Regret minimization in Average Reward Reinforcement Learning
II. ALGORITHMS

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

Master MVA

Take home message

- ☑ Exploration-Exploitation in MDPs
- ☑ Optimistic principle : UCB for MDPs is UCRL
- ☑ Weissman confidence bounds on **p**
- ☑ Building blocks of UCRL: Extended MDP , episodes .
- ☑ Extended Value Iteration.
- ☑ Stopping criterion for EVI
- ☑ Bayesian principle
- ☑ Implementation, python library.

Collaborators




References


Thomas Jaksch, Ronald Ortner, and Peter Auer.
Near-optimal regret bounds for reinforcement learning.
Journal of Machine Learning Research, 11:1563–1600, 2010.


Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric.
Improved analysis of ucrl2 with empirical bernstein inequality.
arXiv preprint arXiv:2007.05456, 2020.


Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi.
Tightening exploration in upper confidence reinforcement learning.
In *International Conference on Machine Learning*, pages 1056–1066. PMLR, 2020.

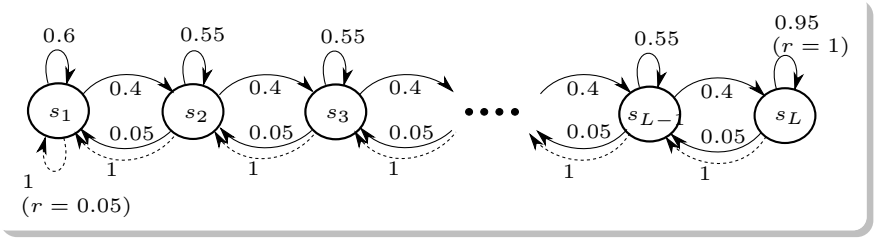
From known to unknown MDPs

☑ We have a proper value iteration procedure valid for the average-value criterion.
☒ Still requires full knowledge of the MDP.

❖ What to do when P, R are unknown ?

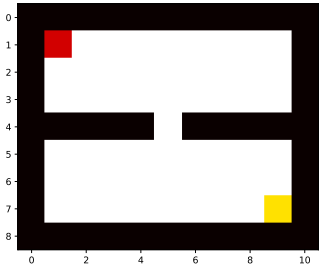
Example: Riverswim MDP

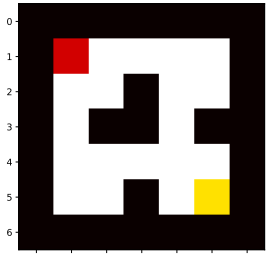
$\mathcal{S} = \{1, 2, \ldots, L\}$, $\mathcal{A} = \{\text{left,right}\}$, $p = 0.4$ to reach right when go right.



Example: Grid-world MDP

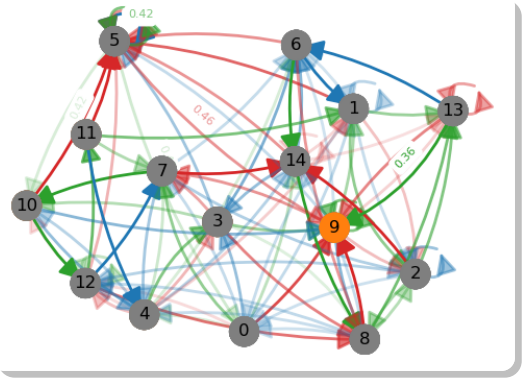
Frozen-lake (stochastic) type mazes.





Example: Randomly Generated MDP

Generic Commmunicating MDP with stochastic transitions.



OUTLINE

1] ε-Greedy policy

2] Optimistic principle

3] ...

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_t = \begin{cases} \arg\max_a Q_{\theta_t}(s_t, a) & \text{w.p. } 1-\epsilon \\ \mathcal{U}(\mathcal{A}) & \text{otherwise} \end{cases}$$

- Q-learning update

$$\theta_{t+1} = (1-\alpha_t)\theta_t + \alpha_t\big(r_t + \gamma \max_{a'} Q_{\theta_t}(s_{t+1}, a') - Q_{\theta_t}(s_t, a)\big)\nabla_\theta Q_{\theta_t}(s_t, a)$$

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_t = \begin{cases} \arg\max_a Q_{\theta_t}(s_t, a) & \text{w.p. } 1-\epsilon \\ \mathcal{U}(\mathcal{A}) & \text{otherwise} \end{cases}$$

- Q-learning update

$$\theta_{t+1} = (1-\alpha_t)\theta_t + \alpha_t\big(r_t + \gamma \max_{a'} Q_{\theta_t}(s_{t+1}, a') - Q_{\theta_t}(s_t, a)\big)\nabla_\theta Q_{\theta_t}(s_t, a)$$

👎 The exploration strategy relies on biased estimates Q_{θ_t}

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_t = \begin{cases} \arg\max_a Q_{\theta_t}(s_t, a) & \text{w.p. } 1-\epsilon \\ \mathcal{U}(\mathcal{A}) & \text{otherwise} \end{cases}$$

- Q-learning update

$$\theta_{t+1} = (1-\alpha_t)\theta_t + \alpha_t\big(r_t + \gamma \max_{a'} Q_{\theta_t}(s_{t+1}, a') - Q_{\theta_t}(s_t, a)\big)\nabla_\theta Q_{\theta_t}(s_t, a)$$

- The exploration strategy relies on biased estimates Q_{θ_t}
- Samples are used once

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_t = \begin{cases} \arg\max_a Q_{\theta_t}(s_t, a) & \text{w.p. } 1-\epsilon \\ \mathcal{U}(\mathcal{A}) & \text{otherwise} \end{cases}$$

- Q-learning update

$$\theta_{t+1} = (1-\alpha_t)\theta_t + \alpha_t\big(r_t + \gamma \max_{a'} Q_{\theta_t}(s_{t+1}, a') - Q_{\theta_t}(s_t, a)\big)\nabla_\theta Q_{\theta_t}(s_t, a)$$

- The exploration strategy relies on **biased** estimates Q_{θ_t}
- Samples are used **once**
- **Dithering effect:** exploration is not effective in covering the state space

What is Wrong with Q-learning with ϵ-greedy?

