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).

2   Preliminaries

MDPs and options. A finite MDP is a tuple $M = \{S, A, p, r\}$ where S is a finite set of states, 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 $\pi : S \to A$ maps states to actions. A (Markov) option is a tuple $o = \{I_o, \beta_o, \pi_o\}$ where $I_o \subset S$ is the set of states where the option can be initiated, $\beta_o : S \to [0, 1]$ is the probability distribution that the option ends in a given state, and $\pi_o : S \to A$ is the policy followed until the option ends. Whenever the set of primitive actions A is replaced by a set of options 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 O, the resulting decision process is an SMDP $M_O = \{S_O, O, p_O, r_O, \tau_O\}$, where $S_O \subseteq S$ is the set of states where options can start and end,*

$$S_O = \left(\cup_{o \in O} \, I_o \right) \bigcup \left(\cup_{o \in O} \{s : \beta_o(s) > 0\} \right),$$

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

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

where $p(s_k = s'|s, \pi_o)$ is the probability of reaching state s' after exactly k steps following policy π_o, $r_O(s, o, s')$ is the distribution of the cumulative reward obtained by executing option o from state s until interruption at s', and $\tau_O(s, o, s')$ 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 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., $\cup_{o \in O}\{s : \beta_o(s) > 0\} \subseteq \cup_{o \in O} I_o$. This also implies that the resulting SMDP M_O is communicating whenever the original MDP M is communicating. Finally, we notice that a stationary deterministic policy constructed on a set of options O may result into a non-stationary policy on the set of actions 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 $M = \{S, A, p, r, \tau\}$, 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 $\overline{\tau}(s, a) = \sum_{s' \in S} \overline{\tau}(s, a, s')p(s'|s, a)$ (resp. $\overline{r}(s, a) = \sum_{s' \in S} \overline{r}(s, a, s')p(s'|s, a)$) 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 $N(t) = \sup\{n \in \mathbb{N}, \sum_{i=1}^{n} \tau_i \leq t\}$ the number of decision steps that occurred before time t. For any policy π and $s \in S$:*

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

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, $\underline{\rho}^{\pi^*}(s) \geq \overline{\rho}^\pi(s)$ and $\overline{\rho}^{\pi^*}(s) = \underline{\rho}^{\pi^*}(s) = \rho^*$.

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

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

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 $s \in S$, and any number of decision steps $n \geq 1$, 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*

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

---

[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

---

[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 $k = 1, 2, ...$ **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)

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

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

$a \in \mathcal{A}$, 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.

Confidence intervals. Unlike in MDPs, we consider

a slightly more general scenario where cumulative rewards and holding times are not bounded but are subexponential 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 $M_{\text{eq}} = \{\mathcal{S}, \mathcal{A}, p_{\text{eq}}, r_{\text{eq}}\}$ with same state and action spaces and such that $\forall s, s' \in \mathcal{S}, \forall a \in \mathcal{A}$:

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

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_{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

$\widetilde{p}_{j+1}(\cdot \mid s, a) \in \text{Arg}\max_{p(\cdot)\in\mathcal{P}_k(s,a)} \{p^\intercal u_j\}$ 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

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

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

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

with arbitrary u_0. Finally, the stopping condition is

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

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*

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

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

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 $\max_{s \in \mathcal{S}, a \in \mathcal{A}} \left\{ \frac{\overline{r}(s,a)}{\overline{\tau}(s,a)} \right\} \leq R_{\max}$ 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 $D(M_{\mathrm{eq}}) = D(M)/\tau$ (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

---

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

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

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 subexponential 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 subexponential 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*

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

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*

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

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] $R_{\max}^{\mathcal{O}} \leq R_{\max}$ 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)

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

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 $R_{\max}^{\mathcal{O}} \geq \max_{s \in \mathcal{S}, a \in \mathcal{A}} \left\{ \frac{\overline{r}(s, a)}{\overline{\tau}(s, a)} \right\}$.
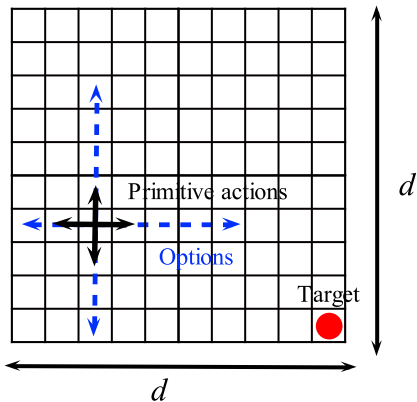
Figure 2: Navigation problem.

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

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

$$\leq \left(1 + 2\frac{m^2}{d} \right) \left(\lim_{n \to \infty} \sqrt{\frac{n}{T_n}} \right),$$

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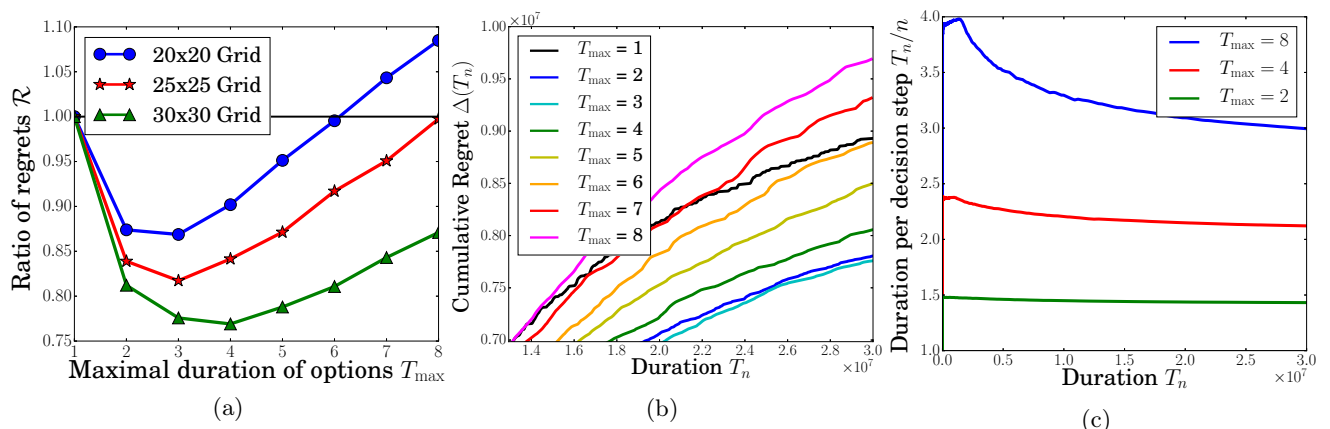
Figure 3: *(a)* Ratio of the regrets with and without options for different values of $T_{\max}$; *(b)* Regret as a function of T_n for a 20x20 grid; *(c)* Evolution of T_n/n for a 20x20 grid.

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.

A   Optimal average reward in discrete and continuous SMDPs: existence and computation

In this section we prove Proposition 2 and Lemma 1. Since extended value iteration is run on SMDP $\widetilde{M}_k^+$ with continuous actions (the choice of the transition probability), in the following we consider both the continuous and discrete case at the same time. In order to have a more rigorous treatment of SMDPs, we introduce further notations from [Puterman, 1994].

A decision rule is a function $d : \mathcal{H} \to \Delta(\mathcal{A})$ where $\mathcal{H}$ is the set of possible histories and $\Delta(\mathcal{A})$ is the set of probability distributions over $\mathcal{A}$. For an SMDP M, we will denote by D_M^{HR} the set of history-dependent randomized decision rules and D_M^{MD} the subset of Markovian deterministic decision rules ($D_M^{MD} \subset D_M^{HR}$). A history-dependent randomized policy is a sequence of elements of D_M^{HR} indexed by the decision steps, i.e., $\pi = (d_1, d_2, ...) \in (D_M^{HR})^{\mathbb{N}}$, while a stationary deterministic policy is a constant sequence of elements of D_M^{MD}: $\pi = (d, d, ...) = d^{\infty}$. The set of history-dependent randomized policies will be denoted Π_M^{HR} and the subset of stationary deterministic policies will be denoted Π_M^{SD}: $\Pi_M^{SD} \subset \Pi_M^{HR}$. We also consider the more general case where the set of available actions may depend on the state, i.e., there exists a set $\mathcal{A}_s$ for each $s \in \mathcal{S}$.

A.1   Optimality criterion

We start by defining the optimality criterion in SMDPs. Unlike for MDPs, where the average reward of a fixed policy is uniquely defined, in SMDPs there are three different definitions that are usually encountered in the literature (see Schäl 1992, Federgruen et al. 1983, and Ros 1970).[6]

Definition 3. *Denote* $N(t) = \sup \left\{ n : n \in \mathbb{N},\ \sum_{i=1}^n \tau_i \le t \right\}$ *the number of decision steps that occured before time t. For any $\pi \in \Pi_M^{HR}$ and $s \in \mathcal{S}$, we define:*

$$\overline{\rho_1}^{\pi}(s) = \limsup_{t \to +\infty} \mathbb{E}^{\pi}\left[\frac{\sum_{i=1}^{N(t)} r_i}{t} \bigg| s_0 = s \right], \quad \underline{\rho_1}^{\pi}(s) = \liminf_{t \to +\infty} \mathbb{E}^{\pi}\left[\frac{\sum_{i=1}^{N(t)} r_i}{t} \bigg| s_0 = s \right] \tag{10}$$

$$\overline{\rho_2}^{\pi}(s) = \limsup_{n \to +\infty} \frac{\mathbb{E}^{\pi}\left[\sum_{i=1}^n r_i \big| s_0 = s \right]}{\mathbb{E}^{\pi}\left[\sum_{i=1}^n \tau_i \big| s_0 = s \right]}, \quad \underline{\rho_2}^{\pi}(s) = \liminf_{n \to +\infty} \frac{\mathbb{E}^{\pi}\left[\sum_{i=1}^n r_i \big| s_0 = s \right]}{\mathbb{E}^{\pi}\left[\sum_{i=1}^n \tau_i \big| s_0 = s \right]} \tag{11}$$

and for any $d^{\infty} \in \Pi_M^{SD}$ and $s \in \mathcal{S}$ we define:

$$\rho_3^{d^{\infty}}(s) = \sum_{\alpha=1}^{\nu(d)} p^d(\alpha|s) g_d(\alpha), \quad with \quad g_d(\alpha) = \frac{\sum_{s \in R_\alpha^d} \mu_\alpha^d(s) \overline{r}(s, d(s))}{\sum_{s \in R_\alpha^d} \mu_\alpha^d(s) \overline{\tau}(s, d(s))} \tag{12}$$

where $\nu(d)$ is the number of positive recurrent classes under d^{∞}, $p^d(\alpha|s)$ is the probability of entering positive recurrent class α starting from s and following policy d^{∞}, R_α^d is the set of states of class α and μ_α^d is the stationary probability distribution of class α.

A.2   Proof of Proposition 2

We say that $(d^*)^{\infty}$ is (3)-average-optimal if for all states $s \in \mathcal{S}$ and decision rules $d \in D_M^{MD}$, $\rho_3^{(d^*)^{\infty}}(s) \ge \rho_3^{d^{\infty}}(s)$. We say that π^* is (1)-average-optimal (respectively (2)-average-optimal) if for all states $s \in \mathcal{S}$ and all $\pi \in \Pi_M^{HR}$, $\underline{\rho_1}^{\pi^*}(s) \ge \overline{\rho_1}^{\pi}(s)$ (respectively $\underline{\rho_2}^{\pi^*}(s) \ge \overline{\rho_2}^{\pi}(s)$). We prove a slightly more general formulation than Proposition 2.

Proposition 4. *If M is communicating and the expected holding times and rewards are finite, then*

- *There exists a stationary policy $\pi^* = (d^*)^{\infty}$ which is (1,2,3)-average-optimal.*

- *All optimal values are equal and constant and we will denote this value by ρ^*:*

$$\forall s \in \mathcal{S}, \ \overline{\rho_1}^{(d^*)^{\infty}}(s) = \underline{\rho_1}^{(d^*)^{\infty}}(s) = \overline{\rho_2}^{(d^*)^{\infty}}(s) = \underline{\rho_2}^{(d^*)^{\infty}}(s) = \rho_3^{(d^*)^{\infty}}(s) = \rho^*.$$

---

[6]Notice that the definition we provide in Eq. 1 of Prop. 2 is ρ_1.

Proof. **Step 1: Optimality equation of a communicating SMDP.** We first recall the average reward optimality equations for a communicating SMDP (Eq. 2)

$$\forall s \in \mathcal{S}, \ u^*(s) = \max_{a \in \mathcal{A}_s} \left\{ \overline{r}(s,a) - \rho^* \overline{\tau}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) u^*(s') \right\} \tag{13}$$

where u^* and ρ^* are the bias (up to additive constant) and average reward respectively. Since we need to analyse both the case where $\mathcal{A}_s$ is finite and the case where $\mathcal{A}_s$ is continuous, we verify that it is appropriate to consider the max instead of sup in the previous expression. For the original SMDP M, $\mathcal{A}_s$ is finite and the maximum is well-defined. For the extended SMDPs $\widetilde{M}_k^+$ considered while computing the optimistic SMDP, $\widetilde{\mathcal{A}}_{k,s}^+$ is compact and $\overline{r}(s,a)$, $\overline{\tau}(s,a)$ and $p(\cdot \mid s,a)$ are continuous in $\widetilde{\mathcal{A}}_{k,s}^+$ by the very definition of $\widetilde{M}_k^+$. The function $\overline{r}(s,a) - \rho^* \overline{\tau}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) u^*(s')$ is thus continuous on $\widetilde{\mathcal{A}}_{k,s}^+$ compact and by Weierstrass theorem, we know that the maximum is reached (i.e., there exists a maximizer). As a result, Eq. 13 is well-defined and we can study the existence and properties of its solutions.

Step 2: Data-transformation (uniformization) of an SMDP. The structure of EVI is based on a data-transformation (also called "uniformization") which turns the initial SMDP M into an "equivalent" MDP $M_{\text{eq}} = \{\mathcal{S}, \mathcal{A}, p_{\text{eq}}, r_{\text{eq}}\}$ defined as in Eq. 4. As a result, we can just apply standard MDP theory to the equivalent MDP. The average optimality equation of M_{eq} is [Puterman, 1994]

$$\forall s \in \mathcal{S}, \ v^*(s) = \max_{a \in \mathcal{A}_s} \left\{ \frac{\overline{r}(s,a)}{\overline{\tau}(s,a)} - g^* + \frac{\tau}{\overline{\tau}(s,a)} \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') + \left(1 - \frac{\tau}{\overline{\tau}(s,a)}\right) v^*(s) \right\} \tag{14}$$

Since $\tau < \tau_{\min}$, every Markov Chain induced by a stationary deterministic policy on M_{eq} is necessarily aperiodic (for any action, the probability of any state to loop on itself is strictly positive). Moreover, since M is assumed to be communicating, M_{eq} is also communicating. The same holds for $\widetilde{M}_{k,\text{eq}}^+$ (i.e., the MDP obtained from the extended SMDP $\widetilde{M}_k^+$ after data transformation). Under these conditions, Eq. 14 has a solution (v^*, g^*) where g^* is the optimal average reward of M_{eq} (respectively $\widetilde{M}_{k,\text{eq}}^+$) and the (stationary deterministic) greedy policy w.r.t. v^* is average-optimal. Moreover, standard value iteration is guaranteed to converge and it can be applied with the stopping condition in Eq. 6 to obtain an ϵ-optimal policy in finitely many steps. This holds for both finite and compact $\mathcal{A}_s$ with continuous $\overline{r}_{\text{eq}}(s,a)$ and $p_{\text{eq}}(s'|s,a)$ (see for example Puterman 1994 and Leizarowitz 2013). It is easy to show that EVI in Eq. 5 is exactly value iteration applied to $\widetilde{M}_{k,\text{eq}}^+$. Finally, Lemma 2 of [Federgruen et al., 1983] (Prop. 3) shows the "equivalence" between M and M_{eq} (respectively $\widetilde{M}_k^+$ and $\widetilde{M}_{k,\text{eq}}^+$): if (v^*, g^*) is a solution to Eq. 14, then $(\tau^{-1} v^*, g^*)$ is a solution to Eq. 13 and conversely. As a result, there exists a solution (u^*, ρ^*) to Eq. 13 for both M and $\widetilde{M}_k^+$.

