Robust Knowledge Transfer in Tiered RL

Jiawei Huang
Department of Computer Science
ETH Zurich
jiawei.huang@inf.ethz.ch

Niao He
Department of Computer Science
ETH Zurich
jiawei.huang@inf.ethz.ch

Abstract

In this paper, we study the Tiered Reinforcement Learning setting, a parallel transfer learning framework, where the goal is to transfer knowledge from the low-tier (source) task to the high-tier (target) task to reduce the exploration risk of the latter while solving the two tasks in parallel. Unlike previous work, we do not assume the low-tier and high-tier tasks share the same dynamics or reward functions, and focus on robust knowledge transfer without prior knowledge on the task similarity. We identify a natural and necessary condition called the "Optimal Value Dominance" for our objective. Under this condition, we propose novel online learning algorithms such that, for the high-tier task, it can achieve constant regret on partial states depending on the task similarity and retain near-optimal regret when the two tasks are dissimilar, while for the low-tier task, it can keep near-optimal without making sacrifice. Moreover, we further study the setting with multiple low-tier tasks, and propose a novel transfer source selection mechanism, which can ensemble the information from all low-tier tasks and allow provable benefits on a much larger state-action space.

1 Introduction

Comparing with individual learning from scratch, transferring knowledge from other similar tasks or side information has been proven to be an effective way to reduce the exploration risk and improve sample efficiency in Reinforcement Learning (RL). Multi-Task RL (MT-RL) [29] and Transfer RL [26, 18, 37] are two mainstream knowledge transfer frameworks; however, both are subject to limitations when dealing with real-world scenarios. MT-RL studies the setting where a set of similar tasks are solved concurrently, and the main objective is to accelerate the learning by sharing information of all tasks together. However, in practice, in many MT-RL scenarios, the tasks are not equally important and we are more interested in the performance of certain tasks. For example, in robot learning, a few robots are more valuable and hard to fix, while the others are cheaper or just simulators. Most existing works on MT-RL treat all tasks equally and focus primarily on the reduction of the total regret of all tasks as a whole [3, 8, 36, 11], with no guarantee of improving a particular task. In contrast, transfer RL distinguishes the priority of different tasks by categorizing them into source and target tasks and aims at transferring the knowledge from source tasks (or some side information like value predictors) to facilitate the learning of target tasks [21, 27, 9, 10]. However, a key assumption in transfer RL is that the source task is completely solved before the learning of the target task, and this is not always practical. For example, in some sim-to-real domain, the source task simulator may require a long time to solve [5], and in some user-interaction scenarios [13], the source and target tasks refer to different user groups and they have to be served simultaneously. In these cases, it's more reasonable to solve the source and target tasks in parallel and transfer the information immediately once available.

Recently, [13] proposed a new "parallel knowledge transfer" framework, called Tiered RL, which is promising to fill the gap. Tiered RL considers the case when a source task M_{Lo} and a target task

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

M_{Hi} are learned in parallel, by two separate algorithms Alg^{Lo} and Alg^{Hi}, and its goal is to reduce the exploration risk and regret in learning M_{Hi} by leveraging knowledge transfer from M_{Lo} to M_{Hi}. [13] showed that under the strong assumption that $M_{\text{Lo}} = M_{\text{Hi}}$, it's possible to achieve constant regret in learning M_{Hi} while keeping regret in M_{Lo} optimal in gap-dependent setting. Yet, their algorithm based on Pessimistic Value Iteration (PVI) [16] can be hardly applied when the assumption $M_{\text{Lo}} = M_{\text{Hi}}$ breaks, given its pure exploitation nature, making their result very restrictive.

In this paper, we study the general Tiered RL setting *without prior knowledge* about how similar the tasks are [1]. The key question we would like to address is: **Can we design algorithms s.t.: (1) Regret in M_{Lo} keeps near-optimal; (2) Regret in M_{Hi} achieves provable benefits when M_{Lo} and M_{Hi} are similar while retaining near-optimal otherwise?** Note for Alg^{Lo}, we expect it to achieve near-optimal regret bounds, which is reasonable since the source task is often important and our results still hold if relaxing it. As for Alg^{Hi}, we expect it to be *robust*, i.e., it can adaptively exploit from M_{Lo} if it is close to M_{Hi}, while avoiding negative transfer in other cases. Notably, our setting strictly generalizes [13] and is much more challenging, for balancing the exploitation from M_{Lo} and exploration from M_{Hi} without prior knowledge of task similarity. We give positive answers to the above question in this paper and demonstrate provable benefits with robust knowledge transfer for Tiered RL framework. Below, we summarize our main contributions in three aspects.

Our first contribution is to identify essential conditions and notions about when and how our objective is achievable. We first provide a mild condition called *Optimal Value Dominance* (OVD), and in Sec. 3, we justify its necessity to our objective by a lower bound result. Our lower bound holds even if M_{Lo} is fully known to the learner, and therefore, it also justifies the necessity of similar assumptions in previous transfer RL literatures [9, 10]. Besides, we introduce the notion of *transferable states* to characterize states on which M_{Hi} is expected to achieve benefits by transferring knowledge from M_{Lo}. We believe those findings also provide useful insights for further works.

As **our second contribution**, in Sec. 4, we propose novel algorithms for Tiered Multi-Armed Bandit (MAB) and Tiered RL, which can achieve robust parallel transfer by balancing between pessimism-based exploitation from M_{Lo} and optimism-based online learning in M_{Hi}. Depending on the similarity between M_{Lo} and M_{Hi}, our algorithms can enjoy constant regret on a proportion of state-action pairs or even on the entire M_{Hi} by leveraging information from M_{Lo}, while timely interrupting negative transfer and retaining near-optimal regret on states dissimilar between two tasks. Moreover, in the bandit setting, our result implies a strictly improved regret bound when $M_{\text{Hi}} = M_{\text{Lo}}$, compared with previous results under the same setting [22, 13].

Beyond the single low-tier task setting, in many real-world scenarios, it's reasonable to assume there are multiple different low-tier tasks $\mathcal{M}_{\text{Lo}} = \{M_{\text{Lo},w}\}_{w=1}^{W}$ available. As **our third contribution**, in Sec. 5, we extend our algorithm to this setting with a new source task selection mechanism. By novel techniques, we show that, even if each $M_{\text{Lo},w}$ may only share similarity with M_{Hi} on a small portion of states, we are able to ensemble the information from each source task together to achieve constant regret on a much larger state-action space "stitched" from the transferable state-action set in each individual $M_{\text{Lo},w}$, at the expense of an additional $\log W$ factor in regret. Besides, our algorithm is still robust to model difference and retains near-optimal regret in general. Although we only study the Tiered RL setting in this paper, we believe our task selection strategy can be applied to standard transfer RL setting [9, 10] when multiple (partially correct) side information or value predictors are provided, which is an interesting direction for future work. Finally, we conduct experiments in toy examples to verify our theoretical results.

1.1 Closely Related Work

For the lack of space, we only discuss closely related work here and defer the rest to Appx. A.2. The most related to us is the Tiered RL framework [13], which was originally motivated by Tiered structure in user-oriented applications. However, they only studied the case when $M_{\text{Hi}} = M_{\text{Lo}}$, which limits the practicability of their algorithms. Although there is a sequence of work studying parallel transfer learning in multi-agent system [25, 19], but they mainly focused on heuristic empirical algorithms and did not have theoretical guarantees.

Transfer RL [37], compared to learning from scratch, can reduces exploration risk of target task by leveraging information in similar source tasks or side information [21, 27, 9, 10]. Comparing with

[1] We defer the detailed framework to Appx. A.1 for completeness, which is the same as Fr. 1 in [13].

transfer RL setting, our parallel transfer setting has some additional challenges. Firstly, M_{Hi} can only leverage estimated value/model/optimal policy from M_{Lo} with uncertainty, which implies we need additional efforts to control failure events with non-zero probability comparing with normal transfer RL setting. Secondly, the constraints on the optimality of $\text{Regret}_K(M_{\text{Lo}})$, although reasonable, restrict the transferable information because in M_{Lo} estimation uncertainty can only be controlled on those states frequently visited. Moreover, none of these previous work studies how to leverage multiple partially correct side information like what we did in Sec. 5. In MT-RL setting, the benefits of leveraging information gathered from other tasks has been observed from both empirical and theoretical works [31, 3, 8, 24, 36, 11]. But MTRL treats each task equally, and the reduction of total regret over all tasks does not directly imply benefits achieved in a particular task.

2 Preliminary and Problem Formulation

Tiered Stochastic MAB and Tiered Episodic Tabular RL In Tiered MAB setting, we consider a low-tier task M_{Lo} and a high-tier task M_{Hi} sharing the arm/action space $\mathcal{A} = \{1, 2, ..., A\}$. By pulling the arm $i \in [A]$ in M_{Lo} or M_{Hi}, the agent can observe a random variable $r_{\text{Lo}}(i)$ or $r_{\text{Hi}}(i) \in [0, 1]$. We will use $\mu_{\text{Lo}}(i) = \mathbb{E}[r_{\text{Lo}}(i)]$ and $\mu_{\text{Hi}}(i) = \mathbb{E}[r_{\text{Hi}}(i)]$ to denote the expected return of the i-th arm in M_{Lo} and M_{Hi}, respectively, and note that it's possible that $\mu_{\text{Lo}}(i) \neq \mu_{\text{Hi}}(i)$.

