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, Simsek 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).
Exploration–Exploitation in MDPs with Options

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 \(\pi : S \rightarrow A\) maps states to actions. A (Markov) option is a tuple
\[
o = \{I_o, \beta_o, \pi_o\}
\]
where \(I_o \subseteq S\) is the set of states where the option can be initiated, \(\beta_o : S \rightarrow [0, 1]\) is the probability distribution that the option ends in a given state, and \(\pi_o : S \rightarrow A\) 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,
\[
S_O = \left( \bigcup_{o \in O} I_o \right) \bigcup \left( \bigcup_{o \in O} \{s : \beta_o(s) > 0\} \right).
\]

\(O\) is the set of available actions, \(p_O(s, o, s')\) is the probability of transition from \(s\) to \(s'\) by taking the policy \(\pi_o\) 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 \(\pi_o\), \(r_O(s, o, s')\) 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 \(\pi_o\)).

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., \(\bigcup_{o \in O} \{s : \beta_o(s) > 0\} \subseteq \bigcup_{o \in O} I_o\). 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 \(\pi(s, a, s')\) (resp. \(\pi(s, a, s')\)) the expectation of \(\tau(s, a, s')\) (resp. \(r(s, a, s')\)) and by \(\pi(s, a) = \sum_{s' \in S} \pi(s, a, s')p(s'|s, a)\) (resp. \(\pi(s, a) = \sum_{s' \in S} \pi(s, a, s')p(s'|s, a)\)) 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 \(\pi\) and \(s \in S\):
\[
\begin{align*}
p^\pi(s) &\overset{def}{=} \limsup_{t \to +\infty} \mathbb{E}_{\pi} \left[ \frac{\sum_{i=1}^{N(t)} r_i}{t} \right| s_0 = s \\
p^\pi(s) &\overset{def}{=} \liminf_{t \to +\infty} \mathbb{E}_{\pi} \left[ \frac{\sum_{i=1}^{N(t)} r_i}{t} \right| s_0 = s.
\end{align*}
\]

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

Finally, we recall the average reward optimality equation for a communicating SMDP
\[
u^*(s) = \max_{a \in A} \left\{ \pi(s, a) - \rho^*\pi(s, a) \right\} + \sum_{s' \in S} p(s'|s, a)\nu^*(s'),
\]
where \(u^*\) and \(\rho^*\) are the bias (up to a constant) and the gain of the optimal policy \(\pi^*\).

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 and \(s_i\) denotes the state reached after \(a_i\) is taken, with \(s_0\) being the initial state. We denote by \((\tau_i(s, a, s'))_{i \in \mathbb{N}^+}\) a sequence of i.i.d. realizations from distribution \(\tau(s, a, s')\) (resp. \(\tau(s, a, s')\)) when the learner explores the SMDP, it observes the sequence \((s_0, s_1, a_1, s_{i+1}, \tau_i(s, a, s'), s_{i+1}, \tau_{i+1}(s_{i+1}, a_{i+1}, s_{i+1}), s_{i+1}, a_{i+1}, s_{i+1}, \ldots)\). 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 \(\mathcal{A}\). Then the total regret of \(\mathcal{A}\) is defined as
\[
\Delta(M, \mathcal{A}, s, n) = \left( \sum_{i=1}^{n} \tau_i \right) \rho^*(M) - \sum_{i=1}^{n} \tau_i.
\]
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 $\rho^*$, the total duration after $n$ decision steps may vary depending on the policy. As a result, when comparing the performance of $\pi^*$ 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 $\rho^*$ (Eq. 1) and Prop. 2 we have

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

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 $(t - \sum_{i=1}^{N(t)} \tau_i) = o(t)$ almost surely since $(t - \sum_{i=1}^{N(t)} \tau_i) \leq \tau_{N(t)+1}$ 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 $\tau_{N(t)+1}/t \to 0$ a.s. and

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

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 $\tilde{M}_k$ in $\mathcal{M}_k$ that maximizes $\rho^*(\tilde{M}_k)$ and the corresponding optimal policy $\tilde{\pi}_k^*$ is computed. To solve this problem, we note that it can be equivalently formulated as finding the optimal policy of an extended SMDP $\tilde{M}_k^+$ obtained by combining all SMDPs in $\mathcal{M}_k$: $\tilde{M}_k^+$ has the same state space and an extended continuous action space $\tilde{A}_k^+$. Choosing an action $a^+ \in \tilde{A}_k^+$ amounts to choosing an action

---

4In this expectation, $N(t)$ is a r.v. depending on $\pi^*$.

5In the MDP literature, the term Bound Parameter MDPs (BPMDPs) [Tewari and Bartlett, 2007] is often used for "extended" MDPs built using confidence intervals on rewards and transition probabilities.

---

Input: Confidence $\delta \in ]0, 1[$, $\mathcal{S}$, $\mathcal{A}$, $b_r, \sigma_r, R_{\text{max}}, \tau_{\text{max}}$ and $\tau_{\text{min}}$.

Initialization: Set $i = 1$, and observe initial state $s_0$.

For episodes $k = 1, 2, \ldots$ do

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 $k$,

   $$N_k(s, a) = \#\{t < i_k : s_t = s, a_t = a\}$$

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

   $$R_k(s, a) = \sum_{i=1}^{i_k-1} r_i \mathbb{1}_{s_i = s, a_i = a}, T_k(s, a) = \sum_{i=1}^{i_k-1} \tau_i \mathbb{1}_{s_i = s, a_i = a}$$

   $$P_k(s, a, s') = \#\{t < i_k : s_t = s, a_t = a, s_{t+1} = s'\}$$

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

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

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

   $$|\tilde{r} - r| \leq \beta_k^r$$

   and $R_{\text{max}}, \tau_{\text{max}} \geq \tilde{\tau}(s, a) \geq 0$

   $$|\tilde{r} - \tilde{r}_k| \leq \beta_k^r$$

   and $\tau_{\text{max}} \geq \tilde{\tau}(s, a) \geq \tilde{\tau}(s, a) / R_{\text{max}}, \tau_{\text{min}}$

   $$||\tilde{p}(\cdot) - \tilde{p}_k(\cdot)||_1 \leq \beta_k^p$$

   and $\sum_{s' \in \mathcal{S}} \tilde{p}(s' \mid s, a) = 1$

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

   $$\tilde{p}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s} \rho(M', \pi, s) - R_{\text{max}} / \sqrt{1/k}$$

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

6. While $\nu_k(s, \tilde{\pi}_k(s)) < \max\{1, N_k(s, \tilde{\pi}_k(s))\}$ do

   (a) Choose action $a_i = \tilde{\pi}_k(s_i)$, 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 $\tilde{r}(s, a)$, a holding time $\tilde{\tau}(s, a)$ and a transition probability $\tilde{p}(\cdot \mid s, a)$ and a transition probability $\tilde{p}_k(\cdot \mid s, a)$ in the confidence intervals. When $a^+$ is executed in $\tilde{M}_k^+$, the probability, the expected reward and the expected holding time of the transition are respectively $\tilde{p}_k(\cdot \mid s, a), \tilde{r}_k(s, a)$ and $\tilde{\tau}_k(s, a)$. Finally, $\tilde{\pi}_k$ 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 redefined 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 sub-exponential r.v. (see Lem. 3). As a result, the confidence intervals used at step 4 are defined as follows. For any state action pair \((s, a)\) and for rewards, transition probabilities, and holding times we define

\[
\beta_k^i(s, a) = \begin{cases} 
\sigma_{r} \sqrt{\frac{14 \log(2 S A t_k/d)}{\max\{1, N_k(s, a)\}}} & \text{if } N_k(s, a) \geq \frac{2k^2}{\sigma_{r}^2} 
\end{cases}
\]

\[
\beta_k^p(s, a) = \frac{14 S \log(2 A t_k/d)}{\max\{1, N_k(s, a)\}}
\]

\[
\beta_k^\rho(s, a) = \frac{\sigma_{r} \sqrt{14 \log(2 S A t_k/d)}}{14 b \max\{1, N_k(s, a)\}}
\]

where \(\sigma_{r}, b_{r}, \sigma_{p}, b_{p}\) are suitable constants characterizing the sub-exponential distributions of rewards and holding times. As a result, the empirical estimates \(\hat{\tau}_k, \hat{\tau}_j, \hat{p}_k\) are \(\pm \beta_k^i(s, a), \beta_k^p(s, a), \beta_k^\rho(s, a)\) 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 \(M_{eq}\) with same state and action spaces and such that \(\forall s, s' \in S, \forall a \in A:\)

\[
\pi_{eq}(s, a) = \frac{\pi(s, a)}{\pi(s, a)}
\]

\[
p_{eq}(s'|s, a) = \frac{\pi}{\pi(s, a)} \left( p(s'|s, a) - \delta_{s, s'} \right) + \delta_{s, s'}
\]

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

**Proposition 3 (Federgruen et al., 1982, 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. \(\pi\) i.e., it is an optimal pair of bias/gain for the original SMDP \(M\).

As a consequence of the equivalence stated in Prop. \(\pi\) computing the optimal policy of an SMDP amounts to computing the optimal policy of the MDP obtained after data transformation (see App. \(\pi\) for more details). Thus, EVI is obtained by applying a value iteration scheme to an MDP \(M_{k, eq}\) equivalent to the extended SMDP \(\tilde{M}_{k, eq}\). We denote the state values of the \(j\)-th iteration by \(u_j(s)\). We also use the vector notation \(u_j = (u_j(s))_{s \in S}\). Similarly, we denote by \(\tilde{\pi}(\cdot | s, a) = (\tilde{\pi}(s' | s, a))_{s' \in S} \) the transition probability vector of state-action pair \((s, a)\). The optimistic reward at episode \(k\) is fixed through the EVI iterations and is obtained as \(\tilde{\tau}_{j+1}(s, a) = \min \{ \tilde{\tau}_k(s, a) + \beta_k^i(s, a), R_{max} \tau_{max} \} \), i.e., by taking the largest possible value compatible with the confidence intervals. At iteration \(j\), the optimistic transition model is obtained as \(\tilde{\tau}_{j+1}(\cdot | s, a) \in \text{Argmax}_{\pi \in \mathcal{P}(S, A)} \{ p u_j \} \) and \(\mathcal{P}(s, a)\) is the set of probability distributions included in the confidence interval defined by \(\beta_k^i(s, a, \delta)\). 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 \(\tilde{\tau}_{j+1}\) as

\[
\tilde{\tau}_{j+1}(s, a) = \min \{ \tau_{max} \text{; } max \{ \tau_{min} \text{; } \tilde{\tau}_k(s, a) \} \text{; } \tilde{\tau}_{j+1}(\cdot | s, a) \text{; } \beta_k^i(s, a) \}.
\]

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

\[
u_{j+1}(s) = \max_{a \in A} \left\{ \frac{\tilde{\tau}_{j+1}(s, a)}{\tilde{\tau}_{j+1}(s, a)} \right\} + u_j(s)
\]

with arbitrary \(u_0\). Finally, the stopping condition is

\[
\max_{s \in S} \{ u_{j+1}(s) - u_j(s) \} - \min_{s \in S} \{ u_{j+1}(s) - u_j(s) \} < \epsilon.
\]

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 \(\epsilon\)-optimal w.r.t. extended SMDP \(\tilde{M}_{k}^+\). 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 \(\tilde{\tau}_k = R_{max}/\sqrt{i_k}\)-optimal w.r.t. max \(M_{eq} \in M_k \rho(M')\).

