Journal of Machine Learning Research 1 (year) pages Submitted 4/00; Published 10/00

Exploration–Exploitation in MDPs with Options

Ronan Fruit RONAN.FRUIT@INRIA.FR
Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
SequeL Team, Inria Lille - Nord Europe

Editor: Gergely Neu, Vicenç Gómez, Csaba Szepesvari

1. Introduction

The option framework (Sutton et al., 1999) is a simple yet powerful model to introduce temporally-extended actions and hierarchies in reinforcement learning (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). While options may significantly improve the performance w.r.t. learning with primitive actions, a theoretical understanding of their actual impact on the learning performance is still fairly limited. Notable exceptions are the sample complexity analysis of approximate value iteration with options (Mann and Mannor, 2014) and the PAC-MDP analysis by Brunskill and Li (2014). In this paper, 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 bound its regret. 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.

2. Preliminaries

MDPs and options. A finite MDP is a tuple $M = \{\mathcal{S}, \mathcal{A}, p, r\}$ where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $p(s'|s,a)$ is the probability of transitioning from state s to state s' when action a is taken, $r(s,a,s')$ is a distribution over rewards obtained when action a is taken in state s and the next state is s'. A (Markov) option is a tuple $o = \{\mathcal{I}_o, \beta_o, \pi_o\}$ where $\mathcal{I}_o \subset \mathcal{S}$ is the set of states where the option can be initiated, $\beta_o : \mathcal{S} \to [0,1]$ s.t. $\beta_o(s)$ is the probability of terminating option o given that the agent reaches state s, and $\pi_o : \mathcal{S} \to \mathcal{A}$ is the policy followed until the option ends. Whenever the set of primitive actions $\mathcal{A}$ is replaced by a set of options $\mathcal{O}$, the resulting decision process is no longer an MDP but it belongs to the family of semi-Markov decision processes (SMDP).

Proposition 1 *[Sutton et al. (1999)] For any MDP M and a set of options $\mathcal{O}$, the resulting decision process is an SMDP $M_{\mathcal{O}} = \{\mathcal{S}_{\mathcal{O}}, \mathcal{O}, p_{\mathcal{O}}, r_{\mathcal{O}}, \tau_{\mathcal{O}}\}$, where $\mathcal{S}_{\mathcal{O}} \subseteq \mathcal{S}$, $p_{\mathcal{O}}(s,o,s')$ is the transition probability from s to s' when o is executed, $r_{\mathcal{O}}(s,o,s')$ is the distribution of the*

cumulative reward obtained by executing option o from state s until interruption at s', and $\tau_{\mathcal{O}}(s,o,s')$ is the distribution of the holding time (i.e., number of steps).

Relying on this mapping, we first study the exploration-exploitation trade-off in a generic SMDP. A thorough discussion on the implications of the regret bounds in SMDPs for the case of learning with options in MDPs is reported in Sect. 5.

Learning in SMDPs. For any SMDP $M = \{\mathcal{S}, \mathcal{A}, p, r, \tau\}$, we denote by $\bar{\tau}(s,a,s')$ (resp. $\bar{r}(s,a,s')$) the expectation of $\tau(s,a,s')$ (resp. $r(s,a,s')$) and by $\bar{\tau}(s,a) = \sum_{s' \in \mathcal{S}} \bar{\tau}(s,a,s') p(s'|s,a)$ (resp. $\bar{r}(s,a) = \sum_{s' \in \mathcal{S}} \bar{r}(s,a,s') p(s'|s,a)$) the expected hodling 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 $N(t) = \sup\{n : n \in \mathbb{N},\ \sum_{i=1}^{n} \tau_i \le t\}$ the number of decision steps that occurred before time t. For any policy π and $s \in \mathcal{S}$, we define:*

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

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

We are now ready to consider the learning problem. For any $i \in \mathbb{N}^*$, a_i denotes the action taken by the agent at the i-th decision step[1] and s_i denotes the state reached after a_i is taken, with s_0 being the initial state. We denote by $(r_i(s,a,s'))_{i \in \mathbb{N}^*}$ (resp. $(\tau_i(s,a,s'))_{i \in \mathbb{N}^*}$) a sequence of i.i.d. realizations from $r(s,a,s')$ (resp. $\tau(s,a,s')$). When the learner explores the SMDP, it observes the sequence $(s_0, \dots, s_i, a_{i+1}, r_{i+1}(s_i, a_{i+1}, s_{i+1}), \tau_{i+1}(s_i, a_{i+1}, s_{i+1}), \dots)$ [2]. The performance of a learning algorithm is measured in terms of its cumulative *regret*.[3]

Definition 3 *For any SMDP M, any starting state $s \in \mathcal{S}$, and any number of decision steps $n \ge 1$, let $\{\tau_i\}_{i=1}^{n}$ be the random holding times observed along the trajectory generated by a learning algorithm $\mathfrak{A}$. Then the total regret of $\mathfrak{A}$ is defined as*

$$\Delta(M, \mathfrak{A}, s, n) = \left(\sum_{i=1}^{n} \tau_i\right) \rho^*(M) - \sum_{i=1}^{n} r_i. \tag{2}$$

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 in Eq. 5. 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. Finally, $\tilde{\pi}_k^*$ is executed until the number of samples from

1. Notice that decision steps are discrete points in time in which an action is started, while the (possibly continuous) holding time is determined by the distribution τ.
2. $r_{i+1}(s_i, a_{i+1}, s_{i+1})$ (respectively $\tau_{i+1}(s_i, a_{i+1}, s_{i+1})$) will be abbreviated r_i (respectively τ_i).
3. This definition reduces to the original definition of regret in MDPs for $\tau_i = 1$.

a state-action pair is doubled. While the structure is similar to UCRL's, the confidence intervals construction and the extended value iteration algorithm need to be redefined.

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 Lemma 8). As a result, the confidence intervals used at step 4 are defined as

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

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

where σ_r, b_r, σ_τ, b_r are suitable constants.

Extended value iteration (EVI). In order to solve EVI, we rely on a data-transformation (also called "uniformization") which turns any SMDP M into an "equivalent" MDP $M' = \{\mathcal{S}, \mathcal{A}, p', r'\}$ with the same state and action spaces and such that

$$\forall s,s' \in \mathcal{S}, \forall a \in \mathcal{A}_s, \begin{cases} \bar{r}'(s,a) = \frac{\bar{r}(s,a)}{\bar{\tau}(s,a)} \\ p'(s'|s,a) = \frac{\tau}{\bar{\tau}(s,a)}\big(p(s'|s,a) - \delta_{s,s'}\big) + \delta_{s,s'} \end{cases} \tag{7}$$

where $\delta_{s,s'} = 0$ if $s \neq s'$ and $\delta_{s,s'} = 1$ otherwise, and τ is an arbitrary non-negative real strictly smaller than $\tau_{\min}$. Lemma 2 of Federgruen et al. (1983) guarantees that if (v^*, g^*) is the optimal pair bias and gain in M' then $(\tau^{-1}v^*, g^*)$ is optimal for the corresponding SMDP M (see Appendix A. Thus, EVI is obtained by applying a value iteration scheme to MDP equivalent to the optimistic plausible SMDP. We denote the state values of the j-th iteration by $u_j(s)$. We also use the vector notation $u_j = (u_j(s))_{s\in\mathcal{S}}$. Similarly, we denote by $\tilde{p}(\cdot \mid s,a) = (\tilde{p}(s' \mid s,a))_{s'\in\mathcal{S}}$ the transition probability vector of state-action pair (s,a). The optimistic reward and holding time are obtained as $\tilde{r}_{j+1}(s,a) = \hat{r}_k(s,a) + \beta_k^r(s,a)$ and

