Provably (More) Sample-Efficient Offline RL with Options

Xiaoyan Hu
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Hong Kong SAR, China
xyhu21@cse.cuhk.edu.hk

Ho-fung Leung
Independent Researcher
Hong Kong SAR, China
ho-fung.leung@outlook.com

Abstract

The options framework yields empirical success in long-horizon planning problems of reinforcement learning (RL). Recent works show that options improves the sample efficiency in *online* RL where the learner can actively explores the environment. However, these results are no longer applicable to scenarios where exploring the environment online is risky, e.g., automated driving and healthcare. In this paper, we provide the first analysis of the sample complexity for offline RL with options, where the agent learns from a dataset without further interaction with the environment. We propose the **PE**ssimistic **V**alue **I**teration for Learning with **O**ptions (PEVIO) algorithm and establish near-optimal suboptimality bounds (with respect to the novel information-theoretic lower bound for offline RL with options) for two popular data-collection procedures, where the first one collects state-option transitions and the second one collects state-action transitions. We show that compared to offline RL with actions, using options not only enjoys a faster finite-time convergence rate (to the optimal value) but also attains a better performance (when either the options are carefully designed or the offline data is limited). Based on these results, we analyze the pros and cons of the data-collection procedures, which may facilitate the selection in practice.

1 Introduction

Planning in long-horizon tasks is challenging in reinforcement learning (RL) (Co-Reyes et al., 2018; Eysenbach et al., 2019; Hoang et al., 2021). A line of study proposes to accelerate learning in these tasks using temporally-extended actions (Fikes et al., 1972; Sacerdoti, 1973; Drescher, 1991; Jiang et al., 2019; Nachum et al., 2019; Machado et al., 2021; Erraqabi et al., 2022). One powerful approach is the *options* framework introduced by Sutton et al. (1999), where the agent interacts with the environment with closed-loop policies called *options*. Empirical success (Tessler et al., 2017; Vezhnevets et al., 2017) shows that options help achieve sample-efficient performance in long-horizon planning problems.

To provide a theoretical guarantee to the options framework, recent works have focused on the sample complexity of RL with options in the *online* setting, where the agent continuously explores the environment and learns a *hierarchical policy* to select options. Brunskill and Li (2014) establish a PAC-like sample complexity of RL with options in the semi-Markov decision processes (SMDPs), where temporally-extended actions are treated as indivisible and unknown units. Later, Fruit and Lazaric (2017) provide the first regret analysis of RL with options under the Markov decision processes (MDPs) framework. While their proposed algorithm attains a sublinear regret, it requires prior knowledge of the environment, which is not usually available in practice. To address this problem, Fruit et al. (2017) propose an algorithm that does not require prior knowledge, yet achieves

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

a near-optimal regret bound. However, these results are inapplicable to many real-world scenarios where online exploration is not allowed. For example, it has been argued that in healthcare (Gottesman et al., 2019) and automated driving (Shalev-Shwartz et al., 2016), learning in an online manner is risky and costly. In these scenarios, *offline* learning, where the agent learns a policy from a dataset, is preferred. We note that there is a line of studies on the sample complexity of offline RL with primitive actions only (i.e., without the use of options) (Levine et al., 2020; Fu et al., 2020; Rashidinejad et al., 2021). Unfortunately, to the best of our knowledge, there have been no results reported on the offline RL with options.

In this paper, we make the following contributions. First, we derive a novel information-theoretic lower bound, which generalizes the one for offline learning with actions. Second, we propose the **PE**ssimistic **V**alue **I**teration for Learning with **O**ptions (PEVIO) algorithm and derive near-optimal suboptimality bounds for two popular data-collection procedures, where the first one collects state-option transitions and the second one collects state-action transitions. More importantly, we show that options facilitate more sample-efficient learning in both the finite-time convergence rate and actual performance. To shed light on offline RL with options in practice, we discuss the pros and cons of both data-collection procedures based on our analysis.

2 Related Work

Learning with Options Building upon the theory of *semi-Markov decision processes* (SMDPs) (Bradtke and Duff, 1994; Mahadevan et al., 1997), Sutton et al. (1999) propose to learn with options. Following their seminal work, learning with options has been widely studied in the function approximation setting (Sorg and Singh, 2010) and hierarchical RL (Igl et al., 2020; Klissarov and Precup, 2021; Wulfmeier et al., 2021). Discovering useful options has also been the subject of extensive research (Stolle and Precup, 2002; Riemer et al., 2018; Mankowitz et al., 2018; Harb et al., 2018; Hiraoka et al., 2019; Bagaria et al., 2021). Despite its empirical success, there have been fairly limited studies on the sample efficiency of learning with options. Brunskill and Li (2014) analyze the sample complexity bound for an RMAX-like algorithm for SMDPs. Fruit and Lazaric (2017) derive the first regret analysis of learning with options. They propose an algorithm that attains sublinear regret in the infinite-horizon average-reward MDP while requiring prior knowledge of the environment. Later, Fruit et al. (2017) remove this requirement.

Offline RL In the offline setting, a dataset that is collected by executing a *behavior policy* in the environment is provided, and the agent is asked to learn a near-optimal policy using only this dataset. A key challenge in offline RL is the insufficient coverage of the dataset (Wang et al., 2021), which is also known as distributional shift (Chen and Jiang, 2019; Levine et al., 2020). To address this problem, the previous study on sample-efficient learning assumes uniform coverage of the dataset (Liu et al., 2018; Chen and Jiang, 2019; Jiang and Huang, 2020; Yang et al., 2020; Xie and Jiang, 2020; Uehara et al., 2020; Qu and Wierman, 2020; Yin et al., 2021). This assumption is relaxed in recent works by pessimism principle (Xie et al., 2021; Rashidinejad et al., 2021; Jin et al., 2021).

3 Preliminaries

3.1 Episodic MDP with Options

Let $\Delta(\mathcal{X})$ denote the probability simplex on space $\mathcal{X}$ and $[N] := \{1, \cdots, N\}$ for any positive integer N. An episodic MDP with options is a sextuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mathcal{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the (primitive) action set, $\mathcal{O}$ the finite set of options, H the length of each episode, $\mathcal{P} = \{P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})\}_{h\in[H]}$ the transition kernel, $r = \{r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]\}_{h\in[H]}$ the deterministic reward function.[1] We define $S := |\mathcal{S}|$, $A := |\mathcal{A}|$, and $O := |\mathcal{O}|$. A (Markov) *option* $o \in \mathcal{O}$ is a pair (π^o, β^o) where $\pi^o = \{\pi^o_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h\in[H]}$ is the option's policy and $\beta^o = \{\beta^o_h : \mathcal{S} \mapsto [0, 1]\}_{h\in[H]}$ is the probability of the option's *termination*. For convenience, we define $\beta^o_{H+1}(s) = 1$ for all $(s, o) \in \mathcal{S} \times \mathcal{O}$, i.e., any option is terminated after the end of an

[1] Our results can be directly generalized to stochastic rewards.

episode. We assume that the initial state s_1 is *fixed*.[2] Upon arriving at state s_h at any timestep $h \in [H]$, if $h = 1$ (at the beginning of an episode), the agent selects option $o_1 \sim \mu_1(\cdot|s_1)$, where $\mu = \{\mu_h : \mathcal{S} \mapsto \Delta(\mathcal{O})\}_{h\in[H]}$ is a *hierarchical policy* to select an option at each state. Otherwise ($h \geq 2$), the agent first terminates option o_{h-1} with probability $\beta_h^{o_{h-1}}(s_h)$. If option o_{h-1} is terminated, she then selects a new option $o_h \sim \mu_h(\cdot|s_h)$ according to the hierarchical policy μ. If option o_{h-1} is not terminated, the agent continues to *use* option o_{h-1} at timestep h, i.e., $o_h = o_{h-1}$. After that, the agent takes action $a_h \sim \pi_h^{o_h}(\cdot|s_h)$, receives a reward $r_h := r_h(s_h, a_h)$, and transits to the next state $s_{h+1} \sim P_h(\cdot|s_h, a_h)$. An episode terminates at timestep $H + 1$. A special case is that an action a is an option o, such that $\pi_h^o(a|s) = 1$ and $\beta_h^o(s) = 1$ for any $(h, s) \in [H] \times \mathcal{S}$. For convenience, we use the notation $\mathcal{O} = \mathcal{A}$ to represent that each option corresponds to an action, which is the case in RL with primitive actions.

To define the Q-function and the value function, we introduce some useful notations.[3] Let $T = \{T_h : \mathcal{S} \times \mathcal{O} \mapsto \Delta(\mathcal{S} \times [H - h + 1])\}_{h\in[H]}$ and $U = \{U_h : \mathcal{S} \times \mathcal{O} \mapsto [0, H]\}_{h\in[H]}$ denote the *option transition function* and the *option utility function*, respectively. Particularly, for any $(h, s, o) \in [H] \times \mathcal{S} \times \mathcal{O}$, the option transition function $T_h(s'|s, o, \tau)$ is the probability that the agent uses option o at state s at timestep h, reaches state s' at timestep $h + \tau$ without terminating option o in these τ timesteps, and finally terminates option o at state s' at timestep $h + \tau$. The option utility function $U_h(s, o)$ is the expected cumulative reward within timesteps that the option is used without being terminated. Given any arbitrary series of functions $\{y_h : \mathcal{S} \mapsto \mathbb{R}\}_{h\in[H]}$, define the operator $[T_h y_{h+\tau}](s, o) := \sum_{s'\in\mathcal{S}} T_h(s'|s, o, \tau) y_{h+\tau}(s')$ for any $(s, o, \tau) \in \mathcal{S} \times \mathcal{O} \times [H - h + 1]$. In the following, we derive the Q-function and the value function for learning with options. (The detailed proof can be found in Appendix B.)

Theorem 1 (Q-function and value function)**.** *For any hierarchical policy μ and $(h, s, o) \in [H] \times \mathcal{S} \times \mathcal{O}$, the Q-function is given by*

$$Q_h^\mu(s, o) := \mathbb{E}_\mu \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, o_h = o \right] = U_h(s, o) + \sum_{\tau\in[H-h+1]} [T_h V_{h+\tau}^\mu](s, o) \tag{1}$$

and the value function is given by

$$V_h^\mu(s) := \mathbb{E}_\mu \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, o_h \sim \mu_h(\cdot|s_h) \right] = \sum_{o\in\mathcal{O}} \mu_h(o|s) Q_h^\mu(s, o) \tag{2}$$

where $V_{H+1}^\mu(s) = Q_{H+1}^\mu(s, o) = 0$ for any $(s, o) \in \mathcal{S} \times \mathcal{O}$.

Intuitively, the first term $U_h(s, o)$ of the Q-function is the expected reward within timesteps that option o is used without being terminated, and the second term $\sum_{\tau\in[H-h+1]} [T_h V_{h+\tau}^\mu](s, o)$ corresponds to the expected reward within timesteps after option o is terminated and a new option is selected according to μ. It can be shown that there exists an optimal (and deterministic) hierarchical policy $\mu^* = \{\mu_h^* : \mathcal{S} \mapsto \mathcal{O}\}_{h\in[H]}$ that attains the optimal value function, i.e., $V_h^*(s) = \sup_\mu V_h^\mu(s)$ for all $(h, s) \in [H] \times \mathcal{S}$ (Sutton et al., 1999).

3.2 Offline RL with Options

We consider learning with options in the offline setting. That is, given a dataset $\mathcal{D}$ that is collected by an experimenter through interacting with the environment, the algorithm outputs a hierarchical policy $\widehat{\mu}$. The sample complexity is measured by the *suboptimality*, i.e., the shortfall in the value function of the hierarchical policy $\widehat{\mu}$ compared to that of the optimal hierarchical policy μ^*, which is given by

$$\text{SubOpt}_\mathcal{D}(\widehat{\mu}, s_1) := V_1^*(s_1) - V_1^{\widehat{\mu}}(s_1) \tag{3}$$

To derive a novel information-theoretic lower bound of $\text{SubOpt}_\mathcal{D}$, we first define some useful notations. For any hierarchical policy μ, we denote by $\theta^\mu = \{\theta_h^\mu : \mathcal{S} \times \mathcal{O} \mapsto [0, 1]\}_{h\in[H]}$ its *state-option*

[2]Note that any H-length episodic MDP with a stochastic initial state is equivalent to an $(H + 1)$-length MDP with a fixed initial state s_0.

[3]The formal definitions can be found in Appendix A.

occupancy measure. That is, $\theta_h^\mu(s,o)$ is the probability that the agent selects a particular option o at state s at timestep h (either when $h=1$ or when the option o_{h-1} used at the timestep $h-1$ is terminated) when following the hierarchical policy μ. With a slight abuse of the notation, we denote by $\theta_h^\mu(s) := \sum_{o\in\mathcal{O}} \theta_h^\mu(s,o)$ the *state occupancy measure* for any $(h,s)\in[H]\times\mathcal{S}$. Further, we define

$$Z_{\mathcal{O}}^\mu := \sum_{h,s} \theta_h^\mu(s),\ \overline{Z}_{\mathcal{O}}^\mu := \sum_{h,s,o} \mathbb{I}[\theta_h^\mu(s,o)>0] \tag{4}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Intuitively, $Z_{\mathcal{O}}^\mu$ is the expected number of timesteps to alternate a new option and $\overline{Z}_{\mathcal{O}}^\mu$ is the maximal number of state-option pairs that can be visited, when following the hierarchical policy μ. The following proposition shows that options facilitate temporal abstraction and reduction of the state space.

Proposition 1. *For any hierarchical policy μ, we have that $Z_{\mathcal{O}}^\mu \le H$. If μ is deterministic, i.e., $\mu=\{\mu_h:\mathcal{S}\mapsto\mathcal{O}\}_{h\in[H]}$, we further have that $\overline{Z}_{\mathcal{O}}^\mu \le HS$. All the above equalities hold when $\mathcal{O}=\mathcal{A}$.*

Next, we derive a novel information-theoretic lower bound of $\text{SubOpt}_{\mathcal{D}}$. The detailed proof can be found in Appendix C.

Theorem 2 (Information-theoretic lower bound)**.** *Let $\rho=\{\rho_h:\mathcal{S}\mapsto\Delta(\mathcal{O})\}_{h\in[H]}$ denote any hierarchical behavior policy to collect the dataset. Define the class of problem instances*

$$\mathcal{M}(C^{\text{option}}, z^*, \overline{z}^*) := \Bigg\{(M,\rho):\ \textit{Exists deterministic } \mu^* \textit{ of an episodic MDP } M$$
$$\textit{such that } \max_{h,s,o}\frac{\theta_h^{\mu^*}(s,o)}{\theta_h^{\rho}(s,o)} \le C^{\text{option}}, Z_{\mathcal{O}}^{\mu^*}\le z^*, \overline{Z}_{\mathcal{O}}^{\mu^*}\le \overline{z}^*\Bigg\}.$$

Suppose that $C^{\text{option}}\ge 2$, $z^*\ge 1$, and $\overline{z}^*\ge\lfloor z^*\rfloor S$, where $\lfloor x\rfloor := \max\{n\in\mathbb{N}: n\le x\}$ is the largest integer no greater than $x\in\mathbb{R}$. Then, there exists an absolute constant c_0 such that for any offline algorithm that outputs a hierarchical policy $\widehat{\mu}$, if the number of episodes

$$K \le \frac{c_0\cdot C^{\text{option}} H z^*\overline{z}^*}{\epsilon^2}$$

then there exists a problem instance $(M,\rho)\in\mathcal{M}(C^{\text{option}}, z^*, \overline{z}^*)$ on which the hierarchical policy $\widehat{\mu}$ suffers from ϵ-suboptimality, that is,

$$\mathbb{E}_M[\text{SubOpt}_{\mathcal{D}_1}(\hat{\mu}, s)] \ge \epsilon$$

where the expectation $\mathbb{E}_M$ is with respect to the randomness during the execution of ρ within MDP M.

Theorem 2 shows that, when dataset $\mathcal{D}$ sufficiently covers the trajectories induced by μ^*, i.e., $\max_{h,s,o}\theta_h^{\mu^*}(s,o)/\theta_h^\rho(s,o)\le C^{\text{option}}$, at least $\Omega(C^{\text{option}} H Z_{\mathcal{O}}^* \overline{Z}_{\mathcal{O}}^*/\epsilon^2)$ episodes are required to learn an ϵ-optimal hierarchical policy from dataset $\mathcal{D}$. Note that when $\mathcal{O}=\mathcal{A}$, it recovers the lower bound $\Omega(H^3 SC^*/\epsilon^2)$ for offline RL with primitive actions, where C^* is the concentrability defined therein.

4 The PEVIO Algorithm

Inspired by the Pessimistic Value Iteration (PEVI) algorithm (Jin et al., 2021), we propose the **PE**ssimistic **V**alue **I**teration for Learning with **O**ptions (PEVIO) in Algorithm 1. Given a dataset $\mathcal{D}$ and the corresponding **O**ffline **O**ption **E**valuation (OOE) subroutine, whose details are specified in Sections 5.1 and 5.2, PEVIO outputs a hierarchical policy $\widehat{\mu}=\{\widehat{\mu}_h:\mathcal{S}\mapsto\Delta(\mathcal{O})\}_{h\in[H]}$.