### 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

\[
D(M) = \max_{s, s' \in S} \min_{\pi} \left\{ \mathbb{E}_{\pi} [T(s'|s_0 = s)] \right\}
\]

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

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 \(s \in S\) and \(a \in A\), we assume that \(\tau_{\text{max}} \geq \tau_{\text{min}} \geq 0\) and
\[
\max_{s,a,s'} \frac{\tau(s,a)}{\tau(s,a')} \leq \frac{\tau_{\text{max}}}{\tau_{\text{min}}} \quad \text{and} \quad \tau_{\text{max}} \text{ known to the learning algorithm. Furthermore,}
\]
we assume that the random variables \((r(s,a,s'))_{s,a,a'}\) and \((\tau(s,a,s'))_{s,a,a'}\) are either 1) sub-Exponential
with constants \((\sigma_r, b_r)\) and \((\sigma_\tau, b_\tau)\), or 2) bounded in \([0, R_{\text{max}} T_{\text{max}}]\) and \([T_{\text{min}}, T_{\text{max}}]\), with \(T_{\text{min}} > 0\). 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 \(s \in S\) and any \(n > 1\), the regret of UCRL-SMDP \(\Delta(M, \mathfrak{A}, s, n)\) is bounded by:

\[
O\left(\left(D \sqrt{S} + C(M, n, \delta)\right) R_{\text{max}} \sqrt{SA n \log \left(\frac{n}{\delta}\right)}\right), \quad (8)
\]

where \(C(M, n, \delta)\) depends on which case of Asm. 1 is considered.

1) sub-Exponential
\[
C(M, n, \delta) = \tau_{\text{max}} + \left(\frac{\sigma_r \vee b_r}{R_{\text{max}}} + \sigma_\tau \vee b_\tau\right) \sqrt{\log \left(\frac{n}{\delta}\right)},
\]

2) bounded
\[
C(M, n, \delta) = T_{\text{max}} + (T_{\text{max}} - T_{\text{min}}).
\]

**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. 4), 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
\(D(M_{eq}) = D(M)/\tau\) (Lem. 4 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
\(R_{\text{max}} = 1\) and bounded holding times, the regret scales as \(O(D \sqrt{An} + T_{\text{max}}\sqrt{SA n})\). The most interesting aspect of this bound is that the extra cost of having
\(\Delta(M, \mathfrak{A}, s, n)\) is \(\Omega(\sqrt{D} + \sqrt{T_{\text{max}}} R_{\text{max}}\sqrt{SA n})\).

**Proof.** Similar to the upper bound, the proof (App. C) is based on Jaksch et al. 2010 but it requires to perturb
the regret smaller than 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 \(\sqrt{T_{\text{max}}}\) in the lower bound.

The bound. Similar to UCRL, this lower bound reveals a gap of \(\sqrt{DS}\) on the first term and \(\sqrt{T_{\text{max}}}\). 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 \(O\) a set of options and let \(M_O\) 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 \(\mathfrak{A}_O\) applied to \(M_O\) 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, $O$ a set of options and $M_O$ the corresponding SMDP. Let $\pi_O$ be any stationary policy on $M_O$ and $\pi$ the equivalent policy on $M$ (not necessarily stationary). For any state $s \in S_O$, any learning algorithm $\mathfrak{A}$, and any number of decision steps $n$ we have $\rho^\pi(M_O, s) = \rho^\pi(M, s)$ and

$$\Delta(M, \mathfrak{A}, T_n) = \Delta(M_O, \mathfrak{A}, n) + T_n (\rho^\pi(M) - \rho^\pi(M_O)).$$

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 $\rho^\pi(M) \geq \rho^\pi(M_O) = \max_{\pi_O} \rho^\pi(M_O)$, where $\pi_O$ is a stationary deterministic policy on $M_O$. Thm. 2 also guarantees that the optimal policy computed in the SMDP $M_O$ (i.e., the policy maximizing $\rho^\pi(M_O, s)$) is indeed the best in the subset of policies that can be expressed in $M$ by using the set of options $O$. In order to use the regret analysis of Thm. 1 we still need to show that Asm. 1 is verified.

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

This result is based on the fact that once an option is executed, we obtain a Markov chain with absorbing states corresponding to the states with non-zero termination probability $\beta_\pi(s)$ and the holding time is the number of visited states before reaching a terminal state. While in general this corresponds to a sub-exponential distribution, whenever the option has a zero probability to reach the same state twice before terminating (i.e., there is no cycle), then the holding times become bounded. Finally, we notice that no intermediate case between sub-exponential and bounded distributions is admissible (e.g., sub-Gaussian). Since these are the two cases considered in Thm. 4 we can directly apply it and obtain the following corollary.

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

$$O \left( D_O \sqrt{S_O} + C(M_O, n, \delta) R_{\max}^O \sqrt{S_O} \log \left( \frac{n}{\delta} \right) \right) + T_n (\rho^*(M) - \rho^*(M_O)),$$

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

$$\Omega \left( \left( \sqrt{D_O} + \sqrt{T_{\max} - T_{\min}} \right) R_{\max}^O \sqrt{S_O} \log(n) \right) + T_n (\rho^*(M) - \rho^*(M_O)).$$

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 $\Delta(M, \mathcal{D}, s, T_n) = O(DS \log \sqrt{T_{n}/D})$. We first notice that $O \leq R_{\max} \leq R_{\max}$ and since $S_O \subseteq S$ we have that $S_O \leq S$. Furthermore, we introduce the following simplifying conditions: 1) $\rho^*(M) = \rho^*(M_O)$ (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 $R(M, n, \delta)$ be the ratio between the regret upper bounds of UCRL-SMDP using options $O$ and UCRL, then we have (up to numerical constants)

$$R(M, n) \leq \frac{D_O \sqrt{S_O} + T_{\max}}{DS \sqrt{T_{\min} \log(T_{n}/\delta)}} \leq \frac{D_O \sqrt{S} + T_{\max}}{DS \sqrt{T_{n}}},$$

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

$$\frac{D_O \sqrt{S} + T_{\max}}{DS \sqrt{T_{\min}}} \leq 1.$$ (9)

In order to have a better grasp on the cases covered by this condition, let $D_O = \alpha D$, with $\alpha \geq 1$. This corresponds to the case when navigating through some $\text{The largest per-step reward in the SMDP is defined as } R_{\max}^O = \max_{s \in S, a \in A} \{ \frac{r(s, o)}{\tau(s, o)} \}.$
states becomes more difficult with options than with primitive actions, thus causing an increase in the diameter. If options are such that \( T_{\text{max}} \leq D\sqrt{S} \) and \( \tau_{\text{min}} > (1 + \alpha)^2 \), then it is easy to see that the condition in Eq. 3 is satisfied. This shows that even when the introduction of options partially disrupt the structure of the original MDP (i.e., \( D_{\text{O}} \geq D \)), 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, \( S_{\text{O}}, Q_{\text{O}} \), and \( T_{\text{max}} \) are in general much smaller than \( S, A, D, \sqrt{S} \) (\( S \) and \( D \) are large in most of interesting applications). Furthermore, \( \tau_{\text{min}} \) is a very loose upper-bound on \( \lim_{n \to +\infty} \frac{T_n}{n} \) and in practice the ratio \( \frac{T_n}{n} \) can take much larger values if \( \tau_{\text{max}} \) 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. 3. 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

\[
\mathcal{R}(M, n) \leq \frac{D\sqrt{S} + C(M, n, \delta)}{D\sqrt{S}} \sqrt{\frac{n}{T_n}},
\]

where \( C(M, n, \delta) = O(\sqrt{\log(n/\delta)}) \). 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 \( R_{\text{max}} \). 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: \text{LEFT}, \text{RIGHT}, \text{UP} and \text{DOWN} (blue arrows in the figure). When \text{LEFT} is taken, primitive action \text{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, \( T_{\text{max}} = m \) and the expected duration is \( \tau := \tau(s, o) = (m+1)/2 \), which reduces to \( T_{\text{max}} = l \) and \( T = (l+1)/2 \) 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 \( S_{\text{O}} = S = d^2 \) and \( A' = A = 4 \) and the optimal average reward \( \rho^*(M) = \rho^*(M') \), while it slightly perturbs the diameter \( D_{\text{O}} \leq D + m(m+1) \) (see App. 4 for further details). Thus, the two problems seem to be as hard to learn. However the (asymptotic) ratio between the regret upper bounds becomes

\[
\lim_{n \to +\infty} \mathcal{R}(M, n) \leq \frac{(2d - 2 + m^2 + m)d + m}{(2d - 2)d} \left( \lim_{n \to +\infty} \sqrt{n} \right) \frac{n}{T_n} \leq \left( 1 + \frac{2n^2}{d} \right) \left( \lim_{n \to +\infty} \sqrt{n} \right) \frac{n}{T_n},
\]

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 \( \tau \) 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 \( (1+2n^2/d)\sqrt{2/(m+1)} \) 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. 6a we empirically validate...
Exploration–Exploitation in MDPs with Options

this finding by studying the ratio between the actual regrets (and not their upper-bounds) as $d$ and $m$ (i.e., $T_{\text{max}}$) 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. 3\textbf{b} we report the cumulative regret and we observe that high values of $T_{\text{max}}$ worsen the learning performances w.r.t. learning without options ($T_{\text{max}} = 1$, plotted in black). Finally, Fig. 3\textbf{c} 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.

\textbf{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 $\pi^*$ 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. \textbf{9} (see App. \textbf{F}).

\section{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. \textbf{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


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 \( \tilde{M}_{k}^{+} \) 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 \cite{Puterman1994}.

