A Provably Efficient Option-Based Algorithm for both High-Level and Low-Level Learning

Gianluca Drappo
DEIB
Politecnico di Milano
Milan, 20133, Italy
gianluca.drappo@polimi.it

Alberto Maria Metelli
DEIB
Politecnico di Milano
Milan, 20133, Italy
albertomaria.metelli@polimi.it

Marcello Restelli
DEIB
Politecnico di Milano
Milan, 20133, Italy
marcello.restelli@polimi.it

Abstract

Hierarchical Reinforcement Learning (HRL) approaches have shown successful results in solving a large variety of complex, structured, long-horizon problems. Nevertheless, a full theoretical understanding of this empirical evidence is currently missing. In the context of the *option* framework, previous works have conceived provably efficient algorithms for the case in which the options are *fixed* and the high-level policy selecting among options only has to be learned. However, the fully realistic scenario in which *both* the high-level and the low-level policies are learned is surprisingly disregarded from a theoretical perspective. This work makes a step towards the understanding of this latter scenario. Focusing on the finite-horizon problem, in this paper, we propose a novel meta-algorithm that alternates between two regret minimization algorithms instanced at different (high and low) temporal abstractions. At the higher level, we look at the problem as a Semi-Markov Decision Process (SMDP), keeping the low-level policies fixed, while at a lower level, we learn the inner option policies by keeping the high-level policy fixed. Then, we specialize the results for a specific choice of algorithms, where we propose a novel provably efficient algorithm for the finite-horizon SMDPs, and we use a state-of-the-art regret minimizer for the options learning. We compare the bounds derived with those of state-of-the-art regret minimization algorithms for non-hierarchical finite-horizon problems. The comparison allows us to characterize the class of problems in which a hierarchical approach is provably preferable, even when a set of pre-trained options is not given.

1 Introduction

Hierarchical Reinforcement Learning [HRL, 20] is a framework in the class of Reinforcement Learning [RL 24] methods that has shown successful results in recent years thanks to its ability to deal with complex, long-horizon, and structured problems [26, 4, 18, 12]. In a large variety of real-world scenarios, a complex task can be decomposed as a concatenation of different sub-tasks that are often solved as a whole to learn the optimal policy. Nevertheless, in several cases, these sub-tasks are not fully coupled, and solving them separately leads to (near)optimal solutions. In these circumstances, a *hierarchical* RL approach could deliver significant benefits w.r.t. the application of *flat* RL algorithms, thanks to its ability to properly exploit the structure of the environment.

16th European Workshop on Reinforcement Learning (EWRL 2023).

A common example in the HRL literature [6] is the *taxi problem*, in which an autonomous agent controls a taxi that has to bring a passenger from a starting point to a destination location. This problem clearly embodies three different tasks: (i) driving, (ii) picking up the passenger, and (iii) dropping off the passenger when the destination is reached. HRL power resides in the explicit exploitation of this inner structure, subdividing the problem into a set of sub-tasks, individually solvable with their own optimal policies, which are then linked sequentially, one after the other. For instance, in the taxi example, the agent would separately learn (i) how to drive, (ii) how to optimally pick the passenger up, and (iii) how to drop her down. Then, it would choose the right sequence of sub-tasks to solve the entire problem. This approach naturally reduces each problem's complexity, being the agent focused on a single objective only, without being affected by other secondary goals. When the problem complexity further increases, HRL deals with it by constructing a hierarchy of sub-tasks depending on their abstraction. In this way, a sub-task could have, in turn, other sub-tasks as action space for its own policy. The root corresponds to the original problem, which has been simplified, becoming the problem of choosing which sub-task to execute first. Then, a sub-task, in turn, could be composed of other more specific sub-tasks, and this structure to follow down to the leaves. Finally, the leaves are the point where the actual state transaction is induced by the so-called *primitive actions* (i.e., actions of the original *flat* MDP on top of which the hierarchy is constructed). It is essential to specify that once a sub-task is selected, the control passes from that level policy to the one below, and this happens for every level. The controller returns to a certain policy only after completing every task below. This introduces the concept of *temporal abstraction* [22], and for what concerns the high-level policy, the action persists for a specific time, resulting in an actual reduction of the original planning horizon.

Recent works have attempted to analyze the theoretical benefits that motivate the great successes of HRL in practice [14, 8, 9, 27, 1]. For simplicity, most of them focus on problems organized in two-level hierarchies, where the high-level policy has control over a set of *pre-trained options* [22] (i.e., a particular formalization of temporally extended actions or sub-tasks), and the options' policies control the actual interaction with the environment throughout the *primitive actions*. The use of this set of fixed options helps to reduce the complexity of particular classes of problems, where the structure enforced by the options does not compromise optimality. For instance, in the *finite-horizon* setting, the usage of the options for training significantly reduces the planning horizon by a value dependent on the expected duration of the options composing the set. This translates into a more efficient dependency on the planning horizon H, which is replaced by a term $d \ll H$, the average per-episode number of options played [1]. On the other hand, even if the same rationale does not straightforwardly apply in *average-reward* problems, for their infinite-horizon nature, [8] and [9] demonstrate advantages in terms of exploration efficiency. In fact, these works show that a set of pre-trained options significantly improves the exploration of infinite-horizon problems, where, with these policies, the agent is able to explore wider regions of the problem faster.

While this clearly motivates the performance improvements empirically experienced in several tasks, it is still obscure when to prefer such approaches in situations where *no pre-trained* supportive policies are available, and, thus, it is required to face the problem from scratch, solving both the high and the low-level training. To the best of our knowledge, [1] provide a primary insight in this direction for the first time. The authors propose a naïve approach for high-and-low-level learning in finite-horizon problems and compare it with a state-of-the-art flat approach to characterize problems in which the former outperforms the latter in complexity. To this end, the authors relax the assumption used by the other approaches by analyzing scenarios in which the only requirement is that the problem presents some structure, in the sense that it can be subdivided into a set of sub-goals, which characterize different sub-tasks, but for which the only information available is the sub-tasks description. This approach, even if in a preliminary manner, provides a first answer to the question above and addresses a more realistic scenario. In fact, if a set of options with fixed policies is a demanding requirement, problem structure discovery is a widely studied topic in the HRL literature [16, 17, 13].

Original Contributions The contributions of this paper can be summarized as follows:

- We propose a novel meta-algorithm, named *High-Level/Low-level Meta-Learning* (HLML), for learning at both the high- and the low-levels by training independent of the regret minimizers chosen for the two levels (Section 3).
- We derive a novel HRL regret minimization algorithm for solving finite-horizon SMDPs, *Options-UCBVI* (O-UCBVI), that extends UCBVI [3], which is the state-of-the-art algorithm for FH-MDP,

and that enjoys an upper bound on the regret of order $\tilde{O}(H\sqrt{SOKd})$, being d the average per-episode number of played options (Section 4).
- We instantiate our meta-algorithm with Options-UCBVI for the *high-level* and UCBVI for the *low-level* (i.e., the options learning) and provide regret guarantees in comparison with UCBVI for solving the *flat* problem. This allows us to characterize specific classes of problems in which the former provide better theoretical guarantees, answering the question *"when to prefer HRL to standard RL, when both high-level and low-level policies are unknown?"* (Section 5).

The proofs of all the results presented in the main paper are reported in the Appendix.

2 Problem Formulation

In this section, we provide the necessary background that will be employed in the subsequent sections.