For Tiered RL, we assume that two tasks $M_{\text{Lo}} = \{\mathcal{S}, \mathcal{A}, H, \mathbb{P}_{\text{Lo}}, r_{\text{Lo}}\}$ and $M_{\text{Hi}} = \{\mathcal{S}, \mathcal{A}, H, \mathbb{P}_{\text{Hi}}, r_{\text{Hi}}\}$ share the finite state $\mathcal{S}$ and action space $\mathcal{A}$ across episode length H (i.e. $\mathcal{S}_h = \mathcal{S}, \mathcal{A}_h = \mathcal{A}$ for any $h \in [H]$), but may have different time-dependent transition and reward functions $\mathbb{P}_{\text{Lo}} = \{\mathbb{P}_{\text{Lo},h}\}_{h=1}^H, r_{\text{Lo}} = \{r_{\text{Lo},h}\}_{h=1}^H$ and $\mathbb{P}_{\text{Hi}} = \{\mathbb{P}_{\text{Hi},h}\}_{h=1}^H, r_{\text{Hi}} = \{r_{\text{Hi},h}\}_{h=1}^H$. W.l.o.g., we assume the initial state s_1 is fixed, and the reward functions r_{Lo} and r_{Hi} are deterministic and bounded by $[0, 1]$. In episodic MDPs, we study the time-dependent policy specified as $\pi := \{\pi_1, ..., \pi_H\}$ with $\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A}_h)$ for all $h \in [H]$, where $\Delta(\mathcal{A}_h)$ denotes the probability simplex over the action space. With a slight abuse of notation, when π_h is a deterministic policy, we use $\pi_h : \mathcal{S}_h \to \mathcal{A}_h$ to denote the deterministic mapping. Besides, we use $Q_h^\pi(s, a) = \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a, \pi]$, $V_h^\pi(s) = \mathbb{E}_{a\sim\pi}[Q_h^\pi(s, a)]$ to denote the value function for π at step $h \in [H]$, and denote $d_h^\pi(\cdot) := \Pr(s_h = \cdot|\pi)$ and $d_h^\pi(\cdot, \cdot) := \Pr(s_h = \cdot, a_h = \cdot|\pi)$ as the state and state-action occupancy w.r.t. policy π. We use π^* to denote the optimal policy, and V_h^*, Q_h^*, d_h^* as a short note when $\pi = \pi^*$. To avoid confusion, we will specify the policy and value functions in M_{Lo} and M_{Hi} by Lo and Hi in subscription, respectively. For example, in M_{Lo} we have $\pi_{\text{Lo}} := \{\pi_{\text{Lo},1}, ..., \pi_{\text{Lo},H}\}$, $Q_{\text{Lo},h}^{\pi_{\text{Lo}}}/V_{\text{Lo},h}^{\pi_{\text{Lo}}}$ and $Q_{\text{Lo},h}^*/V_{\text{Lo},h}^*$, $d_{\text{Lo},h}^{\pi_{\text{Lo}}}, d_{\text{Lo},h}^*$, and similarly for M_{Hi}.

Gap-Dependent Setting Throughout, we focus on gap-dependent setting [17, 23, 34, 7]. Below we introduce the notion of gap for M_{Lo} as an example, and those for M_{Hi} follows similarly. In MAB case, the gap in M_{Lo} w.r.t. arm i is defined as $\Delta_{\text{Lo}}(i) := \max_{j\in[A]} \mu_{\text{Lo}}(j) - \mu_{\text{Lo}}(i), \forall i \in [A]$, and for tabular RL setting, we have $\Delta_{\text{Lo}}(s_h, a_h) := V_{\text{Lo},h}^*(s_h) - Q_{\text{Lo},h}^*(s_h, a_h), \forall h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$. We use $\Delta_{\text{Lo,min}}$ to refer to the minimal gap such that $\Delta(s_h, a_h) \geq \Delta_{\text{Lo,min}}$ for all non-optimal actions a_h, and use $\Delta_{\min} := \min\{\Delta_{\text{Lo,min}}, \Delta_{\text{Hi,min}}\}$ to denote the minimal gap over two tasks. In the gap-dependent setting, we assume $\Delta_{\min} > 0$.

Knowledge Transfer from Multiple Low-Tier Tasks In this case, we assume there are $W > 1$ different source tasks $\mathcal{M}_{\text{Lo}} = \{M_{\text{Lo},w}\}_{w=1}^W$ and all the tasks share the same state and action spaces but may have different transition and reward function. We defer the extended framework for this setting to Appx. A. We specify the task index $w \in [W]$ in sub-scription to distinguish the notation for different source tasks (e.g. $\mathbb{P}_{\text{Lo},w,h}, Q_{\text{Lo},w,h}^{(\cdot)}$). Moreover, we define $\Delta_{\min} := \min\{\Delta_{\text{Lo},1,\text{min}}, ..., \Delta_{\text{Lo},W,\text{min}}, \Delta_{\text{Hi,min}}\}$. For convenience, in the rest of the paper, we use TRL-MST (Tiered RL with Multiple Source Task) as a short abbreviation for this setting.

Performance Measure We use Pseudo-Regret as performance measure: $\text{Regret}_K(M_{\text{Lo}}) := \mathbb{E}\left[\sum_{k=1}^K V_1^*(s_1) - V_1^{\pi_{\text{Lo}}^k}(s_1)\right]$; $\text{Regret}_K(M_{\text{Hi}}) := \mathbb{E}\left[\sum_{k=1}^K V_1^*(s_1) - V_1^{\pi_{\text{Hi}}^k}(s_1)\right]$, where K is the number of iterations, $\{\pi_{\text{Lo}}^k\}_{k=1}^K$ and $\{\pi_{\text{Hi}}^k\}_{k=1}^K$ are generated by the algorithms.

Frequently Used Notations We denote $[n] = \{1, 2, ..., n\}$. Given a transition matrix $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and a function $V : \mathcal{S} \to \mathbb{R}$, we use $\mathbb{P}V(s_h, a_h)$ as a short note of $\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}[V(s')]$. We will use $i_{\text{Lo}}^*/i_{\text{Hi}}^*$ to denote the optimal arm in bandit setting, and $\pi_{\text{Lo}}^*/\pi_{\text{Hi}}^*$ denotes the optimal policy in RL setting. In TRL-MST setting, we use $i_{\text{Lo},w}^*/\pi_{\text{Lo},w}^*$ to distinguish different source tasks.

2.1 Assumptions and Characterization of Transferable States

Throughout the paper, we make several assumptions. The first one is the uniqueness of the optimal policy, which is common in the literature [22, 4].

Assumption A. Both M_{Lo} (or $\{M_{\text{Lo},w}\}_{w=1}^{W}$) and M_{Hi} have unique optimal arms/policies.

Next, we introduce a new concept called "Optimal Value Dominance" (OVD for short), which says that for each state (or at least those states reachable by optimal policy in M_{Lo}), the optimal value of M_{Lo} is an approximate overestimation for the optimal value of M_{Hi}. In Sec. 3, we will use a lower bound to show such a condition is necessary to attain the robust transfer objective.

Assumption B. In single source task setting, we assume M_{Lo} has Optimal Value Dominance (OVD) over M_{Hi}, s.t., $\forall h \in [H]$, for all $s_h \in \mathcal{S}_h$ (or only for those s_h with $d^*_{\text{Lo},h}(s_h) > 0$), we have: $V^*_{\text{Lo},h}(s_h) \geq V^*_{\text{Hi},h}(s_h) - \frac{\Delta_{\min}}{2(H+1)}$. In TRL-MST setting, we assume each $M_{\text{Lo},w}$ has OVD over M_{Hi}.[2]

We remark that Assump. B is a rather mild condition that naturally holds with reward shaping. Note that since $V^*_{\text{Hi},h}(\cdot) \leq H - h$, by shifting the reward function of M_{Lo} to $r'_{\text{Lo},h}(\cdot,\cdot) = r_{\text{Lo},h}(\cdot,\cdot) + 1$, we immediately obtain the OVD property. Even though, such a reward shift may impair the set of transferable states as we will introduce in Def. 2.2. We provide several reasonable settings in Appx. A.3 including identical model [13], small model difference, and known model difference, where Assump. B is satisfied and there exists a non-empty set of transferable states. We also point out that several existing work on transfer RL assumed something similar or even stronger [9, 10], which we defer a thorough discussion to Appx. A.2.

Assumption C. The learner has access to a quantity $\widetilde{\Delta}_{\min}$ satisfying $0 \leq \widetilde{\Delta}_{\min} \leq \Delta_{\min}$.

The final one is about the knowledge of a lower bound of $\Delta_{\min}$, which can always be satisfied by choosing $\widetilde{\Delta}_{\min} = 0$. Nevertheless, it would be more beneficial if the learner has access to some quantity $\widetilde{\Delta}_{\min}$ closer to $\Delta_{\min}$ than 0. As we introduce below, the magnitude of $\widetilde{\Delta}_{\min}$ is related to how we quantify the similarity between M_{Lo} and M_{Hi} and which states we expect to benefit from knowledge transfer. Below we focus on the single source task setting and defer the discussion for TRL-MST setting to Sec. 5.

Definition 2.1 (ε-Close)**.** Task M_{Hi} is ε-close to task M_{Lo} on s_h at step h for some $\varepsilon > 0$, if $V^*_{\text{Lo},h}(s_h) - V^*_{\text{Hi},h}(s_h) \leq \varepsilon$ and $\pi^*_{\text{Hi}}(s_h) = \pi^*_{\text{Lo}}(s_h)$.

Definition 2.2 (λ-Transferable States)**.** State s_h is λ-transferable for some $\lambda > 0$, if $d^*_{\text{Lo}}(s_h) > \lambda$ and M_{Hi} is $\frac{\widetilde{\Delta}_{\min}}{4(H+1)}$-close to M_{Lo} on s_h. The set of λ-transferable states at $h \in [H]$ is denoted as $\mathcal{Z}^{\lambda}_h$.

We regard s_h in M_{Lo} has transferable knowledge to M_{Hi}, if it can be reached by optimal policy in M_{Lo} and the optimal value and action at s_h for two tasks are similar. Here the condition $d^*_{\text{Lo}}(s_h) > 0$ is necessary since in M_{Lo}, only the states reachable by π^*_{Lo} can be explored sufficiently by Alg^{Lo} due to its near optimal regret. Combining with Assump. B, one can observe that the value difference on transferable states are controlled by $|V^*_{\text{Lo},h}(s_h) - V^*_{\text{Hi},h}(s_h)| = O(\frac{\widetilde{\Delta}_{\min}}{H}) \leq O(\frac{\Delta_{\min}}{H})$. As we will show in Thm. 3.2 in Sec. 3, the term $O(\Delta_{\min})$ is indeed unimprovable if we expect robustness.

