_{\mathbf{M}}(s,a) = \mathbf{m}(s,a) + (\mathbf{p}_a \mathbf{b}_\star)(s)$.

Using $\widehat{\star}_t$, $\widehat{\mathbf{m}}_t$, $\widehat{\mathbf{p}}_t$, $\widehat{\mathbf{b}}_{\widehat{\star}_t}$.

Value/Policy iteration

A value iteration scheme (or Policy iteration scheme) is applied to $\widehat{\mathbf{M}}_{|t}$, to find a policy $\widehat{\star}_t$.

▷ This enables to build **estimate** $\varphi_{\widehat{\mathbf{M}}_{|t}}$ of: $\varphi_{\mathbf{M}}(s,a) = \mathbf{m}(s,a) + (\mathbf{p}_a \mathbf{b}_\star)(s)$.

Using $\widehat{\star}_t$, $\widehat{\mathbf{m}}_t$, $\widehat{\mathbf{p}}_t$, $\widehat{\mathbf{b}}_{\widehat{\star}_t}$.

▷ By **ergodicity**, the pairs in the skeleton have enough visits ($\simeq \log^2(t)$), so well-estimated.

▷ In **ergodic** MDPs, for each $\pi \notin \mathcal{O}(\mathbf{M})$, there exists a policy improving over π at Hamming distance 1 of π.

The the IMED-RL index of (s,a) at time t, $\mathbf{H}_{s,a}(t)$, is defined as

$$\mathbf{H}_{s,a}(t) = N_t(s,a)\underline{\mathbf{K}}_{s,a}(\widehat{\mathbf{M}}_{|t}, \varphi^\star_{\widehat{\mathbf{M}}_{|t}}(s)) + \log N_t(s,a)\,,$$

$$\underline{\mathbf{K}}_{s,a}(\mathbf{M},\varphi) = \inf\left\{\mathtt{KL}(\mathbf{r}(x)\otimes\mathbf{p}(x), \tilde{r}\otimes\tilde{p}) : \tilde{m} + \tilde{p}\mathbf{b}_\star > \varphi\right\}$$

where $\varphi^\star_{\mathbf{M}}(s) = \max_{a\in\mathcal{A}_s} \varphi_{\mathbf{M}}(s,a)$ with $\varphi_{\mathbf{M}}(x) = \mathbf{m}(s,a) + (\mathbf{p}_a\mathbf{b}_\star)(s)$.

Indexed Minimum Empirical Divergence for RL

Algorithm 1 `IMED-RL`

Initialisation State s_1

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

  Compute the skeleton $\mathcal{A}_s(t)$ for each s and set $\widehat{\mathbf{M}}_{|t}$

  Run value iteration on $\widehat{\mathbf{M}}_{|t}$ to get $\varphi^\star_{\widehat{\mathbf{M}}_{|t}}$.

  Sample $A_t \in \arg\min_{a \in \mathcal{A}_s} N_t(s,a)\underline{\mathbf{K}}_{s,a}(\widehat{\mathbf{M}}_{|t}, \varphi^\star_{\widehat{\mathbf{M}}_{|t}}(s)) + \log N_t(s,a)$

▷ Note: $\varphi^\star_{\widehat{\mathbf{M}}_{|t}}$ should be good estimate of $\varphi^\star_{\mathbf{M}}$.