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

where the optimistic transition model is obtained as $\tilde{p}_{j+1}(\cdot \mid s,a) \in \operatorname{Arg\,max}_{p(\cdot)\in\mathcal{P}_k(s,a)} \{p^\intercal u_j\}$ and $\mathcal{P}_k(s,a)$ is the set of probability distributions included in the confidence interval defined by $\beta_k^p(s,a)$. This optimization problem can be solved in $O(S)$ operations using the same algorithm as in UCRL. As a result, for any $\tau \in]0, \tau_{\min}[$, EVI proceeds through iterations defined as

$$u_{j+1}(s) = \max_{a\in\mathcal{A}_s}\left\{\frac{\tilde{r}_{j+1}(s,a)}{\tilde{\tau}_{j+1}(s,a)} + \frac{\tau}{\tilde{\tau}_{j+1}(s,a)}\left(\tilde{p}_{j+1}(\cdot \mid s,a)^\intercal u_j - u_j(s)\right)\right\} + u_j(s), \tag{8}$$

with $u_0(s) = 0$ and stopping condition $\max_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\} - \min_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\} < \epsilon$. In particular, we prove the following.

Theorem 4 *If the stopping condition holds at iteration i of EVI, then the greedy policy w.r.t. u_i is ϵ-optimal for the 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 $\epsilon = 1/\sqrt{i_k}$ guarantees that $\tilde{\pi}_k$ is $1/\sqrt{i_k}$-optimal w.r.t. $\max_{\tilde{M}_k \in \mathcal{M}_k} \rho^*(\tilde{M}_k)$.

4. Theoretical Analysis

In this section we report upper and lower bounds on the regret of UCRL-SMDP, their implication to the regret of learning with options in MDPs is postponed to Sec. 5.

We first extend the notion of diameter to the case of SMDP as follows.

Definition 5 *For any SMDP M, we define the diameter $D(M)$ by:*

$$D(M) = \max_{s,s' \in \mathcal{S}} \left\{ \min_{d \in D_M^{MD}} \left\{ \mathbb{E}^{d^\infty} \left[T(s') | s_0 = s \right] \right\} \right\} \tag{9}$$

where $T(s')$ is defined as the first actual time in which s' is encountered, i.e., $T(s') = \inf \left\{ \sum_{i=1}^n \tau_i : n \in \mathbb{N},\ s_n = s' \right\}$ and D_M^{MD} is the set of deterministic Markov Decision rules.

Note that the diameter of an SMDP corresponds to an average *actual* duration and not an average number of decision steps. However, if the SMDP is an MDP the two definitions of diameter coincides. Before reporting the main theoretical results about UCRL-SMDP, we introduce a set of technical assumptions.

Assumption 1 *For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we assume that $\tau_{\max} \geq \bar{\tau}(s,a) \geq \tau_{\min} > 0$ and $\max_{s \in \mathcal{S}, a \in \mathcal{A}_s} \left\{ \frac{\bar{r}(s,a)}{\bar{\tau}(s,a)} \right\} \leq R_{\max}$ with $\tau_{\min}$, $\tau_{\max}$, and $R_{\max}$ known to the learning algorithm. Furthermore, we assume that the random variables $(r(s,a,s'))_{s,a,s'}$ and $(\tau(s,a,s'))_{s,a,s'}$ are either **1)** sub-Exponential with constants (σ_r, b_r) and (σ_τ, b_τ), or **2)** bounded in $[0, R_{\max} T_{\max}]$ and $[T_{\min}, T_{\max}]$, with $T_{\min} > 0$ and $T_{\max}$. 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 6 *With probability of at least $1 - \delta$, it holds that for any initial state $s \in \mathcal{S}$ and any $n > 1$, the regret of* UCRL-SMDP *is bounded as follows:*

$$\Delta(M, \mathfrak{A}, s, n) = O\left(\left(D\sqrt{S} + \mathcal{C}(M, n, \delta) \right) R_{\max} \sqrt{SAn \log\left(\frac{n}{\delta}\right)} \right), \tag{10}$$

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

$$\textit{sub-Exponential} \qquad \mathcal{C}(M, n, \delta) = \tau_{\max} + \left(\frac{\max\{\sigma_r, b_r\}}{R_{\max}} + \max\{\sigma_\tau, b_\tau\} \right) \sqrt{\log\left(\frac{n}{\delta}\right)}, \tag{11}$$

$$\textit{bounded} \qquad \mathcal{C}(M, n, \delta) = T_{\max} + (T_{\max} - T_{\min}). \tag{12}$$

The upper bound is a direct generalization of the bounds 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. In the next section, we discuss how these bounds 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. From SMDPs to MDPs with Options

Let M be an MDP and $\mathcal{O}$ a set of options and let M' be the corresponding SMDP.[4] 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. Finally, 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 $\mathfrak{A}'$ applied to M' 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 of option $N(t)$. We can thus compare the performance of UCRL and UCRL-SMDP when learning in the MDP 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.

Theorem 7 *Let M be an MDP, $\mathcal{O}$ a set of options on M and M' the corresponding SMDP. Let π' be any stationary deterministic policy on M' and π the equivalent policy on M (not necessarily stationary). For any state $s \in \mathcal{S}'$, any learning algorithm $\mathfrak{A}$, and any number of decision steps n we have $\rho^{\pi'}(M', s) = \rho^{\pi}(M, s)$ and*

$$\Delta(M, \mathfrak{A}, s, T_n) = \Delta(M', \mathfrak{A}, s, n) + T_n \left(\rho^*(M) - \rho^*(M')\right). \tag{13}$$

The linear term in the MDP regret is due to the fact that the introduction of options amounts to constraining the space of policies that can be expressed in M and in general $\rho^*(M) \geq \rho^*(M')$. Thm. 7 also guarantees that the optimal policy computed in the SMDP M' (i.e., the policy maximizing $\rho^{\pi}(M', s)$) is indeed the best in the subset of policies that can be expressed in M by using the set of options $\mathcal{O}$. In order to use the regret analysis of Thm. 6, we still need to show that Asm. 1 is verified.