3 Lower Bound Results: Necessary Condition for Robust Transfer

Now we establish lower bounds that show Assump. B is necessary and how the magnitude of $\Delta_{\min}$ restricts the robust transfer objective. The results in this section are based on two-armed Bernoulli bandits for simplicity, and the proofs are deferred to Appx. B. By extending these hard instances to RL case, there is a gap caused by the additional $\frac{1}{H}$ in Assump. B and Def. 2.2, which comes from the requirement of our algorithm design, and we leave it to the future work.

Justification for Assump. B We show that if Assump. B is violated, it is impossible to have algorithms $(\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}})$ to simultaneously achieve constant regret when $M_{\text{Lo}} = M_{\text{Hi}}$, while retaining sub-linear regret for all the cases regardless of the similarity between M_{Lo} and M_{Hi}. Here we require constant regret on $M_{\text{Lo}} = M_{\text{Hi}}$ since we believe it is a minimal expectation to achieve benefits in transfer when

[2]For convenience, we include the bandit setting as a special case with $H = 0$; see also Def. 2.1, 2.2 and F.1.

the two tasks are identical. Intuitively, without Assump. B, even if we know $\mu_{\text{Lo}}(i^*_{\text{Lo}}) = \mu_{\text{Hi}}(i^*_{\text{Lo}})$, we cannot ensure i^*_{Lo} is the optimal arm in M_{Hi}. Then, if $(\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}})$ can achieve constant regret on $M_{\text{Lo}} = M_{\text{Hi}}$, the algorithm must stop exploration on the arm $i \neq i^*_{\text{Lo}}$ after finite steps, and thus, it suffers from linear regret in another instance of M_{Hi} where $i^*_{\text{Lo}} \neq i^*_{\text{Hi}}$.

Moreover, Thm. 3.1 holds even if the learner has full information of M_{Lo}, where the setting degenerates to normal transfer RL since there is no need to explore M_{Lo}. This explains why similar assumptions to Assump. B are considered in previous transfer RL works [9, 10].

Theorem 3.1. *Under the violation of Assump. B, even regardless of the optimality of Alg^{Lo}, for each algorithm pair (Alg^{Lo}, Alg^{Hi}), it cannot simultaneously (1) achieve constant regret for the case when $M_{Lo} = M_{Hi}$ and (2) ensure sub-linear regret in all the other cases.*

$\Delta_{\min}$ in Tolerance Error is Inevitable Next, we show that, if M_{Hi} and M_{Lo} are Δ-close for some $\Delta \geq \frac{\Delta_{\min}}{2}$, in general we cannot expect to achieve constant regret on M_{Hi} by leveraging M_{Lo} without other loss. Similar to Thm. 3.1, the main idea is to construct different instances for M_{Hi} with different optimal arms and cannot be distinguished within finite number of trials.

Theorem 3.2. *[Transferable States are Restricted by $\Delta_{\min}$] Under Assump. B, regardless of the optimality of Alg^{Lo}, given arbitrary $\Delta_{\min}$ and arbitrary $\Delta \in [\frac{\Delta_{\min}}{2}, \Delta_{\min}]$, for each algorithm pair (Alg^{Lo}, Alg^{Hi}), it cannot simultaneously (1) achieve constant regret for the case when M_{Lo} and M_{Hi} with minimal gap $\Delta_{\min}$ are Δ-close, and (2) ensure sub-linear regret in all other cases.*

4 Robust Tiered MAB/RL with Single Source Task

In this section, we study Tiered MAB and Tiered RL when a single low-tier task M_{Lo} is available. The key challenge compared with [13] is that we do not have knowledge about whether M_{Lo} and M_{Hi} are similar or not so the pure exploitation will not work. Instead, the algorithm should be able to identify whether M_{Lo} and M_{Hi} are close enough to transfer by data collected so far, and balance between the exploration by itself and the exploitation from M_{Lo} at the same time.

To overcome the challenge, we identify a state-wise checking event, such that, under Assump. B, if s_h is transferable, the event is true almost all the time, and otherwise, every mistake will reduce the uncertainty so the chance the event holds is limited. By utilizing it, our algorithm can wisely switch between optimistic exploration and pessimistic exploitation and achieve robust transfer. In Sec. 4.1, we start with the MAB setting and illustrate the main idea, and in Sec. 4.2, we generalize our techniques to RL setting, and discuss how to overcome the challenges brought by state transition.

4.1 Robust Transfer in Tiered Multi-Armed Bandits

The algorithm is provided in Alg. 1. We choose Alg^{Lo} as UCB, and Alg^{Hi} as an exploitation-or-UCB style algorithm branched by a checking event in line 7, which is the key step to avoid negative transfer.

Algorithm 1: Robust Tiered MAB

1 **Initilize**: $\alpha > 2$; $N^1_{\text{Lo}}(i)$, $N^1_{\text{Hi}}(i)$, $\widehat{\mu}^1_{\text{Lo}}(i)$, $\widehat{\mu}^1_{\text{Hi}}(i) \leftarrow 0$, $\forall i \in \mathcal{A}$; $f(k) := 1 + 16A^2(k+1)^2$
2 Pull each arm at the beginning A iterations
3 **for** $k = A+1, A+2, ..., K$ **do**
4 $\quad \overline{\mu}^k_{\text{Lo}}(i) \leftarrow \widehat{\mu}^k_{\text{Lo}}(i) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo}}(i)}}, \quad \overline{\pi}^k_{\text{Lo}} \leftarrow \arg\max_i \overline{\mu}^k_{\text{Lo}}(i), \quad \pi^k_{\text{Lo}} \leftarrow \overline{\pi}^k_{\text{Lo}}.$
5 $\quad \underline{\mu}^k_{\text{Lo}}(i) \leftarrow \widehat{\mu}^k_{\text{Lo}}(i) - \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo}}(i)}}, \quad \underline{\pi}^k_{\text{Lo}} \leftarrow \arg\max_i \underline{\mu}^k_{\text{Lo}}(i).$
6 $\quad \overline{\mu}^k_{\text{Hi}}(i) \leftarrow \widehat{\mu}^k_{\text{Hi}}(i) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i)}}, \quad \overline{\pi}^k_{\text{Hi}} \leftarrow \arg\max_i \overline{\mu}^k_{\text{Hi}}(i).$
7 $\quad$ **if** $\underline{\mu}^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) \leq \overline{\mu}^k_{\text{Hi}}(\underline{\pi}^k_{\text{Lo}}) + \varepsilon$ and $N^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) > k/2$ **then** $\pi^k_{\text{Hi}} \leftarrow \underline{\pi}^k_{\text{Lo}}$ **else** $\pi^k_{\text{Hi}} \leftarrow \overline{\pi}^k_{\text{Hi}}$.
8 $\quad$ Interact $M_{\text{Hi}}/M_{\text{Lo}}$ by $\pi^k_{\text{Hi}}/\pi^k_{\text{Lo}}$; Update $N^{k+1}_{\text{Lo}}/N^{k+1}_{\text{Hi}}$ and empirical mean $\widehat{\mu}^{k+1}_{\text{Lo}}/\widehat{\mu}^{k+1}_{\text{Hi}}$.
9 **end**

Key Insights: Separation between Transferable and Non-Transferable Cases To understand our checking event, we consider the following two cases: (1) M_{Lo} and M_{Hi} are ε-close, and (2) $i^*_{\text{Hi}} \neq i^*_{\text{Lo}}$

(in the rest cases we have $i^*_{\text{Hi}} = i^*_{\text{Lo}}$ but $\mu_{\text{Lo}}(i^*_{\text{Lo}}) > \mu_{\text{Hi}}(i^*_{\text{Lo}}) + \varepsilon$, so exploiting from M_{Lo} is harmless). Recall Assump. B, in Case 1, we have $\mu_{\text{Lo}}(i^*_{\text{Lo}}) \le \mu_{\text{Hi}}(i^*_{\text{Lo}}) + \varepsilon$, while in Case 2, with an appropriate choice of ε (e.g. $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4} < \frac{\Delta_{\min}}{4}$), we have $\mu_{\text{Lo}}(i^*_{\text{Lo}}) \ge \mu_{\text{Hi}}(i^*_{\text{Hi}}) - \frac{\Delta_{\min}}{2} \ge \mu_{\text{Hi}}(i^*_{\text{Lo}}) + \varepsilon + \frac{\Delta_{\text{Hi}}(i^*_{\text{Lo}})}{2}$, which reveals the separation between two cases. As a result, if we can construct an uncertainty-based upper bound $\overline{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}})$ for $\mu_{\text{Hi}}(i^*_{\text{Lo}})$, we should expect the event $\mathcal{E} := \mu_{\text{Lo}}(i^*_{\text{Lo}}) < \overline{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}}) + \varepsilon$ almost always be true in Case 1, while in Case 2, everytime $\mathcal{E}$ occurs and Alg^{Hi} takes i^*_{Lo}, the "self-correction" is triggered: the uncertainty is reduced so the estimation $\overline{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}})$ gets closer to $\mu_{\text{Lo}}(i^*_{\text{Lo}})$, and because of the separation between $\mu_{\text{Lo}}(i^*_{\text{Lo}})$ and $\mu_{\text{Hi}}(i^*_{\text{Lo}})$, the number of times that $\mathcal{E}$ is true is limited. The remaining issue is that we do not know $\mu_{\text{Lo}}(i^*_{\text{Lo}})$ and i^*_{Lo}, and we approximate them with LCB value $\underline{\mu}^k_{\text{Lo}}(\cdot)$ and its greedy policy. The additional checking event $N^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) > k/2$ is used to increase the confidence that $\underline{\pi}^k_{\text{Lo}} = i^*_{\text{Lo}}$ once transfer, which also contributes to reducing the regret. Finally, to achieve constant regret, we use α to control the total failure rate to be $\sum_{k=1}^{+\infty} k^{-\Theta(\alpha)} = C$ for some constant C. We summarize the main result in Thm. 4.1 and defer the proof to Appx. C.

