Regret Minimization in MDPs with Options without Prior Knowledge

Ronan Fruit [1] **Matteo Pirotta** [1] **Alessandro Lazaric** [1] **Emma Brunskill** [2]

1. Introduction

Learning how to make good decisions in complex domains almost always requires some form of hierarchical reasoning. One powerful and popular framework for incorporating temporally-extended actions in the context of reinforcement learning is the *options* framework (Sutton et al., 1999). Creating and leveraging options has been the subject of many papers over the last two decades (see e.g., McGovern & Barto, 2001; Şimşek & Barto, 2004; Castro & Precup, 2012; Levy & Shimkin, 2011; Mann et al., 2014) and it has been of particular interest recently in combination with deep reinforcement learning (Tessler et al., 2016). However, incorporating options does not always improve learning efficiency or outcomes as shown by Jong et al. (2008). Therefore, we argue that it is important to build a formal understanding of how and when options may help or hurt reinforcement learning performance.

There has been fairly limited work on formal performance bounds of RL with options. Brunskill & Li (2014) derived sample complexity bounds for an RMax-like algorithm for semi-Markov decision processes (SMDPs) but their analysis cannot be immediately translated into the PAC-MDP sample complexity of learning with options. Fruit & Lazaric (2017) analyzed an SMDP variant of UCRL (Jaksch et al., 2010) and mapped its regret to the regret of learning in the original MDP with options. While their result makes explicit the impact of options on the learning performance, the proposed algorithm (UCRL-SMDP, or SUCRL in short) needs prior knowledge on the parameters of the distributions of cumulative rewards and durations of each option to construct confidence intervals. This strong requirement makes SUCRL not very practical in general, and ill-suited for option discovery.

In this paper we present an extension of SUCRL that combines the semi-Markov decision process view on options and the intrinsic MDP structure underlying their execution to achieve temporal abstraction without relying on unknown parameters. We introduce a transformation mapping each option to an associated irreducible Markov chain and we show that optimistic policies can be computed using only the stationary distributions of the irreducible chains and the SMDP dynamics. We propose an algorithm (Free-SUCRL, or FSUCRL) with provable regret guarantees that computes the stationary distribution of the options' irreducible Markov chains and its confidence intervals through an ad-hoc extended value iteration algorithm.

2. Preliminaries

A finite MDP is a tuple $M = \{\mathcal{S}, \mathcal{A}, p, r\}$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ the set of actions, $p(s'|s,a)$ the probability of transition from state s to state s' using a, $r(s,a)$ is the random reward associated to (s,a) with expectation $\overline{r}(s,a)$. A deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ maps states to actions. We define an option as a tuple $o = \{s_o, \beta_o, \pi_o\}$ where $s_o \in \mathcal{S}$ is the starting state[1], $\pi_o : \mathcal{S} \to \mathcal{A}$ the policy, and $\beta_o : \mathcal{S} \to [0,1]$ the probability of termination. As proved by Sutton et al. (1999), when $\mathcal{A}$ is replaced by a set of options $\mathcal{O}$, the resulting decision process is an SMDP $M_{\mathcal{O}} = \{\mathcal{S}_{\mathcal{O}}, \mathcal{O}, p_{\mathcal{O}}, R_{\mathcal{O}}, \tau_{\mathcal{O}}\}$ where $\mathcal{S}_{\mathcal{O}} \subseteq \mathcal{S}$ is the set of states where options can start and end, $p_{\mathcal{O}}(s'|s,o)$ is the probability of terminating in s' when starting o from s, $R_{\mathcal{O}}(s,o)$ is the cumulative reward and $\tau_{\mathcal{O}}(s,o)$ the duration (i.e., number of actions executed).[2] In the rest of the paper, we assume that options are well defined.

Assumption 1. *The set of options $\mathcal{O}$ is admissible that is, **1)** all options terminate in finite time with probability 1, **2)** in all possible terminal states at least one option can be started, **3)** SMDP $M_{\mathcal{O}}$ is communicating.*

Lem. 3 of Fruit & Lazaric (2017) shows that under Asm. 1, for all $o \in \mathcal{O}$, $R_{\mathcal{O}}(s,o)$ and $\tau_{\mathcal{O}}(s,o)$ have sub-Exponential distributions with parameters $(\sigma_R(o), b_R(o))$ and $(\sigma_\tau(o), b_\tau(o))$ respectively. The maximal expected duration is denoted by $\tau_{\max} = \max_{s,o} \{\overline{\tau}_{\mathcal{O}}(s,o)\}$. Let t denote primitive action steps

[1]INRIA Lille – Nord Europe [2]Stanford University. Correspondence to: Ronan Fruit <ronan.fruit@inria.fr>.