- ϵ-greedy strategy

$$a_t = \begin{cases} \arg\max_a Q_{\theta_t}(s_t, a) & \text{w.p. } 1-\epsilon \\ \mathcal{U}(\mathcal{A}) & \text{otherwise} \end{cases}$$

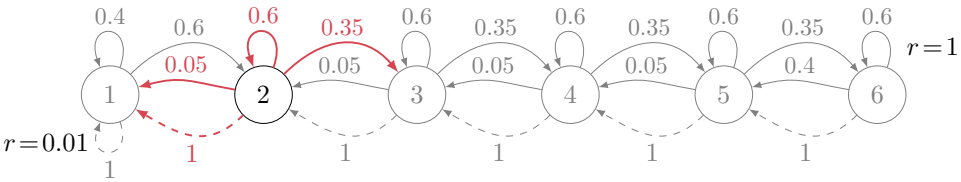
- Q-learning update

$$\theta_{t+1} = (1-\alpha_t)\theta_t + \alpha_t\big(r_t + \gamma \max_{a'} Q_{\theta_t}(s_{t+1}, a') - Q_{\theta_t}(s_t, a)\big)\nabla_\theta Q_{\theta_t}(s_t, a)$$

- The exploration strategy relies on **biased** estimates Q_{θ_t}
- Samples are used **once**
- **Dithering effect:** exploration is not effective in covering the state space
- **Policy shift:** the policy changes at each step

River Swim: Markov Decision Processes

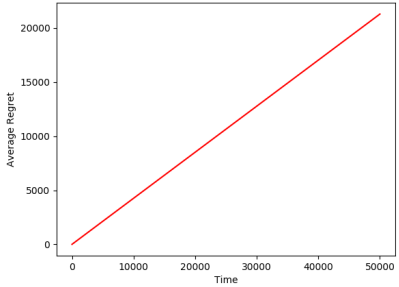
Strehl and Littman [2008]



- $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \{L, R\}$
- $\pi_L(s) = L$, $\pi_R(s) = R$

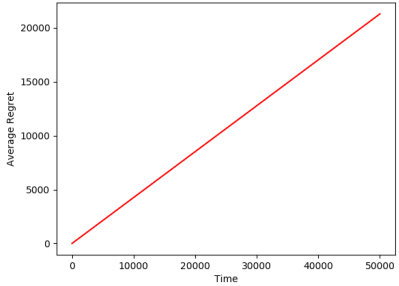
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$



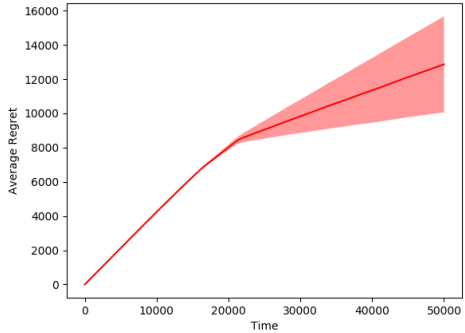
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$



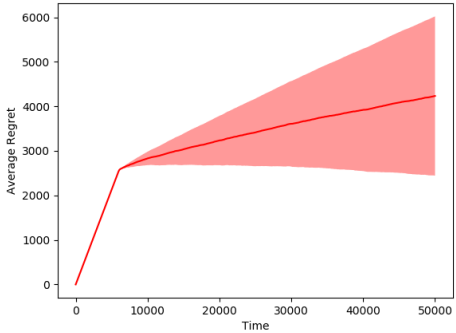
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \frac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$



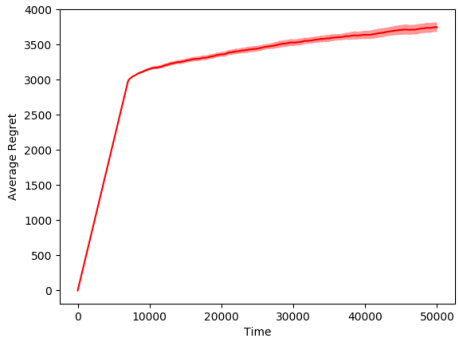
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \dfrac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$
- $\epsilon_t = \begin{cases} 1.0 & t < 6000 \\ \dfrac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$



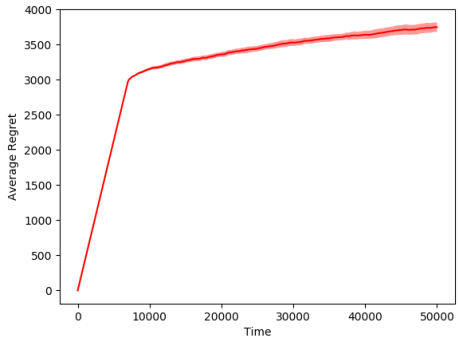
River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \frac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$
- $\epsilon_t = \begin{cases} 1.0 & t < 6000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$
- $\epsilon_t = \begin{cases} 1.0 & t < 7000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$



River Swim: Q-learning w\ ϵ-greedy Exploration

- $\epsilon_t = 1.0$
- $\epsilon_t = 0.5$
- $\epsilon_t = \frac{\epsilon_0}{(N(s_t) - 1000)^{2/3}}$
- $\epsilon_t = \begin{cases} 1.0 & t < 6000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$
- $\epsilon_t = \begin{cases} 1.0 & t < 7000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}} & \text{otherwise} \end{cases}$

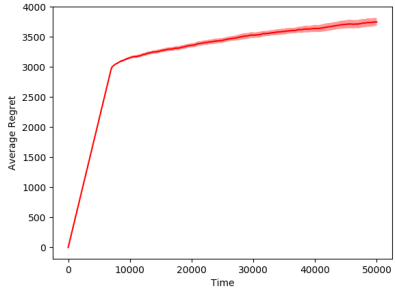


Tuning the ϵ schedule is **difficult and problem dependent**

River Swim: Q-learning w\ ϵ-greedy Exploration

Main drawbacks of Q-learning with ϵ-greedy*

- Q-learning is *model-free*
  - 👎 Inefficient *use* of samples
- ϵ-greedy performs *undirected* exploration
  - 👎 *Non-informative* samples

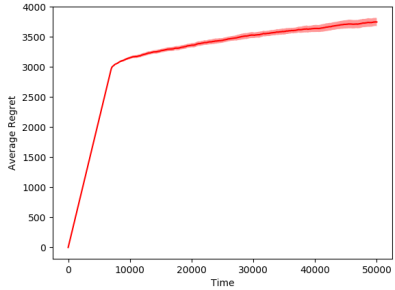


*All of this can be said for large majority for model-free undirected exploration methods

River Swim: Q-learning w\ ϵ-greedy Exploration

Main drawbacks of Q-learning with ϵ-greedy*

- Q-learning is *model-free*
  - 👎 Inefficient *use* of samples
- ϵ-greedy performs *undirected* exploration
  - 👎 *Non-informative* samples



Model-based uncertainty-driven exploration-exploitation

*All of this can be said for large majority for model-free undirected exploration methods

OUTLINE

1] ε-Greedy policy

2] Optimistic principle

3] Bayesian principle

4] ...

Upper confidence Reinforcement Learning

❖ Inspired from UCB for multi-armed bandits:

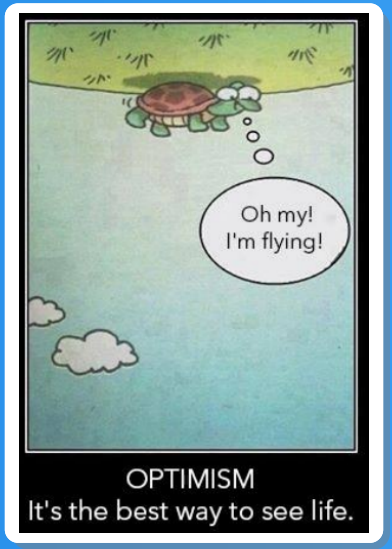
$$a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \max_{\mathbf{m} \in \mathcal{C}_t(a)} \mathbf{m} = \widehat{\mathbf{m}}_t(a) + \sqrt{\frac{2\sigma^2 \log(t^3)}{N_t(a)}}$$

❖ UCRL (Jaksch et al., 2010): A model-based algorithm for undiscounted RL implementing the principle of optimism in the face of uncertainty.