Theorem 4.1. *[Tiered MAB with Single Source Tasks] Under Assump. A, B and C, by running Alg. 1 with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4}$ and $\alpha > 2$, we always have $\textit{Regret}_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log K)$. Moreover, if M_{Hi} and M_{Lo} are $\frac{\widetilde{\Delta}_{\min}}{4}$-close, we have: $\textit{Regret}_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log \frac{A}{\Delta_{\min}})$.*

Comparison with [22, 13] As we can see, our algorithm can automatically achieve constant regret if tasks are similar while retaining near-optimal otherwise. Notably, even when $M_{\text{Hi}} = M_{\text{Lo}}$, our regret bound $\widetilde{O}(\sum_{\Delta_{\text{Hi}}(i)>0} \frac{1}{\Delta_{\text{Hi}}(i)})$ is strictly better than $\widetilde{O}(\sqrt{\frac{A}{\Delta_{\min}}} \sqrt{\sum_{\Delta_{\text{Hi}}(i)>0} \frac{1}{\Delta_{\text{Hi}}(i)}})$ in [22] and $\widetilde{O}(\sum_{\Delta_{\text{Hi}}(i)>0} (A-i)(\frac{1}{\Delta_{\text{Hi}}(i)} - \frac{\Delta_{\text{Hi}}(i)}{\Delta_{\text{Hi}}(i-1)^2}))$ in [13] under the same setting.

4.2 Robust Transfer in Tiered Tabular RL

In this section, we focus on RL setting. We provide the algorithm in Alg. 2, where we defer the details of **ModelLearning** function and the requirements for **Bonus** function to Appx. D.1. In the following, we first highlight our main result.

Theorem 4.2. *[Tiered RL with Single Source Tasks] Under Assump. A, B and C, Cond. D.1 for Alg^{Lo} and Cond. D.3 for **Bonus** function, by running Alg. 2 with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4(H+1)}$, $\alpha > 2$, an arbitrary $\lambda > 0$, we have*

$$\textit{Regret}_K(M_{Hi}) = O\Big(SH \sum_{h=1}^{H} \sum_{(s_h,a_h) \in \mathcal{S}_h \times \mathcal{A}_h \setminus \mathcal{C}^\lambda_h} (\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{Hi}(s_h,a_h)}) \log(SAHK)\Big).$$

Here the set $\mathcal{C}^\lambda_h \subset \mathcal{S}_h \times \mathcal{A}_h$ captures the benefitable state-action pairs to be introduced later. For simplicity, in Thm. 4.2 above, we omit all constant terms independent with K that may include λ^{-1}, $\Delta^{-1}_{\min}$ or $\log 1/d^*_h(\cdot)$. The complete version of Thm. 4.2 can be found in Thm. D.16. As we can see, comparing with pure online learning algorithms [23, 34, 7], in our setting, M_{Hi} only suffers non-constant regret on a subset of the state-action space. The SH factor may be further improved by choosing better **Bonus** functions than our Example D.4 given in Appx. D.1.

Different from the bandit setting, the state transition causes more challenges. In the following, we first explain the algorithm design to highlight how we overcome the difficulties, and then provide the analysis and proof sketch. Detailed proofs can be found in Appx. D.

Technical Challenges and Algorithm Design Similar to MAB setting, for M_{Lo} we choose an arbitrary near-optimal algorithm, and for M_{Hi} we set up a state-wise checking event $\underline{V}^k_{\text{Lo},h}(\cdot) \le \widetilde{Q}^k_{\text{Hi},h}(\cdot, \pi^k_{\text{Lo},h}) + \varepsilon$ in line 15 to determine whether to exploit from M_{Lo} or not. Here $\underline{V}^k_{\text{Lo},h}$ and $\widetilde{Q}^k_{\text{Hi},h}$ serve as lower and upper bounds for V^*_{Lo} and Q^*_{Hi}, and $\underline{V}^k_{\text{Lo},h}$ and $\underline{\pi}^k_{\text{Lo},h}$ are constructed by Pessimistic Value Iteration [16], which can be shown to converge to $V^*_{\text{Lo},h}$ and $\pi^*_{\text{Lo},h}$, respectively. Similar to [13], for the choice of Alg^{Lo} and the bonus term used to construct lower/upper confidence estimation, we consider general algorithm framework under Cond. D.1 and Cond. D.3 in Appx. D.1.

To overcome challenges resulting from state transition, we make two major modifications when moving from MAB to RL setting. First of all, because of the constraint on the optimality of Alg^{Lo},

Algorithm 2: Robust Tiered RL

Input: Ratio $\lambda \in (0,1)$; $\alpha > 2$; Auxiliary functions **Bonus** and **ModelLearning**; Sequence of confidence level $(\delta_k)_{k\geq 1}$ with $\delta_k = 1/SAHk^\alpha$; $\varepsilon := \widetilde{\Delta}_{\min}/4(H+1)$ for some $\widetilde{\Delta}_{\min} \leq \Delta_{\min}$

Initialize: $D^0_{\text{Lo}}/D^0_{\text{Hi}} \leftarrow \{\}$; $\forall k$, $\underline{V}^k_{(\cdot),H+1}, \underline{Q}^k_{(\cdot),H+1}, \widetilde{V}^k_{\text{Hi},H+1}, \widetilde{Q}^k_{\text{Hi},H+1} \leftarrow 0$.

for $k = 1, 2, \ldots$ **do**

  $\pi_{\text{Lo}} \leftarrow \text{Alg}^{\text{Lo}}(D^{k-1}_{\text{Lo}})$;

  $\{\widehat{\mathbb{P}}^k_{\text{Lo},h}\}_{h=1}^H \leftarrow$ **ModelLearning**(D^{k-1}_{Lo}), $\{b^k_{\text{Lo},h}\}_{h=1}^H \leftarrow$ **Bonus**$(D^{k-1}_{\text{Lo}}, \delta_k)$.

  for $h = H, H-1 \ldots, 1$ **do**

    $\underline{Q}^k_{\text{Lo},h}(\cdot,\cdot) \leftarrow \max\{0, r_{\text{Lo},h}(\cdot,\cdot) + \widehat{\mathbb{P}}^k_{\text{Lo},h}\underline{V}^k_{\text{Lo},h+1}(\cdot,\cdot) - b^k_{\text{Lo},h}(\cdot,\cdot)\}$.

    $\underline{V}^k_{\text{Lo},h}(\cdot) = \max_a \underline{Q}^k_{\text{Lo},h}(\cdot,a)$, $\quad \underline{\pi}^k_{\text{Lo},h}(\cdot) \leftarrow \arg\max_a \underline{Q}^k_{\text{Lo},h}(\cdot,a)$.

  end

  $\{\widehat{\mathbb{P}}^k_{\text{Hi},h}\}_{h=1}^H \leftarrow$ **ModelLearning**(D^{k-1}_{Hi}); $\quad \{b^k_{\text{Hi},h}\}_{h=1}^H \leftarrow$ **Bonus**$(D^{k-1}_{\text{Hi}}, \delta_k)$.

  for $h = H, H-1 \ldots, 1$ **do**

    $\underline{Q}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(\cdot,\cdot) \leftarrow \max\{0, r_{\text{Hi},h}(\cdot,\cdot) + \widehat{\mathbb{P}}^k_{\text{Hi},h}\underline{V}^k_{\text{Hi},h+1}(\cdot,\cdot) - b^k_{\text{Hi},h}(\cdot,\cdot)\}$

    $\widetilde{Q}^k_{\text{Hi},h}(\cdot,\cdot) \leftarrow \min\{H, r_{\text{Hi},h}(\cdot,\cdot) + \widehat{\mathbb{P}}^k_{\text{Hi},h}\widetilde{V}^k_{\text{Hi},h+1}(\cdot,\cdot) + b^k_{\text{Hi},h}(\cdot,\cdot)\}$.

    for $s_h \in \mathcal{S}_h$ **do**

      if $\underline{V}^k_{Lo,h}(s_h) \leq \widetilde{Q}^k_{Hi,h}(s_h, \underline{\pi}^k_{Lo,h}) + \varepsilon$ *and* $\max_a N^k_{Lo,h}(s_h,a) > \frac{\lambda}{3}k$ **then**

        $\pi^k_{\text{Hi}}(s_h) \leftarrow \arg\max_a N^k_{\text{Lo},h}(s_h,a)$. // "Trust and Exploit" Branch

      end

      else $\pi^k_{\text{Hi}}(\cdot) \leftarrow \arg\max_a \widetilde{Q}^k_{\text{Hi},h}(\cdot,a)$. // "Explore by itself" Branch ;

      $\widetilde{V}^k_{\text{Hi},h}(s_h) \leftarrow \min\{H, \widetilde{Q}^k_{\text{Hi},h}(s_h,\pi^k_{\text{Hi}}) + \frac{1}{H}(\widetilde{Q}^k_{\text{Hi},h}(s_h,\pi^k_{\text{Hi}}) - \underline{Q}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h,\pi^k_{\text{Hi}}))\}$

      $\underline{V}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h) = \underline{Q}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h,\pi^k_{\text{Hi}})$.

    end

  end

  Deploy $\pi_{\text{Lo}}/\pi_{\text{Hi}}$ to $M_{\text{Lo}}/M_{\text{Hi}}$ and get $\tau^k_{\text{Lo}}/\tau^k_{\text{Hi}}$; and update $D^k_{\text{Lo}}, D^k_{\text{Hi}}$.

end

we cannot expect M_{Lo} to provide useful information on those (s_h, a_h) with $d^*_{\text{Lo}}(s_h,a_h) = 0$ since they will not be explored sufficiently. Therefore, in the checking event in line 15, we include $\max_a N^k_{\text{Lo},h}(s_h,a) > \Theta(\lambda k)$ as a criterion, where λ is a hyper-parameter chosen and input to the algorithm. Intuitively, for all s_h, a_h, we should expect $N^k_{\text{Lo},h}(s_h,a_h) \approx \widetilde{O}(d^*_{\text{Lo}}(s_h,a_h)\cdot k)$ when k is large enough. Therefore, by comparing $N^k_{\text{Lo},h}$ with λk, we can filter out those s_h with $d^*_{\text{Lo}}(s_h) < \lambda$ to avoid harm from inaccurate estimation.

