Exploration Bonus for Regret Minimization in Discrete and Continuous Average Reward MDPs

J. Qian (INRIA Lille), R. Fruit (INRIA Lille), M. Pirotta (FAIR), A. Lazaric (FAIR)

Introduction

What do we do?

- Exploration in continuous MDPs
- With theoretical guarantees

How?

- Using exploration bonus and discretization

Online Learning in MDPs

- **Markov Decision Process** $M = \{S, A, r, p\}$
- **Optimality criterion:** average reward

For any policy π starting from $s \in \mathcal{S}$:

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

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

- **Learning problem:** **cumulative regret minimization**

π^* and g^* are unknown

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

- **Diameter:** [Jaksch et al. 2010]

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

UCRL2-like Exploration

For $k = 1, 2, \dots$

1. Estimation of model and uncertainty
$$\mathcal{M}_k = \{M = (\mathcal{S}, \mathcal{A}, \widetilde{p}, \widetilde{r}) : \widetilde{p}(\cdot|s,a) \in B_k^p(s,a), \; \widetilde{r}(s,a) \in B_k^r(s,a)\}$$
2. Planning for optimistic policy
$$(M_k, \pi_k) = \arg\max_{M\in\mathcal{M}_k} \max_\pi \{g^\pi(M)\}$$
3. Execution of policy π_k
   - execute action $a_t \sim \pi_k$
   - observe reward r_t and next state s_{t+1}

- Estimation: $\mathcal{M}_k$ is the set of plausible MDP such that

$$\|\widetilde{p}(\cdot|s,a) - p(\cdot|s,a)\|_1 \le \beta_k^p(s,a) \approx \sqrt{\frac{SL}{N_k(s,a)}}$$

$$|\widetilde{r}(s,a) - r(s,a)| \le \beta_k^r(s,a) \approx r_{\max}\sqrt{\frac{L}{N_k(s,a)}}$$

- Planning: use Extended Value Iteration

$$v_{n+1}(s) = \widetilde{L}v_n = \max_a \left\{ \max_{\widetilde{r}\in B_k^r(s,a)} \widetilde{r} + \max_{\widetilde{p}\in B_k^p(s,a)} \widetilde{p}^\top v_n \right\}$$

SCAL$^+$: tabular MDP

- Exploration bonus: Used in deep RL [Bellemare et al. 2016, Tang et al. 2017] and/or when the intrinsic horizon is known [Azar et al. 2017, Jin et al. 2018]

For $k = 1, 2, \dots$

1. Estimation of empirical model
$$M_k = (\mathcal{S}, \mathcal{A}, \underbrace{\widehat{p}}_{\text{average transitions}}, \overbrace{\widehat{r}}^{\text{average rewards}} + \underbrace{b_k}_{\text{exploration bonus}})$$
average transitions
$$\widehat{p}_k(s'|s,a) = \frac{N_k(s,a,s')}{\sum_{s'} N_k(s,a,s')}$$
2. Planning for optimistic policy
$$\pi_k = \operatorname*{argmax}_\pi \{g^\pi(M_k)\}$$
3. Execution of policy π_k

- Plan using the empirical MDP
- Bonus is used to recover optimism

$$b_k(s,a) \approx (r_{\max} + c)\sqrt{\frac{L}{N_k(s,a)}} < \beta_k^r(s,a) + c\beta_k^p(s,a)$$

Tighter uncertainty $\implies$ better performance

"equivalent" UCRL2 bonus

⚠ prior knowledge:
$$rng(h^\star) = \max_s h^\star(s) - \min_s h^\star(s) \le c$$

SCAL$^+$: regret

For any MDP such that $rng(h^\star) \le c$, w.p. $1-\delta$

$$R(\mathrm{SCAL}^+, T) = \widetilde{O}\left(c\sqrt{T\sum_{s,a}\Gamma(s,a)}\right)$$

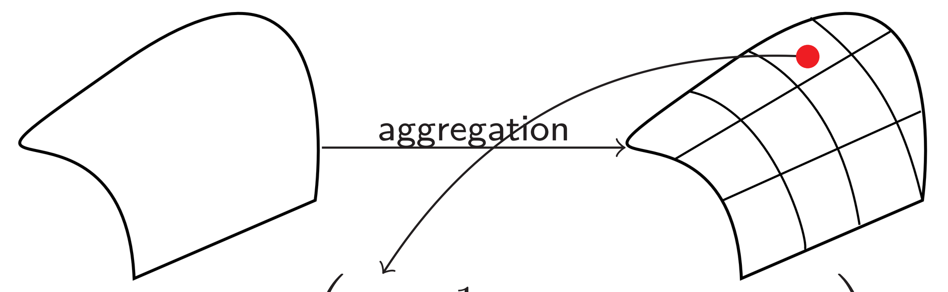
- $\Gamma(s,a) = \|p(\cdot|s,a)\|_0$ (number of next states)
- Worst-case $\widetilde{O}(cS\sqrt{AT})$
  - as UCRL2 except that D is replaced by c

SCCAL$^+$: continuous MDP

- $\mathcal{S}$ is continuous, $\mathcal{A}$ is discrete
- MDP (reward and transitions) is Hölder continuous

$$|r(s,a) - r(s',a)| \le r_{\max} L |s - s'|^\alpha$$
$$\|p(\cdot|s,a) - p(\cdot|s',a)\|_1 \le L|s - s'|^\alpha$$