Lemma 8 *An MDP provided with a set of options is an SMDP where the holding times and rewards are sub-Exponential. Moreover, the holding time of an option is sub-Gaussian if and only if it is almost surely bounded.*

As an immediate consequence of this lemma, we have that $\tau(s, a, s')$ and $r(s, a, s')$ are either bounded or sub-Exponential, which are the two cases considered in Thm. 6. We are now ready to proceed with the comparison of the bounds on the regret of learning with options and primitive actions. We first notice that $R'_{\max} = R_{\max}$ and since $\mathcal{S}' \subset \mathcal{S}$ we have that $S' \leq S$. Furthermore, we introduce the following simplifying conditions: **1)** $\rho^*(M) = \rho^*(M')$ (i.e., the options do not prevent from learning the optimal policy), **2)** $A' \leq A$ (i.e., the number of options is not larger than the number of primitive actions), **3)** $D' = D(M') \leq D(M) = D$ (i.e., the diameter is not increased), **4)** options have bounded holding time (case 2 in Asm. 1). Let $\mathcal{R}(M, n, \delta)$ be the ratio between the regret upper bounds of UCRL-SMDP and UCRL respectively. Then we have (up to numerical constants)[5]

$$\mathcal{R}(M, n, \delta) \leq \left(1 + \frac{T_{\max}}{D\sqrt{S}}\right) \sqrt{\frac{n \log\left(\frac{n}{\delta}\right)}{T_n \log\left(\frac{T_n}{\delta}\right)}} \leq \frac{1}{\sqrt{\tau_{\min}}} \left(1 + \frac{T_{\max}}{D\sqrt{S}}\right),$$

4. We use the *prime* notation to denote SMDP quantities and distinguish them from those related to the MDP with primitive actions.
5. We recall that $\Delta(M, \mathrm{UCRL}, s, T_n) = O(DSR_{\max}\sqrt{AT_n})$.

where we used the fact that $\liminf_{n\to+\infty} \frac{T_n}{n} \geq \tau_{\min}$. If $T_{\max} \leq D\sqrt{S}$ and $\tau_{\min} > 4$ then the above ratio is strictly smaller than 1, thus suggesting that learning with options leads to a smaller regret than learning with primitive actions. While these conditions may look restrictive, in the following we show that it is relatively easy to construct an MDP and a set of options for which they are verified.[6]

We consider the navigation problem in Fig. 2. In any of the d^2 states of the grid except the target, the four cardinal actions are available, each of them being successful with probability 1. If the agent hits a wall then it stays in its current position with probability 1. When the target state is reached, the state is reset to any other state with uniform probability. The reward of any transition is 0 except when the agent leaves the target in which case it equals $R_{\max}$. The optimal policy simply takes the shortest path from any state to the target state. The diameter of the MDP is the longest shortest path in the grid, that is $D = 2d - 2$.

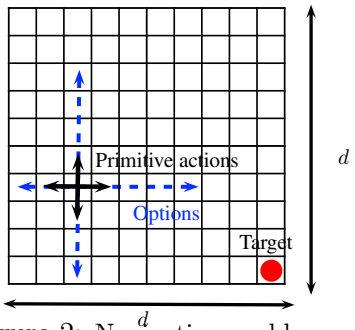

Figure 2: Navigation problem.

Let $d \geq m \geq 1$ be any non-negative integer smaller than d, in every state but the target, we define four macro-actions: *LEFT*, *RIGHT*, *UP* and *DOWN* (blue arrows in the figure). When *LEFT* is taken, primitive action *left* is applied up to m times. For any state s' which is $k \leq m$ steps on the left of the starting state s, we set $\beta_o(s') = 1/(m-k+1)$ (except if a wall is hit) so that the probability of the option to be interrupted after any $k \leq m$ steps is $1/m$. The SMDP formed with this set of options preserves the number of state-action pairs ($S' = S = d^2$ and $A' = A = 4$) while it slightly increases the diameter: $D = 2(d-1)$ and $D' = D + m(m+1)$ (see Appendix E). Thus, the two problems seem to be somehow as hard to learn. However the ratio between the regret upper bounds becomes

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

where we assume $m,\ d \geq 2$. The term $1 + \frac{2m^2}{d}$ can be made arbitrarily close to 1 by increasing the dimension d of the grid. $\tau_{\min} = 1$ in this case because when LEFT is taken in one of the leftmost states for example, the option ends after only one time step. However, for any of the four options $\bar{\tau}(s,o) = (m+1)/2$ in all but md out of d^2 states. The ratio n/T_n asymptotically tends to $2/(m+1)$ as n and d grow. Therefore the term $\limsup_{n\to+\infty}\sqrt{\frac{n}{T_n}}$ can be made arbitrarily small by increasing d and m, thus obtaining a ratio smaller than 1.

Conclusion. Despite its simplicity, the most interesting aspect of this example is that the improvement on the regret is not obtained by reducing the number of state-action pairs, but it is intrinsic in the way options changes the dynamics of the exploration process. 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 is limited by the use of SMDPs which 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.

6. Notice that while conditions **1)**, **3)** and **4)** are indeed in favor of UCRL-SMDP, but S', A', and $T_{\max}$ are in general much smaller than S, A, $D\sqrt{S}$. Furthermore, $\tau_{\min}$ is a very loose upper-bound on $\liminf_{n\to+\infty} \frac{T_n}{n}$ and in practice the ratio $\frac{T_n}{n}$ can take much larger values if $\tau_{\max}$ is large and many options have a high expected holding time. As a result, the set of MDPs and options on which the regret comparison is in favor of UCRL-SMDP is much wider.

References

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

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

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

Input Modeling with Phase-Type Distributions and Markov Models, chapter Chapter 2: Phase-Type Distributions. Springer, 2014.

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

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

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

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1756006.1859902`.

Kfir Y. Levy and Nahum Shimkin. *Unified Inter and Intra Options Learning Using Policy Gradient Methods*, pages 153–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-29946-9. doi: 10.1007/978-3-642-29946-9_17. URL `http://dx.doi.org/10.1007/978-3-642-29946-9_17`.

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

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

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

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

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

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181 – 211, 1999. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/S0004-3702(99)00052-1. URL http://www.sciencedirect.com/science/article/pii/S0004370299000521.

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

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

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

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

$$N_k(s,a) := \#\{\iota < i_k : s_\iota = s, a_\iota = a\} \tag{3}$$

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

$$\begin{aligned} R_k(s,a) := \sum_{\iota=1}^{i_k-1} r_\iota \mathbb{1}_{s_\iota=s, a_\iota=a} \qquad T_k(s,a) := \sum_{\iota=1}^{i_k-1} \tau_\iota \mathbb{1}_{s_\iota=s, a_\iota=a} \\ P_k(s,a,s') := \#\{\iota < i_k : s_\iota = s, a_\iota = a, s_{\iota+1} = s'\} \end{aligned} \tag{4}$$

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

(a) if $N_k(s,a) = 0$: $\hat{\tau}_k(s,a) := \tau_{\max}$ and $\hat{r}_k(s,a) := \tau_{\max} R_{\max}$
(b) if $N_k(s,a) > 0$: $\hat{\tau}_k(s,a) := \frac{T_k(s,a)}{N_k(s,a)}$ and $\hat{r}_k(s,a) := \frac{R_k(s,a)}{N_k(s,a)}$

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

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

$$\begin{aligned} |\tilde{r}(s,a) - \hat{r}_k(s,a)| \le \beta_k^r(s,a) \quad \text{and} \quad R_{\max}\tau_{\max} \ge \tilde{r}(s,a) \ge 0 \\ |\tilde{\tau}(s,a) - \hat{\tau}_k(s,a)| \le \beta_k^\tau(s,a) \quad \text{and} \quad \tau_{\max} \ge \tilde{\tau}(s,a) \ge \tau_{\min} \\ \|\tilde{p}(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a)\|_1 \le \beta_k^p(s,a) \quad \text{and} \quad \sum_{s' \in \mathcal{S}} \tilde{p}(s' \mid s,a) = 1 \end{aligned} \tag{5}$$

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:

$$\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \ge \max_{M' \in \mathcal{M}_k, \pi, s} \rho(M', \pi, s) - \frac{1}{\sqrt{i_k}} \tag{6}$$

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

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

Appendix A. Optimal average reward in SMDPs: existence and computation

The goal of this section is to prove Proposition 2 and Theorem 4. We will address the case of finite and continuous action spaces in parallel.