To estimate the Q-function, given a dataset $\mathcal{D}$, PEVIO first constructs $(\widehat{T},\widehat{U},\Gamma)$ by the OOE subroutine (line 3). Specifically, $\widehat{T}_h$ and $\widehat{U}_h$ are the empirical counterparts of T_h and U_h presented in the Q-function given by Equation (1), respectively. In addition, Γ is a penalty function computed based on dataset $\mathcal{D}$. We remark that the OOE subroutine varies when different data-collecting procedures

Algorithm 1 PEssimistic Value Iteration for Learning with Options (PEVIO)

1: Input: Dataset $\mathcal{D}$ and the corresponding **O**ffline **O**ption **E**valuation (OOE) subroutine.
2: Initialize: $\widehat{Q}_h(s,o) \leftarrow 0, \widehat{V}_h(s) \leftarrow 0, \widehat{V}_{H+1}(s) \leftarrow 0$ for any $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$.
3: $(\widehat{T}, \widehat{U}, \Gamma) \leftarrow \mathrm{OOE}(\mathcal{D})$.
4: **for** $h = H, H-1, \cdots, 1$ **do**
5:     **for** $(s,o) \in \mathcal{S} \times \mathcal{O}$ **do**
6:         $\overline{Q}_h(s,o) \leftarrow \widehat{U}_h(s,o) + \sum_{\tau=1}^{H-h+1} [\widehat{T}_h \widehat{V}_{h+\tau}](s,o) - \Gamma_h(s,o)$.
7:         $\widehat{Q}_h(s,o) \leftarrow \max\{0, \min\{\overline{Q}_h(s,o), H-h+1\}\}$.
8:     **end for**
9:     **for** $s \in \mathcal{S}$ **do**
10:         $\widehat{\mu}_h(\cdot|s) \leftarrow \arg\max_{\mu_h} \langle \widehat{Q}_h(s,\cdot), \mu_h(\cdot|s) \rangle_{\mathcal{O}}$.
11:         $\widehat{V}_h(s) \leftarrow \langle \widehat{Q}_h(s,\cdot), \widehat{\mu}_h(\cdot|s) \rangle_{\mathcal{O}}$.
12:     **end for**
13: **end for**
14: Output: $\widehat{\mu} = \{\widehat{\mu}_h\}_{h\in[H]}$.

are considered and we provide the details in Sections 5.1 and 5.2, respectively. Given $\widehat{U}_h$, $\widehat{T}_h$, and Γ_h, the estimated Q-function $\widehat{Q}_h$ is the derived (lines 6 and 7). Particularly, $\overline{Q}_h$ computed in line 6 can be seen as first replacing U_h and T_h with their empirical counterparts $\widehat{U}_h$, $\widehat{T}_h$ in Equation (1), and then subtracting the penalty function Γ_h. Further, a hierarchical policy $\widehat{\mu}_h$ is constructed greedily with $\widehat{Q}_h$ (line 10), where $\langle f(\cdot), g(\cdot) \rangle_{\mathcal{O}} := \sum_{o\in\mathcal{O}} f(o)g(o)$ for any arbitrary functions f, g defined on $\mathcal{O}$. Finally, given $\widehat{Q}_h$ and $\widehat{\mu}_h$, the corresponding estimated value function $\widehat{V}_h$ is computed (line 11). To analyze the suboptimality of the hierarchical policy $\widehat{\mu}$ output from PEVIO, we first provide the the following definition, which motivates the design of the penalty function Γ.

Definition 1 (ξ-uncertainty quantifier for dataset $\mathcal{D}$)**.** The penalty function $\Gamma = \{\Gamma_h : \mathcal{S} \times \mathcal{O} \mapsto \mathbb{R}^+\}_{h\in[H]}$ output from the OOE subroutine in Algorithm 1 (line 3) is said to be a ξ-uncertainty quantifier with respect to $\mathbb{P}_{\mathcal{D}}$ if the following event

$$\begin{aligned}\mathcal{E} =&\{|\widehat{U}_h(s,o) - U_h(s,o) + \sum_{\tau=1}^{H-h+1} [(\widehat{T}_h - T_h)\widehat{V}_{h+\tau}](s,o)| \\ &\leq \Gamma_h(s,o) \text{ for all } (h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}\}\end{aligned} \tag{5}$$

satisfies that $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$, where $\mathbb{P}_{\mathcal{D}}$ is the joint distribution of the data collecting process.

In other words, the penalty function Γ is a ξ-uncertainty quantifier if it upper bounds the estimation errors in the empirical option transition function $\widehat{T}$ and the empirical option utility function $\widehat{U}$. Next, we show that the suboptimality of $\widehat{\mu}$ output from PEVIO is upper bounded if Γ is a ξ-uncertainty quantifier. (The detailed proof can be found in Appendix D.)

Theorem 3 (Suboptimality of learning with options using dataset $\mathcal{D}$)**.** *Let $\widehat{\mu}$ denote the hierarchical policy output by Algorithm 1. Suppose that $\Gamma = \{\Gamma_h\}_{h\in[H]}$ output from the OOE subroutine is a ξ-uncertainty quantifier. Conditioned on the successful event $\mathcal{E}$ defined in Equation (5), which satisfies that $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$, we have that*

$$\mathrm{SubOpt}_{\mathcal{D}}(\widehat{\mu}, s_1) \leq 2 \sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\Gamma_h(s_h, o_h)|s_1] \tag{6}$$

where $\mathbb{E}_{\mu^*}[g(s_h, o_h)] = \sum_{(s,o)} \theta_h^{\mu^*}(s,o) g(s,o)$ for any $h \in [H]$ and arbitrary function $g : \mathcal{S} \times \mathcal{O} \mapsto \mathbb{R}$.

Remark 1. Since the temporal structure of learning with options is much more complex than learning with actions, PEVIO is significantly different from the algorithms proposed for offline RL with primitive actions, such as PEVI (Jin et al., 2021) or VI-LCB (Xie et al., 2021), despite sharing a similar intuition. First, in terms of the algorithm design, PEVI and VI-LCB estimate (one-step) transition kernel and reward function to compute the Q-function of a state-action pair. However, by Equation (1), the Q-function of a state-option pair depends on multi-step transitions and rewards.

Hence, it is challenging to design the OOE subroutine and analyze the estimated $(\widehat{T}, \widehat{U}, \Gamma)$ for options. Indeed, if the dataset contains (s, o, u) (state-option-utility) tuples, then (T, U, Γ) can be estimated similarly to the case of learning with primitive actions. However, if the dataset contains only (s, a, r) (state-action-reward) tuples, then it remains elusive to estimate and analyze (T, U, Γ). Second, in terms of the suboptimality analysis, previous works on offline RL with primitive actions rely on the extended value difference lemma (Cai et al., 2020, Lemma 4.2), which also depends on the one-step temporal structure of actions and cannot be directly applied to our setting. Hence, to derive Theorem 3, it is non-trivial to generalize the extended value difference lemma to the options framework (See Lemma 8 in the Appendices).

5 Data-Collection and Suboptimality Analysis

In this section, we consider two data-collection procedures that are widely deployed in the options literature. The first one collects state-option-utility tuples (dataset $\mathcal{D}_1$) and similar datasets are utilized in the work of Zhang et al. (2023). The second one collects state-action-reward tuples (dataset $\mathcal{D}_2$) and is studied in a line of works (Ajay et al., 2021; Villecroze et al., 2022; Salter et al., 2022). Intuitively, dataset $\mathcal{D}_1$ requires smaller storage and enables efficient evaluation of the options, while dataset $\mathcal{D}_2$ provides richer information on the environment and even facilitates the evaluation of new options. For each dataset, we design the corresponding OOE subroutine and derive a suboptimality bound for the PEVIO algorithm. Based on these results, we further discuss the advantages and the disadvantages of both data-collection procedures, which sheds light on offline RL with options in practice.

5.1 Learning from State-Option Transitions

We consider dataset $\mathcal{D}_1 := \{(s^k_{t^k_i}, o^k_{t^k_i}, u^k_{t^k_i})\}_{i\in[j^k],k\in[K]}$ consisting of state-option-utility tuples, which is collected by the experimenter's interaction with the environment for K episodes using a hierarchical behavior policy $\rho = \{\rho_h : \mathcal{S} \mapsto \Delta(\mathcal{O})\}_{h\in[H]}$. More precisely, at timestep t^k_i of the kth episode, the experimenter selects a new option $o^k_{t^k_i}$, uses it for $(t^k_{i+1} - t^k_i)$ timesteps, collects a cumulative reward of $u^k_{t^k_i}$ within these $(t^k_{i+1} - t^k_i)$ timesteps, and finally terminates this option at state $s^k_{t^k_{i+1}}$ at timestep t^k_{i+1}. For convenience, we define $t^k_{j^k+1} := H + 1$ for any $k \in [K]$.

Let $a \vee b := \max\{a, b\}$ for any pair of integers $a, b \in \mathbb{N}$. When dataset $\mathcal{D}_1$ is available, the OOE subroutine in Algorithm 1 is given by Subroutine 2. Particularly, Subroutine 2 incorporates the data splitting technique (Xie et al., 2021) (line 2). That is, given dataset $\mathcal{D}_1$, the algorithm randomly splits it into H subdatasets $\{\mathcal{D}_{1,h}\}_{h\in[H]}$. Then $\widehat{T}_h$ and $\widehat{U}_h$ are constructed using subdataset $\mathcal{D}_{1,h}$ (lines 4-7).

To derive the suboptimality for $\widehat{\mu}$ output from PEVIO, we follow the previous study and make a standard assumption on the coverage of dataset $\mathcal{D}_1$.

Assumption 1 (Single hierarchical policy concentrability for dataset $\mathcal{D}_1$)**.** The experimenter collects dataset $\mathcal{D}_1$ by following a hierarchical behavior policy $\rho = \{\rho_h : \mathcal{S} \mapsto \Delta(\mathcal{O})\}_{h\in[H]}$. There exists some deterministic optimal hierarchical policy μ^* such that

$$C_1^{\text{option}} := \max_{h,s,o} \frac{\theta_h^{\mu^*}(s, o)}{\theta_h^{\rho}(s, o)} \tag{7}$$

(with the convention $0/0 = 0$) is finite.

In other words, Assumption 1 states that dataset $\mathcal{D}_1$ sufficiently covers the trajectories of state-option-utility tuples induced by some deterministic optimal hierarchical policy μ^*. We derive an upper bound of $\text{SubOpt}_{\mathcal{D}_1}$ in the following theorem. (The detailed proof can be found in Appendix E.)

Theorem 4 (Suboptimality for dataset $\mathcal{D}_1$)**.** *Under Assumption 1, with probability at least $1 - \xi$, we have that*

$$\text{SubOpt}_{\mathcal{D}_1}(\widehat{\mu}, s_1) \leq \tilde{O}\left(\sqrt{\frac{C_1^{\text{option}} H^3 Z^*_{\mathcal{O}} \overline{Z}^*_{\mathcal{O}}}{K}}\right) \tag{8}$$

Subroutine 2 **O**ffline **O**ption **E**valuation (OOE) for Dataset $\mathcal{D}_1$

1: **Input:** Dataset $\mathcal{D}_1 = \{(s^k_{t^k_i}, o^k_{t^k_i}, u^k_{t^k_i})\}_{i\in[j^k],k\in[K]}$.
2: **Initialize:** Randomly split the dataset $\mathcal{D}$ into H subdatasets $\{\mathcal{D}_{1,h}\}_{h\in[H]}$ with $|\mathcal{D}_{1,h}| = K/H$. More precisely, let $l := \{l_h\}_{h\in[H]}$ be a random partition of the set $[K]$, where $l_h := \{l_{h,j}\}_{j\in[K/H]} \subset [K]$ is uniformly sampled from $[K]$ such that $\cup_{h\in[H]} l_h = [K]$ and $l_h \cap l_{h'} = \emptyset$ for any $h \neq h'$. Then we have that $\mathcal{D}_{1,h} = \{(s^k_{t^k_i}, o^k_{t^k_i}, u^k_{t^k_i})\}_{i\in[j^k],k\in l_h}$ for any $h \in [H]$. Let $n_h(s,o) := \sum_{k\in l_h} \mathbb{I}[h \in \{t^k_i\}_{i\in[j^k]}, s^k_h = s, o^k_h = o]$ denote the number of times that the experimenter selects a particular option o at state s at timestep h in subdataset $\mathcal{D}_{1,h}$.
3: **for** $(h,s,o) \in [H]\times\mathcal{S}\times\mathcal{O}$ **do**
4: **for** $(s',\tau) \in \mathcal{S}\times[H-h+1]$ **do**
5: $\widehat{T}_h(s'|s,o,\tau) \leftarrow \frac{\sum_{k\in l_h} \mathbb{I}[h\in\{t^k_i\}_{i\in[j^k]}, s^k_h=s, o^k_h=o, s^k_{h+\tau}=s']}{1\vee n_h(s,o)}$
6: **end for**
7: $\widehat{U}_h(s,o) \leftarrow \frac{\sum_{k\in l_h} \mathbb{I}[h\in\{t^k_i\}_{i\in[j^k]}, s^k_h=s, o^k_h=o] u^k_h}{1\vee n_h(s,o)}$
8: $\Gamma_h(s,o) \leftarrow \tilde{O}\left(\sqrt{\frac{H^2}{n_h(s,o)\vee 1}}\right)$
9: **end for**
10: **Output:** $(\widehat{T} = \{\widehat{T}_h\}_{h\in[H]}, \widehat{U} = \{\widehat{U}_h\}_{h\in[H]}, \Gamma = \{\Gamma_h\}_{h\in[H]})$.

where $Z^*_{\mathcal{O}} := Z^{\mu^*}_{\mathcal{O}}$ *and* $\overline{Z}^*_{\mathcal{O}} := \overline{Z}^{\mu^*}_{\mathcal{O}}$.

Compared to the lower bound in Theorem 2, suboptimality bound (8) is near-optimal except for an extra factor of H.[4] More importantly, it shows that learning with options enjoys a faster convergence rate to the optimal value than learning with primitive actions. Recall that the VI-LCB algorithm (Xie et al., 2021) that learns with primitive actions attains the suboptimality bound $\tilde{O}(\sqrt{H^5SC^*/K})$, where C^* is the concentrability defined therein. When ignoring the concentrability parameters, the suboptimality bound (8) is smaller since $Z^*_{\mathcal{O}} \le H$ and $\overline{Z}^*_{\mathcal{O}} \le HS$.

Remark 2. While, in the worst case, both $Z^*_{\mathcal{O}}$ and $\overline{Z}^*_{\mathcal{O}}$ can scale with H and HS, respectively, we note that in many long-horizon planning problems, they often scale with the *number of sub-tasks*, which are greatly smaller, especially for tasks that enables temporal abstraction and the reduction of the state space. For example, while the route-planning task of going from City A to City B by transportation takes thousands of primitive actions to finish, it can be efficiently solved by decomposing into the following sub-tasks: (1) going to the airport/train station in City A; (2) taking transportation to City B; and (3) reaching the final destination in City B, for which options are designed. In this case, both $Z^*_{\mathcal{O}}$ and $\overline{Z}^*_{\mathcal{O}}/S$ may only scale as $o(H)$. In other words, options facilitate more sample-efficient learning through temporal abstraction, i.e., sticking to an option until a sub-task is finished. Another concrete example is solving a maze, where options are often designed to move agents to bottleneck states (Şimşek and Barto, 2004; Solway et al., 2014; Machado et al., 2017) that connect different densely connected regions of the state space, e.g., doorways. In this case, while the number of option switches may grow proportionally to H, i.e., $Z^*_{\mathcal{O}}/H = O(1)$, the number of states to switch options can be greatly smaller than S, i.e., $\overline{Z}^*_{\mathcal{O}}/H = o(s)$. That is to say, options help improve the sample complexity by the reduction of the state space.

Further, we show that learning with options attains a better performance than learning with primitive actions, when either the options are carefully designed or the offline data is limited.

Corollary 1 (Better performance)**.** *Let* $\mathrm{TrueSubOpt}_{\mathcal{D}_1}(\widehat{\mu}, s_1) := V^{*,\mathrm{pri}}_1(s_1) - V^{\widehat{\mu}}_1(s_1)$*, where* $V^{*,\mathrm{pri}}$ *is the optimal value function defined for the primitive actions. Ignoring the concentrability parameters, we have that* $\mathrm{TrueSubOpt}_{\mathcal{D}_1}(\widehat{\mu}, s_1) \le \tilde{O}(\sqrt{H^5SC^*/K})$ *attained by the VI-LCB algorithm (Xie et al., 2021), when either the options are carefully designed (i.e.,* $\Delta_{\mathcal{O}}(s_1) := V^{*,\mathrm{pri}}_1(s_1) - V^*_1(s_1) = 0$*) or*

[4] We note that the extra factor H in the suboptimality bound (8) can be reduced by applying the reference-advantage decomposition technique (Xie et al., 2021).

the number K of trajectories in the dataset is

$$o\left(\frac{H^3}{\Delta_{\mathcal{O}}^2}\left(\sqrt{H^2SC^*}-\sqrt{C_1^{\text{option}}Z_{\mathcal{O}}^*\overline{Z}_{\mathcal{O}}^*}\right)_+^2\right)$$

where $(x)_+ := \max\{0, x\}$ *for any* $x \in \mathbb{R}$.

Corollary 1 implies that when data is limited, e.g., in cases where the data collection is highly expensive or risky, learning with options is beneficial since the output hierarchical policy yields a higher value than learning with primitive actions.

5.2 Learning from State-Action Transitions

We consider dataset $\mathcal{D}_2 := \{(s_h^k, a_h^k, r_h^k)\}_{h\in[H],k\in[K]}$ consisting of state-action-reward tuples, which is collected by an experimenter's interaction with the environment for K episodes using any arbitrary behavior policy. That is, the experimenter takes action a_h^k at state s_h^k at timestep h of the kth episode, receives a reward of r_h^k, and transits to state s_{h+1}^k.