A decision rule is a function \( d : \mathcal{H} \to \Delta(A) \) where \( \mathcal{H} \) is the set of possible histories and \( \Delta(A) \) is the set of probability distributions over \( A \). 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 \( \Pi_{M}^{HR} \) and the subset of stationary deterministic policies will be denoted \( \Pi_{M}^{SD} : \Pi_{M}^{SD} \subset \Pi_{M}^{HR} \). We also consider the more general case where the set of available actions may depend on the state, i.e., there exists a set \( A_{s} \) 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 \cite{Schäl1992, Federgruen1983, Ros1970}).

Definition 3. Denote \( N(t) = \sup \{ n : n \in \mathbb{N}, \sum_{i=1}^{n} \tau_{i} \leq t \} \) the number of decision steps that occurred before time \( t \). For any \( \pi \in \Pi_{M}^{HR} \) and \( s \in \mathcal{S} \), we define:

\[
\overline{\rho_{1}}^{\pi}(s) = \limsup_{t \to +\infty} \mathbb{E}^{\pi} \left[ \frac{\sum_{i=1}^{N(t)} r_{i}}{t} \middle| s_{0} = s \right], \quad \underline{\rho_{1}}^{\pi}(s) = \liminf_{t \to +\infty} \mathbb{E}^{\pi} \left[ \frac{\sum_{i=1}^{N(t)} r_{i}}{t} \middle| s_{0} = s \right]
\]

and for any \( d^{\infty} \in \Pi_{M}^{SD} \) and \( s \in \mathcal{S} \) we define:

\[
\rho_{3}^{d^{\infty}}(s) = \sum_{\alpha=1}^{\nu(d)} p^{d}(\alpha|s) g_{d}(\alpha), \quad \text{with} \quad g_{d}(\alpha) = \frac{\sum_{s \in R_{d}^{\alpha}} \mu_{d}^{\pi}(s, \pi_{d}(s))}{\sum_{s \in R_{d}^{\alpha}} \mu_{d}^{\pi}(s, d(s))}
\]

where \( \nu(d) \) is the number of positive recurrent classes under \( d^{\infty} \), \( p^{d}(\alpha|s) \) is the probability of entering positive recurrent class \( \alpha \) starting from \( s \) and following policy \( d^{\infty} \), \( R_{d}^{\alpha} \) is the set of states of class \( \alpha \) and \( \mu_{d}^{\pi} \) is the stationary probability distribution of class \( \alpha \).

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 \( d \in D_{M}^{MD} \), \( \rho_{3}^{(d^{*})^{\infty}}(s) \geq \rho_{3}^{d^{\infty}}(s) \). We say that \( \pi^{*} \) is (1)-average-optimal (respectively (2)-average-optimal) if for all states \( s \in \mathcal{S} \) and all \( \pi \in \Pi_{M}^{HR} \), \( \rho_{1}^{\pi^{*}}(s) \geq \overline{\rho_{1}}^{\pi}(s) \) (respectively \( \rho_{2}^{\pi^{*}}(s) \geq \underline{\rho_{2}}^{\pi}(s) \)). 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 \( \rho^{*} \):

\[
\forall s \in \mathcal{S}, \quad \overline{\rho_{1}}^{(d^{*})^{\infty}}(s) = \rho_{1}^{(d^{*})^{\infty}}(s) = \underline{\rho_{2}}^{(d^{*})^{\infty}}(s) = \rho_{2}^{(d^{*})^{\infty}}(s) = \rho_{3}^{(d^{*})^{\infty}}(s) = \rho^{*}.
\]

Notice that the definition we provide in Eq. 11 of Prop. 2 is \( \rho_{1} \).
Proof. **Step 1: Optimality equation of a communicating SMDP.** We first recall the average reward optimality equations for a communicating SMDP (Eq. 2)

\[ \forall s \in S, \ u^*(s) = \max_{a \in A_s} \left\{ \tau(s, a) - \rho^* \tau(s, a) + \sum_{s' \in S} p(s'|s, a)u^*(s') \right\} \]  

(13)

where \( u^* \) and \( \rho^* \) are the bias (up to additive constant) and average reward respectively. Since we need to analyse both the case where \( A_\pi \) is finite and the case where \( A_\pi \) is continuous, we verify that it is appropriate to consider the max instead of sup in the previous expression. For the original SMDP \( M, A_\pi \) is finite and the maximum is well-defined. For the extended SMDPs \( \tilde{M}_k^* \) considered while computing the optimistic SMDP, \( A_{k,s}^+ \) is continuous and \( \tau(s, a), \tau(s, a) \) and \( p(\cdot | s, a) \) are continuous in \( A_{k,s}^+ \) by the very definition of \( \tilde{M}_k^* \). The function \( \tau(s, a) - \rho^* \tau(s, a) + \sum_{s' \in S} p(s'|s, a)u^*(s') \) is thus continuous on \( A_{k,s}^+ \) 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 into an "equivalent" MDP \( M_{eq} = \{ S, A, p_{eq}, r_{eq} \} \) defined as in Eq. 2. As a result, we can just apply standard MDP theory to the equivalent MDP. The average optimality equation of \( M_{eq} \) is \([\text{Puterman, 1994}]\)

\[ \forall s \in S, \ v^*(s) = \max_{a \in A_s} \left\{ \tau(s, a) g^* + \frac{\tau}{\tau(s, a)} \sum_{s' \in S} p(s'|s, a)v^*(s') + \left( 1 - \frac{\tau}{\tau(s, a)} \right)v^*(s) \right\} \]  

(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 \( \tilde{M}_k^* \) (i.e., the MDP obtained from the extended SMDP \( \tilde{M}_k^* \) after data transformation). Under these conditions, Eq. 13 has a solution \((v^*, g^*)\) where \( g^* \) is the optimal average reward of \( M_{eq} \) (respectively \( \tilde{M}_k^* \)) 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 \( \epsilon \)-optimal policy in finitely many steps. This holds for both finite and compact \( A_\pi \) with continuous \( \tau_{eq}(s, a) \) and \( p_{eq}(s'|s, a) \) (see for example [Puterman, 1994] and [Leizarowitz, 2013]). It is easy to show that EVI in Eq. 5 is exactly value iteration applied to \( \tilde{M}_k^{eq} \). Finally, Lemma 2 of [Federgruen et al., 1983] (Prop. 3) shows the "equivalence" between \( M \) and \( M_{eq} \) (respectively \( \tilde{M}_k^* \) and \( \tilde{M}_k^{eq} \)): if \((v^*, g^*)\) is a solution to Eq. 13 then \((\tau^{-1}v^*, g^*)\) is a solution to Eq. 13 and conversely. As a result, there exists a solution \((u^*, \rho^*)\) to Eq. 13 for both \( M \) and \( \tilde{M}_k^* \).

**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 \( A_\pi \) separately.

**Step 3a: For \( M \) (finite \( A_\pi \)).** 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 \( \overline{\pi_1}(d^*)^\infty(s) \geq \overline{\pi_1}(s) \) for any \( \pi \in \Pi^R_M \) and any \( s \in S \),
2. \( \forall s \in S, \overline{\pi_1}(d^*)^\infty(s) = \rho^* \),

where \((u^*, \rho^*)\) is a solution of Eq. 13. Furthermore, from renewal theory (see e.g., [Tij, 2002] and [Ros, 1970]) we have that \( \forall d^\infty \in \Pi^D_M, \overline{\pi_1}(d^\infty) = \rho_1^\infty = \rho_1^{eq} \) (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]: \( \forall d^\infty \in \Pi^D_M, \forall s \in S, \overline{\pi_2}(d^\infty(s)) = \rho_2^{eq}(s) \) so \( d^* \) is also necessarily (3)-optimal. Finally, by Theorem 7.6 of [Ros, 1970], since \((u^*, \rho^*)\) is a solution of Eq. 13

1. Any greedy policy \((d^*)^\infty\) w.r.t. \( u^* \) is such that \( \overline{\pi_2}(d^*)^\infty(s) \geq \overline{\pi_2}(s) \) for any \( \pi \in \Pi^R_M \) and any \( s \in S \),
2. \( \forall s \in S, \overline{\pi_2}(d^*)^\infty(s) = \rho^* \).
By Theorem 11.4.1 of [Puterman, 1994] we have \( \forall d^\infty \in \Pi_S^{SD}, \ \bar{\rho}^2 = \rho_2^2 = \rho_3^2 \) 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 \( \tilde{M}^+_k \) (compact \( \tilde{A}_{k,s} \) 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 \((u^*, \rho^*)\), 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 \( \tilde{M}^+_k \).

**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 \( \tilde{M}^+_k \), which is also the optimal average reward of \( \tilde{M}^+_k \). 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 \( \epsilon \)-optimal in the equivalent \( \tilde{M}^+_k \). Then, in order to prove Lemma 1 we only need to prove that this policy is also \( \epsilon \)-optimal in the optimistic SMDP \( \tilde{M}^+_k \). [Tu, 2003] shows that for any stationary deterministic policy \( d^\infty \in \Pi_S^{SD} \), the (1)-average reward is the same in the SMDP and the MDP obtained by uniformization, that is

\[
\forall s \in S, \rho_1^d(s) = \rho_2^d(s), \quad (15)
\]

where \( \rho_1^d \) is the optimal gain of \( \tilde{M}^+_k \) and \( \rho_2^d \) is the optimal gain of \( \tilde{M}^+_k \) in the equivalent Markov chain in \( \tilde{M}^+_k \) (or equivalently in \( \tilde{M}^+_k \)), we have: \( \forall s \in S, \rho_1^d(s) = \rho_2^d(s) \) and this value is constant across states (see for example chapter 11 of [Puterman, 1994], Theorem 1.5 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

\[
\forall d^\infty \in \Pi_S^{SD}, \forall s \in S, \ |\rho_2^d(\tilde{M}^+_k) - \rho_2^d(\tilde{M}^+_k, s)| \leq \rho_2^d(\tilde{M}^+_k), s) \text{ and } \rho_2^d(\tilde{M}^+_k, s) \leq \rho_2^d(\tilde{M}^+_k, s)
\]

If we denote by \( d \) the policy returned by EVI and \( \rho^* \) the optimal gain of \( \tilde{M}^+_k \) and \( \tilde{M}^+_k \) we obtain

\[
\forall s \in S, \rho^* - \rho_2^d(\tilde{M}^+_k, s) = \rho_2^d(\tilde{M}^+_k, s) - \rho_2^d(\tilde{M}^+_k, s) + \rho^* - \rho_2^d(\tilde{M}^+_k, s)
\]

For the first inequality we used Eq. (15) and the fact that \( d \) is \( \epsilon \)-optimal in \( \tilde{M}^+_k \). For the second inequality, we used again that \( d \) is \( \epsilon \)-optimal in \( \tilde{M}^+_k \) and we also used the fact that \( \rho_2^d(\tilde{M}^+_k, s) \leq \rho^* \). In conclusion, the policy returned by EVI is (2)-2\( \epsilon \)-optimal. The remaining part of Theorem 1 is thus proved for all optimality criteria.

