Regret minimization in Average Reward Reinforcement Learning
III. EXTENSIONS

Odalric-Ambrym MAILLARD
Centre Inria de l'Université de Lille
*Equipe **SCOOL***
(Sequential, Continual and Online Learning)

Master MVA

References I


Fabien Pesquerel and Odalric-Ambrym Maillard.
Imed-rl: Regret optimal learning of ergodic markov decision processes.
Advances in Neural Information Processing Systems, 35:26363–26374, 2022.


Fabien Pesquerel.
Quantité d'information par unité d'interaction en apprentissage séquentiel stochastique.
2023.


Hassan Saber, Fabien Pesquerel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi.
Logarithmic regret in communicating mdps: Leveraging known dynamics with bandits.
In *Asian Conference on Machine Learning*, pages 1167–1182. PMLR, 2024.

References II

Victor Boone and Odalric-Ambrym Maillard.
The regret lower bound for communicating markov decision processes.
arXiv preprint arXiv:2501.13013, 2025.

Waris Radji and Odalric-Ambrym Maillard.
The confusing instance principle for online linear quadratic control.
Reinforcement Learning Journal, 6, 2025.

Hassan Saber.
Structure adaptation in bandit theory.
PhD thesis, Université de Lille, 2022.

Jie Bian and Vincent YF Tan.
Indexed minimum empirical divergence-based algorithms for linear bandits.
arXiv preprint arXiv:2405.15200, 2024.

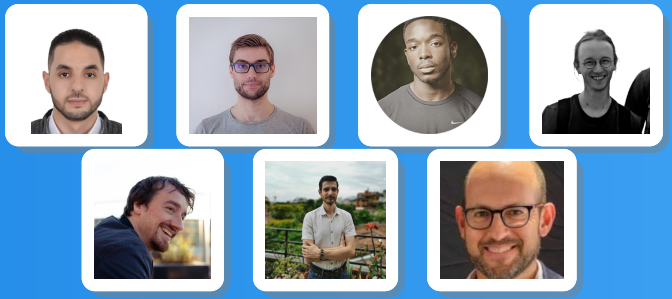
References III


Hassan Saber and Odalric-Ambrym Maillard.
Bandits with multimodal structure.
In *Reinforcement Learning Conference*, 2024.


Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard.
Indexed minimum empirical divergence for unimodal bandits.
Advances in Neural Information Processing Systems, 34:7346–7356, 2021.


Waris Radji and Odalric-Ambrym Maillard.
How Hard is it to Confuse a World Model?
arXiv preprint arXiv:2510.21232, 2025.

Collaborators



Take home message

- ☑ The most confusing instance paradigm
- ☑ Lower bounds for Bandits and Ergodic MDPs.
- ☑ IMED-RL strategy and optimality
- ☑ Extensions to Linear quadratic systems
- ☑ Apply this principle to other setups?

OUTLINE

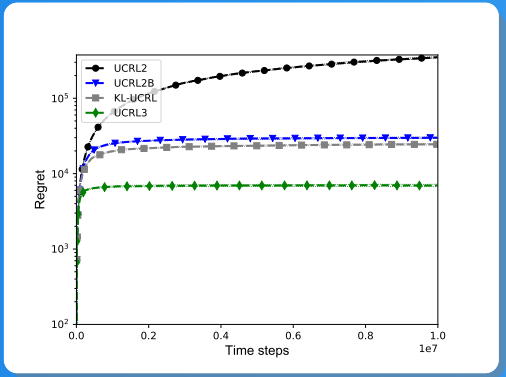
1] The most confusing paradigm.

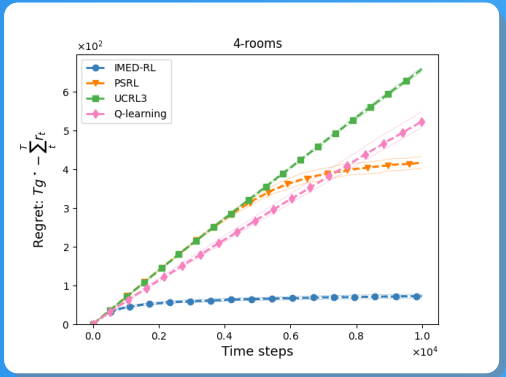
2] IMED-RL

3] ...

The “most confusing instance” paradigm.

Teaser on a 4-room MDP

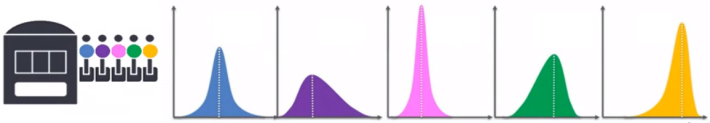




Multi-armed bandits setup

❖ **Multi-armed bandit**

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



At each decision time $t \in \mathbb{N}$:

agent chooses $a_t \in \mathcal{A}$,

agent receives $r_t \sim \mathbf{r}(a_t)$.

Adaptive policy $\pi = (\pi_t)_t$: $a_t \sim \pi_t$ and π_t built from observed trajectory $\tau_{t-1} = (a_1, r_1, \ldots, a_{t-1}, r_{t-1})$.

❖ **Score function**, e.g. the mean, $\mathbf{m} : \mathcal{A} \to [0,1]$

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

Cumulative regret minimization objective

❖ **Cumulative regret**

$$\mathfrak{R}_T(\pi, \mathbf{M}) \quad = \quad \sum_{a \in \mathcal{A}} \mathbb{E}_\pi[N_T(a)]\Delta(a) \text{ where } \begin{cases} \Delta(a) = & \mathbf{m}_\star - \mathbf{m}(a) \\ N_T(a) = & \sum_{t=1}^{T} \mathbb{I}\{a_t = a\} \end{cases}.$$

Classical regret lower bounds: Any consistent* learning must incur at least

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

❖ A **consistent algorithm** must be good simultaneously on many models.

$$\forall \mathbf{M} \in \mathcal{M}, \forall a \notin \mathcal{O}(\mathbf{M}), \forall \alpha \in (0,1) \quad \mathbb{E}[N_T(a)] = o(T^\alpha).$$

Instance dependent and independent settings

What does it mean for Λ to learn?