A.1 Optimality equations of a communicating SMDP

Let's begin with the average reward optimality equations for a communicating SMDP:

$$\forall s \in \mathcal{S},\ u^*(s) = \max_{a \in \mathcal{A}_s} \left\{ \bar{r}(s,a) - \rho^* \bar{\tau}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) u^*(s') \right\} \tag{14}$$

where u^* and ρ^* are the unknowns. Since we need to analyse both the case where $\mathcal{A}_s$ is finite and the case where $\mathcal{A}_s$ is continuous, one might be tempted to think that max should be replaced by sup in equation 14. For the original SMDP M, $\mathcal{A}_s$ is finite and the maximum is well-defined. For the extended SMDPs $\tilde{M}_k$, $\tilde{\mathcal{A}}_s$ is compact and $\bar{r}(s,a)$, $\bar{\tau}(s,a)$ and $p(\cdot \mid s,a)$ are continuous in $\tilde{\mathcal{A}}_s$ by the very definition of $\tilde{M}_k$. The function $\bar{r}(s,a) - \rho^* \bar{\tau}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) u^*(s')$ is thus continuous on $\tilde{\mathcal{A}}_s$ compact and by Weierstrass theorem, we know that the maximum is reached (i.e., there exists a maximizer). As a result, equation 14 is well-defined and we can study the existence and properties of its solutions.

A.2 Existence of a solution of the optimality equations

To prove existence of a solution of 14, we analyse the MDP M' obtained from the communicating SMDP M after data transformation (aka uniformization, see equation 7). According to (Puterman, 1994), the average optimality equations of the transformed MDP are:

$$\forall s \in \mathcal{S},\ v^*(s) = \max_{a \in \mathcal{A}_s} \left\{ \frac{\bar{r}(s,a)}{\bar{\tau}(s,a)} - g^* + \frac{\tau}{\bar{\tau}(s,a)} \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') + \left(1 - \frac{\tau}{\bar{\tau}(s,a)}\right) v^*(s) \right\} \tag{15}$$

As in equation 14, the max in equation 15 is well-defined. Note also that since $\tau < \tau_{\min}$, every Markov Chain induced by a stationary deterministic policy on M' is necessarily aperiodic (for any action, the probability of any state to loop on itself is non-negative). Moreover, since M was assumed to be communicating, M' is also communicating. Note that the same holds for $\tilde{M}'_k$ (MDP obtained from the extended SMDP $\tilde{M}^+_k$ in section 3 after data transformation). It is well know in the literature that under these conditions, equation 15 has a solution (v^*, g^*) where g^* is the optimal average reward of M' (respectively $\tilde{M}'_k$) and the (stationary deterministic) greedy policy w.r.t. v^* is average-optimal. Moreover, Value Iteration converges and can be applied with the stopping condition of section 3 to obtain an ϵ-optimal policy in finitely many steps. This holds for both finite and compact $\mathcal{A}_s$ with continuous $\bar{r}'(s,a)$ and $p'(s'|s,a)$ (see for example (Puterman, 1994) and (Lei, 2013)). EVI 8 is exactly Value Iteration applied to $\tilde{M}'_k$ so the stopping condition is reached in finite time. Finally, Lemma 2 of (Federgruen et al., 1983) shows the "equivalence" between M and M' (respectively $\tilde{M}^+_k$ and $\tilde{M}'_k$): if (v^*, g^*) is a solution to 15 then $(\tau^{-1} v^*, g^*)$ is a solution to 14 and conversely. As a result, there exists a solution (u^*, ρ^*) to 14 for both M' and $\tilde{M}'_k$.

A.3 Existence of deterministic stationary optimal policies

We are now ready to prove the existence of an optimal stationary deterministic stationary policy. Moreover, the optimal value is constant. We denote by Π_M^{HR} the set of (history-dependent randomized) policies and Π_M^{SD} the set of stationary deterministic policies.

For M (finite $\mathcal{A}_s$):
Conditions (L), (F) and (R) of (Schäl, 1992) hold and by their main Theorem:

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

where (u^*, ρ^*) is a solution of 14.
It is known from renewal theory that: $\forall \pi \in \Pi_M^{SD},\ \ \overline{\rho}^{\pi} = \underline{\rho}^{\pi} = \rho^{\pi}$ (in other words, the limit exists for deterministic stationary policies: see (Tij, 2003) and (Ros, 1970)), so π^* is optimal. Proposition 2 is thus proved.

For $\tilde{M}_k^+$ (compact $\tilde{\mathcal{A}}_s$ 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 14 admits a solution (u^*, ρ^*), see above). Because the state space is finite, the rest of the proof is rigorously the same.

A.4 Convergence of EVI

We have seen 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 ϵ-optimal in $\tilde{M}_k'$. Finally, for any stationary deterministic policy $\pi \in \Pi_{\tilde{M}_k^+}^{SD}$, the average reward is the same in the SMDP and the MDP obtained by uniformization: $\forall s \in \mathcal{S}, \rho^{\pi}(\tilde{M}_k^+) = \rho^{\pi}(\tilde{M}_k')$ (see (Tij, 2003)). So the policy returned by EVI is ϵ-optimal in $\tilde{M}_k^+$.

Appendix B. Distribution of the holding time and reward of a finite Markov option (proof of Lemma 8)

We begin with the definition of sub-Exponential and sub-Gaussian random variables. We denote by $\mathbb{R}^+$ and $\mathbb{R}^{+*}$ the set of positive and non-negative reals respectively.

Definition 9 (Wai (2015)) *A random variable X with mean $\mu < +\infty$ is said to be sub-Exponential if one of the following equivalent conditions is satisfied:*

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

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

We will denote: $X \in \text{subExp}(\sigma, b)$.

