

A Related Works

There is a vast literature for provably efficient algorithms for FH-MDP. [Osband & Van Roy \(2016\)](#) proves the lower bound for the regret in the FH-MDP setting, $\Omega(\sqrt{HSAT})$. Then, many works propose algorithms with guarantees that nearly close the problem, i.e., with upper bounds of the same order as the lower bound ([Zanette & Brunskill, 2018](#)). [Azar et al. \(2017\)](#) definitively close the problem by proposing an innovative analysis of an algorithm for which the upper bound, $O(\sqrt{HSAT})$, matches the lower bound in all terms.

Nevertheless, only some works focused on theoretically understanding the benefits of hierarchical reinforcement learning approaches, and most of them consider a known set of pre-trained policies. In [Fruit & Lazaric \(2017\)](#), the authors propose an adaptation of UCRL2 ([Auer et al., 2008](#)) for SMDPs. This work was the first to theoretically compare options instead of primitive actions to learn in SMDPs. It provides both an upper bound for the regret suffered by their algorithm and a lower bound for the general problem. However, it focuses on the average reward setting to study how to possibly induce a more efficient exploration when using a set of fixed options. Differently, we aim to analyze the advantages of using options to reduce the sample complexity of the problem, resorting to the intuition that temporally extended actions can intrinsically reduce the planning horizon in FH-SMDPs, and characterize problems likely to benefit from using HRL even when no prior information about the problem is known, up to its structure. [Fruit et al. \(2017\)](#) is an extension of this work, where the need for prior knowledge of the distribution of cumulative reward and duration of each option is relaxed. However, the setting is identical. Furthermore, [Mann et al. \(2015\)](#) studies the convergence property of Fitted Value Iteration (FVI) using temporally extended actions, showing that a longer options duration and pessimistic value function estimates lead to faster convergence. [Wen et al. \(2020\)](#) demonstrate how patterns and substructures in the MDP provide benefits in terms of planning speed and statistical efficiency. They present a Bayesian approach that exploits this information, analyzing how sub-structure similarities and sub-problems' complexity contribute to the regret of their algorithm. A very recent approach proposed by [Robert et al. \(2024\)](#) studies the sample complexity of a particular sub-class of HRL approaches: the Goal-conditioned one, in which a goal-based problem is structured into a hierarchy of sub-tasks, each with its own sub-goal. They analyzed the best possible performance achievable by the best algorithm in the worst possible problem by adapting to this framework the lower bound on the sample complexity presented by [Dann & Brunskill \(2015\)](#). Nevertheless, this work is not completely related to our framework, which is more general than the goal-conditioned one.

The closest approach in the literature is [Drappo et al. \(2023\)](#). They propose to relax the assumption of having a set of pre-trained options by implementing an Explore-Then-Commit approach ([Lattimore & Szepesvári, 2020](#)), which first learns each options' policy and then exploits an adaptation of UCRL2 to FH-SMDPs ([Auer et al., 2008](#)) to find the optimal policy over options. Nevertheless, they sacrifice optimality to relax this assumption. Indeed, their approach suffer from the standard sub-optimality of Explore-Then-Commit approaches, having a regret scaling with $K^{2/3}$, and additionally is suboptimal in \sqrt{HS} being the high-level algorithm used in the second phase based on UCRL2. Therefore, our approach is the first in the literature able to relax the aforementioned assumption maintaining optimal guarantees.

B Proof of the regret of Options-UCBVI

In this section, we will present the analysis of the upper bound on the regret paid by Options-UCBVI. The analysis will adapt the one of UCBVI [Azar et al. \(2017\)](#) to the FH-SMDP for non-stationary transition models. For simplicity, we will write $o = \mu_k(s, h)$, and $P^{\mu_k}(s', h'|s, h) = P(s', h'|s, \mu_k(s), h)$.

Theorem 3.1. Let \mathcal{SM} be an FH-SMDP with S states and O temporally extended actions (options), known reward,⁶ bounded primitive reward $r^L(s, a, h) \in [0, 1]$. The regret suffered by algorithm Options-UCBVI in K episodes of horizon H is bounded, with probability $1 - \delta$, by:

$$\text{Regret}(\text{O-UCBVI}, K) \leq \tilde{O}\left(H\sqrt{SOKd} + H^3S^2Od + H\sqrt{Kd}\right),$$

where d is the average per-episode number of options played during the execution of the algorithm.

Proof. The Proof follows the same ideas as the proofs of UCBVI for the Bernstein-Freedman exploration bonus. We can write the regret as:

$$\text{Regret}(K) \leq \widetilde{\text{Regret}}(K) \leq \sum_{k=1}^K \tilde{V}^{\mu_k}(s, 1) - V^{\mu_k}(s, 1)$$

Where $\tilde{V}^{\mu_k}(s, 1)$ is the optimistic value function, and $V^{\mu_k}(s, 1)$, is the real value function considering the policy learned at the k^{th} step. Following the analysis of the original paper we can write the regret in terms of the per step regret $\tilde{\Delta}_{hk}(s_{hk})$. Thus,

$$\widetilde{\text{Regret}}(K) \leq \sum_{i=1}^K \sum_{j=1}^H \tilde{\Delta}_{ij}(s_{ij})$$

where the summation over H is composed of d terms, for the temporally extended transitions, where d is a random variable describing the expected number of options played in one episode, refer to the main paper for a more detailed explanation (Section 3).