Idea 1: Sublinear regret is convergence to optimal play. For *all* instance M of MDP,

$$\limsup_{T\to\infty} \frac{1}{T} \operatorname{Reg}(T; \Lambda, M) = 0.$$

Idea 2a: Consistency.

$$\sup_{M\in\mathcal{M}} \limsup_{T\to\infty} \frac{1}{T} \operatorname{Reg}(T; \Lambda, M) = 0$$

$\Downarrow$

instance dependent regret

Idea 2b: Robustness.

$$\limsup_{T\to\infty} \frac{1}{T} \sup_{M\in\mathcal{M}} \operatorname{Reg}(T; \Lambda, M) = 0$$

$\Downarrow$

worst case or minimax
or instance independent regret

Likelihood

Compare **Log-likelihood** of two models $\mathbf{M} = (\mathcal{A}, \mathbf{r})$ and $\tilde{\mathbf{M}} = (\mathcal{A}, \tilde{\mathbf{r}})$ on trajectory $\tau_T = (a_1, r_1, \ldots, a_T, r_T)$:

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

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

❖ **Fundamental inequality** For any test function $h : \mathcal{X}^\infty \to [0, 1]$,

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

where $\mathrm{kl}(x, y) = \mathrm{KL}(\mathrm{Bern}(x), \mathrm{Bern}(y))$.

The Unlikelihood of optimality

Confusing models for sub-optimal action $a \notin \mathcal{O}(\mathbf{M})$

$$\mathcal{C}_a(\mathbf{M}) = \left\{ \tilde{\mathbf{M}} : a \succ \star_{\mathbf{M}} \text{ in } \tilde{\mathbf{M}} \text{ and } \mathbf{M}_{|\mathcal{O}(\mathbf{M})} = \tilde{\mathbf{M}}_{|\mathcal{O}(\mathbf{M})} \right\}$$

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

Unlikelihood of optimality (= confusion cost)

$$\begin{aligned} \mathbb{U}_T(a, \mathbf{M}) &= \inf_{\tilde{\mathbf{M}} \in \mathcal{C}_a(\mathbf{M})} \mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} \right], \\ &= \mathbb{E}_{\mathbf{M}}\Big[N_T(a) \Big] \inf_{\tilde{\mathbf{r}}} \left\{ \mathrm{KL}(\mathbf{r}(a), \tilde{r}) : \tilde{\mathbf{r}} \text{ has mean} > \mathbf{m}_\star \right\} \end{aligned}$$

The larger $\mathbb{U}_T(a, \mathbf{M})$, the most difficult to make a look optimal.

Lower bound for regret mininimization

Combining previous steps, we have for $\tilde{\mathbf{M}}_a \in \mathcal{C}_a(\mathbf{M})$.

$$\mathbb{U}_T(a, \mathbf{M}) = \mathbb{E}_{\mathbf{M}}\Big[N_T(a)\Big]\underline{\mathbf{K}}(\mathbf{r}_a, \mathbf{m}_\star) \geqslant \sup_h \mathrm{kl}\left(\mathbb{E}_{\mathbf{M}}[h(\tau_T)], \mathbb{E}_{\tilde{\mathbf{M}}_a}[h(\tau_T)]\right)$$

Choosing $h(\tau_T) = 1 - N_T(a)/T$, using **consistenty assumption** yields

$$\forall \alpha \in (0,1), \mathrm{kl}\left(\mathbb{E}_{\mathbf{M}}[h(\tau_T)], \mathbb{E}_{\tilde{\mathbf{M}}_a}[h(\tau_T)]\right) \geqslant (1-\alpha)\log(T) + o(\log(T))\,.$$

❖ **Conclusion** : "bad arms" must be pulled often enough.

$$\forall a \notin \mathcal{O}(\mathbf{M}), \quad \liminf_T \frac{\mathbb{E}_{\mathbf{M}}\Big[N_T(a)\Big]}{\log(T)} \geqslant \frac{1}{\underline{\mathbf{K}}(\mathbf{r}_a, \mathbf{m}_\star)}\,.$$

Algorithm design

☒ Strategy 1: Compute MLE $\widehat{\mathbf{M}}_t$,
then pick $a_t \in \mathcal{O}(\widehat{\mathbf{M}}_t)$

vs

☑ Strategy 2: Compute MLE $\widehat{\mathbf{M}}_t$,

For each "bad" a, compute $\inf_{\tilde{\mathbf{M}} \in \mathcal{C}_a(\mathbf{M})} \mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)}\right]$: "track" minimal cost.

Algorithm: Unlikelihood tracking

For MLE $\widehat{\mathbf{M}}_t$, try to match log-frequency of plays $\log(\frac{N_t(a)}{t})$ with $-\mathbb{U}_t(a, \widehat{M}_t)$.

- **Deterministic tracking**: *IMED*, *J. Honda, A. Takemura, JMLR 2015.*

$$\begin{aligned} A_t &\in \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{U}_t(a, \widehat{\mathbf{M}}_t) + \log\left(\frac{N_t(a)}{t}\right). \\ &= \operatorname*{argmin}_{a \in \mathcal{A}} N_T(a) \underline{\mathbf{K}}(\widehat{\mathbf{r}}_t(a), \widehat{\mathbf{m}}_{\star,t}) + \log(N_t(a)). \end{aligned}$$

- **Stochastic tracking**: *Maillard sampling*, *J. Bian, K-S. Jun, AI&STATS 2022, OA.M, PhD, 2011.*

$$A_t \propto \exp\left(-\mathbb{U}_t(a, \widehat{\mathbf{M}}_t)\right)$$

Optimality of unlikelihood tracking

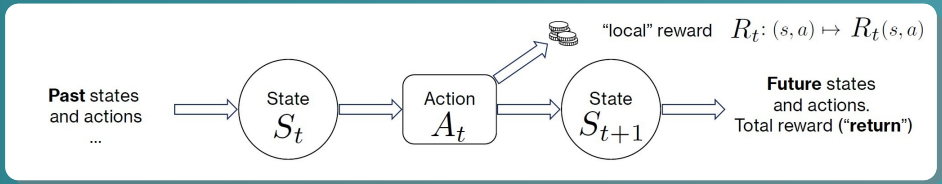
Theorem (Honda & Takemura, 2015) strategy IMED is **optimal** for (regular) models $\mathcal{M}$, $\forall \mathbf{M} \in \mathcal{M}$,

$$\sum_{a \in \mathcal{A}} \frac{\Delta_a}{\underline{\mathbf{K}}(\mathbf{r}_a, \mathbf{m}_\star)} \leqslant \liminf_{T \to \infty} \frac{\mathfrak{R}_T(\pi_{\mathrm{IMED}}, \mathbf{M})}{\log T} \leqslant \limsup_{T \to \infty} \frac{\mathfrak{R}_T(\pi_{\mathrm{IMED}}, \mathbf{M})}{\log T} \leqslant \sum_{a \in \mathcal{A}} \frac{\Delta_a}{\underline{\mathbf{K}}(\mathbf{r}_a, \mathbf{m}_\star)}$$

- ☑ Provably **instance-dependent exact-optimal** for bandit instances.
- ☑ Extension to **structured** bandits & **MDPs**: H. Saber (PhD), F. Pesquerel (PhD).