Finite-Horizon MDPs A Finite-Horizon Markov Decision Process [FH-MDP, 23] is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r^L, p, H)$, where $\mathcal{S}$ is the state space with caridnality S; $\mathcal{A}$ the action space with cardinality A; $r^L : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$ is the reward function, which quantifies the quality $r^L(s, a, h)$ of action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at stage $h \in [H]$; $p : \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S} \to [0, 1]$ is the transition model, defining the probability $p(s'|s, a, h)$ of transitioning to state $s' \in \mathcal{S}$ by taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at stage $h \in [H]$; and $H \in \mathbb{N}$ is the horizon. The behavior of an agent is modeled by a (*low-level*) deterministic policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ that maps a state $s \in \mathcal{S}$ and a stage $h \in [H]$ to a (*low-level* or *primitive*) action $\pi(s, h) \in \mathcal{A}$.

Finite-Horizon Semi-MDPs A Finite-Horizon Semi-Markov Decision Process [FH-SMDP, 1] is the adaptation of Semi-Markov Decision Processes [5] to finite-horizon setting. An FH-SMDP is defined as a tuple $\mathcal{SM} = (\mathcal{S}, \mathcal{O}, p, r^H, H)$, where, as for FH-MDP, $\mathcal{S}$ is the state space, with cardinality S, and H is the horizon. $\mathcal{O}$ is a set of temporally extended actions (*high-level*), with cardinality O. $p : \mathcal{S} \times \mathcal{O} \times [H] \times \mathcal{S} \times [H] \to [0, 1]$ is the transition model, defining the probability $p(s', h'|s, o, h)$ of transitioning to state $s' \in \mathcal{S}$, after $(h - h')$ time steps, $h' \in [H]$, when playing (*high-level*) action $o \in \mathcal{O}$, in state $s \in \mathcal{S}$, and stage $h \in [H]$; $r^H : \mathcal{S} \times \mathcal{O} \times [H] \to [0, H]$ is the (*high-level*) cumulative reward obtained $r^H(s, o, h)$, until the temporally extended (*high-level)* action $o \in \mathcal{O}$ terminates, when selected in state $s \in \mathcal{S}$, at stage $h \in [H]$. Naturally, from the fact that a (*high-level*) action executes for a certain number of *primitive* (*low-level*) steps, the *duration* or *holding time*, $\tau(s, o, h)$, defines the number of primitive steps taken in the environment while a temporally extended action $o \in \mathcal{O}$ is executed. The behavior of an agent is modeled by a deterministic (*high-level*) policy $\mu : \mathcal{S} \times [H] \to \mathcal{O}$ that maps a state and a stage $h \in [H]$ to a (*high-level*) action $\mu(s, h) \in \mathcal{O}$.

Hierarchical Reinforcement Learning builds upon the theory of Semi-MDPs, characterizing the concept of temporally extended action with fundamentally two frameworks [20]: sub-tasks [6] and options [25]. For the sake of this paper, we focus on the options framework.

Options An option [25] is a temporally extended action characterized by three components $o = (\mathcal{I}^o, \beta^o, \pi^o)$. $\mathcal{I}^o \subseteq \mathcal{S} \times [H]$ is the subset of states and stages pairs $(s, h) \in \mathcal{S} \times [H]$ in which the option can start, $\beta^o : \mathcal{S} \times [H] \to [0, 1]$ defines the probability $\beta^o(s, h)$ that an option terminates in state $s \in \mathcal{S}$ and stage $h \in [H]$, and, $\pi^o : \mathcal{S} \times [H] \to \mathcal{A}$ is the deterministic policy executed once an option is selected and until its termination.

Before proceeding, we introduce the following standard assumption.

Assumption 2.1 (Admissible options [9])**.** The set of options $\mathcal{O}$ is assumed *admissible*, i.e. $\forall o \in \mathcal{O}$, $s \in \mathcal{S}$, and $h \in [H]$: $\beta^o(s, h) > 0 \implies \exists o' \in \mathcal{O} : (s, h) \in \mathcal{I}^{o'}$.

The assumption is a minimal requirement for the problem to be well-defined, and it guarantees that whenever an option o stops in a state s at stage h, there always exists another option o' that can start from the state-stage pair (s, h).

Average per-episode duration In the following analysis, we will refer to d [1] as the average per-episode number of decisions taken in an episode of length H.

$$d = \frac{1}{K} \sum_{k=1}^{K} \sum_{o \in \mathcal{O}} \sum_{s \in \mathcal{S}} \sum_{h \in [H]} d_k(s, o, h) \tag{1}$$

where $d_k(s, o, h)$ is the number of times a temporally extended action (or option) o has been selected in state s, in step h, in the episode k of interaction with the environment. This quantity is a random variable, being dependent on the duration of the options, which is a random variable in turn.

Problem Formulation We are given a set of *not pre-trained* options $\mathcal{O}$, i.e., for every option $o \in \mathcal{O}$, the initiation set $\mathcal{I}^o$, and the termination function β^o are *fixed*, while the inner low-level policy π^o has to be learned. We seek to solve the problem of learning *both* the high-level policy μ (selecting options in the FH-SMDP) and the low-level policies π^o (inner to the options) for every $o \in \mathcal{O}$:

$$(\mu^*, \boldsymbol{\pi}^*) \in \operatorname*{argmax}_{\mu, \boldsymbol{\pi}} V^{\mu}_{\boldsymbol{\pi}}(s_1, 1), \tag{2}$$

where $\boldsymbol{\pi} = (\pi^o)_{o \in \mathcal{O}}$ are the low-level policies and μ is the high-level policy, $s_1 \in \mathcal{S}$ is an initial state, and $V^{\mu}_{\boldsymbol{\pi}}$ is the value function, defined for every $(s, h) \in \mathcal{S} \times [H]$ as:

$$V^{\mu}_{\boldsymbol{\pi}}(s, h) \coloneqq \mathop{\mathbb{E}}_{(s', h') \sim p(\cdot | s, \mu(s,h), h)} \Big[r^H(s, \mu(s, h), h) + V^{\mu}_{\boldsymbol{\pi}}(s', h') \Big], \tag{3}$$

$$r^H(s, o, h) \coloneqq \mathop{\mathbb{E}}_{s'' \sim p(\cdot | s, \pi^o(s,h), h)} \Big[r^L(s, \pi^o(s, h), h) + (1 - \beta^o(s'', h+1)) r^H(s'', o, h+1) \Big]. \tag{4}$$

We denote with $V^*_*(s_1, 1) = V^{\mu^*}_{\boldsymbol{\pi}^*}(s_1, 1)$.

Regret The *regret* [3, 28, 8, 1] of an algorithm $\mathfrak{A}$ for the problem defined above is the cumulative value difference over K episodes when playing the high-level policy μ_k and the low-level policies $\boldsymbol{\pi}_k$ at the episode $k \in [K] \coloneqq \{1, \dots, K\}$ instead of the optimal ones:

$$\text{Regret}(\mathfrak{A}, K) \coloneqq \sum_{k=1}^{K} V^*_*(s_1, 1) - V^{\mu_k}_{\boldsymbol{\pi}_k}(s_1, 1)$$

Thus the goal of the algorithm is to play a sequence of policies $\mu_0, \dots, \mu_K$, and $\boldsymbol{\pi}_0, \dots, \boldsymbol{\pi}_K$, such that Regret($\mathfrak{A}, K$) is as small as possible.