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

Definition 10 (Wai (2015)) *A random variable X with mean $\mu < +\infty$ is said to be sub-Gaussian if and only if there exists $\sigma \in \mathbb{R}^+$ such that:*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{for all} \quad \lambda \in \mathbb{R}. \tag{17}$$

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

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

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

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

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

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

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

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

and otherwise: $\mathbb{E}\left[e^{\lambda(\tau-\bar{\tau})}\right] = +\infty$. 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:

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

1. $\rho(Q) = 0$ if and only if all the eigenvalues of Q in $\mathbb{C}$ are 0, if and only if Q is nilpotent ($\exists n > 0$ s.t. $Q^n = 0$). This is because Q can always be triangularized in $\mathbb{C}$: $Q = UTU^{-1}$ where T is upper-triangular with the eigenvalues of Q on the diagonal that is, only zeros if $\rho(Q) = 0$. This implies that $\exists n > 0$ s.t. $T^n = U^{-1}Q^nU = 0 \implies Q^n = 0$ hence Q is nilpotent. The reverse is obviously true: if Q is nilpotent then $\rho(Q) = 0$, (otherwise there would exist $\lambda \neq 0$, $v \neq 0$ and $n > 0$ s.t. $Q^n = 0$ and $Qv = \lambda v \implies Q^nv = \lambda^n v = 0$, 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 $\mathbb{E}\left[e^{\lambda(\tau-\bar{\tau})}\right]$ can not be bounded by a function of the form $\lambda \to e^{\frac{\sigma^2\lambda^2}{2}}$ for $\lambda \geq -\log(\rho(Q))$ so $\tau(s,o,s')$ is not sub-Gaussian (Definition 10). However, since $\rho(Q) < 1$ we can choose $0 < c_0 < -\log(\rho(Q))$ and we have $\mathbb{E}\left[e^{\lambda(\tau-\bar{\tau})}\right] < +\infty$ for all $|\lambda| < c_0$, which implies that $\tau(s,o,s')$ is sub-Gaussian (Definition 9).

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

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

$$\begin{aligned}
\forall \lambda > 0, \quad \mathbb{E}\left[e^{\lambda(r-\bar{r})}\right] &= \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(r-\bar{r})}\middle|\tau = k\right]\mathbb{P}(\tau = k) \leq \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(R_{\max}\tau-\bar{r})}\middle|\tau = k\right]\mathbb{P}(\tau = k) \\
&= \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(R_{\max}k-\bar{r})}\middle|\tau = k\right]\mathbb{P}(\tau = k) \\
&= \sum_{k=1}^{\infty} e^{\lambda(R_{\max}k-\bar{r})}\mathbb{P}(\tau = k) \\
&= e^{\lambda(R_{\max}-\bar{\tau})} e_s^\intercal \left[\sum_{k=0}^{\infty}\left(e^{\lambda R_{\max}}Q\right)^k\right] R e_{s'}
\end{aligned}$$

We can now conclude as we did for $\tau(s,o,s')$: let $0 < c_0 < -\frac{\log(\rho(Q))}{R_{\max}}$, for all $0 < \lambda < c_0$ the quantity $\mathbb{E}\left[e^{\lambda(r-\bar{r})}\right]$ is finite. Note that for $\lambda \leq 0$: $\mathbb{E}\left[e^{\lambda r}\right] \leq 1$ so $\mathbb{E}\left[e^{\lambda(r-\bar{r})}\right] < +\infty$. By Definition 9, $r(s,o,s')$ is sub-Exponential.

Appendix C. Analysis of SMDP-UCRL (proof of Theorem 6)

To prove our bound on the regret we follow the proof of (Jaksch et al., 2010) for MDP-UCRL. Therefore, we will 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.

C.1 Splitting into Episodes

We begin with the Bernstein concentration inequality for sub-Exponential random variables:

Theorem 11 (Bernstein inequality, (Wai, 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:*

$$\begin{aligned} \forall t \geq 0,\ \mathbb{P}\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i \geq t\right) &\leq \begin{cases} e^{-\frac{t^2}{2n\sigma^2}}, & \text{if } 0 \leq t \leq \frac{\sigma^2}{b} \\ e^{-\frac{t}{2b}}, & \text{if } t > \frac{\sigma^2}{b} \end{cases} \\ \mathbb{P}\left(\sum_{i=1}^n X_i + \sum_{i=1}^n \mu_i \leq t\right) &\leq \begin{cases} e^{-\frac{t^2}{2n\sigma^2}}, & \text{if } 0 \leq t \leq \frac{\sigma^2}{b} \\ e^{-\frac{t}{2b}}, & \text{if } t > \frac{\sigma^2}{b} \end{cases} \end{aligned} \tag{18}$$

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

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

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

Conditionally on knowing $(N(s,a))_{s,a}$, the previous sum is equal (in distribution) to a sum of independent random variables with mean $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s} N(s,a)\bar{r}(s,a)$ and from Theorem 11 we have:

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

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

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

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

Similarly, the total holding time satisfies:

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

Lemma 12 *The optimal average reward can be bounded as follows:*

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

Proof It is proved in Appendix A that $\rho^*(M) = \rho^*(M')$ where $\rho^*(M')$ is the optimal average reward of an MDP M' with same state and action spaces as SMDP M and with average rewards of the form $\frac{\bar{r}(s,a)}{\bar{\tau}(s,a)}$. All the rewards of M' are thus bounded by $R_{\max}$ and so $\rho^*(M')$ is necessarily bounded by $R_{\max}$ as well and thus: $\rho^*(M) \leq R_{\max}$. ■

We define the per-episode regret by: $\Delta_k = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s,a)\left(\bar{\tau}(s,a)\rho^* - \bar{r}(s,a)\right)$.
Setting $\gamma_r = \max\left\{\frac{5}{2}b_r, \sqrt{\frac{5}{2}}\sigma_r\right\}$ and $\gamma_\tau = \max\left\{\frac{5}{2}b_\tau, \sqrt{\frac{5}{2}}\sigma_\tau\right\}$, and using a union bound on the previous inequalities we have that with probability at least $1 - \frac{\delta}{12n^{5/4}}$:

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

C.2 Dealing with Failing Confidence Regions

Lemma 13 *For any episode $k \geq 1$, the probability that the true SMDP M is not contained in the set of plausible MDPs $\mathcal{M}_k$ at step i is at most $\frac{\delta}{15i_k^6}$, that is:*

$$\forall k \geq 1,\ \mathbb{P}\left(M \notin \mathcal{M}_k\right) < \frac{\delta}{15i_k^6} \tag{19}$$

Proof This lemma is the SMDP-analogue of Lemma 17 in (Jaksch et al., 2010) and the proof is similar. Using a well-known L^1-concentration inequality for discrete probability

distributions we obtain:

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

In the above inequalities, it is implicitly assumed that the value $N_k(s,a)=n$ is fixed. To be more rigorous, we are bounding the probability of the intersection of event $\{\big\|\tilde{p}(\cdot|s,a)-\hat{p}_k(\cdot|s,a)\big\|_1 \geq \beta_k^p(s,a)\}$ with event $\{N_k(s,a)=n\}$ but we omitted the latter to simplify notations, and we will also omit it in the next inequalities.
Using Bernstein inequality (Theorem 11) and noting that $240 \leq 2^7\left(\frac{SA}{\delta}\right)^6$ for $S,A\geq 2$ and $\delta \leq 1$, we have:

