Exploration–Exploitation in MDPs with Options

Ronan Fruit
Inria Lille - SequeL Team

Alessandro Lazaric
Inria Lille - SequeL Team

Abstract

While a large body of empirical results show that temporally-extended actions and options may significantly affect the learning performance of an agent, the theoretical understanding of how and when options can be beneficial in online reinforcement learning is relatively limited. In this paper, we derive an upper and lower bound on the regret of a variant of UCRL using options. While we first analyze the algorithm in the general case of semi-Markov decision processes (SMDPs), we show how these results can be translated to the specific case of MDPs with options and we illustrate simple scenarios in which the regret of learning with options can be *provably* much smaller than the regret suffered when learning with primitive actions.

1 Introduction

The option framework [Sutton et al., 1999] is a simple yet powerful model to introduce temporally-extended actions and hierarchical structures in reinforcement learning (RL) [Sutton and Barto, 1998]. An important feature of this framework is that Markov decision process (MDP) planning and learning algorithms can be easily extended to accommodate options, thus obtaining algorithms such as option value iteration and Q-learning [Sutton et al., 1999], LSTD [Sorg and Singh, 2010], and actor-critic [Bacon and Precup, 2015]. Temporally extended actions are particularly useful in high dimensional problems that naturally decompose into a hierarchy of subtasks. For instance, Tessler et al. [2016] recently obtained promising results by combining options and deep learning for lifelong learning in the challenging domain of Minecraft. A large body of the literature has then focused on how to automatically construct options that are beneficial to the learning process within a single task or across similar tasks (see e.g., [McGovern and Barto, 2001, Menache et al., 2002, Şimşek and Barto, 2004, Castro and Precup, 2012, Levy and Shimkin, 2011]). An alternative approach is to design an initial set of options and optimize it during the learning process itself (see e.g., interrupting options Mann et al. 2014 and options with exceptions Sairamesh and Ravindran 2012). Despite the empirical evidence of the effectiveness of most of these methods, it is well known that options may as well worsen the performance w.r.t. learning with primitive actions [Jong et al., 2008]. Moreover, most of the proposed methods are heuristic in nature and the theoretical understanding of the actual impact of options on the learning performance is still fairly limited. Notable exceptions are the recent results of Mann and Mannor [2014] and Brunskill and Li [2014]. Nonetheless, Mann and Mannor [2014] rather focus on a batch setting and they derive a sample complexity analysis of approximate value iteration with options. Furthermore, the PAC-SMDP analysis of Brunskill and Li [2014] describes the performance in SMDPs but it cannot be immediately translated into a sample complexity of learning with options in MDPs.

In this paper, we consider the case where a fixed set of options is provided and we study their impact on the learning performance w.r.t. learning without options. In particular, we derive the first regret analysis of learning with options. Relying on the fact that using options in an MDP induces a semi-Markov decision process (SMDP), we first introduce a variant of the UCRL algorithm [Jaksch et al., 2010] for SMDPs and we upper- and lower-bound its regret (sections 3 and 4). While this result is of independent interest for learning in SMDPs, its most interesting aspect is that it can be translated into a regret bound for learning with options in MDPs and it provides a first understanding on the conditions sufficient for a set of options to reduce the regret w.r.t. learning with primitive actions (Sect. 5). Finally, we provide an illustrative example where the empirical results support the theoretical findings (Sect. 6).

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

2 Preliminaries

MDPs and options. A finite MDP is a tuple $M = \{\mathcal{S}, \mathcal{A}, p, r\}$ where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $p(s'|s,a)$ is the probability of transitioning from state s to state s' when action a is taken, $r(s,a,s')$ is a distribution over rewards obtained when action a is taken in state s and the next state is s'. A stationary deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ maps states to actions. A (Markov) option is a tuple $o = \{\mathcal{I}_o, \beta_o, \pi_o\}$ where $\mathcal{I}_o \subset \mathcal{S}$ is the set of states where the option can be initiated, $\beta_o : \mathcal{S} \to [0,1]$ is the probability distribution that the option ends in a given state, and $\pi_o : \mathcal{S} \to \mathcal{A}$ is the policy followed until the option ends. Whenever the set of primitive actions $\mathcal{A}$ is replaced by a set of options $\mathcal{O}$, the resulting decision process is no longer an MDP but it belongs to the family of semi-Markov decision processes (SMDP).

Proposition 1. *[Sutton et al. 1999] For any MDP M and 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,*

$$\mathcal{S}_{\mathcal{O}} = \Big(\cup_{o\in\mathcal{O}} \mathcal{I}_o\Big) \bigcup \Big(\cup_{o\in\mathcal{O}} \{s : \beta_o(s) > 0\}\Big),$$

$\mathcal{O}$ is the set of available actions, $p_{\mathcal{O}}(s,o,s')$ is the probability of transition from s to s' by taking the policy π_o associated to option o, i.e.,

$$p_{\mathcal{O}}(s,o,s') = \sum_{k=1}^{\infty} p(s_k = s'|s,\pi_o)\beta_o(s'),$$

where $p(s_k = s'|s,\pi_o)$ is the probability of reaching state s' after exactly k steps following policy π_o, $r_{\mathcal{O}}(s,o,s')$ is the distribution of the cumulative reward obtained by executing option o from state s until interruption at s', and $\tau_{\mathcal{O}}(s,o,s')$ is the distribution of the holding time (i.e., number of primitive steps executed to go from s to s' by following π_o).

Throughout the rest of the paper, we only consider an "admissible" set of options $\mathcal{O}$ such that all options terminate in finite time with probability 1 and in all possible terminal states there exists at least one option that can start, i.e., $\cup_{o\in\mathcal{O}}\{s : \beta_o(s) > 0\} \subseteq \cup_{o\in\mathcal{O}}\mathcal{I}_o$. This also implies that the resulting SMDP $M_{\mathcal{O}}$ is communicating whenever the original MDP M is communicating. Finally, we notice that a stationary deterministic policy constructed on a set of options $\mathcal{O}$ may result into a non-stationary policy on the set of actions $\mathcal{A}$.

Learning in SMDPs. Relying on this mapping, we first study the exploration-exploitation trade-off in a generic SMDP. A thorough discussion on the implications of the analysis of learning in SMDPs for the case of learning with options in MDPs is reported in Sect. 5. For any SMDP $M = \{\mathcal{S}, \mathcal{A}, p, r, \tau\}$, we denote by $\overline{\tau}(s,a,s')$ (resp. $\overline{r}(s,a,s')$) the expectation of $\tau(s,a,s')$ (resp. $r(s,a,s')$) and by $\overline{\tau}(s,a) = \sum_{s'\in\mathcal{S}} \overline{\tau}(s,a,s')p(s'|s,a)$ (resp. $\overline{r}(s,a) = \sum_{s'\in\mathcal{S}} \overline{r}(s,a,s')p(s'|s,a)$) the expected holding time (resp. cumulative reward) of action a from state s. In the next proposition we define the average-reward performance criterion and we recall the properties of the optimal policy in SMDPs.