Accepted at Lifelong Learning: A Reinforcement Learning Approach Workshop @ICML, Sydney, Australia, 2017.

[1]Restricting the standard initial set to one state s_o is without loss of generality.

[2]Notice that $R_{\mathcal{O}}(s,o)$ (similarly for $\tau_{\mathcal{O}}$) is well defined only when $s = s_o$.

and let i index decision steps at option level. The number of decision steps up to (primitive) step t is $N(t) = \max\{n : T_n \leq t\}$, where $T_n = \sum_{i=1}^{n} \tau_i$ is the number of primitive steps executed over n decision steps and τ_i is the (random) number of steps before the termination of the option chosen at step i. Under Asm. 1 there exists a policy $\pi^* : \mathcal{S} \to \mathcal{O}$ over options that achieves the largest gain (per-step reward)

$$\rho^*_{\mathcal{O}} \overset{def}{=} \max_{\pi} \rho^{\pi}_{\mathcal{O}} = \max_{\pi} \lim_{t \to +\infty} \mathbb{E}^{\pi}\left[\frac{\sum_{i=1}^{N(t)} R_i}{t}\right], \quad (1)$$

where R_i is the reward cumulated by the option executed at step i. The optimal gain also satisfies the optimality equation of an equivalent MDP obtained by data-transformation

$$\rho^*_{\mathcal{O}} = \max_{o \in \mathcal{O}_s} \left\{ \frac{\overline{R}_{\mathcal{O}}(s,o)}{\overline{\tau}_{\mathcal{O}}(s,o)} + \frac{1}{\overline{\tau}_{\mathcal{O}}(s,o)} \left(\sum_{s' \in \mathcal{S}} p_{\mathcal{O}}(s'|s,o) u^*_{\mathcal{O}}(s') - u^*_{\mathcal{O}}(s) \right) \right\}, \quad (2)$$

where $u^*_{\mathcal{O}}$ is an optimal bias and $\mathcal{O}_s$ is the set of options than can be started in s (i.e., $o \in \mathcal{O}_s \Leftrightarrow s_o = s$). In the following sections, we drop the dependency on the option set $\mathcal{O}$ from all previous terms whenever clear from the context. Given the optimal average reward $\rho^*_{\mathcal{O}}$, we evaluate the performance of a learning algorithm $\mathfrak{A}$ by its cumulative (SMDP) regret over n decision steps as $\Delta(\mathfrak{A}, n) = \left(\sum_{i=1}^{n} \tau_i\right) \rho^*_{\mathcal{O}} - \sum_{i=1}^{n} R_i$. Fruit & Lazaric (2017) showed that $\Delta(\mathfrak{A}, n)$ is equal to the MDP regret up to an unavoidable linear "approximation" regret accounting for the difference between the optimal gains of M and $M_{\mathcal{O}}$.

3. Parameter-free SUCRL for Learning with Options

Optimism in SUCRL. At each episode, SUCRL runs a variant of extended value iteration (EVI) (Strehl & Littman, 2008) to solve the "optimistic" version of the data-transformation optimality equation in Eq. 2, i.e.,

$$\widetilde{\rho}^* = \max_{o \in \mathcal{O}_s} \left\{ \max_{\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{\tau}}} \left\{ \frac{\widetilde{R}(s,o)}{\widetilde{\tau}(s,o)} + \frac{1}{\widetilde{\tau}(s,o)} \left(\max_{\widetilde{\boldsymbol{p}}} \left\{ \sum_{s' \in \mathcal{S}} \widetilde{p}(s'|s,o) \widetilde{u}^*(s') \right\} - \widetilde{u}^*(s) \right) \right\} \right\}, \quad (3)$$

where $\widetilde{\boldsymbol{R}}$ and $\widetilde{\boldsymbol{\tau}}$ are the vectors of cumulative rewards and durations for all state-option pairs and they belong to confidence intervals constructed using parameters $(\sigma_R(o), b_R(o))$ and $(\sigma_\tau(o), b_\tau(o))$ (see Sect.3 in (Fruit & Lazaric, 2017) for the exact expression).[3] As a result, without any prior knowledge, such confidence intervals cannot be directly constructed and SUCRL cannot be run. In the following, we see how constructing an irreducible Markov chain (MC) associated to each option avoids this problem.