Adaptation to MDPs

Uncertain Markov Decision Processes $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{r}, \mathbf{p})$.



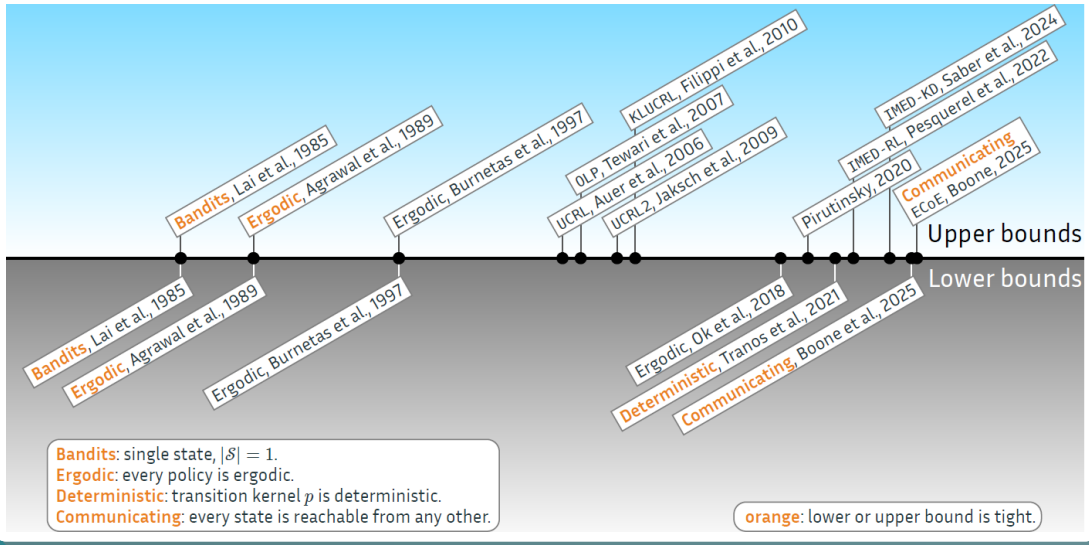
The cumulative reward (value) at time T, starting from an initial state s_1 of policy $\boldsymbol{\pi} = (\pi_t)_t$ is

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

Note: $\lim_{T\to\infty} \frac{1}{T} V_{s_1}(\mathbf{M}, \pi, T) = \mathbf{g}^{\mathbf{M}}(s_1) = (\bar{\mathbf{p}}_\pi \mathbf{m})(s_1)$ for fixed $\boldsymbol{\pi} \equiv \pi$.

Optimize performance of policy (π) vs optimal ($\star_{\mathsf{M}}$) while learning

History of instance-dependent optimality



Likelihood ratios and confusing models

Trajectory of interaction with a system

$$\tau_T = (s_1, a_1, r_1, \ldots, s_T, a_T, r_T, s_{T+1})$$

Compare likelihood of two models $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$, $\tilde{\mathbf{M}} = (\mathcal{S}, \mathcal{A}, \tilde{\mathbf{p}}, \tilde{\mathbf{r}})$.

$$\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} = \sum_{t=1}^{T} \log \frac{\mathbf{r}(s_t, a_t)(r_t)}{\tilde{\mathbf{r}}(s_t, a_t)(r_t)} + \log \frac{\mathbf{p}(s_t, a_t)(s_{t+1})}{\tilde{\mathbf{p}}(s_t, a_t)(s_{t+1})}.$$

Learning: **model unknown**.
Maximise in M: **maximum likelihood**.

The Unlikelihood of optimality

Take **sub-optimal** $\pi \neq \star_{\mathbf{M}}$.

Optimize expectation over **confusing** models: $\mathcal{C}_\pi(\mathbf{M}) = \left\{ \tilde{\mathbf{M}} : \pi \succ \star_{\tilde{\mathbf{M}}} \text{ in } \tilde{\mathbf{M}} \right\}$:

$$\begin{aligned}
\mathbb{U}_T(\pi, \mathbf{M}) &= \inf_{\tilde{\mathbf{M}} \in \mathcal{C}_\pi(\mathbf{M})} \mathbb{E}_{\mathbf{M}}\left[\log \frac{\mathbf{M}(\tau_T)}{\tilde{\mathbf{M}}(\tau_T)} \right], \\
&= \inf_{\tilde{\mathbf{M}} \in \mathcal{C}_\pi(\mathbf{M})} \sum_{x \in \mathcal{S} \times \mathcal{A}} \mathbb{E}_{\mathbf{M}}\left[N_T(x) \right] \Big(\mathrm{KL}(\mathbf{p}(x), \tilde{\mathbf{p}}(x)) + \mathrm{KL}(\mathbf{r}(x), \tilde{\mathbf{r}}(x)) \Big).
\end{aligned}$$

☑ Main tool to derive lower performance bounds.

☒ Possibly **hard** to compute in general (exp. many ways!). Ok in specific MDPs.

Generic lower bound

No explicit bound (until recently): A simplified version is

$$\forall \mathbf{M} \in \mathcal{M}, \quad \liminf_T \frac{\mathfrak{R}_T(\pi, \mathbf{M})}{\ln(T)} \geqslant \overline{\mathbb{C}}(\mathbf{M}, \mathcal{M}) \text{, where}$$

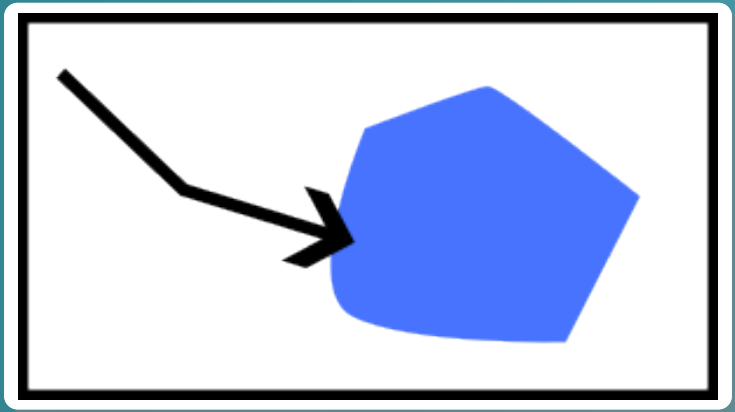
$$\overline{\mathbb{C}}(\mathbf{M}, \mathcal{M}) = \inf_{\kappa \in \mathbb{R}_+^{|\mathcal{X}|}} \left\{ \sum_{c \in \mathcal{C}_\pi \setminus \mathcal{C}_\star} \kappa_c \Delta(c) : \forall \pi, \inf_{\tilde{\mathbf{M}} \in \mathcal{B}(\pi, \mathbf{M}; \mathcal{M})} \sum_{c \in \mathcal{C}} \kappa_c \mathrm{KL}_c(\mathbf{M}, \tilde{\mathbf{M}}) \geqslant 1 \right\}.$$

- ☒ Without further assumption, NP-hard to solve, no **computationally efficient** algorithm can achieve generic optimality.

