Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes

Ronan Fruit (INRIA), Matteo Pirotta (INRIA), Alessandro Lazaric (FAIR)

Inria informatics mathematics
Région Hauts-de-France
facebook Artificial Intelligence Research

Motivations

- Learning in an unknown environment means to balance
  - Exploration
  - Exploitation
- All theoretically-grounded approaches requires prior knowledge
  - This information is hard to get!
  - Limit their applicability
- This is about learning without prior knowledge!!

Online Learning in MDPs

▶ **Markov Decision Process** $M = \{\mathcal{S}, \mathcal{A}, r, p\}$

- states: $\mathcal{S} = \mathcal{S}^{\mathrm{C}} \cup \mathcal{S}^{\mathrm{T}}$
- actions: $A = (A_s)_{s \in \mathcal{S}}$
- mean rewards: $r(s,a)$
- transition probabilities: $p(s'|s,a)$
- *Possible next states:* $\Gamma^{\mathcal{S}} = \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} \|p(\cdot|s,a)\|_0$

communicating set: $\forall s,s' \in \mathcal{S}^{\mathrm{C}}, \exists \pi : \mathbb{P}^\pi(s \to s') > 0$

transient set: $\mathcal{S}^{\mathrm{C}} \cap \mathcal{S}^{\mathrm{T}} = \emptyset$

▶ **Optimality criterion: long-term average reward**

For any policy $\pi \in \Pi^{\mathrm{SR}}(M)$ starting from $s \in \mathcal{S}$:

GAIN: $$g_M^\pi(s) := \lim_{T\to+\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t,a_t)\right]$$

BIAS: $$h_M^\pi(s) := C\text{-}\lim_{T\to+\infty} \mathbb{E}\left[\sum_{t=1}^{T}\left(r(s_t,a_t) - g_M^\pi(s_t)\right)\right]$$

In weakly communicating MDPs:
any optimal policy $\pi^* \in \arg\max_\pi\{g^\pi(s)\}$ has *constant* gain

▶ **Learning problem: cumulative regret minimization**

The true M^* is unknown, thus it is g^*

$$\Delta(\mathfrak{A}, T) = Tg^* - \sum_{t=1}^{T} r_t(s_t,a_t)$$

Asm. 1 The initial state $s_1 \in \mathcal{S}^{\mathrm{C}}$

▶ **Diameter and Span:** [Jaksch et al. 2010; Bartlett and Tewari, 2009]

expected time $s \to s'$

$$D^{\mathcal{S}} = \max_{s,s'\in\mathcal{S}}\left\{\min_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \mathbb{E}_\pi\left[T(s')|s\right]\right\}$$

$$sp_{\mathcal{S}}\{h^*\} = \max_{s\in\mathcal{S}}\{h^*(s)\} - \min_{s\in\mathcal{S}}\{h^*(s)\}$$

- D depends on **all** policies (*global property*)
- $sp_{\mathcal{S}}\{h^*\}$ on only π^*
- $sp_{\mathcal{S}}\{h^*\} \le D$ *(always)*

In weakly communicating MDPs $D = \infty$ but $sp_{\mathcal{S}}\{h^*\} \le \infty$