When dataset $\mathcal{D}_2$ is provided, the OOE subroutine in Algorithm 1 is given by Subroutine 3. Note that one difficulty is that we cannot directly estimate the option transition function and the option utility function from dataset $\mathcal{D}_2$ as it only includes the information of the primitive actions. Hence, Subroutine 3 first constructs the empirical transition kernel $\widehat{P} = \{\widehat{P}_h\}_{h\in[H]}$ and the empirical reward function $\widehat{r} = \{\widehat{r}_h\}_{h\in[H]}$ (lines 4-7), and use them to further construct $\widehat{T}_h$ and $\widehat{U}_h$ (lines 8-20). To

Subroutine 3 Offline Option Evaluation (OOE) for Dataset $\mathcal{D}_2$

1: **Input:** Dataset $\mathcal{D}_2 = \{(s_h^k, a_h^k, r_h^k)\}_{h\in[H],k\in[K]}$.
2: **Initialize:** $\widehat{U}_{H+1}(\cdot) \leftarrow 0$. Let $N_h(s,a) := \sum_{k=1}^K \mathbb{I}[s_h^k = s, a_h^k = a]$ denote the number of visits of state-action pair (s,a) at timestep h in dataset $\mathcal{D}_2$. Function $\phi := \{\phi_h : \mathcal{S}\times\mathcal{O} \mapsto \mathbb{R}\}_{h\in[H]}$ is given by Equation (32) in the Appendices.
3: **for** $h = H, H-1, ..., 1$ **do**
4:     **for** $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**
5:         $\widehat{P}_h(s'|s,a) \leftarrow \frac{\sum_{k\in[K]}\mathbb{I}[s_h^k=s,a_h^k=a,s_{h+1}^k=s']}{1\vee N_h(s,a)}$ for any $s' \in \mathcal{S}$
6:         $\widehat{r}_h(s,a) \leftarrow \mathbb{I}[N_h(s,a) \geq 1]r_h(s,a)$
7:     **end for**
8:     **for** $(s,o,s') \in \mathcal{S}\times\mathcal{O}\times\mathcal{S}$ **do**
9:         $\widehat{T}_h(s'|s,o,1) \leftarrow \beta_{h+1}^o(s')\sum_{a\in\mathcal{A}}\pi_h^o(a|s)\widehat{P}_h(s'|s,a)$
10:         $\overline{\widehat{T}}_h(s'|s,o,1) \leftarrow (1-\beta_{h+1}^o(s'))\sum_{a\in\mathcal{A}}\pi_h^o(a|s)\widehat{P}_h(s'|s,a)$
11:         **for** $l = 2, \cdots, H-h+1$ **do**
12:             $\widehat{T}_h(s'|s,o,l) \leftarrow \sum_a \pi_h^o(a|s)\sum_{s''}\widehat{P}_h(s''|s,a)\left(1-\beta_{h+1}^o(s'')\right)\widehat{T}_{h+1}(s'|s'',o,l-1)$
13:             $\overline{\widehat{T}}_h(s'|s,o,l) \leftarrow \sum_a \pi_h^o(a|s)\sum_{s''}\widehat{P}_h(s''|s,a)\left(1-\beta_{h+1}^o(s'')\right)\overline{\widehat{T}}_{h+1}(s'|s'',o,l-1)$
14:         **end for**
15:     **end for**
16: **end for**
17: **for** $(h,s,o) \in [H]\times\mathcal{S}\times\mathcal{O}$ **do**
18:     $\Gamma_h(s,o) \leftarrow \tilde{O}\left(\sqrt{\sum_{m=h}^H\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}\frac{H^3S}{N_m(s,a)\vee 1}} + H\phi_h(s,o)\right)$
19:     $\widehat{U}_h(s,o) \leftarrow \sum_{a\in\mathcal{A}}\pi_h^o(a|s)\widehat{r}_h(s,a) + \sum_{s'\in\mathcal{S}}\overline{\widehat{T}}_h(s'|s,o,1)\widehat{U}_{h+1}(s',o)$
20: **end for**
21: **Output:** $(\widehat{T} = \{\widehat{T}_h\}_{h\in[H]}, \widehat{U} = \{\widehat{U}_h\}_{h\in[H]}, \Gamma = \{\Gamma_h\}_{h\in[H]})$.

derive the suboptimality, we first define some useful notations. For any $(h,s,o) \in [H]\times\mathcal{S}\times\mathcal{O}$ and $h \leq m \leq H$, we denote by $\mathcal{X}_{h,s,o}^m$ the set of state-action pairs that can be reached at timestep m by using option o at state s and timestep h without being terminated.[5] Further, let $d^\rho := \{d_h^\rho : \mathcal{S}\times\mathcal{A} \mapsto$

[5]For convenience, we assume that $\mathcal{X}_{h,s,o}^m$ is known prior. When $\mathcal{X}_{h,s,o}^m$ is unknown, it can be replaced by a superset that does not require prior knowledge and our results directly follow (See Remark 3 in Appendix F).

$[0,1]\}_{h\in[H]}$ denote the state-action distribution of the behavior policy ρ used by the experimenter. That is, $d_h^\rho(s,a)$ is the probability that the agent takes action a at state s at timestep h. Similarly, we make the following assumption on dataset $\mathcal{D}_2$.

Assumption 2 (Single hierarchical policy concentrability for dataset $\mathcal{D}_2$)**.** The experimenter collects dataset $\mathcal{D}_2$ by following an arbitrary behavior policy ρ. There exists some deterministic optimal hierarchical policy μ^* such that

$$C_2^{\text{option}} := \max_{h,s,o} \sum_{h\le m\le H,(s',a')\in\mathcal{X}_{h,s,o}^m} \frac{\theta_h^{\mu^*}(s,o)}{d_m^\rho(s',a')} \tag{9}$$

(with the convention $0/0=0$) is finite.

Intuitively, Assumption 2 states that dataset $\mathcal{D}_2$ sufficiently covers the trajectories of state-action-reward tuples induced by the optimal hierarchical policy μ^*. Next, we derive an upper bound of $\text{SubOpt}_{\mathcal{D}_2}$ under Assumption 2. The detailed proof can be found in Appendix F.

Theorem 5 (Suboptimality for dataset $\mathcal{D}_2$)**.** *Under Assumption 2, with probability at least $1-\xi$, we have that*

$$\text{SubOpt}_{\mathcal{D}_2}(\widehat{\mu}, s_1) \le \tilde{O}\left(\sqrt{\frac{C_2^{\text{option}} H^3 S Z_{\mathcal{O}}^* \overline{Z}_{\mathcal{O}}^*}{K}} + \frac{H^2 SOC_2^{\text{option}}}{K}\right) \tag{10}$$

which translates to

$$\tilde{O}\left(\sqrt{\frac{C_2^{\text{option}} H^3 S Z_{\mathcal{O}}^* \overline{Z}_{\mathcal{O}}^*}{K}}\right)$$

when K of dataset $\mathcal{D}_2$ is sufficiently large, i.e., $K \ge \tilde{O}(C_2^{\text{option}} H^5 S^9 A^2 O^2/(Z_{\mathcal{O}}^* \overline{Z}_{\mathcal{O}}^*))$, where $Z_{\mathcal{O}}^* = Z_{\mathcal{O}}^{\mu^*}$ and $\overline{Z}_{\mathcal{O}}^* = \overline{Z}_{\mathcal{O}}^{\mu^*}$.

While, in general, suboptimality bound (10) does not compare favorably against the suboptimality $\tilde{O}(\sqrt{H^5 SC^*/K})$ attained by the VI-LCB algorithm that learns with primitive actions, we argue that it can be better in long-horizon problems where the horizon H is much greater than the cardinality of the state space S.

5.3 Further Discussion

We analyze the pros and cons of both data-collection procedures, which sheds light on offline RL with options in practice. Compared to $\mathcal{D}_2$, dataset $\mathcal{D}_1$ requires smaller storage and enjoys faster convergence to the optimal value, which is further illustrated as follows.

- **Storage:** For dataset $\mathcal{D}_2 = \{(s_h^k, a_h^k, r_h^k)\}_{h\in[H],k\in[K]}$, the storage is simply HK. However, for dataset $\mathcal{D}_1 = \{(s_{t_i^k}^k, o_{t_i^k}^k, u_{t_i^k}^k)\}_{i\in[j^k],k\in[K]}$, its expected size is $K\cdot Z_{\mathcal{O}}^\rho \le HK$, where ρ is the hierarchical behavior policy. Therefore, in the case of a small $Z_{\mathcal{O}}^\rho$, dataset $\mathcal{D}_1$ requires much smaller storage than $\mathcal{D}_2$ (on average).
- **Suboptimality:** Ignoring the concentrability, the suboptimality bound (10) for dataset $\mathcal{D}_2$ is worse than the suboptimality bound (8) for dataset $\mathcal{D}_1$ by a factor of $\sqrt{S}$, which is introduced when estimating the option transition function and the option utility function using only the information of the primitive actions.

However, since dataset $\mathcal{D}_2$ contains more information on the environment than $\mathcal{D}_1$, it has a weaker requirement on the behavior (hierarchical) policy and allows the evaluation of new options, which is illustrated as follows.

- **Concentrability:** Recall that the suboptimality bounds for both datasets build upon the sufficient coverage assumptions, i.e., Assumption 1 for $\mathcal{D}_1$ and Assumption 2 for $\mathcal{D}_2$. While they are generally incomparable (as dataset $\mathcal{D}_2$ can be collected by an arbitrary behavior policy), we focus on the case that both datasets are collected by the same hierarchical behavior policy ρ. Particularly, it can be shown that Assumption 2 is weaker than Assumption 1. Indeed, if ρ

covers the trajectories of state-option-utility tuples induced by μ^* (i.e., Assumption 1 holds), then it must have covered the trajectories of state-action-reward tuples induced by μ^* (i.e., Assumption 2 holds). However, the opposite does not hold in general and we provide such an example in Appendix H.

- **Evaluation of New Options:** In the options literature, a popular task is offline option discovery (Ajay et al., 2021; Villecroze et al., 2022), i.e., designing new and useful options from the dataset. Therefore, an important problem is whether these new options can be evaluated through the dataset. We argue that dataset $\mathcal{D}_2$ yields greater flexibility than $\mathcal{D}_1$ in this case. Again, we assume that both datasets are collected by the same hierarchical behavior policy ρ. Unfortunately, one cannot use dataset $\mathcal{D}_1$ to evaluate any (h, s, o) that is not visited by ρ, i.e., $\theta_h^\rho(s, o) = 0$, let along evaluating the new options. However, this is not the case for dataset $\mathcal{D}_2$. In fact, any (h, s, o) can be evaluated if the visiting state-action pairs are also reachable by ρ, i.e., $\sum_{h \le m \le H, (s', a') \in \mathcal{S}^m_{h,s,o}} 1/d^\rho_m(s', a') < \infty$. An interesting problem is how to leverage the results in this paper to facilitate offline option discovery, which we shall research in the future work.

6 Conclusions

In this paper, we provide the first analysis of the sample complexity for offline RL with options. A novel information-theoretic lower bound is established, which generalizes the one for offline RL with actions. We derive near-optimal suboptimality bounds of the **PE**ssimistic **V**alue **I**teration for Learning with **O**ptions (PEVIO) algorithm for two popular data-collection procedures. Our results show that options facilitate more sample-efficient learning than primitive actions in offline RL in both the finite-time convergence rate to the optimal value and the actual performance.

Acknowledgments and Disclosure of Funding

The work presented in this paper was partially supported by a research grant from the Research Grants Council, Hong Kong, China (RGC Ref. No. CUHK 14206820). Part of the research reported in this paper was done when the second author was with The Chinese University of Hong Kong. The authors are grateful to the anonymous reviewers for their helpful comments.

References

Ajay, A., Kumar, A., Agrawal, P., Levine, S., and Nachum, O. (2021). OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*.

Bagaria, A., Senthil, J., Slivinski, M., and Konidaris, G. (2021). Robustly learning composable options in deep reinforcement learning. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2161–2169. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Bradtke, S. and Duff, M. (1994). Reinforcement learning methods for continuous-time Markov decision problems. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Brunskill, E. and Li, L. (2014). PAC-inspired option discovery in lifelong reinforcement learning. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 316–324, Bejing, China. PMLR.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference* *on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1042–1051. PMLR.

Co-Reyes, J., Liu, Y., Gupta, A., Eysenbach, B., Abbeel, P., and Levine, S. (2018). Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1009–1018. PMLR.

Şimşek, O. and Barto, A. G. (2004). Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 95, New York, NY, USA. Association for Computing Machinery.

Drescher, G. L. (1991). *Made-up minds: a constructivist approach to artificial intelligence*. MIT press.

Erraqabi, A., Machado., M. C., Zhao, M., Sukhbaatar, S., Lazaric, A., Ludovic, D., and Bengio, Y. (2022). Temporal abstractions-augmented temporally contrastive learning: An alternative to the Laplacian in RL. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 641–651. PMLR.

Eysenbach, B., Salakhutdinov, R. R., and Levine, S. (2019). Search on the replay buffer: Bridging planning and reinforcement learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fikes, R. E., Hart, P. E., and Nilsson, N. J. (1972). Learning and executing generalized robot plans. *Artificial Intelligence*, 3:251–288.

Fruit, R. and Lazaric, A. (2017). Exploration-Exploitation in MDPs with Options. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 576–584. PMLR.

Fruit, R., Pirotta, M., Lazaric, A., and Brunskill, E. (2017). Regret minimization in mdps with options without prior knowledge. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18.

Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. (2018). When waiting is not an option: Learning options with a deliberation cost. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Hiraoka, T., Imagawa, T., Mori, T., Onishi, T., and Tsuruoka, Y. (2019). Learning robust options by conditional value at risk optimization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hoang, C., Sohn, S., Choi, J., Carvalho, W., and Lee, H. (2021). Successor feature landmarks for long-horizon goal-conditioned reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26963–26975. Curran Associates, Inc.

Igl, M., Gambardella, A., He, J., Nardelli, N., Siddharth, N., Boehmer, W., and Whiteson, S. (2020). Multitask soft option learning. In Peters, J. and Sontag, D., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 969–978. PMLR.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2747–2758. Curran Associates, Inc.

Jiang, Y., Gu, S. S., Murphy, K. P., and Finn, C. (2019). Language as an abstraction for hierarchical deep reinforcement learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2020). Learning adversarial Markov decision processes with bandit feedback and unknown transition. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR.

Klissarov, M. and Precup, D. (2021). Flexible option learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4632–4646. Curran Associates, Inc.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Machado, M. C., Barreto, A., and Precup, D. (2021). Temporal abstraction in reinforcement learning with the successor representation.

Machado, M. C., Bellemare, M. G., and Bowling, M. (2017). A Laplacian framework for option discovery in reinforcement learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2295–2304. PMLR.

Mahadevan, S., Marchalleck, N., Das, T. K., and Gosavi, A. (1997). Self-improving factory simulation using continuous-time average-reward reinforcement learning. In *Proc. 14th International Conference on Machine Learning*, pages 202–210.

Mankowitz, D., Mann, T., Bacon, P.-L., Precup, D., and Mannor, S. (2018). Learning robust options. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Nachum, O., Tang, H., Lu, X., Gu, S., Lee, H., and Levine, S. (2019). Why does hierarchy (sometimes) work so well in reinforcement learning?

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and q-learning. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3185–3205. PMLR.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11702–11716. Curran Associates, Inc.

Riemer, M., Liu, M., and Tesauro, G. (2018). Learning abstract options. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Sacerdoti, E. D. (1973). Planning in a hierarchy of abstraction spaces. *Artif. Intell.*, 5:115–135.

Salter, S., Wulfmeier, M., Tirumala, D., Heess, N., Riedmiller, M., Hadsell, R., and Rao, D. (2022). Mo2: Model-based offline options. In Chandar, S., Pascanu, R., and Precup, D., editors, *Proceedings of the 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 902–919. PMLR.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., and Botvinick, M. M. (2014). Optimal behavioral hierarchy. *PLOS Computational Biology*, 10(8):1–10.

Sorg, J. and Singh, S. (2010). Linear options. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 31–38.

Stolle, M. and Precup, D. (2002). Learning options in reinforcement learning. In *Proceedings of the 5th International Symposium on Abstraction, Reformulation and Approximation*, page 212–223, Berlin, Heidelberg. Springer-Verlag.

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

Tessler, C., Givony, S., Zahavy, T., Mankowitz, D., and Mannor, S. (2017). A deep hierarchical approach to lifelong learning in minecraft. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9659–9668. PMLR.

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). FeUdal networks for hierarchical reinforcement learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3540–3549. PMLR.

Villecroze, V., Braviner, H., Naderian, P., Maddison, C., and Loaiza-Ganem, G. (2022). Bayesian nonparametrics for offline skill discovery. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22284–22299. PMLR.

Wang, R., Foster, D., and Kakade, S. M. (2021). What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*.

Wulfmeier, M., Rao, D., Hafner, R., Lampe, T., Abdolmaleki, A., Hertweck, T., Neunert, M., Tirumala, D., Siegel, N., Heess, N., and Riedmiller, M. (2021). Data-efficient hindsight off-policy option learning. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11340–11350. PMLR.

Xie, T. and Jiang, N. (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In Peters, J. and Sontag, D., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 550–559. PMLR.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27395–27407. Curran Associates, Inc.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized lagrangian. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6551–6561. Curran Associates, Inc.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1567–1575. PMLR.

Zhang, F., Jia, C., Li, Y.-C., Yuan, L., Yu, Y., and Zhang, Z. (2023). Discovering generalizable multi-agent coordination skills from multi-task offline data. In *The Eleventh International Conference on Learning Representations*.

A Notations

A.1 Notations in Section 3

Recall that for any $(h, s, o) \in [H] \times \mathcal{S} \times \mathcal{O}$ and $(s', \tau) \in \mathcal{S} \times [H - h + 1]$, $T_h(s'|s, o, \tau)$ is the probability that the agent uses option o at state s at timestep h, reaches state s' at timestep $h + \tau$ without being terminated option o in these τ timesteps, and finally terminates option o at state s' at timestep $h + \tau$. Hence, it can be recursively defined as

$$T_h(s'|s,o,\tau) := \sum_{a\in\mathcal{A}} \pi_h^o(a|s) \sum_{s''\in\mathcal{S}} P_h(s''|s,a) \left(1 - \beta_{h+1}^o(s'')\right) T_{h+1}(s'|s'',o,\tau-1) \tag{11}$$

where $T_h(s'|s,o,1) := \beta_{h+1}^o(s') \sum_{a\in\mathcal{A}} \pi_h^o(a|s) P_h(s'|s,a)$. It can be shown that $T_h(\cdot|s,o,\cdot)$ is a probability distribution over $[H - h + 1] \times \mathcal{S}$, which is stated in the following lemma. (The detailed proof can be found in Appendix G.1.)

Lemma 1. *Consider $T_h(s'|s,o,\tau)$ defined in Equation (11), it holds that*

$$\sum_{\tau\in[H-h+1]} \sum_{s'\in\mathcal{S}} T_h(s'|s,o,\tau) = 1 \tag{12}$$

for any $(h, s) \in [H] \times \mathcal{S}$.

Recall that for any $(s', l) \in \mathcal{S} \times [H - h + 1]$, $\overline{T}_h(s'|s, o, l)$ is the probability that the agent uses option o at state s at timestep h, reaches state s' at timestep $h + l$ without being terminated option o in these l timesteps, but does *not* terminate option o at state s' at timestep $h + l$. Hence, we can recursively define it by

$$\overline{T}_h(s'|s,o,l) := \sum_{a\in\mathcal{A}} \pi_h^o(a|s) \sum_{s''\in\mathcal{S}} P_h(s''|s,a) \left(1 - \beta_{h+1}^o(s'')\right) \overline{T}_{h+1}(s'|s'',o,l-1) \tag{13}$$

where $\overline{T}_h(s'|s,o,1) := (1 - \beta_{h+1}^o(s')) \sum_{a\in\mathcal{A}} \pi_h(a|s) P_h(s'|s,a)$. Intuitively, $T_h(s'|s,o,\tau)$ can be interpreted as the joint probability of two consecutive and independent events: (i) at state s at timestep h, the agent uses option o for $t < \tau$ timesteps without being terminated this option, reaches some state s_{h+t}, and does not terminate option o at state s_{h+t} at timestep $h + t$ (the probability of this event is $\overline{T}_h(s_{h+t}|s,o,t)$), (ii) at state s_{h+t} at timestep $h + t$, the agent keeps using option o for $(\tau - t)$ timesteps, reaches state s', and then terminates option o at state s' at timestep $h + \tau$ (the probability of this event is $T_{h+t}(s'|s_{h+t}, o, \tau - t)$). Since events (i) and (ii) are independent, the joint probability is therefore $\overline{T}_h(s_{h+t}|s,o,t) T_{h+t}(s'|s_{h+t}, o, \tau - t)$. Similarly, $\overline{T}_h(s'|s,o,\tau)$ can be thought of as the joint probability of two consecutive and independent events: (i′) at state s at timestep h, the agent uses option o for $t < \tau$ timesteps without being terminated this option, reaches some state s_{h+t}, and does not terminate option o at state s_{h+t} at timestep $h + t$ (the probability of this event is $\overline{T}_h(s_{h+t}|s,o,t)$), (ii′) at state s_{h+t} at timestep $h + t$, the agent keeps using option o for $(\tau - t)$ timesteps, reaches state s', and does not terminates option o at state s' at timestep $h + \tau$ (the probability of this event is $\overline{T}_{h+t}(s'|s_{h+t}, o, \tau - t)$). Again, since events (i′) and (ii′) are independent, the joint probability is $\overline{T}_h(s_{h+t}|s,o,t)\overline{T}_{h+t}(s'|s_{h+t}, o, \tau - t)$. We formalize this idea in the following lemma. (The detailed proof can be found in Appendix G.2.)