Now let's define properly the per step regret:

$$\begin{aligned} \tilde{\Delta}_{hk}(s_{ij}) &= \tilde{V}^{\mu_k}(s_{hk}, h) - V^{\mu_k}(s_{hk}, h) \\ &\stackrel{a}{=} [\hat{P}_{hk}^{\mu_k} \tilde{V}^{\mu_k}(s', h')](s_{hk}) + b_{hk} - [P_h^{\mu_k} V^{\mu_k}(s', h')](s_{hk}) \pm [P_h^{\mu_k} \tilde{V}^{\mu_k}(s', h')](s_{hk}) \\ &= [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}) \tilde{V}^{\mu_k}(s', h')](s_{hk}) + b_{hk} + [P_h^{\mu_k} (\tilde{V}^{\mu_k}(s', h') - V^{\mu_k}(s', h'))](s_{hk}) \\ &\quad \pm [\Delta_p V^*(s', h')](s_{hk}) \\ &= [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}) (\tilde{V}^{\mu_k}(s', h') - V^*(s', h'))](s_{hk}) + b_{hk} + P_h^{\mu_k} \tilde{\Delta}_{h',k}(s_{hk}) \\ &\quad + [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}) V^*(s', h')](s_{hk}) \pm \tilde{\Delta}_{h',k}(s') \\ &\stackrel{b}{=} c_{hk} + b_{hk} + e_{hk} + \epsilon_{hk} + \tilde{\Delta}_{h',k}(s') \end{aligned}$$

- (a) By applying the bellman operator considering known reward that simplifies, and where $P_h^{\mu_k} = p(\cdot, \cdot | s_h, \mu_k(s_h), h)$, and $\hat{P}_{hk}^{\mu_k} = \hat{p}(\cdot, \cdot | s_{hk}, \mu_k(s_{hk}), h)$, the estimated transition model at episode k . By applying the bellman operator on the optimistic value function, the bonus term b_{hk} is added to the reward.
- (b) By defining $c_{hk} = [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}) (\tilde{V}^{\mu_k}(s', h') - V^*(s', h'))](s_{hk})$, the correction term, $e_{hk} = [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}) V^*(s', h')](s_{hk})$ the estimation error of the optimal value function, and ϵ_{hk} a martingale difference, defined as $\epsilon_{hk} = \mathcal{M}_t \tilde{\Delta}_{h',k}(s) = P_h^{\mu_k} \tilde{\Delta}_{h',k}(s) - \tilde{\Delta}_{h',k}(s')$, where \mathcal{M}_t is defined as a martingale operator (refer to appendix B.3 of Azar et al. (2017)).

Let us now bound each of these terms separately.

B.1 Bound of the correction term c_{hk}

In this subsection, we bound the correction term

$$\begin{aligned} c_{hk} &= [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}) (\tilde{V}^{\mu_k}(s', h') - V^*(s', h'))](s_{hk}) \\ &\stackrel{a}{=} \sum_{s' \in S} \sum_{h' \in H} (\hat{P}_k^{\mu_k}(s', h' | s_{hk}, h) - P^{\mu_k}(s', h' | s_{hk}, h)) (\tilde{V}^{\mu_k}(s', h') - V^*(s', h')) \end{aligned}$$

⁶The choice of assuming a known reward is for compliance with Azar et al. (2017). Nevertheless, learning the reward function is known to be a negligible task compared to learning the transition model of the environment and, consequently, will not alter the regret order.

$$\begin{aligned}
&\stackrel{b}{\leq} \sum_{s' \in S} \sum_{h' \in H} \left(2\sqrt{\frac{p_{hk}(s')(1-p_{hk}(s'))L}{n_k(s, o, h)}} + \frac{4L}{3n_k(s, o, h)} \right) \tilde{\Delta}_{h'k}(s') \\
&\stackrel{c}{\leq} 2\sqrt{L} \sum_{s' \in S} \sum_{h' \in H} \sqrt{\frac{p_{hk}(s')}{n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s') + \frac{4SH^2L}{3n_k(s, o, h)} \\
&\stackrel{d}{=} 2\sqrt{L} \left(\sum_{(s', h') \in [(s', h')]_{typ}} \sqrt{\frac{p_{hk}(s')}{n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s') \right. \\
&\quad \left. + \sum_{(s', h') \notin [(s', h')]_{typ}} \sqrt{\frac{p_{hk}(s')}{n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s') \right) + \frac{4SH^2L}{3n_k(s, o, h)} \\
&\stackrel{e}{=} 2\sqrt{L} \left(\sum_{(s', h') \in [(s', h')]_{typ}} P^{\mu_k}(s', h' | s_{hk}, h') \sqrt{\frac{1}{p_{hk}(s')n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s') \right. \\
&\quad \left. + \sum_{(s', h') \notin [(s', h')]_{typ}} \sqrt{\frac{p_{hk}(s')n_k(s, o, h)}{n_k(s, o, h)^2}} \tilde{\Delta}_{h'k}(s') \right) + \frac{4SH^2L}{3n_k(s, o, h)} \\
&\stackrel{f}{=} 2\sqrt{L} \left(\bar{\epsilon}_{hk} + \sqrt{\frac{1}{p_{hk}(s')n_k(s, o, h)}} \mathbb{I}((s', h') \in [(s', h')]_{typ}) \tilde{\Delta}_{h'k}(s') \right. \\
&\quad \left. + \sum_{(s', h') \notin [(s', h')]_{typ}} \sqrt{\frac{p_{hk}(s')n_k(s, o, h)}{n_k(s, o, h)^2}} \tilde{\Delta}_{h'k}(s') \right) + \frac{4SH^2L}{3n_k(s, o, h)} \\
&\stackrel{g}{\leq} 2\sqrt{L} \left(\bar{\epsilon}_{hk} + \sqrt{\frac{1}{4LH^2}} \tilde{\Delta}_{h'k}(s') + \frac{SH^2\sqrt{4LH^2}}{n_k(s, o, h)} \right) + \frac{4SH^2L}{3n_k(s, o, h)} \\
&\leq 2\sqrt{L}\bar{\epsilon}_{hk} + \frac{1}{H} \tilde{\Delta}_{h'k}(s') + \frac{4SH^3L}{n_k(s, o, h)} + \frac{4SH^2L}{3n_k(s, o, h)}
\end{aligned}$$