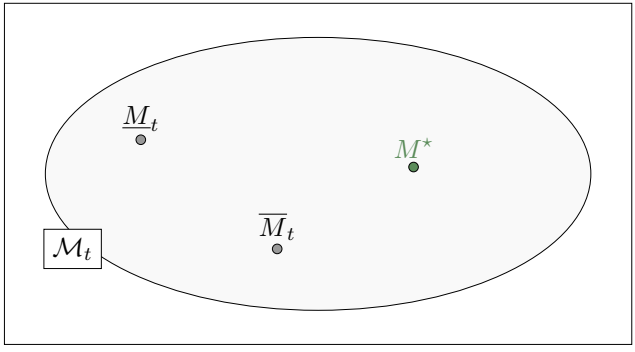
Maintains a set of plausible MDPs (models) by defining high-probability confidence sets for $\mathbf{m}$ and $\mathbf{p}$

Chooses an optimistic model (among models) and an optimistic policy leading to the highest average-reward.

The Optimism principle: Intuition

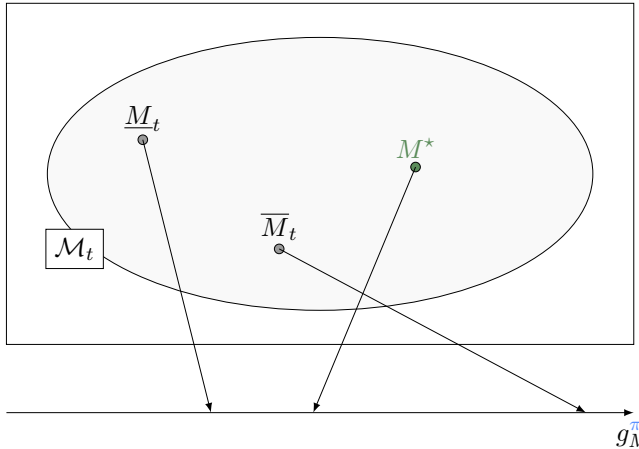


Bounded Parameter MDP: Optimism



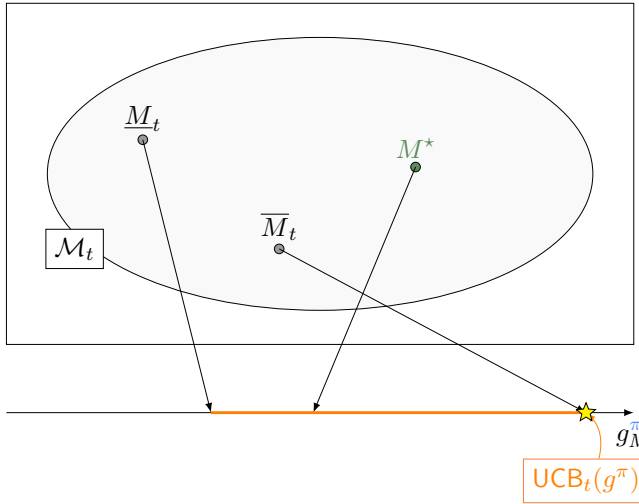
Fix a *policy* π

Bounded Parameter MDP: Optimism



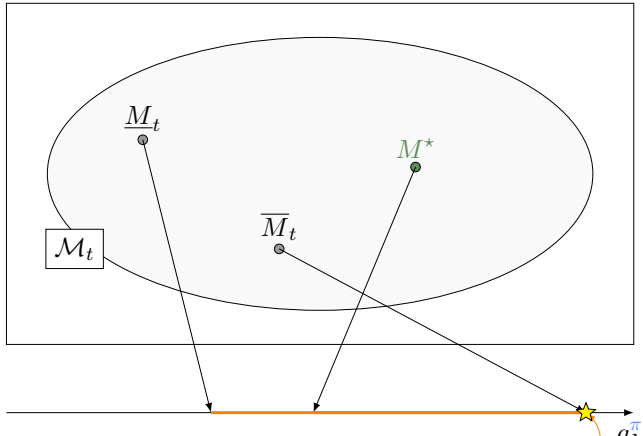
Fix a *policy* π

Bounded Parameter MDP: Optimism



Fix a *policy* π

Bounded Parameter MDP: Optimism



Optimism: $\mathsf{UCB}_t(g^\pi) = \max_{M \in \mathcal{M}_t} g^\pi_M \geq g^\pi_{M^\star}$

$\mathsf{UCB}_t(g^\pi)$

Fix a *policy* π

UCRL principle

UCRL maintains a set of plausible MDPs ($\widetilde{\mathbf{m}}$ denotes the mean of $\widetilde{\mathbf{r}}$)

$$\mathcal{M}_{t,\delta} = \left\{ \widetilde{\mathbf{M}} = (\mathcal{S}, \mathcal{A}, \widetilde{\mathbf{p}}, \widetilde{\mathbf{r}}) : \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \widetilde{\mathbf{p}}(\cdot|s,a) \in \mathcal{C}_{t,\delta}^{\texttt{UCRL}}(s,a) \text{ and } \widetilde{\mathbf{m}}(s,a) \in c_{t,\delta}^{\texttt{UCRL}}(s,a) \right\}.$$

UCRL computes an optimistic policy

$$\overline{\pi}_t^+ = \operatorname*{argmax}_{\pi} \max_{\widetilde{\mathbf{M}} \in \mathcal{M}_{t,\delta}} \mathbf{g}_{\widetilde{\mathbf{M}}}(\pi)$$

UCRL executes $\overline{\pi}_t^+$ until $\mathcal{M}_{t,\delta}$ has changed "enough"

The Optimism Principle: Intuition

Exploration vs. Exploitation

Optimism in Face of Uncertainty

When you are uncertain, consider the **best possible world (reward-wise)**

If the best possible world is **correct**

$\Longrightarrow$ **no regret**

Exploitation

If the best possible world is **wrong**

$\Longrightarrow$ **learn useful information**

Exploration

The Optimism Principle: Intuition

Exploration vs. Exploitation

Optimism in **gain**

Optimism in Face of Uncertainty

When you are uncertain, consider the **best possible world (reward-wise)**

If the best possible world is **correct**

$\Longrightarrow$ **no regret**

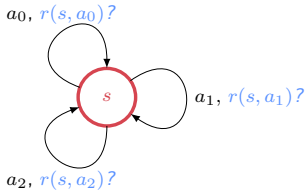
Exploitation

If the best possible world is **wrong**

$\Longrightarrow$ **learn useful information**

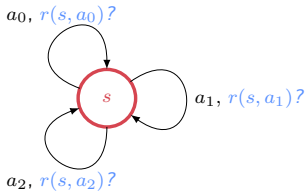
Exploration

Gain Optimism: Example



- Deterministic *policies*:
  - $\pi_0(s) = a_0$
  - $\pi_1(s) = a_1$
  - $\pi_2(s) = a_2$