Lemma 2. *Consider $T_h(s'|s,o,\tau)$ and $\overline{T}_h(s'|s,o,l)$ defined in Equation (11) and Equation (13), respectively, it holds that*

$$T_h(s'|s,o,1) + \overline{T}_h(s'|s,o,1) = \sum_{a\in\mathcal{A}} \pi_h^o(a|s) P_h(s'|s,a) \tag{14}$$

$$T_h(s'|s,o,\tau) = \sum_{s''\in\mathcal{S}} \overline{T}_h(s''|s,o,t) T_{h+t}(s'|s'',o,\tau-t) \tag{15}$$

$$\overline{T}_h(s'|s,o,l) = \sum_{s''\in\mathcal{S}} \overline{T}_h(s''|s,o,t) \overline{T}_{h+t}(s'|s'',o,l-t) \tag{16}$$

for any $\tau \geq 2$ and $t \in [\tau - 1]$.

In addition, recall that $U_h(s,o) = \mathbb{E}[u_h(s,o)]$, where $u_h(s,o)$ is the random cumulative reward within timesteps that the agent keeps using option o from timestep h without being terminated it, provided that the agent is at state s at timestep h. Hence, it can be recursively defined as

$$\begin{aligned}
U_h(s,o) &= \sum_{a\in\mathcal{A}} \pi_h^o(a|s) r_h(s,a) + \sum_{s'\in\mathcal{S}} \overline{T}_h(s'|s,o,1) U_{h+1}(s',o) \\
&= \sum_{a\in\mathcal{A}} \pi_h^o(a|s) r_h(s,a) + \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \overline{T}_h(s'|s,o,1) \pi_{h+1}^o(a'|s') r_{h+1}(s',a') \\
&\quad + \sum_{s''\in\mathcal{S}} \overline{T}_h(s''|s,o,2) U_{h+2}(s'',o) \\
&= \cdots = \sum_{a\in\mathcal{A}} \pi_h^o(a|s) r_h(s,a) + \sum_{l\in[H-h]} \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \overline{T}_h(s'|s,o,l) \pi_{h+l}^o(a'|s') r_{h+l}(s',a')
\end{aligned} \tag{17}$$

where $U_H(s,o) := \sum_{a\in\mathcal{A}} \pi_H^o(a|s) r_H(s,a)$. Finally, for any hierarchical policy μ and an episode starts from state s_1 at the first timestep, we define the state-option occupancy measure θ^μ as follows.

$$\theta_h^\mu(s,o) := \mu_h(o|s) \sum_{\tau\in[h-1]} \sum_{s'\in\mathcal{S}} \sum_{o'\in\mathcal{O}} \theta_{h-\tau}^\mu(s',o') T_{h-\tau}(s|s',o',\tau) \tag{18}$$

where $\theta_1(s,o) := \mathbb{I}[s = s_1]\mu_1(o|s)$.

A.2 Notations in Section 5.2

Recall that $d_h^\rho(s,a)$ is the probability that the agent takes action a at state s at timestep h, given that the episode starts from state s_1 at the first timestep. When the behavior policy $\rho = \{\rho_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h\in[H]}$ is not hierarchical, $d_h^\rho(s,a)$ is given as

$$d_h^\rho(s,a) := \rho_h(a|s) \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} P_{h-1}(s|s',a') d_{h-1}^\rho(s',a') \tag{19}$$

where $d_1^\rho(s,a) = \mathbb{I}[s = s_1]\rho_1(a|s)$. When the behavior policy $\rho = \{\rho_h : \mathcal{S} \mapsto \Delta(\mathcal{O})\}_{h\in[H]}$ is hierarchical, we first define the probability $q_h^\mu(s,o)$ that the agent use option o at state s at timestep h following the hierarchical policy μ, given that the episode starts at state s_1 at the first timestep. That is,

$$q_h^\mu(s,o) := \sum_{s'\in\mathcal{S}} \sum_{o'\in\mathcal{O}} q_{h-1}^\mu(s',o') \sum_{a'\in\mathcal{A}} \pi_{h-1}^{o'}(a'|s') P_h(s|s',a') \Big(\mathbb{I}[o' = o](1 - \beta_h^{o'}(s)) + \beta_h^{o'}(s)\mu_h(o|s) \Big)$$

where $q_1^\mu(s,o) = \mathbb{I}[s = s_1]\mu_1(o|s)$. We note that the difference between $q_h^\mu(s,o)$ and $\theta_h^\mu(s,o)$ defined in Equation (18) is that $q_h^\mu(s,o)$ does not require option o to be newly selected at timestep h, i.e., option o_{h-1} needs not to be terminated. Therefore, $d_h^\rho(s,a)$ for a hierarchical policy ρ is given by

$$d_h^\rho(s,a) := \sum_{o\in\mathcal{O}} q_h^\mu(s,o) \mu_h(a|s)$$

B Proof of Theorem 1

Proof. We decompose the Q-function for any $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$ as follows.

$$\begin{aligned}
Q_h^\mu(s,o) &= \mathbb{E}_\mu\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, o_h = o \right] \\
&= \underbrace{\sum_{a\in\mathcal{A}} \pi_h^o(a|s) r_h(s,a)}_{(7.1)} + \underbrace{\sum_{s'\in\mathcal{S}} \left(T_h(s'|s,o,1) V_{h+1}^\mu(s') + \overline{T}_h(s'|s,o,1) Q_{h+1}^\mu(s',o) \right)}_{(7.2)}
\end{aligned} \tag{20}$$

and we define $V_{H+1}^\mu(s) = Q_{H+1}^\mu(s,o) = 0$ for any $(s,o) \in \mathcal{S} \times \mathcal{O}$. That is to say, the Q-function (20) is the expected return from timestep h to timestep H, provided that option o is used at state s at timestep h, no matter option o_{h-1} used at timestep $h-1$ is terminated or not. Similar to classic RL, the Q-function for learning with options can be decomposed into two parts. Term (7.1) is the (expected) instant reward at timestep h and term (7.2) is the expected return in the rest of the episode (from timestep $h+1$ to timestep H). However, different from the Q-function for classic RL with primitive actions, the expected return in the rest of the episode is not the expectation of the value function over the states at timestep $h+1$. Intuitively, after some state s' is reached at timestep $h+1$ with probability $\sum_{a\in\mathcal{A}}\pi_h^o(a|s)P_h(s'|s,a)$ given that option o is used at state s at timestep h, only one of the following two situations happens. The first situation is that option o is not terminated at state s' at timestep $h+1$ (the probability of this event is $\overline{T}_h(s'|s,o,1)$). As a result, the agent uses option o at state s' at timestep $h+1$, and hence the expected return in the rest of the episode conditioned on this first situation is $Q_{h+1}^{\mu}(s',o)$. The second situation is that option o is terminated at state s' at timestep $h+1$ (the probability of this event is $T_h(s'|s,o,1)$, and by Equation (14) we have that $T_h(s'|s,o,1)=\sum_{a\in\mathcal{A}}\pi_h^o(a|s)P_h(s'|s,a)-\overline{T}_h(s'|s,o,1)$). Consequently, the agent selects a new option according to the hierarchical policy at state s' at timestep $h+1$, and hence the expected return in the rest of the episode conditioned on this second situation is $V_{h+1}^{\mu}(s')$. Note that when $\mathcal{O}=\mathcal{A}$ (In this case, we have that $\overline{T}_h(s'|s,o,1)=0$ hold for any $(h,s,o)\in[H]\times\mathcal{S}\times\mathcal{O}$),[6] Equation (20) recovers the Q-function for episodic MDP that learns with primitive actions, e.g. (Jin et al., 2021, Equation (2.5)). Further, by iteratively applying Equation (20) over $h+1,\cdots,H$, we derive that

$$
\begin{aligned}
Q_h^{\mu}(s,o) = &\sum_{a\in\mathcal{A}}\pi_h^o(a|s)\left(r_h(s,a)+\sum_{s'\in\mathcal{S}}\overline{T}_h(s'|s,o,1)\sum_{a'\in\mathcal{A}}\pi_{h+1}^o(a'|s')r_{h+1}(s',a')\right)\\
&+\sum_{s''\in\mathcal{S}}\underbrace{\sum_{s'\in\mathcal{S}}\overline{T}_h(s'|s,o,1)T_{h+1}(s''|s',o,1)}_{=T_h(s''|s,o,2)\text{ by Equation (15)}}V_{h+2}^{\mu}(s'')\\
&+\sum_{s''\in\mathcal{S}}\underbrace{\sum_{s'\in\mathcal{S}}\overline{T}_h(s'|s,o,1)\overline{T}_{h+1}(s''|s',o,1)}_{=\overline{T}_h(s''|s,o,2)\text{ by Equation (16)}}Q_{h+2}^{\mu}(s'',o)\\
&+\sum_{s'\in\mathcal{S}}T_h(s'|s,o,1)V_{h+1}^{\mu}(s')\\
=&\cdots=U_h(s,o)+\sum_{\tau\in[H-h+1]}[T_hV_{h+\tau}^{\mu}](s,o)
\end{aligned}
\tag{21}
$$

where $[T_hy_{h+\tau}](s,o):=\sum_{s'\in\mathcal{S}}T_h(s'|s,o,\tau)y_{h+\tau}(s')$ is defined for any $(s,o,\tau)\in\mathcal{S}\times\mathcal{O}\times[H-h+1]$ given any arbitrary series of functions $\{y_h:\mathcal{S}\mapsto\mathbb{R}\}_{h\in[H]}$. The value function for any $(h,s)\in[H]\times\mathcal{S}$ can be derived as

$$
V_h^{\mu}(s):=\mathbb{E}_{\mu}\left[\sum_{h'=h}^{H}r_{h'}(s_{h'},a_{h'})\middle|s_h=s,o_h\sim\mu_h(\cdot|s_h)\right]=\sum_{o\in\mathcal{O}}\mu_h(o|s)Q_h^{\mu}(s,o)
$$

which concludes the proof. □

C Proof of Theorem 2

Proof. Our construction of the hard episodic MDP instance is inspired by Xie et al. (2021, Appendix D).

Construction of hard instances. The family of MDPs has $(S+2)$ states, $(2H+1)$ timesteps, A actions, and $(O+1)$ options for any $S,H,O\geq 1$ (The rescaling only affects H,S by at most a multiplicative constant and thus does not affect the result). Each MDP $M_{\mathbf{a}^*}(\mathcal{S},\mathcal{A},\mathcal{O},H,\mathcal{P},r)$ is index by a vector $\mathbf{a}^*=(a_{h,i}^*)\in[A]^{HS}$ and is specified as follows.

[6]The notation $\mathcal{O}=\mathcal{A}$ means that for any $a\in\mathcal{A}$, there exists $o\in\mathcal{O}$ such that $\pi_h^o(a|s)=1$ and $\beta_h^o(s)=1$ for any $(h,s,o)\in[H]\times\mathcal{S}\times\mathcal{O}$. In addition, it holds that $O=|\mathcal{O}|=|\mathcal{A}|=A$.

- The state space is $\mathcal{S} := \{s_i\}_{i\in[S]} \cup \{s_g, s_b\}$. There are S "bandit states" $\{s_i\}_{i\in[S]}$, one "good state" s_g, and one "bad state" s_b. Particularly, s_g and s_b are absorbing states, i.e., $P_h(s_g|s_g, a) = P_h(s_b|s_b, a) = 1$ for all $h \in [2H+1]$ and all $a \in [A]$.
- The action space is $\mathcal{A} := [A]$.
- The option set is $\mathcal{O} := \{o_j\}_{j\in[O]} \cup \{o^*\}$.
  - The option o^* satisfies that $\pi_h^{o^*}(a^*_{h,i}|s_i) = 1$ for any $(h, i) \in [H+1] \times [S]$ and $\pi_h^{o^*}(1|s) = 1$ whenever $h \geq H+1$ or $s \in [s_g, s_b]$, i.e., option o^* always takes action $a^*_{h,i}$ at each bandit state s_i and otherwise, takes action 1.
  - For any $o \neq o^*$, it holds that $\pi_h^o(j|s_i) = 1$ for some $j \neq a^*_{h,i}$ and any $(h, i) \in [\lfloor z^* \rfloor] \times [S]$, i.e., any option other than o^* always takes the suboptimal action at each bandit state at the first $\lfloor z^* \rfloor$ timesteps. And at timestep $\lfloor z^* \rfloor + 1 \leq h \leq H+1$, we have that $\pi_h^o(a^*_{h,i}|s_i) = 1$ for any $o \neq o^*$. In addition, it holds that $\pi_h^o(1|s) = 1$ whenever $h \geq H+1$ or $s \in [s_g, s_b]$.
  - Further, for any $o \in \mathcal{O}$, it holds that $\beta_h^o(s_i) = 1$ for any $(h, i) \in [\lfloor z^* \rfloor] \times [S]$ and $\beta_{h'}^o(s_i) = 0$ for any $([2H+1]\backslash[\lfloor z^* \rfloor]) \times [S]$, i.e., any option is guaranteed to terminate (continue) at any bandit state and at each timestep $h \in [\lfloor z^* \rfloor]$ ($h' \in ([2H+1]\backslash[\lfloor z^* \rfloor])$). In addition, each option is continued at state s_g or s_b, i.e., $\beta_h^o(s_g) = \beta_h^o(s_b) = 0$ for any $(h, o) \in [2H+1] \times \mathcal{O}$.
- Transition kernel $\mathcal{P}$: At the first H timesteps, each bandit state s_i can only transit to s_i itself, s_g, or s_b. The transition probabilities satisfy: (i) $P_h(s_i|s_i, a) = 1 - \frac{1}{H}$ for all $a \in [A]$, (ii) $P_h(s_g|s_i, a) = P_h(s_b|s_i, a) = \frac{1}{2H}$ for any $a \neq a^*_{h,i}$, and (iii) $P_h(s_g|s_i, a^*_{h,i}) = \frac{1}{H}(\frac{1}{2} + \tau), P_h(s_b|s_i, a^*_{h,i}) = \frac{1}{H}(\frac{1}{2} - \tau)$, where τ is a parameter to be determined. At the last $H+1$ timesteps, all bandit states transit to one of s_g and s_b with probability $1/2$ each.
- Reward function r: The bandit states and the bad state s_g do not receive any reward, i.e., $r_h(s_i, a) = r_h(s_g, a) = 0$ for any $(h, i, a) \in [2H+1] \times [S] \times [A]$. The good state does not receive any reward at the first $H+1$ timesteps, while for $h \geq H+2$, it receives a reward 1 regardless of the action taken, i.e., $r_h(s_g, a) = 0$ for any $(h, a) \in [H+1] \times [A]$ and $r_h(s_g, a) = 1$ for any $(h, a) \in ([2H+1]\backslash[H+1]) \times [A]$.
- Initial state distribution is uniform on all bandit states, i.e., $S_1 \sim \text{Unif}\{s_i\}_{i\in[S]}$.

We also let $M_{\mathbf{0}}$ denote the "null" MDP that has the same construction as the above except that there is no "special" action $a^*_{h,i}$, that is, for any $a \in [A]$, it holds that

$$P_h(s_g|s_i, a) = P_h(s_b|s_i, a) = \frac{1}{2H}$$

Optimal hierarchical policy μ^* and hierarchical behavior policy ρ. We define our deterministic hierarchical policy $\mu^* = \{\mu_h^*(s) = o^* \text{ for any } s \in \mathcal{S}\}_{h\in[H]}$. Therefore, we have that $\sum_{h,s,o} \theta_h^{\mu^*}(s, o) \leq \lfloor z^* \rfloor \leq z^*$ and $\sum_{h,s,o} \mathbb{I}[\theta_h^{\mu^*}(s, o) > 0] \leq S\lfloor z^* \rfloor \leq \overline{z}^*$. Let $B = \lfloor C^{\text{option}} \rfloor$ denote the largest integer no greater than C^{option}. The hierarchical behavior policy ρ satisfies that $\rho_h(o|s_i) = \frac{1}{B}$ for any $o \in \{o^*\} \cup \{o_j\}_{j\in[B-1]}$ and all $(h, i) \in [H] \times [S]$ and $\rho_h(o^*|s) = 1$ whenever $h \geq H+1$ or $s \in \{s_g, s_b\}$. It should be obvious that

$$\max_{h,s,o} \frac{\theta_h^{\mu^*}(s, o)}{\theta_h^{\rho}(s, o)} \leq B \leq C^{\text{option}}$$

Since the above statement holds for any arbitrary selection of $\text{a}^* \in [B]^{HS}$. Therefore, the following family of problems is indeed a subset of the class $\mathcal{M}(C^{\text{option}}, z^*, \overline{z}^*)$.

$$\{(M_{\mathbf{a}^*}, \rho) : \mathbf{a}^* \in [K]^{HS}\} \subset \mathcal{M}(C^{\text{option}}, z^*, \overline{z}^*)$$

We denote by ν the uniform (prior) distribution on $[B]^{HS}$, i.e., $\nu(\mathbf{a}^* = \mathbf{a}_0) = 1/B^{HS}$ for all $\mathbf{a}_0 \in [B]^{HS}$. Note that the hierarchical behavior policy ρ is the same for all MDPs in the above family of problems. Define

$$l_{h,i}(\hat{\mu}, \mathbf{a}^*) := \mathbb{P}_{\mathbf{a}^*}(\hat{\mu}_h(s_i) \neq o^*), \quad L(\hat{\mu}, \mathbf{a}^*) := \sum_{h=1}^{\lfloor z^* \rfloor} \sum_{i=1}^{S} l_{h,i}(\hat{\mu}, \mathbf{a}^*)$$

The loss L measures the expected number of (h, i) pairs on which the algorithm fails to identify the best option o^*. A large loss will translate to a high suboptimality bound, which is formalized in the following lemma (The detailed proof can be found in Appendix G.3).