- (a) By considering, for brevity, $P^\mu(s', h' | s, h) = P(s', h' | s, \mu(s), h)$, and summing over all the possible next states and next stages.
- (b) Where for the first term we substitute the difference of transition probabilities with the relative confidence interval (refer to section B.4 on the appendix of [Azar et al. \(2017\)](#)), $|\hat{P}_k^{\mu_k}(s', h' | s_{hk}, h) - P^{\mu_k}(s', h' | s_{hk}, h)| \leq 2\sqrt{\frac{p_{hk}(s')(1-p_{hk}(s'))L}{n_k(s, o, h)}} + \frac{4L}{3n_k(s, o, h)}$, where $p_{hk}(s') = P^{\mu_k}(s', h' | s, h)$. Then we can bound $\tilde{V}^{\mu_k}(s', h') - V^*(s', h')$ with $\tilde{\Delta}_{h'k}(s')$ because $V^*(s', h') \geq V^{\mu_k}(s', h')$ (the true value function of the policy μ_k) by definition.
- (c) Because $(1 - p_{hk}(s')) \leq 1$ and $\tilde{\Delta}_{h'k}(s') \leq H$
- (d) We divide the summation over all the possible next state-stage, in the summation over the pairs contained in the typical pairs and the ones outside the set (the typical episodes are the episodes in which we have smaller regret; refer to the appendix of [Azar et al. \(2017\)](#)).
- (e) We multiply the first term by $\frac{p_{hk}(s')}{p_{hk}(s')}$, and the second by $\frac{n_k(s, o, h)}{n_k(s, o, h)}$.
- (f) We sum and subtract $\sqrt{\frac{\mathbb{I}((s', h') \in [(s', h')]_{typ})}{p_{hk}(s')n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s')$ and apply the martingale operator \mathcal{M} (see (b) in the previous proof). $\bar{\epsilon}_{hk} = P_h^{\mu_k} \sqrt{\frac{\mathbb{I}((s', h') \in [(s', h')]_{typ})}{p_{hk}(s')n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s') + \sqrt{\frac{\mathbb{I}((s', h') \in [(s', h')]_{typ})}{p_{hk}(s')n_k(s, o, h)}} \tilde{\Delta}_{h'k}(s')$.

- (g) For typical next state-stage pairs $n_k(s, o, h)P(s', h'|s, o, h) \geq 2H^2L$, where L is a logarithmic term (We kept the same lower bound of Azar et al. (2017)).

Now, before bounding the estimation error and the exploration bonus, let's rewrite the regret as

$$\begin{aligned} \widetilde{Regret}(K) &= \sum_{i=1}^K \tilde{\Delta}_{1i}(s_1) = \sum_{i=1}^K \sum_{j=1}^H \tilde{\Delta}_{ij}(s_{ij}) \\ &\leq \underbrace{\left(1 + \frac{1}{H}\right)^d}_{\leq e} \sum_{i=1}^K \sum_{j=1}^H \left(b_{hk} + e_{hk} + \epsilon_{hk} + 2\sqrt{L}\bar{\epsilon}_{hk} + \frac{4SH^3L}{n_k(s, o, h)} + \frac{4SH^2L}{3n_k(s, o, h)} \right) \end{aligned}$$

or otherwise omitting the last term which is dominated

$$\widetilde{Regret}(K) \leq \sum_{i=1}^K \sum_{j=1}^H \left(b_{hk} + e_{hk} + \epsilon_{hk} + 2\sqrt{L}\bar{\epsilon}_{hk} + \frac{4SH^3L}{n_k(s, o, h)} \right) \quad (7)$$

B.2 Bound of the estimation error e_{hk}

Let's consider just the typical episodes, the episodes for which the number of visits of state-option-stage pairs is larger than the rest of the episodes.

$$\begin{aligned} \sum_{k=1}^K \sum_{h=1}^H e_{hk} &= \sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) ([\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k}]V^*(s', h'))(s_{hk}) \\ &\stackrel{a}{\leq} \sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \left(2\sqrt{\frac{\mathbb{V}_{hk}^* L}{n_k(s_{hk}, o, h)}} + \frac{4HL}{3n_k(s, o, h)} \right) \\ &\stackrel{b}{\leq} 2\sqrt{L} \sqrt{\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}_{hk}^*} \sqrt{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \frac{1}{n_k(s, o, h)}} \\ &\quad + \sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \frac{4HL}{3n_k(s, o, h)} \\ &\stackrel{c}{\leq} 2\sqrt{L} \left(\sqrt{KH^2 + HdU_{K,1}} + \square \sqrt{H^5KL} + 4/3H^3L \right) \left(\sqrt{2SOdL} \right) + 4/3HSOdL^2 \\ &\stackrel{d}{\leq} \square LH \sqrt{KSOd} + \square Ld \sqrt{HSOU_{K,1}} \end{aligned}$$