Gain Optimism: Example

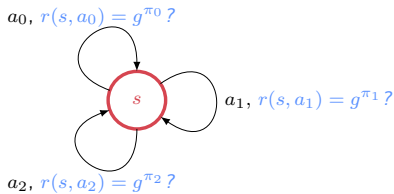


- Deterministic *policies*:
  - $\pi_0(s) = a_0$
  - $\pi_1(s) = a_1$
  - $\pi_2(s) = a_2$

- Optimism

$$\widetilde{\pi} = \arg\max_{\pi_i} \mathsf{UCB}(g^{\pi_i})$$

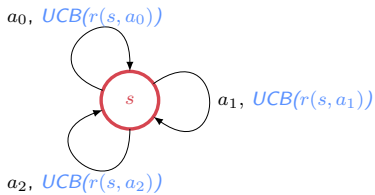
Gain Optimism: Example



- Deterministic *policies*:
  - $\pi_0(s) = a_0$
  - $\pi_1(s) = a_1$
  - $\pi_2(s) = a_2$
- Reward $r(s, a_i) =$ *gain* g^{π_i}
- Optimism

$$\widetilde{\pi} = \arg\max_{\pi_i} \mathsf{UCB}(g^{\pi_i})$$

Gain Optimism: Example



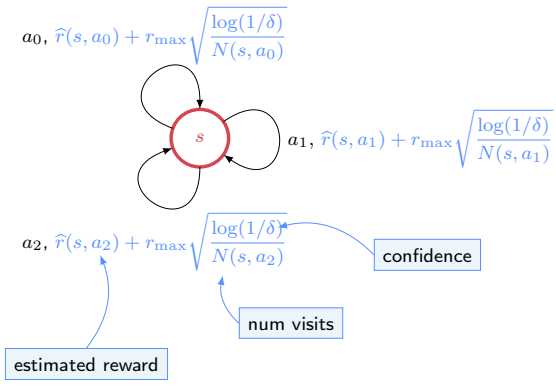
- Deterministic *policies*:
  - $\pi_0(s) = a_0$
  - $\pi_1(s) = a_1$
  - $\pi_2(s) = a_2$
- Reward $r(s, a_i)$ = *gain* g^{π_i}
- Upper confidence bound

$$\text{UCB}(g^{\pi_i}) = \text{UCB}(r(s, a_i))$$

- Optimism

$$\widetilde{\pi} = \arg\max_{\pi_i} \text{UCB}(g^{\pi_i})$$

Gain Optimism: Example



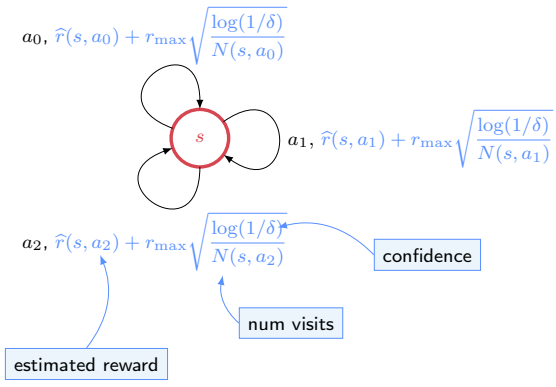
- Deterministic *policies*:
  - $\pi_0(s) = a_0$
  - $\pi_1(s) = a_1$
  - $\pi_2(s) = a_2$
- Reward $r(s, a_i)$ = *gain* g^{π_i}
- Upper confidence bound

$$\text{UCB}(g^{\pi_i}) = \text{UCB}(r(s, a_i))$$

- Optimism

$$\widetilde{\pi} = \arg\max_{\pi_i} \text{UCB}(g^{\pi_i})$$

Gain Optimism: Example



- Deterministic *policies*:
  - $\pi_0(s) = a_0$
  - $\pi_1(s) = a_1$
  - $\pi_2(s) = a_2$
- Reward $r(s, a_i)$ = *gain* g^{π_i}
- Upper confidence bound

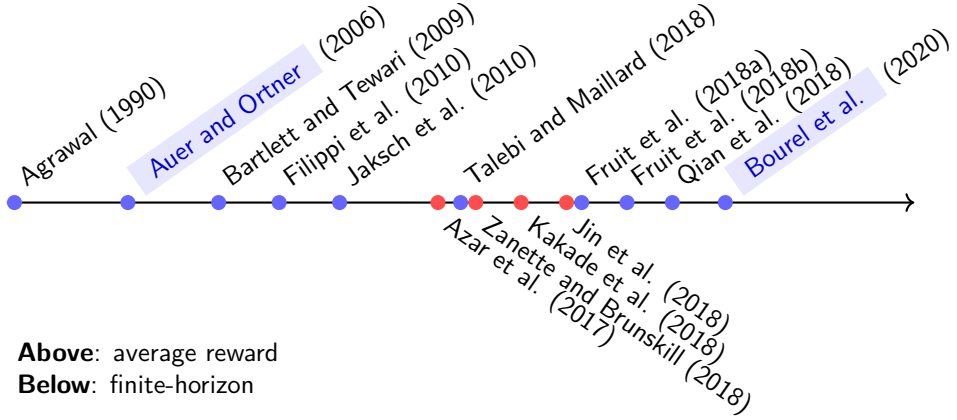
$$\text{UCB}(g^{\pi_i}) = \text{UCB}(r(s, a_i))$$

- Optimism

$$\widetilde{\pi} = \arg\max_{\pi_i} \text{UCB}(g^{\pi_i})$$

👉 UCB algorithm (Bandit)

History of Optimistic RL



Above: average reward
Below: finite-horizon

UCRL Variants

- Different assumptions: known $D(\mathbf{M})$? known horizon T?, etc.
- Different ways to build the confidence sets:
(Hoeffding-wise + Union bounds) vs (Bregman-wise + Time uniform bounds.)
- Different ways to handle support of $\mathbf{p}$
- Different criterion to stop a policy and update computations.

Table of contents

Main questions

❖ Upper Confidence Bounds:
Construct UCB($\mathbf{g}_\pi$) with unknown dynamics $\mathbf{p}$?

❖ Efficient computation:
$\underset{\pi}{\operatorname{argmax}}\,\mathrm{UCB}(\mathbf{g}_\pi)$ over exponentially many policies ?

❖ Efficient "navigation":
exploration as in bandits ?

UCRL2 Confidence sets

❖ Hoeffging/Weissman bounds plus union bound argument

$$c_{t,\delta}^{\texttt{UCRL}}(s,a) = \left\{ \mathbf{m}' \in [0,1] : |\widehat{\mathbf{m}}_t(s,a) - \mathbf{m}'| \leq \sqrt{\frac{3.5 \log(\frac{2SAt}{\delta})}{N_t(s,a)}} \right\},$$

$$\mathcal{C}_{t,\delta}^{\texttt{UCRL}}(s,a) = \left\{ \mathbf{p}' \in \Delta_{\mathcal{S}} : \|\widehat{\mathbf{p}}_t(\cdot|s,a) - \mathbf{p}'\|_1 \leqslant \sqrt{\frac{14S \log(\frac{2At}{\delta})}{N_t(s,a)}} \right\}.$$

❖ Empirical estimates of transition probabilities and rewards:

$$\widehat{\mathbf{m}}_t(s,a) = \frac{\sum_{t'=0}^{t-1} r_{t'} \mathbb{I}\{s_{t'} = s, a_{t'} = a\}}{\max\{N_t(s,a), 1\}} \qquad \widehat{\mathbf{p}}_t(s'|s,a) = \frac{N_t(s,a,s')}{\max\{N_t(s,a), 1\}}$$

$N_t(s,a,s')$: number of visits, up to time t, to (s,a) followed by s'.