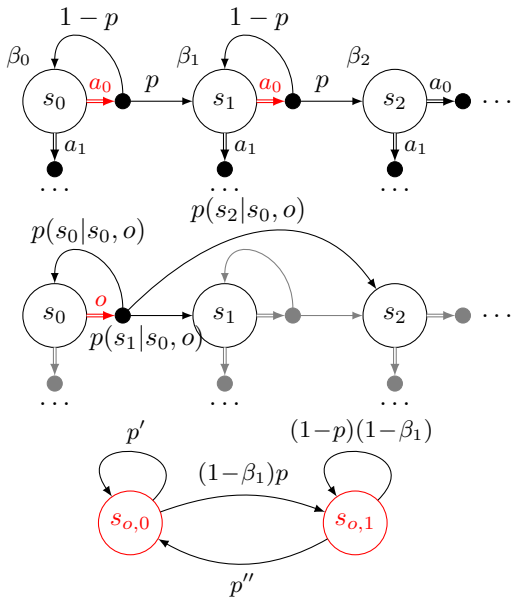

Figure 1. (top) MDP with an option o starting from s_0 and executing a_0 in all states with termination probabilities $\beta_o(s_0) = \beta_0$, $\beta_o(s_1) = \beta_1$ and $\beta_o(s_2) = 1$. *(middle)* SMDP dynamics associated to option o. *(bottom)* Irreducible MC obtained by transforming the associated absorbing MC with $p' = (1-\beta_0)(1-p) + \beta_0(1-p) + p\beta_1$ and $p'' = \beta_1(1-p) + p$.

3.1. Irreducible Markov Chains Associated to Options

A natural way to address SUCRL's limitations is to avoid considering options as atomic operations (as in SMDPs) but take into consideration their inner (MDP) structure. We notice from Eq. 2 that computing the optimal policy only requires computing the ratio $\overline{R}(s,o)/\overline{\tau}(s,o)$ and the inverse $1/\overline{\tau}(s,o)$. We can construct an irreducible MC with transition matrix P_o whose stationary distribution is directly related to these terms. We proceed as illustrated in Fig. 1: all transitions exiting the option are "merged" and "redirected" to the initial state s_o ($s_o = s_0$ in the figure). The set of states of the MC is denoted $\mathcal{S}_o$. To relate $\overline{R}(s,o)/\overline{\tau}(s,o)$ and $1/\overline{\tau}(s,o)$ to P_o we need an additional assumption on the options.

Assumption 2. *For any option $o \in \mathcal{O}$, the starting state s_o is also a terminal state ($\beta_o(s_o) = 1$) and any state $s' \in \mathcal{S}$ with $\beta_o(s') < 1$ is an inner state ($s' \in \mathcal{S}_o$).*

While the first part has a very minor impact on the

[3] Similarly, confidence intervals need to be computed for $\widetilde{\boldsymbol{p}}$, but this does not require any prior knowledge on the SMDP since the transition probabilities naturally belong to the simplex over states.

definition of $\mathcal{O}$, the second part of the assumption guarantees that options are "well designed" as it requires the termination condition to be coherent with the *true* inner states of the option, so that if $\beta_o(s') < 1$ then s' should be indeed reachable by the option. Under Asm. 2, P_o is an irreducible MC as any state can be reached starting from any other state and thus it admits a unique stationary distribution μ_o. We also have the following property.

Lemma 1. *Under Asm. 2, let $\boldsymbol{\mu_o}$ be the unique stationary distribution of the irreducible MC P_o associated to option o, then $\forall s \in \mathcal{S},\ \forall o \in \mathcal{O}_s$,*

$$\frac{1}{\overline{\tau}(s,o)} = \mu_o(s),\ \frac{\overline{R}(s,o)}{\overline{\tau}(s,o)} = \sum_{s' \in \mathcal{S}_o} \overline{r}(s', \pi_o(s'))\mu_o(s') \quad (4)$$

We can apply Lem. 1 to Eq. 3 and obtain the optimistic optimality equation

$$\widetilde{\rho}^* = \max_{o \in \mathcal{O}_s} \left\{ \max_{\widetilde{\boldsymbol{\mu}}_{\boldsymbol{o}}, \widetilde{\boldsymbol{r}}_{\boldsymbol{o}}} \left\{ \sum_{s' \in \mathcal{S}_o} \widetilde{r}_o(s')\, \widetilde{\mu}_o(s') + \widetilde{\mu}_o(s) \Big(\max_{\widetilde{\boldsymbol{b}}_{\boldsymbol{o}}} \{ \widetilde{\boldsymbol{b}}_{\boldsymbol{o}}^{\intercal} \widetilde{\boldsymbol{u}}^{\boldsymbol{*}} \} - \widetilde{u}^*(s) \Big) \right\} \right\}, \quad (5)$$

where $\widetilde{r}_o(s') = \widetilde{r}(s', \pi_o(s'))$ and $\widetilde{\boldsymbol{b}}_{\boldsymbol{o}} = (\widetilde{p}(s'|s,o))_{s' \in \mathcal{S}}$. Estimating μ_o implicitly leverages over the correlation between cumulative reward and duration, which is ignored when estimating $\overline{R}(s,o)$ and $\overline{\tau}(s,o)$ separately.