Proposition 2. *Denote $N(t) = \sup\{n \in \mathbb{N},\ \sum_{i=1}^{n} \tau_i \le t\}$ the number of decision steps that occurred before time t. For any policy π and $s \in \mathcal{S}$:*

$$\begin{aligned} \overline{\rho}^{\pi}(s) &\overset{def}{=} \limsup_{t\to+\infty} \mathbb{E}^{\pi}\left[\frac{\sum_{i=1}^{N(t)} r_i}{t}\middle| s_0 = s\right] \\ \underline{\rho}^{\pi}(s) &\overset{def}{=} \liminf_{t\to+\infty} \mathbb{E}^{\pi}\left[\frac{\sum_{i=1}^{N(t)} r_i}{t}\middle| s_0 = s\right]. \end{aligned} \tag{1}$$

If M is communicating and the expected holding times and reward are finite, there exists a stationary deterministic optimal policy π^* such that for all states s and policies π, $\underline{\rho}^{\pi^*}(s) \ge \overline{\rho}^{\pi}(s)$ and $\overline{\rho}^{\pi^*}(s) = \underline{\rho}^{\pi^*}(s) = \rho^*$.

Finally, we recall the average reward optimality equation for a communicating SMDP

$$u^*(s) = \max_{a\in\mathcal{A}} \Big\{\overline{r}(s,a) - \rho^*\overline{\tau}(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)u^*(s')\Big\}, \tag{2}$$

where u^* and ρ^* are the bias (up to a constant) and the gain of the optimal policy π^*.

We are now ready to consider the learning problem. For any $i \in \mathbb{N}^*$, a_i denotes the action taken by the agent at the i-th decision step[1] and s_i denotes the state reached after a_i is taken, with s_0 being the initial state. We denote by $(r_i(s,a,s'))_{i\in\mathbb{N}^*}$ (resp. $(\tau_i(s,a,s'))_{i\in\mathbb{N}^*}$) a sequence of i.i.d. realizations from distribution $r(s,a,s')$ (resp. $\tau(s,a,s')$). When the learner explores the SMDP, it observes the sequence $(s_0, \ldots, s_i, a_{i+1}, r_{i+1}(s_i,a_{i+1},s_{i+1}), \tau_{i+1}(s_i,a_{i+1},s_{i+1}), \ldots)$. The performance of a learning algorithm is measured in terms of its cumulative *regret*.

Definition 1. *For any SMDP M, any starting state $s \in \mathcal{S}$, and any number of decision steps $n \ge 1$, let $\{\tau_i\}_{i=1}^{n}$ be the random holding times observed along the trajectory generated by a learning algorithm $\mathfrak{A}$. Then the total regret of $\mathfrak{A}$ is defined as*

$$\Delta(M, \mathfrak{A}, s, n) = \Big(\sum_{i=1}^{n} \tau_i\Big)\rho^*(M) - \sum_{i=1}^{n} r_i. \tag{3}$$

[1] Notice that decision steps are discrete points in time in which an action is started, while the (possibly continuous) holding time is determined by the distribution τ.

We first notice that this definition reduces to the standard regret in MDPs for $\tau_i = 1$ (i.e., primitive actions always terminate in one step). The regret measures the difference in cumulative reward obtained by the optimal policy and the learning algorithm. While the performance of the optimal policy is measured by its asymptotic average reward ρ^*, the total duration after n decision steps may vary depending on the policy. As a result, when comparing the performance of π^* after n decision steps, we multiply it by the length of the trajectory executed by the algorithm $\mathfrak{A}$. More formally, from the definition of ρ^* (Eq. 1) and Prop. 2 we have[2]

$$\mathbb{E}^{\pi^*}\left[\sum_{i=1}^{N(t)} r_i \Big| s_0 = s\right] \underset{t\to+\infty}{\sim} \rho^* t + o(t).$$

By introducing the total duration $N(t)$ of $\mathfrak{A}$ we have

$$\rho^* t + o(t) = \rho^*\left(\sum_{i=1}^{N(t)} \tau_i\right) + \rho^*\left(t - \sum_{i=1}^{N(t)} \tau_i\right) + o(t).$$

We observe that $\left(t - \sum_{i=1}^{N(t)} \tau_i\right) = o(t)$ almost surely since $\left(t - \sum_{i=1}^{N(t)} \tau_i\right) \leq \tau_{N(t)+1}$ and $\tau_{N(t)+1}$ is bounded by an almost surely finite (a.s.) random variable since the expected holding time for all state-action pairs is bounded by assumption. So $\tau_{N(t)+1}/t \underset{t\to+\infty}{\to} 0$ a.s. and

$$\mathbb{E}^{\pi^*}\left[\sum_{i=1}^{N(t)} r_i \Big| s_0 = s\right] \underset{t\to+\infty}{\sim} \rho^*\left(\sum_{i=1}^{N(t)} \tau_i\right) + o(t),$$

which justifies the definition of the regret.

3 SMDP-UCRL

In this section we introduce UCRL-SMDP (Fig. 1), a variant of UCRL [Jaksch et al., 2010]. At each episode k, the set of plausible SMDPs $\mathcal{M}_k$ is defined by the current estimates of the SMDP parameters and a set of constraints on the rewards, the holding times and the transition probabilities derived from the confidence intervals. Given $\mathcal{M}_k$, extended value iteration (EVI) finds an SMDP $\widetilde{M}_k \in \mathcal{M}_k$ that maximizes $\rho^*(\widetilde{M}_k)$ and the corresponding optimal policy $\widetilde{\pi}_k^*$ is computed. To solve this problem, we note that it can be equivalently formulated as finding the optimal policy of an extended[3] SMDP $\widetilde{M}_k^+$ obtained by combining all SMDPs in $\mathcal{M}_k$: $\widetilde{M}_k^+$ has the same state space and an extended continuous action space $\widetilde{\mathcal{A}}_k^+$. Choosing an action $a^+ \in \widetilde{\mathcal{A}}_k^+$ amounts to choosing an action $a \in \mathcal{A}$, a reward $\widetilde{r}_k(s,a)$, a holding time $\widetilde{\tau}_k(s,a)$ and a transition probability $\widetilde{p}_k(\cdot \mid s,a)$ in the confidence intervals. When a^+ is executed in $\widetilde{M}_k^+$, the probability, the expected reward and the expected holding time of the transition are respectively $\widetilde{p}_k(\cdot \mid s,a)$, $\widetilde{r}_k(s,a)$ and $\widetilde{\tau}_k(s,a)$. Finally, $\widetilde{\pi}_k^*$ is executed until the number of samples for a state-action pair is doubled. Since the structure is similar to UCRL's, we focus on the elements that need to be rederived for the specific SMDP case: the confidence intervals construction and the extended value iteration algorithm.