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

$$\begin{aligned}
\mathbb{P}\Big(|\bar{r}(s,a)-\hat{r}_k(s,a)| \geq \beta_k^r(s,a)\Big) &= \mathbb{P}\left(|\bar{r}(s,a)-\hat{r}_k(s,a)| \geq \sigma_r\sqrt{\frac{14}{n}\log\left(\frac{2SAi_k}{\delta}\right)}\right)\\
&\leq \mathbb{P}\left(|\bar{r}(s,a)-\hat{r}_k(s,a)| \geq \sigma_r\sqrt{\frac{2}{n}\log\left(\frac{240SAi_k^7}{\delta}\right)}\right)\\
&\leq 2\exp\left(-\frac{n}{2\sigma_r^2}\times\frac{2}{n}\sigma_r^2\log\left(\frac{240SAi_k^7}{\delta}\right)\right)\\
&= \frac{\delta}{120i_k^7 SA}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}\Big(|\bar{r}(s,a)-\hat{r}_k(s,a)| \geq \beta_k^r(s,a)\Big) &= \mathbb{P}\left(|\bar{r}(s,a)-\hat{r}_k(s,a)| \geq \frac{14b_r}{n}\log\left(\frac{2SAi_k}{\delta}\right)\right)\\
&\leq \mathbb{P}\left(|\bar{r}(s,a)-\hat{r}_k(s,a)| \geq \frac{2b_r}{n}\log\left(\frac{240SAi_k^7}{\delta}\right)\right)\\
&\leq 2\exp\left(-\frac{n}{2b_r}\times\frac{2}{n}b_r\log\left(\frac{240SAi_k^7}{\delta}\right)\right)\\
&= \frac{\delta}{120i_k^7 SA}
\end{aligned}$$

Similarly for holding times we have:

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

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

$$\mathbb{P}\Big(|\bar{\tau}(s,a) - \hat{\tau}_k(s,a)| \geq \beta_k^\tau(s,a) \Big) \leq \frac{\delta}{120 i_k^6 SA}$$
$$\mathbb{P}\Big(|\bar{r}(s,a) - \hat{r}_k(s,a)| \geq \beta_k^r(s,a) \Big) \leq \frac{\delta}{120 i_k^6 SA}$$
$$\mathbb{P}\Big(\big\| p(\cdot|s,a) - \hat{p}_k(\cdot|s,a) \big\|_1 \geq \beta_k^p(s,a) \Big) \leq \frac{\delta}{20 i_k^6 SA}$$

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

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

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} N_k(s,a) = i_k - 1 \tag{20}$$

We can thus bound this part of the regret:

$$\begin{aligned}
\sum_{k=1}^m \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} &\leq \sum_{k=1}^m \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \bar{\tau}(s,a) \rho^* \mathbb{1}_{M\notin\mathcal{M}_k} \\
&\leq \tau_{\max}\rho^* \sum_{k=1}^m i_k \mathbb{1}_{M\notin\mathcal{M}_k} = \tau_{\max}\rho^* \sum_{i=1}^n i \sum_{k=1}^m \mathbb{1}_{i=i_k, M\notin\mathcal{M}_k} \\
&\leq \tau_{\max}\rho^* \left(\sum_{i=1}^{\lfloor n^{1/4} \rfloor} i + \sum_{i=\lfloor n^{1/4}\rfloor+1}^n i \sum_{k=1}^m \mathbb{1}_{i=i_k, M\notin\mathcal{M}_k} \right) \\
&\leq \tau_{\max}\rho^* \left(\sqrt{n} + \sum_{i=\lfloor n^{1/4}\rfloor+1}^n i \sum_{k=1}^m \mathbb{1}_{i=i_k, M\notin\mathcal{M}_k} \right)
\end{aligned}$$

where we defined: $\tau_{\max} = \max_{s,a} \bar{\tau}(s,a) < +\infty$.
By Lemma 13, the probability that the second term in the right hand side of the above inequality is stricly greater than 0 is bounded by:

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

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

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

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

Now we assume that $M \in \mathcal{M}_k$ and we start by analysing the regret of a single episode k. By construction, $\tilde{\rho}_k \geq \rho^* - \frac{1}{\sqrt{i_k}}$ hence:

$$\begin{aligned}\Delta_k = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \left(\bar{\tau}(s,a)\rho^* - \bar{r}(s,a)\right) &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \left(\tilde{\tau}_k(s,a)\rho^* - \bar{r}(s,a)\right) \\ &\quad + \rho^* \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \left(\bar{\tau}(s,a) - \tilde{\tau}_k(s,a)\right)\end{aligned}$$

$$\begin{aligned}\implies \Delta_k \leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \left(\tilde{\tau}_k(s,a)\tilde{\rho}_k - \bar{r}(s,a)\right) &+ \rho^* \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \nu_k(s,a) \left(\bar{\tau}(s,a) - \tilde{\tau}_k(s,a)\right) \\ &+ \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \frac{\nu_k(s,a)}{\sqrt{i_k}} \tilde{\tau}_k(s,a)\end{aligned}$$

Lemma 14 *At any iteration $i \geq 0$ of EVI (Extended Value Iteration), the range of the state values is bounded as follows:*

$$\forall i \geq 0, \quad \max_{s\in\mathcal{S}} u_i(s) - \min_{s\in\mathcal{S}} u_i(s) \leq \frac{R_{\max} D(M)}{\tau} \tag{21}$$

Proof In Appendix A it is shown that EVI is in fact a Value Iteration algorithm applied to an MDP $\tilde{M}'_k$ obtained by "uniformizing" SMDP $\tilde{M}^+_k$. Using the same argument as in section 4.3.1 of (Jaksch et al., 2010), we have that: $\forall i \geq 0,\ \max_{s\in\mathcal{S}} u_i(s) - \min_{s\in\mathcal{S}} u_i(s) \leq R_{\max} D(M')$. We thus only need to find a relationship between $D(M)$ and $D(M')$ (M' is the MDP obtained after applying a data transformation on SMDP M). Let $T(s')$ denote the first time at which state s' is reached in M or M':

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

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

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

2. Conversely: $\mathbb{P}_M^\pi\big(T(s')=+\infty|s_0=s\big)=0 \implies \mathbb{P}_{M'}^\pi\big(T(s')=+\infty|s_0=s\big)=0$ in which case both expectations are finite. To prove they are equal up to factor τ, we see the holding time as a "reward" (the true rewards are ignored here). Note that π induces Markov Chains with different dynamics on M and M' (different transition probabilities). We call these Markov Chains MC and MC' 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 $(\bar{s},\pi(\bar{s}))$ have an expected reward equal to $\bar{\tau}(\bar{s},\pi(\bar{s}))$ except $(s',\pi(s'))$ which has a reward of zero. The total expected reward of this Markov Reward Process (MRP denoted MR) starting from s trivially equals $\mathbb{E}_M^\pi\big[T(s')|s_0=s\big]$. Since we assumed that $\mathbb{E}_M^\pi\big[T(s')|s_0=s\big]$ 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 $\bar{s}$ 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 $\big(\overline{T}(\bar{s})\big)_{\bar{s}\in\mathcal{S}}=\big(\mathbb{E}_M^\pi\big[T(s')|s_0=\bar{s}\big]\big)_{\bar{s}\in\mathcal{S}}$ is the unique solution to the system of equations:

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

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

$$\begin{aligned}&\forall \bar{s},\ \overline{T}'(\bar{s})=1+\frac{\tau}{\bar{\tau}(\bar{s},d(\bar{s}))}\sum_{\tilde{s}}p(\tilde{s}|\bar{s},d(\bar{s}))\overline{T}'(\tilde{s})+\left(1-\frac{\tau}{\bar{\tau}(\bar{s},d(\bar{s}))}\right)\overline{T}'(\bar{s})\\ \iff &\forall \bar{s},\ \Big(\tau\overline{T}'(\bar{s})\Big)=\bar{\tau}(\bar{s},d(\bar{s}))+\sum_{\tilde{s}}p(\tilde{s}|\bar{s},d(\bar{s}))\Big(\tau\overline{T}'(\tilde{s})\Big)\end{aligned}$$

By uniqueness of the solution: $\tau\overline{T}'=\overline{T} \implies \tau D(M')=D(M)$.