Now, we need to provide an explicit algorithm to compute the optimistic optimal gain $\widetilde{\rho}^*$ of Eq. 5 and its associated optimistic policy. In the next section, we introduce an algorithm that is guaranteed to compute an ε-optimistic policy.

3.2. SUCRL with Irreducible Markov Chains

The structure of the UCRL-like algorithm for learning with options (called FSUCRL) is reported in Alg. 2. Unlike SUCRL, for each option we do not directly estimate $\overline{R}(s,o)$ and $\overline{\tau}(s,o)$ but we estimate P_o, and the state-action reward $\overline{r}(s,a)$. As in SUCRL we also estimate the SMDP transition probabilities $p(s'|s,o)$. We can compute the respective confidence intervals $\beta_k^P(s,o,s')$, $\beta_k^r(s,a)$ and $\beta_k^p(s,o,s')$ (Hoeffding and empirical Bernstein) without any prior knowledge as (ignoring constants and logarithmic terms)

Input: Confidence $\delta \in]0,1[$, $r_{\max}$, $\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$
For episodes $k = 1, 2, \ldots$ **do**

1. Set $i_k := i$, $t = t_k$ and episode counters $\nu_k(s,a) = 0$, $\nu_k(s,o) = 0$
2. Compute estimates $\widehat{p}_k(s'|s,o)$, $\widehat{P}'_{o,k}$, $\widehat{r}_k(s,a)$ and their confidence intervals in Eq. 6
3. Compute an ϵ_k-approximation of the optimal optimistic policy $\widetilde{\pi}_k$ of Eq. 5
4. **While** $\forall l \in [t+1, t+\tau_i]$, $\nu_k(s_l, a_l) < N_k(s_l, a_l)$ **do**
   (a) Execute option $o_i = \widetilde{\pi}_k(s_i)$, obtain primitive rewards $r_i^1, \ldots, r_i^{\tau_i}$ and visited states $s_i^1, \ldots, s_i^{\tau_i} = s_{i+1}$
   (b) Set $\nu_k(s_i, o_i)$ += 1, i += 1, t += τ_i and $\nu_k(s, \pi_{o_i}(s))$ += 1 for all $s \in \{s_i^1, \ldots, s_i^{\tau_i}\}$
5. Set $N_k(s,o)$ += $\nu_k(s,o)$ and $N_k(s,a)$ += $\nu_k(s,a)$

Figure 2. The general structure of FSUCRL.

$$\beta_k^r(s,a) \propto r_{\max} \sqrt{\frac{1}{N_k(s,a)}}, \quad (6a)$$

$$\beta_k^p(s,o,s') \propto \sqrt{\frac{\widehat{p}_k(s'|s,o)\big(1 - \widehat{p}_k(s'|s,o)\big)}{N_k(s,o)}} + \frac{1}{N_k(s,o)}, \quad (6b)$$

$$\beta_k^P(s,o,s') \propto \sqrt{\frac{\widehat{P}_{o,k}(s,s')\big(1 - \widehat{P}_{o,k}(s,s')\big)}{N_k(s,\pi_o(s))}} + \frac{1}{N_k(s,\pi_o(s))}, \quad (6c)$$

where $N_k(s,a)$ (resp. $N_k(s,o)$) is the number of samples collected at state-action s,a (resp. state-option s,o) up to episode k, Eq. 6a coincides with the one used in UCRL, in Eq. 6b $s = s_o$ and $s' \in \mathcal{S}$, and in Eq. 6c $s, s' \in \mathcal{S}_o$.

To obtain an actual implementation of Alg. 2 we need to define an algorithm to compute an approximation of Eq. 5 (step 3). Similar to UCRL and SUCRL, we define an EVI algorithm starting from a function $u_0(s) = 0$ and computing at each iteration j

$$u_{j+1}(s) = \max_{o \in \mathcal{O}_s} \left\{ \max_{\widetilde{\boldsymbol{\mu}}_{\boldsymbol{o}}} \left\{ \sum_{s' \in \mathcal{S}_o} \widetilde{r}_o(s')\, \widetilde{\mu}_o(s') + \widetilde{\mu}_o(s) \left(\max_{\widetilde{\boldsymbol{b}}_{\boldsymbol{o}}} \left\{ \widetilde{\boldsymbol{b}}_{\boldsymbol{o}}^{\intercal} \boldsymbol{u_j} \right\} - u_j(s) \right) \right\} \right\} + u_j(s), \quad (7)$$

where $\widetilde{r}_o(s')$ is the optimistic state-action reward (i.e., estimate plus the confidence bound of Eq. 6a). Furthermore, we recall that the optimistic transition probability vector $\widetilde{\boldsymbol{b}}_{\boldsymbol{o}}$ can be computed using the algorithm introduced by Dann & Brunskill (2015) (App. A).