Secondly and more importantly, different from MAB setting, besides the error occurred at a particular step, we also need to handle the error accumulated during the back-propagation process of value iteration. In our case, this is reflected by the loss of overestimation when we incorporate selective exploitation into the optimism-based exploration framework. To see this, suppose at some s_h, we have an overestimation on optimal value $Q^*_{\text{Hi},h}$ denoted as $\widetilde{Q}^k_{\text{Hi},h}$. When the checking criterion is satisfied, if we mimic the MAB setting, i.e., assign $\pi^k_{\text{Lo},h}$ to $\pi^k_{\text{Hi},h}$ and update value by $\widetilde{V}^k_{\text{Hi},h}(s_h) \leftarrow \widetilde{Q}^k_{\text{Hi},h}(s_h,\pi^k_{\text{Hi},h})$, when $\pi^*_{\text{Lo}}(s_h) \neq \pi^*_{\text{Hi},h}(s_h)$, $\widetilde{V}^k_{\text{Hi},h}(s_h)$ is no longer guaranteed to be an overestimation for $V^*_{\text{Hi},h}(s_h)$. As $\widetilde{V}^k_{\text{Hi},h}(s_h)$ involves in back-propagation, it will pull down the estimation value for its ancestor states, thus reducing the chance to visit s_h and slowing down the "self-correction process" which works well in MAB setting.

The key insight to overcome such difficulty is that, if the checking event holds yet $\pi^*_{\text{Lo}}(s_h) \neq \pi^*_{\text{Hi},h}(s_h)$, the gap between $\widetilde{Q}^k_{\text{Hi},h}(s_h,\pi^*_{\text{Lo}})$ and $Q^*_{\text{Hi},h}(s_h,\pi^*_{\text{Lo}})$ should not be small, and we can show that $\widetilde{Q}^k_{\text{Hi},h}(s_h,\underline{\pi}^k_{\text{Lo}}) \approx \widetilde{Q}^k_{\text{Hi},h}(s_h,\pi^*_{\text{Lo}}) \geq Q^*_{\text{Hi},h}(s_h,\pi^*_{\text{Lo}}) + \Theta(\frac{H}{H+1}\Delta_{\text{Hi}}(s_h,\pi^*_{\text{Lo}}))$ with the choice of

$\varepsilon = O(\widetilde{\Delta}_{\min}/H)$. Therefore, revising $\widetilde{Q}^k_{\text{Hi},h}(s_h, \underline{\pi}^k_{\text{Lo}})$ by adding $1/H$ of the gap $\widetilde{Q}^k_{\text{Hi},h}(s_h, \underline{\pi}^k_{\text{Lo}}) - Q^*_{\text{Hi},h}(s_h, \pi^*_{\text{Lo}})$ (line 19) is enough to guarantee the overestimation. Lastly, since $Q^*_{\text{Hi},h}(s_h, \pi^*_{\text{Lo}})$ in unknown, we construct an underestimation $\underline{Q}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h, \pi^*_{\text{Lo}})$ and use it instead. As a result, we have the following theorem, where the clip function is defined by $\text{Clip}[x|w] := x \cdot \mathbb{I}[x \geq w]$.

Theorem 4.3. *There exists $k_{ost} = Poly(S, A, H, \lambda^{-1}, \Delta^{-1}_{\min})$, such that, for all $k \geq k_{ost}$, on some event $\mathcal{E}_k$ with $\mathbb{P}(\mathcal{E}_k) \leq 3\delta_k$, we have $Q^*_{Hi,h}(s_h, a_h) \leq \widetilde{Q}^k_{Hi,h}(s_h, a_h)$, $V^*_{Hi,h}(s_h) \leq \widetilde{V}^k_{Hi,h}(s_h), \forall h \in [H]$, $s_h \in \mathcal{S}_h$, $a_h \in \mathcal{A}_h$ and*

$$V^*_{Hi,1}(s_1) - V^{\pi^k_{Hi}}_{Hi,1}(s_1) \leq 2e\mathbb{E}_{\pi^k_{Hi}}\left[\sum_{h=1}^{H} Clip\left[\min\{H, 4b^k_{Hi,h}(s_h, a_h)\}|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{Hi}(s_h, a_h)}{4e}\right]\right]. \quad (1)$$

Benefits of Knowledge Transfer We first take a look at $k \geq k_{ost}$. As implied from Eq. (1), we can upper bound the regret on each s_h, a_h by summing over the RHS of Eq. (1). Note that by Cond. D.3, $b^k_{\text{Hi},h}(s_h, a_h) = O(\frac{\text{Poly}(SAH)\log k}{\sqrt{N^k_{\text{Hi},h}(s_h, a_h)}})$ and $\mathbb{E}[N^k_{\text{Hi},h}(s_h, a_h)] = \sum_{k'=1}^{k-1} d^{\pi^k_{\text{Hi}}}(s_h, a_h)$, we can establish the near-optimal regret bound with similar techniques in [23] regardless of the similarity between M_{Hi} and M_{Lo}. Moreover, because of the knowledge transfer from M_{Lo}, we can achieve better regret bounds for M_{Hi}. In the following, we characterize three subclasses of state-action pairs, on which Alg^{Hi} only suffers constant regret. *First of all*, for those $s_h \in \mathcal{Z}^\lambda_h$, we can expect the checking event almost always hold for arbitrary k. Hence, when k is large enough, $\pi^k_{\text{Hi},h}(s_h) = \underline{\pi}^k_{\text{Lo},h}(s_h) \approx \pi^*_{\text{Hi},h}(s_h)$, implying Alg^{Hi} will almost never take sub-optimal actions at s_h since then. We denote this first subclass as $\mathcal{C}^{1,\lambda}_h := \{(s_h, a_h)|s_h \in \mathcal{Z}^\lambda_h, a_h \neq \pi^*_{\text{Hi},h}(s_h)\}$. *Secondly*, note that, given a state s_h, if all possible trajectories starting from s_1 to s_h have overlap with $\mathcal{C}^{1,\lambda}_{h'}$ for some $h' \in [h-1]$, when k is large enough, π^k_{Hi} will almost have no chance to reach s_h and will not suffer the regret at s_h. For convenience, we define function $\text{Block}(\{\mathcal{C}^{1,\lambda}_{h'}\}^{h-1}_{h'=1}, s_h)$ which takes `True` for those states described above, and takes `False` for the others. Then, we define the second subclass by $\mathcal{C}^{2,\lambda}_h := \{(s_h, a_h)|\text{Block}(\{\mathcal{C}^{1,\lambda}_{h'}\}^{h-1}_{h'=1}, s_h) = \texttt{True}, s_h \notin \mathcal{Z}^\lambda_h, a_h \in \mathcal{A}_h\}$. *Finally*, for those s_h, a_h with $d^*_{\text{Hi}}(s_h, a_h) > 0$, we can show $N^k_{\text{Hi},h}(s_h, a_h) \approx d^*_{\text{Hi}}(s_h, a_h)k$. Therefore, $b^k_{\text{Hi},h}(s_h, a_h) \propto \log k/\sqrt{N^k_{\text{Hi},h}(s_h, a_h)}$ in Eq. (1) will decay and the clipping operator will take effect, which leads to constant regret. This third subclass is denoted by $\mathcal{C}^*_h := \{(s_h, a_h)|d^*_{\text{Hi}}(s_h, a_h) > 0\}$. Based on the above discussion, we define $\mathcal{C}^\lambda_h := \mathcal{C}^{\lambda,1}_h \cup \mathcal{C}^{\lambda,2}_h \cup \mathcal{C}^*_h$ to be the benefitable states set in Thm. 4.2.

For $k \leq k_{ost}$, for the lack of overestimation, we simply use H to upper bound the value gap $V^* - V^{\pi^k_{\text{Hi}}}$. This results in a $\text{Poly}(S, A, H, \lambda^{-1}, \Delta^{-1}_{\min})$ burn-in term, which was omitted in Thm. 4.2 since it is independent with K. Besides, by the definition of k_{ost} in Thm. 4.3, we can see the trade-off of choosing λ: a smaller λ can enlarge $\mathcal{C}^\lambda_h$ so we have constant regret on more state-action pairs, while it also results in the delay of overestimation by the larger k_{ost}.

Constant Regret in the Entire MDP We may expect constant regret in the entire M_{Hi} in some special cases. Note that, if $\forall h \in [H], \forall s_h$ with $d^*_{\text{Hi}}(s_h) > 0$, $s_h \in \mathcal{Z}^\lambda_h$, we have $\mathcal{C}^\lambda_h = \mathcal{S}_h \times \mathcal{A}_h$, $\text{Regret}_K(M_{\text{Hi}})$ will be independent w.r.t. K. From this perspective, if λ is chosen appropriately, e.g. $\lambda \leq \min_{s_h} d^*_{\text{Hi}}(s_h)$, we can recover the constant regret under the setting $M_{\text{Lo}} = M_{\text{Hi}}$ in [13].

Choice of λ In this paper, we do not treat λ as a parameter to optimize. In practice, without prior knowledge about $\max_{s_h} d^*_{\text{Lo}}(s_h)$, one may choose $\lambda = O(1/S)$ to ensure some chance that transferable states exist, since there exists at least some states satisfying $d^*_{\text{Lo}}(s_h) \geq 1/S$.

5 Robust Tiered MAB/RL with Multiple Low-Tier Tasks

Now, we focus on the case when a source task set $\mathcal{M}_{\text{Lo}} := \{M_{\text{Lo},w}\}^W_{w=1}$ is available (see Frw. 5 in Appx. A). Our objective is to achieve benefits on those states s_h as long as there exists some task $w \in [W]$ such that $M_{\text{Lo},w}$ and M_{Hi} are close on s_h, while retaining near-optimal regret in other cases under Assump. B. The key challenge comparing with single task case is that, Alg^{Hi} has to identify for each state which task in $\mathcal{M}_{\text{Lo}}$ is the appropriate one to leverage. The main novelty and contribution

in this section is a task selection mechanism we call *"Trust till Failure"*, which can automatically adapt to the similar task if it exists. We first highlight the main results for MAB and RL setting.