- (a) Using Bernstein Inequality. $\mathbb{V}_{hk}^* = \text{Var}_{(s', h') \sim P^{\mu_k}(\cdot|s, h)}(V^*(s', h'))$ (Remember the meaning of P^{μ_k})
- (b) Using Cauchy-Schwartz inequality
- (c) Summing and subtracting $\mathbb{V}_{hk}^{\mu_k} = \text{Var}_{(s', h') \sim P^{\mu_k}(\cdot|s, h)}(V^{\mu_k}(s', h'))$ the variance of the next state-stage pair value function, inside the first square root, and then using Lemma D.2 and D.3. For the second square root and the additional term, we just use a pigeon-hole argument (Lemma D.1). We ignore the numerical constant represented as \square .
- (d) Because for typical episodes $K \geq H^2L^2S^2Od$ and thus we consider only the dominant terms.

B.3 Bound of the martingale differences ϵ_{hk} and $\bar{\epsilon}_{hk}$

$$\sum_{k=1}^K \sum_{h=1}^H \epsilon_{hk} \leq H\sqrt{dKL}$$

$$\sum_{k=1}^K \sum_{h=1}^H \bar{\epsilon}_{hk} \leq \sqrt{dK}$$

These results follow the same proofs of the original paper, thus considering the same event \mathcal{E} to hold. The only difference is that the summation over H is a summation of d elements, and thus, $(H - h)$ is at most d in this case for the effect of the temporally extended actions.

B.4 Second-order term

Let's now see the upper bound on the second-order term, which will be useful for the upper bound on the exploration bonus.

By applying the pigeon-hole principle (Lemma D.1).

$$\sum_{k=1}^K \sum_{h=1}^H \frac{4SH^3L}{n_k(s, o, h)} \leq \square H^3 S^2 OL^2 d$$

B.5 Bound of the exploration bonus b_{hk}

Before bounding the sum, we need to define the exploration bonus. We will consider an adaptation to temporally extended actions and non-stationary transitions of the same bonus presented in the original paper of UCBVI Azar et al. (2017). However, to make the definition clearer, let us motivate the need for this term.

Given that the optimistic value function \tilde{V}^{μ_k} is an upper bound of the true value function V^* , we can not guarantee the same for the relative empirical variance. Hence, if the empirical variance of \tilde{V}^{μ_k} is an upper bound on the empirical variance of V^* . Nonetheless, it is possible to prove that when the two value functions are sufficiently close to each other, the same applies to their empirical variance.

Let's resort to Lemma 2 of Azar et al. (2017),

$$\hat{\mathbb{V}}_{hk}^* \leq 2\hat{\mathbb{V}}_{hk} + 2 \mathbb{V}_{(s', h') \sim \hat{P}^{\mu_k}}(\tilde{V}(s', h') - V^*(s', h')) \leq 2\hat{\mathbb{V}}_{hk} + 2\hat{P}^{\mu_k}(\tilde{V}(s', h') - V^*(s', h'))^2$$

where $\hat{\mathbb{V}}_{hk}^* = \mathbb{V}_{(s', h') \sim P^{\mu_k}(\cdot | s, h)}(V^*(s', h'))$ and $\hat{\mathbb{V}}_{hk} = \mathbb{V}_{(s', h') \sim \hat{P}^{\mu_k}}(\tilde{V}^{\mu_k}(s, h))$.

We need this term to be of the same order as the estimation error e_{hk} , and thus we can say that

$$b_{hk} \sim [(\hat{P}_{hk}^{\mu_k} - P_h^{\mu_k})V^*(s', h')](s_{hk})$$

This time, however, we use the Empirical-Bernstein inequality Maurer & Pontil (2009) because we need the empirical variance to appear.

$$b_{hk} \leq \left(2\sqrt{\frac{\hat{\mathbb{V}}_{hk}^* L}{n_k(s, o, h)}} + \frac{14HL}{3n_k(s, o, h)} \right)$$

By applying Lemma 2 to this equation and substituting $\hat{\mathbb{V}}_{hk}^*$ we get the same form of bonus of Azar et al. (2017).

$$b_{hk} = \sqrt{\frac{8L \mathbb{V}_{(s', h') \sim \hat{P}_k^{\mu_k}(\cdot | s, h)}(\tilde{V}^{\mu_k}(s', h'))}{n_k(s, o, h)}} + \frac{14HL}{3n_k(s, o, h)} + \sqrt{\frac{8 \sum_{s', h'} \hat{P}_k^{\mu_k}(s', h' | s, h) [\min(b'_{h'k}, H^2)]}{n_k(s, o, h)}}$$

in which b'_{hk} stands for the upper bound on the square root of the difference between the optimistic value function in the next state-stage pair, and the optimal value function in the same next state-stage.

