Tightening Exploration in Upper Confidence Reinforcement Learning

Hippolyte Bourel (Inria)

Odalric-Ambrym Maillard (Inria)

Mohammad Sadegh Talebi (University of Copenhagen)

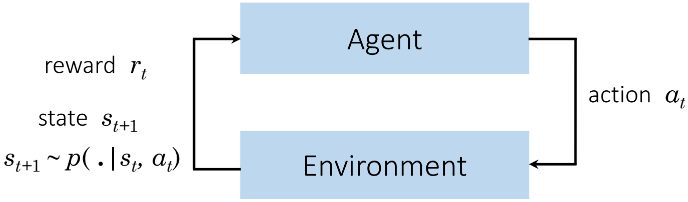
Undiscounted RL: MDP Model

We consider reinforcement learning (RL), where the environment is modeled as an undiscounted Markov Decision Process (MDP).

Undiscounted MDP $M = (\mathcal{S}, \mathcal{A}, p, \mu)$:

- State-space $\mathcal{S}$ with cardinality S
- Action-space $\mathcal{A}$ with cardinality A
- Transition function p: Selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$ leads to a transition to s' with probability $p(s'|s, a)$.
- Reward function μ: Selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$ gives $r(s, a)$ with mean $\mu(s, a)$.

Undiscounted RL: MDP Model



p and μ are unknown, and the goal is to maximize $\sum_{t=1}^{T} r_t$.

We consider communicating (or finite-diameter) MDPs

- Diameter (Jaksch et al., 2010): Captures the maximal shortest-path between any pair of states.

Undiscounted RL: Regret

Regret: The difference between the cumulative reward of an optimal policy $\star$ and that gathered by the learner:

$$\mathfrak{R}(T) := Tg^\star - \sum_{t=1}^{T} r_t$$

where $g^\star$ is the average-reward (gain) of an optimal policy.

Undiscounted RL: Regret

Regret: The difference between the cumulative reward of an optimal policy $\star$ and that gathered by the learner:

$$\mathfrak{R}(T) := Tg^\star - \sum_{t=1}^{T} r_t$$

where $g^\star$ is the average-reward (gain) of an optimal policy.

Alternatively, the objective of the learner is to minimize the regret.

The key difficulty to do so is to balance *exploration* vs. *exploitation*:

- Play the best action so far, . . .
- . . . or rather explore a different action?

Outline

1. Background: UCRL2
2. UCRL3
3. Regret Analysis
4. Numerical Experiments
5. Conclusion

Notations

Under a given algorithm, we define:

- $N_t(s, a)$: number of visits, up to time t, to (s, a).
- $N_t(s, a, s')$: number of visits, up to time t, to (s, a) followed by a visit to s'.
- Empirical estimates of transition probabilities and rewards:

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

$$\widehat{p}_t(s'|s, a) = \frac{N_t(s, a, s')}{\max\{N_t(s, a), 1\}}$$

UCRL2

UCRL2 (Jaksch et al., 2010): A model-based algorithm for undiscounted RL implementing the principle of optimism in the face of uncertainty.

- Mainstains a set of plausible MDPs (models) by defining high-probability confidence sets for μ and p
- Chooses an optimistic model (among models) and an optimistic policy leading to the highest average-reward.

UCRL2

UCRL2 (Jaksch et al., 2010): A model-based algorithm for undiscounted RL implementing the principle of optimism in the face of uncertainty.

At time t, UCRL2 considers the set $\mathcal{M}_{t,\delta}$ of candidate MDPs $M' = (\mathcal{S}, \mathcal{A}, \mu', p')$ satisfying: For all s, a,

$$\left\|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\right\|_1 \le \sqrt{\frac{14S}{N_t(s,a)} \log\left(\frac{2At}{\delta}\right)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \le \sqrt{\frac{7}{2N_t(s,a)} \log\left(\frac{2SAt}{\delta}\right)}$$

UCRL2

UCRL2 (Jaksch et al., 2010): A model-based algorithm for undiscounted RL implementing the principle of optimism in the face of uncertainty.

At time t, UCRL2 considers the set $\mathcal{M}_{t,\delta}$ of candidate MDPs $M' = (\mathcal{S}, \mathcal{A}, \mu', p')$ satisfying: For all s, a,

$$\left\|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\right\|_1 \le \sqrt{\frac{14S}{N_t(s,a)} \log\left(\frac{2At}{\delta}\right)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \le \sqrt{\frac{7}{2N_t(s,a)} \log\left(\frac{2SAt}{\delta}\right)}$$

$\Longrightarrow$ With probability, $M \in \mathcal{M}_{t,\delta}$.

UCRL2

UCRL2 (Jaksch et al., 2010): A model-based algorithm for undiscounted RL implementing the principle of optimism in the face of uncertainty.

- For any communicating MDP with S states, A actions, and diameter D, UCRL2 satisfies

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

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

UCRL2

UCRL2 (Jaksch et al., 2010): A model-based algorithm for undiscounted RL implementing the principle of optimism in the face of uncertainty.

- For any communicating MDP with S states, A actions, and diameter D, UCRL2 satisfies

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

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

UCRL2 and its variants do not perform epmirically well despite their strong regret guarantees.

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.