Theorem 5.1. *[Tiered MAB with Multiple Source Tasks] Under Assump. A, B, and C, by running Alg. 3 with $\mathcal{M}_{Lo} = \{M_{Lo,w}\}_{w=1}^{W}$ and M_{Lo}, with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4}$ and $\alpha > 2$, we always have: $Regret_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log(WK))$. Moreover, if at least one task in $\mathcal{M}_{Lo}$ is $\frac{\widetilde{\Delta}_{\min}}{4}$-close to M_{Hi}, we further have: $Regret_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log \frac{AW}{\Delta_{\min}})$.*

Theorem 5.2. *[Tiered RL with Multiple Source Tasks] Under Assump. A, B, C, and Cond. D.3, F.4, by running Alg. 7 in Appx. F.1 with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4(H+1)}$, $\alpha > 2$ and any $\lambda > 0$, we have*

$$Regret_K(M_{Hi}) = O\Big(SH \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}_h^{\lambda,[W]}} (\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{Hi}(s_h,a_h)}) \log(SAHWK)\Big).$$

Algorithm 3: Robust Tiered MAB with Multiple Source Tasks

1 **Initilize**: $\alpha > 2$; $N_{\text{Lo}}^1(i)$, $N_{\text{Hi}}^1(i)$, $\widehat{\mu}_{\text{Lo}}^1(i)$, $\widehat{\mu}_{\text{Hi}}^1(i) \leftarrow 0$, $\forall i \in \mathcal{A}$; $f(k) := 1 + 16TA^2(k+1)^2$
2 Pull each arm at the beginning A iterations. Set $w^A \leftarrow$ `Null`.
3 **for** $k = A+1, 2, ..., K$ **do**
4 **for** $w = 1, 2..., W$ **do**
5 $\overline{\mu}_{\text{Lo},w}^k(i) \leftarrow \widehat{\mu}_{\text{Lo},w}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Lo},w}(i)}}$, $\overline{\pi}_{\text{Lo},w}^k \leftarrow \arg\max_i \overline{\mu}_{\text{Lo},w}^k(i)$, $\pi_{\text{Lo},w}^k \leftarrow \overline{\pi}_{\text{Lo},w}^k$.
6 $\underline{\mu}_{\text{Lo},w}^k(i) \leftarrow \widehat{\mu}_{\text{Lo},w}^k(i) - \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Lo},w}(i)}}$, $\underline{\pi}_{\text{Lo},w}^k \leftarrow \arg\max_i \underline{\mu}_{\text{Lo},w}^k(i)$.
7 **end**
8 $\overline{\mu}_{\text{Hi}}^k(i) \leftarrow \widehat{\mu}_{\text{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i)}}$, $\overline{\pi}_{\text{Hi}}^k \leftarrow \arg\max_i \overline{\mu}_{\text{Hi}}^k(i)$.
9 $\mathcal{I}^k \leftarrow \{w \in [W] | \underline{\mu}_{\text{Lo},w}^k(\underline{\pi}_{\text{Lo},w}^k) \le \overline{\mu}_{\text{Hi}}^k(\underline{\pi}_{\text{Lo},w}^k) + \varepsilon \text{ and } N_{\text{Lo},w}^k(\underline{\pi}_{\text{Lo},w}^k) > k/2\}$
10 **if** $\mathcal{I}^k = \emptyset$ **then** $w^k \leftarrow$ `Null`, $\pi_{\text{Hi}}^k \leftarrow \overline{\pi}_{\text{Hi}}^k$;
11 **else**
12 **if** $w^{k-1} \neq$ `Null` *and* $w^{k-1} \in \mathcal{I}^k$ **then** $w^k \leftarrow w^{k-1}$, $\pi_{\text{Hi}}^k \leftarrow \underline{\pi}_{\text{Lo},w^k}^k$;
13 **else if** $w^{k-1} \neq$ `Null` *and* $\exists w \in \mathcal{I}^k$ *such that* $\pi_{Hi}^{k-1} = \arg\max_i N_{Lo,w}^k(i)$ **then**
14 $w^k \leftarrow w$, $\pi_{\text{Hi}}^k \leftarrow \underline{\pi}_{\text{Lo},w}^k$
15 **end**
16 **else** $w^k \sim \text{Unif}(\mathcal{I}^k)$, $\pi_{\text{Hi}}^k \leftarrow \underline{\pi}_{\text{Lo},w^k}^k$;
17 **end**
18 Interact with $M_{\text{Hi}}/\{M_{\text{Lo},}\}_{w=1}^W$ by $\pi_{\text{Hi}}^k/\{\pi_{\text{Lo},w}^k\}_{w=1}^W$;
19 Update $\{N_{\text{Lo},w}^{k+1}\}_{w=1}^W, N_{\text{Hi}}^{k+1}$ and empirical mean $\{\widehat{\mu}_{\text{Lo},w}^{k+1}\}_{w=1}^W, \widehat{\mu}_{\text{Hi}}^{k+1}$ for each arm.
20 **end**

For the lack of space, in the following, we only analyze the bandit setting to explain the key idea of our task selection strategy. For the RL setting, we defer to Appx. F the algorithm Alg. 7, detailed version of Thm. 5.2 (Thm. D.16), defintion of transferable set $\mathcal{C}_h^{\lambda,[W]}$ (Def. F.2), and technical details.

Algorithm Design and Proof Sketch for Bandit Setting The algorithm is provided in Alg. 3. Comparing with Alg. 1 in single task setting, the main difference is the task selection strategy from line 9 to line 17. We first examine each source task with a checking event similar to single task setting, and collect those feasible tasks passing the test to $\mathcal{I}^k$. Intuitively, for those M_{Lo,w^*} close to M_{Lo}, we expect $M_{\text{Lo},w} \in \mathcal{I}^k$ holds almostly for arbitrary $k > 0$, while for the other $M_{\text{Lo},w'}$, if it takes the position of w^k, following $M_{\text{Lo},w'}$ will reduce the uncertainty and it will be ruled out from $\mathcal{I}^k$, eventually. So we expect w^k can "escape" from dissimilar source tasks but be absorbed to the similar task if exists. Therefore, if $\mathcal{I}^k$ is empty, Alg^{Hi} will do exploration by itself. Otherwise, we choose one from $\mathcal{I}^k$ to transfer the action until it fails on the checking event. However, for any ε the algorithm chosen, those "marginally similar" source tasks (denoted as $M_{\text{Lo},\widetilde{w}}$), which are ε'-close to M_{Hi} for some ε' only slightly larger than ε, may cause some trouble. Because the checking event

will finally eliminate $M_{\text{Lo},\widetilde{w}}$ since they are not ε-close, but it may occupy the position w^k for a long time before elimination, especially when ε' is extremely close to ε. After eliminating $M_{\text{Lo},\widetilde{w}}$, Alg^{Hi} needs to re-select one from $\mathcal{I}^k$. Now since other sub-optimal arms in M_{Hi} haven't been chosen for a long time and the confidence level $\delta_k = O(1/k^\alpha)$ is decreasing, $\mathcal{I}^k$ will include those dissimilar tasks again, which causes difficulty to identify the true similar task. To solve this issue, once the previous trusted task fails, we give priority to the task recommending the same action as the previous one (line 14). As a result, since $M_{\text{Lo},\widetilde{w}}$ and M_{Lo,w^*} share the optimal action, after the elimination of $M_{\text{Lo},\widetilde{w}}$, we can expect w^k to only switch among those tasks $M_{\text{Lo},w}$ with $\pi^*_{\text{Lo},w} = \pi^*_{\text{Hi}}$. We highlight this technical novelty to Lem. 5.3 below, and defer all the proofs to Appx. E.

Lemma 5.3. *[Absorbing to Similar Task] Under Assump. A, B and C, there exists a constant c^*, s.t., if there exists at least one $w^* \in [W]$ such that M_{Lo,w^*} is $\frac{\widetilde{\Delta}_{\min}}{4}$-close to M_{Hi}, by running Alg. 3 with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4}$ and $\alpha > 2$, for any $k \geq k^* := c^* \frac{\alpha A}{\Delta^2_{\min}} \log \frac{\alpha AW}{\Delta_{\min}}$, we have $\Pr(\pi^k_{Hi} \neq i^*_{Hi}) = O(\frac{A}{k^{2\alpha-2}})$.*

6 Experiments

In this section, we evaluate our most representative algorithm, Alg. 7, in multiple source tasks setting.

Experiments Setting[3] We set $S = A = 3$ and $H = 5$. The details for construction of source and target tasks are defered to Appx. G. We adapt StrongEuler in [23] as online learning algorithm to solve source tasks, and use the bonus function in [23] as the bonus function in our Alg. 7. We evaluate our algorithm when $W = 0, 1, 2, 5$, where $W = 0$ means the high-tier task M_{Hi} is simply solved by normal online learning method (StrongEuler) without any parallel knowledge transfer. We choose $\lambda = 0.3 \approx 1/S$ in Alg. 7, and in the MDP instance we test, across all $S \cdot H = 15$ states, for $W = 1, 2, 5$, the number of transferable states would be 6, 9 and 13, respectively.

We choose iteration number $K = 1e7$, where we start the transfer since $k = 5e5$ to avoid large "burn-in" terms. As we can see, after the transfer starts, the regret in target task will suddenly increase for a while, because the target task has to make some mistakes and learn from it as a result of the model uncertainty. However, because of our algorithm design, the negative transfer will terminate after a very short period. As predicted by our theory, by adding more and more source tasks which can introduce new transferable states, the target task will suffer less and less regret.

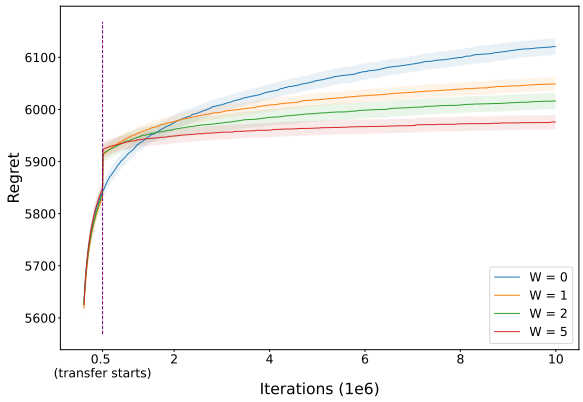

Figure 1: **Regret in the Target Task given Multiple Source Tasks** We report the result when W source tasks are available with $W = 0, 1, 2, 5$. The shadows indicate 96% confidence interval.

