Regret Minimization in MDPs with Options without Prior Knowledge

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, Emma Brunskill

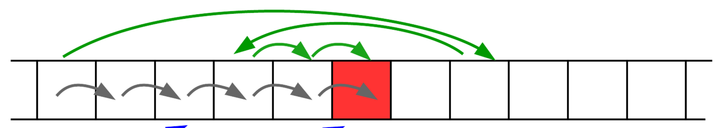
Motivations

- **"Flat" RL**: difficult to learn *complex behaviours* (eg, sequence of subgoals) ⇒ Humans abstract from **low-level actions**
- **Hierarchical RL**: decompose large problems into smaller ones by imposing constraints on value function or **policy**
- Possible implementation: **options** [Sutton et al., 1999]
- Empirical observations: introducing options in an MDP can **speed up** learning but can also be **harmful** [Jong et al., 2008].

⇒ Lack of theoretical motivation and understanding of options

Theoretical analysis of learning with options

- Adding options does not just reduce the space of stationary policies, the **exploration** is also greatly affected



- Navigability and temporal abstraction can be preserved in an UCRL-like algorithm when prior knowledge about options is given [Fruit and Lazaric, 2017]

Can we preserve temporal abstraction without prior knowledge?

Temporal Abstraction in MDPs

Markov Decision Process

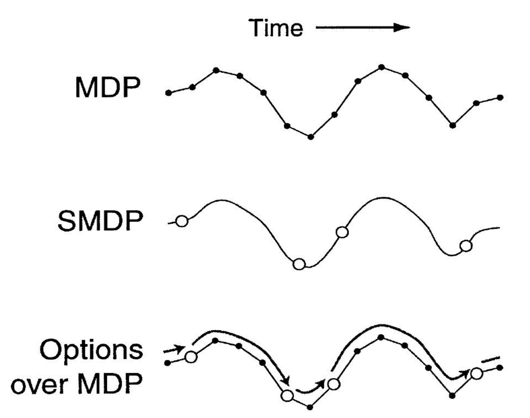
$M = \{\mathcal{S}, \mathcal{A}, r, p\}$

- $\mathcal{S}$: states
- $\mathcal{A} = (\mathcal{A}_s)_{s\in\mathcal{S}}$: actions
- $r(s,a)$: rewards
- $p(s'|s,a)$: transition probabilities

The Option Framework

(Markov) Option $o = \{s_o, \beta_o, \pi_o\}$

- $s_o \in \mathcal{S}$: initial state,
- π_o: stationary policy,
- $\beta_o : \mathcal{S} \to [0,1]$: termination.



T_n/n: **"temporal abstraction factor"**

n decision steps ∘

T_n time steps •

Proposition 1. A set of options $\mathcal{O}$ defined on an MDP M induces a **semi-MDP** $M' = \{\mathcal{S}_\mathcal{O}, \mathcal{O}, R_\mathcal{O}, p_\mathcal{O}, \tau_\mathcal{O}\}$ where the cumulative reward $R_\mathcal{O}$ and holding time $\tau_\mathcal{O}$ of every option $o \in \mathcal{O}$ are sub-exponential random variables with parameters $\sigma_R(o)$ and $\sigma_\tau(o)$.

Online Learning for SMDPs

Optimality criterion

An optimal policy π^* **maximizes the long-term average reward**

$$\rho^*(M') = \max_\pi \left\{ \lim_{n\to+\infty} \mathbb{E}^\pi\left[\frac{R_n}{T_n}\right]\right\};\quad T_n = \sum_{i=1}^n \tau_i;\quad R_n = \sum_{i=1}^n r_i$$

Performance Measure

The learning agent aims at **minimizing the** $\boxed{\text{cumulative regret}}$

$$\Delta(M', n) = \rho^*(M')T_n - R_n$$

Diameter of an SMDP

$$D(M') = \max_{s,s'\in\mathcal{S}} \left\{ \min_{d\in D_M^{MD}} \left\{ \mathbb{E}^{d^\infty}\left[T(s')|s_0 = s\right]\right\}\right\}$$

$$\text{where } T(s') = \inf\left\{\sum_{i=1}^n \tau_i : n\in\mathbb{N},\ s_n = s'\right\}$$

Free-Parameter SMDP-UCRL

SMDP optimistic optimality equation:

$$\widetilde{\rho}^*_{\mathrm{SUCRL}} = \max_{o\in\mathcal{O}_s}\left\{ \max_{\widetilde{\boldsymbol{R}},\widetilde{\boldsymbol{\tau}}} \left\{ \frac{\widetilde{R}(s,o)}{\widetilde{\tau}(s,o)} + \frac{1}{\widetilde{\tau}(s,o)}\left(\max_{\widetilde{\boldsymbol{p}}}\left\{\sum_{s'\in\mathcal{S}} \widetilde{p}(s'|s,o)\widetilde{u}^*(s')\right\} - \widetilde{u}^*(s)\right)\right\}\right\}$$