- ☑ $\mathcal{C}_\pi = \{(s, a) \in \mathcal{C} : \overline{\mathbf{p}}_\pi(s)\pi(a|s) > 0\}$ set of recurrent pairs under π.

Recurrent pairs

- $\mathcal{C}_\pi = \{(s,a) \in \mathcal{C} : \overline{\mathbf{p}}_\pi(s)\pi(a|s) > 0\}$ set of recurrent pairs under π.



- ☑ Ergodic : all states are recurrent under any policy.

Simplifications under Ergodic assumption

☑ Lower bound are explicit for the same reason: (Agrawal, 90) (Burnetas & Katehakis, 97), (Graves & Lai, 97). On ergodic MDPs $\mathcal{M}$,

$$\liminf_{T} \frac{\overline{\mathfrak{R}}_T(\pi, \mathbf{M})}{\ln(T)} \geqslant \sum_{(s,a)\in\mathcal{C}} \frac{\Delta(s,a)}{\underline{\mathbf{K}}_c\left(\mathbf{M}, \mathbf{Q}^\star_{\mathbf{M}}(s)\right)}$$

☑ Reduction : Unlikelihood for a single-pair perturbation $\star_{s,a}$

$$\inf_{\tilde{M}\in\mathcal{B}(\star_{s,a},\mathbf{M})} \sum_{c\in\mathcal{C}} \mathbb{E}_{\mathbf{M}}\Big[N_T(c)\Big]\mathrm{KL}_c(\mathbf{M},\tilde{\mathbf{M}}) = \mathbb{E}_{\mathbf{M}}\Big[N_T(s,a)\Big]\underline{\mathbf{K}}_c\left(\mathbf{M}, \mathbf{Q}^\star_{\mathbf{M}}(s)\right)$$

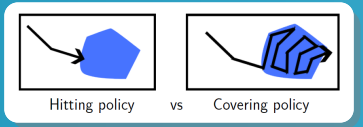
where $\underline{\mathbf{K}}_c\left(\mathbf{M}, \mathbf{Q}\right)$ is reminiscent of bandits and computable !

Teaser: Beyond Ergodic MDPs

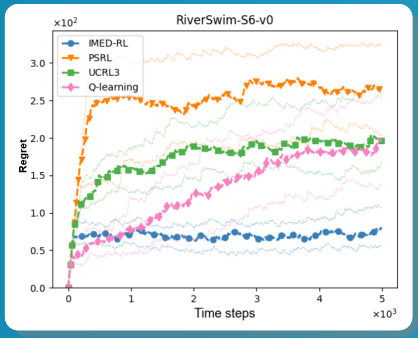
H. Saber (PhD), F. Pesquerel (PhD), S. Talebi (U. Copenhagen)

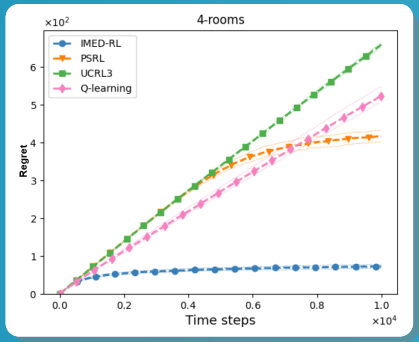
Challenge:

- ≈ computation of unlikelihood.
- ☑ navigation challenge: covering time of recurrent sets.
- ☑ both sound, computable, competitive in practice.



Regret novel strategy (blue) **IMED-RL** against state-of-the-art:





Teaser: The generic lower bound

☑ Very recently (Boone & M., 2026), we uncovered the delicate structure of the lower bound in generic MDPs.

The Instance-Dependent World

State-of-the-art of Regret Minimization: *(theoretical side)*

Setting	Lower bound	Upper bound
Minimax	$\Omega\left(\sqrt{\mathrm{sp}(b^\star)SAT}\right)$ (Jaksch et al., 2009)	$\mathrm{O}\left(\sqrt{\mathrm{sp}(b^\star)SAT\log(T)}\right)$ (Boone, Zhang, 2024)
Instance dependent	$K(M)\log(T)$ (Boone, Maillard, 2025)	$K(M)\log(T)+\mathrm{o}(\log(T))$ (Boone, 2025)

as soon as M is a **communicating** environment.

- (TBP) Both bounds cannot be reached simulatenously by the same algorithm (**no Best-of-Both-Worlds**)

OUTLINE

1] The most confusing paradigm.

2] IMED-RL

3] Extensions

4] ...

IMED-RL

❖ For $\widetilde{\mathbf{r}} \otimes \widetilde{\mathbf{p}} \in \mathcal{P}(\mathbb{R} \times \mathcal{S})$, where $\widetilde{\mathbf{r}}$ has mean $\widetilde{\mathbf{m}}$, we introduce

$$T[\widetilde{\mathbf{r}} \otimes \widetilde{\mathbf{p}}](\mathbf{b}^{\mathbf{M}}) \quad = \quad \widetilde{\mathbf{m}} + \widetilde{\mathbf{p}}\mathbf{b}^{\mathbf{M}}$$

Poisson equation, with gain $\mathbf{g}^{\mathbf{M}}$ and bias $\mathbf{b}^{\mathbf{M}}$ functions.

$$\mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s) = \max_{a \in \mathcal{A}_s} \left\{ \mathbf{m}(s,a) + \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s,a)\mathbf{b}^{\mathbf{M}}(s') \right\}. \qquad (2)$$

We denote the sub-optimality gap of a pair (s,a) by

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

Critical pairs

❖ The sub-optimality cost of a sub-optimal state-action pair $(s,a) \in \mathcal{X}_{\mathbf{M}}$ is defined as $\mathbf{K}_{s,a}(\mathbf{M}) \stackrel{\text{def}}{=} \mathbf{K}_{s,a}(\mathbf{M}, \gamma_s(\mathbf{M}))$ where

$$\mathbf{K}_{s,a}(\mathbf{M},\gamma) = \inf_{\substack{\widetilde{\mathbf{r}} \in \mathcal{F}_{s,a} \\ \widetilde{\mathbf{p}} \in \mathcal{P}(\mathcal{S})}} \left\{ \mathrm{KL}\left(\mathbf{r}(s,a) \otimes \mathbf{p}(\cdot|s,a), \widetilde{\mathbf{r}} \otimes \widetilde{\mathbf{p}}\right) \ : \ T[\widetilde{\mathbf{r}} \otimes \widetilde{\mathbf{p}}](\mathbf{b}^{\mathbf{M}}) > \gamma \right\}$$

where $\gamma_s(\mathbf{M}) = \mathbf{g}^{\mathbf{M}} + \mathbf{b}^{\mathbf{M}}(s) = \max_{a \in \mathcal{A}_s} T[\mathbf{r}(s,a) \otimes \mathbf{p}(s,a)](\mathbf{b}^{\mathbf{M}})(s)$.