3 Meta-Algorithm for High-and-Low-level Training

In this section, we introduce the first contribution of this work, consisting of a meta-algorithm, *High-Level/Low-Level Meta-Learning* (HLML), that alternates between high- and low-level learning.

HLML presented in Algorithm 1, takes as input two regret minimizers $\mathfrak{A}^H$ and $\mathfrak{A}^L$ designed for learning in the FH-SMDP (i.e., at a high level, learning μ^*) and in the FH-MDP (i.e., at a low level, learning $\boldsymbol{\pi}^*$), respectively. The meta-algorithm operates in N *stages*. In stage $n \in [N] \coloneqq \{1, \dots, N\}$, we run the high-level regret minimizer $\mathfrak{A}^H$ for K^H_n episodes, keeping the low-level policies $\boldsymbol{\pi}_{n-1} = (\pi^o_{n-1})_{o \in \mathcal{O}}$ fixed. Algorithm $\mathfrak{A}^H$ will output the high-level policy μ_n which is chosen uniformly at random among the $\mu_{n,1}, \dots, \mu_{n,K^H_n}$ played during its execution in the stage. Then, the control moves to the low level, and we run the low-level regret minimizer $\mathfrak{A}^L$ for K^L_n episodes, keeping the high-level policy μ_n fixed. Algorithm $\mathfrak{A}^L$ will output the low-level policies $\boldsymbol{\pi}_n$ chosen uniformly at random among the ones $\boldsymbol{\pi}_{n,1}, \dots, \boldsymbol{\pi}_{n,K^L_n}$ played during its execution in the stage. The meta-algorithm, then, moves to the next stage $n + 1$, passing back the control to the high level, and the process continues. The schedule of the number of episodes $(K^H_n, K^L_n)_{n \in [N]}$ must satisfy that $\sum_{n=1}^{N} K^H_n + K^L_n = K$.

The key feature of our meta-algorithm is that when the high-level algorithm $\mathfrak{A}^H$ is running in stage n the low-level (inner-option) policies $\boldsymbol{\pi}_{n-1}$ are kept fixed. Therefore, $\mathfrak{A}^H$ is actually performing regret minimization in an FH-SMDP, enjoying the corresponding regret guarantees, for converging to the optimal high-level policy for the fixed options $\mathcal{O}$. This allows us to solve the common non-stationarity issues that arise when two learning processes are carried out in parallel. Clearly, such a high-level policy will not necessarily be μ^*, since we are not guaranteed that the low-level policies $\boldsymbol{\pi}_{n-1}$ are optimal for the corresponding options. This is the reason why the execution of $\mathfrak{A}^H$ is stopped after K^H_n episodes and, within the same stage n, we proceed running the low-level regret minimizer $\mathfrak{A}^L$, before continuing learning at the high-level. Similarly, in this phase, $\mathfrak{A}^L$ is acting on the flat MDP with the goal of learning the inner policy π^o_n for each of the options $o \in \mathcal{O}$. This amounts to solving for each option $o \in \mathcal{O}$ a single FH-MDP formalized as $\mathcal{M}_o = (\mathcal{S}_o, \mathcal{A}_o, p, r_o, H_o)$ where $\mathcal{S}_o \subseteq \mathcal{S}$, $\mathcal{A}_o \subseteq \mathcal{A}$, $H_o \leq H$, meaning that each option operates on a restricted portion of the original problem and for a specific fixed horizon H_o (induced by $\mathcal{I}^o$ and β^o). This time the high-level policy is kept

Algorithm 1 High-Level/Low-level Meta-Learning (HLML)

Input: FH-SMDP regret minimizer $\mathfrak{A}^H$, FH-MDP regret minimizer $\mathfrak{A}^L$, episode schedules $(K_n^H, K_n^L)_{n=1}^N$
Arbitrarily initialize μ_0 and $\boldsymbol{\pi}_0$
for $n = 1, \dots, N$ **do**
  Run $\mathfrak{A}^H$ on the FH-SMDP for K_n^H episodes playing the sequence of high-level policies $\mu_{n,1}, \dots, \mu_{n,K_n^H}$
  Fix the high-level policy $\mu_n = \mu_{n,X}$ where $X \sim \mathrm{Uni}([K_n^H])$
  Run $\mathfrak{A}^L$ on the FH-MDP for K_n^L episodes playing the sequence of low-level policies $\boldsymbol{\pi}_{n,1}, \dots, \boldsymbol{\pi}_{n,K_n^L}$
  Fix the low-level policies $\boldsymbol{\pi}_{n-1} = \boldsymbol{\pi}_{n-1,Y}$ with $Y \sim \mathrm{Uni}([K_n^L])$
end for
return $(\mu_N, \boldsymbol{\pi}_N)$

fixed, and consequently, its effect is enforcing a specific exploration that determines a particular option visitation.

In principle, solving such FH-MDPs $\mathcal{M}_o$ can be as complex as solving the original problem $\mathcal{M}$ with a flat approach. This is expected since the advantages of a hierarchical approach emerge when a certain *structure* on the original problem is present. This is particularly evident if we think of the convergence of the learning process of the low-level policies, which could potentially end up in a different optimum than the one reached by a flat approach in that same portion of the problem because the latter would have a complete scope over the whole problem. For this reason, a further assumption over the structure of the problem is required, similar to the one presented in [1].

Assumption 3.1. For any optimal high-level policy μ^*, let $\mathcal{O}_{\mu^*}$ the set of options played by μ^* and for $o \in \mathcal{O}_{\mu^*}$, let Π_o^* the set of optimal low-level policies form the joint optimization. Let $\Pi_o^\#$ be the set of optimal low-level policies from the local optimization $(\pi_o^\# \in \mathrm{argmax}_{a \in \mathcal{A}} Q^{*,o}(s,a) \forall s \in \mathcal{S}_o)$. It holds

$$\Pi_o^\# \subseteq \Pi_o^* \tag{5}$$

This assumption ensures that the optimal inner-option policies π_o^*, on a portion of the original MDP $\mathcal{M}_o$ induced by an options $o \in \mathcal{O}$, selected by the optimal SMDP policy μ^*, do not differ from an optimal policy π^* of the flat problem. This way, we can safely learn in the FH-MDPs $\mathcal{M}_o$ knowing that the learned policy will be "a portion" of the optimal policy π^* in the flat FH-MDP. This assumption, seemingly demanding, is the first one, to the best of our knowledge, that attempts to characterize a structural property of the FH-MDPs that is suitable for being addressed by means of a hierarchical approach. Indeed, if Assumption **??** is violated, it means that the inner-option learning deviates from the process of learning the optimal policy in the flat MDP, possibly preventing the convergence to the optimal policy in the hierarchical architecture. An example of a scenario in which this assumption is valid is the taxi problem described above. For instance, from a starting point A to destination B, the optimal driving policy (i.e., the one solving the subtask (i)) does not differ if the problem is considered a whole or a smaller one that includes just the neighborhood of the two points.