Exploits σ_R and σ_τ to compute $\widetilde{\boldsymbol{R}}$ and $\widetilde{\boldsymbol{\tau}}$

$$|\widetilde{r}_k(s,a) - \widehat{r}_k(s,a)| \lesssim \alpha_r\sigma_r\sqrt{\frac{1}{N_k(s,a)}}$$

$$\|\widetilde{p}_k(s'|s,o) - \widehat{p}_k(s'|s,o)\|_1 \lesssim \alpha_p\sqrt{\frac{1}{N_k(s,o)}}$$

$$|\widetilde{\tau}_k(o) - \widehat{\tau}_k(o)| \lesssim \alpha_\tau\sigma_\tau\sqrt{\frac{1}{N_k(s,o)}}$$

Visit counters of (s,a) or (s,o)

Remove prior knowledge

An option can be seen as an **absorbing Markov Chain** by merging initial and absorbing states $\Longrightarrow$ **Irreducible MC**

Irreducible Transformation

τ: first-*hitting*-time of $\{s_3, s_4\}$
$R = \sum_{t=1}^\tau r(s_t, \pi_o(s_t))$

τ: First-*return*-time in s_0

$$\Longrightarrow \begin{cases} \overline{\tau} = 1/\mu(s_0) \\ \overline{R}/\overline{\tau} = r^\intercal \mu \end{cases}$$

Irreducible optimistic optimality equation:

$$\widetilde{\rho}^*_{\mathrm{FSUCRL}} = \max_{o\in\mathcal{O}_s}\left\{ \max_{\widetilde{\boldsymbol{\mu}}_o,\widetilde{\boldsymbol{r}}_o} \left\{ \sum_{s'\in\mathcal{S}_o} \widetilde{r}_o(s')\widetilde{\mu}_o(s') + \widetilde{\mu}_o(s)\left(\max_{\widetilde{\boldsymbol{b}}_o}\left\{\widetilde{\boldsymbol{b}}_o^\intercal\widetilde{\boldsymbol{u}}^*\right\} - \widetilde{u}^*(s)\right)\right\}\right\}$$

Exploits knowledge acquired over time

$$|\widetilde{r}_k(s,a) - \widehat{r}_k(s,a)| \lesssim \alpha_r r_{\max}\sqrt{\frac{1}{N_k(s,a)}}$$

$$|\widetilde{p}_k(s'|s,o) - \widehat{p}_k(s'|s,o)| \lesssim \sqrt{\frac{\widehat{p}_k(s'|s,o)\big(1-\widehat{p}_k(s'|s,o)\big)}{N_k(s,o)}} + \frac{\alpha_p}{N_k(s,o)}$$

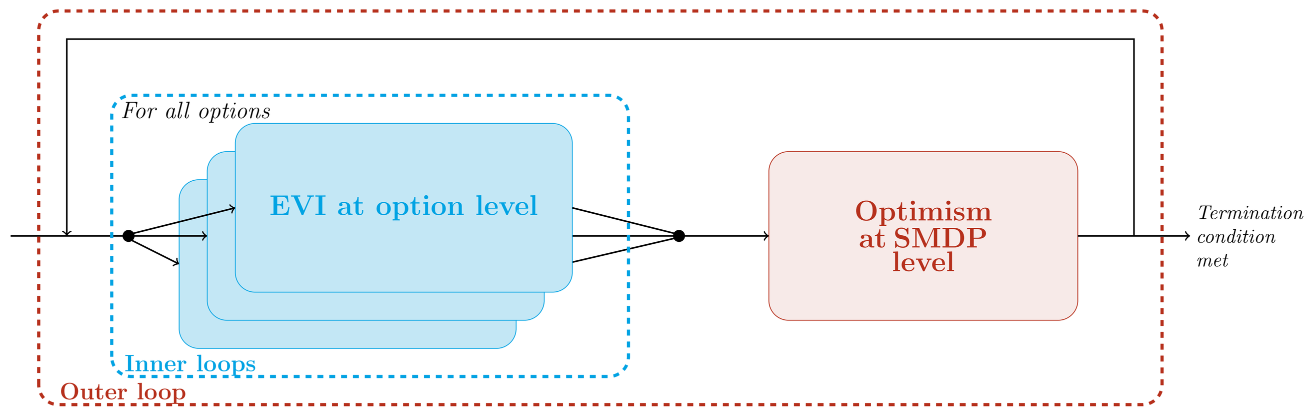
$$|\widetilde{P}'_{o,k}(s'|s) - \widehat{P}'_{o,k}(s'|s)| \lesssim \sqrt{\frac{\widehat{P}_{o,k}(s,s')\big(1-\widehat{P}_{o,k}(s,s')\big)}{N_k(s,\pi_o(s))}} + \frac{\alpha_{mc}}{N_k(s,\pi_o(s))}$$

How do we solve the problem?