References

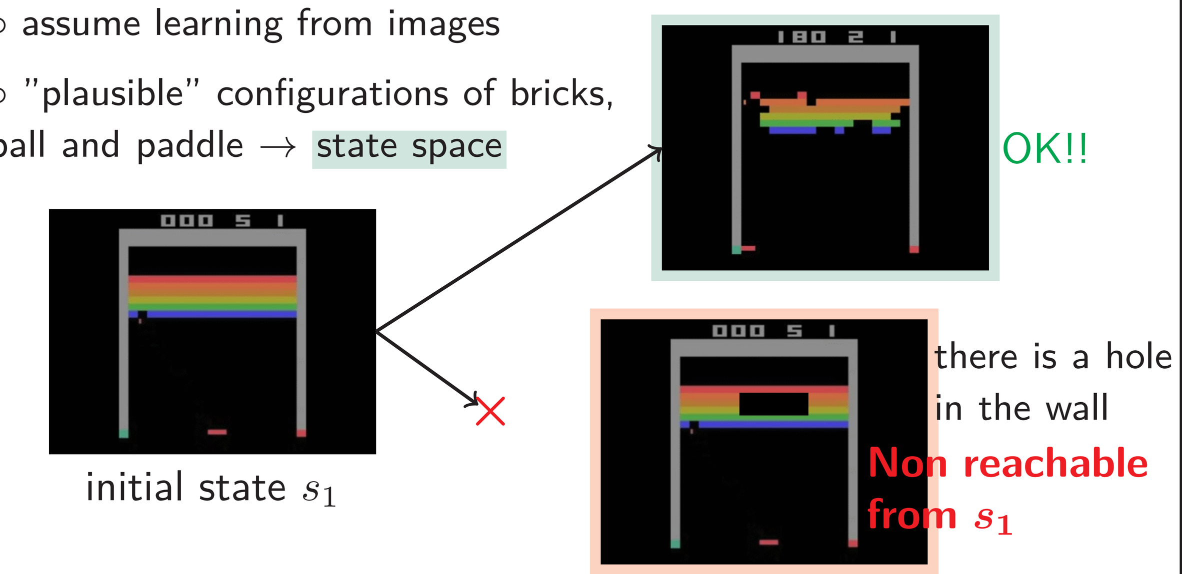
- Jaksch, Ortner, and Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 2010.
- Bartlett and Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In UAI 2009.
- Agrawal and Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In NIPS 2017, 2017.
- Fruit, Pirotta, Lazaric and Ortner. Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning. In ICML 2018, 2018.
- Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. J. Artif. Intell. Res., 13:227-303, 2000.

Prior Knowledge & Misspecified states

- UCRL and OPT-PSRL assume communicating MDPs
  ⇒ all states are reachable
  ⇒ $D < +\infty$ **regret:** $\widetilde{O}\left(D^{\mathcal{S}}\sqrt{\Gamma^{\mathcal{S}}SAT}\right)$ / $\widetilde{O}\left(D^{\mathcal{S}}\sqrt{SAT}\right)$
- reasonable! but rarely verified in practice

BREAKOUT EXAMPLE

- assume learning from images
- "plausible" configurations of bricks, ball and paddle → state space



- In weakly communicating or misspecified problems ($D = +\infty$) UCRL and OPT-PSRL ⇝ linear regret
- REGAL.C and SCAL exploits the knowledge $sp_{\mathcal{S}}\{h^*_{M^*}\} \le H$ implicitly "removes" non-reachable states able to learn in weakly comm. MDPs **regret:** $\widetilde{O}\left(H\sqrt{\Gamma^{\mathcal{S}}SAT}\right)$

⚠ knowing H not easier than designing well-specified states

*similar assumptions in Bayesian regret

Truncated Upper-Confidence for Reinforcement Learning (TUCRL)

Plain OFU (e.g., UCRL) executed on

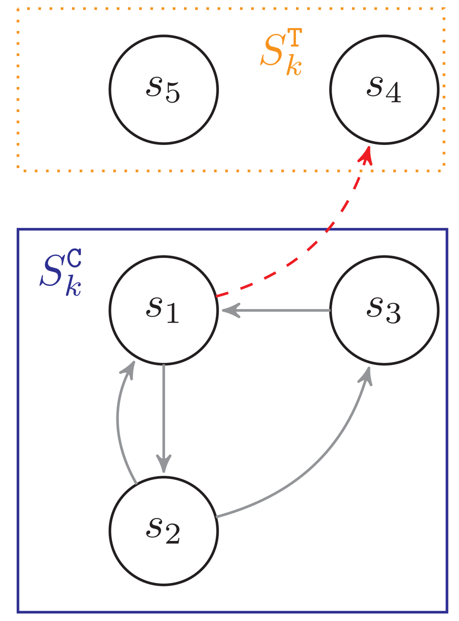
- $\mathcal{S}$ and $\mathcal{S}^{\mathrm{T}} \neq \emptyset$ is over-exploring (*linear regret*)
- $\mathcal{S}_k^{\mathrm{C}} \subseteq \mathcal{S}^{\mathrm{C}}$ may under-exploration (*linear regret*)