Step 3: Existence of deterministic stationary optimal policy. We are now ready to prove the existence of a deterministic stationary policy that is (1,2,3)-optimal and that the corresponding optimal value is constant and equal in all three cases. We consider the case of finite and continuous $\mathcal{A}_s$ separately.

Step 3a: For M (finite $\mathcal{A}_s$). Since conditions (L), (F) and (R) of [Schäl, 1992] hold, we can apply their main theorem and obtain that

1. Any greedy policy $(d^*)^\infty$ w.r.t. u^* is such that $\overline{\rho_1}^{(d^*)^\infty}(s) \geq \overline{\rho_1}^\pi(s)$ for any $\pi \in \Pi_M^{HR}$ and any $s \in \mathcal{S}$,

2. $\forall s \in \mathcal{S}, \ \overline{\rho_1}^{(d^*)^\infty}(s) = \rho^*$,

where (u^*, ρ^*) is a solution of Eq. 13. Furthermore, from renewal theory (see e.g., Tij 2003 and [Ros, 1970]) we have that $\forall d^\infty \in \Pi_M^{SD}$, $\overline{\rho_1}^{d^\infty} = \underline{\rho_1}^{d^\infty} = \rho_1^{d^\infty}$ (i.e., the limit exists for deterministic stationary policies), thus we can conclude that $(d^*)^\infty$ is (1)-optimal. Furthermore, by Lemma 2.7 of [Schäl, 1992]: $\forall d^\infty \in \Pi_M^{SD}$, $\forall s \in \mathcal{S}, \ \overline{\rho_1}^{d^\infty}(s) = \rho_3^{d^\infty}(s)$ so d^* is also necessarily (3)-optimal. Finally, by Theorem 7.6 of [Ros, 1970], since (u^*, ρ^*) is a solution of Eq. 13:

1. Any greedy policy $(d^*)^\infty$ w.r.t. u^* is such that $\overline{\rho_2}^{(d^*)^\infty}(s) \geq \overline{\rho_2}^\pi(s)$ for any $\pi \in \Pi_M^{HR}$ and any $s \in \mathcal{S}$,

2. $\forall s \in \mathcal{S}, \overline{\rho_2}^{(d^*)^\infty}(s) = \rho^*$.

By Theorem 11.4.1 of [Puterman, 1994] we have $\forall d^\infty \in \Pi_M^{SD}$, $\overline{\rho_2} = \underline{\rho_2} = \rho_2^{d^\infty}$ and thus d^* is also (2)-optimal. This concludes the proof for the finite case, which proves the statement of Prop. 2.

Step 3b: For $\widetilde{M}_k^+$ (compact $\widetilde{\mathcal{A}}_{k,s}^+$ with continuous rewards, holding times and transition probabilities). The proof is almost the same as with discrete action spaces. The only difference is that we can't apply the Theorem of [Schäl, 1992] because conditions (R) and (C*) do not hold in general. However, we can use Propositions 5.4 and 5.5 of [Schäl, 1992] and we have the same result as in the discrete case (assumptions (L), (C), (P) and (I) hold in our case and we know that the optimality equation 13 admits a solution (u^*, ρ^*), see above). Since the state space is finite, the rest of the proof is rigorously the same (the Theorems and Lemmas still applies). This guarantees the same statement as Prop. 2 but for the optimistic SMDP $\widetilde{M}_k^+$. $\qquad\square$

A.3 Proof of Lemma 1

Proof. From the proof of Prop. 2, we already have that EVI converges towards the optimal average reward of $\widetilde{M}_{k,\text{eq}}^+$, which is also the optimal average reward of $\widetilde{M}_k^+$. We also know that the stopping criterion is met in a finite number of steps and that the greedy policy when the stopping criterion holds is ϵ-optimal in the equivalent $\widetilde{M}_{k,\text{eq}}^+$. Then, in order to prove Lemma 1, we only need to prove that this policy is also ϵ-optimal in the optimistic SMDP $\widetilde{M}_k^+$. Tij [2003] shows that for any stationary deterministic policy $d^\infty \in \Pi_{\widetilde{M}_k^+}^{SD}$, the (1)-average reward is the same in the SMDP and the MDP obtained by uniformization, that is

$$\forall s \in \mathcal{S}, \rho_1^{d^\infty}(\widetilde{M}_k^+) = \rho^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+).$$

Then it immediately follows that the policy returned by EVI is (1)-ϵ-optimal in $\widetilde{M}_k^+$ and since $\forall d^\infty \in \Pi_M^{SD}$, $\rho_1^{d^\infty} = \rho_3^{d^\infty}$, it is also (3)-$\epsilon$-optimal. Note that for any deterministic stationary policy $d \in \Pi_{\widetilde{M}_k^+}^{SD}$ defining a unichain Markov chain in $\widetilde{M}_{k,\text{eq}}^+$ (or equivalently in $\widetilde{M}_k^+$), we have: $\forall s \in \mathcal{S}, \rho_1^{d^\infty}(s) = \rho_2^{d^\infty}(s)$ and this value is constant across states (see for example chapter 11 of [Puterman, 1994], Theorem 7.5 of [Ros, 1970] or [Liu and Zhao, 2004]). However, in the general case, this equality does not hold (see Example 2.1 of [Liu and Zhao, 2004]). Nevertheless, by Theorem 3.1 of [Liu and Zhao, 2004] we have

$$\forall d^\infty \in \Pi_{\widetilde{M}_k^+}^{SD}, \ \forall s \in \mathcal{S}, \ \left|\rho_2^{d^\infty}(\widetilde{M}_k^+, s) - \rho^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+, s)\right| \leq \rho_{\max}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) - \rho_{\min}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+)$$
$$\text{where} \ \ \rho_{\max}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) = \max_{s \in \mathcal{S}} \rho^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+, s) \ \ \text{and} \ \ \rho_{\min}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) = \min_{s \in \mathcal{S}} \rho^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+, s). \tag{15}$$

If we denote by d the policy returned by EVI and ρ^* the optimal gain of $\widetilde{M}_{k,\text{eq}}^+$ and $\widetilde{M}_k^+$ we obtain

$$\forall s \in \mathcal{S}, \quad \rho^* - \rho_2^{d^\infty}(\widetilde{M}_k^+, s) = \rho^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+, s) - \rho_2^{d^\infty}(\widetilde{M}_k^+, s) + \rho^* - \rho^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+, s)$$
$$\leq \rho_{\max}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) - \rho_{\min}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) + \epsilon$$
$$= \rho_{\max}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) - \rho^* + \rho^* - \rho_{\min}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) + \epsilon \leq 2\epsilon.$$

For the first inequality we used Eq. 15 and the fact that d is ϵ-optimal in $\widetilde{M}_{k,\text{eq}}^+$. For the second inequality, we used again that d is ϵ-optimal in $\widetilde{M}_{k,\text{eq}}^+$ and we also used the fact that $\rho_{\max}^{d^\infty}(\widetilde{M}_{k,\text{eq}}^+) \leq \rho^*$. In conclusion, the policy returned by EVI is (2)-2ϵ-optimal. The remaining part of Theorem 1 is thus proved for all optimality criteria. $\qquad\square$

By Theorem 8.3.2 of [Puterman, 1994], we know that there exists an optimal policy $\widetilde{d}^*$ of MDP $\widetilde{M}_{k,\text{eq}}^+$ that yields a unichain Markov Chain (i.e., a Markov Chain with a single positive recurrent class). The Markov Chain induced by $\widetilde{d}^*$ in $\widetilde{M}_k^+$ is thus also unichain and moreover: $\rho_1^{(\widetilde{d}^*)^\infty}(\widetilde{M}_k^+) = \rho^{(\widetilde{d}^*)^\infty}(\widetilde{M}_{k,\text{eq}}^+) = \rho^*(\widetilde{M}_{k,\text{eq}}^+) = \rho^*(\widetilde{M}_k^+)$. We have seen that for any policy $d \in \Pi_{\widetilde{M}_k^+}^{SD}$ yielding a unichain Markov Chain $\rho_1^{d^\infty}(\widetilde{M}_k^+) = \rho_2^{d^\infty}(\widetilde{M}_k^+)$ and so in particular, it is true for $\widetilde{d}^*$. Therefore, there exists a policy of $\widetilde{M}_k^+$ which yields a unichain Markov Chain and which is (1)-optimal, (2)-optimal and (3)-optimal. This explains why the optimal gain is the same for criteria (1) and (2) but EVI must be run with a different accuracy to insure ϵ-accuracy (the Markov Chain induced by the policy returned by EVI is not necessarily unichain).

B   Analysis of SMDP-UCRL (proof of Theorem 1)

The proof follows the same steps as in [Jaksch et al., 2010]. Therefore, in the following we only emphasize the differences between SMDPs and MDPs and we refer to [Jaksch et al., 2010] for the parts of the proof which are similar.

B.1   Splitting into Episodes

We first recall the definition of sub-exponential random variables.

Definition 4 (Wainwright [2015]). *A random variable X with mean $\mu < +\infty$ is said to be sub-exponential, if one of the following equivalent conditions is satisfied:*

1. *(Laplace transform condition) There exists $(\sigma, b) \in \mathbb{R}^+ \times \mathbb{R}^{+*}$ such that:*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2 \lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{b}. \tag{16}$$

   In this case, we say that X is sub-exponential of parameters σ, b and we denote it by $X \in \mathrm{subExp}(\sigma, b)$.

2. *There exists $c_0 > 0$ such that $\mathbb{E}[e^{\lambda(X-\mu)}] < +\infty$ for all $|\lambda| \leq c_0$.*

In order to define the confidence intervals, we use the Bernstein concentration inequality for sub-exponential random variables.

Proposition 5 (Bernstein inequality, [Wainwright, 2015]). *Let $(X_i)_{1 \leq i \leq n}$ be a collection of independent sub-Exponential random variables s.t. $\forall i \in \{1, ..., n\}$, $X_i \in \mathrm{subExp}(\sigma_i, b_i)$ and $\mathbb{E}[X_i] = \mu_i$. We have the following concentration inequalities:*

$$\forall t \geq 0, \ \mathbb{P}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i \geq t\right) \leq \begin{cases} e^{-\frac{t^2}{2n\sigma^2}}, & \text{if } 0 \leq t \leq \frac{\sigma^2}{b} \\ e^{-\frac{t}{2b}}, & \text{if } t > \frac{\sigma^2}{b} \end{cases}$$

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \mu_i \leq t\right) \leq \begin{cases} e^{-\frac{t^2}{2n\sigma^2}}, & \text{if } 0 \leq t \leq \frac{\sigma^2}{b} \\ e^{-\frac{t}{2b}}, & \text{if } t > \frac{\sigma^2}{b} \end{cases} \tag{17}$$

where $\sigma = \sqrt{\frac{\sum_{i=1}^{n} \sigma_i^2}{n}}$ and $b = \max_{1 \leq i \leq n}\{b_i\}$.

Denoting by $N(s, a)$ the state-action counts we have

$$\sum_{i=1}^{n} r_i(s_{i-1}, a_{i-1}) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \sum_{j=1}^{N(s,a)} r_{k_j}(s, a).$$

Conditionally on knowing $(N(s, a))_{s,a}$, the previous sum is equal (in distribution) to a sum of independent random variables with mean $\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} N(s, a)\overline{r}(s, a)$ and from Prop. 5 we have

$$\mathbb{P}\left(\sum_{i=1}^{n} r_i \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} N(s, a)\overline{r}(s, a) - \sigma_r \sqrt{\frac{5}{2} n \log\left(\frac{13n}{\delta}\right)} \middle| (N(s, a))_{s,a}\right) \leq \left(\frac{\delta}{13n}\right)^{5/4} \leq \frac{\delta}{24n^{5/4}},$$

$$\text{if } n \geq \frac{5b_r^2}{2\sigma_r^2} \log\left(\frac{13n}{\delta}\right)$$

$$\mathbb{P}\left(\sum_{i=1}^{n} r_i \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} N(s, a)\overline{r}(s, a) - \frac{5}{2} b_r \log\left(\frac{13n}{\delta}\right) \middle| (N(s, a))_{s,a}\right) \leq \left(\frac{\delta}{13n}\right)^{5/4} \leq \frac{\delta}{24n^{5/4}},$$

$$\text{if } n \leq \frac{5b_r^2}{2\sigma_r^2} \log\left(\frac{13n}{\delta}\right)$$

Similarly, the total holding time satisfies

$$\mathbb{P}\left(\sum_{i=1}^{n}\tau_i \geq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}N(s,a)\overline{\tau}(s,a) + \sigma_\tau\sqrt{\frac{5}{2}n\log\left(\frac{13n}{\delta}\right)}\,\Bigg|\,(N(s,a))_{s,a}\right) \leq \left(\frac{\delta}{13n}\right)^{5/4} \leq \frac{\delta}{24n^{5/4}},$$

$$\text{if } n \geq \frac{5b_\tau^2}{2\sigma_\tau^2}\log\left(\frac{13n}{\delta}\right)$$

$$\mathbb{P}\left(\sum_{i=1}^{n}\tau_i \geq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}N(s,a)\overline{\tau}(s,a) + \frac{5}{2}b_\tau\log\left(\frac{13n}{\delta}\right)\,\Bigg|\,(N(s,a))_{s,a}\right) \leq \left(\frac{\delta}{13n}\right)^{5/4} \leq \frac{\delta}{24n^{5/4}},$$

$$\text{if } n \leq \frac{5b_\tau^2}{2\sigma_\tau^2}\log\left(\frac{13n}{\delta}\right)$$

Lemma 4. *The optimal average reward can be bounded as follows:*

$$\rho^*(M) \leq \max_{s\in\mathcal{S},a\in\mathcal{A}_s}\left\{\frac{\overline{r}(s,a)}{\overline{\tau}(s,a)}\right\} \leq R_{\max}.$$

Proof. In App. A we prove that $\rho^*(M) = \rho^*(M_{\text{eq}})$ where $\rho^*(M_{\text{eq}})$ is the optimal average reward of an MDP M_{eq} with same state and action spaces as SMDP M and with average rewards of the form $\frac{\overline{r}(s,a)}{\overline{\tau}(s,a)}$. All the rewards of M_{eq} are thus bounded by $R_{\max}$ and so $\rho^*(M_{\text{eq}})$ is necessarily bounded by $R_{\max}$ as well and thus: $\rho^*(M) \leq R_{\max}$. $\square$

We are now ready to split the regret over episodes. We define the per-episode regret as

$$\Delta_k = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s,a)\left(\overline{\tau}(s,a)\rho^* - \overline{r}(s,a)\right).$$

Setting $\gamma_r = \max\left\{\frac{5}{2}b_r, \sqrt{\frac{5}{2}}\sigma_r\right\}$ and $\gamma_\tau = \max\left\{\frac{5}{2}b_\tau, \sqrt{\frac{5}{2}}\sigma_\tau\right\}$, and using a union bound on the previous inequalities we have that with probability at least $1 - \frac{\delta}{12n^{5/4}}$

$$\Delta(M,\mathfrak{A},s,n) \leq \sum_{k=1}^{n}\Delta_k + (\gamma_r + \gamma_\tau R_{\max})\log\left(\frac{13n}{\delta}\right)\sqrt{n}$$

B.2 Dealing with Failing Confidence Regions

Lemma 5. *For any episode $k \geq 1$, the probability that the true SMDP M is not contained in the set of plausible MDPs $\mathcal{M}_k$ at step i is at most $\frac{\delta}{15i_k^6}$, that is:*

$$\forall k \geq 1, \ \mathbb{P}\left(M \notin \mathcal{M}_k\right) < \frac{\delta}{15i_k^6} \tag{18}$$

Proof. This lemma is the SMDP-analogue of Lemma 17 in [Jaksch et al., 2010] and the proof is similar. Using an ℓ_1-concentration inequality for discrete probability distributions we obtain

$$\begin{aligned}
\mathbb{P}\left(\left\|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \geq \beta_k^p(s,a)\right) &= \mathbb{P}\left(\left\|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \geq \sqrt{\frac{14S}{n}\log\left(\frac{2Ai_k}{\delta}\right)}\right) \\
&\leq \mathbb{P}\left(\left\|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \geq \sqrt{\frac{2}{n}\log\left(\frac{2^S 20SAi_k^7}{\delta}\right)}\right) \\
&\leq 2^S \exp\left(-\frac{n}{2} \times \frac{2}{n}\log\left(\frac{2^S 20SAi_k^7}{\delta}\right)\right) \\
&= \frac{\delta}{20i_k^7 SA}
\end{aligned}$$