Theorem (Asymptotic Optimality)

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

$$\lim_{T \to +\infty} \frac{\overline{\mathbf{R}}_{T,s_1}(\mathit{IMED\text{-}RL}; \mathbf{M})}{\log T} \leqslant \sum_{s,a \in \mathcal{X}} \frac{\Delta_{s,a}(\mathbf{M})}{\underline{\mathbf{K}}_{s,a}(\mathbf{M})}.$$

Numerical illustrations

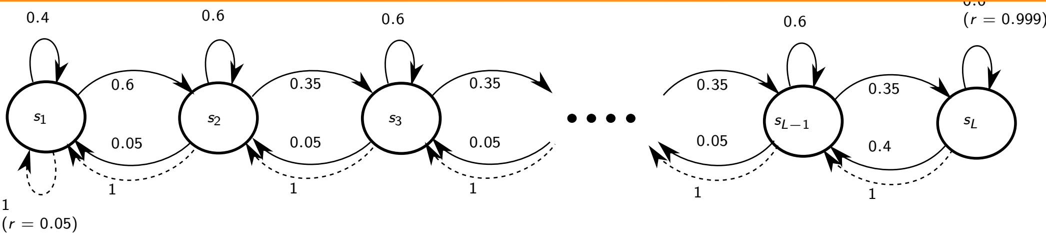
- Ergodic Environment
- Communicating Environment

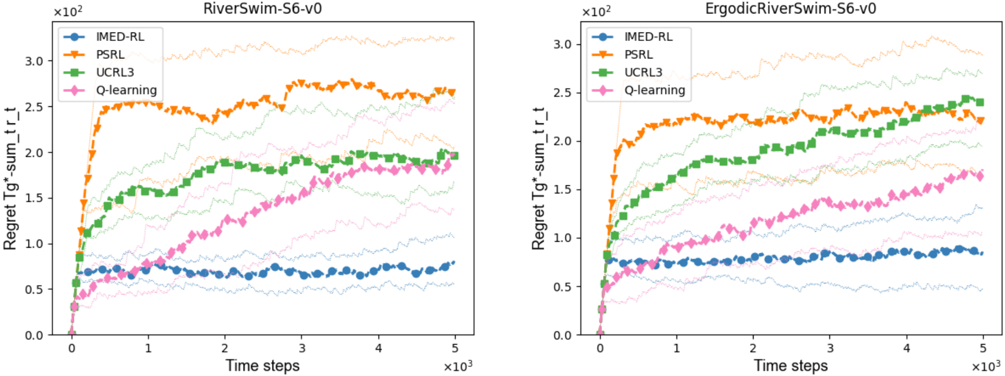
From a communicating only MDP, by mixing its transition $\mathbf{p}$ with that obtained from playing a uniform policy, formally

$$\mathbf{p}_\varepsilon(\cdot|s,a) = (1-\varepsilon)\mathbf{p}(\cdot|s,a) + \varepsilon \sum_{a'\in\mathcal{A}_s} \mathbf{p}(\cdot|s,a')/|\mathcal{A}_s|,$$

for an arbitrarily small $\varepsilon > 0$ one obtain an ergodic MDP.
Experimentally, we take $\varepsilon = 10^{-3}$.

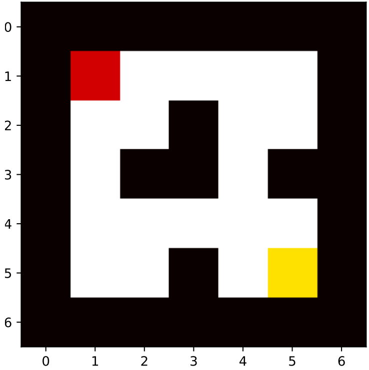
RiverSwim

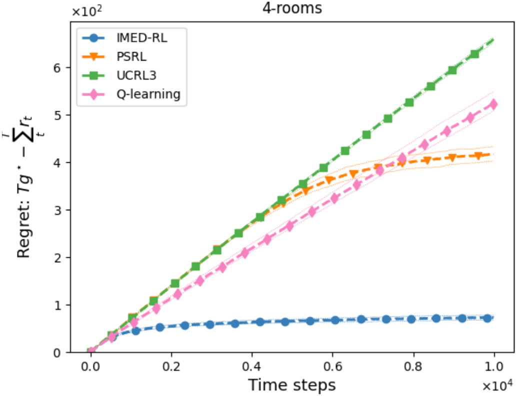




Average regret and quantiles (0.1 ; 0.9) on a communicating (left) and an ergodic (right) 6-states RiverSwim