Lemma 3. *For any $\mathbf{a}^* \in [B]^{HS}$ and any offline algorithm that outputs a deterministic hierarchical policy $\hat{\mu}$, we have that*

$$\mathbb{E}_{M_{\mathbf{a}^*}}\left[V^*_{1,M_{\mathbf{a}^*}} - V^{\hat{\mu}}_{1,M_{\mathbf{a}^*}}\right] \geq \frac{\tau}{3S} \cdot L(\hat{\mu}, \mathbf{a}^*) \tag{22}$$

Next, we devote to establishing the following inequality, which lower bounds $L(\hat{\mu}, \mathbf{a}^*)$.

$$\mathbb{E}_{\mathbf{a}^* \sim \nu}[L(\hat{\mu}, \mathbf{a}^*)] \geq \lfloor z^* \rfloor S \left(\frac{1}{2} - \sqrt{\frac{2\tau^2 K}{BHS}}\right) \tag{23}$$

Once Inequality (23) is established, then if $\sqrt{2\tau^2 K/BHS} \leq \frac{1}{4}$, we have that

$$\mathbb{E}_{\mathbf{a}^* \sim \nu}[L(\hat{\mu}, \mathbf{a}^*)] \geq \frac{\lfloor z^* \rfloor S}{4}$$

Plugging in Inequality (22), we can further derive that

$$\mathbb{E}_{\mathbf{a}^* \sim \nu}[L(\hat{\mu}, \mathbf{a}^*)] \geq \frac{\tau}{3S} \frac{\lfloor z^* \rfloor S}{4} = \frac{\tau \lfloor z^* \rfloor}{12}$$

Let $\tau = 12\epsilon / \lfloor z^* \rfloor$ where $\epsilon \leq 1/12$. Then, if

$$K \leq \frac{BHS}{32\tau^2} = \frac{BHS(\lfloor z^* \rfloor)^2}{32 \cdot 12^2 \epsilon^2} = c_0 \cdot \frac{C^{\text{option}} H z^* \overline{z}^*}{\epsilon^2}$$

Hence, the suboptimality satisfies that

$$\mathbb{E}_{\mathbf{a}^* \sim \nu}[L(\hat{\mu}, \mathbf{a}^*)] \geq \epsilon$$

and we conclude the proof of Theorem 2. Therefore, the rest of this section is to establish Inequality (23). By slightly modifying the proof of (Xie et al., 2021, Theorem 3), we have that

$$\mathbb{E}_{\mathbf{a}^* \sim \nu}[l_{h,i}(\hat{\mu}, \mathbf{a}^*)] \geq \frac{1}{2} - \sqrt{\mathbb{E}_{\mathbb{P}_{\mathbf{o}}, \mathcal{A}}[N_h(s_i)] \cdot 2\tau^2 / HB}$$

Then, summing the preceding bound over all (h, i), we have that for the hierarchical policy $\hat{\mu}$ output by any offline algorithm,

$$\begin{aligned} &\mathbb{E}_{\mathbf{a}^* \sim \nu}[L(\hat{\mu}, \mathbf{a}^*)] \\ \geq &\lfloor z^* \rfloor S \left(\frac{1}{2} - \sqrt{\frac{1}{\lfloor z^* \rfloor S} \sum_{h=1}^{\lfloor z^* \rfloor} \sum_{i=1}^{S} \mathbb{E}_{\mathbb{P}_{\mathbf{o}}, \mathcal{A}}[N_h(s_i)] \cdot 2\tau^2 / HB}\right) = \lfloor z^* \rfloor S \left(\frac{1}{2} - \sqrt{\frac{2\tau^2 K}{BHS}}\right) \end{aligned}$$

which concludes the proof. $\square$

D Proof of Theorem 3

Proof. Our proof relies on the following lemma, which decomposes the suboptimality (3) into three terms that can be analyzed separately. (The detailed proof can be found in Appendix G.4.)

Lemma 4 (Decomposition of suboptimality)**.** *Let $\widehat{\mu} = \{\widehat{\mu}_h\}_{h\in[H]}$, $\widehat{Q} = \{\widehat{Q}_h\}_{h\in[H]}$, $\widehat{V} = \{\widehat{V}_h\}_{h\in[H]}$ denote the hierarchical policy, the estimated Q-function, and the corresponding estimated value function output from Algorithm 1, respectively. We have that*

$$\begin{aligned}
&\text{SubOpt}_{\mathcal{D}}(\widehat{\mu}, s_1) \\
=& \underbrace{-\sum_{h=1}^{H} \mathbb{E}_{\widehat{\mu}}[\iota_h(s_h, o_h)|s_1]}_{\text{(i)}} + \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\iota_h(s_h, o_h)|s_1]}_{\text{(ii)}} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\langle \widehat{Q}_h(s_h, \cdot), \mu_h^*(\cdot|s_h) - \widehat{\mu}_h(\cdot|s_h)\rangle_{\mathcal{O}}|s_1]}_{\text{(iii)}}
\end{aligned} \tag{24}$$

where μ^* is the optimal hierarchical policy and

$$\iota_h(s, o) = U_h(s, o) + \sum_{\tau\in[H-h+1]} [T_h \widehat{V}_{h+\tau}](s, o) - \widehat{Q}_h(s, o) \tag{25}$$

is defined for any $(h, s, o) \in [H] \times \mathcal{S} \times \mathcal{O}$.

The following lemma states that ι is non-negative. (The detailed proof can be found in Appendix G.6.)

Lemma 5 (Pessimism for Learning with Options in General MDP Using Dataset $\mathcal{D}$)**.** *Suppose that $\{\Gamma_h\}_{h\in[H]}$ in Algorithm 1 are ξ-uncertainty quantifiers. Under $\mathcal{E}$ defined in Equation (5), which satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \xi$, we have*

$$0 \leq \iota_h(s, o) \leq 2\Gamma_h(s, o)$$

for all $(h, s, o) \in [H] \times \mathcal{S} \times \mathcal{O}$.

Therefore, the first term Equation (24) non-positive. By Lemma 4 and the fact that $\widehat{\mu}$ is greedy with respect to $\widehat{Q}_h$ (which implies the third term in Equation (24) is non-positive) for all $h \in [H]$, we have that

$$\text{SubOpt}_{\mathcal{D}}(\widehat{\mu}, s_1) \leq 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\Gamma_h(s_h, o_h)|s_1]$$

which concludes the proof. □

E Proof of Theorem 4

Proof. We first show that Γ defined in line 5 of Subroutine 2 is a ξ-uncertainty quantifier. (The detailed proof can be found in Appendix G.7.)

Lemma 6. *Given a dataset $\mathcal{D}_1$, we have that*

$$\Gamma_h = O\left(\sqrt{\frac{H^2}{n_h(s, o) \vee 1} \log\left(\frac{HSO}{\xi}\right)}\right) \tag{26}$$

for any $(h, s, o) \in [H] \times \mathcal{S} \times \mathcal{O}$ is a ξ-uncertainty quantifier.

Hence, by Theorem 3 and Lemma 6, we have that with probability at least $1 - \xi$

$$\begin{aligned}
\text{SubOpt}_{\mathcal{D}_1}(\widehat{\mu}, s_1) \leq& \tilde{O}\left(\sum_{h=1}^{H} \mathbb{E}_{\mu^*}\left[\sqrt{\frac{H^2}{n_h(s, o) \vee 1}}\middle| s_1\right]\right) \\
\leq& \tilde{O}\left(\sum_{h=1}^{H} \mathbb{E}_{\mu^*}\left[\sqrt{\frac{H^3}{K\theta_h^\rho(s, o)}}\middle| s_1\right]\right)
\end{aligned} \tag{27}$$

$$
\begin{aligned}
&\leq \tilde{O}\left(\sum_{h=1}^{H}\sum_{(s,o)\in\mathcal{S}\times\mathcal{O}}\theta_h^{\mu^*}(s,o)\sqrt{\frac{H^3}{K\theta_h^{\rho}(s,o)}}\right)\\
&\leq \tilde{O}\left(\sum_{h=1}^{H}\sum_{(s,o)\in\mathcal{S}\times\mathcal{O}}\sqrt{\frac{\theta_h^{\mu^*}(s,o)C_1^{\text{option}}H^3}{K}}\right) && (28)\\
&\leq \sqrt{\frac{C_1^{\text{option}}H^3}{K}}\cdot\tilde{O}\left(\sqrt{\sum_{h=1}^{H}\sum_{s,o}\mathbb{I}[\theta_h^{\mu^*}(s,o)>0]}\cdot\sqrt{\sum_{h=1}^{H}\sum_{s,o}\theta_h^{\mu^*}(s,o)}\right) && (29)\\
&= \tilde{O}\left(\sqrt{\frac{C_1^{\text{option}}H^3\overline{Z}_{\mathcal{O}}^*Z_{\mathcal{O}}^*}{K}}\right)
\end{aligned}
$$

where $Z_{\mathcal{O}}^* = \sum_{h\in[H]}\sum_{s\in\mathcal{S}}\sum_{o\in\mathcal{O}}\theta_h^{\mu^*}(s,o)$ and $\overline{Z}_{\mathcal{O}}^* = \sum_{h\in[H]}\sum_{s\in\mathcal{S}}\sum_{o\in\mathcal{O}}\mathbb{I}[\theta_h^{\mu^*}(s,o) > 0]$. Here, Inequality (27) follows from (Xie et al., 2021, Lemma B.1) with $n = K, p = \theta_h^{\rho}(s,o)$. Inequality (28) holds by Assumption 1. Inequality (29) is implied by Cauchy-Schwarz Inequality. Therefore, we conclude the proof. ☐

F Proof of Theorem 5

Proof. To begin with, we first define the set $\mathcal{X}_{h,s,o}^m$, which is the set of state-action pairs that can be reached at timestep m by using option o at state s and timestep h without being terminated. For $m > h$, $\mathcal{X}_{h,s,o}^m$ is given by

$$
\mathcal{X}_{h,s,o}^m := \{(s',a') : \overline{T}_h(s'|s,o,m-h)\cdot\pi_m^o(a'|s') > 0\} \tag{30}
$$

where $\overline{T}_h(s'|s,o,l)$ is defined in Equation (13). Particularly, we define $\mathcal{X}_{h,s,o}^h := \{(s,a') : \pi_h^o(a'|s) > 0\}$.

Remark 3. In the proof, we assume that $\mathcal{X}_{h,s,o}^m$ is known prior for convenience. However, this assumption can be relaxed since one can replace $\mathcal{X}_{h,s,o}^m$ with its superset $\overline{\mathcal{X}}_{h,s,o}^m := \{(s',a') : \pi_m^o(a'|s') > 0\}$, which does not require prior knowledge, and our results follow directly.

First, we show that Γ defined in Line 18 of Subroutine 3 is a ξ-uncertainty quantifier. (The detailed proof can be found in Appendix G.8.)

Lemma 7. *Given a dataset $\mathcal{D}_2$, we have that*

$$
\Gamma_h(s,o) = \tilde{O}\left(H\sqrt{\sum_{m=h}^{H}\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}\frac{HS}{N_m(s,a)\vee 1}} + H\phi_h(s,o)\right) \tag{31}
$$

where

$$
\begin{aligned}
\phi_h(s,o) :=& HS^2\sum_{h\leq m<t\leq H}\sqrt{\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\frac{1}{N_m(s_m,a_m)\vee 1}}\cdot\sqrt{\sum_{(s_t',a_t')\in\mathcal{X}_{h,s,o}^t}\frac{1}{N_t(s_t',a_t')\vee 1}}\\
&+ HS^3\sum_{h\leq m<t\leq H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m,(s_t',a_t')\in\mathcal{X}_{h,s,o}^t}\left(\frac{1}{N_m(s_m,a_m)\vee 1}+\frac{1}{N_t(s_t',a_t')\vee 1}\right)\\
&+ \sum_{m=h}^{H}\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}\frac{S}{N_m(s,a)\vee 1}
\end{aligned}
\tag{32}
$$

is a ξ-uncertainty quantifier, where $\{\widehat{V}_{h+1}\}_{h\in[H]}$ are obtained in line 8 of Algorithm 1.

By Theorem 3 and Lemma 7, we have that with probability at least $1-\xi$

$$
\begin{aligned}
&\text{SubOpt}_{\mathcal{D}_2}(\widehat{\mu}, s_1) \\
\leq& \underbrace{\tilde{O}\left(\sum_{h=1}^{H} \mathbb{E}_{\mu^*}\left[H\sqrt{\sum_{h\leq m\leq H}\sum_{s,a}\frac{HS}{N_m(s,a)\vee 1}}\middle| s_1\right]\right)}_{\text{(F. i)}} + \underbrace{\tilde{O}\left(\sum_{h=1}^{H}\mathbb{E}_{\mu^*}\left[H\phi_h(s,o)|s_1\right]\right)}_{\text{(F. ii)}}
\end{aligned}
$$

Recall that $\mathcal{X}^m_{h,s,o}$ is the set of state-action pairs that can be reached at timestep $m \geq h$ by using option o at state s and timestep h without being terminated. Particularly, we have that $\mathcal{X}^h_{h,s,o} = \{(s', a')|s' = s, \pi^o_h(a'|s') > 0\}$. For term (F. i), we have that

$$
\begin{aligned}
&\sqrt{H^3S}\cdot\tilde{O}\left(\sum_{h=1}^{H}\mathbb{E}_{\mu^*}\left[\sqrt{\sum_{h\leq m\leq H}\sum_{(s,a)\in\mathcal{X}^m_{h,s,o}}\frac{1}{N_m(s,a)\vee 1}}\middle| s_1\right]\right) \\
\leq&\sqrt{H^3S}\cdot\tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\sqrt{\frac{\theta^{\mu^*}_h(s,o)}{K}}\cdot\sqrt{\sum_{h\leq m\leq H}\sum_{(s,a)\in\mathcal{X}^m_{h,s,o}}\frac{\theta^{\mu^*}_h(s,o)}{d^\rho_m(s',a')}}\right) &(33)\\
\leq&\sqrt{C^{\text{option}}_2 H^3 S}\cdot\tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\sqrt{\frac{\theta^{\mu^*}_h(s,o)}{K}}\right) &(34)\\
\leq&\sqrt{C^{\text{option}}_2 H^3 S}\cdot\tilde{O}\left(\sqrt{\sum_{h,s,o}\mathbb{I}[\theta^{\mu^*}_h(s,o)>0]}\cdot\sqrt{\frac{\sum_{h=1}^H\sum_{s,o}\theta^{\mu^*}_h(s,o)}{K}}\right)\\
\leq&\tilde{O}\left(\sqrt{\frac{C^{\text{option}}_2 H^3 S Z^*_{\mathcal{O}}\overline{Z}^*_{\mathcal{O}}}{K}}\right) &(35)
\end{aligned}
$$

where Inequality (33) holds by the similar analysis in deriving Inequality (27), Inequality (34) follows from the definition of C^{option}_2 in Equation (9), and the last inequality holds by the definition of $\overline{Z}^*_{\mathcal{O}}$ and $Z^*_{\mathcal{O}}$. We note that since we do not split dataset $\mathcal{D}_2$, a factor of $\sqrt{H}$ is saved in deriving Inequality (33). For term (F. ii) and by the definition of ϕ in Equation (32), we have that

$$
\begin{aligned}
&\tilde{O}\left(\sum_{h=1}^{H}\mathbb{E}_{\mu^*}\left[H\phi_h(s,o)|s_1\right]\right)\\
\leq& HS\cdot\tilde{O}\left(\mathbb{E}_{\mu^*}\left[\sum_{h\leq m\leq H}\sum_{(s,a)\in\mathcal{X}^m_{h,s,o}}\frac{1}{N_m(s,a)\vee 1}\middle| s_1\right]\right)\\
&+H^2S^3\cdot\tilde{O}\left(\sum_{h=1}^{H}\mathbb{E}_{\mu^*}\left[\sum_{h\leq m<t\leq H}\sum_{(s_m,a_m)\in\mathcal{X}^m_{h,s,o},(s'_t,a'_t)\in\mathcal{X}^t_{h,s,o}}\frac{1}{N_m(s_m,a_m)\vee 1}\middle| s_1\right]\right)\\
&+H^2S^3\cdot\tilde{O}\left(\sum_{h=1}^{H}\mathbb{E}_{\mu^*}\left[\sum_{h\leq m<t\leq H}\sum_{(s_m,a_m)\in\mathcal{X}^m_{h,s,o},(s'_t,a'_t)\in\mathcal{X}^t_{h,s,o}}\frac{1}{N_t(s'_t,a'_t)\vee 1}\middle| s_1\right]\right)\\
&+H^2S^2\cdot\tilde{O}\left(\mathbb{E}_{\mu^*}\left[\sum_{h\leq m<t\leq H}\sqrt{\sum_{(s_m,a_m)\in\mathcal{X}^m_{h,s,o}}\frac{1}{N_m(s_m,a_m)\vee 1}}\cdot\sqrt{\sum_{(s'_t,a'_t)\in\mathcal{X}^t_{h,s,o}}\frac{1}{N_t(s'_t,a'_t)\vee 1}}\middle| s_1\right]\right)\\
\leq& H^2S^3\cdot\tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\theta^{\mu^*}_h(s,o)\sum_{h\leq m<t\leq H}\sum_{(s_m,a_m)\in\mathcal{X}^m_{h,s,o},(s'_t,a'_t)\in\mathcal{X}^t_{h,s,o}}\frac{1}{Kd^\rho_m(s_m,a_m)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ H^2S^3 \cdot \tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\theta_h^{\mu^*}(s,o) \sum_{h\le m<t\le H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m,(s_t',a_t')\in\mathcal{X}_{h,s,o}^t} \frac{1}{Kd_t^\rho(s_t',a_t')}\right)\\
&+ HS \cdot \tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\theta_h^{\mu^*}(s,o) \sum_{h\le m\le H}\sum_{(s',a')\in\mathcal{X}_{h,s,o}^m} \frac{1}{Kd_m^\rho(s',a')}\right)\\
&+ H^2S^2 \cdot \tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\theta_h^{\mu^*}(s,o) \sum_{h\le m<t\le H}\sqrt{\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m} \frac{1}{Kd_m^\rho(s_m,a_m)}}\cdot\sqrt{\sum_{(s_t',a_t')\in\mathcal{X}_{h,s,o}^t} \frac{1}{Kd_t^\rho(s_t',a_t')}}\right)\\
\le& C_2^{\text{option}}H^2S^3\cdot\tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\frac{1}{K}\sum_{h<t\le H,(s_t',a_t')\in\mathcal{X}_{h,s,o}^t} 1\right) + C_2^{\text{option}}H^2S^3\cdot\tilde{O}\left(\sum_{h=1}^{H}\sum_{s,o}\frac{1}{K}\sum_{h\le m\le H,(s_m,a_m)\in\mathcal{X}_{h,s,o}^m} 1\right)\\
&+\tilde{O}\left(\frac{C_2^{\text{option}}H^2S^2O}{K}\right)+\tilde{O}\left(\frac{C_2^{\text{option}}H^4S^3O}{K}\right)\\
\le&\tilde{O}\left(\frac{C_2^{\text{option}}H^4S^5AO}{K}\right)
\end{aligned}
\tag{36}
$$

where the last-second inequality holds by

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{s,o}\theta_h^{\mu^*}(s,o) \sum_{h\le m<t\le H}\sqrt{\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m} \frac{1}{Kd_m^\rho(s_m,a_m)}}\cdot\sqrt{\sum_{(s_t',a_t')\in\mathcal{X}_{h,s,o}^t} \frac{1}{Kd_t^\rho(s_t',a_t')}}\\
\le&\sum_{h=1}^{H}\sum_{s,o}\frac{1}{K}\sqrt{H\sum_{h\le m\le H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\frac{\theta_h^{\mu^*}(s,o)}{d_m^\rho(s_m,a_m)}}\sqrt{H\sum_{h<t\le H}\sum_{(s_t',a_t')\in\mathcal{X}_{h,s,o}^t}\frac{\theta_h^{\mu^*}(s,o)}{d_t^\rho(s_t',a_t')}}\\
\le&\sum_{h=1}^{H}\sum_{s,o}\frac{HC_2^{\text{option}}}{K}=\frac{H^2SOC_2^{\text{option}}}{K}
\end{aligned}
$$