The last thing to do to properly define the bonus is express b'_{hk} in our scenario. Let's write

$$\tilde{V}(s', h') - V^*(s', h') \leq \sqrt{b'_{hk}}$$

and consider that b'_{hk} has to be appropriate to guarantee an adaptation of Lemma 16 of Azar et al. (2017), in which the second inequality applies if $\sqrt{N'_{hk}(s)} \geq 2500H^2S^2AL^2$, which is the second order term for standard UCBVI, given that $N'_{hk}(s) \geq H^2S^2AL^2$ for good episodes. Therefore, in

our scenario, we need that

$$\sqrt{b'_{hk}} \left(\sum_o n_k(s, o, h) \right) \geq \square H^4 S^2 O L^2 \geq \square H^3 S^2 O L^2 d$$

where the r.h.s of the equation above is the second-order term in our case. Thus, considering that $\sum_o n_k(s, o, h) \leq K$, and $K \geq H^3 L^2 S^2 O \geq H^2 L^2 S^2 O d$ for typical episodes, we have:

$$b'_{hk} = \frac{100^2 H^5 S^2 L^2 O}{\sum_o n_k(s, o, h)}$$

When considering the bound for the next state-stage pair $b'_{h'k}$, we simply refer to the visit count of the next state and next stage $n_k(s', o, h')$. The numerical constant 100^2 is derived analogously to [Azar et al. \(2017\)](#).

Let's now analyze the summation of this term, considering, as for e_{hk} , just the typical episodes.

$$\begin{aligned} \sum_{k=1}^K \sum_{h=1}^H b_{hk} &= \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \left(\sqrt{\frac{8L \text{Var}_{(s', h') \sim \hat{P}_k^{\mu_k}(\cdot|s, h)}(\tilde{V}^{\mu_k}(s', h'))}{n_k(s, o, h)}} + \frac{14HL}{3n_k(s, o, h)} \right)}_{(ft)} \\ &\quad + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \sqrt{\frac{8 \sum_{s', h'} \hat{P}_k^{\mu_k}(s', h'|s, h) [\min(b'_{h'k}, H^2)]}{n_k(s, o, h)}}}_{(st)} \end{aligned}$$

We separately analyze the first two terms and then the last.

The analysis of (ft) follows the same concept as the analysis conducted for the estimation error e_{hk} where instead of using Lemma D.3 we use Lemma D.4

$$\begin{aligned} (ft) &\stackrel{a}{\leq} \sqrt{8L} \left(\sqrt{KH^2 + \square HdU_{K,1} + \square H^2 S d \sqrt{KLO} + 4/3H^3 L} \right) (\sqrt{SOdL}) + 14/3HSOdL^2 \\ &\stackrel{b}{\leq} \sqrt{8L} \left(\sqrt{KH^2 + \square HdU_{K,1}} \right) (\sqrt{SOdL}) + 14/3HSOdL^2 \\ &\leq \square LH \sqrt{KSOd} + \square Ld \sqrt{HSOU_{K,1}} \end{aligned}$$

- (a) As we said above, we follow the same concept of point (c) of the proof of the upper bound of e_{hk} . In this case, we use Lemma D.4 instead of Lemma D.3.
- (b) Because for typical episodes $K \geq H^2 L^2 S^2 O d$ and thus we consider only the dominant terms.

Regarding the second term (st) adapting the proofs of [Azar et al. \(2017\)](#), we will focus only on the last term $(k)(h)$, which results in a term of the same order of the second-order term already analyzed, the other two terms are upper bounded by the main terms.

$$\begin{aligned} (st) &\stackrel{a}{\leq} \sqrt{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) b'_{h'k}} \sqrt{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \frac{1}{n_k(s, o, h)}} \\ &\stackrel{b}{\leq} \sqrt{H^5 S^2 L^2 O} \sqrt{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \frac{1}{n_k(s', o, h')}} \sqrt{\sum_{k=1}^K \sum_{h=1}^H \mathbb{I}(k \in [k]_{typ}) \frac{1}{n_k(s, o, h)}} \\ &\stackrel{c}{\leq} \sqrt{H^5 S^2 L^2 O} (\sqrt{SOdL})^2 \\ &= H^2 S^2 L^2 \sqrt{O^3 H d^2} \\ &\stackrel{d}{\leq} H^3 S^2 L^2 O d \end{aligned}$$

- (a) Considering only the $(k)(h)$ of the original proof and applying Cauchy-Schwartz inequality.

- (b) By substituting b'_{hk} in the equation.
- (c) By applying two times Lemma D.1.
- (d) If $O \leq H$.

To conclude the summation of exploration bonuses

$$\sum_{k=1}^K \sum_{h=1}^H b_{hk} \leq \square LH\sqrt{KS Od} + \square Ld\sqrt{HSOU_{K,1}} + H^3 S^2 L^2 Od$$

neglecting smaller order terms.