In the inequalities above, it is implicitly assumed that the value $N_k(s,a) = n$ is fixed. To be more rigorous, we are bounding the probability of the intersection of event $\{\|\widetilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \geq \beta_k^p(s,a)\}$ with event $\{N_k(s,a) = n\}$ but we omitted the latter to simplify notations, and we will also omit it in the next inequalities. Using Bernstein inequality (Prop. 5) and noting that $240 \leq 2^7 \left(\frac{SA}{\delta}\right)^6$ for $S, A \geq 2$ and $\delta \leq 1$, we have:

- If $n \geq \frac{2b_r^2}{\sigma_r^2} \log\left(\frac{240SAi_k^7}{\delta}\right)$:

$$\mathbb{P}\Big(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \beta_k^r(s,a)\Big) = \mathbb{P}\left(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \sigma_r\sqrt{\frac{14}{n}\log\left(\frac{2SAi_k}{\delta}\right)}\right)$$

$$\leq \mathbb{P}\left(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \sigma_r\sqrt{\frac{2}{n}\log\left(\frac{240SAi_k^7}{\delta}\right)}\right)$$

$$\leq 2\exp\left(-\frac{n}{2\sigma_r^2} \times \frac{2}{n}\sigma_r^2\log\left(\frac{240SAi_k^7}{\delta}\right)\right)$$

$$= \frac{\delta}{120i_k^7SA}$$

- If $n < \frac{2b_r^2}{\sigma_r^2} \log\left(\frac{240SAi_k^7}{\delta}\right)$:

$$\mathbb{P}\Big(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \beta_k^r(s,a)\Big) = \mathbb{P}\left(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \frac{14b_r}{n}\log\left(\frac{2SAi_k}{\delta}\right)\right)$$

$$\leq \mathbb{P}\left(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \frac{2b_r}{n}\log\left(\frac{240SAi_k^7}{\delta}\right)\right)$$

$$\leq 2\exp\left(-\frac{n}{2b_r} \times \frac{2}{n}b_r\log\left(\frac{240SAi_k^7}{\delta}\right)\right)$$

$$= \frac{\delta}{120i_k^7SA}$$

Similarly for holding times we have:

$$\mathbb{P}\Big(|\overline{\tau}(s,a) - \hat{\tau}_k(s,a)| \geq \beta_k^\tau(s,a)\Big) \leq \frac{\delta}{120i_k^7SA}$$

Note that when there hasn't been any observation, the confidence intervals trivially hold with probability 1. Moreover, $N_k(s,a) < i_k$ by the stopping condition of an episode. Taking a union bound over all possible values of $N_k(s,a)$ yields:

$$\mathbb{P}\Big(|\overline{\tau}(s,a) - \hat{\tau}_k(s,a)| \geq \beta_k^\tau(s,a)\Big) \leq \frac{\delta}{120i_k^6SA}$$

$$\mathbb{P}\Big(|\overline{r}(s,a) - \hat{r}_k(s,a)| \geq \beta_k^r(s,a)\Big) \leq \frac{\delta}{120i_k^6SA}$$

$$\mathbb{P}\Big(\|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \geq \beta_k^p(s,a)\Big) \leq \frac{\delta}{20i_k^6SA}$$

Summing over all state-action pairs: $\mathbb{P}\left(M \notin \mathcal{M}_k\right) < \frac{\delta}{15i_k^6}$. $\qquad\square$

We now consider the regret of episodes in which the set of plausible SMDPs $\mathcal{M}_k$ does not contain the true SMDP M: $\sum_{k=1}^m \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k}$. By the stopping criterion for episode k (except for episodes where $\nu_k(s,a) = 1$ and $N_k(s,a) = 0$ for which $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s,a) = 1 \leq i_k$):

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s,a) \leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}N_k(s,a) = i_k - 1 \tag{19}$$

We can thus bound this part of the regret:

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq \sum_{k=1}^{m} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \nu_k(s,a) \overline{\tau}(s,a) \rho^* \mathbb{1}_{M \notin \mathcal{M}_k}$$

$$\leq \tau_{\max} \rho^* \sum_{k=1}^{m} i_k \mathbb{1}_{M \notin \mathcal{M}_k} = \tau_{\max} \rho^* \sum_{i=1}^{n} i \sum_{k=1}^{m} \mathbb{1}_{i=i_k, M \notin \mathcal{M}_k}$$

$$\leq \tau_{\max} \rho^* \left(\sum_{i=1}^{\lfloor n^{1/4} \rfloor} i + \sum_{i=\lfloor n^{1/4} \rfloor+1}^{n} i \sum_{k=1}^{m} \mathbb{1}_{i=i_k, M \notin \mathcal{M}_k} \right)$$

$$\leq \tau_{\max} \rho^* \left(\sqrt{n} + \sum_{i=\lfloor n^{1/4} \rfloor+1}^{n} i \sum_{k=1}^{m} \mathbb{1}_{i=i_k, M \notin \mathcal{M}_k} \right)$$

where we defined: $\tau_{\max} = \max_{s,a} \overline{\tau}(s,a) < +\infty$.
By Lemma 5, the probability that the second term in the right hand side of the above inequality is non-zero is bounded by

$$\sum_{i=\lfloor n^{1/4} \rfloor}^{n} \frac{\delta}{15 i^6} \leq \frac{\delta}{15 n^{6/4}} + \int_{n^{1/4}}^{+\infty} \frac{\delta}{15 x^6} \, \mathrm{d}x \leq \frac{\delta}{12 n^{5/4}}.$$

In other words, with probability at least $1 - \frac{\delta}{12 n^{5/4}}$:

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq \tau_{\max} R_{\max} \sqrt{n}.$$

B.3 Episodes with $M \in \mathcal{M}_k$

Now we assume that $M \in \mathcal{M}_k$ and we start by analysing the regret of a single episode k. By construction, $R_{\max} \geq \widetilde{\rho}_k \geq \rho^* - \frac{R_{\max}}{\sqrt{i_k}}$ hence:

$$\Delta_k = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \nu_k(s,a) \left(\overline{\tau}(s,a) \rho^* - \overline{r}(s,a) \right) \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \nu_k(s,a) \left(\overline{\tau}(s,a) \widetilde{\rho}_k - \overline{r}(s,a) \right) + R_{\max} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \frac{\nu_k(s,a)}{\sqrt{i_k}} \overline{\tau}(s,a)$$

$$\implies \Delta_k \leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \nu_k(s,a) \left(\widetilde{\tau}_k(s,a) \widetilde{\rho}_k - \overline{r}(s,a) \right) + R_{\max} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \nu_k(s,a) \left(\overline{\tau}(s,a) - \widetilde{\tau}_k(s,a) \right)$$

$$+ R_{\max} \tau_{\max} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \frac{\nu_k(s,a)}{\sqrt{i_k}}$$

We now need two results about the extended value iteration algorithm.

Lemma 6. *At any iteration $i \geq 0$ of EVI (extended value iteration), the range of the state values is bounded as follows,*

$$\forall i \geq 0, \quad \max_{s \in \mathcal{S}} u_i(s) - \min_{s \in \mathcal{S}} u_i(s) \leq \frac{R_{\max} D(M)}{\tau}, \tag{20}$$

where $R_{\max}$ is an upper-bound on the per-step reward $\overline{r}(s,a)/\overline{\tau}(s,a)$, τ is the parameter used in the uniformization of the SMDP M and $D(M)$ is its diameter (Def. 2).

Proof. In Appendix A we show that EVI is value iteration applied to the equivalent MDP $\widetilde{M}_{k,\mathrm{eq}}^+$ obtained by "uniformizing" the extended SMDP $\widetilde{M}_k^+$. Thus, we focus on any SMDP M and its equivalent MDP M_{eq}. Using the same argument as in section 4.3.1 of [Jaksch et al., 2010], we have that: $\forall i \geq 0, \max_{s \in \mathcal{S}} u_i(s) - \min_{s \in \mathcal{S}} u_i(s) \leq$

$R_{\max}D(M_{\text{eq}})$ since all rewards of M_{eq} are bounded by $R_{\max}$ whenever the average reward in M is bounded by $R_{\max}$. Thus we need to find a relationship between $D(M)$ and $D(M_{\text{eq}})$. Let $T(s')$ denote the first time at which state s' is reached in M or M_{eq}, that is

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

$$\text{In MDP } M_{\text{eq}}: \ T(s') = \inf\left\{n : n \in \mathbb{N}, \ s_n = s'\right\}.$$

We prove that $\forall s, s' \in \mathcal{S}$, $\forall \pi \in \Pi_M^{SD} = \Pi_{M'}^{SD}$, $\mathbb{E}_M^{\pi}[T(s')|s_0 = s] = \tau\mathbb{E}_{M'}^{\pi}[T(s')|s_0 = s]$. We consider two cases:

1. If $\mathbb{P}_M^{\pi}(T(s') = +\infty|s_0 = s) > 0$ then necessarily $\mathbb{E}_M^{\pi}[T(s')|s_0 = s] = +\infty$.
   Moreover: $\mathbb{P}_M^{\pi}(T(s') = +\infty|s_0 = s) > 0 \implies \mathbb{P}_{M_{\text{eq}}}^{\pi}(T(s') = +\infty|s_0 = s) > 0$ and so $\mathbb{E}_{M_{\text{eq}}}^{\pi}[T(s')|s_0 = s] = +\infty = \frac{1}{\tau}\mathbb{E}_M^{\pi}[T(s')|s_0 = s]$.

2. Conversely: $\mathbb{P}_M^{\pi}(T(s') = +\infty|s_0 = s) = 0 \implies \mathbb{P}_{M_{\text{eq}}}^{\pi}(T(s') = +\infty|s_0 = s) = 0$ in which case both expectations are finite. To prove they are equal up to factor τ, we see the holding time as a "reward" (the true rewards are ignored here). Note that any policy π induces Markov chains with different dynamics on M and M_{eq} (different transition probabilities). We call these Markov chains MC and MC_{eq} respectively. Suppose we modify MC as follows: all states that are not reachable from s are ignored, all other states are unchanged except s' that is assumed to be absorbing (i.e., $\pi(s')$ is an action that loops on s' with probability 1). Furthermore, we build a Markov reward process MR with the same dynamics as MC and such that all transitions $(\overline{s}, \pi(\overline{s}))$ have an expected reward equal to $\overline{\tau}(\overline{s}, \pi(\overline{s}))$ except $(s', \pi(s'))$ which has a reward of zero. The total expected reward of this Markov reward process (MRP denoted MR) starting from s trivially equals $\mathbb{E}_M^{\pi}[T(s')|s_0 = s]$. Since we assumed that $\mathbb{E}_M^{\pi}[T(s')|s_0 = s]$ is finite, and because all states of MR are reachable from s (the other states were ignored), s' is reached with probability 1 no matter which starting state $\overline{s}$ of MR is chosen (or in other words, even though we ignored some states, the transition matrix of MR is stochastic −and not sub-stochastic− and has a single recurrent class consisting of the absorbing state s'). By [Puterman, 1994], the vector $(\overline{T}(\overline{s}))_{\overline{s}\in\mathcal{S}} = (\mathbb{E}_M^{\pi}[T(s')|s_0 = \overline{s}])_{\overline{s}\in\mathcal{S}}$ is the unique solution to the system of equations

$$\forall \overline{s}, \ \overline{T}(\overline{s}) = \overline{\tau}(\overline{s}, d(\overline{s})) + \sum_{\widetilde{s}} p(\widetilde{s}|\overline{s}, d(\overline{s}))\overline{T}(\widetilde{s}).$$

Applying the same transformation to MC_{eq} and assigning a reward of 1 to all transitions but $(s', \pi(s'))$ (which has reward 0) in order to build MR_{eq}, we deduce that the vector $(\overline{T}_{\text{eq}}(\overline{s}))_{\overline{s}\in\mathcal{S}} = (\mathbb{E}_{M_{\text{eq}}}^{\pi}[T(s')|s_0 = \overline{s}])_{\overline{s}\in\mathcal{S}}$ is the unique solution to the system of equations

$$\forall \overline{s}, \ \overline{T}_{\text{eq}}(\overline{s}) = 1 + \frac{\tau}{\overline{\tau}(\overline{s}, d(\overline{s}))}\sum_{\widetilde{s}} p(\widetilde{s}|\overline{s}, d(\overline{s}))\overline{T}_{\text{eq}}(\widetilde{s}) + \left(1 - \frac{\tau}{\overline{\tau}(\overline{s}, d(\overline{s}))}\right)\overline{T}_{\text{eq}}(\overline{s})$$

$$\iff \forall \overline{s}, \ (\tau\overline{T}_{\text{eq}}(\overline{s})) = \overline{\tau}(\overline{s}, d(\overline{s})) + \sum_{\widetilde{s}} p(\widetilde{s}|\overline{s}, d(\overline{s}))(\tau\overline{T}_{\text{eq}}(\widetilde{s})).$$

By uniqueness of the solution: $\tau\overline{T}_{\text{eq}} = \overline{T} \implies \tau D(M_{\text{eq}}) = D(M)$.

$\square$

Lemma 7. *If the convergence criterion of EVI hold at iteration i, then:*

$$\forall s \in \mathcal{S}, \ \left|u_{i+1}(s) - u_i(s) - \widetilde{\rho}_k\right| \leq \frac{1}{\sqrt{i_k}} \tag{21}$$

Proof. We introduce the following quantities

$$M_i = \max_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\}, \quad m_i = \min_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\}, \quad \epsilon = \frac{1}{\sqrt{i_k}}.$$

Since EVI is just value iteration applied to MDP M'_k, Theorem 8.5.6 of [Puterman, 1994] hold and we have:

$$\frac{1}{2}(M_i + m_i) \geq \widetilde{\rho}_k - \frac{\epsilon}{2} \iff m_i \geq \widetilde{\rho}_k - \frac{\epsilon}{2} - \frac{1}{2}(M_i - m_i) \implies m_i \geq \widetilde{\rho}_k - \epsilon$$

$$\frac{1}{2}(M_i + m_i) - \widetilde{\rho}_k \leq \frac{\epsilon}{2} \iff M_i \leq \widetilde{\rho}_k + \frac{\epsilon}{2} + \frac{1}{2}(M_i - m_i) \implies M_i \leq \widetilde{\rho}_k + \epsilon.$$

In conclusion:

$$\forall s \in \mathcal{S}, \quad \frac{-1}{\sqrt{i_k}} \leq u_{i+1}(s) - u_i(s) - \widetilde{\rho}_k \leq \frac{1}{\sqrt{i_k}}.$$

$\square$

Based on Lemma 7, Eq. 21, and optimiality equation Eq. 14, we have:

$$\forall s \in \mathcal{S}, \quad \left| \left(\widetilde{\rho}_k - \frac{\widetilde{r}_k(s, \widetilde{\pi}_k(s))}{\widetilde{\tau}_k(s, \widetilde{\pi}_k(s))} \right) - \left(\sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) u_i(s') - u_i(s) \right) \frac{\tau}{\widetilde{\tau}_k(s, \widetilde{\pi}_k(s))} \right| \leq \frac{1}{\sqrt{i_k}} \tag{22}$$

Setting $r_k = \left(\widetilde{r}_k(s, \widetilde{\pi}_k(s)) \right)_{s \in \mathcal{S}}$ to be the column vector of rewards under policy $\widetilde{\pi}_k$, $\widetilde{P}_k = \left(\widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) \right)_{s,s' \in \mathcal{S}}$ the transition matrix and $v_k = \left(\nu_k(s, \widetilde{\pi}_k(s)) \right)_{s \in \mathcal{S}}$ the row vector of visit counts for each state and the corresponding action chosen by $\widetilde{\pi}_k$. We will use the fact that $a \neq \widetilde{\pi}_k(s) \implies \nu_k(s, a) = 0$.

$$\Delta_k \leq \sum_{s,a} \nu_k(s, a) \left(\widetilde{\tau}_k(s, a) \widetilde{\rho}_k - \overline{r}(s, a) \right) + R_{\max} \sum_{s,a} \nu_k(s, a) \left(\overline{\tau}(s, a) - \widetilde{\tau}_k(s, a) \right) + R_{\max} \tau_{\max} \sum_{s,a} \frac{\nu_k(s, a)}{\sqrt{i_k}}$$

$$= \sum_{s,a} \nu_k(s, a) \left(\widetilde{\tau}_k(s, a) \widetilde{\rho}_k - \widetilde{r}_k(s, a) \right) + \sum_{s,a} \nu_k(s, a) \left(\widetilde{r}_k(s, a) - \overline{r}(s, a) \right) + R_{\max} \tau_{\max} \sum_{s,a} \frac{\nu_k(s, a)}{\sqrt{i_k}}$$

$$+ R_{\max} \sum_{s,a} \nu_k(s, a) \left(\overline{\tau}(s, a) - \widetilde{\tau}_k(s, a) \right)$$