Combining Inequalities (35) and (36), when $K \ge \tilde{O}(C_2^{\text{option}}H^5S^9A^2O^2/(Z_{\mathcal{O}}^*\overline{Z}_{\mathcal{O}}^*))$, we conclude the proof. ☐

G Proofs of Auxiliary Lemmas

G.1 Proof of Lemma 1

Proof. We prove this lemma by backward induction. Recall that $\beta_{H+1}^o(s)=1$ for any $(s,o)\in\mathcal{S}\times\mathcal{O}$. At timestep H, for any $s\in\mathcal{S}$, we have that

$$
\begin{aligned}
\sum_{\tau\in[H-H+1]}\sum_{s'\in\mathcal{S}}T_H(s'|s,o,\tau)&=\sum_{s'\in\mathcal{S}}T_H(s'|s,o,1)\\
&=\sum_{s'\in\mathcal{S}}\beta_{H+1}^o(s')\sum_{a\in\mathcal{A}}\pi_H^o(a|s)P_H(s'|s,a)\\
&=\sum_{a\in\mathcal{A}}\pi_H^o(a|s)\sum_{s'\in\mathcal{S}}P_H(s'|s,a)=1
\end{aligned}
$$

Next, assume that Equation (12) holds for timestep $h+1$. At timestep h, for any $s \in \mathcal{S}$, it holds that

$$
\begin{aligned}
&\sum_{\tau \in [H-h+1]} \sum_{s' \in \mathcal{S}} T_h(s'|s,o,\tau) \\
&= \sum_{s' \in \mathcal{S}} T_h(s'|s,o,1) + \sum_{\tau=2}^{H-h+1} \sum_{s' \in \mathcal{S}} T_h(s'|s,o,\tau) \\
&= \sum_{s' \in \mathcal{S}} \beta^o_{h+1}(s') \sum_{a \in \mathcal{A}} \pi^o_h(a|s) P_h(s'|s,a) \\
&\quad + \sum_{a \in \mathcal{A}} \pi^o_h(a|s) \sum_{s'' \in \mathcal{S}} P_h(s''|s,a) \left(1 - \beta^o_{h+1}(s'')\right) \underbrace{\sum_{\tau=2}^{H-h+1} \sum_{s' \in \mathcal{S}} T_{h+1}(s'|s'',o,\tau-1)}_{=1} \\
&= \sum_{s' \in \mathcal{S}} \beta^o_{h+1}(s') \sum_{a \in \mathcal{A}} \pi^o_h(a|s) P_h(s'|s,a) + \sum_{a \in \mathcal{A}} \pi^o_h(a|s) \sum_{s'' \in \mathcal{S}} P_h(s''|s,a) \left(1 - \beta^o_{h+1}(s'')\right) \\
&= \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi^o_h(a|s) P_h(s'|s,a) = 1
\end{aligned}
$$

which concludes the proof. $\square$

G.2 Proof of Lemma 2

Proof. Equation (14) can be easily derived by the definitions of $T_h(s'|s,o,1)$ and $\overline{T}_h(s'|s,o,1)$. We first prove Equation (16) by backward induction on t. When $t = \tau - 1$, by the definition of $\overline{T}_h(s'|s,o,\tau)$ in Equation (13), we have that for any $\tau \geq 2$,

$$
\begin{aligned}
\overline{T}_h(s'|s,o,\tau) &= \sum_{s'' \in \mathcal{S}} \underbrace{\sum_{a \in \mathcal{A}} \pi^o_h(a|s) P_h(s''|s,a) \left(1 - \beta^o_{h+1}(s'')\right)}_{=\overline{T}_h(s''|s,o,1)} \overline{T}_{h+1}(s'|s'',o,\tau-1) \\
&= \sum_{s'' \in \mathcal{S}} \overline{T}_h(s''|s,o,1) \overline{T}_{h+1}(s'|s'',o,\tau-1)
\end{aligned}
$$

Assume that Equation (16) holds for any $\tau \geq 2$ and $2 \leq t+1 \leq \tau - 1$, we have that

$$
\begin{aligned}
\overline{T}_h(s'|s,o,\tau) &= \sum_{s'' \in \mathcal{S}} \overline{T}_h(s''|s,o,t+1) \overline{T}_{h+t+1}(s'|s'',o,\tau-t-1) \\
&= \sum_{s''' \in \mathcal{S}} \overline{T}_h(s'''|s,o,t) \underbrace{\sum_{s'' \in \mathcal{S}} \overline{T}_{h+t}(s''|s''',o,1) \overline{T}_{h+t+1}(s'|s'',o,\tau-t-1)}_{=\overline{T}_{h+t}(s'|s''',o,\tau-t)} \\
&= \sum_{s'' \in \mathcal{S}} \overline{T}_h(s''|s,o,t) \overline{T}_{h+t}(s'|s'',o,\tau-t)
\end{aligned}
$$

which concludes the proof of Equation (16). Next, we prove Equation (15) by backward induction on t. When $t = \tau - 1$, by the definition of $T_h(s'|s,o,\tau)$ in Equation (11), we have that for any $\tau \geq 2$

$$
\begin{aligned}
T_h(s'|s,o,\tau) &= \sum_{s'' \in \mathcal{S}} \underbrace{\sum_{a \in \mathcal{A}} \pi^o_h(a|s) P_h(s''|s,a) \left(1 - \beta^o_{h+1}(s'')\right)}_{=\overline{T}_h(s''|s,o,1)} T_{h+1}(s'|s'',o,\tau-1) \\
&= \sum_{s'' \in \mathcal{S}} \overline{T}_h(s''|s,o,1) T_{h+1}(s'|s'',o,\tau-1)
\end{aligned}
$$

Assume that Equation (15) holds for any $\tau \geq 2$ and $t+1 \in [\tau-1]$, we have that

$$
\begin{aligned}
T_h(s'|s,o,\tau) &= \sum_{s''\in\mathcal{S}} \overline{T}_h(s''|s,o,t+1)T_{h+t+1}(s'|s'',o,\tau-t-1) \\
&= \sum_{s'''\in\mathcal{S}} \overline{T}_h(s'''|s,o,t) \underbrace{\sum_{s''\in\mathcal{S}} \overline{T}_{h+t}(s''|s''',o,1)T_{h+t+1}(s'|s'',o,\tau-t-1)}_{=T_{h+t}(s'|s''',o,\tau-t)} \qquad (37) \\
&= \sum_{s''\in\mathcal{S}} \overline{T}_h(s''|s,o,t)T_{h+t}(s'|s'',o,\tau-t)
\end{aligned}
$$

where Equation (37) holds by Equation (16) and we concludes the proof. □

G.3 Proof of Lemma 3

Proof. Fix any $\mathbf{a}^* \in [B]^{HS}$, by construction of our MDP $M_{\mathbf{a}^*}$, only the good state s_g receives a reward of 1 starting at timestep $h \in \{H+2, \cdots, 2H+1\}$. Along each trajectory, there will be exactly one transition from the bandit states $\{s_i\}$ to either the good state s_g or the bad state s_b. This transition can happen at timestep $h = H+1$ but with the same transition probability regardless of the policy. In addition, the state occupancy measure $\theta_h^\mu(s_i) := \sum_o \theta_h^\mu(s_i,o) = \frac{1}{S}(1-\frac{1}{H})^{h-1} =: \theta_h(s_i)$ (for $h \leq \lfloor z^* \rfloor$) does not depend on the hierarchical policy μ. Note that any option is terminated at any bandit state at the first $\lfloor z^* \rfloor$ timesteps. Therefore, we have that

$$
\begin{aligned}
&V^*_{1,M_{\mathbf{a}^*}} - V^{\hat{\mu}}_{1,M_{\mathbf{a}^*}} \\
=&\sum_{h=1}^{\lfloor z^*\rfloor}\sum_{i=1}^{S} \theta_h(s_i)\cdot\left[\frac{1}{H}\left(\frac{1}{2}+\tau\right)-\frac{1}{2H}\right]\cdot \mathbb{I}[\hat{\mu}_h(s_i)\neq o^*]\cdot H \\
=&\sum_{h=1}^{\lfloor z^*\rfloor}\sum_{i=1}^{S} \frac{1}{S}\left(1-\frac{1}{H}\right)^{h-1}\tau\cdot \mathbb{I}[\hat{\mu}_h(s_i)\neq o^*]
\end{aligned}
$$

Taking the expectation with respect to the execution of the behavior policy ρ within the MDP $M_{\mathbf{a}^*}$, we have that

$$
\begin{aligned}
\mathbb{E}_{M_{\mathbf{a}^*}}\left[V^*_{1,M_{\mathbf{a}^*}} - V^{\hat{\mu}}_{1,M_{\mathbf{a}^*}}\right] &= \sum_{h=1}^{\lfloor z^*\rfloor}\sum_{i=1}^{S} \frac{1}{S}\left(1-\frac{1}{H}\right)^{h-1}\tau\cdot \underbrace{\mathbb{P}_{\mathbf{a}^*}(\hat{\mu}_h(s_i)\neq o^*)}_{=l_{h,i}(\hat{\mu},\mathbf{a}^*)} \\
\geq &\frac{\tau}{3S}\cdot\sum_{h=1}^{\lfloor z^*\rfloor}\sum_{i=1}^{S} l_{h,i}(\hat{\mu},\mathbf{a}^*) = \frac{\tau}{3S}\cdot L(\hat{\mu},\mathbf{a}^*)
\end{aligned}
$$

which concludes the proof. □

G.4 Proof of Lemma 4

Proof. We first provide the extended value difference lemma for options, which generalizes the extended value difference lemma (Cai et al., 2020, Lemma 4.2). (The detailed proof can be found in Appendix G.5.)

Lemma 8 (Extended value difference for learning with options)**.** *Let $\mu = \{\mu_h\}_{h\in[H]}$ and $\mu' = \{\mu'_h\}_{h\in[H]}$ be any two hierarchical policies and let $\{\widehat{Q}_h : \mathcal{S}\times\mathcal{O}\mapsto\mathbb{R}^+\}_{h\in[H]}$ be any estimated Q-function. For all $h\in[H]$, we define the estimated value function $\widehat{V}_h : \mathcal{S}\mapsto\mathbb{R}$ by setting $\widehat{V}_h(s) = \langle\widehat{Q}_h(s,\cdot),\mu_h(\cdot|s)\rangle_{\mathcal{O}}$ for all $s\in\mathcal{S}$. For all $s\in\mathcal{S}$, we have*

$$
\begin{aligned}
&\widehat{V}_1(s) - V_1^{\mu'}(s) \\
=&\sum_{h=1}^{H}\mathbb{E}_{\mu'}[\langle\widehat{Q}_h(s_h,\cdot),\mu_h(\cdot|s_h)-\mu'_h(\cdot|s_h)\rangle_{\mathcal{O}}|s_1] - \sum_{h=1}^{H}\mathbb{E}_{\mu'}[\iota_h(s_h,o_h)|s_1]
\end{aligned} \qquad (38)
$$

where $\mathbb{E}_{\mu'}[f(s_h)|s_1] := \sum_{s'\in\mathcal{S}}\theta_h^{\mu'}(s')f(s')$ and $\mathbb{E}_{\mu'}[g(s_h,o_h)] := \sum_{(s',o')}\theta_h^{\mu'}(s',o')g(s',o')$ for any $h\in[H]$ and arbitrary function $f:\mathcal{S}\mapsto\mathbb{R}$ and $g:\mathcal{S}\times\mathcal{O}\mapsto\mathbb{R}$.

We decompose the suboptimality (3) into two terms.

$$\text{SubOpt}_{\mathcal{D}}(\widehat{\mu}, s_1) = \underbrace{\Big(V_1^*(s) - \widehat{V}_1(s)\Big)}_{\text{(i)}} + \underbrace{\Big(\widehat{V}_1(s) - V_1^{\widehat{\mu}}(s)\Big)}_{\text{(ii)}} \tag{39}$$

Term (i). Applying Lemma 8 with $\mu = \widehat{\mu}, \mu' = \mu^*$ and $\{\widehat{Q}_h\}_{h\in[H]}$ being the estimated Q-functions constructed by the meta-algorithm (and taking the inverse in both sides), we have that

$$V_1^*(s) - \widehat{V}_1(s) = \sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\langle \widehat{Q}_h(s_h, \cdot), \mu_h^*(\cdot|s_h) - \widehat{\mu}_h(\cdot|s_h)\rangle_{\mathcal{O}}|s_1] + \sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\iota_h(s_h, o_h)|s_1] \tag{40}$$

Term (ii). Similarly, applying Lemma 8 with $\mu = \mu' = \widehat{\mu}$ and $\{\widehat{Q}_h\}_{h\in[H]}$ being the estimated Q-functions constructed by the meta-algorithm, we have that

$$\widehat{V}_1(s) - V_1^{\widehat{\mu}}(s) = -\sum_{h=1}^{H} \mathbb{E}_{\widehat{\mu}}[\iota_h(s_h, o_h)|s_1] \tag{41}$$

Combining Equation (40) and (41), we decompose the suboptimality (39) as follows

$$\begin{aligned}\text{SubOpt}_{\mathcal{D}}(\widehat{\mu}, s_1) = &- \sum_{h=1}^{H} \mathbb{E}_{\widehat{\mu}}[\iota_h(s_h, o_h)|s_1] + \sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\iota_h(s_h, o_h)|s_1] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\mu^*}[\langle \widehat{Q}_h(s_h, \cdot), \mu_h^*(\cdot|s_h) - \widehat{\mu}_h(\cdot|s_h)\rangle_{\mathcal{O}}|s_1]\end{aligned}$$

which concludes the proof. □

G.5 Proof of Lemma 8

Proof. By the definition of $\widehat{V}_h(s) = \langle \widehat{Q}_h(s, \cdot), \mu_h(\cdot|s)\rangle_{\mathcal{O}}$ and $V_h^{\mu'} = \langle Q_h^{\mu'}(s, \cdot), \mu_h'(\cdot|s)\rangle_{\mathcal{O}}$, we have that

$$\begin{aligned}&\widehat{V}_h(s) - V_h^{\mu'}(s) = \langle \widehat{Q}_h(s, \cdot), \mu_h(\cdot|s)\rangle_{\mathcal{O}} - \langle Q_h^{\mu'}(s, \cdot), \mu_h'(\cdot|s)\rangle_{\mathcal{O}} \\ =&\langle \widehat{Q}_h(s, \cdot), \mu_h(\cdot|s) - \mu_h'(\cdot|s)\rangle_{\mathcal{O}} + \langle \widehat{Q}_h(s, \cdot) - Q_h^{\mu'}(s, \cdot), \mu_h'(\cdot|s)\rangle_{\mathcal{O}}\end{aligned}$$

By the definition of the model prediction error in Equation (25), we have that

$$Q_h^{\mu'} = U_h + \sum_{\tau=1}^{H-h+1} [T_h V_{h+\tau}^{\mu'}], \quad \widehat{Q}_h = U_h + \sum_{\tau=1}^{H-h+1} [T_h \widehat{V}_{h+\tau}] - \iota_h$$

which implies that

$$\widehat{Q}_h - Q_h^{\mu'} = \sum_{\tau=1}^{H-h+1} [T_h(\widehat{V}_{h+\tau} - V_{h+\tau}^{\mu'})] - \iota_h$$

That is, we have that

$$\begin{aligned}\widehat{V}_h(s) - V_h^{\mu'}(s) = &\sum_{\tau=1}^{H-h+1} \langle [T_h(\widehat{V}_{h+\tau} - V_{h+\tau}^{\mu'})](s, \cdot), \mu_h'(\cdot|s)\rangle_{\mathcal{O}} \\ &- \langle \iota_h(s, \cdot), \mu_h'(\cdot|s)\rangle_{\mathcal{O}} + \langle \widehat{Q}_h(s, \cdot), \mu_h(\cdot|s) - \mu_h'(\cdot|s)\rangle_{\mathcal{O}}\end{aligned} \tag{42}$$

Recall that for any hierarchical policy μ and episode that starts from state s_1 at the first timestep, the state-option occupancy measure θ^μ is given by

$$\theta_h^\mu(s, o) = \mu_h(o|s) \sum_{\tau\in[h-1]} \sum_{s'\in\mathcal{S}} \sum_{o'\in\mathcal{O}} \theta_{h-\tau}^\mu(s', o') T_{h-\tau}(s|s', o', \tau)$$

where $\theta_1(s, o) = \mathbb{I}[s = s_1]\mu_1(o|s)$. Let $q_h^\mu(s'|s, \tau) := \sum_{o\in\mathcal{O}} \mu_h(o|s) T_h(s'|s, o, \tau)$ is the probability that the agent selects a new option o according to the hierarchical policy μ at state s at timestep h

(either when $h = 1$ or when the option used at timestep $(h-1)$ is terminated if $h \geq 2$), uses option o for τ timesteps, and terminates option o at state s' at timestep $h+\tau$. We have that

$$\theta_h^{\mu}(s) = \sum_{o \in \mathcal{O}} \theta_h^{\mu}(s,o) = \sum_{\{(t_i, s_{t_i})\}_{i=1}^{j} \in \mathcal{L}_{h,s}} \prod_{i=1}^{j-1} q_{t_i}^{\mu}(s_{t_{i+1}} | s_{t_i}, t_{i+1} - t_i)$$