[2] In this expectation, $N(t)$ is a r.v. depending on π^*.

[3] In the MDP literature, the term Bounded Parameter MDPs (BPMDPs) [Tewari and Bartlett, 2007] is often used for "extended" MDPs built using confidence intervals on rewards and transition probabilities.

Input: Confidence $\delta \in]0,1[$, $\mathcal{S}$, $\mathcal{A}, b_r, \sigma_r, b_\tau, \sigma_\tau, R_{\max}, \tau_{\max}$ and $\tau_{\min}$.
Initialization: Set $i = 1$, and observe initial state s_0.

For episodes $k = 1, 2, \ldots$ **do**
Initialize episode k:

1. Set the start step of episode k, $i_k := i$
2. For all (s,a) initialize the counter for episode k, $\nu_k(s,a) := 0$ and set counter prior to episode k,

$$N_k(s,a) = \#\{\iota < i_k : s_\iota = s, a_\iota = a\}$$

3. For s, s', a set the accumulated rewards, duration and transition counts prior to episode k,

$$R_k(s,a) = \sum_{\iota=1}^{i_k-1} r_\iota \mathbb{1}_{s_\iota=s, a_\iota=a}, T_k(s,a) = \sum_{\iota=1}^{i_k-1} \tau_\iota \mathbb{1}_{s_\iota=s, a_\iota=a}$$
$$P_k(s,a,s') = \#\{\iota < i_k : s_\iota = s, a_\iota = a, s_{\iota+1} = s'\}$$

Compute estimates $\hat{p}_k(s' \mid s,a) := \frac{P_k(s,a,s')}{\max\{1, N_k(s,a)\}}$ and $\hat{\tau}_k(s,a) := \frac{T_k(s,a)}{N_k(s,a)}$ and $\hat{r}_k(s,a) := \frac{R_k(s,a)}{N_k(s,a)}$

Compute policy $\widetilde{\pi}_k$:

4. Let $\mathcal{M}_k$ be the set of all SMDPs with states and actions as in M, and with transition probabilities $\widetilde{p}$, rewards $\widetilde{r}$, and holding time $\widetilde{\tau}$ such that for any (s,a)

$$|\widetilde{r} - \hat{r}_k| \leq \beta_k^r \quad \text{and} \quad R_{\max}\tau_{\max} \geq \widetilde{r}(s,a) \geq 0$$
$$|\widetilde{\tau} - \hat{\tau}_k| \leq \beta_k^\tau \quad \text{and} \quad \tau_{\max} \geq \widetilde{\tau}(s,a) \geq \widetilde{r}(s,a)/R_{\max}, \tau_{\min}$$
$$\|\widetilde{p}(\cdot) - \hat{p}_k(\cdot)\|_1 \leq \beta_k^p \quad \text{and} \quad \sum_{s' \in \mathcal{S}} \widetilde{p}(s' \mid s,a) = 1$$

5. Use *extended value iteration* (EVI) to find a policy $\widetilde{\pi}_k$ and an optimistic SMDP $\widetilde{M}_k \in \mathcal{M}_k$ such that:

$$\widetilde{\rho}_k := \min_s \rho(\widetilde{M}_k, \widetilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s} \rho(M', \pi, s) - \frac{R_{\max}}{\sqrt{i_k}}$$

Execute policy $\widetilde{\pi}_k$:

6. **While** $\nu_k(s_i, \widetilde{\pi}_k(s_i)) < \max\{1, N_k(s_i, \widetilde{\pi}_k(s_i))\}$ **do**
   (a) Choose action $a_i = \widetilde{\pi}_k(s_i)$, obtain reward r_i, and observe next state s_{i+1}
   (b) Update $\nu_k(s_i, a_i) := \nu_k(s_i, a_i) + 1$ and set $i = i + 1$

Figure 1: UCRL-SMDP

Confidence intervals. Unlike in MDPs, we consider

a slightly more general scenario where cumulative rewards and holding times are not bounded but are sub-exponential r.v. (see Lem. 3). As a result, the confidence intervals used at step 4 are defined as follows. For any state action pair (s,a) and for rewards, transition probabilities, and holding times we define

$$\beta_k^r(s,a)=\begin{cases}\sigma_r\sqrt{\frac{14\log(2SAi_k/\delta)}{\max\{1,N_k(s,a)\}}}, & \text{if } N_k(s,a)\geq\frac{2b_r^2}{\sigma_r^2}\log\left(\frac{240SAi_k^7}{\delta}\right)\\ 14b_r\frac{\log(2SAi_k/\delta)}{\max\{1,N_k(s,a)\}}, & \text{otherwise}\end{cases}$$

$$\beta_k^p(s,a)=\sqrt{\frac{14S\log(2Ai_k/\delta)}{\max\{1,N_k(s,a)\}}},$$

$$\beta_k^\tau(s,a)=\begin{cases}\sigma_\tau\sqrt{\frac{14\log(2SAi_k/\delta)}{\max\{1,N_k(s,a)\}}}, & \text{if } N_k(s,a)\geq\frac{2b_\tau^2}{\sigma_\tau^2}\log\left(\frac{240SAi_k^7}{\delta}\right)\\ 14b_\tau\frac{\log(2SAi_k/\delta)}{\max\{1,N_k(s,a)\}}, & \text{otherwise}\end{cases}$$

where σ_r, b_r, σ_τ, b_r are suitable constants characterizing the sub-exponential distributions of rewards and holding times. As a result, the empirical estimates $\hat{r}_k$, $\hat{\tau}_k$, and $\hat{p}_k$ are $\pm\beta_k^r(s,a),\beta_k^\tau(s,a),\beta_k^p(s,a)$ away from the true values.

Extended value iteration (EVI). We rely on a data-transformation (also called "uniformization") that turns an SMDP M into an "equivalent" MDP $M_{\rm eq}=\{\mathcal{S},\mathcal{A},p_{\rm eq},r_{\rm eq}\}$ with same state and action spaces and such that $\forall s,s'\in\mathcal{S},\forall a\in\mathcal{A}$:

$$\begin{aligned}\overline{r}_{\rm eq}(s,a)&=\frac{\overline{r}(s,a)}{\overline{\tau}(s,a)}\\ p_{\rm eq}(s'|s,a)&=\frac{\tau}{\overline{\tau}(s,a)}\big(p(s'|s,a)-\delta_{s,s'}\big)+\delta_{s,s'}\end{aligned}\tag{4}$$

where $\delta_{s,s'}=0$ if $s\neq s'$ and $\delta_{s,s'}=1$ otherwise, and τ is an arbitrary non-negative real such that $\tau<\tau_{\min}$. $M_{\rm eq}$ enjoys the following equivalence property.