Nested extended value iteration. Our implementation of Alg. 2 builds on the observation that the maximum over $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{o}}$ in Eq. 7 can be seen as the opti-

mization of the average reward (gain)

$$\rho_o^*(u_j) = \max_{\widetilde{\boldsymbol{\mu}}_{\boldsymbol{o}}} \left\{ \sum_{s' \in \mathcal{S}_o} \zeta_o(s') \widetilde{\mu}_o(s') \right\}, \tag{8}$$

where ζ_o is defined as $\zeta_o(s_o) = \widetilde{r}_o(s_o) + \max_{\widetilde{\boldsymbol{b}}_{\boldsymbol{o}}} \{\widetilde{\boldsymbol{b}}_{\boldsymbol{o}}^\intercal \boldsymbol{u_j}\} - u_j(s_o)$ and $\zeta_o(s) = \widetilde{r}_o(s)$ for $s \neq s_o$. Eq. 8 is indeed the optimal gain of a bounded-parameter MDP with states $\mathcal{S}_o$, an action space composed of the option action (i.e., $\pi_o(s)$), and transitions $\widetilde{P}_o$ in the confidence intervals of Eq. 6c, and thus we can write its optimality equation

$$\rho_o^*(u_j) = \max_{\widetilde{P}_o} \left\{ \zeta_o(s) + \sum_{s'} \widetilde{P}_o(s,s') w_o^*(s') \right\} - w_o^*(s), \tag{9}$$

where w_o^* is an optimal bias. For any input function v we can compute $\rho_o^*(v)$ by using EVI on the bounded-parameter MDP, thus avoiding to explicitly construct confidence intervals on $\widetilde{\mu}_o$. As a result, we obtain two nested EVI algorithms where, starting from an initial bias function $v_0(s) = 0$,[4] at any iteration j we set the bias function of the inner EVI to $w_{j,0}^o(s) = 0$ and we compute

$$w_{j,l+1}^o(s') = \max_{\widetilde{P}_o} \left\{ \zeta_o(s) + \widetilde{\boldsymbol{P}}_{\boldsymbol{o}}(\cdot|\boldsymbol{s'})^\intercal \boldsymbol{w_{j,l}^o} \right\}, \tag{10}$$

until the stopping condition $l_j^o = \inf\{l \geq 0 : \mathrm{sp}\{\boldsymbol{w_{j,l+1}^o} - \boldsymbol{w_{j,l}^o}\} \leq \varepsilon_j\}$ is met, where $(\varepsilon_j)_{j \geq 0}$ is a vanishing sequence. As $w_{j,l+1}^o - w_{j,l}^o$ converges to $\rho_o^*(v_j)$ with l, the outer EVI becomes

$$v_{j+1}(s) = \max_{o \in \mathcal{O}_s} \left\{ g\big(\boldsymbol{w_{j,l_j^o+1}^o} - \boldsymbol{w_{j,l_j^o}^o}\big) \right\} + v_j(s), \tag{11}$$

where $g : \boldsymbol{v} \mapsto \frac{1}{2}(\max\{\boldsymbol{v}\} + \min\{\boldsymbol{v}\})$. It can be shown that this nested scheme converges to the solution of Eq. 5 and reaches ε-accuracy if the algorithm is stopped when $\mathrm{sp}\{\boldsymbol{v_{j+1}} - \boldsymbol{v_j}\} + \varepsilon_j \leq \varepsilon$. One of the interesting features of this algorithm is its hierarchical structure. Nested EVI is operating on two different time scales by iteratively considering every option as an independent optimistic planning sub-problem (EVI of Eq. 10) and gathering all the results into a higher level planning problem (EVI of Eq. 11). This idea is at the core of the hierarchical approach in RL, but it is not always present in the algorithmic structure, while nested EVI naturally arises from decomposing Eq. 7 in two value iteration algorithms.