❖ $\mathcal{C}(\mathbf{M}) = \{(s,a) : 0 < \mathbf{K}_{s,a}(\mathbf{M}) < \infty\}$: pairs that could be confused for an optimal one at the price of paying displacement cost $\mathbf{K}_{s,a}(\mathbf{M})$.

Assumption

- **Communicating** : $\forall s, s', \exists \pi, t \in \mathbb{N} : \mathbf{p}_\pi^t(s'|s) > 0.$

- **Ergodic** : $\forall s, s', \forall \pi, \exists t \in \mathbb{N} : \mathbf{p}_\pi^t(s'|s) > 0.$

Ergodic is more restrictive. Ensures all policies visits all states infinitely often.

Lower bounds

Regret lower bound (Burnetas & Katehakis, 1997)

Let $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{p}, \mathbf{r})$ be an ergodic MDP with bounded rewards. For all uniformly consistent learning algorithm π,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\pi,\mathbf{M}}\left[N_{s,a}(T)\right]}{\log T} \geqslant \frac{1}{\mathbf{K}_{s,a}\left(\mathbf{M}\right)} \tag{3}$$

with the convention that $1/\infty = 0$. The regret lower bound is

$$\liminf_{T\to\infty} \frac{\mathcal{R}_{\pi}\left(\mathbf{M}, T\right)}{\log T} \geqslant \sum_{(s,a)\in\mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}\left(\mathbf{M}\right)}{\mathbf{K}_{s,a}\left(\mathbf{M}\right)} \tag{4}$$

IMED-RL ingredients

❖ The skeleton is the subset of state-action pairs considered sampled enough at time t: $\mathcal{A}(t) = (\mathcal{A}_s(t))_{s\in\mathcal{S}}$ where

$$\mathcal{A}_s(t) = \left\{ a \in \mathcal{A}_s \ : \ N_{s,a}(t) \geqslant \log^2 \left(\max_{a' \in \mathcal{A}_s} N_{s,a'}(t) \right) \right\}. \tag{5}$$

❖ Empirical MDP $\widehat{\mathbf{M}}_t = (\mathcal{S}, \mathcal{A}(t), \widehat{\mathbf{p}}_t, \widehat{\mathbf{r}}_t)$ restricted to the skeleton.

Empirical threshold $\widehat{\gamma}_s(t) = \max_{a\in\mathcal{A}_s} T[\widehat{\mathbf{r}}(s,a) \otimes \mathbf{p}(s,a)](\mathbf{b}^{\widehat{\mathbf{M}}_t})$.

The **IMED-RL index** of state-action pair $(s,a) \in \mathcal{X}_{\mathbf{M}}$ is

$$I_{s,a}(t) \quad = \quad N_{s,a}(t) \widehat{\mathbf{K}}_{s,a} \left(\widehat{\mathbf{M}}_t, \widehat{\gamma}_s(t) \right) + \log N_{s,a}(t) \, .$$

Algorithm: For all t, sample $a_t \in \arg\min_{a\in\mathcal{A}_{s_t}} I_{s,a}(t)$.

Asymptotic optimality

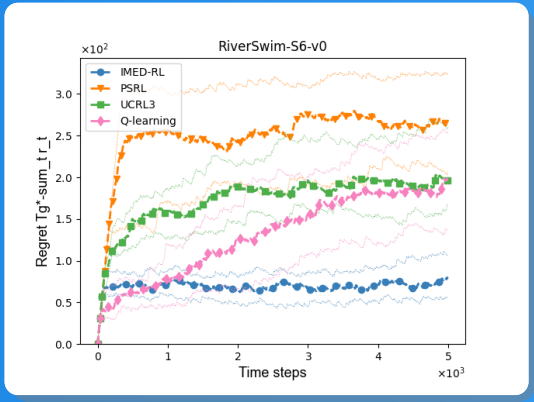
IMED-RL is asymptotically optimal, for ergodic MDPs, that is,

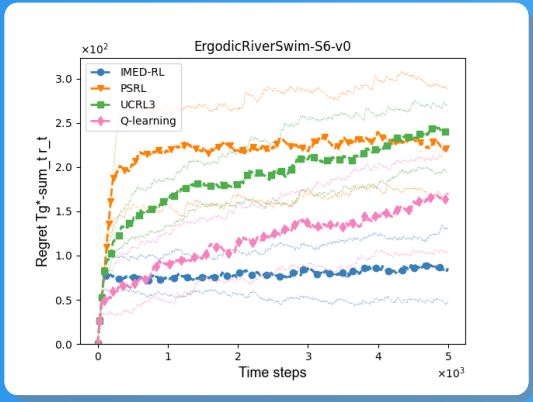
$$\lim_{T \to +\infty} \frac{\mathcal{R}_{IMED-RL}(\mathbf{M}, T)}{\log T} \leqslant \sum_{(s,a) \in \mathcal{C}(\mathbf{M})} \frac{\Delta_{s,a}(\mathbf{M})}{\mathbf{K}_{s,a}(\mathbf{M})}. \tag{7}$$

Also, the strategy is fast (no complicated optimization step!): Average runtime (second) on 8×8 grid-world.

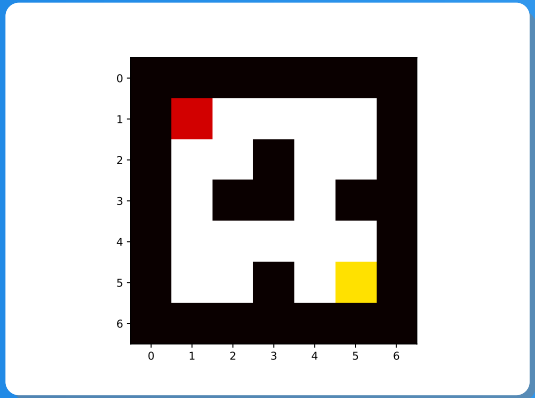
IMED-RL	PSRL	UCRL3	Q-learning
1.82	0.75	6.36	0.03