FSUCRL

1: **for** episodes $k = 1, 2, \dots$ **do**
2: Compute ε_k-approximation of the optimal policy $\widetilde{\pi}_k$ giving $\widetilde{\rho}^*_{\mathrm{FSUCRL}}$
– Use *Extended Value Iteration* for each option
– Use *optimism* at option level
3: Execute $\widetilde{\pi}_k$ to acquire new experience
4: **end for**

Nested Extended Value Iteration



Theoretical Results

Regret bound in an SMDP [Fruit and Lazaric, 2017]

$$\Delta(M', T_n) = O\left(\left(D'\sqrt{S'} + \mathcal{C}(M', n, \delta)\right)\sqrt{S'A'n\log\left(\frac{n}{\delta}\right)}\right)$$

where $\mathcal{C}(M', n, \delta)$ depends on knowing $\sigma_R(o)$ and $\sigma_\tau(o)$.

Special case: Bound for MDPs [Jaksch et al., 2010]:

$$\Delta(M, T_n) = O\left(DS\sqrt{AT_n\log\left(\frac{T_n}{\delta}\right)}\right)$$

Advantages of temporal abstraction: $n \ll T_n \Rightarrow \Delta(M', T_n) \le \Delta(M, T_n)$ if $D' \approx D$

Theorem 1. The regret of FSUCRL in SMDP M' is w.h.p.

$$\boldsymbol{\Delta(M', T_n) = O}\left(\left(D'\sqrt{S'} + \mathcal{C}(\mu^*, \kappa^*, \tau_{\max})\right)\sqrt{S'On\log\left(\frac{n}{\delta}\right)}\right)$$

$$\mu^* = \min_{o\in\mathcal{O}, s\in\mathcal{S}_o} \underbrace{\mu_o(s)}_{\text{(stationary distribution of MC)}} \qquad \kappa^* = \max_{o\in\mathcal{O}} \underbrace{\kappa_o}_{\text{(condition number of } P'_o)}$$

$$\tau_{\max} = \max_{o\in\mathcal{O}, s\in\mathcal{S}} \bar{\tau}(s,o)$$

- To assess the potential benefit of options, we study the ratio of the regrets SMDP/MDP:

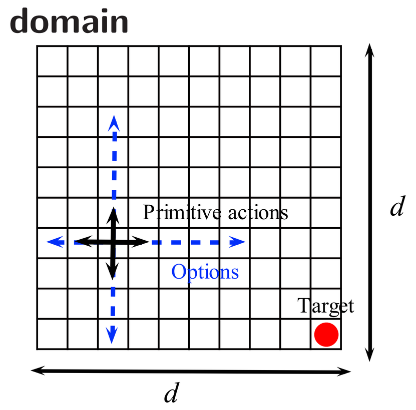
$$\mathcal{R} = \frac{\Delta(M, \textbf{SMDP-UCRL}, T_n)}{\Delta(M, \textbf{MDP-UCRL}, T_n)} \overset{?}{\le} 1$$

- Proxy for $\mathcal{R}$: ratio of the upper-bounds

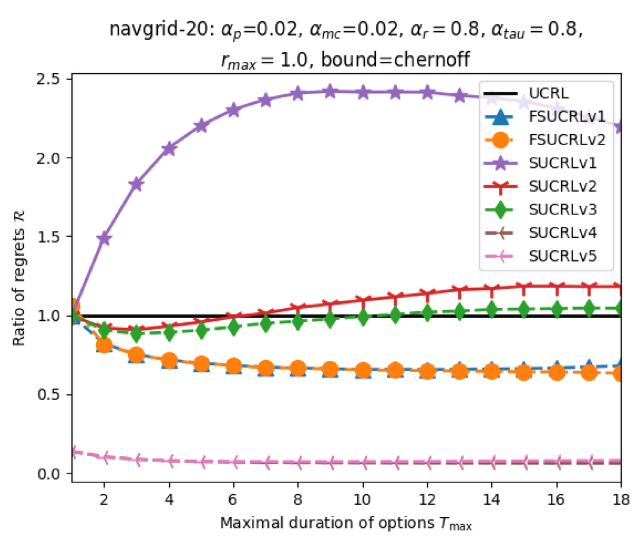
$$\widetilde{\mathcal{R}} \sim \frac{D'}{D}\sqrt{\frac{O}{A}}\sqrt{\frac{n}{T_n}}$$