4. Theoretical Analysis

Before presenting the guarantees for FSUCRL, we recall the definition of diameter of M and $M_{\mathcal{O}}$:

$$D = \max_{s,s' \in \mathcal{S}} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}\big[\tau_\pi(s,s')\big],$$
$$D_{\mathcal{O}} = \max_{s,s' \in \mathcal{S}_{\mathcal{O}}} \min_{\pi : \mathcal{S}_{\mathcal{O}} \to \mathcal{O}} \mathbb{E}\big[\tau_\pi(s,s')\big],$$

where $\tau_\pi(s,s')$ is the (random) number of primitive actions to move from s to s' following policy π. We also associate to each option o a *pseudo-diameter* $\widetilde{D}_o = (\mathrm{sp}(r_o)\kappa_o^1 + r_{\max}\tau_o\kappa_o^\infty)/\sqrt{\min_{s \in \mathcal{S}_o} \mu_o(s)}$, where κ_o^1 and κ_o^∞ are condition numbers of the irreducible MC P_o associated to option o (for the ℓ_1 and ℓ_∞-norm respectively (Cho & Meyer, 2001)), $\tau_o = \overline{\tau}(s_o, o)$ and $\mathrm{sp}(r_o)$ is the span of the rewards in the inner states of the option. We proved the following bound

Theorem 1. *Let M be a communicating MDP with reward bounded by $r_{\max} = 1$ and let $\mathcal{O}$ be a set of options satisfying Asm. 1 and 2 such that $\sigma_R(o) \leq \sigma_R$, $\sigma_\tau(o) \leq \sigma_\tau$. We also define $B_{\mathcal{O}} = \max_{s,o} supp(p(\cdot|s,o))$ (resp. $B = \max_{s,a} supp(p(\cdot|s,a))$) as the largest support of the SMDP (resp. MDP) dynamics and $\widetilde{D}_{\mathcal{O}} = \max_o \widetilde{D}_o$. Then its regret is bounded as (ignoring constants and logarithmic terms)*

$$\Delta(\mathrm{FSUCRL}, n) = \widetilde{O}\Bigg(\underbrace{D_{\mathcal{O}}\sqrt{B_{\mathcal{O}} S O n}}_{\Delta_p} + \underbrace{(\sigma_R + \sigma_\tau)\sqrt{n}}_{\Delta_{R,\tau}} + \underbrace{\sqrt{SAT_n} + \widetilde{D}_{\mathcal{O}}\sqrt{BSOT_n}}_{\Delta_\mu} \Bigg) \tag{12}$$

Comparison to UCRL. The regret bound for UCRL is $D\sqrt{BSAT_n}$. As for SUCRL, the main term Δ_p in the regret of FSUCRL scales as $\sqrt{n}$ while UCRL scales as $\sqrt{T_n}$: this is the effect (and potential benefit) of temporal abstraction. There are also additive terms $\Delta_{R,\tau}$ and Δ_μ characterizing the complexity of learning the options. While in general this additional regret may be large, we show empirically in the next section that FSUCRL can perform much better than both UCRL and SUCRL. Moreover, like SUCRL, FSUCRL can benefit from a reduction of the state-action space ($SO < SA$). But Eq. 12 reveals that options can also improve the learning speed when $B_{\mathcal{O}} < B$. This can lead to a huge improvement e.g., when options are designed so as to reach a specific goal (the transition probability may be almost deterministic in this case, even if the transitions at primitive actions are very stochastic). This potential advantage extends the results of Fruit & Lazaric (2017) and comes from the use of Bernstein (instead of Chernoff) confidence intervals on $p_{\mathcal{O}}$. This also matches the intuition of what are "good" options often present in the literature.

5. Numerical Simulations

In this section we compare the regret of FSUCRL to SUCRL and UCRL to verify the impact of removing

[4] We use v_j instead of u_j since the errors in the inner EVI generate a sequence of functions different from $\{u_j\}$.

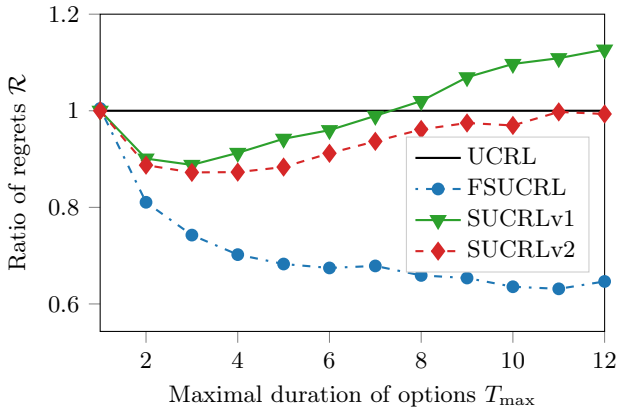

Figure 3. Regret after $1.2 \cdot 10^8$ steps normalized w.r.t. UCRL for different option durations in a 20x20 grid-world.