Numerical performances





Numerical performances



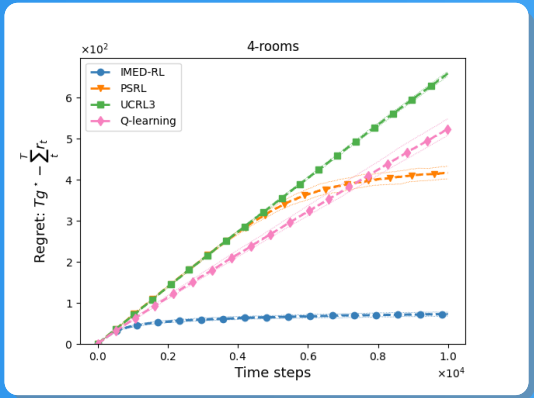
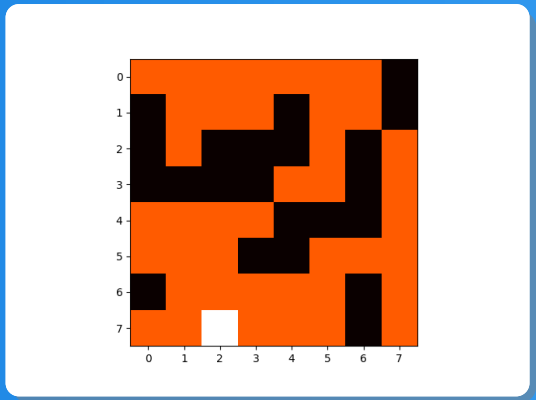


Figure: Average regret and quantiles (0.1 and 0.9) curves of algorithms (right) corresponding to learning on a 4-room (left) grid-world environment, with 20 states: the starting state is shown in red, and the rewarding state is shown in yellow. From the yellow state, all actions bring the learner to the red state. Other transitions are noisy as in a *frozen-lake* environment.

Numerical performances



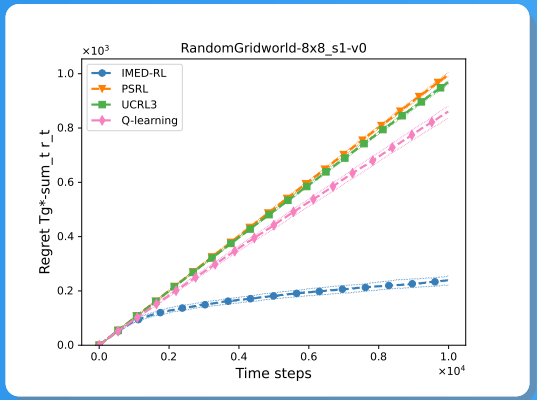


Figure: Average regret and quantiles (0.1 and 0.9) curves of algorithms (right) in a randomly generated grid-world (8x8 grid, 4 actions) with reward 0.99 in white state (right).

State-of-the-Art Regret Bounds

Algorithm	Regret bound
`UCRL` (Jaksch et al., 2010)	$\mathcal{O}\left(DS\sqrt{AT\log(T/\delta)}\right)$
`KLUCRL` (Filippi et al., 2010)	$\mathcal{O}\left(DS\sqrt{AT\log(\log(T)/\delta)}\right)$
`KLUCRL` (Talebi et al., 2018)	$\mathcal{O}\left(\left[D+\sqrt{S\sum_{s,a}\max(\mathbb{V}_{s,a},1)}\right]\sqrt{T\log(\log(T)/\delta)}\right)$
`SCAL`$^+$ (Qian et al., 2019)	$\mathcal{O}\left(D\sqrt{\sum_{s,a}K_{s,a}T\log(T/\delta)}\right)$
`UCRLB` (Fruit et al., 2019)	$\mathcal{O}\left(\sqrt{D\sum_{s,a}K_{s,a}T\log(T)\log(T/\delta)}\right)$
`UCRLnew` **(This Paper)**	$\mathcal{O}\left(\left(D+\sqrt{\sum_{s,a}\max(D_s^2L_{s,a},1)}\right)\sqrt{T\log(T/\delta)}\right)$
Lower Bound (Jaksch et al., 2010)	$\Omega\left(\sqrt{DSAT}\right)$

OUTLINE

Beyond Ergodicity

❖ Ergodicity ensured that all states are visited enough, no matter which policy is played: good estimate of all transitions and rewards.

☒ Most studied MDPs are not ergodic

☒ In non-ergodic MDPs, some state-actions will not be sampled often!

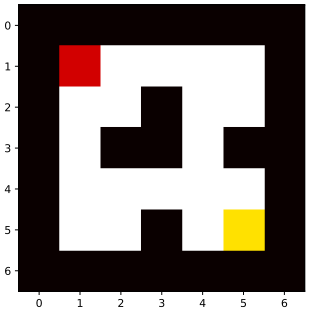
❖ ☑ Policy improvement : In ergodic MDPs, for all sub-optimal policy π, there exists single pair perturbation π' with larger gain $\mathbf{g}_{\pi'} > \mathbf{g}_{\pi}$.

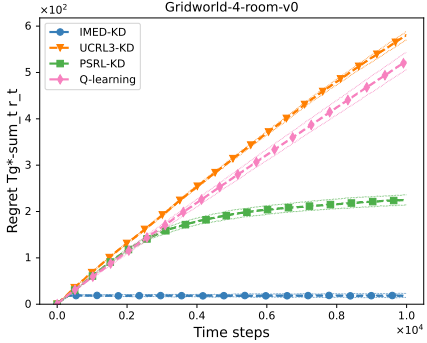
❖ ☒ In non-ergodic MDPs, single state perturbation may not be enough to improve a policy: $\forall \pi \notin \mathcal{O}(\mathbf{M}), \exists \pi' : \texttt{hamming}(\pi, \pi') \leqslant 1$, s.t. $\mathbf{g}_{\pi'} > \mathbf{g}_{\pi}$ may not hold.

Hassan Saber, Fabien Pesquerel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi.
Logarithmic regret in communicating mdps: Leveraging known dynamics.
In *Asian Conference on Machine Learning*, pages 1167–1182. PMLR, 2024.

Non-ergodic MDP Experiments

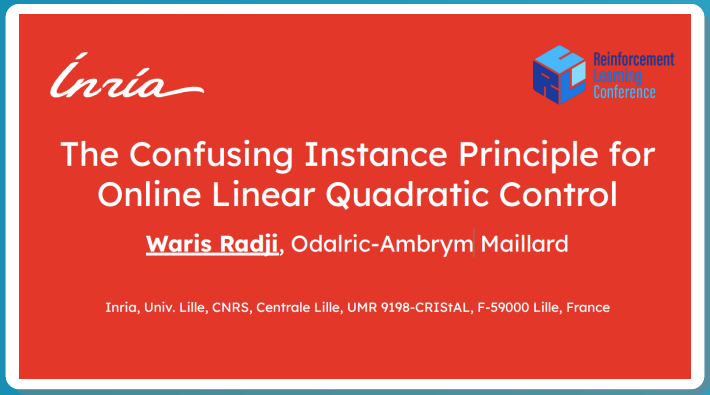
Teaser for IMED-KD, assuming **r** is unknown, and **p** is known.