We will now upper-bound the four terms of the right-hand side of the above inequality. Setting $c_r = \max\{14b_r, \sqrt{14}\sigma_r\}$ and $c_\tau = \max\{14b_\tau, \sqrt{14}\sigma_\tau\}$ we have:

$$\widetilde{r}_k(s, a) - \overline{r}(s, a) \leq \left| \widetilde{r}_k(s, a) - \hat{r}_k(s, a) \right| + \left| \hat{r}_k(s, a) - \overline{r}(s, a) \right| \leq 2\beta_k^r(s, a) \leq 2c_r \frac{\log(2SAi_k/\delta)}{\sqrt{\max\{1, N_k(s, a)\}}}$$

$$\overline{\tau}(s, a) - \widetilde{\tau}_k(s, a) \leq \left| \widetilde{\tau}_k(s, a) - \hat{\tau}_k(s, a) \right| + \left| \hat{\tau}_k(s, a) - \overline{\tau}(s, a) \right| \leq 2\beta_k^\tau(s, a) \leq 2c_\tau \frac{\log(2SAi_k/\delta)}{\sqrt{\max\{1, N_k(s, a)\}}}$$

Finally, using 22 and noting that $\widetilde{\tau}_k(s, a) \leq \tau_{\max}$ (by construction) we obtain:

$$\widetilde{\tau}_k(s, a) \widetilde{\rho}_k - \widetilde{r}_k(s, a) \leq \frac{R_{\max} \tau_{\max}}{\sqrt{i_k}} + \tau \left(\sum_{s' \in \mathcal{S}} \widetilde{p}_k(s'|s, \widetilde{\pi}_k(s)) u_i(s') - u_i(s) \right), \quad \text{if } a = \widetilde{\pi}_k(s)$$

$$\implies \sum_{s,a} \nu_k(s, a) \left(\widetilde{\tau}_k(s, a) \widetilde{\rho}_k - \widetilde{r}_k(s, a) \right) \leq R_{\max} \tau_{\max} \sum_{s,a} \frac{\nu_k(s, a)}{\sqrt{i_k}} + \tau \left(v_k \left(\widetilde{P}_k - I \right) u_i \right)$$

where i is the iteration at which the stopping condition of EVI holds. Defining the column vector w_k by:

$$w_k(s) = u_i(s) - \frac{\min_{s \in \mathcal{S}} u_i(s) + \max_{s \in \mathcal{S}} u_i(s)}{2}$$

and since the rows of $\widetilde{P}_k$ sum to one, we have: $v_k\big(\widetilde{P}_k - I\big)u_i = v_k\big(\widetilde{P}_k - I\big)w_k$. Moreover, by Lemma 6: $\|w_k\|_\infty \leq \frac{R_{\max}D}{2\tau}$. Noting that $\max\{1, N_k(s,a)\} \leq i_k \leq n$ we get:

$$\Delta_k \leq \tau\Big(v_k\big(\widetilde{P}_k - I\big)w_k\Big) + 2\left(R_{\max}\tau_{\max} + (c_r + R_{\max}c_\tau)\log\left(\frac{2SAn}{\delta}\right)\right)\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}$$

Using exactly the same arguments as in Jaksch et al. [2010], it is trivial to prove that with probability at least $1 - \frac{\delta}{12n^{5/4}}$:

$$\sum_{k=1}^{m} v_k\big(\widetilde{P}_k - I\big)w_k\mathbb{1}_{M\in\mathcal{M}_k} \leq \frac{R_{\max}D}{\tau}\left[\sqrt{14S\log\left(\frac{2An}{\delta}\right)}\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} + \sqrt{\frac{5}{2}n\log\left(\frac{8n}{\delta}\right)}\right.$$
$$\left. + SA\log_2\left(\frac{8n}{SA}\right)\right]$$

Lemma 8. *Consider a sequence of positive reals $(z_k)_k$ and define: $\forall k,\ Z_k = \max\{1, \sum_{i=1}^{k} z_i\}$. Assuming that $0 \leq z_k \leq Z_{k-1}$ we have:*

$$\forall n \geq 1,\ \sum_{k=1}^{n}\frac{z_k}{\sqrt{Z_{k-1}}} \leq \big(\sqrt{2}+1\big)\sqrt{Z_n}$$

Proof. See Appendix C.3 of [Jaksch et al., 2010]. $\qquad\square$

Using Lemma 8 we get:

$$\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \leq \big(\sqrt{2}+1\big)\sum_{s,a}\sqrt{N(s,a)}$$

By Jensen's inequality we thus have:

$$\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \leq \big(\sqrt{2}+1\big)\sqrt{SAn}$$

In conclusion, when $M \in \mathcal{M}_k$, with probability at least $1 - \frac{\delta}{12n^{5/4}}$:

$$\sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\in\mathcal{M}_k} \leq R_{\max}D\sqrt{\frac{5}{2}n\log\left(\frac{8n}{\delta}\right)} + R_{\max}DSA\log_2\left(\frac{8n}{SA}\right) + \big(\sqrt{2}+1\big)\left[2R_{\max}\tau_{\max}\right.$$
$$\left. + 2(c_r + R_{\max}c_\tau)\log\left(\frac{2SAn}{\delta}\right) + R_{\max}D\sqrt{14S\log\left(\frac{2An}{\delta}\right)}\right]\sqrt{SAn}$$

B.4 Computing the final bound

Gathering all previous inequalities, we have that with probability at least $1 - \frac{3\delta}{12n^{5/4}} = 1 - \frac{\delta}{4n^{5/4}}$:

$$\Delta(M, \mathfrak{A}, s, n) \leq (\gamma_r + \gamma_\tau R_{\max})\log\left(\frac{13n}{\delta}\right)\sqrt{n} + \tau_{\max}R_{\max}\sqrt{n} + R_{\max}D\sqrt{\frac{5}{2}n\log\left(\frac{8n}{\delta}\right)}$$
$$+ \big(\sqrt{2}+1\big)\left[2\tau_{\max} + 2(c_r + R_{\max}c_\tau)\log\left(\frac{2SAn}{\delta}\right) + R_{\max}D\sqrt{14S\log\left(\frac{2An}{\delta}\right)}\right]\sqrt{SAn}$$
$$+ R_{\max}DSA\log_2\left(\frac{8n}{SA}\right)$$

In [Jaksch et al., 2010] (see Appendix C.4), it is shown that when $n > 34A \log\left(\frac{n}{\delta}\right)$:

$$DSA \log_2\left(\frac{8n}{SA}\right) < \frac{2}{34} DS\sqrt{An \log\left(\frac{n}{\delta}\right)}, \text{ and } \log\left(\frac{2An}{\delta}\right) \leq 2\log\left(\frac{n}{\delta}\right)$$

and moreover if $n > S \log\left(\frac{n}{\delta}\right)$ and $A \geq 2$ (if $A = 1$ the regret is zero):

$$\frac{n^2}{\delta^2} \geq \frac{nS \log\left(\frac{n}{\delta}\right)}{\delta} \geq \frac{nS}{\delta} \implies \frac{n^2 A^2}{\delta^2} \geq \frac{2SAn}{\delta} \implies 4\log\left(\frac{n}{\delta}\right) \geq 2\log\left(\frac{An}{\delta}\right) \geq \log\left(\frac{2SAn}{\delta}\right)$$

$$\implies \Delta(M, \mathfrak{A}, s, n) = O\left(\left(D\sqrt{S} + \tau_{\max} + \left(\frac{C_r}{R_{\max}} + C_\tau\right)\sqrt{\log\left(\frac{n}{\delta}\right)}\right) R_{\max}\sqrt{SAn \log\left(\frac{n}{\delta}\right)}\right)$$

where $C_r = \max\{b_r, \sigma_r\}$ and $C_\tau = \max\{b_\tau, \sigma_\tau\}$.

Note that if $n \leq 34A \log\left(\frac{n}{\delta}\right)$ then we trivially have:

$$\sum_{k=1}^{m} \Delta_k \leq \tau_{\max} R_{\max} n = \tau_{\max} R_{\max}(\sqrt{n})^2 \leq 34\tau_{\max} R_{\max}\sqrt{An \log\left(\frac{n}{\delta}\right)}$$

and if $n \leq S \log\left(\frac{n}{\delta}\right)$:

$$\sum_{k=1}^{m} \Delta_k \leq \tau_{\max} R_{\max} n = \tau_{\max} R_{\max}(\sqrt{n})^2 \leq \tau_{\max} R_{\max}\sqrt{Sn \log\left(\frac{n}{\delta}\right)}$$

and thus the previous bound on the whole regret still holds. Taking a union bound over all possible values of $n \geq 1$ we have that with probability at least $1 - \delta$:

$$\forall n \geq 1, \ \Delta(M, \mathfrak{A}, s, n) = O\left(\left(D\sqrt{S} + \tau_{\max} + \left(\frac{C_r}{R_{\max}} + C_\tau\right)\sqrt{\log\left(\frac{n}{\delta}\right)}\right) R_{\max}\sqrt{SAn \log\left(\frac{n}{\delta}\right)}\right).$$

The derivation for the case of bounded holding times is exactly the same with different concentration inequalities applied to estimates $\widehat{\tau}(s, a)$. Note that all the terms in the upper bound are very similar to those appearing in the derivation of the upper bound for MDPs, thus the constants in the big O are very close. This justifies the analysis of the ratio between the two upper bounds in Sect. 5.

C The Lower Bound (Theorem 2)

C.1 Lower Bound for SMDPs

We will derive the lower bound by applying the same techniques as in the proof of the lower bound for MDPs (section 6 of [Jaksch et al., 2010]). We first consider the two-state SMDP M' depicted in Fig. 4. Since by assumption $\frac{D}{12} > T_{\min}$ and $\frac{T_{\max}}{3} > T_{\min}$, let $\overline{\tau} \in \left]T_{\min}, \min\left\{\frac{D}{12}, \frac{T_{\max}}{3}\right\}\right]$. Define $p = \frac{\overline{\tau}}{T_{\max}}$ and $\delta = \frac{4\overline{\tau}}{D}$. By definition of $\overline{\tau}$ we have: $p, \delta \leq \frac{1}{3}$. There are $A' = \lfloor\frac{A-1}{2}\rfloor$ actions available in each state of M'. We assume that $\forall(s, a, s') \in \mathcal{S} \times \mathcal{A}_s \times \mathcal{S}$ and $\forall i \geq 0$, $r_i(s, a, s')$ and $\tau_i(s, a, s')$ are independent. We also assume that $\forall i \geq 0$, $\tau_i(s_{i-1}, a_i, s_i)$ and $r_i(s_{i-1}, a_i, s_i)$ are independent of the next state s_i and we write: $\tau_i(s_{i-1}, a_i)$ and $r_i(s_{i-1}, a_i)$. For each action a in $\mathcal{A}_{s_0}$, $r(s_0, a) = 0$ and $\tau(s, a) \sim \tau$ where τ is a r.v. defined in Table 1. Moreover, for all actions a but a specific action a_0^*, $p(s_1|s_0, a) = \delta$ whereas $p(s_1|s_0, a_0^*) = \delta + \epsilon$ for some $0 < \epsilon < \delta$ specified later in the proof. For all actions a in $\mathcal{A}_{s_1}$, $p(s_0|s_1, a) = \delta$ and $\tau(s_1, a) \sim \tau$. Finally, $r(s_1, a) \sim r$ for all actions a except a_1^* for which $r(s_1, a_1^*) \sim r^*$, where r and r^* are r.v. defined in Table 1 where $0 < \eta < p$ will be defined later in the proof. Note that since $\eta < \frac{\overline{\tau}}{T_{\max}}$, we have: $\overline{r} \leq \overline{\tau} R_{\max}$ and $\overline{r}^* \leq \overline{\tau} R_{\max}$ which satisfies the definition of $R_{\max}$ given in assumption 1. We denote $\mathbb{E}[\cdot|s]$ the expectation conditionally on starting in state s.

Let's define $T(s_1) = \inf\{t : s_t = s_1\}$ the first time in which s_1 is encountered. $\forall d^\infty \in \Pi_{M'}^{SD}$ such that $d(s_0) \neq a_0^*$
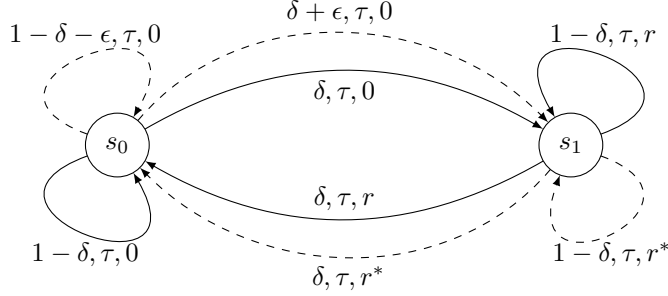
Figure 4: The two-state SMDP M' for the lower-bound on SMDPs. The two special actions a_0^* and a_1^* are shown as dashed lines.

we have:

$$
\begin{aligned}
\mathbb{E}^{d^\infty}[T(s_1)|s_0] &= \mathbb{E}^{d^\infty}\left[\sum_{n=1}^{+\infty}\left[\left(\sum_{i=1}^{n}\tau_i(s_{i-1},a_i,s_i)\right)\left(\prod_{j=0}^{n-1}\mathbb{1}_{s_j=s_0}\right)\mathbb{1}_{s_n=s_1}\right]\middle| s_0\right]\\
&= \sum_{n=1}^{+\infty}\mathbb{E}^{d^\infty}\left[\left(\sum_{i=1}^{n}\tau_i(s_0,d(s_0))\right)\left(\prod_{j=0}^{n-1}\mathbb{1}_{s_j=s_0}\right)\mathbb{1}_{s_n=s_1}\middle| s_0\right]\\
&= \sum_{n=1}^{+\infty}\mathbb{E}^{d^\infty}\left[\sum_{i=1}^{n}\tau_i(s_0,d(s_0))\middle| s_0\right]\mathbb{E}^{d^\infty}\left[\left(\prod_{j=0}^{n-1}\mathbb{1}_{s_j=s_0}\right)\mathbb{1}_{s_n=s_1}\middle| s_0\right]\\
&= \sum_{n=1}^{+\infty}n\overline{\tau}\times\delta(1-\delta)^{n-1} = \frac{\overline{\tau}}{\delta}
\end{aligned}
$$

We used the fact that τ_i is independent of the next state s_i and that $\overline{\tau} = \mathbb{E}[\tau]$. We can compute $\mathbb{E}^{d^\infty}[T(s_1)|s_0]$ with $d(s_0) = a_0^*$ and $\mathbb{E}^{d^\infty}[T(s_0)|s_1]$ (for both $d(s_0) \neq a_0^*$ and $d(s_0) = a_0^*$) similarly. The diameter of SMDP M' is thus:

$$
D' = \max\left\{\min\left\{\frac{\overline{\tau}}{\delta},\frac{\overline{\tau}}{\delta+\epsilon}\right\}, \min\left\{\frac{\overline{\tau}}{\delta},\frac{\overline{\tau}}{\delta}\right\}\right\} = \max\left\{\frac{\overline{\tau}}{\delta+\epsilon},\frac{\overline{\tau}}{\delta}\right\} = \frac{\overline{\tau}}{\delta}
$$

Any policy $d^\infty \in \Pi_{M'}^{SD}$ induces a recurrent Markov Chain on M'. Let's denote by P_d^* the limiting matrix of this Markov Chain. We know (see [Puterman, 1994]) that $P_d^* = e\mu_d^\mathsf{T}$ where $\mu_d = (1-p_d,p_d)^\mathsf{T}$ is the stationary distribution of the recurrent Markov Chain. The probability p_d can take only two different values: $p_d = \frac{1}{2}$ if and only if $d(s_0) \neq a_0^*$, and $p_d = \frac{\delta+\epsilon}{2\delta+\epsilon}$ if and only if $d(s_0) = a_0^*$. Using criterion 3 of Definition 3, the gain yielded by d^∞ has the form:

$$
\rho_d = \frac{p_d\overline{X}}{(1-p_d)\overline{\tau}+p_d\overline{\tau}} = \frac{p_d\overline{X}}{\overline{\tau}}
$$

where $X = r$ if and only if $d(s_1) \neq a_1^*$, and $X = r^*$ if and only if $d(s_1) = a_1^*$. Since $\overline{r}^* > \overline{r}$, the optimal decision rule d^* must satisfy: $d^*(s_1) = a_1^*$. Similarly, since $\frac{\delta+\epsilon}{2\delta+\epsilon} > \frac{1}{2}$ we must have: $d^*(s_0) = a_0^*$. The optimal gain is thus:

$$
\rho^* = \frac{R_{\max}}{2}\times\frac{(\delta+\epsilon)(\overline{\tau}+\eta T_{\max})}{(2\delta+\epsilon)\overline{\tau}}
$$