Empirical estimate of $\mathcal{S}^{\mathrm{C}}$

$$\mathcal{S}_k^{\mathrm{C}} := \left\{s \in \mathcal{S} \,\Big|\, \sum_{a\in\mathcal{A}_s} N_k(s,a) > 0\right\} \cup \{s_{t_k}\}$$

s_{t_k}: starting state of episode k

★ TUCRL learns in weakly communicating or misspecified problems WITHOUT PRIOR KNOWLEDGE



Does this transition $s_1 \to s_4$ exists? Is $s_4 \in \mathcal{S}_k^{\mathrm{T}}$? Should be enabled it at episode k? Explore or not?

At episode k we know:
$\widehat{p}_k(s'|s,a)$ (empirical mean) and $|\widetilde{p}(s'|s,a) - \widehat{p}(s'|s,a)| \le \beta_{p,k}^{sas'}$ (confidence interval)

TUCRL idea:
"guess" a lower bound ρ_k to the trans. probabilities
$\widehat{p}(s'|s,a) + \beta_{p,k}^{sas'} < \rho_k \Rightarrow s \to s'$ FORBIDDEN
Maximum probability given conf. interval

⚠ $(\rho_k)_k$ should be non-increasing!!

WHY DOES TUCRL WORK?

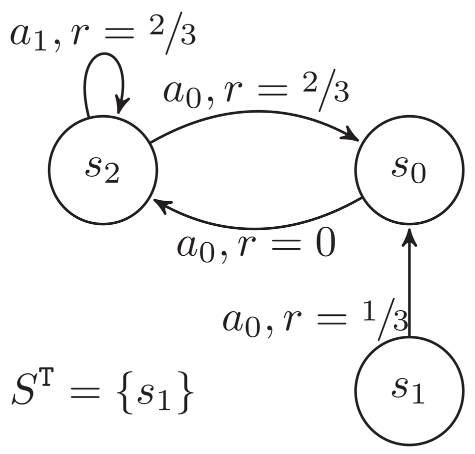
- reconsiders transitions periodically (i.e., decrease ρ_k)
  ⇒ avoid *under-exploration*
- exploits Bernstein confidence interval:
  $$\beta_{p,k}^{sas'} = \sqrt{\frac{\alpha\sigma_k^2(s'|s,a)}{N_k(s,a)}} + \frac{\beta}{N_k(s,a)}$$
  - $(s,a) \to s'$ not observed
    ⇒ $\sigma_k^2 = 0$ and $\widehat{p}_k = 0$
    ⇒ $\beta/N_k(s,a) < \rho_k$
  - fast shrinking $\approx O(1/N_k(s,a))$
- we set $\rho_k = O(SA/t_k)$ equivalent to removing transitions s.t.:
  $N_k(s,a) > \sqrt{t_k/SA} \Rightarrow \widetilde{p}(s'|s,a) = 0, \forall s' \in \mathcal{S}_k^{\mathrm{T}}$