Here, $\mathcal{L}_{h,s}$ contains any possible sequence of timestep-state pair $\{(t_1, s_{t_1}), (t_2, s_{t_2}), ..., (t_j, s_{t_j})\}$ with a (random) length of $j \leq h$, where $1 = t_1 < t_2 < ... < t_j \leq H$. For such a sequence, the agent selects a new option at state s_{t_i} at timestep t_i (either when $t_i = 1$ or when the option used at timestep $(t_i - 1)$ is terminated if $t_i \geq 2$), uses it for $(t_{i+1} - t_i)$ timesteps, and terminates this option at timestep t_{i+1}. It also holds that $(t_1, s_{t_1}) = (1, s_1)$ and $(t_j, s_{t_j}) = (h, s)$. Recursively expanding Equation (42), we obtain that

$$\begin{aligned}
&\widehat{V}_1(s) - V_1^{\mu'}(s) \\
=& \sum_{\tau=1}^{H} \langle [T_1(\widehat{V}_{1+\tau} - V_{1+\tau}^{\mu'})](s,\cdot), \mu_1'(\cdot|s)\rangle_{\mathcal{O}} - \langle \iota_1(s,\cdot), \mu_1'(\cdot|s)\rangle_{\mathcal{O}} + \langle \widehat{Q}_1(s,\cdot), \mu_1(\cdot|s) - \mu_1'(\cdot|s)\rangle_{\mathcal{O}} \\
=& \sum_{o\in\mathcal{O}} \mu_1'(o|s) \sum_{\tau=1}^{H} \sum_{s'\in\mathcal{S}} T_1(s'|s,o,\tau)\left(\widehat{V}_{1+\tau}(s') - V_{1+\tau}^{\mu'}(s')\right) - \langle \iota_1(s,\cdot), \mu_1'(\cdot|s)\rangle_{\mathcal{O}} + \langle \widehat{Q}_1(s,\cdot), \mu_1(\cdot|s) - \mu_1'(\cdot|s)\rangle_{\mathcal{O}} \\
=& \sum_{\tau=1}^{H} \sum_{s'\in\mathcal{S}} \underbrace{\left(\sum_{o\in\mathcal{O}} \mu_1'(o|s) T_1(s'|s,o,\tau)\right)}_{:= q_1^{\mu'}(s'|s,\tau)} \left(\sum_{\tau'=1}^{H-\tau} \langle [T_{1+\tau}(\widehat{V}_{1+\tau+\tau'} - V_{1+\tau+\tau'}^{\mu'})](s,\cdot), \mu_{1+\tau}'(\cdot|s')\rangle_{\mathcal{O}}\right. \\
&\qquad \left. - \langle \iota_{1+\tau}(s',\cdot), \mu_{1+\tau}'(\cdot|s')\rangle_{\mathcal{O}} + \langle \widehat{Q}_{1+\tau}(s',\cdot), \mu_{1+\tau}(\cdot|s') - \mu_{1+\tau}'(\cdot|s')\rangle_{\mathcal{O}}\right) \\
&- \langle \iota_1(s,\cdot), \mu_1'(\cdot|s)\rangle_{\mathcal{O}} + \langle \widehat{Q}_1(s,\cdot), \mu_1(\cdot|s) - \mu_1'(\cdot|s)\rangle_{\mathcal{O}} \\
=& \sum_{\tau=1}^{H} \sum_{s'\in\mathcal{S}} q_1^{\mu'}(s'|s,\tau) \sum_{\tau'=1}^{H-\tau} \langle [T_{1+\tau}(\widehat{V}_{1+\tau+\tau'} - V_{1+\tau+\tau'}^{\mu'})](s,\cdot), \mu_{1+\tau}'(\cdot|s')\rangle_{\mathcal{O}} \\
&+ \left(\langle \widehat{Q}_1(s,\cdot), \mu_1(\cdot|s) - \mu_1'(\cdot|s)\rangle_{\mathcal{O}} + \sum_{\tau=1}^{H} \sum_{s'\in\mathcal{S}} q_1^{\mu'}(s'|s,\tau) \langle \widehat{Q}_{1+\tau}(s',\cdot), \mu_{1+\tau}(\cdot|s') - \mu_{1+\tau}'(\cdot|s')\rangle_{\mathcal{O}}\right) \\
&- \left(\langle \iota_1(s,\cdot), \mu_1'(\cdot|s)\rangle_{\mathcal{O}} + \sum_{\tau=1}^{H} \sum_{s'\in\mathcal{S}} q_1^{\mu'}(s'|s,\tau) \langle \iota_{1+\tau}(s',\cdot), \mu_{1+\tau}'(\cdot|s')\rangle_{\mathcal{O}}\right) \\
=& \cdots = \sum_{h=1}^{H} \sum_{s\in\mathcal{S}} \theta_h^{\mu'}(s) \langle \widehat{Q}_h(s,\cdot), \mu_h(\cdot|s) - \mu_h'(\cdot|s)\rangle_{\mathcal{O}} - \sum_{h=1}^{H} \sum_{s\in\mathcal{S}} \theta_h^{\mu'}(s) \langle \iota_h(s,\cdot), \mu_h'(\cdot|s)\rangle_{\mathcal{O}} \\
=& \sum_{h=1}^{H} \mathbb{E}_{\mu'}[\langle \widehat{Q}_h(s_h,\cdot), \mu_h(\cdot|s_h) - \mu_h'(\cdot|s_h)\rangle_{\mathcal{O}} | s_1] - \sum_{h=1}^{H} \mathbb{E}_{\mu'}[\iota_h(s_h, o_h)|s_1]
\end{aligned}$$

which concludes the proof. □

G.6 Proof of Lemma 5

Proof. We first show that conditioned on the successful event $\mathcal{E}$ defined in Equation (5), the model evaluation error $\{\iota_h\}_{h\in[H]}$ is non-negative. Indeed, by the definition of ι_h in Equation (25), if $\overline{Q}_h(s,o) < 0$ (which implies $\widehat{Q}_h(s,o) = 0$ by line 6 of Algorithm 1), we have

$$\iota_h(s,o) = U_h(s,o) + \sum_{\tau=1}^{H-h+1} [T_h \widehat{V}_{h+\tau}](s,o) - \widehat{Q}_h(s,o) = U_h(s,o) + \sum_{\tau=1}^{H-h+1} [T_h \widehat{V}_{h+\tau}](s,o) \geq 0$$

Otherwise, if $\overline{Q}_h(s,o) \geq 0$, we have

$$\begin{aligned}\iota_h(s,o) \geq & U_h(s,o) + \sum_{\tau=1}^{H-h+1}[T_h \widehat{V}_{h+\tau}](s,o) - \overline{Q}_h(s,o) \\ = & U_h(s,o) - \widehat{U}_h(s,o) + \sum_{\tau=1}^{H-h+1}[(T_h - \widehat{T}_h)\widehat{V}_{h+\tau}](s,o) + \Gamma_h(s,o) \geq 0\end{aligned}$$

Conditioned on the successful event $\mathcal{E}$, we have that

$$\begin{aligned}\overline{Q}_h(s,o) = & \widehat{U}_h(s,o) + \sum_{\tau=1}^{H-h+1}[\widehat{T}_h \widehat{V}_{h+\tau}](s,o) - \Gamma_h(s,o) \\ \leq & U_h(s,o) + \sum_{\tau=1}^{H-h+1}[T_h \widehat{V}_{h+\tau}](s,o) \leq H-h+1\end{aligned}$$

Hence, we obtain that

$$\widehat{Q}_h(s,o) = \min\{\overline{Q}_h(s,o), H-h+1\}^+ = \max\{\overline{Q}_h(s,o), 0\} \geq \overline{Q}_h(s,o)$$

Therefore, we finally derive that

$$\begin{aligned}\iota_h(s,o) = & U_h(s,o) + \sum_{\tau=1}^{H-h+1}[T_h \widehat{V}_{h+\tau}](s,o) - \widehat{Q}_h(s,o) \\ \leq & U_h(s,o) + \sum_{\tau=1}^{H-h+1}[T_h \widehat{V}_{h+\tau}](s,o) - \overline{Q}_h(s,o) \\ = & U_h(s,o) - \widehat{U}_h(s,o) + \sum_{\tau=1}^{H-h+1}[(T_h - \widehat{T}_h)\widehat{V}_{h+\tau}](s,o) + \Gamma_h(s,o) \leq 2\Gamma_h(s,o)\end{aligned}$$

which concludes the proof. □

G.7 Proof of Lemma 6

Proof. It suffices to show that for all $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$, it holds that with probability at least $1-\xi$

$$\left|\widehat{U}_h(s,o) - U_h(s,o)\right| + \left|\sum_{\tau=1}^{H-h+1}[(\widehat{T}_h - T_h)\widehat{V}_{h+\tau}](s,o)\right| \leq \Gamma_h(s,o) \tag{43}$$

where $\widehat{T}_h(\cdot,\cdot)$ and $\widehat{U}_h(\cdot,\cdot)$ are given in lines 6 and 7 in Subroutine 2, respectively, $\Gamma_h(\cdot,\cdot)$ is defined in Equation (26), and $\widehat{V}_h(\cdot)$ is defined in line 8 of Algorithm 1. Recall that $n_h(s,o)$ is the number of visits to state-option pair (s,o) at the hth timestep in subdataset $\mathcal{D}_{1,h}$. By Hoeffding's Inequality and noting that $U_h(\cdot,\cdot) \in [0, H-h+1]$, for any $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$, it holds with probability at least $1-p$ that

$$\left|\widehat{U}_h(s,o) - U_h(s,o)\right| \leq O\left(\sqrt{\frac{H^2}{n_h(s,o) \vee 1}\log\left(\frac{1}{p}\right)}\right) \tag{44}$$

By Hoeffding's Inequality and noting that $\widehat{T}_h$ only depends on $\mathcal{D}_{1,h}$ and $\widehat{V}_{h+\tau}$ only depends on $\mathcal{D}_{1,t>h}$, for any $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$, we have that with probability at least $1-p$

$$\left|\sum_{\tau=1}^{H-h+1}[(\widehat{T}_h - T_h)\widehat{V}_{h+\tau}](s,o)\right| \leq O\left(\sqrt{\frac{H^2}{n_h(s,o) \vee 1}\log\left(\frac{1}{p}\right)}\right) \tag{45}$$

Therefore, applying a union bound over $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$ and letting $p \leftarrow \frac{\xi}{2HSO}$, we conclude the proof. □

G.8 Proof of Lemma 7

Proof. Our proof relies on the following lemma. (The detailed proof can be found in Appendix G.9.)

Lemma 9. *With probability at least $1-\xi$, for any $(h,s,o) \in [H] \times \mathcal{S} \times \mathcal{O}$, it holds that*

$$\sum_{(s',\tau)\in\mathcal{S}\times[H-h+1]} \left|\widehat{T}_h(s'|s,o,\tau) - T_h(s'|s,o,\tau)\right|$$
$$\leq O\left(\sqrt{\sum_{m=h}^{H} \sum_{(s,a)\in\mathcal{X}_{h,s,o}^m} \frac{HS}{N_m(s,a)\vee 1} \log\left(\frac{HSA}{\xi}\right)} + \phi_h(s,o)\log\left(\frac{HSA}{\xi}\right)\right)$$

where ϕ_h is defined in Equation (32) in the proof.

Similar to the proof of Lemma 6, it suffices to show that for all $(h,s,o) \in [H]\times\mathcal{S}\times\mathcal{O}$, it holds that with probability at least $1-\xi$

$$\left|\widehat{U}_h(s,o) - U_h(s,o)\right| + \left|\sum_{\tau=1}^{H-h+1} [(\widehat{T}_h - T_h)\widehat{V}_{h+\tau}](s,o)\right| \leq \Gamma_h(s,o) \tag{46}$$

where $\widehat{T}_h(\cdot,\cdot)$ and $\widehat{U}_h(\cdot,\cdot)$ are constructed by replacing P_h, r_h, and $\overline{T}_h$ with their empirical counterparts in Equation (11) and Equation (17), respectively, $\Gamma_h(\cdot,\cdot)$ is defined in Equation (31), and $\widehat{V}_h(\cdot)$ is defined in line 8 of Algorithm 1. First, recall that

$$U_h(s,o) = \sum_{a\in\mathcal{A}} \pi_h^o(a|s) r_h(s,a) + \sum_{l\in[H-h]}\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} \overline{T}_h(s'|s,o,l)\pi_{h+l}^o(a'|s')r_{h+l}(s',a')$$

Therefore, we have that with probability at least $1-\xi$,

$$\begin{aligned}
&\left|\widehat{U}_h(s,o) - U_h(s,o)\right| \\
\leq & \sum_{a\in\mathcal{A}} \pi_h^o(a|s)|r_h(s,a) - \widehat{r}_h(s,a)| \\
&+ \sum_{l\in[H-h]}\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} \pi_{h+l}^o(a'|s')\left|\widehat{\overline{T}}_h(s'|s,o,l)\widehat{r}_{h+l}(s',a') - \overline{T}_h(s'|s,o,l)r_{h+l}(s',a')\right| \\
\leq & \sum_{a\in\mathcal{A}} \pi_h^o(a|s)|r_h(s,a) - \widehat{r}_h(s,a)| + \sum_{l\in[H-h]}\sum_{(s',a')\in\mathcal{X}_{h,s,o}^l}\sum_{a'\in\mathcal{A}} \pi_{h+l}^o(a'|s')|r_{h+l}(s',a') - \widehat{r}_{h+l}(s',a')| \\
\leq & O\left(\sum_{m=h}^{H}\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m} \pi_m^o(a|s)\sqrt{\frac{1}{N_m(s,a)\vee 1}\log\left(\frac{HSA}{\xi}\right)}\right) \\
\leq & O\left(\sqrt{\left(\sum_{m=h}^{H}\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}(\pi_m^o(a|s))^2\right)\left(\sum_{m=h}^{H}\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}\frac{1}{N_m(s,a)\vee 1}\log\left(\frac{HSA}{\xi}\right)\right)}\right) \\
\leq & O\left(\sqrt{\sum_{m=h}^{H}\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}\frac{HS}{N_m(s,a)\vee 1}\log\left(\frac{HSA}{\xi}\right)}\right)
\end{aligned}$$

where the last second inequality holds by Cauchy-Schwarz inequality and the last inequality follows from the fact that $\sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}(\pi_m^o(a|s))^2 \leq \sum_{(s,a)\in\mathcal{X}_{h,s,o}^m}\pi_m^o(a|s) \leq S$. In addition, by Lemma 9, we have that

$$\left|\sum_{\tau=1}^{H-h+1}[(\widehat{T}_h - T_h)\widehat{V}_{h+\tau}](s,o)\right|$$

$$\leq H \sum_{\tau \in [H-h+1]} \sum_{s' \in \mathcal{S}} \left| \widehat{T}_h(s'|s,o,\tau) - T_h(s'|s,o,\tau) \right|$$

$$\leq O\left(H \sqrt{\sum_{m=h}^{H} \sum_{(s,a)\in \mathcal{X}^m_{h,s,o}} \frac{HS}{N_m(s,a)\vee 1} \log\left(\frac{HSA}{\xi}\right) + H\phi_h(s,o)\log\left(\frac{HSA}{\xi}\right)} \right)$$

which concludes the proof. □

G.9 Proof of Lemma 9

Proof. We adopt a similar analysis to the proof of (Jin et al., 2020, Lemma 4). By the construction of $\widehat{T}_h(s'|s,o,\tau)$ in Line 12 of Subroutine 3, we have that

$$\begin{aligned}
&\widehat{T}_h(s'|s,o,\tau) \\
&= \sum_{a\in\mathcal{A}} \pi^o_h(a|s) \sum_{s''\in\mathcal{S}} \widehat{P}_{h+\tau-1}(s''|s,a)\left(1-\beta^o_{h+1}(s'')\right)\widehat{T}_{h+1}(s'|s'',o,\tau-1) \\
&= \beta^o_{h+\tau}(s') \sum_{\{s_t,a_t\}_{t=h}^{h+\tau-1}} \prod_{t=h}^{h+\tau-1} \left(1-\beta^o_t(s_t)\right)\pi^o_h(a_t|s_t) \prod_{t=h}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t)
\end{aligned}$$

where $s_h = s, s_{h+\tau} = s'$. Define $\eta^o_h(a|s) := (1-\beta^o_h(s))\pi^o_h(a|s)$. Hence, we have

$$\begin{aligned}
&\widehat{T}_h(s'|s,o,\tau) - T_h(s'|s,o,\tau) \\
=&\beta^o_{h+\tau}(s') \sum_{\{s_t,a_t\}_{t=h}^{h+\tau-1}} \prod_{t=h}^{h+\tau-1} \eta^o_t(a_t|s_t) \underbrace{\left(\prod_{t=h}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t) - \prod_{t=h}^{h+\tau-1} P_t(s_{t+1}|s_t,a_t)\right)}_{(*)}
\end{aligned} \tag{47}$$

Consider any such trajectory $\{s_t,a_t\}_{t=h}^{h+\tau-1}$. To bound term (*) by the error in estimating the one-step transition kernel, i.e., $|P_h(s'|s,a) - \widehat{P}_h(s'|s,a)|$, we add and subtract $\tau - 1$ terms and rewrite it as

$$\begin{aligned}
(*) =& \underbrace{\prod_{t=h}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t)}_{\text{(i)}} - \prod_{t=h}^{h+\tau-1} P_t(s_{t+1}|s_t,a_t) \underbrace{\pm \sum_{m=h+1}^{h+\tau-1}\prod_{t=h}^{m-1} P_t(s_{t+1}|s_t,a_t) \prod_{t=m}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t)}_{\text{(ii)}} \\
=& \underbrace{(\widehat{P}_h(s_{h+1}|s_h,a_h) \pm P_h(s_{h+1}|s_h,a_h)) \prod_{t=h+1}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t)}_{\text{(i)}} \\
& \underbrace{\pm P_h(s_{h+1}|s_h,a_h) \prod_{t=h+1}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t) \pm \sum_{m=h+2}^{h+\tau-1}\prod_{t=h}^{m-1} P_t(s_{t+1}|s_t,a_t) \prod_{t=m}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t)}_{\text{(ii)}} \\
& - \prod_{t=h}^{h+\tau-1} P_t(s_{t+1}|s_t,a_t) \\
=&(\widehat{P}_h(s_{h+1}|s_h,a_h) - P_h(s_{h+1}|s_h,a_h)) \prod_{t=h+1}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t) \\
& + P_h(s_{h+1}|s_h,a_h)\left(\widehat{P}_{h+1}(s_{h+2}|s_{h+1},a_{h+1}) \pm P_{h+1}(s_{h+2}|s_{h+1},a_{h+1})\right) \prod_{t=h+2}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t) \\
& \pm \sum_{m=h+2}^{h+\tau-1}\prod_{t=h}^{m-1} P_t(s_{t+1}|s_t,a_t) \prod_{t=m}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t) - \prod_{t=h}^{h+\tau-1} P_t(s_{t+1}|s_t,a_t)
\end{aligned}$$

$$
\begin{aligned}
&= \cdots \\
&= \sum_{m=h}^{h+\tau-1} \left(\widehat{P}_m(s_{m+1}|s_m,a_m) - P_m(s_{m+1}|s_m,a_m)\right) \prod_{t=h}^{m-1} P_t(s_{t+1}|s_t,a_t) \prod_{t=m+1}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t)
\end{aligned} \tag{48}
$$

The following lemma shows the error of the empirical transition kernel.