The actual SMDP M that we will use to prove the lower bound is built by considering $k = \lfloor\frac{S}{2}\rfloor$ copies of the two-state SMDP M', where only one of the copies has such "good" actions a_0^* and a_1^* (all the other copies have the exact same number of actions as M', but all actions are identical). $A'+1$ additional actions with deterministic transitions are introduced in every s_0-state. The reward for each of those actions is zero and the

R.v. X	$X_{\min}$	$X_{\max}$	$\mathbb{P}(X = X_{\min})$	$\mathbb{P}(X = X_{\max})$	$\mathbb{E}[X]$
τ	$T_{\min}$	$T_{\max}$	> 0	> 0	$\overline{\tau}$
r	0	$\frac{1}{2}R_{\max}T_{\max}$	$1 - p$	p	$\frac{1}{2}\overline{\tau}R_{\max}$
r^*	0	$\frac{1}{2}R_{\max}T_{\max}$	$1 - p - \eta$	$p + \eta$	$\frac{1}{2}\left(\overline{\tau}R_{\max} + \eta T_{\max}R_{\max}\right)$

Table 1: Definition of random variables τ, r and r^*.

holding time is $T_{\min}$. These actions connect the s_0-states of the k copies in a A'-ary tree structure on the s_0 states: one action goes toward the root and A' actions goes toward the leaves (the same structure is described in section 6 of [Jaksch et al., 2010], see Figure 6 for a representation of the A'-ary tree). The diameter of such an SMDP M is at most $2\left(\frac{D}{4} + T_{\min}\lceil\log_{A'} k\rceil\right) \leq D$. All holding times of M are in $[T_{\min}, T_{\max}]$, and all rewards are in $[0, \frac{1}{2}R_{\max}T_{\max}]$. For all $(s, a) \in \mathcal{S} \times \mathcal{A}_s$ we have: $\overline{r}(s, a) \leq \overline{\tau}(s, a)R_{\max}$. Moreover, M has at most S states and A actions per state.

For the analysis, we will study the simpler-to-learn SMDP M'' where all s_0-states are merged together as well as all s_1-states. The "merged" s_0-state is set to be the initial state. M'' is thus equivalent to the two-state SMDP M' with kA' available actions in both s_0 and s_1. Let's assume that the learning algorithm $\mathfrak{A}$ used is fixed. The probability distribution of the stochastic process $(s_0, a_0, \tau_0, r_0, s_1, ...)$ is denoted:

- $\mathbb{P}_{a_0, a_1}$ when (a_0, a_1) are the best actions in respectively s_0 and s_1,
- $\mathbb{P}_*$ when the pair (a_0, a_1) identifying the best actions is first chosen uniformly at random from $\{1, ..., kA'\} \times \{1, ..., kA'\}$ before algorithm $\mathfrak{A}$ starts,
- $\mathbb{P}_{\text{unif0}, a1}$ when the pair a_1 is the best action in s_1 and $\epsilon = 0$ (no-optimal actions in s_0),
- $\mathbb{P}_{a0, \text{unif1}}$ when the pair a_0 is the best action in s_0 and $\eta = 0$ (no-optimal actions in s_1).

By construction, it is trivial to see that: $\mathbb{E}_*[\Delta(M, \mathfrak{A}, s, n)] \geq \mathbb{E}_*[\Delta(M'', \mathfrak{A}, s, n)]$ (M'' is easier to learn). We will show that $\mathbb{E}_*[\Delta(M'', \mathfrak{A}, s_0, n)] = \Omega\left(\left(\sqrt{D'} + \sqrt{T_{\max}}\right) R_{\max}\sqrt{kA'n}\right)$ and the same result can be proved with initial state s_1 using similar arguments. This will necessarily imply that there exists at least one choice of pair (a_0, a_1) for which, for all states s we have: $\mathbb{E}_{a_0, a_1}[\Delta(M, \mathfrak{A}, s, n)] = \Omega\left(\left(\sqrt{D} + \sqrt{T_{\max}}\right) R_{\max}\sqrt{SAn}\right)$.

As already argued by [Jaksch et al., 2010], for the analysis it is sufficient to consider algorithms with deterministic strategies for choosing actions.

We assume algorithm $\mathfrak{A}$ is run for n decision steps, which means that $n + 1$ states are visited in total (including the last state in which no action is taken). Let's denote by N_0 and N_1 the number of visits in states s_0 and s_1 respectively, last state excluded. For any $(a, \overline{a}) \in \mathcal{A}_{s_0} \times \mathcal{A}_{s_1}$, let's denote by N_0^a and $N_1^{\overline{a}}$ the respective number of times actions a and $\overline{a}$ are taken. Finally, let's denote by N_0^* and N_1^* the respective number of times best actions in s_0 and s_1 are taken. The SMDP M'' has the same transition probabilities as the MDP considered by [Jaksch et al., 2010] and we can use their proof to show that for any choice of best actions (a_0, a_1):

$$\mathbb{E}_{\text{unif0}, a_1}[N_1] \geq \frac{n}{2} - \frac{1}{2\delta} = \frac{n}{2} - \frac{D'}{2\overline{\tau}}$$

$$\mathbb{E}_{a_0, a_1}[N_1] \leq \frac{n}{2} + \frac{\mathbb{E}_{a_0, a_1}[N_0^*]\epsilon}{\delta} + \frac{1}{2\delta} = \frac{n}{2} + \frac{\mathbb{E}_{a_0, a_1}[N_0^{a_0}]\epsilon D'}{2\overline{\tau}} + \frac{D'}{2\overline{\tau}}$$

The regret is defined as:

$$\mathbb{E}_*[\Delta(M'', \mathfrak{A}, s_0, n)] = \mathbb{E}_*\left[\sum_{i=1}^{n} \tau_i(s_{i-1}, a_i, s_i)\right]\rho^* - \mathbb{E}_*\left[\sum_{i=1}^{n} r_i(s_{i-1}, a_i, s_i)\right]$$

where the total duration is simply:

$$\mathbb{E}_*\left[\sum_{i=1}^{n} \tau_i(s_{i-1}, a_i, s_i)\right] = \sum_{i=1}^{n} \mathbb{E}_*\left[\tau_i(s_{i-1}, a_i)\right] = n\overline{\tau}$$

and the cumulated reward is given by:

$$\mathbb{E}_* \left[\sum_{i=1}^n r_i(s_{i-1}, a_i, s_i) \right] = \sum_{i=1}^n \mathbb{E}_* \left[r_i(s_1, a_1^*) \mathbb{1}_{s_{i-1}=s_1, a_i=a_1^*} + r_i(s_1, a_i \neq a_1^*) \mathbb{1}_{s_{i-1}=s_1, a_i \neq a_1^*} \right]$$

$$= \mathbb{E}_* \left[\overline{r}^* \left(\sum_{i=0}^{n-1} \mathbb{1}_{s_i=s_1, a_{i+1}=a_1^*} \right) + \overline{r} \left(\sum_{i=0}^{n-1} \mathbb{1}_{s_i=s_1, a_{i+1}\neq a_1^*} \right) \right]$$

$$= \mathbb{E}_* \left[\overline{r} (N_1 - N_1^*) + \overline{r}^* N_1^* \right]$$

$$= \mathbb{E}_* \left[\overline{r} N_1 + (\overline{r}^* - \overline{r}) N_1^* \right]$$

$$= \frac{R_{\max}}{2} \overline{\tau} \mathbb{E}_* [N_1] + \eta \frac{R_{\max}}{2} T_{\max} \mathbb{E}_* [N_1^*]$$

hence the formula:

$$\mathbb{E}_* [\Delta(M'', \mathfrak{A}, s_0, n)] = \overline{\tau} \left(n\rho^* - \frac{R_{\max}}{2} \mathbb{E}_* [N_1] \right) - \eta \frac{R_{\max}}{2} T_{\max} \mathbb{E}_* [N_1^*] \tag{23}$$

Lemma 9. *Let* $f : \{s_0, s_1\}^{n+1} \times \{T_{\min}, T_{\max}\}^n \times \{0, \frac{1}{2} R_{\max} T_{\max}\}^n \to [0, M]$ *be any function defined on state/reward sequence* $(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \in \{s_0, s_1\}^{n+1} \times \{T_{\min}, T_{\max}\}^n \times \{0, \frac{1}{2} R_{\max} T_{\max}\}^n$ *observed in the SMDP* M''. *Then for any* $n \geq 1$, *any* $0 \leq \delta \leq \frac{1}{2}$, *any* $0 \leq \epsilon \leq 1 - 2\delta$, *any* $0 \leq p \leq \frac{1}{2}$, *any* $0 \leq \eta \leq 1 - 2p$, *and any* $(a_0, a_1) \in \{1, ..., kA'\} \times \{1, ..., kA'\}$:

$$\left| \mathbb{E}_{a_0, a_1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] - \mathbb{E}_{unif0, a_1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] \right| \leq \frac{M}{2} \frac{\epsilon}{\sqrt{\delta}} \sqrt{2 \mathbb{E}_{unif0, a_1} [N_0^{a_0}]}$$

$$\left| \mathbb{E}_{a_0, a_1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] - \mathbb{E}_{a_0, unif1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] \right| \leq \frac{M}{2} \frac{\eta}{\sqrt{p}} \sqrt{2 \mathbb{E}_{a_0, unif1} [N_1^{a_1}]}$$

Proof. We refer the reader to Appendix E of [Jaksch et al., 2010] where a similar Lemma is proved for the MDP-analogue of SMDP M''. In the following, we will only stress the main difference with the proof in [Jaksch et al., 2010]. We know from information theory that:

$$\left| \mathbb{E}_{a_0, a_1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] - \mathbb{E}_{\text{unif0}, a_1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] \right| \leq \frac{M}{2} \sqrt{2 \log(2) KL \left(\mathbb{P}_{\text{unif0}, a_1} \| \mathbb{P}_{a_0, a_1} \right)}$$

$$\left| \mathbb{E}_{a_0, a_1} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] - \mathbb{E}_{a_0, \text{unif1}} \left[f(\boldsymbol{s^{n+1}}, \boldsymbol{\tau^n}, \boldsymbol{r^n}) \right] \right| \leq \frac{M}{2} \sqrt{2 \log(2) KL \left(\mathbb{P}_{a_0, \text{unif1}} \| \mathbb{P}_{a_0, a_1} \right)}$$

By the chain rule of Kullback–Leibler divergences, it holds that:

$$KL \left(\mathbb{P}_{a_0, \text{unif1}} \| \mathbb{P}_{a_0, a_1} \right) = \sum_{i=1}^n KL \left(\mathbb{P}_{a_0, \text{unif1}}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}}) \| \mathbb{P}_{a_0, a_1}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}}) \right)$$

where $KL \left(\mathbb{P}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}}) \| \mathbb{Q}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}})) = $

$$\sum_{\boldsymbol{s^i} \in \mathcal{S}^i, \boldsymbol{\tau^i} \in \mathcal{T}, \boldsymbol{r^i} \in \mathcal{R}^i} \mathbb{P}(\boldsymbol{s^i}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^i}) \log_2 \left(\frac{\mathbb{P}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}})}{\mathbb{Q}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}})} \right)$$

with $\mathcal{S} = \{s_0, s_1\}$, $\mathcal{T} = \{T_{\min}, T_{\max}\}$ and $\mathcal{R} = \{0, \frac{1}{2} R_{\max} T_{\max}\}$. The same holds for $\mathbb{P}_{\text{unif0}, a_1}$.

Similarly to [Jaksch et al., 2010] and using the independence between s_i, τ_i and r_i, we obtain:

$$KL \left(\mathbb{P}_{\text{unif0}, a_1}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}}) \| \mathbb{P}_{a_0, a_1}(s_i, \tau_i, r_i | \boldsymbol{s^{i-1}}, \boldsymbol{\tau^{i-1}}, \boldsymbol{r^{i-1}}) \right)$$

$$= \mathbb{P}_{\text{unif0}, a_1}(s_{i-1} = s_0, a_i = a_0) \sum_{s' \in \mathcal{S}, \tau' \in \mathcal{T}, r' \in \mathcal{R}} \mathbb{P}_{\text{unif0}, a_1}(s', \tau', r' | s_0, a_0) \log_2 \left(\frac{\mathbb{P}_{\text{unif0}, a_1}(s', \tau', r' | s_0, a_0)}{\mathbb{P}_{a_0, a_1}(s', \tau', r' | s_0, a_0)} \right)$$

$$= \mathbb{P}_{\text{unif0}, a_1}(s_{i-1} = s_0, a_i = a_0) \sum_{s' \in \mathcal{S}} \mathbb{P}_{\text{unif0}, a_1}(s' | s_0, a_0) \log_2 \left(\frac{\mathbb{P}_{\text{unif0}, a_1}(s' | s_0, a_0)}{\mathbb{P}_{a_0, a_1}(s' | s_0, a_0)} \right)$$

$$= \mathbb{P}_{\text{unif0}, a_1}(s_{i-1} = s_0, a_i = a_0) \left(\delta \log_2 \left(\frac{\delta}{\delta + \epsilon} \right) + (1 - \delta) \log_2 \left(\frac{1 - \delta}{1 - \delta - \epsilon} \right) \right)$$

$$KL\left(\mathbb{P}_{a_0,\mathrm{unif1}}(s_i,\tau_i,r_i|\boldsymbol{s^{i-1}},\boldsymbol{\tau^{i-1}},\boldsymbol{r^{i-1}})\|\mathbb{P}_{a_0,a_1}(s_i,\tau_i,r_i|\boldsymbol{s^{i-1}},\boldsymbol{\tau^{i-1}},\boldsymbol{r^{i-1}})\right)$$

$$= \mathbb{P}_{a_0,\mathrm{unif1}}(s_{i-1}=s_1,a_i=a_1)\sum_{s'\in\mathcal{S},\tau'\in\mathcal{T},r'\in\mathcal{R}}\mathbb{P}_{a_0,\mathrm{unif1}}(s',\tau',r'|s_1,a_1)\log_2\left(\frac{\mathbb{P}_{a_0,\mathrm{unif1}}(s',\tau',r'|s_1,a_1)}{\mathbb{P}_{a_0,a_1}(s',\tau',r'|s_1,a_1)}\right)$$

$$= \mathbb{P}_{a_0,\mathrm{unif1}}(s_{i-1}=s_1,a_i=a_1)\sum_{r'\in\mathcal{R}}\mathbb{P}_{a_0,\mathrm{unif1}}(r'|s_1,a_1)\log_2\left(\frac{\mathbb{P}_{a_0,\mathrm{unif1}}(r'|s_1,a_1)}{\mathbb{P}_{a_0,a_1}(r'|s_1,a_1)}\right)$$

$$= \mathbb{P}_{a_0,\mathrm{unif1}}(s_{i-1}=s_1,a_i=a_1)\left(p\log_2\left(\frac{p}{p+\eta}\right)+(1-p)\log_2\left(\frac{1-p}{1-p-\eta}\right)\right)$$

Using Lemma 20 of [Jaksch et al., 2010] we have that under conditions $0\le\delta\le\frac{1}{2}$, $0\le\epsilon\le1-2\delta$, $0\le p\le\frac{1}{2}$, and $0\le\eta\le1-2p$ the following inequalities hold:

$$\delta\log_2\left(\frac{\delta}{\delta+\epsilon}\right)+(1-\delta)\log_2\left(\frac{1-\delta}{1-\delta-\epsilon}\right)\le\frac{\epsilon^2}{\delta\log(2)}$$

$$\text{and }\ p\log_2\left(\frac{p}{p+\eta}\right)+(1-p)\log_2\left(\frac{1-p}{1-p-\eta}\right)\le\frac{\eta^2}{p\log(2)}$$

which concludes the proof. $\qquad\square$

Note that by assumption: $\epsilon\le\delta\le\frac{1}{3}\le1-2\delta$ and $\eta\le p\le\frac{1}{3}\le1-2p$.