Extension to LQR systems

W. Radji & O. Maillard *RLC 2025*



The most confusing instance paradigm

1] Each arm competes against current **empirical best arm**.

2] For each arm, what is the smallest change to the environment that makes it **optimal**?

3] **Unlikelihood of optimality**: the infimum of KL divergence to make challenger optimal.

4]**Confusing instance**: The environment corresponding to the infimum.

5]**Strategy**: For each candidate find the most confusing instance, evaluate the unlikelihood and select the candidate with the lowest such.

Why Linear systems?

- Testing MED ideas **beyond bandits and tabular MDPs**
- LQR: the simplest continuous control problem

Linear Dynamics: $x_{t+1} = Ax_t + Bu_t$
Agent policy: $u_t = -Kx_t$
Quadratic cost: $c_t = x_t^\top Q x_t + u_t^\top R u_t$

Known dynamics $\rightarrow$ closed-form optimal policy via Algebraic Riccati Equation
Unknown dynamics $\rightarrow$ model-based RL problem.

Counts in linear spaces

Linear bandit setting: $\mathbf{m}(a) = \theta^\top a$, $r_t = \mathbf{m}(a_t) + \xi_t$.

- Estimate $\widehat{\mathbf{m}}_t(a) = \widehat{\theta}_t^\top a$ where

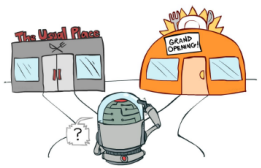
$$\widehat{\theta}_t = G_t^{-1} \sum_{s=1}^{t} a_s r_s \text{ with } G_t = \sum_{s=1}^{t} a_s a_s^\top + \lambda I$$

$$\mathbb{P}\left(|\widehat{\mathbf{m}}_t(a) - \mathbf{m}(a)| > \|\varphi(a)\|_{G_t^{-1}} \left[\sqrt{\lambda}\|\theta\| + \sqrt{\log\left(\frac{\det(G_t)}{\det(\lambda I)}\right) + 2\log\left(\frac{1}{\delta}\right)} \right] \right) \leqslant \delta$$

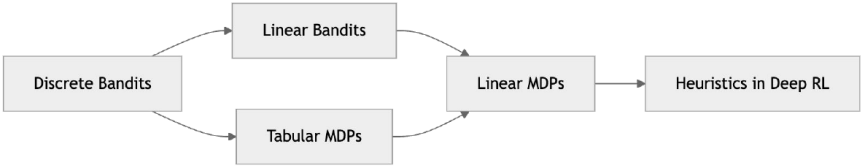
- Generalizes bound $\frac{1}{\sqrt{N_t(a)}}\sqrt{2\log(t/\delta)}$: $N_t(a)$ becomes $\frac{1}{\|a\|^2_{G_t^{-1}}}$

Context: From bandits to continuous RL

Exploration: Try new things to learn more.
Exploitation: Use what you know to get rewards.
The tradeoff: Try something new or stick with what works?

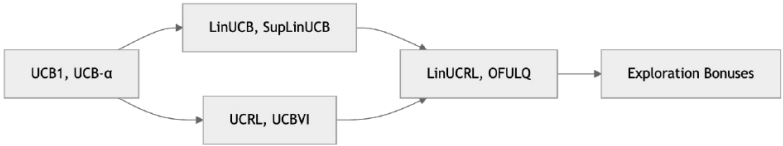


The exploration-exploitation dilemma in RL is rooted in bandits.

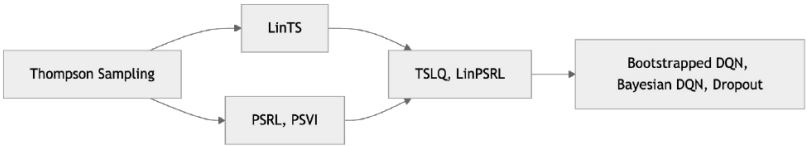


Context: Classical paradigms

Optimism in Face of Uncertainty: Hope for the best when unsure

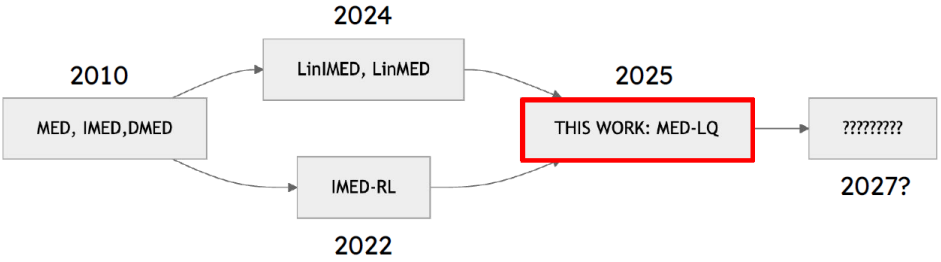


Thompson Sampling: Pick actions by sampling from posterior beliefs.

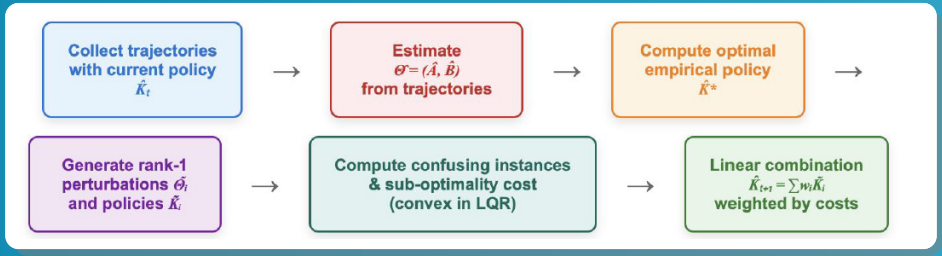


MED for Linear quadratic control

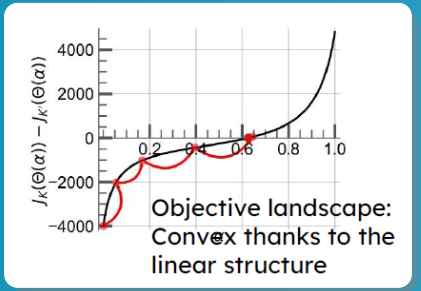
- Regret lower bound: Minimal amount of exploration needed to solve a problem.
- **M**inimum **E**mpirical **D**ivergence, is directly derived from this regret lower bound.