TUCRL ALGORITHM

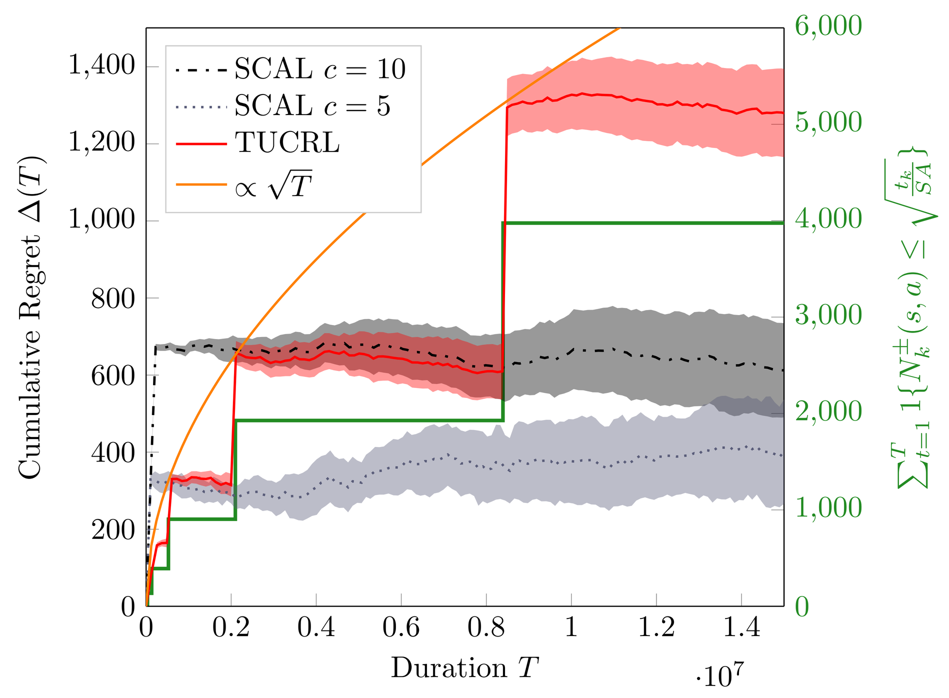
For episode $k = 1, 2, \ldots$

1. **Confidence set:** TUCRL builds $\mathcal{M}_k = \{M = (\mathcal{S}, \mathcal{A}, \widetilde{r}, \widetilde{p}) : \widetilde{r}(s,a) \in B_{r,k}(s,a),\ \widetilde{p}(\cdot|s,a) \in B_{p,k}(s,a)\}$ with:
$$B_{p,k}(s,a) = \{\widetilde{p}(\cdot|s,a) : \|\widetilde{p}(\cdot|s,a) - \widehat{p}(\cdot|s,a)\|_1 \le \sum_{s'} \beta_{p,k}^{sas'}\}$$
$$\cap \{\widetilde{p}(\cdot|s,a) : N_k(s,a) > \sqrt{t_k/SA} \Rightarrow \forall s' \in \mathcal{S}_k^{\mathrm{T}},\ \widetilde{p}(s'|s,a) = 0\}$$
2. **Planning:** TUCRL solves $(\widetilde{M}_k, \widetilde{\pi}_k) = \arg\max_{M\in\mathcal{M}_k,\pi}\{g_M^\pi\}$
3. **Execution:** of policy $\widetilde{\pi}_k$ in the true MDP