We can bound $\mathbb{E}_*[N_1]$ using Lemma 9 as is done in [Jaksch et al., 2010]. This is because N_1 can be written as a function of $(\boldsymbol{s^{n+1}},\boldsymbol{\tau^n},\boldsymbol{r^n})$. Since the computations are rigorously the same except that $\delta=\frac{\overline{\tau}}{D'}$ instead of $\frac{1}{D'}$, we give the results without any further details:

$$\mathbb{E}_*[N_1]\le\frac{n}{2}+\frac{D'}{2\overline{\tau}}+\frac{\epsilon nD'}{2\overline{\tau}kA'}+\frac{\epsilon D'^2}{2\overline{\tau}^2kA'}+\frac{\epsilon^2nD'}{2\overline{\tau}kA'}\sqrt{\frac{D'kA'n}{\overline{\tau}}}+\frac{\epsilon^2nD'^2}{2\overline{\tau}^2kA'}\sqrt{kA'}$$

Taking into account the fact that by assumption $n\ge DSA\ge16D'kA'$ we get:

$$\mathbb{E}_*[N_1]\le\frac{n}{2}+\frac{D'}{2\overline{\tau}}+\frac{\epsilon nD'}{\overline{\tau}}\left(\frac{1}{2kA'}+\frac{1}{32\overline{\tau}k^2A'^2}\right)+\frac{\epsilon^2nD'}{\overline{\tau}kA'}\sqrt{\frac{D'kA'n}{\overline{\tau}}}\left(\frac{1}{2}+\frac{1}{8\sqrt{kA'}}\right) \tag{24}$$

Given that $\delta\ge\epsilon\ge0$ we have the following inequality:

$$\rho^*-\frac{R_{\max}}{4}=\frac{R_{\max}}{2}\times\frac{\epsilon\overline{\tau}+(\delta+\epsilon)\eta T_{\max}}{2(2\delta+\epsilon)\overline{\tau}}\ge\frac{R_{\max}}{2}\times\left(\frac{\epsilon D'}{6\overline{\tau}}+\frac{\eta T_{\max}}{6\overline{\tau}}\right) \tag{25}$$

Applying Lemma 9 (N_1^* is a function of $(\boldsymbol{s^{n+1}},\boldsymbol{\tau^n},\boldsymbol{r^n})$) and Jensen's inequality we get $\forall a_0\in\{1,...,kA'\}$:

$$\mathbb{E}_*[N_1^*]=\frac{1}{kA'}\sum_{a_1=1}^{kA'}\mathbb{E}_{a_0,a_1}[N_1^{a_1}]\le\frac{\mathbb{E}_{a_0,\mathrm{unif1}}[N_1]}{kA'}+\frac{n}{2kA'}\eta\sqrt{\frac{2kA'T_{\max}}{\overline{\tau}}\mathbb{E}_{a_0,\mathrm{unif1}}[N_1]}$$

We will now derive an upper-bound on $\mathbb{E}_{a_0,\mathrm{unif1}}[N_1]$.

Lemma 10. *Let $(u_n)_{n\in\mathbb{N}}\in\mathbb{R}^{\mathbb{N}}$ be any real sequence satisfying the following arithmetico-geometric recurrence relation:*

$$\forall n\in\mathbb{N},\quad u_{n+1}\ge qu_n+r\quad\text{where }(q,r)\in\mathbb{R}\setminus\{1\}\times\mathbb{R}$$

Then we have that:

$$\forall n\in\mathbb{N},\quad u_n\ge\left(u_0-\frac{r}{1-q}\right)\alpha^n+\frac{r}{1-q}$$

Proof. Defining sequence $v_n = u_n - \frac{r}{1-q}$, we have:

$$\forall n \in \mathbb{N}, \quad v_{n+1} = u_{n+1} - \frac{r}{1-q} \geq qu_n + r - \frac{r}{1-q} = q\left(u_n - \frac{r}{1-q}\right) = qv_n$$

By trivial induction we get: $\forall n \in \mathbb{N}, \quad v_n \geq v_0 q^n$. The result follows by replacing v_n by $u_n - \frac{r}{1-q}$. $\square$

By the law of total probability and since $\mathbb{P}_{a_0,\text{unif1}}(s_i = s_0) + \mathbb{P}_{a_0,\text{unif1}}(s_i = s_1) = 1$ we have:

$$\begin{aligned}
\forall i \geq 0, \quad \mathbb{P}_{a_0,\text{unif1}}(s_{i+1} = s_0) =& \mathbb{P}_{a_0,\text{unif1}}(s_{i+1} = s_0 | s_i = s_0) \mathbb{P}_{a_0,\text{unif1}}(s_i = s_0) \\
&+ \mathbb{P}_{a_0,\text{unif1}}(s_{i+1} = s_0 | s_i = s_1) \mathbb{P}_{a_0,\text{unif1}}(s_i = s_1) \\
=& \mathbb{P}_{a_0,\text{unif1}}(s_i = s_0)\left(\mathbb{P}_{a_0,\text{unif1}}(s_{i+1} = s_0 | s_i = s_0)\right. \\
&\left.- \mathbb{P}_{a_0,\text{unif1}}(s_{i+1} = s_0 | s_i = s_1)\right) + \mathbb{P}_{a_0,\text{unif1}}(s_{i+1} = s_0 | s_i = s_1) \\
\geq& \mathbb{P}_{a_0,\text{unif1}}(s_i = s_0)(1 - 2\delta - \epsilon) + \delta
\end{aligned}$$

Since the initial state is s_0 and $2\delta + \epsilon \leq 3\delta$ we have by Lemma 10:

$$\forall i \geq 0, \quad \mathbb{P}_{a_0,\text{unif1}}(s_i = s_0) \geq \left(1 - \frac{\delta}{2\delta + \epsilon}\right)(1 - 2\delta - \epsilon)^i + \frac{\delta}{2\delta + \epsilon} \geq \frac{1}{3}$$

$$\implies \mathbb{E}_{a_0,\text{unif1}}[N_0] = \sum_{i=0}^{n} \mathbb{P}_{a_0,\text{unif1}}(s_i = s_0) \geq \frac{n}{3}$$

$$\implies \mathbb{E}_{a_0,\text{unif1}}[N_1] \leq \frac{2n}{3}$$

$$\implies \mathbb{E}_*[N_1^*] \leq \frac{2n}{3kA'} + \frac{n\eta}{kA'}\sqrt{\frac{T_{\max}nkA'}{3\overline{\tau}}}$$

Hence the bound:

$$\eta T_{\max} \mathbb{E}_*[N_1^*] \leq \frac{2n\eta T_{\max}}{3kA'} + \frac{n\eta^2 T_{\max}}{kA'}\sqrt{\frac{T_{\max}nkA'}{3\overline{\tau}}} \tag{26}$$

By setting $\epsilon = c\sqrt{\frac{kA'}{nD'}}$ and $\eta = \kappa\sqrt{\frac{kA'}{nT_{\max}}}$ and incorporating inequalities 24, 25 and 26 into inequality 23 we obtain:

$$\begin{aligned}
\mathbb{E}_*[\Delta(M'', \mathfrak{A}, s_0, n)] \geq& \left[\frac{c}{6} - \frac{c}{2kA'} - \frac{c}{32k^2A'^2\overline{\tau}} - \frac{c^2}{2\sqrt{\overline{\tau}}} - \frac{c^2}{8\sqrt{kA'\overline{\tau}}} - \frac{1}{8kA'}\right]\frac{R_{\max}}{2}\sqrt{D'kA'n} \\
&+ \left[\frac{\kappa}{6} - \frac{2\kappa}{3kA'} - \frac{\kappa^2}{\sqrt{3\overline{\tau}}}\right]\frac{R_{\max}}{2}\sqrt{T_{\max}kA'n} \\
\geq& \left[\frac{c}{6} - \frac{c}{2kA'} - \frac{c}{32k^2A'^2} - \frac{c^2}{2} - \frac{c^2}{8\sqrt{kA'}} - \frac{1}{8kA'}\right]\frac{R_{\max}}{2}\sqrt{D'kA'n} \\
&+ \left[\frac{\kappa}{6} - \frac{2\kappa}{3kA'} - \frac{\kappa^2}{\sqrt{3}}\right]\frac{R_{\max}}{2}\sqrt{T_{\max}kA'n}
\end{aligned}$$

For the second inequality, we used the fact that $\overline{\tau} > T_{\min} \geq 1$ (by assumption). If c and κ are sufficiently small, then the conditions of lemma 9 are indeed satisfied and the above polynomials in c and κ are non-negative. For example, if $c = \kappa = \frac{1}{5}$:

$$n \geq DSA \geq 16D'kA' \implies \epsilon = \frac{1}{5}\sqrt{\frac{kA'}{nD'}} \leq \frac{\delta}{20} < \delta$$

$$n \geq T_{\max}SA \geq 4T_{\max}kA' \implies \eta = \frac{1}{5}\sqrt{\frac{kA'}{nT_{\max}}} \leq \frac{p}{10} < p$$

$$\text{and } \mathbb{E}_*[\Delta(M'', \mathfrak{A}, s_0, n)] \geq 0.0015 \times \left(\sqrt{D'} + \sqrt{T_{\max}}\right)R_{\max}\sqrt{kA'n}$$
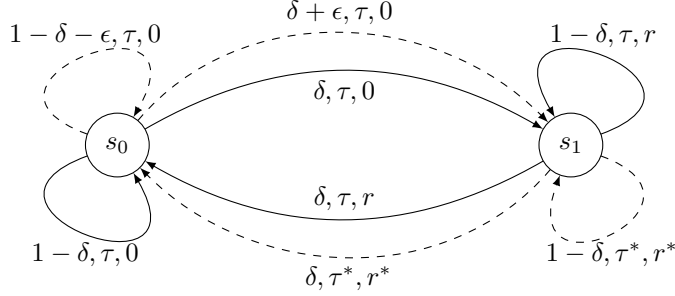
Figure 5: The two-state SMDP M' for the lower-bound on MDPs with options. The two special actions a_0^* and a_1^* are shown as dashed lines.

C.2 Lower Bound for MDPs with options

We first note that SMDP M' depicted in Fig. 4 cannot be converted into an MDP with options. This is due to the fact that $\tau_i(s_{i-1}, a_i, s_i)$ and $r_i(s_{i-1}, a_i, s_i)$ were assumed to be independent and the fact that $\mathbb{P}(r^* = \frac{1}{2}R_{\max}T_{\max}) > \mathbb{P}(\tau = T_{\max})$. However, it is possible to prove a slightly smaller lower bound for a family of SMDPs that can be transformed into an equivalent MDP with options. We will first present the SMDPs and the lower bound and then we will describe how to transform them into an MDP with options.

The family of SMDPs is constructed in the same way as previously except that we use a slightly different SMDP M', represented on Fig. 5 with the random variables given in Table 2. We take: $p = \frac{1}{2(T_{\max} - T_{\min})}$ and $\delta = \frac{4\overline{\tau}}{D}$. By assumption $p < \frac{1}{3}$ and $\delta < \frac{1}{2}$. As before, we assume that for all i, $\tau_i(s_{i-1}, a_i, s_i)$ and $r_i(s_{i-1}, a_i, s_i)$ are independent of the next state s_i. But the main difference with the previous lower bound is that we assume that r_i and τ_i are strongly correlated, namely: $r_i(s_{i-1}, a_i, s_i) = R_{\max}\mathbb{1}_{\{s_{i-1}=s_1\}}\tau_i(s_{i-1}, a_i)$. We assume $\epsilon \le \delta$ and $\eta \le 1 - 2p$. The optimal gain of M' is reached when a_0^* and a_1^* are chosen in s_0 and s_1 respectively and is equal to:

$$\rho^* = R_{\max} \times \frac{(\delta + \epsilon)\left(\overline{\tau} + \eta(T_{\max} - T_{\min})\right)}{(2\delta + \epsilon)\overline{\tau} + (\delta + \epsilon)\left(T_{\max} - T_{\min}\right)}$$

By adapting the proof of the lower bound for general SMDPs, we can obtain the following result:

$$
\begin{aligned}
\mathbb{E}_*[\Delta(M'', \mathfrak{A}, s_0, n)] &\ge \left[\frac{c}{6} - \frac{c}{2kA'} - \frac{c}{32k^2A'^2\overline{\tau}} - \frac{c^2}{2\sqrt{\overline{\tau}}} - \frac{c^2}{8\sqrt{kA'\overline{\tau}}} - \frac{1}{8kA'}\right] R_{\max}\sqrt{D'kA'n} \\
&+ \left[\frac{\kappa}{6} - \frac{2\kappa}{3kA'} - \frac{\kappa^2}{\sqrt{3/2}}\right] R_{\max}\sqrt{(T_{\max} - T_{\min})kA'n} \\
&\ge \left[\frac{c}{6} - \frac{c}{2kA'} - \frac{c}{32k^2A'^2} - \frac{c^2}{2} - \frac{c^2}{8\sqrt{kA'}} - \frac{1}{8kA'}\right] R_{\max}\sqrt{D'kA'n} \\
&+ \left[\frac{\kappa}{6} - \frac{2\kappa}{3kA'} - \frac{\kappa^2}{\sqrt{3/2}}\right] R_{\max}\sqrt{(T_{\max} - T_{\min})kA'n}
\end{aligned}
$$

by setting $\epsilon = c\sqrt{\frac{kA'}{nD'}}$ and $\eta = \kappa\sqrt{\frac{kA'}{n(T_{\max}-T_{\min})}}$. It is then possible to tune c and κ so that ϵ and η satisfy the constraints and:

$$\mathbb{E}_*[\Delta(M'', \mathfrak{A}, s_0, n)] = \Omega\left(\left(\sqrt{D'} + \sqrt{T_{\max} - T_{\min}}\right) R_{\max}\sqrt{kA'n}\right)$$

SMDPs M' and M'' can be transformed into equivalent MDPs with options. We illustrate this transformation on Fig. 6 for an action $a_1 \in \mathcal{A}_{s_1}$ different than a_1^*. The same method can be applied for the other actions. The idea consists in adding new states (blank states in Fig. 6) and primitive actions between those states. Note that the states added are just "hidden" states from which no option can be started. Thus, they should not be counted in the number of states for the lower bound. In our example it is sufficient to consider primitive actions with constant (i.e., deterministic) reward. On Fig. 6 we give the probabilities of each primitive action when $a_1 \in \mathcal{A}_{s_1} \setminus \{a_1^*\}$ is executed.

R.v. X	$X_{\min}$	$X_{\max}$	$\mathbb{P}(X = X_{\min})$	$\mathbb{P}(X = X_{\max})$	$\mathbb{E}[X]$
τ	$T_{\min}$	$T_{\max}$	$1-p$	p	$\overline{\tau} = T_{\min} + p(T_{\max} - T_{\min})$
τ^*	$T_{\min}$	$T_{\max}$	$1-p-\eta$	$p+\eta$	$\overline{\tau} + \eta(T_{\max} - T_{\min})$
r	$R_{\max}T_{\min}$	$R_{\max}T_{\max}$	$1-p$	p	$\overline{\tau}R_{\max}$
r^*	$R_{\max}T_{\min}$	$R_{\max}T_{\max}$	$1-p-\eta$	$p+\eta$	$\overline{\tau}R_{\max} + \eta R_{\max}(T_{\max} - T_{\min})$

Table 2: Definition of random variables τ, τ^*, r and r^*.
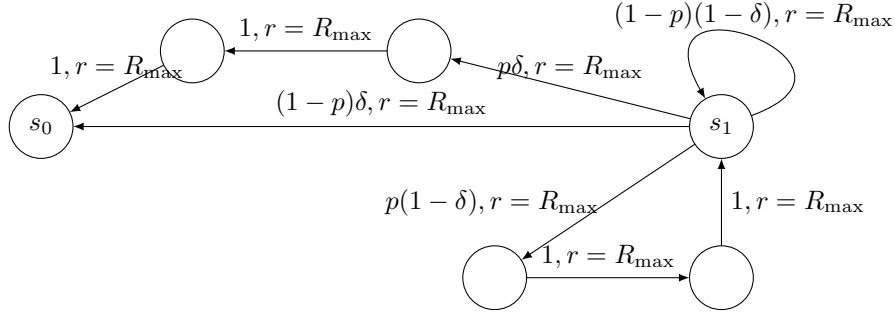
Figure 6: Decomposition of an action $a_1 \in \mathcal{A}_{s_1} \setminus \{a_1^*\}$ into a set of primitive actions in an MDP. In this example, $T_{\max} = 3$ and $T_{\min} = 1$.