Proposition 3 ([Federgruen et al., 1983], Lemma 2)**.** *If (v^*,g^*) is an optimal pair of bias and gain in M_{eq} then $(\tau^{-1}v^*,g^*)$ is a solution to Eq. 2, i.e., it is an optimal pair of bias/gain for the original SMDP M.*

As a consequence of the equivalence stated in Prop. 3, computing the optimal policy of an SMDP amounts to computing the optimal policy of the MDP obtained after data transformation (see App. A for more details). Thus, EVI is obtained by applying a value iteration scheme to an MDP $\widetilde{M}_{k,eq}^+$ equivalent to the extended SMDP $\widetilde{M}_k^+$. We denote the state values of the j-th iteration by $u_j(s)$. We also use the vector notation $u_j=(u_j(s))_{s\in\mathcal{S}}$. Similarly, we denote by $\widetilde{p}(\cdot\mid s,a)=(\widetilde{p}(s'\mid s,a))_{s'\in\mathcal{S}}$ the transition probability vector of state-action pair (s,a). The optimistic reward at episode k is fixed through the EVI iterations and it is obtained as $\widetilde{r}_{j+1}(s,a)=\min\{\hat{r}_k(s,a)+\beta_k^r(s,a);R_{\max}\tau_{\max}\}$, i.e., by taking the largest possible value compatible with the confidence intervals. At iteration j, the optimistic transition model is obtained as $\widetilde{p}_{j+1}(\cdot\mid s,a)\in\operatorname{Arg\,max}_{p(\cdot)\in\mathcal{P}_k(s,a)}\{p^\intercal u_j\}$ and $\mathcal{P}_k(s,a)$ is the set of probability distributions included in the confidence interval defined by $\beta_k^p(s,a)$. This optimization problem can be solved in $O(S)$ operations using the same algorithm as in UCRL. Finally, the optimistic holding time depends on u_j and the optimistic transition model $\widetilde{p}_{j+1}$ as

$$\widetilde{\tau}_{j+1}(s,a)=\min\Big\{\tau_{\max};\max\big\{\tau_{\min};\hat{\tau}_k(s,a)$$
$$-\operatorname{sgn}\big[\widetilde{r}_{j+1}(s,a)+\tau\big(\widetilde{p}_{j+1}(\cdot|s,a)^\intercal u_j-u_j(s)\big)\big]\beta_k^\tau(s,a)\big\}\Big\},$$

The min and max insure that $\widetilde{\tau}_{j+1}$ ranges between $\tau_{\min}$ and $\tau_{\max}$. When the term $\widetilde{r}_{j+1}(s,a)+\big(\widetilde{p}_{j+1}(\cdot|s,a)^\intercal u_j-u_j(s)\big)$ is positive (resp. negative), $\widetilde{\tau}_{j+1}(s,a)$ is set to the minimum (resp. largest) possible value compatible with its confidence intervals so as to maximize the right-hand side of Eq. 5 below. As a result, for any $\tau\in]0,\tau_{\min}[$, EVI is applied to an MDP equivalent to the extended SMDP $\widetilde{M}_k^+$ generated over iterations as

$$u_{j+1}(s)=\max_{a\in\mathcal{A}}\left\{\frac{\widetilde{r}_{j+1}(s,a)}{\widetilde{\tau}_{j+1}(s,a)}+\frac{\tau}{\widetilde{\tau}_{j+1}(s,a)}\Big(\widetilde{p}_{j+1}(\cdot\mid s,a)^\intercal u_j-u_j(s)\Big)\right\}+u_j(s)\tag{5}$$

with arbitrary u_0. Finally, the stopping condition is

$$\max_{s\in\mathcal{S}}\{u_{i+1}(s)-u_i(s)\}-\min_{s\in\mathcal{S}}\{u_{i+1}(s)-u_i(s)\}<\epsilon.\tag{6}$$

We prove the following.

Lemma 1. *If the stopping condition holds at iteration i of EVI, then the greedy policy w.r.t. u_i is ϵ-optimal w.r.t. extended SMDP $\widetilde{M}_k^+$. The stopping condition is always reached in a finite number of steps.*

As a result, we can conclude that running EVI at each episode k with an accuracy parameter $\epsilon=R_{\max}/\sqrt{i_k}$ guarantees that $\widetilde{\pi}_k$ is $R_{\max}/\sqrt{i_k}$-optimal w.r.t. $\max_{M'\in\mathcal{M}_k}\rho^*(M')$.

4 Regret Analysis

In this section we report upper and lower bounds on the regret of UCRL-SMDP. We first extend the notion of diameter to the case of SMDP as follows.

Definition 2. *For any SMDP M, we define the diameter $D(M)$ by*

$$D(M)=\max_{s,s'\in\mathcal{S}}\left\{\min_\pi\left\{\mathbb{E}^\pi\big[T(s')|s_0=s\big]\right\}\right\}\tag{7}$$

where $T(s')$ is the first time in which s' is encountered, i.e., $T(s')=\inf\big\{\sum_{i=1}^n\tau_i:n\in\mathbb{N},\ s_n=s'\big\}$.

Note that the diameter of an SMDP corresponds to an average *actual* duration and not an average number of

decision steps. However, if the SMDP is an MDP the two definitions of diameter coincides. Before reporting the main theoretical results about UCRL-SMDP, we introduce a set of technical assumptions.

Assumption 1. *For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we assume that $\tau_{\max} \geq \overline{\tau}(s,a) \geq \tau_{\min} > 0$ and $\max_{s\in\mathcal{S},a\in\mathcal{A}} \left\{\frac{\overline{\tau}(s,a)}{\overline{\tau}(s,a)}\right\} \leq R_{\max}$ with $\tau_{\min}$, $\tau_{\max}$, and $R_{\max}$ known to the learning algorithm. Furthermore, we assume that the random variables $(r(s,a,s'))_{s,a,s'}$ and $(\tau(s,a,s'))_{s,a,s'}$ are either **1)** sub-Exponential with constants (σ_r, b_r) and (σ_τ, b_τ), or **2)** bounded in $[0, R_{\max}T_{\max}]$ and $[T_{\min}, T_{\max}]$, with $T_{\min} > 0$. We also assume that the constants characterizing the distributions are known to the learning agent.*

We are now ready to introduce our main result.