Theoretical Analysis As described above, in each stage $n \in [N]$, the learning process alternates between the high- and the low-level learning problems, keeping the other fixed. This induces a bias in both optimizations. To make this clear, we provide a convenient decomposition of the regret, which highlights the contributions of the two phases of learning in each stage:

$$\mathrm{Regret}(\mathrm{HLML}, K) = \sum_{n=1}^{N} \left(\sum_{k=1}^{K_n^H} \underbrace{V_*^*(s_1,1) - V_{\boldsymbol{\pi}_{n-1}}^{\mu_{n,k}}(s_1,1)}_{\text{Regret during high-level learning}} + \sum_{k=1}^{K_n^L} \underbrace{V_*^*(s_1,1) - V_{\boldsymbol{\pi}_{n,k}}^{\mu_n}(s_1,1)}_{\text{Regret during low-level learning}} \right), \tag{6}$$

where $\mu_{n,k}$ and $\boldsymbol{\pi}_{n,k}$ are the high-level policy and the low-level policies played by the corresponding algorithms $\mathfrak{A}^H$ and $\mathfrak{A}^L$ at episode k of phase n. Unfortunately, the two terms in Equation (6) cannot be directly bounded in terms of the properties of the regret minimization algorithms $\mathfrak{A}^H$ and $\mathfrak{A}^L$. This is because each of them, as explained above, will converge to the corresponding high/low-level optimal policy, given that the other-level policy is fixed. Thus, further elaboration is needed to

highlight the bias terms:

$$\underbrace{V^*_*(s_1,1) - V^{\mu_{n,k}}_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Regret during high-level learning}} = \underbrace{V^*_*(s_1,1) - V^*_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Bias of not playing } \boldsymbol{\pi}^*} + \underbrace{V^*_{\boldsymbol{\pi}_{n-1}}(s_1,1) - V^{\mu_{n,k}}_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Regret of } \mathfrak{A}^H} \tag{7}$$

$$\underbrace{V^*_*(s_1,1) - V^{\mu_n}_{\boldsymbol{\pi}_{n,k}}(s_1,1)}_{\text{Regret during low-level learning}} = \underbrace{V^*_*(s_1,1) - V^{\mu_n}_*(s_1,1)}_{\text{Bias of not playing } \mu^*} + \underbrace{V^{\mu_n}_*(s_1,1) - V^{\mu_n}_{\boldsymbol{\pi}_{n,k}}(s_1,1)}_{\text{Regret of } \mathfrak{A}^L}, \tag{8}$$

Thus, the regrets of the two phases (low- and high-level learning) are decomposed into a proper *regret* term and a *bias* term, which accounts for the fact that the other level is kept fixed. The regret terms can be easily managed by resorting to the properties of the regret minimizers $\mathfrak{A}^H$ and $\mathfrak{A}^L$. Concerning the bias terms, the high level corresponds to the value difference between playing the current low-level policies $\boldsymbol{\pi}_{n-1}$ compared to playing the optimal ones $\boldsymbol{\pi}^*$. Symmetrically, for the low level, this bias translates into the value difference between playing the current high-level policy μ_n compared to the optimal one μ^*. From a technical perspective, we decide to upper bound the bias terms with the proper regret terms at the price of introducing a *concentrability* coefficient for accounting of the distribution shift, as shown in the following result.

Lemma 3.2. *Let us define the concentrability coefficients:*

$$C^H \coloneqq \max_{n\in[N]} \inf_{\mu^* \text{ optimal}} \max_{(s,h)\in\mathcal{S}\times[H]} \frac{d^{\mu^*}_{s_1,1}(s,h)}{d^{\mu_n}_{s_1,1}(s,h)}, \tag{9}$$

$$C^L \coloneqq \max_{n\in[N]} \max_{o\in\mathcal{O}} \inf_{\pi^*_o \text{ optimal}} \max_{(s,h)\in\mathcal{I}^o} \max_{(s',h')\in\mathcal{S}_o\times[H_o]} \frac{d^{\pi^*_o}_{s,h}(s',h')}{d^{\pi^o_{n-1}}_{s,h}(s',h')}. \tag{10}$$

Then, it holds that:

$$\underbrace{V^*_*(s_1,1) - V^*_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Bias of not playing } \boldsymbol{\pi}^*} \leq C^H \Big(\underbrace{V^{\mu_n}_*(s_1,1) - V^{\mu_n}_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Regret of } \mathfrak{A}^L} \Big), \tag{11}$$

$$\underbrace{V^*_*(s_1,1) - V^{\mu_n}_*(s_1,1)}_{\text{Bias of not playing } \mu^*} \leq C^L \Big(\underbrace{V^*_{\boldsymbol{\pi}_{n-1}}(s_1,1) - V^{\mu_n}_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Regret of } \mathfrak{A}^H} \Big). \tag{12}$$

We are finally ready to state the main theoretical guarantees on the regret of our meta-algorithm. To this end, we assume that the individual low- and high-level regret minimizers enjoy suitable convergence properties, and, as a consequence, we derive the regret guarantees of the meta-algorithm.

Theorem 3.3. *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 **??**, Algorithm 1 when run with the episode schedule $(K^H_n, K^L_n)_{n=1}^N$ such that $\sum_{n=1}^N K^L_n + K^H_n = K$, suffers regret bounded by:*

$$\textit{Regret}(\textit{HLML}, K) \leq \sum_{n=1}^{N} (C^H + 1) R^L(K^L_n) + (C^L + 1) R^H(K^H_n). \tag{13}$$

Some observations are in order. First, we relate the regret of the meta-algorithm in terms of the regret suffered by the individual regret minimizers $\mathfrak{A}^H$ and $\mathfrak{A}^L$. It is worth noting that, for the sake of the analysis, we are assuming that whenever each algorithm starts running, all the data collected in the previous stages are discarded in order to remove inconvenient dependencies among the stages. Clearly, a practical version of the algorithm might save data (especially when learning at the low level) to be reused to further improve the estimates. Second, we can now appreciate the role of Assumption **??**. Indeed, in order to be able to converge at a low level to the optimal inner-option policies $\boldsymbol{\pi}^*$ (as in Equation 2), it must happen that the low-level regret minimizer $\mathfrak{A}^L$ performs an optimization that is compliant with what would have happened if solving the original flat MDP. Finally, let us note that our result is instanced for a generic choice of the schedule $(K^H_n, K^L_n)_{n=1}^N$. In the subsequent section, we will show that, for specific choices of $\mathfrak{A}^L$ and $\mathfrak{A}^H$, an exponential schedule allows achieving desirable regret guarantees.

4 Options-UCBVI

Algorithm 2 Options-UCBVI

Input: $\mathcal{S}, \mathcal{O}, H, K$
Initialize μ_0 arbitrarily, $Q_1(s,o,h) = 0$ for all $(s,o,h) \in \mathcal{S} \times \mathcal{O} \times [H]$, $L = \log(5SOKH/\delta)$, $\mathcal{D}^H \leftarrow \{\}$
for $k = 1, \dots, K$ **do**
  Compute $n_k(s,o,h) = \sum_{(x,y,z) \in \mathcal{D}^H} \mathbb{1}\{x = s, y = o, z = h\}$
  Estimate $\hat{P}_k(s',h'|s,o,h) = \frac{1}{\max 1, n_k(s,o,h)} \sum_{(x,y,z,w,u) \in \mathcal{D}^H} \mathbb{1}\{(x,y,z,w,u) = (s,o,h,s',h')\}$
  Set $Q_k(s,o,H+1) = 0$ for all $(s,o,h) \in \mathcal{S} \times \mathcal{O} \times [H]$
  for $h = H, \dots, 1$ **do**
    for $(s,o) \in \mathcal{S} \times \mathcal{O}$ **do**
      for $h' = h+1 \dots H+1$ **do**
        Compute