D   Distribution of the holding time and reward of a finite Markov option (Lem. 3)

We denote by $\mathbb{R}^+$ and $\mathbb{R}^{+*}$ the set of positive and non-negative reals respectively. For the definition of sub-exponential random variables, we refer to Def. 4, while sub-Gaussian random variables are defined as follows.

Definition 5 (Wainwright [2015]). *A random variable X with mean $\mu < +\infty$ is said to be sub-Gaussian if and only if there exists $\sigma \in \mathbb{R}^+$ such that:*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{for all} \quad \lambda \in \mathbb{R}. \tag{27}$$

A finite[7] Markov option can be seen as an absorbing Markov Chain together with a reward process (i.e., a finite Markov option can be seen as an absorbing Markov Reward Process). To see this we add a new state $\widetilde{s}$ for every state s for which $\beta_o(s) > 0$. We then add a transitions from s to $\widetilde{s}$ with probability $\beta_o(s) > 0$ and reward 0, and we add a self-loop on $\widetilde{s}$ with probability 1 and reward 0 ($\widetilde{s}$ is an absorbing state). The Markov Reward Process obtained is indeed absorbing since we assumed the option to be a.s. finite, and it is equivalent to the original option (same reward and holding time). Let's denote by P the transition matrix of the Markov Chain. In canonical form we have:

$$P = \begin{bmatrix} Q & R \\ 0 & I_r \end{bmatrix}$$

where r is the number of absorbing states, I_r is the identity matrix of dimension r, Q is the transition matrix between non-absorbing states and R the transition matrix from non-absorbing to absorbing states. If the option is a.s. finite then Q is necessarily (strictly) sub-stochastic ($Qe \leq e$ where $e = (1, ..., 1)^\intercal$ and $\exists j$ s.t. $(Qe)_j < 1$) and irreducible (no recurrent class). It is well-known that such a matrix has a spectral radius strictly smaller than 1 ($\rho(Q) < 1$) and thus $I - Q$ is invertible (where I is the identity matrix). The holding time $\tau(s, o, s')$ of any option o is defined as the first time absorbing state s' is reached starting from state s: $\inf\{n \geq 1 : s_n = s' \text{ with } s_0 = s\}$ where $(s_n)_n$ is the sequence of states in the absorbing Markov Chain defined by o. It is well-known in the literature [Peter Buchholz, 2014] that this type of stopping times have Discrete Phase-Type distributions, with probability mass function given by:

$$\forall k \in \mathbb{N}^*, \quad \mathbb{P}(\tau(s, a, s') = k) = e_s^\intercal Q^{k-1} R e_{s'}$$

---

[7]Note that if at least one option is not (almost surely) finite, the learning agent can potentially be stuck executing that option forever and the problem is ill-posed.

where $e_s = (0, 0, ...0, 1, 0, ..., 0)^\intercal$ is a vector of all zeros except in state s where it equals 1. These distributions generalize the geometric distribution (defined in dimension 1) to higher dimensions. The Laplace transform can be computed as follows (we simplify notations and denote: $\tau \leftarrow \tau(s, o, s')$ and $\overline{\tau} \leftarrow \overline{\tau}(s, o, s') = \mathbb{E}[\tau(s, o, s')]$):

$$\mathbb{E}\left[e^{\lambda(\tau - \overline{\tau})}\right] = \sum_{k=1}^{\infty} e^{\lambda(k - \overline{\tau})} e_s^\intercal Q^{k-1} R e_{s'} = e^{\lambda(1 - \overline{\tau})} e_s^\intercal \left[\sum_{k=0}^{\infty} \left(e^\lambda Q\right)^k\right] R e_{s'}$$

The term $\sum_{k=0}^{\infty} \left(e^\lambda Q\right)^k$ is finite if and only if $e^\lambda \rho(Q) < 1$, in which case we have:

$$\mathbb{E}\left[e^{\lambda(\tau - \overline{\tau})}\right] = e^{\lambda(1 - \overline{\tau})} e_s^\intercal \left(I - e^\lambda Q\right)^{-1} R e_{s'}$$

and otherwise: $\mathbb{E}\left[e^{\lambda(\tau - \overline{\tau})}\right] = +\infty$. Note that $e^\lambda \rho(Q) < 1$ if and only if either $\lambda < -\log\left(\rho(Q)\right)$ or $\rho(Q) = 0$. We will now analyse the two cases separately:

1. $\rho(Q) = 0$ if and only if all the eigenvalues of Q in $\mathbb{C}$ are 0, if and only if Q is nilpotent ($\exists n > 0$ s.t. $Q^n = 0$). This is because Q can always be triangularized in $\mathbb{C}$: $Q = UTU^{-1}$ where T is upper-triangular with the eigenvalues of Q on the diagonal that is, only zeros if $\rho(Q) = 0$. This implies that $\exists n > 0$ s.t. $T^n = U^{-1}Q^n U = 0 \implies Q^n = 0$ hence Q is nilpotent. The reverse is obviously true: if Q is nilpotent then $\rho(Q) = 0$, (otherwise there would exist $\lambda \neq 0$, $v \neq 0$ and $n > 0$ s.t. $Q^n = 0$ and $Qv = \lambda v \implies Q^n v = \lambda^n v = 0$, which is absurd). By definition, matrix Q is nilpotent of order n if and only if the Markov Chain reaches an absorbing state in at most n steps (a.s.). In conclusion, $\rho(Q) = 0$ if and only if the option is almost surely bounded. This happens if and only if there is no cycle in the option (with probability 1, every non-absorbing state is visited at most once).

2. In the case where $\rho(Q) > 0$: it is clear that $\mathbb{E}\left[e^{\lambda(\tau - \overline{\tau})}\right]$ can not be bounded by a function of the form $\lambda \to e^{\frac{\sigma^2 \lambda^2}{2}}$ for $\lambda \geq -\log\left(\rho(Q)\right)$ so $\tau(s, o, s')$ is not sub-Gaussian (Definition 5). However, since $\rho(Q) < 1$ we can choose $0 < c_0 < -\log\left(\rho(Q)\right)$ and we have $\mathbb{E}\left[e^{\lambda(\tau - \overline{\tau})}\right] < +\infty$ for all $|\lambda| < c_0$, which implies that $\tau(s, o, s')$ is sub-exponential (Definition 4).

In conclusion, either option o contains inner-loops (some states are visited several times with non-zero probability) in which case the distribution of $\tau(s, o, s')$ is sub-Exponential but not sub-Gaussian, or o has no inner-loop in which case o is bounded (and thus sub-Gaussian). There is no other alternative.

The distribution of rewards $r(s, o, s')$ is not as simple: the reward of an option is the sum of all micro-rewards obtained at every time step before the option ends, and every micro-reward earned at each time step can have a different distribution. The only constraint is that all micro-rewards should be (a.s.) bounded between 0 and $R_{\max}$. As a result, if $\tau(s, o, s')$ is a.s. bounded (by let's say $T_{\max}$) then $r(s, o, s')$ is also a.s. bounded (by $R_{\max} T_{\max}$). But if $\tau(s, o, s')$ is unbounded then $r(s, o, s')$ may still be bounded if for example, all micro-rewards are 0. If however all micro-rewards are equal to $R_{\max}$ then $r(s, o, s')$ has a discrete phase-type distribution just like $\tau(s, o, s')$. $r(s, o, s')$ can thus be unbounded (and even not sub-Gaussian). However, we will show that $r(s, o, s')$ is always sub-Exponential. Using the law of total expectations and the fact that $\mathbb{P}\left(r \leq R_{\max}\tau\right) = 1$ we have:

$$\forall \lambda > 0, \quad \mathbb{E}\left[e^{\lambda(r - \overline{r})}\right] = \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(r - \overline{r})}\big|\tau = k\right] \mathbb{P}(\tau = k) \leq \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(R_{\max}\tau - \overline{r})}\big|\tau = k\right] \mathbb{P}(\tau = k)$$

$$= \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(R_{\max}k - \overline{r})}\big|\tau = k\right] \mathbb{P}(\tau = k)$$

$$= \sum_{k=1}^{\infty} e^{\lambda(R_{\max}k - \overline{r})} \mathbb{P}(\tau = k)$$

$$= e^{\lambda(R_{\max} - \overline{r})} e_s^\intercal \left[\sum_{k=0}^{\infty} \left(e^{\lambda R_{\max}} Q\right)^k\right] R e_{s'}$$

We can now conclude as we did for $\tau(s, o, s')$: let $0 < c_0 < -\frac{\log(\rho(Q))}{R_{\max}}$, for all $0 < \lambda < c_0$ the quantity $\mathbb{E}\left[e^{\lambda(r - \overline{r})}\right]$ is finite. Note that for $\lambda \leq 0$: $\mathbb{E}\left[e^{\lambda r}\right] \leq 1$ so $\mathbb{E}\left[e^{\lambda(r - \overline{r})}\right] < +\infty$. By Definition 4, $r(s, o, s')$ is sub-Exponential.

E   Equivalent policies in an MDP with options and the induced SMDP (Lem. 2)

We consider the original MDP M and the SMDP $M_{\mathcal{O}}$ induced by the set of options $\mathcal{O}$. By definition of $M_{\mathcal{O}}$, the reward of an option is equal to the sum of the rewards of all the primitive actions taken until the option terminates (when the option is executed in M). Therefore $\sum_{i=1}^{n} r_{\mathcal{O}}^i = \sum_{i=1}^{N(T_n)} r_{\mathcal{O}}^i = \sum_{t=1}^{T_n} r_t$ and:

$$
\begin{aligned}
\Delta(M, \mathfrak{A}, s, T_n) &= T_n \rho^*(M) - \sum_{t=1}^{T_n} r_t \\
&= T_n \rho^*(M') + T_n \left(\rho^*(M) - \rho^*(M_{\mathcal{O}}) \right) - \sum_{i=1}^{n} r_{\mathcal{O}}^i \\
&= \Delta(M_{\mathcal{O}}, \mathfrak{A}, s, n) + T_n \left(\rho^*(M) - \rho^*(M_{\mathcal{O}}) \right)
\end{aligned}
\tag{28}
$$

The second part of Lem. 2 is thus proved. We now define the (finite-time) average reward in the two processes

$$
\forall T \in \mathbb{N}^*, \quad \rho^\pi(M, s, T) = \mathbb{E}_M^\pi \left[\left. \frac{\sum_{t=1}^{T} r_t}{T} \right| s_0 = s \right]
$$

$$
\forall T' \in \mathbb{R}^{+*}, \quad \rho^{\pi_{\mathcal{O}}}(M_{\mathcal{O}}, s, T') = \mathbb{E}_{M_{\mathcal{O}}}^{\pi_{\mathcal{O}}} \left[\left. \frac{\sum_{i=1}^{N(T')} r_{\mathcal{O}}^i}{T'} \right| s_0 = s \right].
$$

The limit $\lim_{n \to +\infty} T_n = +\infty$ since the sequence $(T_n)_{n \in \mathbb{N}^*}$ is strictly increasing and unbounded (at least one primitive action is executed before the option ends: $\forall n \geq 1, T_{n+1} \geq T_n + 1$). Moreover, $\lim_{T' \to +\infty} \rho^{\pi_{\mathcal{O}}}(M_{\mathcal{O}}, s, T')$ exists since $\pi_{\mathcal{O}}$ is stationary and deterministic (see appendix A) and by composition of the limit we have

$$
\lim_{n \to +\infty} \rho^{\pi_{\mathcal{O}}}(M_{\mathcal{O}}, s, T_n) = \lim_{T' \to +\infty} \rho^{\pi_{\mathcal{O}}}(M_{\mathcal{O}}, s, T') = \rho^{\pi_{\mathcal{O}}}(M_{\mathcal{O}}, s)
$$

The limit $\lim_{T \to +\infty} \rho^\pi(M, s, T)$ also exists. To see this, we can build an augmented MDP equivalent to M where the state and actions encountered in two different options are duplicated (see section 3 of [Levy and Shimkin, 2012]). The equivalence between the original and augmented MDPs is in the strong sense: for any optional policy, the corresponding policy in the augmented MDP yields exactly the same reward for any finite horizon. In the augmented MDP, policy π is stationary deterministic and we know from MDP theory [Puterman, 1994] that the corresponding average reward exists. We also have:

$$
\forall n \geq 1, \quad \mathbb{E}_M^\pi \left[\left. \frac{\sum_{t=1}^{T_n} r_t}{T_n} \right| s_0 = s \right] = \mathbb{E}_{M_{\mathcal{O}}}^{\pi_{\mathcal{O}}} \left[\left. \frac{\sum_{i=1}^{n} r_{\mathcal{O}}^i}{T_n} \right| s_0 = s \right]
$$

$$
\implies \rho^{\pi_{\mathcal{O}}}(M_{\mathcal{O}}, s) = \lim_{n \to +\infty} \rho^\pi(M, s, T_n) = \lim_{T \to +\infty} \rho^\pi(M, s, T) = \rho^\pi(M, s)
$$

The first part of Lem. 2 is thus proved. Finally, we know from the literature [Puterman, 1994] that there exists a stationary deterministic optimal policy in the augmented MDP and thus there also exists a stationary deterministic optional policy (a policy using only options in $\mathcal{O}$) in the original MDP. As a result, $\rho^*(M_{\mathcal{O}})$ is the maximal average reward achievable in M using only options in $\mathcal{O}$. In conclusion, the linear term $T_n \left(\rho^*(M) - \rho^*(M_{\mathcal{O}}) \right)$ in equation 28 is the minimal asymptotic regret incurred if the learning agent decides to only use options. This linear loss is unavoidable.

F   Details of the illustrative experiments

In this section we will detail the experiments described in section 6.

Terminating Condition   Let's denote the current state by s_0 and for all $k \in \{1...m\}$, denote by s_k the state which is k steps on the left to s_0. Assume option $LEFT$ is taken in s_0. By definition, once s_k is reached, the probability of ending the option is given by $\beta_o(s_k) = 1/(m - k + 1)$. Since all transitions in the MDP have probability 1 (except at the target), the probability of ending in exactly k steps can be computed as follows:
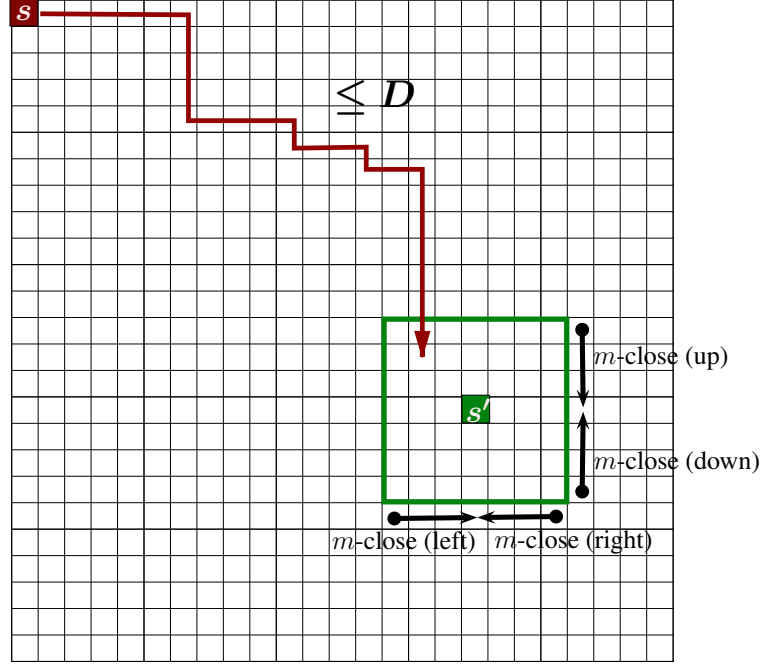
Figure 7: Upper bound on the diameter of the SMDP used in the experiments.