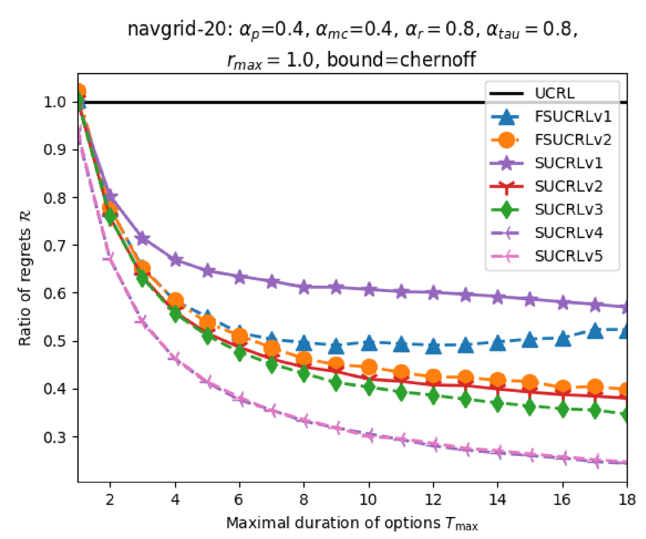
Experimental Results

Gridworld domain



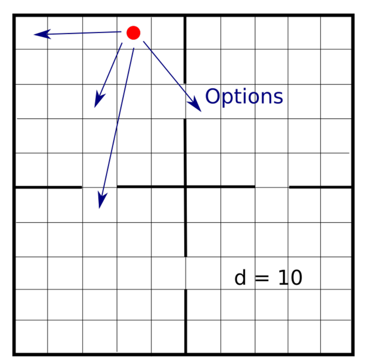
- The options are a.s. bounded $\tau \le T_{\max}$



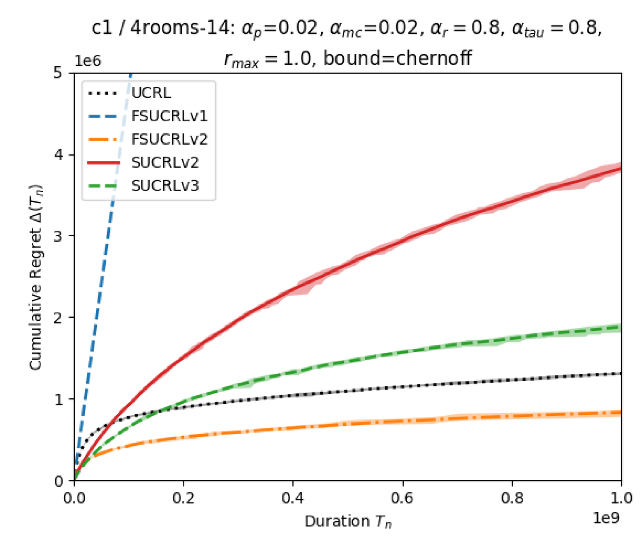


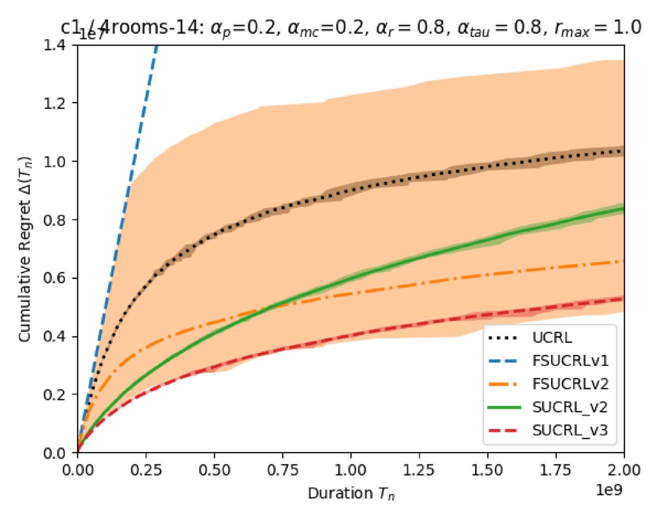
- FSUCRL always outperforms UCRL

Four Rooms Domain



- Size of state/action space is preserved





Conclusion

Take away message

- Temporal abstraction can be achieved without prior knowledge
  – better regret due to exploitation of correlations
  – but navigability within an option should not be compromised too much

Future work

- Can we leverage on these insights to design better options?
  – single task and transfer settings

References

[Brunskill et al. (2014)] Brunskill, E. and Li, L. Pac-inspired option discovery in lifelong RL. In *ICML*, 2014.

[Jaksch et al. (2010)] Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for RL. In *JMLR*, 2010.

[Jong et al. (2008)] Jong, N. K., Hester, T., and Stone, P. The utility of temporal abstraction in RL. In *AAMAS*, 2008.

[Sutton et al. (1999)] Sutton, R., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in RL. In *AI*, 1999.

[Fruit and Lazaric. (2017)] Fruit, R., and Lazaric, A., Exploration-Exploitation in MDPs with Options. In *AISTATS*, 2017.