B.6 Summing all the terms

Finally, we can combine all the terms analyzed separately back into Equation (7), and we will get:

$$\begin{aligned} \widetilde{\text{Regret}}(K) &\leq \square LH\sqrt{KS Od} + \square Ld\sqrt{HSOU_{K,1}} + \square H^3 S^2 L^2 Od + H\sqrt{dKL} \\ &\stackrel{(a)}{\leq} \square LH\sqrt{KS Od} + \square HSL^2 Od^2 + \square H^3 S^2 L^2 Od + H\sqrt{dKL} \\ &\leq \square LH\sqrt{KS Od} + \square H^3 S^2 L^2 Od + H\sqrt{dKL} \end{aligned}$$

where (a) results by solving for $U_{K,1}$, and this completes the proof, ignoring the numeric constants replaced by \square . \square

Remark: The term d is a random variable, being the duration of each option a random variable itself. However, as shown in Drappo et al. (2023), it is possible to bound this value when we have options with duration $\tau_{\min} \leq \tau_o \leq \tau_{\max}$, resorting to *renewal processes* theory (Pinelis, 2019) with

$$d \leq \sqrt{\frac{32H(\tau_{\max} - \tau_{\min}) \log(2/\delta)}{\min_{o \in \mathcal{O}} \mathbb{E}[\tau_o]^3}} + \frac{H}{\min_{o \in \mathcal{O}} \mathbb{E}[\tau_o]}.$$

holding with probability at least $1 - \delta$.

This term is bounded by the ratio between the horizon H and the expected duration of the shorter option composing the set, plus a confidence interval accounting for the stochasticity of the duration.

C Proof of Theorem 4.3

In this section, we will provide a detailed proof of Theorem 4.3.

As described in the main paper, the meta-algorithm alternates between two regret minimizers, UCBVI and Options-UCBVI, for N stages at two levels of temporal abstraction of the problem. While learning on one level, the policies of the second are kept fixed for all episodes on the stage.

Initially, we will keep the analysis general for any pair of regret minimizers, $\mathfrak{A}^L, \mathfrak{A}^H$ - where the former is the regret minimizer used for the low-level and the latter the one used for the high-level.

Before proceeding, we introduce Lemma 4.2, which relates the regret paid by the regret minimizer of one level to the bias introduced in the learning of the other level.

Lemma 4.2. *Let us define the concentrability coefficients:*

$$\begin{aligned} C^H &:= \max_{n \in [N]} \inf_{\mu^*} \max_{\text{optimal}(s,h) \in \mathcal{S} \times [H]} \frac{d_{s_1,1}^{\mu^*}(s,h)}{d_{s_1,1}^{\mu_n}(s,h)}, \\ C^L &:= \max_{n \in [N]} \max_{o \in \mathcal{O}} \inf_{\pi_o^*} \max_{\text{optimal}(s,h) \in \mathcal{I}^o} \max_{(s',h') \in \mathcal{S}_o \times [H_o]} \frac{d_{s,h}^{\pi_o^*}(s',h')}{d_{s,h}^{\pi_{n-1}^o}(s',h')}. \end{aligned}$$

Then, it holds that:

$$\underbrace{V^*(s_1, 1) - V_{\pi_{n-1}}^*(s_1, 1)}_{\text{Bias of not playing } \pi^*} \leq C^H \underbrace{\left(V_{\pi_n}^{\mu_n}(s_1, 1) - V_{\pi_{n-1}}^{\mu_n}(s_1, 1) \right)}_{\text{Regret of low-level algorithm}},$$

$$\underbrace{V_{\star}^*(s_1, 1) - V_{\star}^{\mu_n}(s_1, 1)}_{\text{Bias of not playing } \mu^*} \leq C^L \underbrace{\left(V_{\pi_{n-1}}^*(s_1, 1) - V_{\pi_{n-1}}^{\mu_n}(s_1, 1) \right)}_{\text{Regret of high-level algorithm}}.$$

where μ^* is the optimal high-level policy (SMDP), and π_o^* is the optimal policy of a single option o (low-level optimal policy).

Proof. Let us write the bias of a level for the stage $n \in [N]$ as β_n , respectively specialized as β_n^H for the high-level bias and β_n^L for the low-level bias.

$$\begin{aligned} \beta_n^H &= V_{\star}^*(s_1, 1) - V_{\pi_{n-1}}^*(s_1, 1) \\ &\stackrel{a}{=} \mathbb{E}_{(s,h) \sim d_{s_1,1}^{\mu^*}} [R_{\pi^*}(s, h) - R_{\pi_{n-1}}(s, h)] \\ &\stackrel{b}{=} \mathbb{E}_{(s,h) \sim d_{s_1,1}^{\mu_n}} \left[\frac{d_{s_1,1}^{\mu^*}(s, h)}{d_{s_1,1}^{\mu_n}(s, h)} (R_{\pi^*}(s, h) - R_{\pi_{n-1}}(s, h)) \right] \\ &\stackrel{c}{\leq} \max_{n \in [N]} \inf_{\mu^*} \max_{\text{optimal } (s,h) \in S \times [H]} \frac{d_{s_1,1}^{\mu^*}(s, h)}{d_{s_1,1}^{\mu_n}(s, h)} (V_{\star}^{\mu_n}(s_1, 1) - V_{\pi_{n-1}}^{\mu_n}(s_1, 1)) \\ &\stackrel{d}{\leq} C^H (V_{\star}^{\mu_n}(s_1, 1) - V_{\pi_{n-1}}^{\mu_n}(s_1, 1)) \end{aligned}$$

- (a) We can write the difference in value as the difference in return of the two option policies, where R_{π^*} and $R_{\pi_{n-1}}$ are respectively the return obtained by playing the optimal options policies, and the return obtained by playing the options policies learned up to the previous step, and the state-stage pairs (s, h) are sampled from the distribution of visit induced by the policy μ^* .
- (b) Using an *importance-sampling* argument, we can change the exploration policy by adding the *importance weighting* term $\frac{d_{s_1,1}^{\mu^*}(s, h)}{d_{s_1,1}^{\mu_n}(s, h)}$
- (c) Substituting the expectation with the *sup* over the states and stages, the *inf* over the possible optimal exploration policies, and maximizing for all possible n stages.
- (d) Substituting the first term with the constant C^H , defined above.