Refined concentration inequalities

Laplace-Hoeffding inequality for $[0,1]$-bounded r.v. and stopping time τ:

$$\mathbb{P}\left(|\widehat{\mathbf{m}}_\tau - \mathbf{m}| \geqslant \sqrt{\frac{(1+\frac{1}{\tau})\log(2\sqrt{\tau+1}/\delta)}{2\tau}}\right) \leqslant \delta\,.$$

Weissman inequality for probability distribution on discrete set $\mathcal{S}$:

$$\mathbb{P}\left(\|\widehat{p}_\tau - p\|_1 \geqslant \sqrt{\frac{2(1+\frac{1}{\tau})\log\left(\sqrt{\tau+1}\frac{2^{|\mathcal{S}|}-2}{\delta}\right)}{\tau}}\right) \leqslant \delta\,.$$

Proof using method of types for discrete measures:

$$\mathbb{P}\left(\|\widehat{p}_\tau - p\|_1 \geqslant \varepsilon\right) \leqslant \sum_{B\subset\mathcal{S}} \mathbb{P}\left(p_\tau(B) - p(B) \geqslant \frac{1}{2}\varepsilon\right).$$

- $2^{|\mathcal{S}|}-2$ non trivial sets.
- Each term: Bernoulli concentration, handled e.g. by Hoeffding.

Table of contents

EVI principle

❖ UCRL computes an optimistic policy

$$\overline{\pi}_t^+ = \operatorname*{argmax}_{\pi} \max_{\widetilde{\mathbf{M}} \in \mathcal{M}_{t,\delta}} \mathbf{g}_{\widetilde{\mathbf{M}}}(\pi)$$

❖ In pratice, this is approximated using Extended Value Iteration (EVI): build a near-optimal policy π_t^+ and MDP $\widetilde{\mathbf{M}}_t$ such that

$$\mathbf{g}_{\pi_t^+}^{\widetilde{\mathbf{M}}_t} \geqslant \max_{\pi} \max\{\mathbf{g}_{\mathbf{M}}(\pi) : \mathbf{M} \in \mathcal{M}_{t,\delta}\} - \varepsilon_t \text{ where } \varepsilon_t = \tfrac{1}{\sqrt{t}}.$$

Extended Value Iteration

Value iteration on an extended MDP

Value iteration for unknown $\mathbf{m}, \mathbf{p}$:

$$u_{n+1}(s) = \max_{(a,\mathbf{m},\mathbf{p}) \in \mathcal{A} \times C_{t,\delta}^{\mathrm{UCRL}}(s,a) \times \mathcal{C}_{t,\delta}^{\mathrm{UCRL}}(s,a)} \left\{ \mathbf{m} + \mathbf{p} u_n \right\} \quad \text{(Extended actions)}$$

$$= \max_{a \in \mathcal{A}} \left\{ \max_{\mathbf{m} \in C_{t,\delta}^{\mathrm{UCRL}}(s,a)} \mathbf{m} + \max_{\mathbf{p} \in \mathcal{C}_{t,\delta}^{\mathrm{UCRL}}(s,a)} \mathbf{p} u_n \right\} \quad \text{(Optimistic model)}$$

$$\pi_{n+1} = \text{Greedy with respect to } v_n.$$

Extended Value Iteration

Require: ε_t

Let $u_0 \equiv 0, u_{-1} \equiv -\infty,\ n = 0$

while $\mathbb{S}(u_n - u_{n-1}) > \varepsilon_t$ **do**

Compute $\begin{cases} \mathbf{m}^+ : s, a \mapsto \max\{\mathbf{m}' : \mathbf{m}' \in c^{\mathtt{UCRL}_{t,\delta}(s,a)}\} \\ \mathbf{p}_n^+ : s, a \mapsto \operatorname{argmax}\{\mathbf{p}' u_n : \mathbf{p}' \in \mathcal{C}^{\mathtt{UCRL}_{t,\delta}(s,a)}\} \end{cases}$

Update $\begin{cases} u_{n+1}(s) = \max\{\mathbf{m}^+(s,a) + (\mathbf{p}_n^+ u_n)(s,a) \colon a \in \mathcal{A}\} \\ \pi_{n+1}^+(s) \in \operatorname{Argmax}\{\mathbf{m}^+(s,a) + (\mathbf{p}_n^+ u_n)(s,a) \colon a \in \mathcal{A}\} \end{cases}$

$n = n + 1$

end while

Extended MDP

[Strehl and Littman, 2008, Jaksch et al., 2010]

Theorem (Bounded parameter MDP $\iff$ Extended MDP)

Let $\mathcal{M}_t^+ := \langle \mathcal{S}, \mathcal{A}_t^+, r^+, p^+ \rangle$ *be an extended MDP such that*

$$\mathcal{A}_t^+(s) = \mathcal{A}(s) \times B_t^r(s,a) \times B_t^p(s,a)$$

with $a^+ = (a, r, p) \in \mathcal{A}_t^+(s),\ r^+(s,a^+) = r,\ p^+(\cdot|s,a^+) = p.$

Continuous **compact** action space

Then the optimal gain of $\mathcal{M}_t^+$ *satisfies*

$$g^*_{\mathcal{M}_t^+} := \max_\pi \left\{ \max_{M \in \mathcal{M}_t} g_M^\pi \right\}$$

Let $\pi_t^+ = \arg\max_\pi g^\pi_{\mathcal{M}_t^+}$, *then*

$$\pi_t = \arg\max_\pi \left\{ \max_{M \in \mathcal{M}_t} g_M^\pi \right\} \text{ s.t. } \pi_t(s) = \pi_t^+(s)[a]$$

Extended MDP

[Strehl and Littman, 2008, Jaksch et al., 2010]

Theorem (Bounded parameter MDP $\iff$ Extended MDP)

Let $\mathcal{M}_t^+ := \langle \mathcal{S}, \mathcal{A}_t^+, r^+, p^+ \rangle$ *be an extended MDP such that*

$$\mathcal{A}_t^+(s) = \mathcal{A}(s) \times B_t^r(s,a) \times B_t^p(s,a)$$

with $a^+ =$ *Abuse of notation*: $\mathcal{M}_t$ denotes the extended MDP

ompact
pace

Then the optimal gain of $\mathcal{M}_t^+$ *satisfies*

$$g^*_{\mathcal{M}_t^+} := \max_\pi \left\{ \max_{M \in \mathcal{M}_t} g^\pi_M \right\}$$