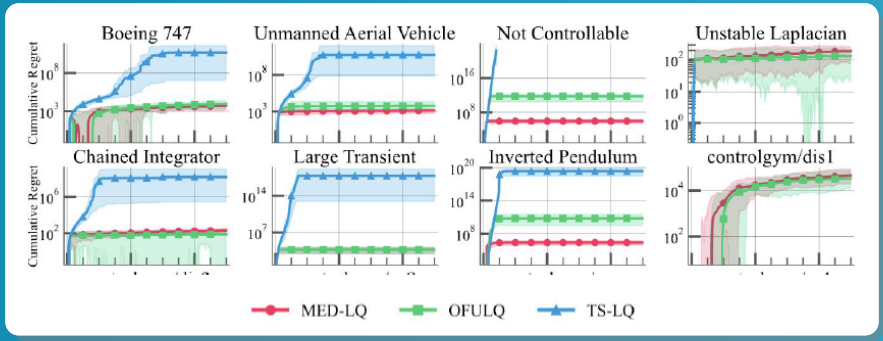
MED-LQ principle



Interpolation enables convex/concave optimization problem:



ControlGym environments testbed



Beyond?

Waris Radji and Odalric-Ambrym Maillard.
How Hard is it to Confuse a World Model?
arXiv preprint arXiv:2510.21232, 2025.

OUTLINE

1] The most confusing paradigm.

2] IMED-RL

3] Extensions

4] Sim-to-real gap in Exp. Sciences

Real-life Experimental Science is challenging

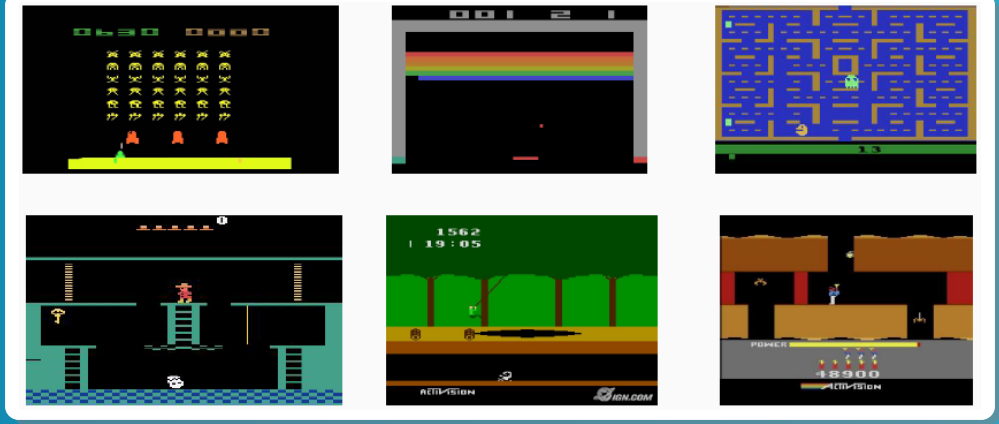
Reinforcement learning: learning behavior from trial and error

In practice:

- Requires unsustainable number of trials
- Real world cost for failed trials

Experimental Science: medicine, physics, agriculture, ecology, urban planning, etc.

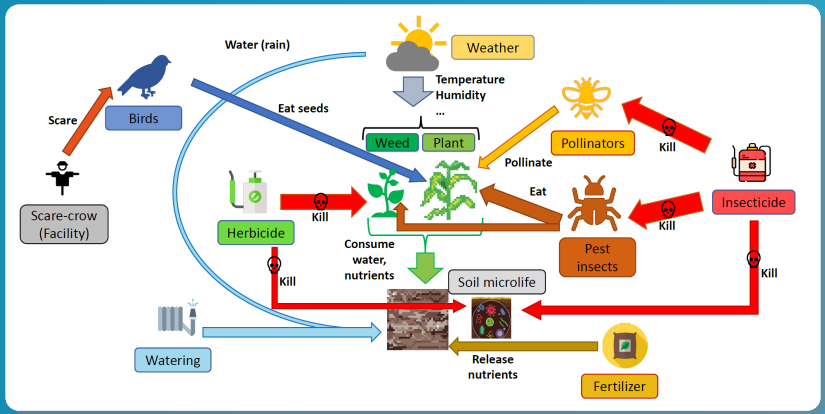
Atari environments



AgroSystems environments

AgroEcoGym framework

- ❖ Several **Entities** interact modeled by phy/chi/bio- **Processes**.
- ☒ **Huge** amount of processes described in literature.
- ☒ **Risk-averse**, **Non-parametric**, **Batch interaction**, **Autonomous**, etc.



RL for AgroEcosystems

Risk-averse, Non-parametric

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard.
Optimal thompson sampling strategies for support-aware cvar bandits.
In *International Conference on Machine Learning*, pages 716–726. PMLR, 2021.

Dorian Baudry, Emilie Kaufmann, and Odalric-Ambrym Maillard.
Sub-sampling for efficient non-parametric bandit exploration.
Advances in Neural Information Processing Systems, 33:5468–5478, 2020.

Sumit Vashishtha and Odalric-Ambrym Maillard.
Leveraging priors on distribution functions for multi-arm bandits
Reinforcement Learning Journal, 2025

Batch, Autonomous dynamic

- Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, pages 5065–5073. PMLR, 2021.
- Orlane Rossini, Meritxell Vinyals, Alice Cleynen, Benoîte de Saporta, and Régis Sabbadin. Bayes-adaptive impulse control of piecewise-deterministic markov processes. 2025.

RL for AgroEcosystems

RL for agriculture


Romain Gautron, Odalric-Ambrym Maillard, Philippe Preux, Marc Corbeels, and Régis Sabbadin.
Reinforcement learning for crop management support: Review, prospects and challenges.
Computers and Electronics in Agriculture, 200:107182, 2022.


Romain Gautron, Emilio J Padrón, Philippe Preux, Julien Bigot, Odalric-Ambrym Maillard, and David Emukpere.
Gym-dssat: a crop model turned into a reinforcement learning environment.
arXiv e-prints, pages arXiv–2207, 2022.


Odalric-Ambrym Maillard, Timothée Mathieu, and Debabrota Basu.
Farm-gym: A modular reinforcement learning platform for stochastic agronomic games.
In *2nd AAAI Workshop on AI for Agriculture and Food Systems*, 2023.

OUTLINE

1] The most confusing paradigm.

2] IMED-RL

3] Extensions

4] Sim-to-real gap in Exp. Sciences

Self-check

- ☑ The most confusing instance paradigm
- ☑ Lower bounds for Bandits and Ergodic MDPs.
- ☑ IMED-RL strategy and optimality
- ☑ Extensions to Linear quadratic systems
- ☑ Apply this principle to other setups?

MERCI