$$b_{hk}(s,o) = \sqrt{\frac{8L \, \mathbb{V}\mathrm{ar}_{(s',h') \sim \hat{P}_k(\cdot|s,o,h)}[\tilde{V}^{\mu_k}(s',h')]}{n_k(s,o,h)}} + \frac{14HL}{3n_k(s,o,h)} + \sqrt{\frac{8 \sum_{(s',h') \in \mathcal{S} \times [H]} \hat{P}_k(s',h'|s,o,h) \left[\min\left(\frac{100^2 H^5 S^2 O L^2}{\sum_o n_k(s',o,h')}\right), H^2\right]}{n_k(s,o,h)}}$$

$$Q_k(s,o,h) = r(s,o,h) + \sum_{(s',h') \in \mathcal{S} \times [H]} \hat{P}_k(s'h'|s,o,h) \tilde{V}^{\mu_k}(s',h') + b_{hk}(s,o)$$

$$\tilde{V}^{\mu_k}(s,h) = \min\left\{H - (h'-1), \max_{o \in \mathcal{O}} Q_k(s,o,h')\right\}$$

      end for
    end for
  end for
  $\mu_k(s,h) = \text{argmax}_{o \in O} Q_k(s,o,h)$
  $s \leftarrow s_1$
  while $h < H$ **do**
    Play option $o = \mu_k(s,h)$, and observe (s',h')
    Update $\mathcal{D}^H \leftarrow \mathcal{D}^H \cup \{(s,o,h,s',h')\}$
    $s \leftarrow s'$, $h \leftarrow h'$
  end while
end for

The adaptation to the finite-horizon setting of SMDP has been recently presented by [1], yet no provably efficient algorithm has been proposed. Therefore, in this section, we introduce a novel approach, *Options-UCBVI* (O-UCBVI), which builds upon UCBVI [3], that exploits a set of given options $\mathcal{O}$ to learn the optimal FH-SMDP policy μ^*. UCBVI is a model-based algorithm that implements *optimism in the face of uncertainty* by adding a confidence exploration bonus on the empirical Bellman operator. However, it is not directly applicable to FH-SMDPs. Indeed, in FH-SMDP, contrary to FH-MDP, there is an additional stochasticity for the uncertain duration of the temporally extended actions. Thus, it is not possible to directly apply the standard backward induction present in both versions of UCBVI. Intuitively the number of steps for which the procedure is repeated is unknown, or more precisely, is a random variable itself that depends on the duration of the temporally extended actions (or options, for our case) played in one episode.

For this reason, we have to change the algorithm by introducing a variable called $d \leq H$ [1], which is the average per-episode number of options that are selected in an episode of horizon H. This element will play a significant role in the analysis. As shown by [1], resorting to the *renewal processes* theory [21], it is possible to compute an upper bound when we have options with duration $\tau_{\min} \leq \tau(s,o,h) \leq \tau_{\max}$ holding with probability at least $1 - \delta$:

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

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.

Up to this crucial change, the O-UCBVI follows the same philosophy as UCBVI-BF [3], as shown in Algorithm 2. From a technical perspective, we modified the exploration bonus to deal with the non-stationary transition models and the set of given options, with their temporally extended nature. In particular, we focused on the version using the *Bernstein-Freedman* [7, 15] bonus in order

to achieve tight regret guarantees. The key intuitions behind the analysis are to directly maintain confidence intervals on the optimal value function and the use of Empirical-Bernstein [15] with a correction bonus to guarantee that the empirical variance is an upper bound on the variance of the true value function in the next state. We follow the same intuition in our analysis, and we end up demonstrating the following regret guarantee.

Theorem 4.1. *Let $\mathcal{SM}$ be an FH-SMDP with S states and O temporally extended actions (options), known reward,*[1]*, 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:*[2]

$$\textit{Regret}(\textit{O-UCBVI}, K) \leq \tilde{O}\left(H\sqrt{SOKd} + H^3S^2Od + H\sqrt{Kd}\right), \tag{14}$$

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

For $T \geq H^4S^3Od$ this bound translates into a regret bound of $\tilde{O}(H\sqrt{SOKd})$. The differences with the regret of UCBVI-FH, that scales with $\tilde{O}(\sqrt{HSAT})$ (being $T = KH$), are the additional $\sqrt{H}$, coming from the non-stationarity of the transition model, and the d term, that results from the nature of the problem solved, being an FH-SMDP and not an FH-MDP.[3] This result also highlights the performance improvement brought by the set of fixed options in the finite-horizon problems, as shown in [1]. The regret scales with $\sqrt{Kd}$ instead of $\sqrt{KH}$ as in the *flat* version. Since $d \ll H$, Options-UCBVI suffers smaller regret than its flat counterpart when fixed options are given. Furthermore, we can see that the result is a generalization of the flat case. Indeed, the upper bound is tight in its dominating term also when considering $\mathcal{O} = \mathcal{A}$ and $d = H$, i.e., running Options-UCBVI on the flat MDP. Nevertheless, because of the fact that the inner-option policies π^o are not learned, Options-UCBVI only partially answers our original question.

5 High-and-Low-Level Provably Efficient Learning

We are now ready to provide a complete algorithm able to learn both the high-level and the low-level policies in a provably efficient way. We instantiate our meta-algorithm HLML presented in Section 3 with Options-UCBVI presented in Section 4 as the high-level regret minimizer $\mathfrak{A}^H$ and an original version of UCBVI-FH as the low-level regret minimizer $\mathfrak{A}^L$. In order to achieve tight regret guarantees, we need to accurately select the schedule of the number of episodes K_n^H and K_n^L, namely, we duplicate the number of episodes when moving from one stage n to the next one $n+1$:

$$\forall n \in [N]: \qquad K_n^H = K_n^L = \lfloor 2^{n-1} \rfloor \qquad \text{where} \qquad N = \lfloor \log_2(2K+1) \rfloor. \tag{15}$$

Given this schedule, we can prove the following regret bound.

Corollary 5.1. *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 1 when instanced with $\mathfrak{A}^H$=O-UCBVI and $\mathfrak{A}^L$=UCBVI-FH, run with the episode schedule as in Equation* (15)*, and having where $H_O = \max_{o\in\mathcal{O}} H_o$, is bounded with probability at least $1-\delta$ by:*

$$\textit{Regret}(\textit{HLML}, K) \leq \tilde{O}\left(C^L \underbrace{H\sqrt{SOKd}}_{\textit{High-level regret}} + C^H \underbrace{H_O\sqrt{SAH_OK}}_{\textit{Low-level regret}}\right). \tag{16}$$

This result, as expected, is composed of the sum of the regrets suffered by the regret minimizers at the two levels, weighted by the *concentrability coefficients* C^H and C^L, coming from the direct application of Theorem 3.3. While the first term is exactly the regret paid by Options-UCBVI , the second is an upper bound of the total regret paid for the O options learning. In fact, in the analysis, instead of considering O different UCBVI, one for each option, with K_n^L/O episodes each, we assume to have a single algorithm running in the worst problem, i.e., the one with the longest horizon

[1] The choice of assuming a known reward is for compliance with [3]. 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.

[2] $\tilde{O}$ neglects logarithmic terms.

[3] This additional $\sqrt{H}$ term is well-known to be tight even in standard FH-MDPs when the transition model is non-stationary. The non-stationarity of the transition model is unavoidable in the Semi-Markov setting due to the different durations of the temporally extended actions.