- SCCAL$^+$ combines SCAL$^+$ with state aggregation



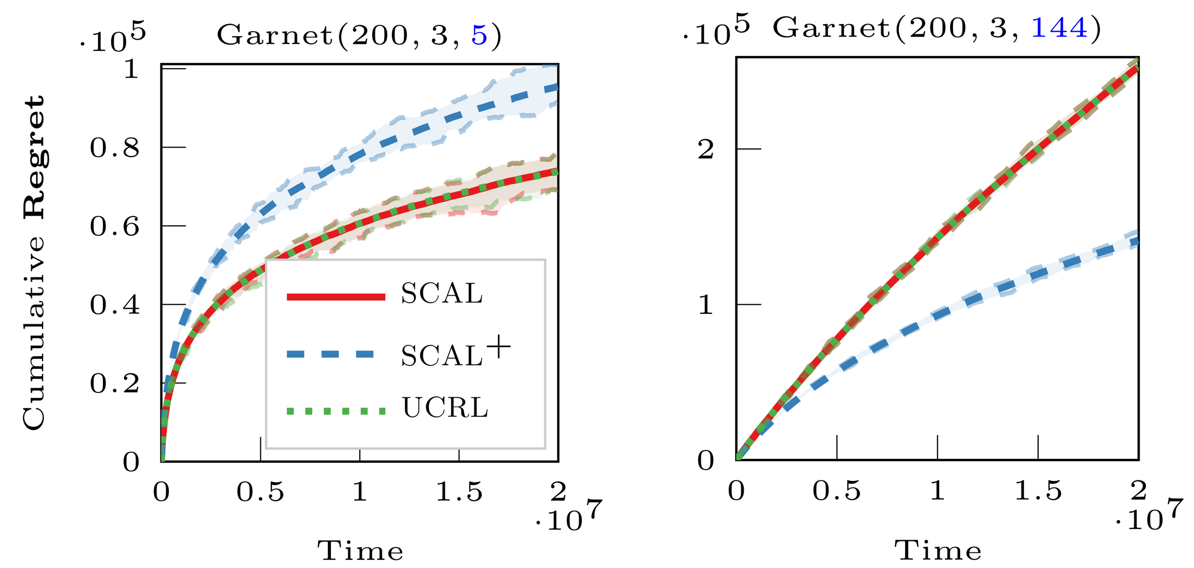
$$b(I,a) \approx (c + r_{\max})\left(\frac{1}{\sqrt{N_k(I,a)}} + \underbrace{LM_S^{-\alpha}}_{\text{bias}}\right)$$

Regret: for any Hölder MDP w.p. $1-\delta$

$$R(\mathrm{SCCAL}^+, T) = \widetilde{O}\left(cL\sqrt{A}T^{(\alpha+2)/(2\alpha+2)}\right)$$

Numerical Results

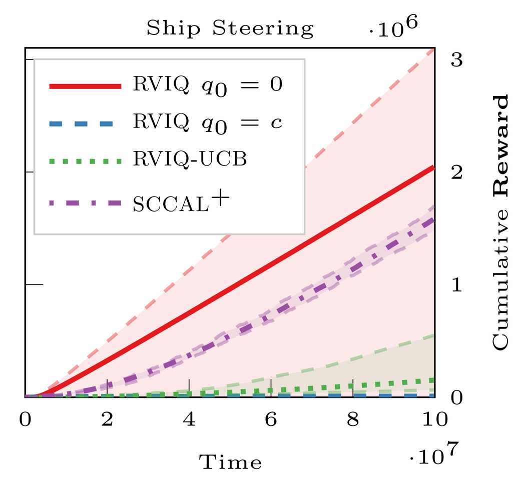
Garnet: tabular MDP



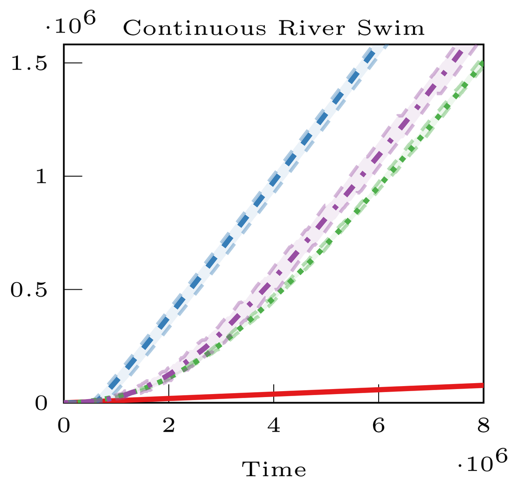
- SCAL$^+$ can indeed outperform UCRL2

– Continuous MDPs –

Ship Steering:



RiverSwim:



- First implementable algorithm with guarantees in cont. MDPs (lots details are missing here, see paper)
- More stable than model-free version

Prior Knowledge on the Bias Span

- Provides a sense of what is realizable in the true MDP
- Avoids over-optimism
- Necessary to define the exploration bonus

Implicit in other settings (infinite-horizon discounted, finite-horizon)

$$|\widetilde{p}(\cdot|s,a) - \widehat{p}(\cdot|s,a)| \le \|\widetilde{p}(\cdot|s,a) - \widehat{p}(\cdot|s,a)\|_1 \|h^\star\|_\infty$$

Setting	MDP parameter	Horizon	Knowledge	Exploration Bonus
infinite-horizon discounted	γ	$\frac{1}{1-\gamma}$	$\|v^\star\|_\infty \le \frac{r_{\max}}{1-\gamma}$	$\widetilde{\Theta}\left(\frac{r_{\max}}{1-\gamma}\sqrt{\frac{1}{N_k(s,a)}}\right)$
finite-horizon	H	H	$\|v^\star\|_\infty \le r_{\max}H$	$r_{\max}H\sqrt{\frac{1}{N_k(s,a)}}$
average reward	?	$+\infty$	$rng(h^\star) \le c$ ⚠ assumption	$\widetilde{\Theta}\left(c\sqrt{\frac{1}{N_k(s,a)}}\right)$