Theorem 1. *With probability of at least $1-\delta$, it holds that for any initial state $s \in \mathcal{S}$ and any $n > 1$, the regret of* UCRL-SMDP *$\Delta(M, \mathfrak{A}, s, n)$ is bounded by:*

$$O\left(\left(D\sqrt{S} + \mathcal{C}(M,n,\delta)\right) R_{\max}\sqrt{SAn\log\left(\frac{n}{\delta}\right)}\right), \quad (8)$$

where $\mathcal{C}(M,n,\delta)$ depends on which case of Asm. 1 is considered [4]

***1)** sub-Exponential*

$$\mathcal{C}(M,n,\delta) = \tau_{\max} + \left(\frac{\sigma_r \vee b_r}{R_{\max}} + \sigma_\tau \vee b_\tau\right)\sqrt{\log\left(\frac{n}{\delta}\right)},$$

***2)** bounded*

$$\mathcal{C}(M,n,\delta) = T_{\max} + (T_{\max} - T_{\min}).$$

Proof. The proof (App. B) follows similar steps as in [Jaksch et al., 2010]. Apart from adapting the concentration inequalities to sub-exponential r.v. and deriving the guarantees about EVI applied to the equivalent MDP M_{eq} (Lem. 1), one of the key aspects of the proof is to show that the learning complexity is actually determined by the diameter $D(M)$ in Eq. 2. As for the analysis of EVI, we rely on the data-transformation and we show that the span of u_j (Eq. 5) can be bounded by the diameter of M_{eq}, which is related to the diameter of the original SMDP as $D(M_{\text{eq}}) = D(M)/\tau$ (Lem. 6 in App. B).

The bound. The upper bound is a direct generalization of the result derived by Jaksch et al. [2010] for UCRL in MDPs. In fact, whenever the SMDP reduces to an MDP (i.e., each action takes exactly one step to execute), then $n = T$ and the regret, the diameter, and the bounds are the same as for UCRL. If we consider $R_{\max} = 1$ and bounded holding times, the regret scales as $\widetilde{O}(DS\sqrt{An} + T_{\max}\sqrt{SAn})$. The most interesting aspect of this bound is that the extra cost of having actions with random duration is only partially additive rather than multiplicative (as it happens e.g., with the per-step reward $R_{\max}$). This shows that errors in estimating the holding times do not get amplified by the diameter D and number of states S as much as it happens for errors in reward and dynamics. This is confirmed in the following lower bound.

Theorem 2. *For any algorithm $\mathfrak{A}$, any integers $S, A \geq 10$, any reals $T_{\max} \geq 3T_{\min} \geq 3$, $R_{\max} > 0$, $D > \max\{20T_{\min}\log_A(S), 12T_{\min}\}$, and for $n \geq \max\{D, T_{\max}\}SA$, there is an SMDP M with at most S states, A actions, and diameter D, with holding times in $[T_{\min}, T_{\max}]$ and rewards in $\left[0, \frac{1}{2}R_{\max}T_{\max}\right]$ satisfying $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}_s$, $\overline{r}(s,a) \leq R_{\max}\overline{\tau}(s,a)$, such that for any initial state $s \in \mathcal{S}$ the expected regret of $\mathfrak{A}$ after n decision steps is lower-bounded by:*

$$\mathbb{E}\left[\Delta(M, \mathfrak{A}, s, n)\right] = \Omega\left(\left(\sqrt{D} + \sqrt{T_{\max}}\right)R_{\max}\sqrt{SAn}\right)$$

Proof. Similar to the upper bound, the proof (App. C) is based on [Jaksch et al., 2010] but it requires to perturb transition probabilities and rewards at the same time to create a family of SMDPs with different optimal policies that are difficult to discriminate. The contributions of the two perturbations can be made independent. More precisely, the lower bound is obtained by designing SMDPs where learning to distinguish between "good" and "bad" transition probabilities and learning to distinguish between "good" and "bad" rewards are two independent problems, leading to two additive terms $\sqrt{D}$ and $\sqrt{T_{\max}}$ in the lower bound.

The bound. Similar to UCRL, this lower bound reveals a gap of $\sqrt{DS}$ on the first term and $\sqrt{T_{\max}}$. While closing this gap remains a challenging open question, it is a problem beyond the scope of this paper.

In the next section, we discuss how these results can be used to bound the regret of options in MDPs and what are the conditions that make the regret smaller than using UCRL on primitive actions.

5 Regret in MDPs with Options

Let M be an MDP and $\mathcal{O}$ a set of options and let $M_{\mathcal{O}}$ be the corresponding SMDP obtained from Prop. 1. We index time steps (i.e., time at primitive action level) by t and decision steps (i.e., time at option level) by i. We denote by $N(t)$ the total number of decision steps that occurred before time t. Given n decision steps, we denote by $T_n = \sum_{i=1}^{n} \tau_i$ the number of time steps elapsed after the execution of the n first options so that $N(T_n) = n$. Any SMDP-learning algorithm $\mathfrak{A}_{\mathcal{O}}$ applied to $M_{\mathcal{O}}$ can be interpreted as a learning algorithm $\mathfrak{A}$ on M so that at each time step t, $\mathfrak{A}$ selects an action of M based on the policy associated to the

[4] We denote $\max\{a,b\} = a \vee b$.

option started at decision step $N(t)$. We can thus compare the performance of UCRL and UCRL-SMDP when learning in M. We first need to relate the notion of average reward and regret used in the analysis of UCRL-SMDP to the original counterparts in MDPs.

Lemma 2. *Let M be an MDP, $\mathcal{O}$ a set of options and $M_\mathcal{O}$ the corresponding SMDP. Let $\pi_\mathcal{O}$ be any stationary policy on $M_\mathcal{O}$ and π the equivalent policy on M (not necessarily stationary). For any state $s \in \mathcal{S}_\mathcal{O}$, any learning algorithm $\mathfrak{A}$, and any number of decision steps n we have $\rho^{\pi_\mathcal{O}}(M_\mathcal{O}, s) = \rho^{\pi}(M, s)$ and*

$$\Delta(M, \mathfrak{A}, T_n) = \Delta(M_\mathcal{O}, \mathfrak{A}, n) + T_n\left(\rho^*(M) - \rho^*(M_\mathcal{O})\right).$$

The linear regret term is due to the fact that the introduction of options amounts to constraining the space of policies that can be expressed in M. As a result, in general we have $\rho^*(M) \geq \rho^*(M_\mathcal{O}) = \max_{\pi_\mathcal{O}} \rho^{\pi_\mathcal{O}}(M_\mathcal{O})$, where $\pi_\mathcal{O}$ is a stationary deterministic policy on $M_\mathcal{O}$. Thm. 2 also guarantees that the optimal policy computed in the SMDP $M_\mathcal{O}$ (i.e., the policy maximizing $\rho^{\pi_\mathcal{O}}(M_\mathcal{O}, s)$) is indeed the best in the subset of policies that can be expressed in M by using the set of options $\mathcal{O}$. In order to use the regret analysis of Thm. 1, we still need to show that Asm. 1 is verified.