7 Conclusion and Future Work

In this paper, we study how to do robust parallel transfer RL when single or multiple source tasks are avilable, without knowledge on models' similarity. The possible future directions include relaxing assumptions, better strategies to leveraging multiple source tasks, and identifying mild structural assumptions allowing for more aggressive transfer, and we defer to Appx. A.4 for more details.

[3] Code is available at `https://github.com/jiaweihhuang/Robust-Tiered-RL`

Acknowledgments and Disclosure of Funding

The authors would like to thank Andreas Krause for valuable discussion. The work is supported by ETH research grant and Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343.

References

[1] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

[2] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[3] Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning, 2013.

[4] Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. *arXiv preprint arXiv:2203.13935*, 2022.

[5] HeeSun Choi, Cindy Crump, Christian Duriez, Asher Elmquist, Gregory Hager, David Han, Frank Hearl, Jessica Hodgins, Abhinandan Jain, Frederick Leve, et al. On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. *Proceedings of the National Academy of Sciences*, 118(1):e1907856118, 2021.

[6] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning, 2017.

[7] Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=CU8qQMhB3dh`.

[8] Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, Jan Peters, et al. Sharing knowledge in multi-task deep reinforcement learning. In *8th International Conference on Learning Representations,{ICLR} 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, pages 1–11. OpenReview. net, 2020.

[9] Noah Golowich and Ankur Moitra. Can q-learning be improved with advice? In *Conference on Learning Theory*, pages 4548–4619. PMLR, 2022.

[10] Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham M. Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=D-X3kH-BkpN`.

[11] Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pages 4349–4358. PMLR, 2021.

[12] Jiawei Huang, Jinglin Chen, Li Zhao, Tao Qin, Nan Jiang, and Tie-Yan Liu. Towards deployment-efficient reinforcement learning: Lower bound and optimality. *arXiv preprint arXiv:2202.06450*, 2022.

[13] Jiawei Huang, Li Zhao, Tao Qin, Wei Chen, Nan Jiang, and Tie-Yan Liu. Tiered reinforcement learning: Pessimism in the face of uncertainty and constant regret. *arXiv preprint arXiv:2205.12418*, 2022.

[14] Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

[15] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

[16] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[17] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[18] Alessandro Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.

[19] Yongyuan Liang and Bangwei Li. Parallel knowledge transfer in multi-agent reinforcement learning, 2020.

[20] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

[21] Timothy A Mann and Yoonsuck Choe. Directed exploration in reinforcement learning with transferred knowledge. In *European Workshop on Reinforcement Learning*, pages 59–76. PMLR, 2013.

[22] Chloé Rouyer and Yevgeny Seldin. Tsallis-inf for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, pages 3227–3249. PMLR, 2020.

[23] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning*, pages 9767–9779. PMLR, 2021.

[25] Adam Taylor, Ivana Dusparic, Maxime Guériau, and Siobhán Clarke. Parallel transfer learning in multi-agent systems: What, when and how to transfer? In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[26] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

[27] Volodymyr Tkachuk, Sriram Ganapathi Subramanian, and Matthew E. Taylor. The effect of q-function reuse on the total regret of tabular, model-free, reinforcement learning, 2021.

[28] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

[29] Nelson Vithayathil Varghese and Qusay H Mahmoud. A survey of multi-task deep reinforcement learning. *Electronics*, 9(9):1363, 2020.

[30] Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. *arXiv preprint arXiv:2211.04974*, 2022.

[31] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*, pages 1015–1022, 2007.

[32] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[33] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[34] Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472. PMLR, 2021.

[35] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

[36] Chicheng Zhang and Zhi Wang. Provably efficient multi-task reinforcement learning with model transfer. *Advances in Neural Information Processing Systems*, 34:19771–19783, 2021.

[37] Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020.

Contents

A Extended Introduction

A.1 Tiered-RL Framework

Algorithm 4: The Tiered RL Framework with Single Low-Tier Task

Initialize $D_{\text{Lo}}^1, D_{\text{Hi}}^1 \leftarrow \{\}$.
for $k = 1, 2, ..., K$ **do**
  $\pi_{\text{Lo}}^k \leftarrow \text{Alg}^{\text{Lo}}(D_{\text{Lo}}^k)$; π_{Lo}^k interacts with M_{Lo}, and collect data τ_{Lo}^k; $D_{\text{Lo}}^{k+1} = D_{\text{Lo}}^k \cup \{\tau_{\text{Lo}}^k\}$
  $\pi_{\text{Hi}}^k \leftarrow \text{Alg}^{\text{Hi}}(D_{\text{Hi}}^k)$; π_{Hi}^k interacts with M_{Hi}, and collect data τ_{Hi}^k; $D_{\text{Hi}}^{k+1} = D_{\text{Hi}}^k \cup \{\tau_{\text{Hi}}^k\}$.
end

Algorithm 5: The Tiered RL Framework with Multiple Low-Tier Tasks

Initialize $D_{\text{Lo}}^1 \leftarrow \{\}$; $D_{\text{Hi},w}^1 \leftarrow \{\}$, $\forall w \in [W]$.
for $k = 1, 2, ..., K$ **do**
  for $w \in [W]$ **do**
    $\pi_{\text{Lo},w}^k \leftarrow \text{Alg}^{\text{Lo}}(D_{\text{Hi},w}^k)$; $\pi_{\text{Lo},w}^k$ interacts with $M_{\text{Lo},w}$, and collect data $\tau_{\text{Lo},w}^k$.
    $D_{\text{Lo},w}^{k+1} = D_{\text{Hi}}^k \cup \{\tau_{\text{Lo},w}^k\}$.
  end
  $\pi_{\text{Hi}}^k \leftarrow \text{Alg}^{\text{Hi}}(D_{\text{Hi}}^k)$; π_{Hi}^k interacts with M_{Hi}, and collect data τ_{Hi}^k; $D_{\text{Hi}}^{k+1} = D_{\text{Hi}}^k \cup \{\tau_{\text{Hi}}^k\}$.
end

A.2 Other Related Works

Online and Offline RL In normal online RL/MAB setting, the learner targets at actively explore the environment while balancing the trade-off between the exploration and exploitation [17, 1, 23, 7, 2, 15] Differently, motivated by many real world scenarios where historical data are available, offline RL considers how to do pure exploitation given the pre-collected dataset without additional exploration and new information collection, and theoretical works mainly focus on methods for sufficiently exploitation [35, 28, 32, 16, 20, 14, 4].

Recently, there is also a line of work studying the settings lying between pure online and offline RL, such as hybrid setting where offline data is available for online exploration [33, 30], and efficient batched exploration with limited policy deployments [12]. Tiered RL framework can be regarded as another approach to bridging the online and offline setting, where we do online learning in the high-tier task with a gradually updated dataset from low-tier task for reference.

Detailed Comparison with Previous Transfer RL Paper about Assumptions The most recent works in transfer RL are [9] and [10]. In general we are not comparable because of our different settings, but we can observe some similarity of our assumptions and the way to capture the transferable states.

[9] considered the case where a predicted Q-function $\{\widetilde{Q}_h(\cdot,\cdot)\}_{h=1}^H$ is provided for each state action pair, which can be regarded as $\{Q_{\text{Lo},h}^*(\cdot,\cdot)\}_{h=1}^H$ in our setting. The key assumption in their paper is the "approximate distillation" condition (Def. 3.1 in [9]), which assumed that for each s_h, there exists $a_h \in \mathcal{A}_h$ such that $\Delta(s_h, a_h) + \max\{0, Q_h^*(s_h, a_h) - \widetilde{Q}_h(s_h, a_h)\} \le \varepsilon$. However, according to Eq. (2) in their Thm. 3.1, there is an $\varepsilon' TH = 4\varepsilon(H+1)TH$ term in the regret of their algorithm (where T is the episode number). Therefore, in order to achieve regret sub-linear to T, they need $\varepsilon = T^{-\alpha}$ for some $\alpha > 0$. As $T \to +\infty$, we have $\varepsilon \to 0$, then their "approximate distillation" condition will reduce to $V_{\text{Lo},h}^*(s_h) \ge V_{\text{Hi},h}^*(s_h)$, which is a stronger version of our OVD condition in Assump. B.

As for [10], the authors assumed that there is a value function and parameter β such that $\beta\widetilde{V}_h(s_h)$ (i.e. $V_{\text{Lo}}^*(s_h)$ in our setting) forms an overestimation for $V_h^*(s_h)$ in target task, which is also similar to our Assump. B. Besides, although they didn't make it explicitly, to achieve provable benefits, they also require such a overestimation $\beta\widetilde{V}_h(s_h)$ should not deviate too far away from the true value $V_h^*(s_h)$. To see this, in Sec. 5.1 of [10], they use $V_h^*(s) \ge \Delta + \widetilde{Q}_h^u(s_h, a_h)$ to characterize state-action pairs

with regret reduction, where $\widetilde{Q}^u_h(s_h, a_h) := \mathbb{E}_{s'}[r(s_h, a_h) + \beta \widetilde{V}_{h+1}(s')]$, and $\beta\widetilde{V}_{h+1}$ has to stay close to $V^*_h(s)$ for such condition to be realizable.

Finally, both [9, 10] assumed the condition holds for each state action pair, while our Assump. B can only require the overestimation on states reachable by optimal policy.

A.3 Examples for Assump. B

Example A.1 (Identical Model [13])**.** For arbitrary $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$, $r_{\text{Lo},h}(s_h, a_h) = r_{\text{Hi},h}(s_h, a_h)$, $\mathbb{P}_{\text{Hi},h}(\cdot|s_h, a_h) = \mathbb{P}_{\text{Lo},h}(\cdot|s_h, a_h)$.