By Theorem 8.3.2 of [Puterman, 1994], we know that there exists an optimal policy \( \tilde{d}^* \) of MDP \( \tilde{M}^+_k \), that yields a unichain Markov Chain (i.e., a Markov Chain with a single positive recurrent class). The Markov Chain induced by \( \tilde{d}^* \) in \( \tilde{M}^+_k \) is thus also unichain and moreover: \( \rho_1^{(\tilde{d})^\infty}(\tilde{M}^+_k) = \rho_2^{(\tilde{d})^\infty}(\tilde{M}^+_k, s) = \rho_2^{(\tilde{d})^\infty}(\tilde{M}^+_k) = \rho^*(\tilde{M}^+_k) \). We have seen that for any policy \( d \in \Pi_S^{SD} \) yielding a unichain Markov Chain \( \rho_1^d(\tilde{M}^+_k) = \rho_2^d(\tilde{M}^+_k) \) and so in particular, it is true for \( \tilde{d}^* \). Therefore, there exists a policy of \( \tilde{M}^+_k \) 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 \( \epsilon \)-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:

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

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

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. $\forall i \in \{1, ..., n\}$, $X_i \in \text{subExp}(\sigma_i, b_i)$ and $\mathbb{E}[X_i] = \mu_i$. We have the following concentration inequalities:

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

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

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

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

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

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

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

$$P \left( \sum_{i=1}^{n} r_i \leq \sum_{s \in S} \sum_{a \in A_s} N(s, a) \mathbb{P}(s, a) - \frac{5}{2} b_r \log \left( \frac{13n}{\delta} \right) \right) \leq \left( \frac{\delta}{13n} \right)^{5/4} \leq \frac{\delta}{24 n^{5/4}},$$

if $n \leq \frac{5b^2}{2 \sigma_r^2} \log \left( \frac{13n}{\delta} \right)$
Similarly, the total holding time satisfies
\[
P \left( \sum_{i=1}^{n} \tau_i \geq \sum_{s \in S} \sum_{a \in A_s} N(s,a) \tau(s,a) + \sigma_\tau \sqrt{\frac{5}{2} \tau n \log \left( \frac{13n}{\delta} \right)} \right) \leq \left( \frac{\delta}{13n} \right)^{5/4} \leq \frac{\delta}{24n^{5/4}},
\]
if \( n \geq \frac{5b^2}{2\sigma^2} \log \left( \frac{13n}{\delta} \right) \)
\[
P \left( \sum_{i=1}^{n} \tau_i \geq \sum_{s \in S} \sum_{a \in A_s} N(s,a) \tau(s,a) + \frac{5}{2} b_\tau \log \left( \frac{13n}{\delta} \right) \right) \leq \left( \frac{\delta}{13n} \right)^{5/4} \leq \frac{\delta}{24n^{5/4}},
\]
if \( n \leq \frac{5b^2}{2\sigma^2} \log \left( \frac{13n}{\delta} \right) \)

**Lemma 4.** The optimal average reward can be bounded as follows:
\[
\rho^*(M) \leq \max_{s \in S, a \in A_s} \left\{ \frac{\tau(s,a)}{\tau(s,a)} \right\} \leq R_{\max}.
\]

*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 \( M_{\text{eq}} \) with same state and action spaces as SMDP \( M \) and with average rewards of the form \( \tau(s,a) \). All the rewards of \( M_{\text{eq}} \) are thus bounded by \( R_{\max} \) and so \( \rho^*(M_{\text{eq}}) \) is necessarily bounded by \( R_{\max} \) as well and thus: \( \rho^*(M) \leq R_{\max} \).

We are now ready to split the regret over episodes. We define the per-episode regret as
\[
\Delta_k = \sum_{s \in S} \sum_{a \in A_s} \nu_k(s,a) (\tau(s,a) \rho^* - \tau(s,a)).
\]
Setting \( \gamma_r = \max \left\{ \frac{5}{2} b_\tau, \sqrt{\frac{5}{2} \sigma_\tau} \right\} \) 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 \( 1 - \frac{\delta}{12n^{\gamma_\tau}} \)
\[
\Delta(M, M, s, n) \leq \sum_{k=1}^{n} \Delta_k + (\gamma_r + \gamma_\tau R_{\max}) \log \left( \frac{13n}{\delta} \right) \sqrt{n}
\]

**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 \( M_k \) at step \( i \) is at most \( \frac{\delta}{15i_k^5} \), that is:
\[
\forall k \geq 1, \quad P( M \notin M_k ) < \frac{\delta}{15i_k^5}
\]

*Proof.* This lemma is the SMDP-analogue of Lemma 17 in [Jaksch et al., 2010] and the proof is similar. Using an \( \ell_1 \)-concentration inequality for discrete probability distributions we obtain
\[
P \left( \|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \geq \beta_k(s,a) \right) = P \left( \|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \geq \sqrt{\frac{14S}{n} \log \left( \frac{24\delta}{\delta} \right)} \right)
\]
\[
\leq P \left( \|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \geq \sqrt{\frac{2}{n} \log \left( \frac{2^S 20SA_i^2}{\delta} \right)} \right)
\]
\[
\leq 2^S \exp \left( -\frac{n}{2} \times \frac{2}{n} \log \left( \frac{2^S 20SA_i^2}{\delta} \right) \right)
\]
\[
= \frac{\delta}{20i_k^5SA}
\]
We now consider the regret of episodes in which the set of plausible SMDPs \( \mathcal{M} \) are bounding the probability of the intersection of event \( \{ \| \hat{p}(s, a) - \bar{p}_k(s, a) \| \geq \beta_k^p(s, a) \} \) 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. \( \text{[5]} \)) 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 \( n \geq \frac{2\delta^2}{\sigma_r^2} \log \left( \frac{240SAi_k}{\delta} \right) \):
  \[
  P \left( |\hat{r}(s, a) - \bar{r}(s, a)| \geq \beta_k^r(s, a) \right) = P \left( |\hat{r}(s, a) - \bar{r}(s, a)| \geq \sigma_r \sqrt{n \log \left( \frac{2SAi_k}{\delta} \right)} \right) 
  \leq P \left( |\hat{r}(s, a) - \bar{r}(s, a)| \geq \sigma_r \sqrt{n \log \left( \frac{240SAi_k^2}{\delta} \right)} \right) 
  \leq 2 \exp \left( \frac{-n^2}{2\sigma_r^2} \right) \times \frac{2^3 \sigma_r^2 \log \left( \frac{240SAi_k^2}{\delta} \right)}{2} 
  = \frac{\delta}{120i_k^2SA}
  
- If \( n < \frac{2\delta^2}{\sigma_r^2} \log \left( \frac{240SAi_k}{\delta} \right) \):
  \[
  P \left( |\hat{r}(s, a) - \bar{r}(s, a)| \geq \beta_k^r(s, a) \right) = P \left( |\hat{r}(s, a) - \bar{r}(s, a)| \geq \frac{2\beta_r}{n \log \left( \frac{2SAi_k}{\delta} \right)} \right) 
  \leq P \left( |\hat{r}(s, a) - \bar{r}(s, a)| \geq \frac{2\beta_r}{n \log \left( \frac{240SAi_k^2}{\delta} \right)} \right) 
  \leq 2 \exp \left( \frac{-n^2}{2\beta_r} \right) \times \frac{2 \beta_r \log \left( \frac{240SAi_k^2}{\delta} \right)}{2} 
  = \frac{\delta}{120i_k^2SA}
  
Similarly for holding times we have:

\[
P \left( |\tau(s, a) - \hat{\tau}(s, a)| \geq \beta_k^\tau(s, a) \right) \leq \frac{\delta}{120i_k^2SA}
\]

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:

\[
P \left( |\tau(s, a) - \hat{\tau}(s, a)| \geq \beta_k^\tau(s, a) \right) \leq \frac{\delta}{120i_k^2SA}
\]

\[
P \left( |\hat{\tau}(s, a) - \bar{\tau}(s, a)| \geq \beta_k^\tau(s, a) \right) \leq \frac{\delta}{120i_k^2SA}
\]

\[
P \left( |\tau(s, a) - \hat{\tau}(s, a)| \geq \beta_k^\tau(s, a) \right) \leq \frac{\delta}{20i_k^2SA}
\]

Summing over all state-action pairs: \( P(M \notin \mathcal{M}_k) < \frac{\delta}{150i_k} \). \( \Box \)

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

\[
\sum_{s \in S} \sum_{a \in A_s} \nu_k(s, a) \leq \sum_{s \in S} \sum_{a \in A_s} N_k(s, a) = i_k - 1
\] (19)
We can thus bound this part of the regret:

\[
\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \not\in M_k} \leq \sum_{k=1}^{m} \sum_{s \in S} \sum_{a \in A_s} \nu_k(s, a) \tau(s, a) \rho^* \mathbb{1}_{M \not\in M_k}
\]

\[
\leq \tau_{\max} \rho^* \sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{i=k, M \not\in M_k}
\]

\[
\leq \tau_{\max} \rho^* \left( \frac{n}{\gamma} + \sum_{i=\lceil n/\gamma \rceil + 1}^{n} \sum_{k=1}^{m} \mathbb{1}_{i=k, M \not\in M_k} \right)
\]

where we defined: \( \tau_{\max} = \max_{s,a} \tau(s,a) < +\infty \).