Lemma 3. *An MDP provided with a set of options is an SMDP where the holding times and rewards $\tau(s, o, s')$ and $r(s, o, s')$ are distributed as sub-exponential random variables. Moreover, the holding time of an option is sub-Gaussian if and only if it is almost surely bounded.*

This result is based on the fact that once an option is executed, we obtain a Markov chain with absorbing states corresponding to the states with non-zero termination probability $\beta_o(s)$ and the holding time is the number of visited states before reaching a terminal state. While in general this corresponds to a sub-exponential distribution, whenever the option has a zero probability to reach the same state twice before terminating (i.e., there is no cycle), then the holding times become bounded. Finally, we notice that no intermediate case between sub-exponential and bounded distributions is admissible (e.g., sub-Gaussian). Since these are the two cases considered in Thm. 1, we can directly apply it and obtain the following corollary.

Corollary 1. *For any MDP $M = \{\mathcal{S}, \mathcal{A}, p, r\}$ with $r(s, a, s') \in [0, R_{\max}]$ and a set of options $\mathcal{O}$, consider the resulting SMDP $M_\mathcal{O} = \{\mathcal{S}_\mathcal{O}, \mathcal{A}_\mathcal{O}, p_\mathcal{O}, r_\mathcal{O}, \tau_\mathcal{O}\}$. Then with probability of at least $1 - \delta$, it holds that for any initial state $s \in \mathcal{S}$ and any $n > 1$, the regret of* UCRL-SMDP *in the original MDP is bounded as*

$$O\Big(\big(D_\mathcal{O}\sqrt{S_\mathcal{O}} + \mathcal{C}(M_\mathcal{O}, n, \delta)\big) R^\mathcal{O}_{\max} \sqrt{S_\mathcal{O} O n \log\left(\frac{n}{\delta}\right)}\Big) + T_n\left(\rho^*(M) - \rho^*(M_\mathcal{O})\right),$$

where O is the number of options.

We can also show that the lower bound holds for MDPs with options as well. More precisely, it is possible to create an MDP and a set of options such that the lower bound is slightly smaller than that of Thm. 2.

Corollary 2. *Under the same assumptions as in Theorem 2, there exists an MDP with options such that the regret of any algorithm is lower-bounded as*

$$\Omega\left(\left(\sqrt{D_\mathcal{O}} + \sqrt{T_{\max} - T_{\min}}\ \right) R^\mathcal{O}_{\max} \sqrt{S_\mathcal{O} O n}\right) + T_n\left(\rho^*(M) - \rho^*(M_\mathcal{O})\right).$$

This shows that MDPs with options are slightly easier to learn than SMDPs. This is due to the fact that in SMDPs resulting from MDPs with options rewards and holding times are strictly correlated (i.e., $r(s, o, s') \leq R_{\max}\tau(s, o, s')$ a.s. and not just in expectation for all (s, o, s')).

We are now ready to proceed with the comparison of the bounds on the regret of learning with options and primitive actions. We recall that for UCRL $\Delta(M, \text{UCRL}, s, T_n) = \widetilde{O}(DSR_{\max}\sqrt{AT_n})$. We first notice that[5] $R^\mathcal{O}_{\max} \leq R_{\max}$ and since $\mathcal{S}_\mathcal{O} \subseteq \mathcal{S}$ we have that $S_\mathcal{O} \leq S$. Furthermore, we introduce the following simplifying conditions: **1)** $\rho^*(M) = \rho^*(M_\mathcal{O})$ (i.e., the options do not prevent from learning the optimal policy), **2)** $O \leq A$ (i.e., the number of options is not larger than the number of primitive actions), **3)** options have bounded holding time (case 2 in Asm. 1). While in general comparing upper bounds is potentially loose, we notice that both upper-bounds are derived using similar techniques and thus they would be "similarly" loose and they both have almost matching worst-case lower bounds. Let $\mathcal{R}(M, n, \delta)$ be the ratio between the regret upper bounds of UCRL-SMDP using options $\mathcal{O}$ and UCRL, then we have (up to numerical constants)

$$\mathcal{R}(M, n) \leq \frac{\left(D_\mathcal{O}\sqrt{S_\mathcal{O}} + T_{\max}\right)\sqrt{S_\mathcal{O} O n \log(n/\delta)}}{DS\sqrt{AT_n \log(T_n/\delta)}} \leq \frac{D_\mathcal{O}\sqrt{S} + T_{\max}}{D\sqrt{S}}\sqrt{\frac{n}{T_n}},$$

where we used $n \leq T_n$ to simplify the logarithmic terms. Since $\liminf_{n\to+\infty} \frac{T_n}{n} \geq \tau_{\min}$, then the previous expression gives an (asymptotic) sufficient condition for reducing the regret when using options, that is

$$\frac{D_\mathcal{O}\sqrt{S} + T_{\max}}{D\sqrt{S\tau_{\min}}} \leq 1. \tag{9}$$

In order to have a better grasp on the cases covered by this condition, let $D_\mathcal{O} = \alpha D$, with $\alpha \geq 1$. This corresponds to the case when navigating through some

[5] The largest per-step reward in the SMDP is defined as $R^\mathcal{O}_{\max} \geq \max_{s\in\mathcal{S}, a\in\mathcal{A}} \left\{\frac{\overline{r}(s,a)}{\overline{\tau}(s,a)}\right\}$.