We will not consider the proof of the second inequality because it follows the same passages. \square

Given this Lemma, we can provide a general result for any choice of $\mathfrak{A}^L, \mathfrak{A}^H$, and any choice of scheduling.

Lemma C.1. *Let \mathfrak{A}^H and \mathfrak{A}^L be two regret minimizers that suffer regret bounded $R^H(K)$ and $R^L(K)$ when run for K episodes. Then, under Assumption 4.1, Algorithm 2 when run with the episode schedule $(K_n^H, K_n^L)_{n=1}^N$ such that $\sum_{n=1}^N K_n^L + K_n^H = K$, suffers regret bounded by:*

$$R(\text{HLML}, K) \leq \sum_{n=1}^N \left((C^H + 1) R^L(K_n^L) + (C^L + 1) R^H(K_n^H) \right).$$

Proof. We can write the regret of the two-phase algorithm as a summation of the regret of the high-level and the regret of the low-level as expressed by Equation (3) in the main paper.

$$\begin{aligned} \text{Regret}(\text{HLML}, K) &= \sum_{n=1}^N \left(\sum_{k=1}^{K_n^H} (V_{\star}^*(s_1, 1) - V_{\pi_{n-1}^k}^{\mu_n}(s_1, 1)) + \sum_{k=1}^{K_n^L} (V_{\star}^*(s_1, 1) - V_{\pi_{n,k}}^{\mu_n}(s_1, 1)) \right) \\ &\stackrel{a}{=} \sum_{n=1}^N (\beta_n^H + R^H(K_n^H) + \beta_n^L + R^L(K_n^L)) \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{n=1}^b (C^H R^L(K_{n-1}^L) + R^H(K_n^H) + C^L R^H(K_{n-1}^H) + R^L(K_n^L)) \\
&\leq \sum_{n=1}^c (C^H + 1) R^L(K_n^L) + (C^L + 1) R^H(K_n^H).
\end{aligned}$$

- (a) We can decompose the two terms of the summation as shown in Equations (4) and (5), and then for shortness, use β_n to express the bias of the two levels at the n^{th} stage, and $R(K_n)$ for the regret of the two regret minimizers, $\mathfrak{A}^L, \mathfrak{A}^H$, at the n^{th} stage.
- (b) By applying Lemma 4.2 for the two general regret minimizers.
- (c) Clearly the sum of $n - 1$ is smaller than the sum of n terms, thus we can upper bound $R^L(K_{n-1}^L)$ with $R^L(K_n^L)$, and the same for $R^H(K_{n-1}^H)$.

And with the last step, we conclude the proof. \square

Now we can specialize Lemma C.1 for UCBVI for the options learning and Options-UCBVI for the high-level, and we get:

Theorem 4.3. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, H)$ be an FH-MDP and let \mathcal{O} be a set of options to be learned inducing the FH-MDPs $\mathcal{M}_o = (\mathcal{S}_o, \mathcal{A}_o, p, r_o, H_o)$ for $o \in \mathcal{O}$. The regret suffered by Algorithm 2 under Assumption 4.1, episode schedule as in Equation (2), and where $H_O = \max_{o \in \mathcal{O}} H_o$, is bounded with probability at least $1 - \delta$ by:*

$$R(\text{HLML}, K) \leq \tilde{O} \left(\underbrace{C^L \frac{H \sqrt{S O K d}}{}}_{\text{High-Level Regret}} + \underbrace{C^H \frac{H_O \sqrt{O S A K H_O}}{}}_{\text{Low-Level Regret}} \right).$$

Proof. For the option learning procedure, we instantiate a UCBVI algorithm for each sub-MDP \mathcal{M}_o , and for the $n - th$ phase, we paid a regret proportional to:

$$\begin{aligned}
\sum_{k=1}^{K_n^L} R_{o_{k_k}} &= \sum_o \sum_{j=1}^{K_o} R_{oj} \\
&\stackrel{a}{=} \sum_o H_o \sqrt{S_o A_o K_o H_o} \\
&\stackrel{b}{\leq} H_O \sqrt{S A H_O} \sum_o \sqrt{K_o} \\
&\stackrel{c}{\leq} H_O \sqrt{S A H_O} \sqrt{O \sum_o K_o} \\
&= H_O \sqrt{O S A H_O K_n^L}
\end{aligned}$$

where $R_{o_{k_k}}$ is the regret paid for running the option o_k in the $k - th$ episode and K_o are the episodes given to that option o . With (a), we just write the regret of running UCBVI on K_o episodes. In the passage (b), we upper bound to the worst possible sub-MDP, \mathcal{M}_o , where for the state space and the action space, we have the cardinalities of the primitive MDP, and we have an episode duration $H_O = \max_o H_o$. In the next inequality (c), we use the Cauchy-Schwartz inequality, and being $\sum_o K_o = K_n^L$ the last equality holds. Therefore, by considering just the dominant term of the two upper bounds of regret, we can write

$$\begin{aligned}
R_{K_n^L}^L &= \text{Regret-UCBVI} \leq \tilde{O} \left(H_O \sqrt{O S A K_n^L H_O} \right) \\
R_{K_n^H}^H &= \text{Regret-O-UCBVI} \leq \tilde{O} \left(H \sqrt{S O K_n^H d} \right)
\end{aligned}$$