By Lemma \ref{lem:non_zero_second_term} the probability that the second term in the right hand side of the above inequality is non-zero is bounded by

\[
\sum_{i=\lceil n/\gamma \rceil}^{n} \frac{\delta}{15n^6} \leq \frac{\delta}{15n^{6/4}} + \int_{\lceil n/\gamma \rceil}^{+\infty} \frac{\delta}{15x^6} \, dx \leq \frac{\delta}{12n^{3/4}}.
\]

In other words, with probability at least \( 1 - \frac{\delta}{12n^{3/4}} \):

\[
\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \not\in M_k} \leq \tau_{\max} R_{\max} \sqrt{n}.
\]

### B.3 Episodes with \( M \in M_k \)

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

\[
\Delta_k = \sum_{s \in S} \sum_{a \in A_s} \nu_k(s, a) (\tau(s, a) \rho^* - \overline{\tau}(s, a)) \leq \sum_{s \in S} \sum_{a \in A_s} \nu_k(s, a) (\tau(s, a) \bar{\rho}_k - \overline{\tau}(s, a)) + R_{\max} \sum_{s \in S} \sum_{a \in A_s} \frac{\nu_k(s, a)}{\sqrt{k}} \tau(s, a)
\]

\[
\implies \Delta_k \leq \sum_{s \in S} \sum_{a \in A_s} \nu_k(s, a) (\bar{\tau}_k(s, a) \bar{\rho}_k - \overline{\tau}(s, a)) + R_{\max} \sum_{s \in S} \sum_{a \in A_s} \nu_k(s, a) (\overline{\tau}(s, a) - \bar{\tau}_k(s, a)) + R_{\max} \tau_{\max} \sum_{s \in S} \sum_{a \in A_s} \frac{\nu_k(s, a)}{\sqrt{k}} \tau(s, a)
\]

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,

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

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

**Proof.** In Appendix \ref{app:uniformization} we show that EVI is value iteration applied to the equivalent MDP \( \tilde{M}_{k,\text{eq}} \) obtained by “uniformizing” the extended SMDP \( \tilde{M}^+_k \). Thus, we focus on any SMDP \( M \) and its equivalent MDP \( M_{\text{eq}} \). Using the same argument as in section 4.3.1 of \cite{Jaksch2010}, we have that: \( \forall i \geq 0, \max_{s \in S} u_i(s) - \min_{s \in S} u_i(s) \leq \frac{R_{\max} D(M)}{\tau} \).
We prove that \( \forall s, s' \in S, \forall \pi \in \Pi^\text{SP}_M = \Pi^\text{SP}_M, \ E^\pi_M[T(s')|s_0 = s] = \tau E^\pi_M[T(s')|s_0 = s] \). 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: \( \mathbb{P}_M^\pi(T(s') = +\infty|s_0 = s) > 0 \implies \mathbb{P}_M^\pi(T(s') = +\infty|s_0 = s) > 0 \) and so \( \mathbb{E}_M^\pi[T(s')|s_0 = s] = +\infty = \frac{1}{\tau} \mathbb{E}_M^\pi[T(s')|s_0 = s] \).

2. Conversely: \( \mathbb{P}_M^\pi(T(s') = +\infty|s_0 = s) = 0 \implies \mathbb{P}_M^\pi(T(s') = +\infty|s_0 = s) = 0 \) in which case both expectations are finite. To prove they are equal up to factor \( \tau \), we see the holding time as a “reward” (the true rewards are ignored here). Note that any policy \( \pi \) induces Markov chains with different dynamics on \( M \) and \( M_{\text{eq}} \) (different transition probabilities). We call these Markov chains \( MC \) and \( MC_{\text{eq}} \) 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 \( (s, \pi(s)) \) have an expected reward equal to \( \mathbb{E}(\pi) \) 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 \( s \) 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 \( (\mathbb{T}(\pi))_{\pi \in S} = (\mathbb{E}_M^\pi[T(s')|s_0 = \pi])_{\pi \in S} \) is the unique solution to the system of equations

\[
\forall \pi, \ \mathbb{T}(\pi) = \mathbb{E}(\pi) + \sum_{\bar{s}} p(\bar{s}|\pi) \mathbb{T}(\bar{s}).
\]

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

\[
\forall \pi, \ \mathbb{T}_{\text{eq}}(\pi) = 1 + \frac{\tau}{\mathbb{T}(\pi)} \sum_{\bar{s}} p(\bar{s}|\pi) \mathbb{T}_{\text{eq}}(\bar{s}) + \left( 1 - \frac{\tau}{\mathbb{T}(\pi)} \right) \mathbb{T}(\pi)
\]

\[
\iff \forall \pi, \ \left( \mathbb{T}_{\text{eq}}(\pi) = \mathbb{T}(\pi) + \sum_{\bar{s}} p(\bar{s}|\pi) \mathbb{T}_{\text{eq}}(\bar{s}) \right) \implies \tau \mathbb{T}_{\text{eq}} = \mathbb{T} \implies \tau D(M_{\text{eq}}) = D(M).
\]

By uniqueness of the solution: \( \tau D(M_{\text{eq}}) = D(M) \).

\[\square\]

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

\[
\forall s \in S, \ |u_{i+1}(s) - u_i(s) - \bar{\rho}_k| \leq \frac{1}{\sqrt{k}} \tag{21}
\]

**Proof.** We introduce the following quantities

\[
M_i = \max_{s \in S} \{ u_{i+1}(s) - u_i(s) \}, \ m_i = \min_{s \in S} \{ u_{i+1}(s) - u_i(s) \}, \ \epsilon = \frac{1}{\sqrt{i}}
\]
We will now upper-bound the four terms of the right-hand side of the above inequality. Setting \(i\)
\[
\frac{1}{2}(M_i + m_i) \geq \tilde{p}_k - \frac{\epsilon}{2} \iff m_i \geq \tilde{p}_k - \frac{\epsilon}{2} \iff \frac{1}{2}(M_i - m_i) \iff m_i \geq \tilde{p}_k - \epsilon
\]
\[
\frac{1}{2}(M_i + m_i) - \tilde{p}_k \leq \frac{\epsilon}{2} \iff M_i \leq \tilde{p}_k + \frac{\epsilon}{2} \iff \frac{1}{2}(M_i - m_i) \iff M_i \leq \tilde{p}_k + \epsilon.
\]

In conclusion:
\[
\forall s \in S, \quad \frac{-1}{\sqrt{k}} \leq u_{i+1}(s) - u_i(s) - \tilde{p}_k \leq \frac{1}{\sqrt{k}}.
\]

Based on Lemma 7, Eq. 21, and optimality equation Eq. 14, we have:
\[
\forall s \in S, \quad \left| \left( \tilde{P}_k^* - \tilde{P}_k(s, \pi_k(s)) \right) - \left( \sum_{s' \in S} \tilde{P}_k(s'|s, \pi_k(s))u_i(s') - u_i(s) \right) \right| \leq \frac{\tau}{\tilde{P}_k(s, \pi_k(s))} \leq \frac{1}{\sqrt{k}}.
\]

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

\[
\Delta_k \leq \sum_{s,a} \nu_k(s,a) \left( \tilde{P}_k(s,a)\tilde{p}_k - \tilde{P}(s,a) \right) + R_{\max} \sum_{s,a} \nu_k(s,a) \left( \tilde{P}(s,a) - \tilde{P}_k(s,a) \right) + R_{\max} \tau_{\max} \sum_{s,a} \nu_k(s,a) \frac{\sqrt{v_k}}{\sqrt{k}}
\]
\[
= \sum_{s,a} \nu_k(s,a) \left( \tilde{P}_k(s,a)\tilde{p}_k - \tilde{P}_k(s,a) \right) + \sum_{s,a} \nu_k(s,a) \left( \tilde{P}_k(s,a) - \tilde{P}(s,a) \right) + R_{\max} \tau_{\max} \sum_{s,a} \nu_k(s,a) \frac{\sqrt{v_k}}{\sqrt{k}}
\]
\[
= \sum_{s,a} \nu_k(s,a) \left( \tilde{P}_k(s,a)\tilde{p}_k - \tilde{P}_k(s,a) \right) + R_{\max} \tau_{\max} \sum_{s,a} \nu_k(s,a) \frac{\sqrt{v_k}}{\sqrt{k}}
\]

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

\[
\tilde{r}_k(s,a) - \tilde{P}(s,a) \leq |\tilde{r}_k(s,a) - \tilde{P}_k(s,a)| + |\tilde{P}_k(s,a) - \tilde{P}(s,a)| \leq 2\beta_k^2 \leq 2c_r \frac{\log(2SAI_k/\delta)}{\max\{1, N_k(s,a)\}}
\]

\[
\tilde{P}(s,a) - \tilde{P}_k(s,a) \leq |\tilde{P}(s,a) - \tilde{P}_k(s,a)| + |\tilde{P}_k(s,a) - \tilde{P}_k(s,a)| \leq 2\beta_k^2 \leq 2c_\tau \frac{\log(2SAI_k/\delta)}{\max\{1, N_k(s,a)\}}
\]

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

\[
\tilde{P}_k(s,a)\tilde{p}_k - \tilde{P}_k(s,a) \leq R_{\max} \tau_{\max} \sqrt{v_k} + \tau \left( \sum_{s' \in S} \tilde{P}_k(s'|s, \pi_k(s))u_i(s') - u_i(s) \right), \quad \text{if} \ a = \pi_k(s)
\]

\[
\implies \sum_{s,a} \nu_k(s,a) \left( \tilde{P}_k(s,a)\tilde{p}_k - \tilde{P}_k(s,a) \right) \leq R_{\max} \tau_{\max} \sum_{s,a} \nu_k(s,a) \sqrt{v_k} + \tau \left( v_k(\tilde{P}_k - I)u_i \right)
\]

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

\[
w_k(s) = u_i(s) - \frac{\min_{s \in S} u_i(s) + \max_{s \in S} u_i(s)}{2}
\]
and since the rows of $\tilde{P}_k$ sum to one, we have: $v_k(\tilde{P}_k - I)u_i = v_k(\tilde{P}_k - I)w_k$. 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:

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

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

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

**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:

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

**Proof.** See Appendix C.3 of Jaksch et al. [2010].

Using Lemma 8 we get:

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

By Jensen’s inequality we thus have:

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

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

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

**B.4 Computing the final bound**

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

$$\Delta(M, \mathfrak{A}, s, n) \leq (\gamma_r + \gamma_rR_{\max}) \log \left(\frac{13n}{\delta}\right) \sqrt{n} + \tau_{\max}R_{\max} \sqrt{n} + R_{\max}D \sqrt{\frac{5}{2}n \log \left(\frac{8n}{\delta}\right)}$$

$$+ (\sqrt{2} + 1) \left[2\tau_{\max} + 2(c_r + R_{\max}c_r) \log \left(\frac{2SA_n}{\delta}\right) + R_{\max}D \sqrt{14S \log \left(\frac{2An}{\delta}\right)} \right] \sqrt{SA}n$$

$$+ R_{\max}DSA \log_2 \left(\frac{8n}{SA}\right)$$
In [Jaksch et al., 2010] (see Appendix C.4), it is shown that when \( n > 34A \log \left( \frac{n}{\delta} \right) \):

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

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

\[
\frac{n^2}{\delta^2} \geq \frac{nS \log \left( \frac{n}{\delta} \right)}{\delta} \geq \frac{nS}{\delta} \Rightarrow \frac{n^2A^2}{\delta^2} \geq 2SA \frac{n}{\delta} \Rightarrow 4 \log \left( \frac{n}{\delta} \right) \geq 2 \log \left( \frac{2An}{\delta} \right) \geq \log \left( 2SA \frac{n}{\delta} \right)
\]

\[
\Rightarrow \Delta(M, \mathfrak{A}, s, n) = O \left( \left( D \sqrt{S} + \tau_{\max} + \frac{C_r}{R_{\max}} + C_r \right) \sqrt{\log \left( \frac{n}{\delta} \right)} \right) R_{\max} \sqrt{SA} \log \left( \frac{n}{\delta} \right)
\]

where \( C_r = \max \{ b_r, \sigma_r \} \) and \( C_r = \max \{ b_r, \sigma_r \} \).