- If $k = 1$:

$$\mathbb{P}(\tau = 1) = \beta_o(s_1) = \frac{1}{m}$$

- If $k \geq 1$:

$$
\begin{aligned}
\mathbb{P}(\tau = k) &= \left(\prod_{i=1}^{k-1} (1 - \beta_o(s_i)) \right) \times \beta_o(s_k) \\
&= \left(\prod_{i=1}^{k-1} \left(1 - \frac{1}{m-i+1} \right) \right) \times \frac{1}{m-k+1} \\
&= \left(\prod_{i=1}^{k-1} \left(\frac{m-i}{m-i+1} \right) \right) \times \frac{1}{m-k+1} = \frac{1}{m}
\end{aligned}
$$

By symmetry, the other options ($RIGHT$, UP and $DOWN$) have the same holding time.

Expected Holding Time Based on the previous result, we can easily compute the expected holding time:

$$\mathbb{E}[\tau] = \sum_{k=1}^{m} k \cdot \mathbb{P}(\tau = k) = \frac{1}{m} \sum_{k=1}^{m} k = \frac{m+1}{2}$$

Diameter Let s and s' be two distinct states in the grid. With the options defined above, the expected shortest path from s to s' is obtained if in each visited state on the way to s', we choose an option that goes in the direction of s'. For example, if s is the state located in the top left corner of the grid and s' is the target, the expected shortest path is obtained when either $RIGHT$ or $DOWN$ is taken in every state. With this policy, the expected time to get m-close to s' both horizontally and vertically is trivially bounded by D (red path on Fig. 7). Once we are m-close to s' (green square on Fig. 7, $m = 3$ on this example), we will potentially start cycling until we reach s'. On Fig. 8, we give an example (in one dimension) of a possible path before reaching s' once in an m-close state (the green arrows represent the successive transitions, and $m = 3$ on this example). Since all
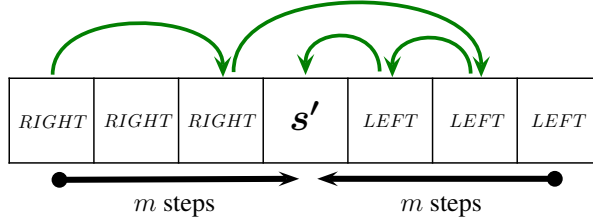
Figure 8: Behaviour of the agent in m-close states.

options end after at most m time steps, once we are m-close to s', we stay m-close with the chosen policy. The expected time it takes to reach s' once we are m-close to it is $m(m+1)/2$ both horizontally and vertically. To prove this, we need to solve a linear system. For all $i \in \{1...m-1\}$, denote by τ_i the time it takes to go from s to the i-th state to the left (respectively right, up or down) when the option chosen is left (respectively right, up or down). The value is the same in all directions by symmetry. We can express the τ_i as follows:

$$\tau_1 = \frac{1}{m} + \frac{1}{m}(2 + \tau_1) + \cdots + \frac{1}{m}(m + \tau_{m-1})$$
$$\tau_2 = \frac{1}{m} \times 2 + \frac{1}{m}(1 + \tau_1) + \frac{1}{m}(3 + \tau_1) + \cdots + \frac{1}{m}(m + \tau_{m-2})$$
$$\tau_3 = \frac{1}{m} \times 3 + \frac{1}{m}(1 + \tau_2) + \frac{1}{m}(2 + \tau_1) + \frac{1}{m}(4 + \tau_1) + \cdots + \frac{1}{m}(m + \tau_{m-3}) \tag{29}$$
$$\cdots$$
$$\tau_i = \frac{m+1}{2} + \frac{1}{m}\sum_{j=1}^{i-1}\tau_j + \frac{1}{m}\sum_{j=1}^{m-i}\tau_j$$

With probability $1/m$, the next state after executing the option is 1 step to the left of s and the value of τ_1 is then 1 . With probability $1/m$ the next state is 2 steps to the left of s and so s' is now located 1 step to the right of the new state: the value of τ_1 is thus $2 + \tau_1$. With probability $1/m$ the next state is 3 steps to the left of s and and so s' is now located 2 steps to the right of the new state: the value of τ_1 is thus $3 + \tau_2$. And so on and so forth. What we used here is basically the law of total expectations where the partition of events is the set of all possible states reached after executing the option only once. The same thing can be done for $\tau_2 \ldots \tau_{m-1}$. It is trivial to verify that the only solution of the linear system in equation 29 is: $\tau_i = m(m+1)/2, \forall i \in \{1...m-1\}$. This results is rather intuitive: m corresponds to the expected number of times the option needs to be executed to end up in the desired state s' whereas $(m+1)/2$ is the expected duration at every decision step. The simplicity of this result comes from the symmetry of the problem: every time an option is executed, we stay m-close to s' and the probability to exactly reach s' is always $1/m$. So in this sense, we have i.i.d. Bernoulli trials where the probability of success is $1/m$. The expected time to reach s' when we start in an m-close state both horizontally and vertically is thus $2 \times m(m+1)/2 = m(m+1)$. Therefore, the expected time to go from s to s' is always bounded by $D + m(m+1)$.

On Fig. 9 we illustrate what happens when the options are deterministic i.e., when they terminate after exactly m time steps. On this example we chose $m = 3$. If we start from state s_0, only the green states can be reached without resorting to the restart triggered by the target state, whereas if we start from s_1 only the blue states can be reached [8] (and similarly for the white states). Let's assume that we want to go from a green state to a blue state. The only way to do so is to go to the target and "hope" to end up in one of the blue states after the restart (we recall that the restart state is chosen randomly with equi-probability). The shortest path to go from any state to the target is bounded by D and the probability to restart in a state with the desired colour is $1/m$ ($1/m^2$ in dimension 2). We can thus upper bound the diameter of the SMDP $M_{\mathcal{O}}$ by the expected time needed to go from s_0 to s_1 in the SMDP of Fig. 10, that is: $D_{\mathcal{O}} \leq D(1 + m^2)$. This bound is tight (up to a

---

[8]Here we slightly simplified the problem. In reality, due to the truncation operated by the walls and if we assume s_0 to be the leftmost state, it is possible to go from s_1 to s_0 in one time step by executing $LEFT$. But for $m \geq 3$ there will always be pairs of states that cannot be reached from each other without a restart. If s_0 is the leftmost state, this is the case for s_1 and the white state adjacent to it on Fig. 9. So the proof remains valid.
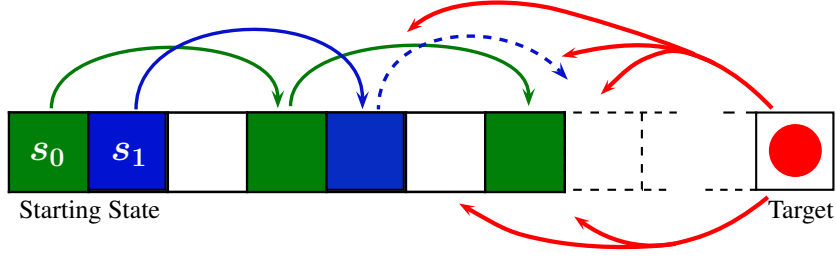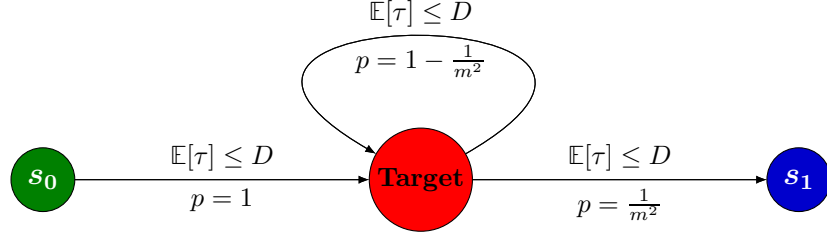
Figure 9: Behaviour of the agent with deterministic options.

Figure 10: Upper bound on the diameter of $M_{\mathcal{O}}$ for deterministic options.

constant factor) since the average time to go from any state chosen at random with equi-probability to the target is exactly $D/2$ in the 2-dimensional grid.

Optimality Since the target state is located in a corner of the grid, the shortest path to go from any state to the target is equally long in the original MDP and the MDP with options. As a result, the optimal average rewards are also equal (i.e., there exists an optimal policy using only options *LEFT RIGHT*, *UP* and *DOWN* which consists in applying only *RIGHT* or *DOWN*).

Asymptotic behaviour We will now analyse the behaviour of the ratio $\frac{n}{T_n}$ using results on martingales.

Theorem 3 (Martingale Strong Law of Large Numbers, [Vovk et al., 2005]). *Let $X_1, ..., X_n$ be a martingale difference sequence w.r.t. a filtration $\mathcal{F}_0, \mathcal{F}_1, ..., \mathcal{F}_n$ and let $A_1, ..., A_n$ be an increasing predictable sequence w.r.t. the same filtration with $A_1 > 0$ and $\lim_{n \to +\infty} A_n = +\infty$ almost surely. If:*

$$\sum_{i=1}^{+\infty} \frac{\mathbb{E}[X_i^2 | \mathcal{F}_{i-1}]}{A_i^2} < +\infty \quad a.s.$$

then:

$$\frac{1}{A_n} \sum_{i=1}^{n} X_i \xrightarrow[n \to +\infty]{} 0 \quad a.s.$$

Let's take $X_i = \tau_i - \overline{\tau}_i$ (where $\overline{\tau}_i = \overline{\tau}_i(s_{i-1}, a_i)$) and $\mathcal{F}_i = \sigma(s_0, a_1, \tau_1, r_1, ..., s_i, a_{i+1})$. The sequence $(X_i)_{i \leq 1}$ is a martingale difference because $\mathbb{E}[X_i] < +\infty$ and $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0$. Since $(\tau(s, a, s'))_{s,a,s'}$ are sub-Exponential, all moments are finite and it is well known from the literature that the variance is bounded by the sub-Exponential constant σ_τ^2 hence: $\mathbb{E}[X_i^2 | \mathcal{F}_{i-1}] < \sigma_\tau^2$. If in addition we take $A_i = i$ then the conditions of Theorem 3 are satisfied and thus:

$$\frac{T_n}{n} - \frac{\overline{T_n}}{n} \xrightarrow[n \to +\infty]{} 0 \quad a.s.$$

where $\overline{T_n} = \mathbb{E}[T_n]$. By definition: $\tau_{\max} n \geq \overline{T_n} \geq \tau_{\min} n$ hence: $\forall \epsilon > 0, \ \exists N_\epsilon > 0$ s.t. $\forall n \geq N_\epsilon$:

$$\left| \frac{T_n}{n} - \frac{\overline{T_n}}{n} \right| \leq \epsilon \ \text{a.s.} \implies \tau_{\min} - \epsilon \leq \frac{\overline{T_n}}{n} - \epsilon \leq \frac{T_n}{n} \leq \epsilon + \frac{\overline{T_n}}{n} \leq \epsilon + \tau_{\max}$$

and so: $\liminf_{n\to+\infty} \frac{T_n}{n} \geq \tau_{\min}$ and $\limsup_{n\to+\infty} \frac{T_n}{n} \leq \tau_{\max}$ a.s. Finally:

$$\frac{\log\{\frac{n}{\delta}\}}{\log\{\frac{T_n}{\delta}\}} = \frac{\log\{\frac{n}{\delta}\}}{\log\{\frac{T_n-\overline{T_n}}{n} + \frac{\overline{T_n}}{n}\} + \log\{\frac{n}{\delta}\}} \leq \frac{\log\{\frac{n}{\delta}\}}{\log\{\frac{T_n-\overline{T_n}}{n} + \tau_{\min}\} + \log\{\frac{n}{\delta}\}} \xrightarrow[n\to+\infty]{} 1$$

In the general case of sub-Exponential rewards and holding times our results provide no theoretical evidence of the advantage of introducing options due to the fact that $\mathcal{C}(M', n, \delta)$ scales as $\sqrt{\log(n)}$:

$$\lim_{n\to+\infty} \mathcal{R}(M, n, \delta) = +\infty \quad \text{a.s.}$$

but if the rewards and holding times are bounded we have:

$$\limsup_{n\to+\infty} \mathcal{R}(M, n, \delta) \leq \frac{1}{\sqrt{\tau_{\min}}} \left(1 + \frac{T_{\max}}{D\sqrt{S}}\right) \quad \text{a.s.}$$

Note that $\tau_{\min}$ 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 higher values if $\tau_{\max}$ is big and many options have a high expected holding time.

Tightness of the upper bounds On Fig. 11a, we plot the expected theoretical values taken by the ratio of the regrets according to our upper bounds (formula given in Sect. 6). On Fig. 11b however, we plot the empirical values of the ratios in our experiments (same graph as on Fig. 3a). We can see that the curves have similar shapes. In particular, they reach their respective minima for the same value of $T_{\max}$ and the value of these minimum is below 1 (meaning that learning with options is more efficient than with primitive actions in this case). Moreover, the theoretical ratios are upper-bounding the empirical ones for all values of $T_{\max}$. We can conclude that the ratio of the upper bounds is a good proxy for the true ratio in this example.
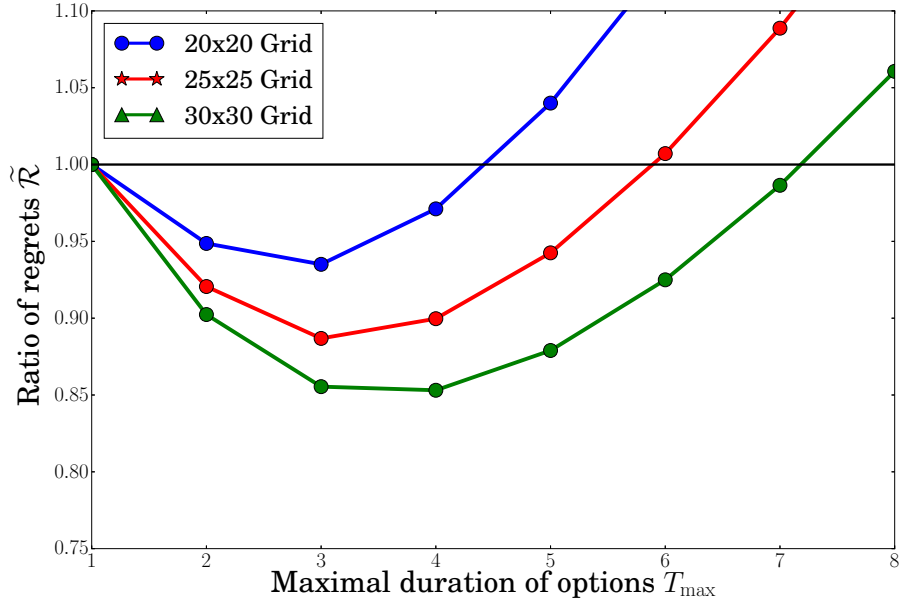
(a)
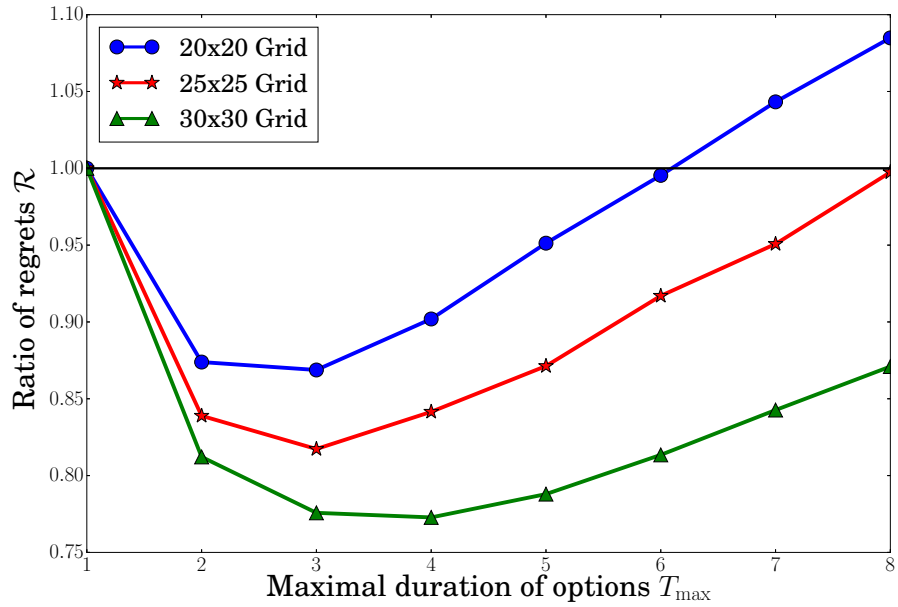
(b)

Figure 11: *(a)* Theoretical ratios of the regrets with and without options for different values of $T_{\max}$; *(b)* Empirical ratios of the regrets with and without options for different values of $T_{\max}$.