■

Lemma 15 *If the convergence criterion of EVI hold at iteration i, then:*

$$\forall s\in\mathcal{S},\quad \big|u_{i+1}(s)-u_i(s)-\tilde{\rho}_k\big|\leq\frac{1}{\sqrt{i_k}} \tag{22}$$

Proof Let's define the following notations:

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

Since EVI is just Value Iteration applied to MDP M'_k, Theorem 8.5.6 of (Puterman, 1994) hold and we have:

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

In conclusion:

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

■

Based on Lemma 15 and equation 22 we have:

$$\forall s \in \mathcal{S}, \ \left| \left(\tilde{\rho}_k - \frac{\tilde{r}_k(s, \tilde{\pi}_k(s))}{\tilde{\tau}_k(s, \tilde{\pi}_k(s))} \right) - \left(\sum_{s' \in \mathcal{S}} \tilde{p}_k(s'|s, \tilde{\pi}_k(s)) u_i(s') - u_i(s) \right) \frac{\tau}{\tilde{\tau}_k(s, \tilde{\pi}_k(s))} \right| \leq \frac{1}{\sqrt{i_k}} \tag{23}$$

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

$$\begin{aligned}
\Delta_k &\leq \sum_{s,a} \nu_k(s,a) \left(\tilde{\tau}_k(s,a)\tilde{\rho}_k - \bar{r}(s,a)\right) + \rho^* \sum_{s,a} \nu_k(s,a) \left(\bar{\tau}(s,a) - \tilde{\tau}_k(s,a)\right) + \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{i_k}} \tilde{\tau}_k(s,a) \\
&= \sum_{s,a} \nu_k(s,a) \left(\tilde{\tau}_k(s,a)\tilde{\rho}_k - \tilde{r}_k(s,a)\right) + \sum_{s,a} \nu_k(s,a) \left(\tilde{r}_k(s,a) - \bar{r}(s,a)\right) + \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{i_k}} \tilde{\tau}_k(s,a) \\
&\qquad + \rho^* \sum_{s,a} \nu_k(s,a) \left(\bar{\tau}(s,a) - \tilde{\tau}_k(s,a)\right)
\end{aligned}$$

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

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

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

and symmetrically:

$$\tilde{\tau}_k(s,a) - \bar{\tau}(s,a) \leq 2c_\tau \frac{\log(2SAi_k/\delta)}{\sqrt{\max\{1, N_k(s,a)\}}} \implies \tilde{\tau}_k(s,a) \leq \tau_{\max} + 2c_\tau \frac{\log(2SAi_k/\delta)}{\sqrt{\max\{1, N_k(s,a)\}}}$$

Finally, using 23 we obtain:

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

$$\implies \sum_{s,a}\nu_k(s,a)\left(\tilde{\tau}_k(s,a)\tilde{\rho}_k - \tilde{r}_k(s,a)\right) \leq \sum_{s,a}\nu_k(s,a)\left(\frac{\tau_{\max}}{\sqrt{i_k}} + 2c_\tau\frac{\log(2SAi_k/\delta)}{\sqrt{\max\{1,N_k(s,a)\}}}\right) + \tau\Big(v_k\big(\tilde{P}_k - I\big)u_i\Big)$$

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

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

and since the rows of $\tilde{P}_k$ sum to one, we have: $v_k\big(\tilde{P}_k - I\big)u_i = v_k\big(\tilde{P}_k - I\big)w_k$. Moreover, by Lemma 14: $\|w_k\|_\infty \leq \frac{R_{\max}D}{2\tau}$. Noting that $\max\{1,N_k(s,a)\} \leq i_k \leq n$ we get:

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

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

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

Lemma 16 *Consider a sequence of positive reals $(z_k)_k$ and define: $\forall k,\ Z_k = \max\{1, \sum_{i=1}^{k} z_k\}$. Assuming that $0 \leq z_k \leq Z_k$ we have:*

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

Proof See Appendix C.3 of (Jaksch et al., 2010). ■

Using Lemma 16 we get:

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

By Jensen's inequality we thus have:

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

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

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

C.4 Computing the final bound

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix D. Equivalent policies in an MDP with options and the induced SMDP (proof of Theorem 7)

By definition of M', the reward of an option is equal to the sum of the rewards of all the primitive actions taken until the option ends (when the option is executed in M). Therefore $\sum_{i=1}^{n} r'_i = \sum_{i=1}^{N(T_n)} r'_i = \sum_{t=1}^{T_n} r_t$ and:

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

Let's now define:

$$\forall T \in \mathbb{N}^*, \quad \rho^{\pi}(M, s, T) = \mathbb{E}_M^{\pi}\left[\frac{\sum_{t=1}^{T} r_t}{T}\middle| s_0 = s\right]$$
$$\forall T' \in \mathbb{R}^{+*}, \quad \rho^{\pi'}(M', s, T') = \mathbb{E}_{M'}^{\pi'}\left[\frac{\sum_{i=1}^{N(T')} r'_i}{T'}\middle| s_0 = s\right]$$

$\lim_{n\to+\infty} T_n = +\infty$ because 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, $\lim_{T'\to+\infty} \rho^{\pi'}(M', s, T')$ exists since π' is stationary and deterministic (see appendix A) and by composition of the limit we have:

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

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)). In the augmented MDP, policy π is stationary deterministic and we know from MDP theory (Puterman, 1994) that the corresponding average reward exists. We also have:

$$\forall n \geq 1, \quad \mathbb{E}_M^{\pi}\left[\frac{\sum_{t=1}^{T_n} r_t}{T_n}\middle| s_0 = s\right] = \mathbb{E}_{M'}^{\pi'}\left[\frac{\sum_{i=1}^{n} r'_i}{T_n}\middle| s_0 = s\right]$$
$$\implies \rho_1^{\pi'}(M', s) = \lim_{n\to+\infty} \rho^{\pi}(M, s, T_n) = \lim_{T\to+\infty} \rho^{\pi}(M, s, T) = \rho^{\pi}(M, s)$$