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

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

UCRL3: Confidence Set for p

For each pair (s, a), define

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

- $C^1_{t,\delta}(s,a,s')$ 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: 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

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

$$C^1_{t,\delta}(s,a,s') = \left\{ \lambda : |\widehat{p}_t(s'|s,a) - \lambda| \le \sqrt{\frac{2\lambda(1-\lambda)\ell_{N_t(s,a)}\big(\frac{\delta}{2S^2A}\big)}{N_t(s,a)}} + \frac{\ell_{N_t(s,a)}\big(\frac{\delta}{2S^2A}\big)}{3N_t(s,a)} \right\}$$

where $\ell_n(\delta) = \eta \log\left(\frac{\log(n)\log(\eta n)}{\log^2(\eta)\delta}\right)$ with $\eta > 1$ (an arbitrary choice).

UCRL3: Confidence Set for p

For each pair (s,a), define

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

UCRL3: Confidence Set for p

For each pair (s,a), define

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

$$C^2_{t,\delta}(s,a,s') = \left\{ \lambda : -\sqrt{\underline{g}(\lambda)} \le \frac{\widehat{p}_t(s'|s,a) - \lambda}{\beta_{N_t(s,a)}\left(\frac{\delta}{2SA}\right)} \le \sqrt{\underline{g}(\lambda)} \right\}$$

where $g(\lambda) = \frac{1/2-\lambda}{\log(1/\lambda - 1)}$ and $\underline{g}(\lambda) = \begin{cases} g(\lambda) & \text{if } \lambda < 0.5 \\ \lambda(1-\lambda) & \text{else} \end{cases}$, and

$$\beta_n(\delta) := \sqrt{\frac{2(1+\frac{1}{n})\log(\sqrt{n+1}/\delta)}{n}}.$$

UCRL3: Set of Models

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

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

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

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

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 $s_1, s_2 \in \cup_{a\in\mathcal{A}} \mathrm{supp}\big(p(\cdot|s,a)\big)$ 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

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

- 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,

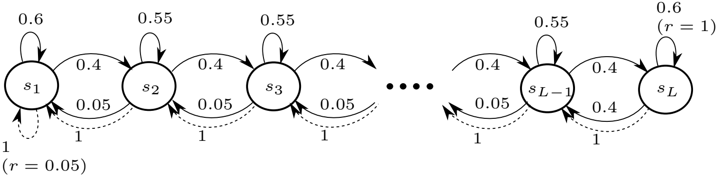
$$\mathfrak{R}(T) \leq \widetilde{\mathcal{O}}\Big(\Big[\sqrt{\sum\nolimits_{s,a} \max(D_s^2 K_{s,a}, 1)} + D\Big]\sqrt{T}\Big).$$

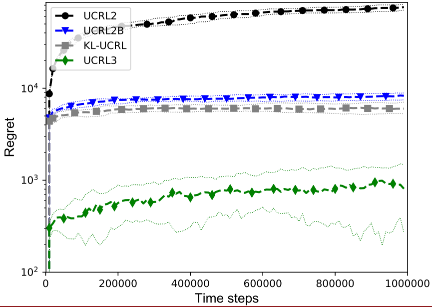
State-of-the-Art Regret Bounds

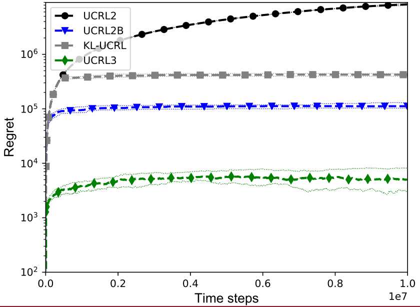
Algorithm	Regret bound
UCRL2 (Jaksch et al., 2010)	$\mathcal{O}\Big(DS\sqrt{AT\log(T/\delta)}\Big)$
KL-UCRL (Filippi et al., 2010)	$\mathcal{O}\Big(DS\sqrt{AT\log(\log(T)/\delta)}\Big)$
KL-UCRL (Talebi et al., 2018)	$\mathcal{O}\Big(\Big[D+\sqrt{S\sum_{s,a}\max(\mathbb{V}_{s,a},1)}\Big]\sqrt{T\log(\log(T)/\delta)}\Big)$
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)	$\Omega(\sqrt{DSAT})$

Numerical Experiments

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

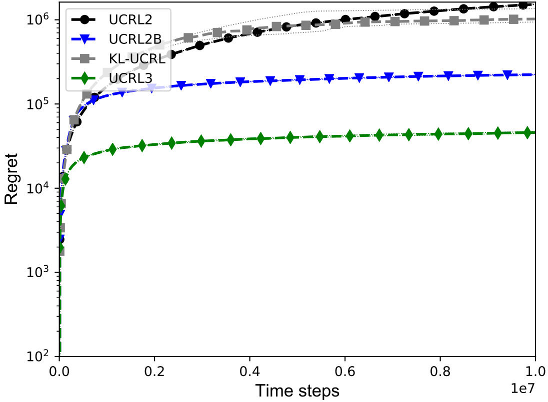






Numerical Experiments

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



Conclusions and Future Work

We introduced UCRL3 for average-reward RL in communicating MDPs:

- A novel variant of UCRL2 using (i) tight and time-uniform confidence sets, and (ii) a novel approach for planning.
- Beats all existing variants of UCRL2 in practice yet enjoying a better regret bound.

Future Work:

- Closing the gap between upper and lower bounds
- Problem-dependent regret lower and upper bounds for average-reward RL