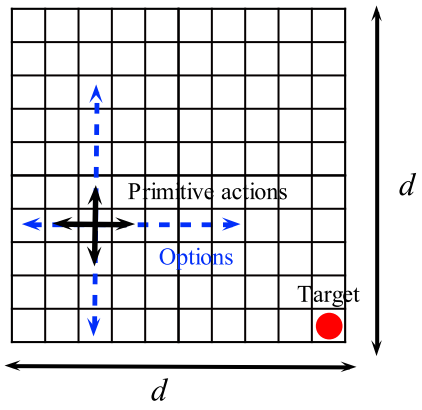


Figure 2: Navigation problem.

states becomes more difficult with options than with primitive actions, thus causing an increase in the diameter. If options are such that $T_{\max} \leq D\sqrt{S}$ and $\tau_{\min} > (1+\alpha)^2$, then it is easy to see that the condition in Eq. 9 is satisfied. This shows that even when the introduction of options partially disrupt the structure of the original MDP (i.e., $D_{\mathcal{O}} \geq D$), it is enough to choose options which are long enough (but not too much) to guarantee an improvement in the regret. Notice that while conditions **1)** and **2)** are indeed in favor of UCRL-SMDP, $S_{\mathcal{O}}$, O, and $T_{\max}$ are in general much smaller than S, A, $D\sqrt{S}$ (S and D are large in most of interesting applications). Furthermore, $\tau_{\min}$ is a very loose upper-bound on $\liminf_{n\to+\infty} \frac{T_n}{n}$ and in practice the ratio $\frac{T_n}{n}$ can take much larger values if $\tau_{\max}$ is large and many options have a high expected holding time. As a result, the set of MDPs and options on which the regret comparison is in favor of UCRL-SMDP is much wider than the one defined in Eq. 9. Nonetheless, as illustrated in Lem. 3, the case of options with bounded holding times is quite restrictive since it requires the absence of self-loops during the execution of an option. If we reproduce the same comparison in the general case of sub-exponential holding times, then the ratio between the regret upper bounds becomes

$$\mathcal{R}(M,n) \leq \frac{D_{\mathcal{O}}\sqrt{S} + \mathcal{C}(M,n,\delta)}{D\sqrt{S}} \sqrt{\frac{n}{T_n}},$$

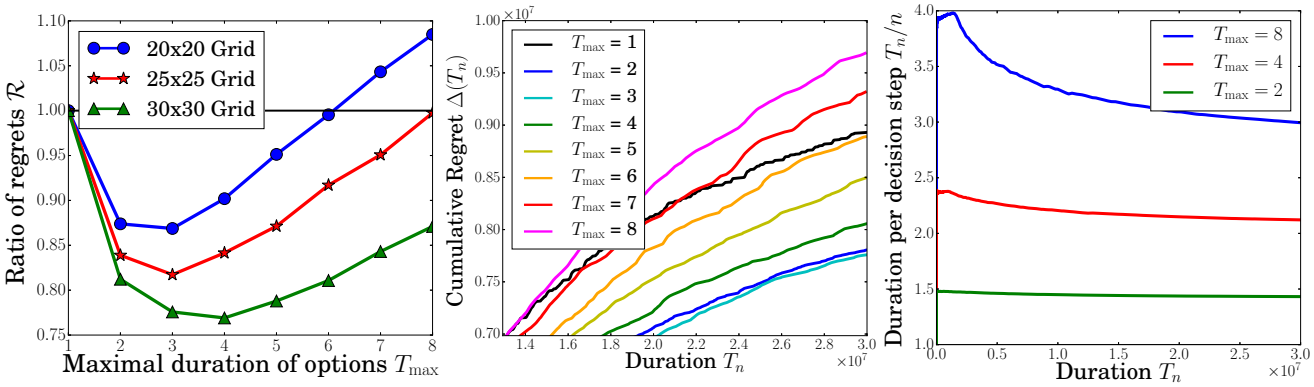
where $\mathcal{C}(M,n,\delta) = O(\sqrt{\log(n/\delta)})$. As a result, as n increases, the ratio is always greater than 1, thus showing that in this case the regret of UCRL-SMDP is asymptotically worse than UCRL. Whether this is an artefact of the proof or it is an intrinsic weakness of options, it remains an open question.

6 Illustrative Experiment

We consider the navigation problem in Fig. 2. In any of the d^2 states of the grid except the target, the four cardinal actions are available, each of them being successful with probability 1. If the agent hits a wall then it stays in its current position with probability 1. When the target state is reached, the state is reset to any other state with uniform probability. The reward of any transition is 0 except when the agent leaves the target in which case it equals $R_{\max}$. The optimal policy simply takes the shortest path from any state to the target state. The diameter of the MDP is the longest shortest path in the grid, that is $D = 2d-2$. Let m be any non-negative integer smaller than d and in every state but the target we define four macro-actions: *LEFT*, *RIGHT*, *UP* and *DOWN* (blue arrows in the figure). When *LEFT* is taken, primitive action *left* is applied up to m times (similar for the other three options). For any state s' which is $k \leq m$ steps on the left of the starting state s, we set $\beta_o(s') = 1/(m-k+1)$ so that the probability of the option to be interrupted after any $k \leq m$ steps is $1/m$. If the starting state s is l steps close to the left border with $l < m$ then we set $\beta_o(s') = 1/(l-k+1)$ for any state s' which is $k \leq l$ steps on the left. As a result, for all options started m steps far from any wall, $T_{\max} = m$ and the expected duration is $\tau := \tau(s,o) = (m+1)/2$, which reduces to $T_{\max} = l$ and $\tau = (l+1)/2$ for an option started $l < m$ step from the wall and moving towards it. More precisely, all options have an expected duration of $\tau(s,o) = \tau$ in all but in md states, which is small compared to the total number of d^2 states. The SMDP formed with this set of options preserves the number of state-action pairs $S_{\mathcal{O}} = S = d^2$ and $A' = A = 4$ and the optimal average reward $\rho^*(M) = \rho^*(M')$, while it slightly perturbs the diameter $D_{\mathcal{O}} \leq D + m(m+1)$ (see App. F for further details). Thus, the two problems seem to be as hard to learn. However the (asymptotic) ratio between the regret upper bounds becomes

$$\begin{aligned}\lim_{n\to\infty} \mathcal{R}(M,n) &\leq \frac{(2d-2+m^2+m)d+m}{(2d-2)d} \left(\lim_{n\to\infty} \sqrt{\frac{n}{T_n}}\right) \\ &\leq \left(1+2\frac{m^2}{d}\right)\left(\lim_{n\to\infty} \sqrt{\frac{n}{T_n}}\right),\end{aligned}$$

where we assume $m, d \geq 2$. While a rigorous analysis of the ratio between the number of option decision steps n and number of primitive actions T_n is difficult, we notice that as d increases w.r.t. m, the chance of executing options close to a wall decreases, since for any option only md out of d^2 states will lead to a duration smaller than τ and thus we can conclude that n/T_n tends to $1/\tau = 2/(m+1)$ as n and d grow. As a result, the ratio would reduce to $(1+2m^2/d)\sqrt{2/(m+1)}$ that is smaller than 1 for a wide range of values for m and d. Finally, the ratio is (asymptotically in d) minimized by $m \approx \sqrt{d}$, which gives $\mathcal{R}(M,n) = O(d^{-1/4})$, thus showing that as d increases there is always an appropriate choice of m for which learning with options becomes significantly better than learning with primitive actions. In Fig. 3a we empirically validate



(a) (b) (c)