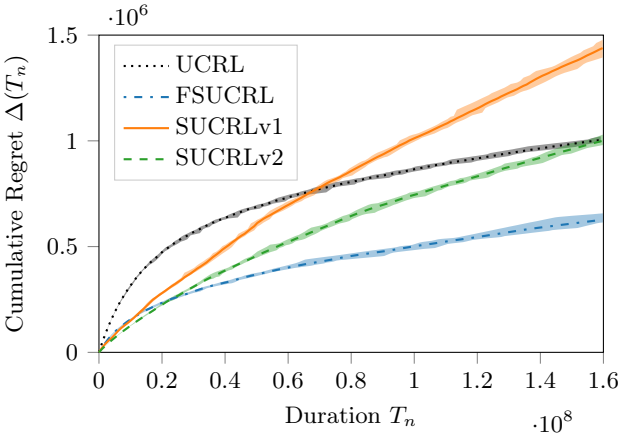

Figure 4. Evolution of the regret as T_n increases for a 14x14 four-room maze.

prior knowledge about options and estimating their structure through the irreducible MC transformation.

5.1. Temporal abstraction

We first consider the toy domain analysed by Fruit & Lazaric (2017) that was specifically designed to show the advantage of temporal abstraction. It is an instance of a grid-world navigation problem where the 4 cardinal actions are replaced by 4 cardinal options with various maximal duration $T_{\max}$.[5] The optimal policy is the shortest path to a target state that triggers a random restart in the grid. The reward is zero everywhere except at the target where it is $r_{\max} = 1$. To be able to reproduce the results of Fruit & Lazaric (2017), we ran our algorithms with Hoeffding confidence bounds for the ℓ_1-deviation of the empirical distribution (implying that B and $B_{\mathcal{O}}$ have no impact in our simulations).

Interpretation of the results. Regret On Fig. 3 we plot the value of the ratio $\mathcal{R} = \Delta(\mathfrak{A}, n)/\Delta(\text{UCRL}, n)$ where $n = N(1.2 \cdot 10^8)$ (with $N(t) = \max\{n : T_n \leq t\}$) and $\mathfrak{A} \in \{\text{SUCRL}, \text{FSUCRL}\}$ with different sets of options characterized by the maximal duration $T_{\max}$. When the ratio is smaller than 1, $\mathfrak{A}$ performs better than UCRL and conversely. Trivially, when $T_{\max} = 1$, all algorithms are equivalent to UCRL. The value of n is big enough for all algorithms to have explored the environment extensively: for $t \geq 1.2 \cdot 10^8$ the regret increases only logarithmically and the value of the ratio is stable. When comparing FSUCRL to UCRL, we empirically observe that the advantage of temporal abstraction is indeed preserved when removing the knowledge of the parameters of the option (blue curve on Fig. 3). This shows that the benefit of temporal abstraction is not just a mere artefact of prior knowledge on the options: it can be achieved without any additional information w.r.t. UCRL. The two versions of SUCRL plotted on Fig. 3 differ in the amount of prior knowledge given to the algorithm: SUCRLv1 receives $r_{\max}$, $\tau_{\max}$, $\max_o\{\sigma_\tau(o)\}$ and $\max_o\{b_\tau(o)\} = 0$, while SUCRLv2 uses $r_{\max}$, $(\tau_o)_{o\in\mathcal{O}}$, $(\sigma_\tau(o))_{o\in\mathcal{O}}$ (option dependent quantities) and $\max_o\{b_\tau(o)\} = 0$.[6] As expected, the more prior knowledge, the better the regret (curves green and red on Fig. 3). Unlike FSUCRL, SUCRL is highly sensitive to the knowledge we have on the distributions of $R_{\mathcal{O}}$ and $\tau_{\mathcal{O}}$. In particular, if our knowledge on $R_{\mathcal{O}}$ and $\tau_{\mathcal{O}}$ is very loose, SUCRL can even perform worse than UCRL for all values of $T_{\max}$. Although we expect SUCRL to perform better than FSUCRL due to the additional knowledge provided to the algorithm, the fact that the blue curve is always below all other curves can be explained by the fact that FSUCRL not only exploits correlations between options sharing state-action pairs (by collecting samples at action level and not at option level like SUCRL), but it also leverages over the correlation between $R_{\mathcal{O}}$ and $\tau_{\mathcal{O}}$ within a single option (by being optimistic on the ratio $\overline{R}_{\mathcal{O}}/\overline{\tau}_{\mathcal{O}}$ directly through the stationary distribution instead of $\overline{R}_{\mathcal{O}}$ and $\overline{\tau}_{\mathcal{O}}$ separately as in SUCRL).

5.2. Four-room maze

We now consider the famous four-room environment introduced by Sutton et al. (1999) with the four cardinal actions having a probability 0.2 of failure (uniformly in any other direction). The grid-world is a square of dimension 14x14 with every room being a square of dimension 7x7. Each room has exactly two exit doors. In every state of every room, we define four options: two are leading to the two exit doors, one is

[5] $T_{\max}$ is the maximal *actual* duration as opposed to the *maximal* expected duration $\tau_{\max} \leq T_{\max}$.