Let $\pi_t^+ = \arg\max_\pi g^\pi_{\mathcal{M}_t^+}$, *then*

$$\pi_t = \arg\max_\pi \left\{ \max_{M \in \mathcal{M}_t} g^\pi_M \right\} \text{ s.t. } \pi_t(s) = \pi_t^+(s)[a]$$

Efficient Exploration

❖ UCRL maintains a set of plausible MDPs ($\widetilde{\mathbf{m}}$ denotes the mean of $\widetilde{\mathbf{r}}$)

❖ UCRL computes an approximate-optimistic policy $\overline{\pi}_t^+$

❖ Episodes : Only recompute policies at time $(t_k)_k$ (hence $\pi_k^+ := \pi_{t_k}^+$) s.t. $t_1 = 1$, and

$$\forall k > 1,\ t_k = \min\left\{t > t_{k-1} : \max_{s,a} \frac{n_{t_{k-1}:t}(s,a)}{N_{t_{k-1}}(s,a)} \geqslant 1\right\},$$

where $n_{t_1:t_2}(s,a)$ is the number of observations of (s,a) between time t_1 and t_2.

❖ Keep the same policy π_k^+ for some time, instead of changing immediately. (Same idea in Deep RL but this is 2010!)

Table of contents

Regret upper bound of UCRL

Regret of UCRL2 (Jaksch et al., 2010)

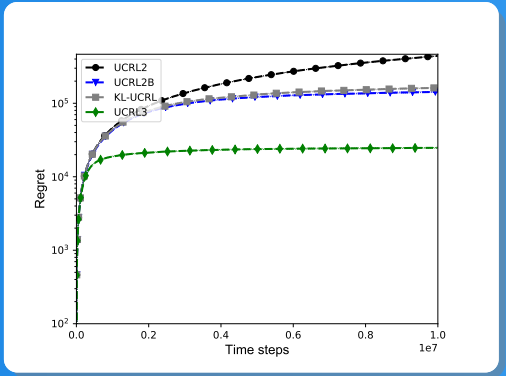
For any communicating MDP with S states, A actions, and diameter D, UCRL satisfies

$$\mathfrak{R}(T) \leq 34\mathbf{m}_{\max} DS\sqrt{AT\log(T/\delta)} \quad \text{w.p. at least } 1-\delta.$$

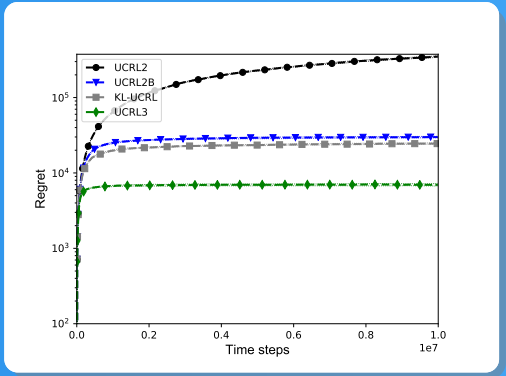
Minimax lower bound (Jaksch et al., 2010): $\Omega(\sqrt{DSAT})$

❖ Vanilla UCRL does not perform empirically great despite their strong regret guarantees (e.g. loose confidence bounds)

Results for UCRL and variants



2-room MDP



4-room MDP

UCRL improvements

☑ We can refine conf. bounds replacing $\|.\|_1$ bounds on distributions p, with KL balls, or term-wise intervals on each $p(s)$.

☑ We can refine the conf. bounds of UCRL, using time-uniform concentration bounds.

☑ We can suggest a less conservative alternative navigation criterion

☑ We can revisit the way sparse transitions are handled (important in practice).

☑ ...

Refined Confidence Bounds

- UCRL2 with *Bernstein bounds* (instead of Hoeffding/Weissman):

$$R(T, M^\star, \text{UCRL2B}) = \mathcal{O}\left(\sqrt{D\Gamma SAT \log\left(\frac{T}{\delta}\right) \log(T)}\right)$$

- 👎 Still not matching the lower bound!
- 👍 For most MPDs: $\Gamma \ll S$

Refined Confidence Bounds

- UCRL2 with *Bernstein bounds* (instead of Hoeffding/Weissman):

$$R(T, M^\star, \text{UCRL2B}) = \mathcal{O}\left(\sqrt{D\Gamma SAT \log\left(\frac{T}{\delta}\right) \log(T)}\right)$$

  - 👎 Still not matching the lower bound!
  - 👍 For most MPDs: $\Gamma \ll S$

- *Kullback-Leibler* UCRL [Filippi et al., 2010, Talebi and Maillard, 2018]:

$$R(T, M^\star, \text{UCRL-KL}) = \mathcal{O}\left(\sqrt{\underbrace{\sum_{s,a} \mathbb{V}_{X \sim p^\star(\cdot|s,a)}\left(h^\star_{M^\star}(X)\right)}_{\leq D^2 SA} ST \log\left(\frac{T}{\delta}\right)} + D\sqrt{T}\right)$$

  - 👎 Only for ergodic MPDs!

Infinite Diameter (weakly communicating MDPs)

- *Known* bound on the optimal bias span $C \geq \mathsf{sp}(h^\star_{M^\star})$
  [Bartlett and Tewari, 2009, Fruit et al., 2018b]

$$R(T, M^\star, \mathsf{SCAL}) = \mathcal{O}\left(\sqrt{C\Gamma SAT \log\left(\frac{T}{\delta}\right)\log(T)}\right)$$

👎 Requires prior knowledge!

Infinite Diameter (weakly communicating MDPs)

- *Known* bound on the optimal bias span $C \geq \mathsf{sp}(h^\star_{M^\star})$
  [Bartlett and Tewari, 2009, Fruit et al., 2018b]

$$R(T, M^\star, \mathsf{SCAL}) = \mathcal{O}\left(\sqrt{C\Gamma SAT \log\left(\frac{T}{\delta}\right)\log(T)}\right)$$

  👎 Requires prior knowledge!

- No prior knowledge: TUCRL [Fruit et al., 2018a]:

$$R(T, M^\star, \mathsf{SCAL}) = \mathcal{O}\left(\sqrt{D_{\mathsf{com}} S_{\mathsf{com}} \Gamma AT \log\left(\frac{T}{\delta}\right)\log(T)}\right)$$

  👎 Never achieves *logarithmic* regret! Intrinsic limitation of the setting!

Table of contents

UCRL3

Our main contribution is UCRL3, a new algorithm for average-reward RL.

UCRL3 is a variant of UCRL2, combining the following key elements:

- Tight and element-wise confidence intervals for transition function p
  - Intersection of time-uniform Bernstein and sub-Gaussian Bernoulli concentration for each $p(s'|s,a)$
- A modified planning algorithm, called EVI-NOSS, to compute a near-optimistic policy.

To simplify the presentation, we assume that μ is known.