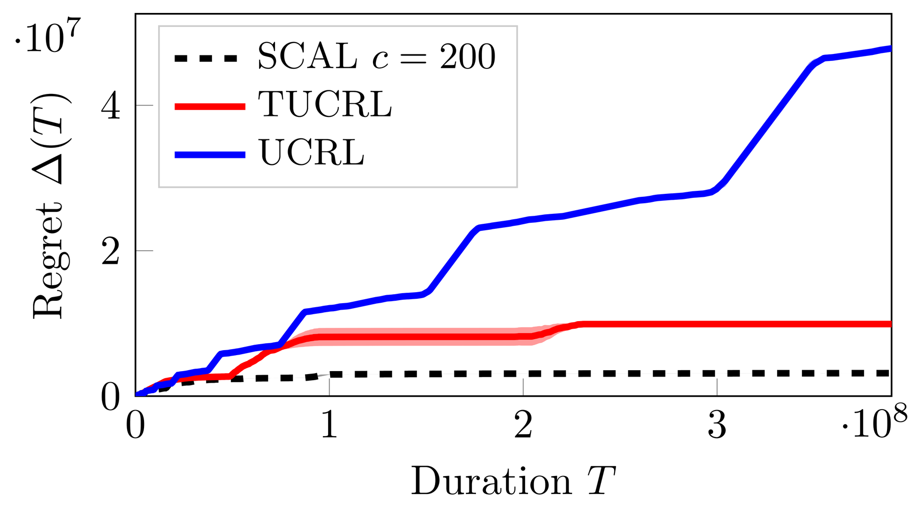
Numerical Experiments



$D = +\infty$ (weakly com.)
- UCRL linear regret
- SCAL logarithmic regret
- TUCRL square-root regret



TAXI PROBLEM [Dietterich, 2000]
OpenAI gym implementation ⇒ misspecified states



Regret of TUCRL

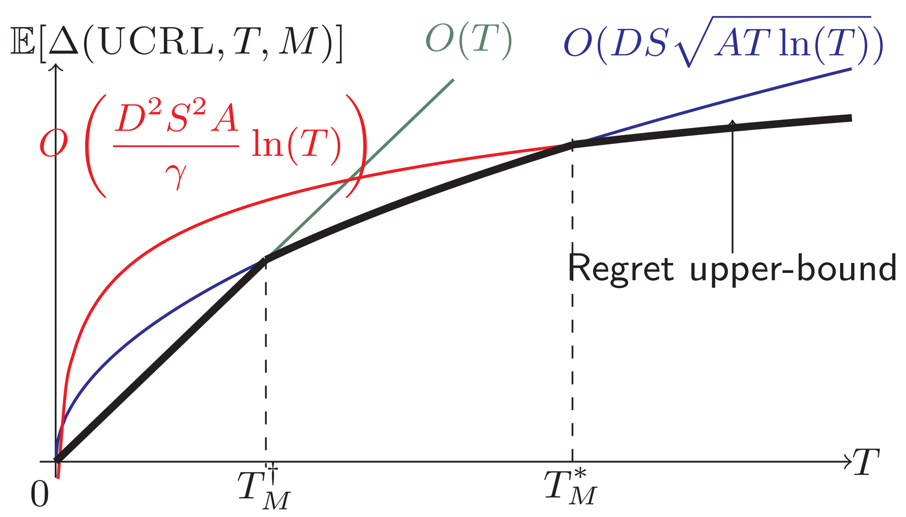
$$\Delta(\text{TUCRL}, T) = \widetilde{O}\left(D^{\mathrm{C}}\sqrt{\Gamma^{\mathrm{C}}S^{\mathrm{C}}AT} + \left(D^{\mathrm{C}}\right)^2 S^3 A\right)$$

- Adaptability to communicating part
  $D^{\mathrm{C}} := D^{\mathcal{S}^{\mathrm{C}}}$, $\Gamma^{\mathrm{C}} := \Gamma^{\mathcal{S}^{\mathrm{C}}}$ and S^{C}
- Regret due to the early stage where TUCRL suffers linear regret

If M^* is communicating
First term: same as UCRL
Second term: bigger than UCRL by a factor $S^{\mathrm{c}}/\Gamma^{\mathrm{c}}$

Exploration-Exploitation Dilemma with infinite diameter

UCRL regret "regimes"



efficient algorithm: time $T_M^\dagger$ to achieve sublinear regret is *polynomial* in the parameters of the MDP

In communicating MDPs, UCRL achieves sublinear regret in $T_M^\dagger = \widetilde{O}\left((D^{\mathcal{S}})^2\Gamma^{\mathcal{S}}SA\right)$
⟹ *efficient algorithm*

Impossibility result

Without prior knowledge, any efficient learning algorithm must satisfy $T_M^* = +\infty$ when M has infinite diameter (i.e., it cannot achieve logarithmic regret)

No logarithmic regret without prior knowledge!