Appendix E. Details of the illustrative experiments

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

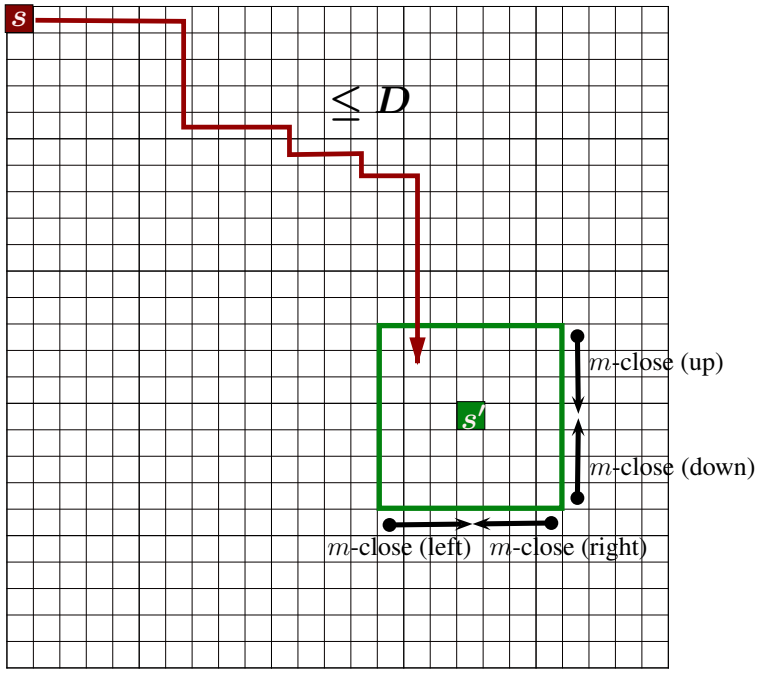

Figure 3: Upper bound on the diameter of the SMDP used in the experiments.

- If $k = 1$:

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

- If $k \geq 1$:

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

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

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

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

Diameter Let s and s' be two distinct states in the grid. With the options defined above, the expected shortest path from s to s' is obtained if in each visited state on the way to s', we choose an option that goes in the direction of s'. For example, if s is the state located in the top left corner of the grid and s' is the target, the expected shortest path is obtained when either *RIGHT* or *DOWN* is taken in every state. With this policy, the expected time to get m-close to s' both horizontally and vertically is trivially bounded by D (red path on Fig. 3). Once we are m-close to s' (green square on Fig. 3, $m = 3$ on this example), we will potentially start cycling until we reach s'. On Fig. 4, 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 options end after at most m time steps, once we are m-close to s', we stay m-close with the chosen policy. The expected time it takes to reach s' once we are m-close to it is $m(m+1)/2$ both horizontally and vertically. To prove this, we need to solve a linear system. For all $i \in \{1...m-1\}$, denote by τ_i the time it takes to go from s to the i-th state to the left (respectively right, up or down) when the option chosen is left (respectively right, up or down). The value is the same in all directions by symmetry. We can express the τ_i as follows:

$$
\begin{aligned}
\tau_1 &= \frac{1}{m} + \frac{1}{m}(2+\tau_1) + \cdots + \frac{1}{m}(m+\tau_{m-1}) \\
\tau_2 &= \frac{1}{m} \times 2 + \frac{1}{m}(1+\tau_1) + \frac{1}{m}(3+\tau_1) + \cdots + \frac{1}{m}(m+\tau_{m-2}) \\
\tau_3 &= \frac{1}{m} \times 3 + \frac{1}{m}(1+\tau_2) + \frac{1}{m}(2+\tau_1) + \frac{1}{m}(4+\tau_1) + \cdots + \frac{1}{m}(m+\tau_{m-3}) \\
&\cdots \\
\tau_i &= \frac{m+1}{2} + \frac{1}{m}\sum_{j=1}^{i-1}\tau_j + \frac{1}{m}\sum_{j=1}^{m-i}\tau_j
\end{aligned}
\tag{24}
$$

With probability $1/m$, the next state after executing the option is 1 step to the left of s and the value of τ_1 is then 1 . With probability $1/m$ the next state is 2 steps to the left of s and so s' is now located 1 step to the right of the new state: the value of τ_1 is thus $2 + \tau_1$. With probability $1/m$ the next state is 3 steps to the left of s and and so s' is now located 2 steps to the right of the new state: the value of τ_1 is thus $3 + \tau_2$. And so on and so forth. What we used here is basically the law of total expectations where the partition of events is the set of all possible states reached after executing the option only once. The same thing can be done for $\tau_2 \dots \tau_{m-1}$. It is trivial to verify that the only solution of the linear system in equation 24 is: $\tau_i = m(m+1)/2, \forall i \in \{1...m-1\}$. This results is rather intuitive: m corresponds to the expected number of times the option needs to be executed to end up in the desired state s' whereas $(m+1)/2$ is the expected duration at every decision step. The simplicity of this result comes from the symmetry of the problem: every time an option is executed, we stay m-close to s' and the probability to exactly reach s' is always $1/m$. So in this sense, we have i.i.d. Bernoulli trials where the probability of success is $1/m$. The expected time to reach s' when we start in an m-close state both horizontally and vertically is thus $2 \times m(m+1)/2 = m(m+1)$. Therefore, the expected time to go from s to s' is always bounded by $D + m(m+1)$.

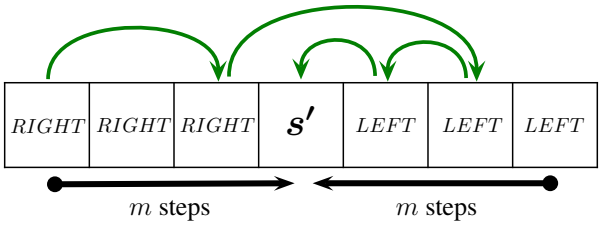

Figure 4: Behaviour of the agent in m-close states.

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

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

Theorem 17 (Martingale Strong Law of Large Numbers, (Vovk et al., 2005)) *Let $X_1, ..., X_n$ be a martingale difference sequence w.r.t. a filtration $\mathcal{F}_0, \mathcal{F}_1, ..., \mathcal{F}_n$ and let $A_1, ..., A_n$ be an increasing predictable sequence w.r.t. the same filtration with $A_1 > 0$ and $\lim_{n\to+\infty} A_n = +\infty$ almost surely. If:*

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

then:

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

Let's take $X_i = \tau_i - \overline{\tau}_i$ (where $\overline{\tau}_i = \overline{\tau}_i(s_{i-1}, a_i)$) and $\mathcal{F}_i = \sigma\left(s_0, a_1, \tau_1, r_1, ..., s_i, a_{i+1}\right)$. The sequence $(X_i)_{i\leq 1}$ is a martingale difference because $\mathbb{E}[X_i] < +\infty$ and $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$. Since $(\tau(s, a, s'))_{s,a,s'}$ are sub-Exponential, all moments are finite and it is well known from the literature that the variance is bounded by the sub-Exponential constant σ_τ^2 hence: $\mathbb{E}[X_i^2|\mathcal{F}_{i-1}] < \sigma_\tau^2$. If in addition we take $A_i = i$ then the conditions of Theorem 17 are satisfied and thus:

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

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

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

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

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

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

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

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

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

Note that $\tau_{\min}$ is a very loose upper-bound on $\liminf_{n\to+\infty} \frac{T_n}{n}$ and in practice the ratio $\frac{T_n}{n}$ can take much higher values if $\tau_{\max}$ is big and many options have a high expected holding time.