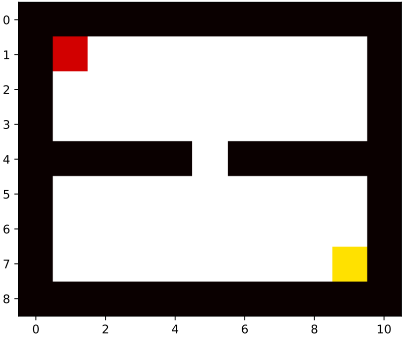
IMED-RL

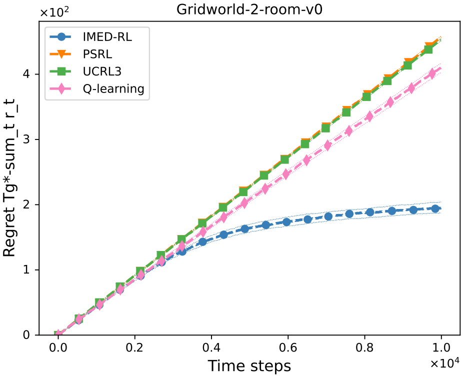




*Average **regret** in the 4-rooms environment*

2-rooms





*Average **regret** and quantiles (0.1 and 0.9) curves in the 2-rooms environment*

Table of contents

Back to IMED-RL

Algorithm 2 `IMED-RL`

Initialisation State s_1

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

Compute the skeleton $\mathcal{A}_s(t)$ for each s and set $\widehat{\mathbf{M}}_{|t}$

Run value iteration on $\widehat{\mathbf{M}}_{|t}$ to get $\varphi^\star_{\widehat{\mathbf{M}}_{|t}}$.

Sample $A_t \in \arg\min_{a \in \mathcal{A}_s} N_t(s,a)\underline{\mathbf{K}}_{s,a}(\widehat{\mathbf{M}}_{|t}, \varphi^\star_{\widehat{\mathbf{M}}_{|t}}(s)) + \log N_t(s,a)$

- Optimization problem $\underline{\mathbf{K}}_{s,a}$ computed at each step: find an **iterative approach** (reusing previous computation)?
- Compute $N_t(s,a)$, $\underline{\mathbf{K}}_{s,a}$ for continuous s, a: Introduce encoder for s, a, density estimation, etc.

Introduction

Achievable performances

Learning algorithm

Extensions

Beyond discrete MDPs

Beyond ergodic MDPs

Challenges

Recurrent classes

- ▷ Perturbation of $\star$ at (s,a) may completely change the recurrent class: $\mathcal{X}_\star$ vs $\mathcal{X}_{\star_{s,a}}$
- ▷ (s,a) may not be recurrent: find confusing instance by changing $\mathbf{M}$ at other pairs.

$$L_T(\mathbf{M},\tilde{\mathbf{M}}) \quad = \sum_{x\in\mathcal{X}_{\star_{s,a}}\setminus\mathcal{X}_\star} \mathbb{E}_{\mathbf{M}}\Big[N_T(x)\Big]\mathrm{KL}_x(\mathbf{M},\tilde{\mathbf{M}})\,.$$

- ▷ NO clear simplification in general.

Estimation

- ▷ Some states may not be visited frequently, even by an optimal policy
- ▷ Estimates of $\mathbf{r}(x)$, $\mathbf{p}(x)$ may not converge for such $x \notin \mathcal{X}_\star$!
- ▷ Study case with known support of $\mathbf{p}$?