[6] We computed $\sigma_\tau(o)$ based on the analytical formula relating $\sigma_\tau(o)$ to the dynamics of o. Moreover, given the knowledge of $r_{\max}$, τ_o and $\sigma_\tau(o)$, the tightest bound on $\sigma_R(o)$ is: $\sigma_R(o) \leq r_{\max}\sqrt{\tau_o + \sigma_\tau(o)^2}$. In this specific problem $\max_o\{b_R(o)\} = \max_o\{b_\tau(o)\} = 0$.

leading to the center of the room, and the last one leads to the unique corner of the grid in the room. Thus, the number of state-options is slightly bigger than the number of state-actions. The optimal policy takes the shortest path to the target state which is located in one of the 4 corners of the grid and the rewards are the same as in the previous experiment. Once the target is reached, the next state is chosen uniformly at random in the grid. Like in the previous experiments, we ran our algorithms with Hoeffding confidence bounds for the ℓ_1-deviation of the empirical distribution.

Interpretation of the results. On Fig. 4, we plot the regret $\Delta(\mathfrak{A}, n)$ as a function of T_n for $\mathfrak{A} \in \{\text{UCRL}, \text{SUCRL}, \text{FSUCRL}\}$. The two versions of SUCRL are exactly the same as in the previous experiments: SUCRLv1 uses $\max_o\{\sigma_\tau(o)\}$ while SUCRLv2 uses $(\sigma_\tau(o))_{o\in\mathcal{O}}$. On this example, both versions of SUCRL fail to beat UCRL. However, FSUCRL has nearly half the regret of UCRL.

In both experiments, UCRL and FSUCRL had similar running times meaning that the improvement in cumulative regret is not at the expense of the computational complexity.

Since FSUCRL does not require strong prior knowledge about options and its regret bound is partially computable, we believe the results of this paper could be used as a basis to construct more principled option discovery algorithms that explicitly optimize the exploration-exploitation performance of the learning algorithm. Finally, the hierarchical structure of FSUCRLv2 calls for further investigation on how to generalize the approach to more than two levels, e.g., options-over-options...

References

Brunskill, Emma and Li, Lihong. PAC-inspired Option Discovery in Lifelong Reinforcement Learning. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *JMLR Proceedings*, pp. 316–324. JMLR.org, 2014.

Castro, Pablo Samuel and Precup, Doina. Automatic construction of temporally extended actions for mdps using bisimulation metrics. In *Proceedings of the 9th European Conference on Recent Advances in Reinforcement Learning*, EWRL'11, pp. 140–152, Berlin, Heidelberg, 2012. Springer-Verlag.

Cho, Grace E. and Meyer, Carl D. Comparison of perturbation bounds for the stationary distribution of a markov chain. *Linear Algebra and its Applications*, 335(1):137 – 150, 2001.

Şimşek, Özgür and Barto, Andrew G. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, 2004.

Dann, Christoph and Brunskill, Emma. Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pp. 2818–2826, Cambridge, MA, USA, 2015. MIT Press.

Fruit, Ronan and Lazaric, Alessandro. Exploration–exploitation in mdps with options. In *Proceedings of Machine Learning Research*, volume 54: Artificial Intelligence and Statistics, 20-22 April 2017, Fort Lauderdale, FL, USA, pp. 576–584, 2017.

Jaksch, Thomas, Ortner, Ronald, and Auer, Peter. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010.

Jong, Nicholas K. Hester, Todd, and Stone, Peter. The utility of temporal abstraction in reinforcement learning. In *The Seventh International Joint Conference on Autonomous Agents and Multiagent Systems*, May 2008.

Levy, Kfir Y. and Shimkin, Nahum. Unified inter and intra options learning using policy gradient methods. In Sanner, Scott and Hutter, Marcus (eds.), *EWRL*, volume 7188 of *Lecture Notes in Computer Science*, pp. 153–164. Springer, 2011.

Mann, Timothy Arthur, Mankowitz, Daniel J., and Mannor, Shie. Time-regularized interrupting options (TRIO). In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1350–1358, 2014.

McGovern, Amy and Barto, Andrew G. Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 361–368, 2001.

Strehl, Alexander L. and Littman, Michael L. An analysis of model-based interval estimation for markov decision processes. *J. Comput. Syst. Sci.*, 74(8):1309–1331, December 2008.

Sutton, Richard S., Precup, Doina, and Singh, Satinder. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181 – 211, 1999.

Tessler, Chen, Givony, Shahar, Zahavy, Tom, Mankowitz, Daniel J., and Mannor, Shie. A deep hierarchical approach to lifelong learning in minecraft. *CoRR*, abs/1604.07255, 2016.