UCRL3: Confidence Set for p

For each pair (s, a), define

$$\mathcal{C}_{t,\delta}(s,a) := \left\{ q \in \Delta_{\mathcal{S}} : q(s') \in \underbrace{C^1_{t,\delta}(s,a,s')}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s,a,s')}_{\text{sub-Gaussian}} \text{ for all } s' \right\}$$

UCRL3: Confidence Set for p

For each pair (s, a), define

$$\mathcal{C}_{t,\delta}(s,a) := \left\{ q \in \Delta_{\mathcal{S}} : q(s') \in \underbrace{C^1_{t,\delta}(s,a,s')}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s,a,s')}_{\text{sub-Gaussian}} \text{ for all } s' \right\}$$

- $C^1_{t,\delta}(s,a,s')$ is defined using Bernstein's concentration modified using **a peeling technique**.
- $C^2_{t,\delta}(s,a,s')$ is obtained by leveraging sub-Gaussianity of Bernoulli distributions combined with **the method of mixtures**.

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs:

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a) \text{ for all } (s,a) \Big\}$$

Lemma (Time-uniform confidence bounds)

For any MDP M with transition function p, for all $\delta \in (0,1)$, it holds

$$\mathbb{P}\big(\exists t \in \mathbb{N}, M \notin \mathcal{M}_{t,\delta}\big) \leq \delta.$$

UCRL3: Revisiting EVI

- To compute an optimistic policy (i.e., planning) in UCRL2 is done by EVI as a subroutine, which involves solving

$$\max \left\{ \sum_{x \in \mathcal{S}} p'(x) u_n(x) : p' \in \mathcal{C}_{t,\delta}(s,a) \right\}$$

  where u_n is a value function (at iteration n of EVI)
- EVI outputs a *conservative* policy (hence introducing unnecessary exploration), in particular when transition function p has a sparse support.
- UCRL3 remedies this issue by combining EVI with an adaptive support selection procedure.

UCRL3: Revisiting EVI

More specifically, at each iteration n of EVI:

- We first compute $\widetilde{S}_{s,a} \subset \mathcal{S}$, an approximation of the support of $p(\cdot|s,a)$, using NOSS (Algorithm 2 in the paper).
- Then, we solve

$$\max \left\{ \sum_{x \in \mathcal{S}} p'(x) u_n(x) : p' \in \mathcal{C}_{t,\delta}(s,a) \text{ and } \operatorname{supp}(p') = \widetilde{S}_{s,a} \right\}$$

This combined algorithm is called EVI-NOSS and outputs a near-optimistic policy.

For the complete pseudo-code of UCRL3, we refer to the paper.

UCRL3: Local Diameter

Definition (Local Diameter of State s)

Consider state $s \in \mathcal{S}$. For $s_1, s_2 \in \cup_{a \in \mathcal{A}} \text{supp}\big(p(\cdot|s,a)\big)$ with $s_1 \neq s_2$, let $T^\pi(s_1, s_2)$ denote the number of steps it takes to get to s_2 starting from s_1 and following policy π. Then, the local diameter of MDP M for s is defined as

$$D_s := \max_{s_1, s_2 \in \cup_a \text{supp}(p(\cdot|s,a))} \min_\pi \mathbb{E}[T^\pi(s_1, s_2)].$$

- D_s refines the (global) diameter (Jaksch et al., 2010).
- For all s, $D_s \leq D$, and for some states $D_s \ll D$.

UCRL3: Regret

Theorem (Regret of UCRL3)

With probability higher than $1-\delta$, uniformly over all $T \geq 3$, the regret under UCRL3 satisfies:

$$\mathfrak{R}(T) \leq \mathcal{O}\Big(\Big[\sqrt{\sum\nolimits_{s,a} \max(D_s^2 L_{s,a}, 1)} + D\Big]\sqrt{T\log(T/\delta)}\Big),$$

where $L_{s,a} := \big(\sum_{x\in\mathcal{S}} \sqrt{p(x|s,a)\big(1-p(x|s,a)\big)}\big)^2$ *denotes the local effective support of* (s,a).

UCRL3: Regret

Theorem (Regret of UCRL3)

With probability higher than $1-\delta$, uniformly over all $T \geq 3$, the regret under UCRL3 satisfies:

$$\mathfrak{R}(T) \leq \mathcal{O}\Big(\Big[\sqrt{\sum\nolimits_{s,a} \max(D_s^2 L_{s,a}, 1)} + D\Big]\sqrt{T\log(T/\delta)}\Big),$$

where $L_{s,a} := \big(\sum_{x\in\mathcal{S}} \sqrt{p(x|s,a)\big(1-p(x|s,a)\big)}\big)^2$ *denotes the local effective support of* (s,a).

Note that $L_{s,a} \leq K_{s,a} - 1$ (with $K_{s,a} := |\mathrm{supp}\big(p(\cdot|s,a)\big)|$). Hence,

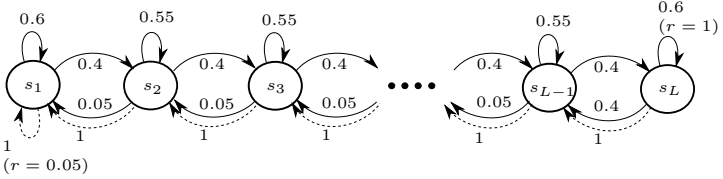
$$\mathfrak{R}(T) \leq \widetilde{\mathcal{O}}\Big(\Big[\sqrt{\sum\nolimits_{s,a} \max(D_s^2 K_{s,a}, 1)} + D\Big]\sqrt{T}\Big).$$

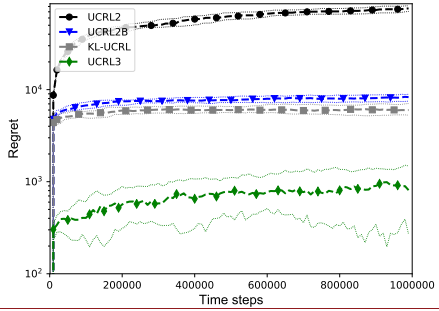
State-of-the-Art Regret Bounds

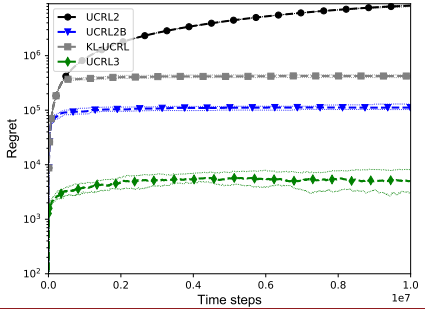
Algorithm	Regret bound
UCRL2 (Jaksch et al., 2010)	$\mathcal{O}\Big(DS\sqrt{AT\log(T/\delta)}\Big)$
KL-UCRL (Filippi et al., 2010)	$\mathcal{O}\Big(DS\sqrt{AT\log(\log(T)/\delta)}\Big)$
KL-UCRL (Talebi et al., 2018)	$\mathcal{O}\Big(\Big[D+\sqrt{S\sum_{s,a}\max(\mathbb{V}_{s,a},1)}\Big]\sqrt{T\log(\log(T)/\delta)}\Big)$
SCAL$^+$ (Qian et al., 2019)	$\mathcal{O}\Big(D\sqrt{\sum_{s,a}K_{s,a}T\log(T/\delta)}\Big)$
UCRL2B (Fruit et al., 2019)	$\mathcal{O}\Big(\sqrt{D\sum_{s,a}K_{s,a}T\log(T)\log(T/\delta)}\Big)$
UCRL3 **(This Paper)**	$\mathcal{O}\Big(\big(D+\sqrt{\sum_{s,a}\max(D_s^2L_{s,a},1)}\big)\sqrt{T\log(T/\delta)}\Big)$
Lower Bound (Jaksch et al., 2010)	$\Omega\big(\sqrt{DSAT}\big)$