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

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

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 \( 1 - \delta \):

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

The derivation for the case of bounded holding times is exactly the same with different concentration inequalities applied to estimates \( \hat{\tau}(s,a) \). 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. C.

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 \( \frac{B}{12} > T_{\min} \) and \( \frac{B}{3} > T_{\min} \), let \( T = \left[ T_{\min}, \min \left\{ \frac{B}{12}, \frac{B}{3} \right\} \right] \). Define \( p = \frac{T}{T_{\max}} \) and \( \delta = \frac{1}{T_{\max}} \). By definition of \( T \) we have: \( p, \delta \leq \frac{1}{4} \). There are \( A' \) actions available in each state of \( M' \). We assume that \( \forall (s,a,a') \in S \times A \times S \) actions available in each state of \( M' \) are independent. We also assume that \( \forall i \geq 0, \tau_i(s_{i-1},a_{i-1},s_i) \) and \( r_i(s_{i-1},a_{i-1},s_i) \) are independent of the next state \( s_i \) and we write: \( \tau_i(s_{i-1},a_{i-1},s_i) \sim \tau \), \( r_i(s_{i-1},a_{i-1},s_i) \sim \tau \), where \( \tau \) is a r.v. defined in Table 4. Moreover, for all actions \( a \) a specific action \( a_0 \) and \( p(s|s_0,a_0) = \delta \) whereas \( p(s|s_0,a_0) = \delta + \epsilon \) for some \( 0 < \epsilon < \delta \) specified later in the proof. For all actions \( a \) in \( A_{s_0} \), \( r(s_0,a) = 0 \) and \( \tau(s,a) \sim \tau \), where \( \tau \) is a r.v. defined in Table 4. Moreover, for all actions \( a \) a specific action \( a_0 \) and \( p(s|s_0,a_0) = \delta \) whereas \( p(s|s_0,a_0) = \delta + \epsilon \) for some \( 0 < \epsilon < \delta \) specified later in the proof.

Note that since \( \eta < \frac{1}{T_{\max}} \), we have: \( T \leq TR_{\max} \) and \( T \leq \tau R_{\max} \) which satisfies the definition of \( R_{\max} \) given in assumption 1. We denote \( E[|s|] \) 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. \( \forall d^\infty \in \Pi_{M'} \), such that \( d(s_0) \neq a_0^* \)
we have:

$$E^{d^\infty}[T(s_1)|s_0] = E^{d^\infty}\left[ \sum_{n=1}^{+\infty} \left( \prod_{i=1}^{n} \tau_i(s_{i-1}, a_{i}, s_{i}) \right) \left( \prod_{j=0}^{n-1} 1_{s_j=s_0} \right) 1_{s_n=s_1} \right] s_0$$

$$= \sum_{n=1}^{+\infty} E^{d^\infty} \left[ \sum_{i=1}^{n} \tau_i(s_0, d(s_0)) \prod_{j=0}^{n-1} 1_{s_j=s_0} \right] 1_{s_n=s_1}$$

$$= \sum_{n=1}^{+\infty} E^{d^\infty} \left[ \sum_{i=1}^{n} \tau_i(s_0, d(s_0)) \right] E^{d^\infty} \left[ \prod_{j=0}^{n-1} 1_{s_j=s_0} \right] 1_{s_n=s_1}$$

$$= \sum_{n=1}^{+\infty} n \tau \times \delta(1-\delta)^{n-1} = \frac{\tau}{\delta}$$

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

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

Any policy $d^\infty \in \Pi^{S^D}$ 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 $P^*_d = \epsilon \mu_d^T$ where $\mu_d = (1-p_d, p_d)^T$ is the stationary distribution of the recurrent Markov Chain. The probability $p_d$ can take only two different values: $p_d = \frac{1}{\tau}$ 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 $d^\infty$ has the form:

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

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 $\tau^* > \tau$, the optimal decision rule $d^*$ must satisfy: $d^*(s_1) = a_1^*$. Similarly, since $\frac{X + \epsilon}{\tau} \geq \frac{R_{\max}}{2\epsilon}$ we must have: $d^*(s_0) = a_0^*$. The optimal gain is thus:

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

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
Exploration–Exploitation in MDPs with Options

<table>
<thead>
<tr>
<th>R.v.</th>
<th>X</th>
<th>$X_{\text{min}}$</th>
<th>$X_{\text{max}}$</th>
<th>$P(X = X_{\text{min}})$</th>
<th>$P(X = X_{\text{max}})$</th>
<th>$\mathbb{E}[X]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>0</td>
<td>$T_{\text{min}}$</td>
<td>$T_{\text{max}}$</td>
<td>$&gt; 0$</td>
<td>$&gt; 0$</td>
<td>$\frac{1}{\tau}R_{\text{max}}$</td>
</tr>
<tr>
<td>$r$</td>
<td>0</td>
<td>$\frac{1}{2}R_{\text{max}}T_{\text{max}}$</td>
<td>$1 - p$</td>
<td>$p$</td>
<td>$p + \eta$</td>
<td>$\frac{1}{\tau}R_{\text{max}}$</td>
</tr>
<tr>
<td>$r^*$</td>
<td>0</td>
<td>$\frac{1}{2}R_{\text{max}}T_{\text{max}}$</td>
<td>$1 - p - \eta$</td>
<td>$p$</td>
<td>$p + \eta$</td>
<td>$\frac{1}{\tau}R_{\text{max}}$</td>
</tr>
</tbody>
</table>

Table 1: Definition of random variables $\tau$, $r$ and $r^*$.

holding time is $T_{\text{min}}$. 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 go 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 $2\left(\frac{D}{\tau} + T_{\text{min}}\log A' k\right) \leq D$. All holding times of $M$ are in $[T_{\text{min}}, T_{\text{max}}]$, and all rewards are in $[0, \frac{1}{2}R_{\text{max}}T_{\text{max}}]$. For all $(s, a) \in S \times A$, we have: $\tau(s, a) \leq \tau(s, a)R_{\text{max}}$. 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 $A$ used is fixed. The probability distribution of the stochastic process $(s_0, a_0, \tau_0, r_0, s_1, ...)$ is denoted:

- $P_{a_0,a_1}$ when $(a_0, a_1)$ are the best actions in respectively $s_0$ and $s_1$,
- $P_*$ 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 $A$ starts,
- $P_{\text{uni0, a1}}$ when the pair $a_1$ is the best action in $s_1$ and $\epsilon = 0$ (no-optimal actions in $s_0$),
- $P_{a_0, \text{uni0}}$ 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: $\mathbb{E}_*[\Delta(M, A, s, n)] \geq \mathbb{E}_*[\Delta(M'', A, s, n)]$ ($M''$ is easier to learn). We will show that $\mathbb{E}_*[\Delta(M'', A, s, n)] = \Omega(\left(\sqrt{D'} + \sqrt{T_{\text{max}}}\right) R_{\text{max}}\sqrt{kA'n})$ 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: $\mathbb{E}_{a_0, a_1}[\Delta(M, A, s, n)] = \Omega\left((\sqrt{D'} + \sqrt{T_{\text{max}}}\right) R_{\text{max}}\sqrt{SA'n})$.

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 $A$ 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 $(a, \overline{a}) \in A_{s_0} \times A_{s_1}$, let’s denote by $N_{a_0}^{\overline{a}}$ and $N_{a_1}^{\overline{a}}$ the respective number of times actions $a$ and $\overline{a}$ are taken. Finally, let’s denote by $N_0^n$ and $N_1^n$ 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 $(a_0, a_1)$:

$$
\mathbb{E}_{\text{uni}0, a_1}[N_1] \geq \frac{n}{2} - \frac{1}{2\delta} = \frac{n}{2} - \frac{D'}{2\tau}
$$

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

The regret is defined as:

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

where the total duration is simply:

$$
\mathbb{E}_* \left[ \sum_{i=1}^{n} \tau_i(s_{i-1}, a_i, s_i) \right] = \sum_{i=1}^{n} \mathbb{E}_* \left[ \tau_i(s_{i-1}, a_i) \right] = n\tau
$$
and the cumulated reward is given by:

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

\[
= \mathbb{E}_s \left[ \tau^* \left( \sum_{i=0}^{n-1} \mathbb{I}_{s_i=s_{i+1}, a_i=a_i^*} \right) + \tau \left( \sum_{i=0}^{n-1} \mathbb{I}_{s_i=s_{i+1}, a_i \neq a_i^*} \right) \right]
\]

\[
= \mathbb{E}_s \left[ \tau (N_1 - N_1^\ast) + \tau^* N_1^\ast \right]
\]

\[
= \mathbb{E}_s \left[ \tau N_1 + (\tau^* - \tau) N_1^\ast \right]
\]

\[
= \frac{R_{\text{max}}}{2} \tau \mathbb{E}_s \left[ N_1 \right] + \frac{R_{\text{max}}}{2} T_{\text{max}} \mathbb{E}_s \left[ N_1^\ast \right]
\]

hence the formula:

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

**Lemma 9.** Let \( f : \{s_0, s_1\}^{n+1} \times \{T_{\text{min}}, T_{\text{max}}\}^n \times \{0, \frac{1}{2} R_{\text{max}} T_{\text{max}}\}^n \to [0, M] \) be any function defined on state/reward sequence \( (s^{n+1}, r^n, n) \in \{s_0, s_1\}^{n+1} \times \{T_{\text{min}}, T_{\text{max}}\}^n \times \{0, \frac{1}{2} R_{\text{max}} T_{\text{max}}\}^n \) 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 \( (a_0, a_1) \in \{1, \ldots, kA'\} \times \{1, \ldots, kA'\} \):

\[
\left| \mathbb{E}_{a_0, a_1} \left[ f(s^{n+1}, r^n) \right] - \mathbb{E}_{\text{unif}, a_1} \left[ f(s^{n+1}, r^n) \right] \right| \leq \frac{M}{2} \sqrt{2} \frac{2 \log(2) KL(P_{\text{unif}, a_1}||P_{a_0, a_1})}{\sqrt{\delta} \mathbb{E}_{\text{unif}, a_1} \left[ N_0^{\alpha} \right]} \]

\[
\left| \mathbb{E}_{a_0, a_1} \left[ f(s^{n+1}, r^n) \right] - \mathbb{E}_{\text{unif}, a_1} \left[ f(s^{n+1}, r^n) \right] \right| \leq \frac{M}{2} \sqrt{2} \frac{2 \log(2) KL(P_{a_0, a_1}||P_{\text{unif}, a_1})}{\sqrt{\delta} \mathbb{E}_{a_0, \text{unif}} \left[ N_1^{\alpha} \right]} \]

**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:

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

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

\[
KL(P_{a_0, a_1}||P_{\text{unif}, a_1}) = \sum_{i=1}^{n} KL \left( P_{a_0, a_1}(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) || P_{\text{unif}, a_1}(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) \right)
\]

where

\[
KL \left( P(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) || Q(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) \right) = \sum_{s_i \in \mathcal{S}, \tau_i \in \mathcal{T}, r_i \in \mathcal{R}} P(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) \log \frac{P(s_i, \tau_i, r_i | s^{i-1}, r^{i-1})}{Q(s_i, \tau_i, r_i | s^{i-1}, r^{i-1})}
\]

with \( \mathcal{S} = \{s_0, s_1\} \), \( \mathcal{T} = \{T_{\text{min}}, T_{\text{max}}\} \) and \( \mathcal{R} = \{0, \frac{1}{2} R_{\text{max}} T_{\text{max}}\} \). The same holds for \( P_{\text{unif}, a_1} \).