Generic lower bound

▷ **Finite time lower bound** For any **uniformly η-good** strategy on $\mathfrak{M}$, and any MDP $\mathbf{M} \in \mathfrak{M}$, it holds for all T such that $\eta_T \leqslant \frac{1}{2|\mathcal{X}|}$,

$$\overline{\mathfrak{R}}_{T,s_1}(\pi; \mathbf{M}) \quad \geqslant \quad \overline{\mathbb{C}}_{T,s_1}(\mathbf{M}, \mathfrak{M}) \,,$$

where we introduced

$$\overline{\mathbb{C}}_{T,s_1}(\mathbf{M},\mathfrak{M}) = \inf_{n \in T \cdot \mathcal{P}_{|\mathcal{X}|}} \Bigg\{ \sum_{x\in\mathcal{X}} n_x \Delta(x) - \mathbb{S}(b_\star) : \forall \pi, \inf_{\tilde{\mathbf{M}} \in \overline{\mathcal{B}}_{s_1}(\pi,\mathbf{M};\mathfrak{M})} \sum_{x\in\mathcal{X}} \frac{n_x \mathrm{KL}_x(\mathbf{M},\tilde{\mathbf{M}})}{\mathtt{kl}\Big(1-\eta_T|\mathcal{X}|, \eta_T|\mathcal{X}|\Big)}$$

Generic lower bound

▷ **Finite time lower bound** For any **uniformly η-good** strategy on $\mathfrak{M}$, and any MDP $\mathbf{M} \in \mathfrak{M}$, it holds for all T such that $\eta_T \leqslant \frac{1}{2|\mathcal{X}|}$,

$$\overline{\mathfrak{R}}_{T,s_1}(\pi;\mathbf{M}) \quad \geqslant \quad \overline{\mathbb{C}}_{T,s_1}(\mathbf{M},\mathfrak{M}),$$

where we introduced

$$\overline{\mathbb{C}}_{T,s_1}(\mathbf{M},\mathfrak{M}) = \inf_{n \in T \cdot \mathcal{P}_{|\mathcal{X}|}} \left\{ \sum_{x \in \mathcal{X}} n_x \Delta(x) - \mathbb{S}(b_\star) : \forall \pi, \inf_{\tilde{\mathbf{M}} \in \overline{\mathcal{B}}_{s_1}(\pi,\mathbf{M};\mathfrak{M})} \sum_{x \in \mathcal{X}} \frac{n_x \mathrm{KL}_x(\mathbf{M},\tilde{\mathbf{M}})}{\mathtt{kl}\Big(1-\eta_T|\mathcal{X}|, \eta_T|\mathcal{X}|\Big)}\right.$$

▷ **Asymptotic lower bound** Furthemore, for $\eta_T = O(T^{\alpha-1})$ for all $\alpha \in (0,1)$, asymptotically, we get

$$\liminf_T \frac{\overline{\mathfrak{R}}_{T,s_1}(\pi,\mathbf{M})}{\ln(T)} \quad \geqslant \inf_{\kappa \in \mathbb{R}_+^{|\mathcal{X}|}} \left\{ \sum_{x \in \mathcal{X}} \kappa_x \Delta(x) : \forall \pi, \inf_{\tilde{\mathbf{M}} \in \overline{\mathcal{B}}_{s_1}(\pi,\mathbf{M};\mathfrak{M})} \sum_{x \in \mathcal{X}} \kappa_x \mathrm{KL}_x(\mathbf{M},\tilde{\mathbf{M}}) \geqslant 1 \right\}$$

Complexity

▷ Computing **confusing MDP** hence cost implies searching over **all places** where MDP could be modified ($|\mathcal{X}_{\star_{s,a}} \setminus \mathcal{X}_\star|$, possibly $O(SA)$ pairs). Conjecture: cost is exponential in $|\mathcal{X}_{\star_{s,a}} \setminus \mathcal{X}_\star|$.

▷ In non-unichain MDPs, there may be no improving policies over π at Hamming distance 1 of π.

"The more applied you go, the stronger theory you need"

MERCI

odalricambrym.maillard@inria.fr

odalricambrymmaillard.wordpress.com