Figure 3: *(a)* Ratio of the regrets with and without options for different values of $T_{\max}$; *(b)* Regret as a function of T_n for a 20x20 grid; *(c)* Evolution of T_n/n for a 20x20 grid.

this finding by studying the ratio between the actual regrets (and not their upper-bounds) as d and m (i.e., $T_{\max}$) vary, and with a fixed value of T_n that is chosen big enough for every d. As expected, for a fixed value of d, the ratio $\mathcal{R}$ first decreases as m increases, reaches a minimum and starts increasing to eventually exceed 1. As d increases, the value of the minimum decreases, while the optimal choice of m increases. This behaviour matches the theory, which suggests that the optimal choice for m increases as $O(\sqrt{d})$. In Fig. 3b we report the cumulative regret and we observe that high values of $T_{\max}$ worsen the learning performances w.r.t. learning without options ($T_{\max} = 1$, plotted in black). Finally, Fig. 3c shows that, as n tends to infinity, T_n/n tends to converge to $(m+1)/2$ when $m \ll d$, whereas it converges to slightly smaller values when m is close to d because of the truncations operated by walls.

Discussion. Despite its simplicity, the most interesting aspect of this example is that the improvement on the regret is not obtained by trivially reducing the number of state-action pairs, but it is intrinsic in the way options change the dynamics of the exploration process. The two key elements in designing a successful set of options $\mathcal{O}$ is to preserve the average reward of the optimal policy and the diameter. The former is often a weaker condition than the latter. In this example, we achieved both conditions by designing a set $\mathcal{O}$ where the termination conditions allow any option to end after only one step. This preserves the diameter of the original MDP (up to an additive constant), since the agent can still navigate at the level of granularity of primitive actions. Consider a slightly different set of options $\mathcal{O}'$, where each option moves exactly by m steps (no intermediate interruption). The number of steps to the target remains unchanged from any state and thus we can achieve the optimal performance. Nonetheless, having π^* in the set of policies that can be represented with $\mathcal{O}'$ does not guarantee that the UCRL-SMDP would be as efficient in learning the optimal policy as UCRL. In fact, the expected number of steps needed to go from a state s to an adjacent state s' may significantly increase. Despite being only one primitive action apart, there may be no sequence of options that allows to reach s' from s without relying on the random restart triggered by the target state. A careful analysis of this case shows that the diameter is as large as $D_{\mathcal{O}'} = D(1+m^2)$ and there exists no value of m that satisfies Eq. 9 (see App. F).

7 Conclusions

We derived upper and lower bounds on the regret of learning in SMDPs and we showed how these results apply to learning with options in MDPs. Comparing the regret bounds of UCRL-SMDP with UCRL, we provided sufficient conditions on the set of options and the MDP (i.e., similar diameter and average reward) to reduce the regret w.r.t. learning with primitive actions. To the best of our knowledge, this is the first attempt of explaining when and how options affect the learning performance. Nonetheless, we believe that this result leaves space for improvements. In fact, Prop. 1 implies that the class of SMDPs is a strict superset of MDPs with options. This suggests that a more effective analysis could be done by leveraging the specific structure of MDPs with options rather than moving to the more general model of SMDPs. This may actually remove the additional $\sqrt{\log(n/\delta)}$ factor appearing because of sub-exponential distributions in the UCRL-SMDP regret. An interesting direction of research is to use this theoretical result to provide a more explicit and quantitative objective function for option discovery, in the line of what is done in [Brunskill and Li, 2014]. Finally, it would be interesting to extend the current analysis to more sophisticated hierarchical approaches to RL such as MAXQ [Dietterich, 2000].

References

Applied Probability Models with Optimization Applications, chapter 7: Semi Markov Decision Processes. Dover Publications, INC., New York, 1970.

A First Course in Stochastic Models, chapter 7: Semi Markov Decision Processes. Wiley, 2003.

Pierre-Luc Bacon and Doina Precup. The option-critic architecture. In *NIPS'15 Deep Reinforcement Learning Workshop*, 2015.

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

Pablo Samuel Castro and Doina Precup. 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, pages 140–152, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-29945-2. doi: 10.1007/978-3-642-29946-9_16.

Özgür Şimşek and Andrew G. Barto. 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.

Thomas G. Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303, 2000.

A. Federgruen, P.J. Schweitzer, and H.C. Tijms. Denumerable undiscounted semi-markov decision processes with unbounded rewards. *Mathematics of Operations Research*, 1983.

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

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

Arie Leizarowitz. *Decision and Control in Management Science: Essays in Honor of Alain Haurie*, chapter 5: On Optimal Policies of Multichain Finite State Compact Action Markov Decision Processes. Springer Science and Business Media, 2013.

Kfir Y. Levy and Nahum Shimkin. Unified inter and intra options learning using policy gradient methods. In Scott Sanner and Marcus Hutter, editors, *EWRL*, volume 7188 of *Lecture Notes in Computer Science*, pages 153–164. Springer, 2011. ISBN 978-3-642-29945-2.

Kfir Y. Levy and Nahum Shimkin. *Recent Advances in Reinforcement Learning: 9th European Workshop, EWRL 2011*, chapter Unified Inter and Intra Options Learning Using Policy Gradient Methods, pages 153–164. Springer Berlin Heidelberg, 2012.

Jianyong Liu and Xiaobo Zhao. On average reward semi-markov decision processes with a general multichain structure. *Math. Oper. Res.*, 29(2):339–352, 2004.

Timothy A. Mann and Shie Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

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

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

Ishai Menache, Shie Mannor, and Nahum Shimkin. Q-cut - dynamic discovery of sub-goals in reinforcement learning. In *Proceedings of the 13th European Conference on Machine Learning*, 2002.

Iryna Felko Peter Buchholz, Jan Kriege. *Input Modeling with Phase-Type Distributions and Markov Models*, chapter Phase-Type Distributions. Springer, 2014.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

Munu Sairamesh and Balaraman Ravindran. Options with exceptions. In *Proceedings of the 9th European Conference on Recent Advances in Reinforcement Learning*, EWRL'11, pages 165–176, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-29945-2. doi: 10.1007/978-3-642-29946-9_18.

M. Schäl. On the second optimality equation for semi-markov decision models. *Mathematics of Operations Research*, 1992.

Jonathan Sorg and Satinder P. Singh. Linear Options. In *AAMAS*, pages 31–38, 2010.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

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

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

Ambuj Tewari and Peter L. Bartlett. *Bounded Parameter Markov Decision Processes with Average Reward Criterion*, pages 263–277. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-72927-3. doi: 10.1007/978-3-540-72927-3_20.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387001522.

Martin Wainwright. *Course on Mathematical Statistics*, chapter 2: Basic tail and concentration bounds. University of California at Berkeley, Department of Statistics, 2015.