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

\[
KL \left( P_{\text{unif}, a_1}(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) || P_{a_0, a_1}(s_i, \tau_i, r_i | s^{i-1}, r^{i-1}) \right)
\]

\[
= P_{\text{unif}, a_1}(s_{i-1} = s_0, a_i = a_0) \sum_{s' \in \mathcal{S}, \tau' \in \mathcal{T}, r' \in \mathcal{R}} P_{\text{unif}, a_1}(s', \tau', r' | s_0, a_0) \log \frac{P_{\text{unif}, a_1}(s', \tau', r' | s_0, a_0)}{P_{a_0, a_1}(s', \tau', r' | s_0, a_0)}
\]

\[
= P_{\text{unif}, a_1}(s_{i-1} = s_0, a_i = a_0) \sum_{s' \in \mathcal{S}} P_{\text{unif}, a_1}(s' | s_0, a_0) \log \frac{P_{\text{unif}, a_1}(s' | s_0, a_0)}{P_{a_0, a_1}(s' | s_0, a_0)}
\]

\[
= P_{\text{unif}, a_1}(s_{i-1} = s_0, a_i = a_0) \left( \delta \log_2 \left( \frac{\delta}{\delta + \epsilon} \right) + (1 - \delta) \log_2 \left( \frac{1 - \delta}{1 - \delta - \epsilon} \right) \right)
\]
Then we have that:

We can bound the error using Lemma 9 from [Jaksch et al., 2010] which concludes the proof.

Taking into account the fact that by assumption we have the following inequality:

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

which concludes the proof.

Note that by assumption: \( \epsilon \leq \delta \leq \frac{1}{2} < 1 - 2\delta \) and \( \eta \leq p \leq \frac{1}{2} < 1 - 2p. \)

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 \( (s^{n+1}, \tau^n, r^n) \). Since the computations are rigorously the same except that \( \delta = \frac{1}{3\tau} \) instead of \( \frac{1}{2\tau} \), we give the results without any further details:

Taking into account the fact that by assumption \( n \geq DSA \geq 16D'kA' \) we get:

Given that \( \delta \geq \epsilon \geq 0 \) we have the following inequality:

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

We will now derive an upper-bound on \( \mathbb{E}_{a_0, \text{unif}} [N_1]. \)

**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:

Then we have that:

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

\[
\forall n \in \mathbb{N}, \quad u_n \geq \left( u_0 - \frac{r}{1-q} \right) a^n + \frac{r}{1-q}
\]
Proof. Defining sequence \( v_n = u_n - \frac{r}{1-q} \), we have:

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

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 \( u_n - \frac{r}{1-q} \).

By the law of total probability and since \( P_{a_0, \text{unif}}(s_i = s_0) + P_{a_0, \text{unif}}(s_i = s_1) = 1 \) we have:

\[
\forall i \geq 0, \quad P_{a_0, \text{unif}}(s_{i+1} = s_0) = P_{a_0, \text{unif}}(s_{i+1} = s_0 | s_i = s_0) P_{a_0, \text{unif}}(s_i = s_0) + P_{a_0, \text{unif}}(s_{i+1} = s_0 | s_i = s_1) P_{a_0, \text{unif}}(s_i = s_1)
\]

\[
= P_{a_0, \text{unif}}(s_i = s_0) (P_{a_0, \text{unif}}(s_{i+1} = s_0 | s_i = s_1) + P_{a_0, \text{unif}}(s_{i+1} = s_0 | s_i = s_1))
\]

\[
\geq P_{a_0, \text{unif}}(s_i = s_0) (1 - 2 \delta - \epsilon) + \delta
\]

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

\[
\forall i \geq 0, \quad P_{a_0, \text{unif}}(s_i = s_0) \geq \left( 1 - \frac{\delta}{2 \delta + \epsilon} \right) (1 - 2 \delta - \epsilon)^i + \frac{\delta}{2 \delta + \epsilon} \geq \frac{1}{3}
\]

\[
\Rightarrow E_{a_0, \text{unif}}[N_0] = \sum_{i=0}^{n} P_{a_0, \text{unif}}(s_i = s_0) \geq \frac{n}{3}
\]

\[
\Rightarrow E_{a_0, \text{unif}}[N_1] \leq \frac{2n}{3} + \eta \frac{\sqrt{T_{\text{max}} n k A'}}{3 \tau}
\]

Hence the bound:

\[
\eta T_{\text{max}} E_{a_0, \text{unif}}[N_1] \leq \frac{2n T_{\text{max}}}{3 k A'} + \frac{\eta}{3} \frac{T_{\text{max}} n k A'}{3 \tau} \tag{26}
\]

By setting \( \epsilon = c \sqrt{\frac{k A'}{n D'}} \) and \( \eta = \kappa \sqrt{\frac{n D'}{k A'}} \) and incorporating inequalities \( 21 \) and \( 23 \) into inequality \( 25 \) we obtain:

\[
E_{a_0, \text{unif}}[\Delta(M'', A, s_0, n)] \geq \left[ \frac{c}{6} - \frac{c}{2 k A'} - \frac{c}{32 k^2 A'^2} \right] \frac{c^2}{2} + \frac{c^2}{2} - \frac{c^2}{8 k A'} - \frac{1}{8 k A'} \right] R_{\text{max}} \frac{\sqrt{k A'} n}{3 \tau}
\]

For the second inequality, we used the fact that \( \tau > T_{\text{min}} \leq 1 \) (by assumption). If \( c \) and \( k \) are sufficiently small, then the conditions of lemma \( 9 \) are indeed satisfied and the above polynomials in \( c \) and \( k \) are non-negative. For example, if \( c = k = \frac{1}{4} \):

\[
n \geq D S A \geq 16 D' k A' \quad \Rightarrow \quad \epsilon = \frac{1}{5} \sqrt{\frac{k A'}{n D'}} \leq \frac{\delta}{20} < \delta
\]

\[
n \geq T_{\text{max}} S A \geq 4 T_{\text{max}} k A' \quad \Rightarrow \quad \eta = \frac{1}{5} \sqrt{\frac{k A'}{n T_{\text{max}}}} \leq \frac{p}{10} < p
\]

and \( E_{a_0, \text{unif}}[\Delta(M'', A, s_0, n)] \geq 0.0015 \times \left( \sqrt{D'} + \sqrt{T_{\text{max}}} \right) R_{\text{max}} \sqrt{k A'} n \)
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 $P(\tau^* = \frac{1}{2}R_{\max}T_{\max}) > P(\tau = T_{\max})$. 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: $\rho \leq \frac{1}{2}R_{\max}T_{\max}$, $\delta = \frac{1}{2}R_{\max}T_{\max}$.

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 $s_i$. The main difference with the previous lower bound is that we assume that $r_i$ and $\tau_i$ are strongly correlated, namely: $r_i(s_{i-1}, a_i, s_i) = R_{\max}\mathbbm{1}_{\{s_{i-1}=s_1\}}\tau_i(s_{i-1}, a_i)$. We assume $\epsilon < \delta$ and $\eta \leq 1 - 2p$. 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:

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

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

$$\mathbb{E}_a[\Delta(M'', A, s_0, n)] \geq \left[ \frac{\kappa}{6} - \frac{c}{2kA'} - \frac{\kappa^2}{32k^2A'^2} \right] R_{\max} \sqrt{D'kA'n}$$

$$+ \left[ \frac{\kappa}{6} - \frac{c}{2kA'} - \frac{\kappa^2}{32k^2A'^2} \right] R_{\max} \sqrt{(T_{\max} - T_{\min})kA'n}$$

$$\geq \left[ \frac{\kappa}{6} - \frac{c}{2kA'} - \frac{\kappa^2}{32k^2A'^2} \right] R_{\max} \sqrt{(T_{\max} - T_{\min})kA'n}$$

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

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

SMDPs $M'$ and $M''$ can be transformed into equivalent MDPs with options. We illustrate this transformation on Fig. 6 for an action $a_1 \in A_{s_1}$ 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 $a_1 \in A_{s_1} \setminus \{a_1^*\}$ is executed.
Table 2: Definition of random variables \(\tau, \tau^*, r\) and \(r^*\).

![Decomposition of an action](image)

Figure 6: Decomposition of an action \(a_1 \in A_s \setminus \{a_1^*\}\) into a set of primitive actions in an MDP. In this example, \(T_{\text{max}} = 3\) and \(T_{\text{min}} = 1\).