H_o, for the total amount of episodes K_n^L. Besides, the dependency on the entire state and action space is motivated by the estimated transition model that is kept common to all the options.

At this point, it is possible to properly characterize the class of problems more efficiently solvable with this HRL approach instead of a *flat* one. We can do so by relating the regret of Equation (16), with regret paid by UCBVI in the original MDP with non-stationary transitions [4]. Let us consider a particular case for which $H_O = \alpha H$, with $0 < \alpha < 1$, we can write:

$$\frac{C^L H\sqrt{SOKd} + C^H H_O\sqrt{SAKH_O}}{H\sqrt{SAKH}} = C^L\sqrt{\frac{Od}{AH}} + C^H\sqrt{\alpha^3} \tag{17}$$

Therefore, considering the r.h.s of Equation (17), the classes of problems for which this HRL approach will suffer less regret than the *flat* approach are problems that guarantee to have this ratio smaller than 1, and that has a structure compliant to Assumption **??**. Under the assumption that the effect of the concentrability coefficients is negligible, there is a clear advantage of using the hierarchical approach when $Od \ll AH$ and, since $d \leq H$ by definition, for sure when $O \ll A$, i.e., when the number of options is smaller than the number of primitive actions. Of course, given the presence of C^L and C^H, this advantage gets mitigated by the magnitude of these constants.

6 Related Works

There is a vast literature for provably efficient algorithms for FH-MDP. [19] 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 [28]. [3] 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 [8], the authors propose an adaptation of UCRL2 [2] 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. [9] 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, [14] 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. [27] 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.

The closest approach in the literature is [1]. They propose to relax the assumption of having a set of pre-trained options and try to characterize the problems solvable more efficiently with an HRL approach, comparing their results with the finite-horizon version of UCRL2 [2, 10]. The authors propose an Explore-Then-Commit approach [11] for finite horizon problems, which takes the miss-specified options-set as input, then learns each option policy, and, after a defined number of episodes, exploits the options to solve an FH-SMDP with a custom algorithm proposed by the authors, inspired by UCRL2 [2]. They theoretically analyze the algorithm's performance and provide an upper bound on the regret suffered. Thus, they characterize the classes of problems more efficiently solvable by HRL, comparing the regret with the *flat* version of UCRL2 for finite horizon problems [10]. Nevertheless, this approach presents some limitations. First of all, due to its Explore-Then-Commit nature, it suffers from a regret of the order of $K^{2/3}$, which does not match the regret of the state-of-the-art algorithm for standard RL. Furthermore, the algorithm proposed for the FH-SMDP

[4] While in general comparing upper bounds is potentially loose, we notice that both upper-bounds are derived using similar techniques, and thus they would be "similarly" loose

setting is not optimal in all the main terms composing the regret, and it is compared to FH-UCRL, which in turn is suboptimal in $\sqrt{HS}$.

7 Conclusions

In this paper, we investigated the problem of learning the inner-option policies together with learning the high-level policy in an HRL setting based on the options framework. We first provided a novel meta-algorithm HLML based on the alternation between high- and low-level learning whose theoretical guarantees depend on those of the individual regret minimizers employed at the two levels under particular structural assumptions of the problem. This assumption represents the first attempt to characterize the structure that an MDP should have to make a *hierarchical* RL approach provably convenient compared to a *flat* one. Then, we develop Options-UCBVI, a novel provably efficient algorithm for learning in finite-horizon SMDPs enjoying favorable regret guarantees, which become nearly tight when applied to standard FH-MDPs. By combining Options-UCBVI and the standard UCBVI-FH algorithm in the framework of our meta-algorithm, we succeeded in achieving sublinear regret for learning at both (high and low) levels, also showing the advantages over the resolution of the FH-MDP with a flat approach. One of the main limitations of the approach lies in the need for the concentrability coefficients in the analysis of the meta-algorithm. Future works should investigate further in this direction to understand whether this represents an artifact of our analysis, a limitation of the algorithm, or an inherent challenge of the setting.

References

[1] Anonymous. An option-dependent analysis of regret minimization algorithms in finite-horizon semi-markov decision processes. 2023.

[2] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[3] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[4] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[5] M. Baykal-Gürsoy. Semi-markov decision processes. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

[6] T. G. Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

[7] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[8] R. Fruit and A. Lazaric. Exploration-exploitation in mdps with options. In *Artificial intelligence and statistics*, pages 576–584. PMLR, 2017.

[9] R. Fruit, M. Pirotta, A. Lazaric, and E. Brunskill. Regret minimization in mdps with options without prior knowledge. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] M. Ghavamzadeh, A. Lazaric, and M. Pirotta. Exploration in reinforcement learning. Tutorial at AAAI'20, 2020.

[11] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[12] A. Levy, G. Konidaris, R. Platt, and K. Saenko. Learning multi-level hierarchies with hindsight. In *Proceedings of International Conference on Learning Representations*, 2019.

[13] M. C. Machado, M. G. Bellemare, and M. Bowling. A laplacian framework for option discovery in reinforcement learning. In *International Conference on Machine Learning*, pages 2295–2304. PMLR, 2017.

[14] T. A. Mann, S. Mannor, and D. Precup. Approximate value iteration with temporally extended actions. *Journal of Artificial Intelligence Research*, 53:375–438, 2015.

[15] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[16] A. McGovern and A. G. Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. 2001.

[17] I. Menache, S. Mannor, and N. Shimkin. Q-cut—dynamic discovery of sub-goals in reinforcement learning. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 295–306. Springer, 2002.

[18] O. Nachum, S. S. Gu, H. Lee, and S. Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[19] I. Osband and B. Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

[20] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

[21] I. Pinelis. Dkw type inequality for renewal processes. MathOverflow, 2019.

[22] D. Precup and R. S. Sutton. Multi-time models for temporally abstract planning. *Advances in neural information processing systems*, 10, 1997.

[23] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[24] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[25] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[26] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pages 3540–3549. PMLR, 2017.

[27] Z. Wen, D. Precup, M. Ibrahimi, A. Barreto, B. Van Roy, and S. Singh. On efficiency in hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 33:6708–6718, 2020.

[28] A. Zanette and E. Brunskill. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *International Conference on Machine Learning*, pages 5747–5755. PMLR, 2018.

A Proof of Theorem 3.3

In this section, we will provide detailed proof of Theorem 3.3. As described in the main paper, the meta-algorithm alternates two regret minimizers $\mathfrak{A}^L, \mathfrak{A}^H$ for N stages at two levels of temporal abstractions of the problem. While learning on one level, the policies of the second are kept fixed for all the episodes on the stage.

First of all, we introduce Lemma 3.2, which relates the regret paid by the regret minimizer of one level with the bias introduced in the learning of the other level.