Now by directly substituting these results in Lemma C.1 and considering the scheduling proposed in Equation (2), we can rewrite the regret of the meta-algorithm as:

$$\begin{aligned}
\text{Regret}(\text{HLML}, K) &\leq \tilde{O} \left(\sum_{n=1}^N \left((C^H + 1)H_O \sqrt{O S A H_O 2^n} + (C^L + 1)H \sqrt{S O d 2^n} \right) \right) \\
&= \tilde{O} \left(\left((C^H + 1)H_O \sqrt{O S A H_O} + (C^L + 1)H \sqrt{S O d} \right) \sum_{n=1}^N \sqrt{2^n} \right) \\
&= \tilde{O} \left(\left((C^H + 1)H_O \sqrt{O S A H_O} + (C^L + 1)H \sqrt{S O d} \right) 2\sqrt{2} \sum_{n=0}^{N/2} 2^n \right) \\
&= \tilde{O} \left(\left((C^H + 1)H_O \sqrt{O S A H_O} + (C^L + 1)H \sqrt{S O d} \right) \left(2\sqrt{2}(2^{N/2+1} - 1) \right) \right) \\
&\stackrel{a}{\propto} \tilde{O} \left(\left(C^H H_O \sqrt{O S A H_O} + C^L H \sqrt{S O d} \right) 2^{(\log_2(K))/2} \right) \\
&\leq \tilde{O} \left(\left(C^H H_O \sqrt{O S A H_O} + C^L H \sqrt{S O d} \right) \sqrt{K} \right)
\end{aligned}$$

Where all the passages follow algebraic operations, except for (a) in which we neglect all the numerical constants and we consider that $K = 2 \sum_{n=1}^N 2^{n-1} = 2^{N+1} - 1$ and thus, $N = \log_2(K)$. The last passage concludes the proof. \square

D Useful Lemmas

Lemma D.1. *Considering $n_k(s, o, h)$ the number of visits of the triple (s, o, h) up to episode k , and $[k]_{typ}$ the typical episodes for which $n_k(s, o, h)$ is sufficiently large, the following holds true:*

$$\sum_{k=1}^K \mathbb{I}(k \in [k]_{typ}) \sum_{h=1}^H \frac{1}{n_k(s, o, h)} \leq dSO \ln(Kd)$$

Proof.

$$\begin{aligned}
\sum_{k=1}^K \mathbb{I}(k \in [k]_{typ}) \sum_{h=1}^H \frac{1}{n_k(s, o, h)} &\stackrel{a}{\leq} \sum_{(s,o) \in S \times O} \sum_{h \in [d]}^{n_K(s,o,h)} \frac{1}{n} \\
&\stackrel{b}{\leq} dSO \sum_{n=1}^{Kd} \frac{1}{n} \\
&\stackrel{c}{\leq} dSO \ln(3Kd)
\end{aligned}$$

- (a) Considering $n_k(s, o, h)$ for the whole state space and options space, and considering the summation over H bounded by d elements, for the temporal extension of the actions.
- (b) Considering that the maximum number of (s, o, h) visited until episode K is bounded by Kd
- (c) Considering the rate of divergence of the harmonic series $\sum_{i=1}^n \frac{1}{i} \sim \ln(n)$

\square

The following lemmas are adaptations to SMDPs of Lemma 8, 9, and 10 of the paper of the UCBVI paper Azar et al. (2017). We consider to have the same good event \mathbb{E} and $\Omega_{k,h}$.

Lemma D.2. *Let $k \in [K]$ and $h \in [H]$. Then under the event \mathbb{E} and $\Omega_{k,h}$ of the original paper, the following hold*

$$\sum_{i=1}^k \sum_{j=h}^H \mathbb{V}_{i,j'}^\mu \leq KH^2 + 2\sqrt{H^5 KL} + 4d^3/3L$$

Proof. The proof follows the same passages of the proof of Lemma 8 in Azar et al. (2017), where j' is the next stage after a temporally extended transition. \square

Lemma D.3. *Let $k \in [K]$ and $h \in [H]$. Then under the event \mathbb{E} and $\Omega_{k,h}$ of the original paper, the following hold*

$$\sum_{i=1}^k \sum_{j=h}^H \left(\mathbb{V}_{i,j'}^* - \mathbb{V}_{i,j'}^\mu \right) \leq 2HdU_k + 4H^2\sqrt{HKL} + 4d^3/3L$$

Proof. The proof follows the same passages of the proof of Lemma 9 in Azar et al. (2017), where j' is the next stage after a temporally extended transition. \square

Lemma D.4. *Let $k \in [K]$ and $h \in [H]$. Then under the event \mathbb{E} and $\Omega_{k,h}$ of the original paper, the following hold*

$$\sum_{i=1}^k \sum_{j=h}^H \left(\hat{\mathbb{V}}_{i,j'} - \mathbb{V}_{i,j'}^\mu \right) \leq \square HdU_{k,1} + \square H^2 S \square d^2 KLO$$

Proof. The proof follows the same passages of the proof of Lemma 10 in Azar et al. (2017), where j' is the next stage after a temporally extended transition. More precisely, what changes is the application of the pigeon hole principle (Lemma D.1). \square