## 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:

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

Equation (27)

A finite 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 \(s^*\) for every state \(s\) for which \(\beta_o(s) > 0\). We then add a transitions from \(s\) to \(s^*\) with probability \(\beta_o(s) > 0\) and reward 0, and we add a self-loop on \(s^*\) with probability 1 and reward 0 (\(s^*\) 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 (\(Q e \leq e\) where \(e = (1, ..., 1)^\top\) and \(\exists j\) s.t. \((Q e)_j < 1\) 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:

\[
\forall k \in \mathbb{N}^+, \quad P(\tau(s, a, s') = k) = e^{\frac{Q}{s}}Q^{k-1}Re_{s'}
\]

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 \( e_s = (0, 0, \ldots, 1, 0, \ldots, 0)^T \) 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 \( \tau \leftarrow \tau(s, o, s') = \mathbb{E}[\tau(s, o, s')] \)):

\[
\mathbb{E}\left[ e^{\lambda(\tau-\tau)} \right] = \sum_{k=1}^{\infty} e^{\lambda(1-\tau)} e^k \mathbb{E}[Q^{k-1}] \mathbb{E}_s \mathbb{E}_{s'} = e^{\lambda(1-\tau)} e^\tau \sum_{k=0}^{\infty} (e^\lambda Q)^k \mathbb{E}_s \mathbb{E}_{s'}
\]

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

\[
\mathbb{E}\left[ e^{\lambda(\tau-\tau)} \right] = e^{\lambda(1-\tau)} e^\tau (I - e^\lambda Q)^{-1} \mathbb{E}_s \mathbb{E}_{s'}
\]

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

1. \( \rho(Q) = 0 \) if and only if all the eigenvalues of \( Q \) are 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. \( T^n = U^{-1}Q^nU = 0 \Rightarrow Q^n = 0 \) 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 \( Qv = \lambda v \Rightarrow Q^n v = \lambda^n v = 0 \), 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 \( \mathbb{E}\left[ e^{\lambda(\tau-\tau)} \right] \) can not be bounded by a function of the form \( \lambda \mapsto e^{\frac{\lambda^2}{c_0}} \) for \( \lambda \geq -\log(\rho(Q)) \) so \( \tau(s, o, s') \) is not sub-Gaussian (Definition 4). However, since \( \rho(Q) < 1 \) we can choose \( 0 < c_0 < -\log(\rho(Q)) \) and we have \( \mathbb{E}\left[ e^{\lambda(\tau-\tau)} \right] < +\infty \) 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 \( P(r \leq R_{\max}) = 1 \) we have:

\[
\forall \lambda > 0, \quad \mathbb{E}\left[ e^{\lambda(r-\tau)} \right] = \sum_{k=1}^{\infty} \mathbb{E}\left[ e^{\lambda(r-\tau)} \tau = k \right] P(\tau = k) = \sum_{k=1}^{\infty} \mathbb{E}\left[ e^{\lambda(R_{\max}T-\tau)} \tau = k \right] P(\tau = k) = \sum_{k=1}^{\infty} e^{\lambda(R_{\max}T-k)} P(\tau = k) = e^{\lambda(R_{\max}T)} e^T \sum_{k=0}^{\infty} (e^{\lambda R_{\max}Q})^k \mathbb{E}_s \mathbb{E}_{s'}
\]

We can now conclude as we did for \( \tau(s, o, s') \): let \( 0 < c_0 < -\log(\rho(Q)) R_{\max} \), for all \( 0 < \lambda < c_0 \) the quantity \( \mathbb{E}\left[ e^{\lambda(r-\tau)} \right] \) is finite. Note that for \( \lambda \leq 0 \): \( \mathbb{E}\left[ e^{\lambda r} \right] \leq 1 \) so \( \mathbb{E}\left[ e^{\lambda(r-\tau)} \right] < +\infty \). 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 $M_O$ induced by the set of options $O$. By definition of $M_O$, 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 $\sum_{i=1}^n r_O^i = \sum_{i=1}^{N(T_n)} r_O^i = \sum_{i=1}^{T_n} r_i$ and:

$$\Delta(M, A, s, T_n) = T_n \rho^*(M) - \sum_{t=1}^{T_n} r_t$$

$$= T_n \rho^*(M') + T_n (\rho^*(M) - \rho^*(M_O)) - \sum_{i=1}^n r_O^i$$

$$= \Delta(M_O, A, s, n) + T_n (\rho^*(M) - \rho^*(M_O))$$

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

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

$$\forall T' \in \mathbb{R}^+, \quad \rho^{\pi_O}(M_O, s, T') = \mathbb{E}_{M_O}^{\pi_O} \left[ \frac{\sum_{i=1}^{N(T')} r_O^i}{T'} \middle| s_0 = s \right].$$

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, $\lim_{T' \to +\infty} \rho^{\pi_O}(M_O, s, T')$ exists since $\pi_O$ is stationary and deterministic (see appendix A) and by composition of the limit we have

$$\lim_{n \to +\infty} \rho^{\pi_O}(M_O, s, T_n) = \lim_{T' \to +\infty} \rho^{\pi_O}(M_O, s, T') = \rho^{\pi_O}(M_O, s)$$

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 $\pi$ is stationary deterministic and we know from MDP theory [Puterman, 1994] that the corresponding average reward exists. We also have:

$$\forall n \geq 1, \quad \mathbb{E}_M^\pi \left[ \frac{\sum_{t=1}^T r_t}{T} \middle| s_0 = s \right] = \mathbb{E}_{M_O}^{\pi_O} \left[ \frac{\sum_{i=1}^{N(T')} r_O^i}{T'} \middle| s_0 = s \right]$$

$$\implies \rho^{\pi_O}(M_O, s) = \lim_{n \to +\infty} \rho^\pi(M, s, T_n) = \lim_{T' \to +\infty} \rho^\pi(M, s, T) = \rho^\pi(M, s)$$

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 $O$) in the original MDP. As a result, $\rho^\pi(M_O)$ is the maximal average reward achievable in $M$ using only options in $O$. In conclusion, the linear term $T_n (\rho^\pi(M) - \rho^\pi(M_O))$ 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 $\beta(s_k) = 1/(m - k + 1)$. 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:
• If $k = 1$:

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

• If $k \geq 1$:

$$P(\tau = k) = \left( \prod_{i=1}^{k-1} (1 - \beta_o(s_i)) \right) \times \beta_o(s_k)$$

$$= \left( \prod_{i=1}^{k-1} \left( 1 - \frac{1}{m - i + 1} \right) \right) \times \frac{1}{m - k + 1}$$

$$= \left( \prod_{i=1}^{k-1} \left( \frac{m - i}{m - i + 1} \right) \right) \times \frac{1}{m - k + 1} = \frac{1}{m}$$

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:

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

**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
With probability 1, this result is rather intuitive: without resorting to the restart triggered by the target state, whereas if we start 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 \( \tau_i \) as follows:

\[
\begin{align*}
\tau_1 &= \frac{1}{m} + \frac{1}{m^2} (2 + \tau_1) + \cdots + \frac{1}{m} (m + \tau_{m-1}) \\
\tau_2 &= \frac{1}{m} \times 2 + \frac{1}{m^2} (1 + \tau_1) + \frac{1}{m} (3 + \tau_1) + \cdots + \frac{1}{m} (m + \tau_{m-2}) \\
\tau_3 &= \frac{1}{m} \times 3 + \frac{1}{m^2} (1 + \tau_2) + \frac{1}{m} (2 + \tau_1) + \frac{1}{m} (4 + \tau_1) + \cdots + \frac{1}{m} (m + \tau_{m-3}) \\
&\vdots \\
\tau_i &= \frac{m + 1}{2} + \frac{1}{m} \sum_{j=1}^{i-1} \tau_j + \frac{1}{m} \sum_{j=1}^{m-i} \tau_j
\end{align*}
\]

With probability 1, the next state after executing the option is 1 step to the left of \( s \) and the value of \( \tau_1 \) is then 1. With probability 1, 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 \( \tau_1 \) is thus \( 2 + \tau_1 \). With probability 1, the next state is 3 steps to the left of \( s \) and so \( s' \) is now located 2 steps to the right of the new state: the value of \( \tau_1 \) 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. It is trivial to verify that the only solution of the linear system in equation 29 is: \( \tau_i = m(m+1)/2, \forall i \in \{1, \ldots, m - 1\} \).

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. So in this sense, we have i.i.d. Bernoulli trials where the probability of success is 1. The expected time to reach \( r \) 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. 8 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 (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 \( 1/m \). The expected time needed to go from \( s_0 \) to \( s_1 \) in the SMDP is bounded by \( D + m(m+1) \).

---

\( ^{8} \)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 \( \text{LEFT} \). 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. 8. So the proof remains valid.
Exploration–Exploitation in MDPs with Options

Figure 9: Behaviour of the agent with deterministic options.

\[ \mathbb{E}[\tau] \leq D \]

\[ p = 1 - \frac{1}{m^2} \]

Figure 10: Upper bound on the diameter of \( M_O \) 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 \( \text{LEFT}, \text{RIGHT}, \text{UP} \) and \( \text{DOWN} \)).

**Asymptotic behaviour** We will now analyse the behaviour of the ratio \( n \) 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 \( F_0, F_1, ..., F_n \) 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:

\[ \sum_{i=1}^{+\infty} \mathbb{E}[X_i^2|F_{i-1}] < +\infty \quad \text{a.s.} \]

then:

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

Let’s take \( X_i = \tau_i - \tau_{i-1} \) (where \( \tau_i = \tau_i(s_{i-1}, a_i) \)) and \( F_i = \sigma(s_0, a_1, \tau_1, ..., s_i, a_{i+1}) \). The sequence \( (X_i)_{i<1} \) is a martingale difference because \( \mathbb{E}[X_i] < +\infty \) and \( \mathbb{E}[X_i|F_{i-1}] = 0 \). 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 \( \sigma^2 \) hence: \( \mathbb{E}[X_i^2|F_{i-1}] < \sigma_i^2 \). If in addition we take \( A_i = i \) then the conditions of Theorem 3 are satisfied and thus:

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

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

\[ \left| \frac{T_n}{n} - \frac{T_n}{n} \right| \leq \epsilon \quad \text{a.s.} \implies \tau_{\text{min}} - \epsilon \leq \frac{T_n}{n} \leq \tau_{\text{max}} + \epsilon \leq \frac{T_n}{n} \leq \tau_{\text{max}} \]
and so: \( \liminf_{n \to +\infty} \frac{T_n}{n} \geq \tau_{\text{min}} \) and \( \limsup_{n \to +\infty} \frac{T_n}{n} \leq \tau_{\text{max}} \) a.s. Finally:

\[
\log\left\{ \frac{T_n}{\delta} \right\} = \log\left\{ \frac{T_n}{n} + \frac{T_n}{\delta} \right\} \leq \log\left\{ \frac{T_n}{n} + \tau_{\text{min}} \right\} + \log\left\{ \frac{n}{\delta} \right\}, \quad n \to +\infty
\]

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 \( C(M', n, \delta) \) scales as \( \sqrt{\log(n)} \):

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

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

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

Note that \( \tau_{\text{min}} \) 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 \( \tau_{\text{max}} \) 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 \( T_{\text{max}} \) 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 \( T_{\text{max}} \). We can conclude that the ratio of the upper bounds is a good proxy for the true ratio in this example.
Figure 11: (a) Theoretical ratios of the regrets with and without options for different values of $T_{\text{max}}$; (b) Empirical ratios of the regrets with and without options for different values of $T_{\text{max}}$. 