Lemma 3.2. *Let us define the concentrability coefficients:*

$$C^H \coloneqq \max_{n\in[N]} \inf_{\mu^* \text{ optimal}} \max_{(s,h)\in\mathcal{S}\times[H]} \frac{d^{\mu^*}_{s_1,1}(s,h)}{d^{\mu_n}_{s_1,1}(s,h)}, \tag{9}$$

$$C^L \coloneqq \max_{n\in[N]} \max_{o\in\mathcal{O}} \inf_{\pi^*_o \text{ optimal}} \max_{(s,h)\in\mathcal{I}^o} \max_{(s',h')\in\mathcal{S}_o\times[H_o]} \frac{d^{\pi^*_o}_{s,h}(s',h')}{d^{\pi^o_{n-1}}_{s,h}(s',h')}. \tag{10}$$

Then, it holds that:

$$\underbrace{V^*_*(s_1,1) - V^*_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Bias of not playing } \boldsymbol{\pi}^*} \leq C^H \Big(\underbrace{V^{\mu_n}_*(s_1,1) - V^{\mu_n}_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Regret of } \mathfrak{A}^L} \Big), \tag{11}$$

$$\underbrace{V^*_*(s_1,1) - V^{\mu_n}_*(s_1,1)}_{\text{Bias of not playing } \mu^*} \leq C^L \Big(\underbrace{V^*_{\boldsymbol{\pi}_{n-1}}(s_1,1) - V^{\mu_n}_{\boldsymbol{\pi}_{n-1}}(s_1,1)}_{\text{Regret of } \mathfrak{A}^H} \Big). \tag{12}$$

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^H_n for the high-level bias and β^L_n for the low-level bias.

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

(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 $\frac{d^{\mu^*}_{s_1,1}(s,h)}{d^{\mu_n}_{s_1,1}(s,h)}$

(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. □

The proof of theorem 3.3 directly follows from the previous lemma

Theorem 3.3. *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* **??***, Algorithm 1 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:*

$$\textit{Regret}(\textit{HLML}, K) \leq \sum_{n=1}^{N} (C^H + 1) R^L(K_n^L) + (C^L + 1) R^H(K_n^H). \tag{13}$$

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 (6) in the main paper.

$$\begin{aligned}
\text{Regret}(\text{HLML}, K) &= \sum_{n=1}^{N} \left(\sum_{k=1}^{K_n^H} \left(V_*^*(s_1, 1) - V_{\boldsymbol{\pi}_{n-1}}^{\mu_{n,k}}(s_1, 1) \right) + \sum_{k=1}^{K_n^L} \left(V_*^*(s_1, 1) - V_{\boldsymbol{\pi}_{n,k}}^{\mu_n}(s_1, 1) \right) \right) \\
&\overset{a}{=} \sum_{n=1}^{N} \left(\beta_n^H + R^H(K_n^H) + \beta_n^L + R^L(K_n^L) \right) \\
&\overset{b}{\leq} \sum_{n=1}^{N} \left(C^H R^L(K_{n-1}^L) + R^H(K_n^H) + C^L R^H(K_{n-1}^H) + R^L(K_n^L) \right) \\
&\overset{c}{\leq} \sum_{n=1}^{N} (C^H + 1) R^L(K_n^L) + (C^L + 1) R^H(K_n^H).
\end{aligned}$$

(a) We can decompose the two terms of the summation as shown in Equations (7) and (8), 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 3.2.

(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. □

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 [3] 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 4.1. *Let $\mathcal{SM}$ be an FH-SMDP with S states and O temporally extended actions (options), known reward,*[5]*, 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:*[6]

$$\textit{Regret}(\textit{O-UCBVI}, K) \leq \tilde{O}\left(H\sqrt{SOKd} + H^3 S^2 O d + H\sqrt{Kd} \right), \tag{14}$$

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:

$$Regret(K) \leq \widetilde{Regret}(K) \leq \sum^{K} \tilde{V}^{\mu_k}(s, 1) - V^{\mu_k}(s, 1) \tag{18}$$

[5]The choice of assuming a known reward is for compliance with [3]. 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.

[6]$\tilde{O}$ neglects logarithmic terms.

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,

$$\widetilde{Regret}(K) \leq \sum_{i=1}^{K}\sum_{j=1}^{H}\tilde{\Delta}_{ij}(s_{ij}) \tag{19}$$

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 4).
Now let's define properly the per step regret:

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

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

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

$$\begin{aligned}
c_{hk} &= [(\hat{P}^{\mu_k}_{hk} - P^{\mu_k}_{h})(\tilde{V}^{\mu_k}(s',h') - V^{*}(s',h')](s_{hk}) \\
&\overset{a}{=} \sum_{s'\in S}\sum_{h'\in H}(\hat{P}^{\mu_k}_{k}(s',h'|s_{hk},h) - P^{\mu_k}(s',h'|s_{hk},h))(\tilde{V}^{\mu_k}(s',h') - V^{*}(s',h')) \\
&\overset{b}{\leq} \sum_{s'\in S}\sum_{h'\in H}\left(2\sqrt{\frac{p_{hk}(s')(1-p_{hk}(s'))L}{n_k(s,o,h)}} + \frac{4L}{3n_k(s,o,h)}\right)\tilde{\Delta}_{h'k}(s') \\
&\overset{c}{\leq} 2\sqrt{L}\sum_{s'\in S}\sum_{h'\in H}\sqrt{\frac{p_{hk}(s')}{n_k(s,o,h)}}\tilde{\Delta}_{h'k}(s') + \frac{4SH^2L}{3n_k(s,o,h)} \\
&\overset{d}{=} 2\sqrt{L}\Bigg(\sum_{(s',h')\in[(s',h')]_{typ}}\sqrt{\frac{p_{hk}(s')}{n_k(s,o,h)}}\tilde{\Delta}_{h'k}(s') \\
&\quad + \sum_{(s',h')\notin[(s',h')]_{typ}}\sqrt{\frac{p_{hk}(s')}{n_k(s,o,h)}}\tilde{\Delta}_{h'k}(s')\Bigg) + \frac{4SH^2L}{3n_k(s,o,h)}
\end{aligned}$$

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

(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 [3]), $\left|\hat{P}_k^{\mu_k}(s',h'|s_{hk},h) - P^{\mu_k}(s',h'|s_{hk},h)\right| \leq 2\sqrt{\frac{p_{hk}(s')(1-p_{hk}(s'))L}{n_k(s,o,h)}} + \frac{4L}{3n_k(s,o,h)}$, 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 [3]).

(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 [3]).

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

$$
\widetilde{Regret}(K) \leq \sum_{i=1}^{K}\sum_{j=1}^{H}\left(b_{hk} + e_{hk} + \epsilon_{hk} + 2\sqrt{L}\bar{\epsilon}_{hk} + \frac{4SH^3L}{n_k(s,o,h)}\right) \tag{20}
$$

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.

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

(a) Using Bernstein Inequality. $\mathbb{V}_{hk}^* = \mathbb{V}\mathrm{ar}_{(s',h')\sim P^{\mu_k}(\cdot|s,h)}(V^*(s',h'))$ (Remember the meaning of P^{μ_k})

(b) Using Cauchy-Schwartz inequality

(c) Summing and subtracting $\mathbb{V}_{hk}^{\mu_k} = \mathbb{V}\mathrm{ar}_{(s',h')\sim P^{\mu_k}(\cdot|s,h)}(V^{\mu_k}(s',h'))$ 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 $\bar{\epsilon}_{hk}$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\epsilon_{hk} \leq H\sqrt{dKL} \tag{21}$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\epsilon}_{hk} \leq \sqrt{dK} \tag{22}$$

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).