Lemma 10. *By the empirical Bernstein Inequality, with probability at least* $1-4p$*, it holds that for any* $(h,s,a,s') \in [H] \times \mathcal{S}^2 \times \mathcal{A}$

$$
\left|P_h(s'|s,a) - \widehat{P}_h(s'|s,a)\right| \leq \sqrt{\frac{2\widehat{P}_h(s'|s,a)(1-\widehat{P}_h(s'|s,a))}{(N_h(s,a)-1)\vee 1}\log\left(\frac{1}{p}\right)} + \frac{7\log\left(\frac{1}{p}\right)}{3((N_h(s,a)-1)\vee 1)}
$$

which implies

$$
\left|P_h(s'|s,a) - \widehat{P}_h(s'|s,a)\right| \leq \epsilon_h(s'|s,a) \tag{49}
$$

for any $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$*, where* $N_h(s,a)$ *is the number of visits to the state-action pair* (s,a) *at the hth timestep in Dataset* $\mathcal{D}_2$ *and*

$$
\epsilon_h(s'|s,a) := \min\left\{1, O\left(\sqrt{\frac{P_h(s'|s,a)}{N_h(s,a)\vee 1}\log\left(\frac{HSA}{p}\right)} + \frac{\log\left(\frac{HSA}{p}\right)}{N_h(s,a)\vee 1}\right)\right\} \tag{50}
$$

Proof. See the proof of (Jin et al., 2020, Lemmas 2 and 8). □

Combining Inequalities (47), (48), and (50), we further derive that

$$
\begin{aligned}
&\left|\widehat{T}_h(s'|s,o,\tau) - T_h(s'|s,o,\tau)\right| \\
\leq &\beta^o_{h+\tau}(s') \sum_{\{s_t,a_t\}_{t=h}^{h+\tau-1}} \prod_{t=h}^{h+\tau-1} \eta^o_t(a_t|s_t) \sum_{m=h}^{h+\tau-1} \epsilon_m(s_{m+1}|s_m,a_m) \prod_{t=h}^{m-1} P_t(s_{t+1}|s_t,a_t) \prod_{t=m+1}^{h+\tau-1} \widehat{P}_t(s_{t+1}|s_t,a_t) \\
= &\sum_{m=h}^{h+\tau-1} \sum_{\{s_t,a_t\}_{t=h}^{h+\tau-1}} \epsilon_m(s_{m+1}|s_m,a_m) \left(\eta^o_m(a_m|s_m) \prod_{t=h}^{m-1} \eta^o_t(a_t|s_t) P_t(s_{t+1}|s_t,a_t)\right) \\
&\qquad \cdot \left(\beta^o_{h+\tau}(s') \prod_{t=m+1}^{h+\tau-1} \eta^o_t(a_t|s_t)\widehat{P}_t(s_{t+1}|s_t,a_t)\right) \\
= &\sum_{m=h}^{h+\tau-1} \sum_{s_m,a_m,s_{m+1}} \epsilon_m(s_{m+1}|s_m,a_m) \left(\sum_{\{s_t,a_t\}_{t=h}^{m-1}} \eta^o_m(a_m|s_m) \prod_{t=h}^{m-1} \eta^o_t(a_t|s_t) P_t(s_{t+1}|s_t,a_t)\right) \\
&\qquad \cdot \left(\sum_{a_{m+1}} \sum_{\{s_t,a_t\}_{t=m+2}^{h+\tau-1}} \beta^o_{h+\tau}(s') \prod_{t=m+1}^{h+\tau-1} \eta^o_t(a_t|s_t)\widehat{P}_t(s_{t+1}|s_t,a_t)\right) \\
= &\sum_{m=h}^{h+\tau-1} \sum_{s_m,a_m,s_{m+1}} \epsilon_m(s_{m+1}|s_m,a_m) \overline{T}_h(s_m|s,o,m-h) \widehat{T}_h(s'|s_{m+1},o,h+\tau-(m+1))
\end{aligned} \tag{51}
$$

Similarly, we have that

$$
\begin{aligned}
&|\widehat{T}_h(s'|s_{m+1},o,h+\tau-(m+1)) - T_h(s'|s_{m+1},o,h+\tau-(m+1))| \\
\leq &\sum_{t=m+1}^{h+\tau-1} \sum_{s'_t,a'_t,s'_{t+1}} \epsilon_t(s'_{t+1}|s'_t,a'_t) \overline{T}_h(s'_t|s_{m+1},o,t-(m+1)) \widehat{T}_h(s'|s'_t,o,h+\tau-t)
\end{aligned}
$$

$$\leq \beta_{h+\tau}^{o}(s') \sum_{t=m+1}^{h+\tau-1} \sum_{s'_t, a'_t, s'_{t+1}} \epsilon_t(s'_{t+1}|s'_t, a'_t) \overline{T}_h(s'_t|s_{m+1}, o, t-(m+1)) \tag{52}$$

for any $m \in \{h, \cdots, h+\tau-1\}$, where the last inequality holds by $\widehat{T}_h(s'|s'_t, o, h+\tau-t) \leq \beta_{h+\tau}^{o}(s')$. Let $w_m := (s_m, a_m, s_{m+1})$. Combining Inequality (51) and (52), we derive that

$$\begin{aligned}
&\sum_{(s',\tau)\in\mathcal{S}\times[H-h+1]} \left|\widehat{T}_h(s'|s,o,\tau) - T_h(s'|s,o,\tau)\right| \\
\leq& \sum_{s',\tau} \sum_{m=h}^{h+\tau-1} \sum_{w_m} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h) T_h(s'|s_{m+1}, o, h+\tau-(m+1)) \\
&+ \sum_{s',\tau} \sum_{m=h}^{h+\tau-1} \sum_{w_m} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h) \\
&\qquad \cdot \left(\beta_{h+\tau}^{o}(s') \sum_{t=m+1}^{h+\tau-1} \sum_{w'_t} \epsilon(s'_{t+1}|s'_t, a'_t) \overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1)) \right) \\
=& \sum_{m=h}^{H} \sum_{w_m} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h) \underbrace{\sum_{s',\tau} T_h(s'|s_{m+1}, o, h+\tau-(m+1))}_{=1 \text{ by Lemma 1}} \\
&+ \sum_{m=h}^{H} \sum_{w_m} \sum_{t=m+1}^{d-1} \sum_{w'_t} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h) \epsilon(s'_{t+1}|s'_t, a'_t) \overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1)) \\
&\qquad \cdot \underbrace{\left(\sum_{s',\tau} \beta_{h+\tau}^{o}(s') \right)}_{\leq HS} \\
\leq& \underbrace{\sum_{m=h}^{H} \sum_{w_m} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h)}_{B_1} \\
&+ HS \underbrace{\sum_{h\leq m<t\leq H} \sum_{w_m, w'_t} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h) \epsilon_t(s'_{t+1}|s'_t, a'_t) \overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}_{B_2}
\end{aligned} \tag{53}$$

Step 1. Bounding term B_1. Recall that from Equation (50)

$$\epsilon_h(s'|s,a) = \min\left\{1, O\left(\sqrt{\frac{P_h(s'|s,a)}{N_h(s,a)\vee 1} \log\left(\frac{HSA}{p}\right)} + \frac{\log\left(\frac{HSA}{p}\right)}{N_h(s,a)\vee 1}\right)\right\}$$

We have that

$$\begin{aligned}
B_1 =& \sum_{m=h}^{H} \sum_{w_m} \epsilon_m(s_{m+1}|s_m, a_m) \overline{T}_h(s_m, a_m|s, o, m-h) \\
=& O\Bigg(\sum_{m=h}^{H} \sum_{w_m} \overline{T}_h(s_m, a_m|s, o, m-h) \sqrt{\frac{P_m(s_{m+1}|s_m, a_m)}{N_m(s_m, a_m)\vee 1} \log\left(\frac{HSA}{p}\right)} \\
&+ \sum_{m=h}^{H} \sum_{w_m} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)}{N_m(s_m, a_m)\vee 1} \log\left(\frac{HSA}{p}\right) \Bigg)
\end{aligned}$$

For the first term, applying Cauchy-Schwarz inequality yields

$$
\begin{aligned}
&\sum_{m=h}^{H}\sum_{w_m}\overline{T}_h(s_m,a_m|s,o,m-h)\sqrt{\frac{P_m(s_{m+1}|s_m,a_m)}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)}\\
&\le\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\overline{T}_h(s_m,a_m|s,o,m-h)\sum_{s_{m+1}\in\mathcal{S}}\sqrt{\frac{P_m(s_{m+1}|s_m,a_m)}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)}\\
&\le\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\overline{T}_h(s_m,a_m|s,o,m-h)\sqrt{\left(\sum_{s_{m+1}\in\mathcal{S}}1\right)\left(\frac{\sum_{s_{m+1}\in\mathcal{S}}P_m(s_{m+1}|s_m,a_m)}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)\right)}\\
&=\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\overline{T}_h(s_m,a_m|s,o,m-h)\sqrt{\frac{S}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)}\\
&\le\sqrt{\left(\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}(\overline{T}_h(s_m,a_m|s,o,m-h))^2\right)\left(\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\frac{S}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)\right)}\\
&\le\sqrt{\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\frac{HS}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)}
\end{aligned}
$$

where the last inequality holds by $\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}(\overline{T}_h(s_m,a_m|s,o,m-h))^2 \le \sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\overline{T}_h(s_m,a_m|s,o,m-h) \le 1$, i.e., it is the probability that the agent does not terminate the option at timestep m. Hence, we obtain that

$$
\begin{aligned}
B_1 \le O\Bigg(&\sqrt{\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\frac{HS}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)}\\
&+\sum_{m=h}^{H}\sum_{(s_m,a_m)\in\mathcal{X}_{h,s,o}^m}\frac{S}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)\Bigg)
\end{aligned}
\tag{54}
$$

Step 2. Bounding term B_2. Since $\epsilon_h, \overline{T}_h \le 1$, we have that

$$
\begin{aligned}
B_2 \le &\sum_{h\le m<t\le H}\sum_{w_m,w_t'}\Bigg(\sqrt{\frac{P_m(s_{m+1}|s_m,a_m)}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)}\overline{T}_h(s_m,a_m|s,o,m-h)\\
&\qquad\cdot\sqrt{\frac{P_t(s_{t+1}|s_t,a_t)}{N_t(s_t',a_t')\vee 1}\log\left(\frac{HSA}{p}\right)}\overline{T}_h(s_t',a_t'|s_{m+1},o,t-(m+1))\Bigg)\\
&+\sum_{h\le m<t\le H}\sum_{w_m,w_t'}\frac{\overline{T}_h(s_m,a_m|s,o,m-h)}{N_m(s_m,a_m)\vee 1}\log\left(\frac{HSA}{p}\right)\\
&+\sum_{h\le m<t\le H}\sum_{w_m,w_t'}\frac{\overline{T}_h(s_t',a_t'|s_{m+1},o,t-(m+1))}{N_t(s_t',a_t')\vee 1}\log\left(\frac{HSA}{p}\right)
\end{aligned}
\tag{55}
$$

Applying the Cauchy-Schwartz inequality, the first term of Inequality (55) can be written as

$$
\sum_{h\le m<t\le H}\sum_{w_m,w_t'}\log\left(\frac{HSA}{p}\right)\sqrt{\frac{\overline{T}_h(s_m,a_m|s,o,m-h)P_t(s_{t+1}'|s_t',a_t')\overline{T}_h(s_t',a_t'|s_{m+1},o,t-(m+1))}{N_m(s_m,a_m)\vee 1}}
$$

$$
\begin{aligned}
&\cdot \sqrt{\frac{\overline{T}_h(s_m, a_m|s, o, m-h)P_m(s_{m+1}|s_m, a_m)\overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}{N_t(s'_t, a'_t) \vee 1}} \\
\leq &\sum_{h \leq m < t \leq H} \log\left(\frac{HSA}{p}\right) \sqrt{\sum_{w_m, w'_t} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)P_t(s'_{t+1}|s'_t, a'_t)\overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}{N_m(s_m, a_m) \vee 1}} \\
&\cdot \sqrt{\sum_{w_m, w'_t} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)P_m(s_{m+1}|s_m, a_m)\overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}{N_t(s'_t, a'_t) \vee 1}} \\
\leq &\sum_{h \leq m < t \leq H} \log\left(\frac{HSA}{p}\right) \sqrt{S \sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)}{N_m(s_m, a_m) \vee 1}} \cdot \sqrt{S \sum_{(s'_t, a'_t) \in \mathcal{X}^t_{h,s,o}} \frac{\overline{T}_h(s'_t, a'_t|s, o, t-h)}{N_t(s'_t, a'_t) \vee 1}}
\end{aligned} \tag{56}
$$

where the last inequality holds by

$$
\begin{aligned}
&\sum_{w_m, w'_t} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)P_t(s'_{t+1}|s'_t, a'_t)\overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}{N_m(s_m, a_m) \vee 1} \\
=&\sum_{w_m} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)}{N_m(s_m, a_m) \vee 1} \underbrace{\sum_{s'_t, a'_t} \overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}_{\leq 1} \underbrace{\sum_{s'_{t+1}} P_t(s'_{t+1}|s'_t, a'_t)}_{=1} \\
\leq &S \sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)}{N_m(s_m, a_m) \vee 1}
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{w_m, w'_t} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)P_m(s_{m+1}|s_m, a_m)\overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}{N_t(s'_t, a'_t) \vee 1} \\
\leq &S \sum_{(s'_t, a'_t) \in \mathcal{X}^t_{h,s,o}} \frac{\overline{T}_h(s'_t, a'_t|s, o, t-h)}{N_t(s'_t, a'_t) \vee 1}
\end{aligned}
$$

by the same analysis.

In addition, the second term of Inequality (55) can be bounded by

$$
\begin{aligned}
&\sum_{h \leq m < t \leq H} \sum_{w_m, w'_t} \frac{\overline{T}_h(s_m, a_m|s, o, m-h)}{N_m(s_m, a_m) \vee 1} \log\left(\frac{HSA}{p}\right) \\
\leq &\sum_{h \leq m < t \leq H} \sum_{w_m, w'_t} \frac{1}{N_m(s_m, a_m) \vee 1} \log\left(\frac{HSA}{p}\right) \\
=&S^2 \sum_{h \leq m < t \leq H} \sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}, (s'_t, a'_t) \in \mathcal{X}^t_{h,s,o}} \frac{1}{N_m(s_m, a_m) \vee 1} \log\left(\frac{HSA}{p}\right)
\end{aligned} \tag{57}
$$

By the exact analysis, the third term of Inequality (55) can be bounded by

$$
\begin{aligned}
&\sum_{h \leq m < t \leq H} \sum_{w_m, w'_t} \frac{\overline{T}_h(s'_t, a'_t|s_{m+1}, o, t-(m+1))}{N_t(s'_t, a'_t) \vee 1} \log\left(\frac{1}{p}\right) \\
\leq &S^2 \sum_{h \leq m < t \leq H} \sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}, (s'_t, a'_t) \in \mathcal{X}^t_{h,s,o}} \frac{1}{N_t(s'_t, a'_t) \vee 1} \log\left(\frac{HSA}{p}\right)
\end{aligned} \tag{58}
$$

Combining Inequalities (56), (57), and (58), we have that

$$\begin{aligned} B_2 \leq & S \sum_{h \leq m < t \leq H} \sqrt{\sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}} \frac{1}{N_m(s_m, a_m) \vee 1}} \cdot \sqrt{\sum_{(s'_t, a'_t) \in \mathcal{X}^t_{h,s,o}} \frac{1}{N_t(s'_t, a'_t) \vee 1}} \\ & + S^2 \sum_{h \leq m < t \leq H} \sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}, (s'_t, a'_t) \in \mathcal{X}^t_{h,s,o}} \left(\frac{1}{N_m(s_m, a_m) \vee 1} + \frac{1}{N_t(s'_t, a'_t) \vee 1} \right) \end{aligned} \tag{59}$$

Step 3. Putting all together. Therefore, combining Inequalities (53), (54), and (59), we derive that

$$\begin{aligned} & \sum_{(s', \tau) \in \mathcal{S} \times [H-h+1]} \left| \widehat{T}_h(s'|s, o, \tau) - T_h(s'|s, o, \tau) \right| \\ \leq & B_1 + HS \cdot B_2 \\ \leq & \sqrt{\sum_{m=h}^{H} \sum_{(s_m, a_m) \in \mathcal{X}^m_{h,s,o}} \frac{HS}{N_m(s_m, a_m) \vee 1} \log\left(\frac{HSA}{p}\right)} + \phi_h(s, o) \log\left(\frac{HSA}{p}\right) \end{aligned}$$

where ϕ_h is defined in Equation (32). $\square$

H Counterexample

Example 1. We consider an episodic MDP with the following structures: (1) $P_{H-1}(x|s, a) = 1$ for some $x \in \mathcal{S}$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$, i.e., the agent is guaranteed to arrive state x at the Hth timestep using any hierarchical policy, (2) the set of actions $\mathcal{A} = \{a^1, a^2\}$, which satisfies that $r_H(x, a^1) = 1$ and $r_H(x, a^2) = 0$, (3) the set of options $\mathcal{O} = \{o_1, o_2\}$, which satisfies that: (3.i) o_1 and o_2 are exactly the same before timestep H, i.e., $\pi_h^{o_1}(a|s) = \pi_h^{o_2}(a|s), \beta_h^{o_1}(s) = \beta_h^{o_2}(s)$ for any $(h, s, a) \in [H-1] \times \mathcal{S} \times \mathcal{A}$, (3.ii) at state x at timestep H, option o_1 always takes action a^1 while option o_2 takes action a^2 with probability $\epsilon \in (0, 1)$, i.e., $\pi_H^{o_1}(a^1|x) = 1$, $\pi_H^{o_2}(a^1|x) = 1 - \epsilon$, $\pi_H^{o_2}(a^2|x) = \epsilon$, (3.iii) the agent guarantees to terminate option o_{h-1} and select a new option at timestep H, i.e., $\beta_H^{o_1}(x) = \beta_H^{o_2}(x) = 1$. It can be easily seen that $\mu_H^*(x) = o_1$. If the hierarchical behavior policy ρ to collect the dataset $\mathcal{D}_1$ always selects o_2 in the Hth timestep, i.e., $\rho_H(x) = o_2$, then $\widehat{\mu}_H(x) = o_2$ since no information of o_1 is provided. Therefore, it holds that $C_1^{\text{option}} = \infty$ while $C_2^{\text{option}} < \infty$.