Numerical Experiments

UCRL3 vs. existing algorithms in RiverSwim: $L=6$ (left) vs. $L=25$ (right)

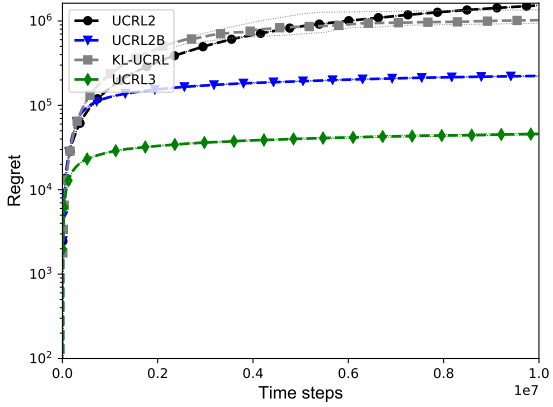






Numerical Experiments

UCRL3 vs. existing algorithms in a 100-state randomly generated MDP using Garnet (Bhatnagar et al., 2009)



Extensions

- We used a VI approach: what about PI, MPI instead?
- What if MDP transitions are not arbitrary but structured ?

Chowdhury, Gopalan, Maillard.
Reinforcement Learning in Parametric MDPs with Exponential Families.
AISTATS, 2021.

- What are generic lower bounds ?
- What about function approximation? large/continuous states space?

OUTLINE

1] ε-Greedy policy

2] Optimistic principle

3] Bayesian principle

4] Numerical experimentation

5] ...

Posterior sampling reinforcement learning

•From observations, build **posterior** on reward function (using e.g. Beta, Gaussian) $\pi^{\mathbf{r}}$ and on transitions $\pi^{\mathbf{p}}$ using Dirichlet.

Dirichlet $\pi^{\mathbf{p}}(s,a) = \mathrm{Dir}(n_t(s,a,s_1),\ldots,n_t(s,a,s_{|S|}))$

• Sample an MDP $\tilde{\mathbf{M}}$: $\forall s,a, \tilde{\mathbf{r}}(s,a) \sim \pi^{\mathbf{r}}(s,a), \tilde{\mathbf{p}}(\cdot|s,a) \sim \pi^{\mathbf{p}}(s,a)$

•Compute optimal policy $\tilde{\pi}^\star$ for $\tilde{\mathbf{M}}$ using e.g. Value-Iteration.

•Execute $\tilde{\pi}^\star$ until navigation criterion is met : e.g. nb of observations on one (s,a) as doubled.

Posterior sampling reinforcement learning

•From observations, build **posterior** on reward function (using e.g. Beta, Gaussian) $\pi^{\mathbf{r}}$ and on transitions $\pi^{\mathbf{p}}$ using Dirichlet.

Dirichlet $\quad \pi^{\mathbf{p}}(s,a) = \mathrm{Dir}(n_t(s,a,s_1),\dots,n_t(s,a,s_{|S|}))$

• Sample an MDP $\tilde{\mathbf{M}}$: $\quad \forall s,a, \tilde{\mathbf{r}}(s,a) \sim \pi^{\mathbf{r}}(s,a), \tilde{\mathbf{p}}(\cdot|s,a) \sim \pi^{\mathbf{p}}(s,a)$

•Compute optimal policy $\tilde{\pi}^\star$ for $\tilde{\mathbf{M}}$ using e.g. Value-Iteration.

•Execute $\tilde{\pi}^\star$ until navigation criterion is met : e.g. nb of observations on one (s,a) as doubled.

❖ **Careful:** Initial analysis wrong, due to lack of optimism . Requires correction.
(Osband, von Roy 2016) https://arxiv.org/pdf/1608.02731.pdf
(Agrawal, Jia 2017) https://dl.acm.org/doi/abs/10.5555/3294771.3294884
If $N_t(s,a) < \eta$, $\tilde{\mathbf{p}}(\cdot|s,a)$ is chosen optimistically.

OUTLINE

Python library

Github https://github.com/StatisticalRL

- Environments https://github.com/StatisticalRL/environments

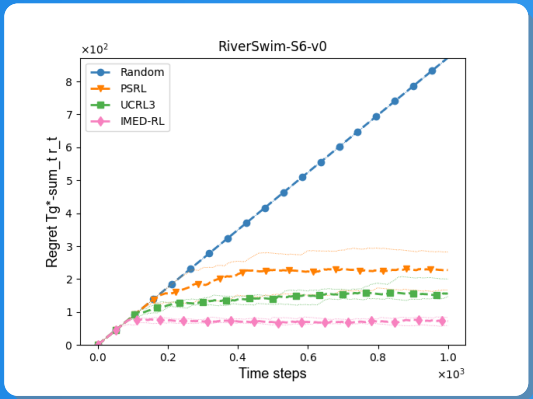
MDPs:
random-rich
ergodic-random-rich
random-12
random-small
random-small-sparse
random-100
three-state
nasty
river-swim-6
ergo-river-swim-6
ergo-river-swim-25
river-swim-25

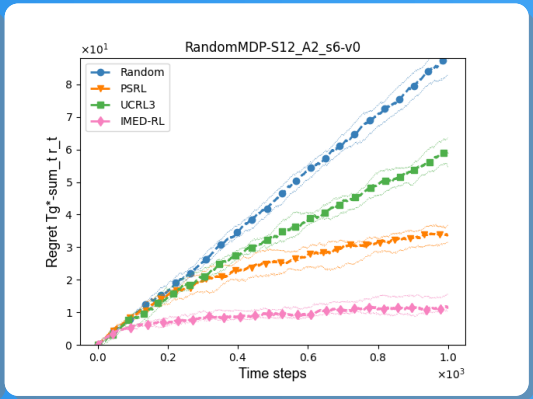
MABS:
mab-bernoulli
mab-gaussian
mab-binomial
mab-batch-quantized

GRIDWORLD MDPs:
grid-random-1616
grid-random-1212
grid-random-88
grid-2-room
grid-4-room

Showtime

Python `StatisticalRL.experiments.src/genericxp.py`





OUTLINE

Self-check

- ☑ Exploration-Exploitation in MDPs
- ☑ Optimistic principle : UCB for MDPs is UCRL
- ☑ Weissman confidence bounds on **p**
- ☑ Building blocks of UCRL: Extended MDP , episodes .
- ☑ Extended Value Iteration.
- ☑ Stopping criterion for EVI
- ☑ Differences between UCRL2 and UCRL3?
- ☑ Bayesian principle (cautious)
- ☑ Implementation, python library.

MERCI