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{4SH^3L}{n_k(s,o,h)} \leq \square H^3S^2OL^2d \tag{23}$$

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 [3]. However, to make the definition clearer, let us motivate the need for this term.
Given that the optimistic value function $\tilde{V}^{\mu_k}$ 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 $\tilde{V}^{\mu_k}$ 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 [3],

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

where $\hat{\mathbb{V}}^*_{hk} = \mathbb{V}\text{ar}_{(s',h')\sim P^{\mu_k}(\cdot|s,h)}(V^*(s',h'))$ and $\hat{\mathbb{V}}_{hk} = \mathbb{V}\text{ar}_{(s',h')\sim\hat{P}^{\mu_k}_k}(\tilde{V}^{\mu_k}(s,h))$.
We need this term to be of the same order as the estimation error e_{hk}, and thus we can say that

$$b_{hk} \sim [(\hat{P}^{\mu_k}_{hk} - P^{\mu_k}_h)V^*(s',h')](s_{hk}) \tag{24}$$

This time, however, we use the Empirical-Bernstein inequality [15] because we need the empirical variance to appear.

$$b_{hk} \leq \left(2\sqrt{\frac{\hat{\mathbb{V}}^*_{hk}L}{n_k(s,o,h)}} + \frac{14HL}{3n_k(s,o,h)}\right) \tag{25}$$

By applying Lemma 2 to this equation and substituting $\hat{\mathbb{V}}^*_{hk}$ we get the same form of bonus of [3].

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

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

$$\tilde{V}(s',h') - V^*(s',h') \leq \sqrt{b'_{hk}} \tag{26}$$

and consider that b'_{hk} has to be appropriate to guarantee an adaptation of Lemma 16 of [3], 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

$$\sqrt{b'_{hk}}\left(\sum_o n_k(s,o,h)\right) \geq \square H^4S^2OL^2 \geq \square H^3S^2OL^2d \tag{27}$$

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 $K \geq H^3L^2S^2O \geq H^2L^2S^2Od$ for typical episodes, we have:

$$b'_{hk} = \frac{100^2H^5S^2L^2O}{\sum_o n_k(s,o,h)} \tag{28}$$

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 [3].

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

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

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

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

(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 $K \geq H^2L^2S^2Od$ and thus we consider only the dominant terms.

Regarding the second term (st) adapting the proofs of [3], 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.

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

(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

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_{hk} \leq \square LH\sqrt{KSOd}+\square Ld\sqrt{HSOU_{K,1}}+H^3S^2L^2Od \tag{29}$$

neglecting smaller order terms.

B.6 Summing all the terms

Finally, we can combine all the terms analyzed separately back into Equation (20), and we will get:

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

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

C Proof of Corollary 5.1

In this section, we will proof Corollary 5.1, which is the specialization of Theorem 3.3 using Options-UCBVI as regret minimizer for the high-level problem, and UCBVI for the options learning.

Corollary 5.1. *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 1 when instanced with $\mathfrak{A}^H$=O-UCBVI and $\mathfrak{A}^L$=UCBVI-FH, run with the episode schedule as in Equation* (15)*, and having where $H_O = \max_{o\in\mathcal{O}} H_o$, is bounded with probability at least $1-\delta$ by:*

$$Regret(\textbf{HLML}, K) \leq \tilde{O}\left(C^L \underbrace{H\sqrt{SOKd}}_{\textit{High-level regret}} + C^H \underbrace{H_O\sqrt{SAH_OK}}_{\textit{Low-level regret}}\right). \tag{16}$$

Proof. For the option learning procedure, we instantiate a UCBVI algorithm for each sub-MDP $\mathcal{M}_o$, and we execute each method for K_n^L/O episodes. For the sake of the analysis, the regret paid by these O different learning procedures is upper bounded by the regret paid after the total amount of episodes provided for the low-level learning, K_n^L, in the worst possible instance of sub-MDP, the one with the longest possible horizon $H_O = \max_{o\in\mathcal{O}} H_o$. Moreover, the transition probability estimate is shared across all the options that are estimating the true transition probability p of the original MDP. Thus, we can consider a dependency on the entire state and action space. Therefore, by considering just the dominant term of the two upper bounds of regret, we can write

$$\begin{aligned}R^L_{K^L_n} &= \textit{Regret-UCBVI} \leq \tilde{O}\left(H_O\sqrt{SAK_n^L H_O}\right) \\ R^H_{K^H_n} &= \textit{Regret-O-UCBVI} \leq \tilde{O}\left(H\sqrt{SOK_n^H d}\right)\end{aligned}$$

Now by directly substituting these results in Theorem 3.3 and considering the scheduling proposed in Equation (15), we can rewrite the regret of the meta-algorithm as:

$$\begin{aligned}Regret(\text{HLML}, K) &\leq \tilde{O}\left(\sum_{n=1}^{N}\left((C^H+1)H_O\sqrt{SAH_O 2^n} + (C^L+1)H\sqrt{SOd2^n}\right)\right) \\ &= \tilde{O}\left(\left((C^H+1)H_O\sqrt{SAH_O} + (C^L+1)H\sqrt{SOd}\right)\sum_{n=1}^{N}\sqrt{2^n}\right) \\ &= \tilde{O}\left(\left((C^H+1)H_O\sqrt{SAH_O} + (C^L+1)H\sqrt{SOd}\right)2\sqrt{2}\sum_{n=0}^{N/2}2^n\right) \\ &= \tilde{O}\left(\left((C^H+1)H_O\sqrt{SAH_O} + (C^L+1)H\sqrt{SOd}\right)\left(2\sqrt{2}(2^{N/2+1}-1)\right)\right) \\ &\overset{a}{\propto} \tilde{O}\left(\left(C^H H_O\sqrt{SAH_O} + C^L H\sqrt{SOd}\right)2^{(\log_2(K))/2}\right) \\ &\leq \tilde{O}\left(\left(C^H H_O\sqrt{SAH_O} + C^L H\sqrt{SOd}\right)\sqrt{K}\right)\end{aligned}$$

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

D Useful Lemmas

Lemma D.1 (Pigeon-hole argument)**.** *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:*

$$\sum_{k=1}^{K} \mathbb{I}(k \in [k]_{typ}) \sum_{h=1}^{H} \frac{1}{n_k(s, o, h)} \leq dSO \ln(Kd) \tag{30}$$

Proof.

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

(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 $\sum_{i=1}^{n} \frac{1}{i} \sim \ln(n)$

□

The following lemmas are adaptations to SMDPs of Lemma 8, 9, and 10 of the paper of the UCBVI paper [3]. 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*

$$\sum_{i=1}^{k} \sum_{j=h}^{H} \mathbb{V}^{\mu}_{i,j'} \leq KH^2 + 2\sqrt{H^5 KL} + 4d^3/3L \tag{31}$$

Proof. The proof follows the same passages of the proof of Lemma 8 in [3], where j' is the next stage after a temporally extended transition. □

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*

$$\sum_{i=1}^{k} \sum_{j=h}^{H} \left(\mathbb{V}^{*}_{i,j'} - \mathbb{V}^{\mu}_{i,j'} \right) \leq 2HdU_k + 4H^2\sqrt{HKL} + 4d^3/3L \tag{32}$$

Proof. The proof follows the same passages of the proof of Lemma 9 in [3], where j' is the next stage after a temporally extended transition. □

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*

$$\sum_{i=1}^{k} \sum_{j=h}^{H} \left(\hat{\mathbb{V}}_{i,j'} - \mathbb{V}^{\mu}_{i,j'} \right) \leq \square H d U_{k,1} + \square H^2 S \square d^2 K L O \tag{33}$$

Proof. The proof follows the same passages of the proof of Lemma 10 in [3], 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). □