Example A.2 (Small Model Error)**.** For arbitrary $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$, $|r_{\text{Lo},h}(s_h, a_h) - r_{\text{Hi},h}(s_h, a_h)| \le \frac{\Delta_{\min}}{4H(H+1)}$, $\|\mathbb{P}_{\text{Hi},h}(\cdot|s_h, a_h) - \mathbb{P}_{\text{Lo},h}(\cdot|s_h, a_h)\|_1 \le \frac{\Delta_{\min}}{4H^2(H+1)}$.

Example A.3 (Known Model Difference)**.** Suppose there exists known quantities ξ_r and $\xi_{\mathbb{P}}$ such that, for arbitrary $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$:

$$|r_{\text{Lo},h}(s_h, a_h) - r_{\text{Hi},h}(s_h, a_h)| \le \xi_r, \quad \|\mathbb{P}_{\text{Hi},h}(\cdot|s_h, a_h) - \mathbb{P}_{\text{Lo},h}(\cdot|s_h, a_h)\|_1 \le \xi_{\mathbb{P}}.$$

Then, one can revise the reward function of M_{Lo} to r'_{Lo} defined by $r'_{\text{Lo},h}(s_h, a_h) = r_{\text{Lo},h}(s_h, a_h) + \xi_r + (H-h)\xi_{\mathbb{P}}$, and the new MDP $M_{\text{Lo}'} = \{\mathcal{S}, \mathcal{A}, \mathbb{P}_{\text{Lo}}, r'_{\text{Lo}}, H\}$ has optimal value dominance on M_{Hi}.

Proofs for Examples Above Ex. A.1 is obvious, we just prove the rest two. First of all, for arbitrary $h \in [H]$ and $s_h \in \mathcal{S}_h$, we should have:

$$\begin{aligned}
&V^*_{\text{Hi},h}(s_h) - V^*_{\text{Lo},h}(s_h) \le Q^*_{\text{Hi},h}(s_h, \pi^*_{\text{Hi}}) - Q^*_{\text{Lo},h}(s_h, \pi^*_{\text{Hi}}) \\
=&r_{\text{Hi},h}(s_h, \pi^*_{\text{Hi}}) - r_{\text{Lo},h}(s_h, \pi^*_{\text{Hi}}) + (\mathbb{P}_{\text{Hi},h} - \mathbb{P}_{\text{Lo},h})V^*_{\text{Hi},h+1}(s_h, \pi^*_{\text{Hi}}) + \mathbb{P}_{\text{Lo},h}(V^*_{\text{Hi},h+1} - V^*_{\text{Lo},h+1})(s_h, \pi^*_{\text{Hi}}) \\
\le&r_{\text{Hi},h}(s_h, \pi^*_{\text{Hi}}) - r_{\text{Lo},h}(s_h, \pi^*_{\text{Hi}}) + (\mathbb{P}_{\text{Hi},h} - \mathbb{P}_{\text{Lo},h})V^*_{\text{Hi},h+1}(s_h, \pi^*_{\text{Hi}}) + \mathbb{P}_{\text{Lo},h}(V^*_{\text{Hi},h+1} - Q^*_{\text{Lo},h+1}(\cdot, \pi^*_{\text{Hi}}))(s_h, \pi^*_{\text{Hi}}) \\
\le&... \\
\le&\mathbb{E}_{M_{\text{Lo}}, \pi^*_{\text{Hi}}}[\sum_{h'=h}^{H} r_{\text{Hi},h}(s_{h'}, a_{h'}) - r_{\text{Lo},h'}(s_{h'}, a_{h'}) + (\mathbb{P}_{\text{Hi},h'} - \mathbb{P}_{\text{Lo},h'})V^*_{\text{Hi},h'+1}(s_{h'}, a_{h'})|s_h] \\
\le&\mathbb{E}_{M_{\text{Lo}}, \pi^*_{\text{Hi}}}[\sum_{h'=h}^{H} |r_{\text{Hi},h}(s_{h'}, a_{h'}) - r_{\text{Lo},h'}(s_{h'}, a_{h'})| + (H-h) \cdot \|\mathbb{P}_{\text{Hi},h'}(\cdot|s_{h'}, a_{h'}) - \mathbb{P}_{\text{Lo},h'}(\cdot|s_{h'}, a_{h'})\|_1|s_h],
\end{aligned}$$

Therefore, in Example A.2, we should expect:

$$V^*_{\text{Hi},h}(s_h) - V^*_{\text{Lo},h}(s_h) \le (H-h) \cdot \frac{\Delta_{\min}}{4H(H+1)} + (H-h) \cdot (H-h) \cdot \frac{\Delta_{\min}}{4H^2(H+1)} \le \frac{\Delta_{\min}}{2(H+1)}.$$

Besides, for Example A.3, we have:

$$\begin{aligned}
V^*_{\text{Hi},h}(s_h) - V^*_{\text{Lo}',h}(s_h) \le&\mathbb{E}_{M_{\text{Lo}}, \pi^*_{\text{Hi}}}[\sum_{h'=h}^{H} r_{\text{Hi},h}(s_{h'}, a_{h'}) - r_{\text{Lo},h'}(s_{h'}, a_{h'}) + \xi_r \\
&+ (H-h)\xi_{\mathbb{P}} + (H-h) \cdot \|\mathbb{P}_{\text{Hi},h'}(\cdot|s_{h'}, a_{h'}) - \mathbb{P}_{\text{Lo},h'}(\cdot|s_{h'}, a_{h'})\|_1|s_h] \\
\le&0
\end{aligned}$$

Therefore, both Example A.2 and A.3 satisfy Assump. B.

A.4 Detailed Discussion on Open Problems

We believe there are many interesting directions to follow in the future and highlight in three aspects:

First of all, we conjecture our unique optimal policy assumption can be relaxed and the $O(\frac{1}{H})$ factor in Def. 2.2 and Def. F.1 can be removed by advanced techniques. It's also important to study how to get rid of lower bound knowledge in $\Delta_{\min}$.

Secondly, in TRL-MST setting, for those s_h such that there are multiple source tasks $M_{\text{Lo},w_1}, M_{\text{Lo},w_2}, ..., M_{\text{Lo},w_j} \in \mathcal{M}_{\text{Lo}}$ close to M_{Hi}, beyond the constant regret, one may consider to

integrate the information from those source tasks together to further accelerate the learning; Moreover, although we do not make it explicitly, it's possible to combine our techniques in Sec. 5 with existing MT-RL algorithms to develop algorithms with guarantees about the reduction on not only the total regret but also some specific tasks.

Finally, although we show in Sec. 3 that robust transfer objective requires OVD assumption, and the model difference tolerance is at most $O(\Delta_{\min})$, we conjecture that, there might exists milder assumptions about the structure of source and target tasks and the prior knowledge about it, which may eliminate out our hard instance. Additionaly, in some cases, it may be reasonable to relax the objective by allowing some chance of negative transfer in part of target tasks. Then, we can do more aggressive transfer without too much concern on the algorithm's overall performance. These potential directions are left for future work.

B Proofs for Lower Bound

Theorem 3.1. *Under the violation of Assump. B, even regardless of the optimality of Alg^{Lo}, for each algorithm pair (Alg^{Lo}, Alg^{Hi}), it cannot simultaneously (1) achieve constant regret for the case when $M_{Lo} = M_{Hi}$ and (2) ensure sub-linear regret in all the other cases.*

Proof. Consider the two-armed bandit setting. Given arbitrary $\Delta, \mu \in (0, 1)$ satisfying $0 < \mu - \Delta < \mu < \mu + \Delta$, we can construct two two-armed Bernoulli bandit problem M and M' such that:

$$\mu_M(1) = \mu_{M'}(1) = \mu; \quad \mu_M(2) = \mu - \Delta; \quad \mu_{M'}(2) = \mu + \Delta.$$

We choose M to be the low-tier task, i.e. $M_{\text{Lo}} = M$, and choose M and M' to be the high-tier task. Note that the minimal gap in M and M' is Δ, and $\mu_M(1) = \mu_{M'}(2) - \Delta < \mu_{M'}(2) - \frac{\Delta}{2}$, which implies M_{Lo} does not have optimal value dominance on M_{Hi} when $M_{\text{Lo}} = M$ and $M_{\text{Hi}} = M'$.

Now, we consider the following learning process: the learner will get access to the low-tier task $M_{\text{Lo}} = M$, and the high-tier task M_{Hi} will be uniformly randomly selected between M and M', while the learner does not know which it is. Without loss of generality, we consider deterministic algorithms Alg^{Lo} and Alg^{Hi} (since one can first generate the randomness before the learning process), i.e. for arbitrary step k, given the interaction history $\tau_k := (a^1_{\text{Hi}}, r^1_{\text{Hi}}, a^1_{\text{Lo}}, r^1_{\text{Lo}}, ..., a^{k-1}_{\text{Hi}}, r^{k-1}_{\text{Hi}}, a^{k-1}_{\text{Lo}}, r^{k-1}_{\text{Lo}})$, the policy $(\pi^k_{\text{Lo}}, \pi^k_{\text{Hi}})$ produced by $\text{Alg}^{\text{Lo}}(\tau_k)$ and $\text{Alg}^{\text{Hi}}(\tau_k)$ is fixed, where $a^k_{\text{Lo}}, r^k_{\text{Lo}}$ (or $a^k_{\text{Hi}}, r^k_{\text{Hi}}$) denotes the arm pulled and the reward observed in task M_{Lo} (or M_{Hi}) at iteration k.

In the following, we will use $\Pr^{M_{\text{Lo}}, M_{\text{Hi}}}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\cdot)$ to denote the probability if the learner use algorithm pair $(\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}})$ and solve task pair $(M_{\text{Lo}}, M_{\text{Hi}})$. Note that the pseudo-regret of Alg^{Hi} when $M_{\text{Hi}} = M$ can be written as:

$$\text{Regret}_K(M_{\text{Hi}}; M_{\text{Hi}} = M) = \sum_{\tau_K : \Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_K) > 0} \Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_K) N_{\text{Hi}}(2; \tau_K) \Delta. \tag{2}$$

where $N_{\text{Hi}}(i; \tau_K)$ denotes the number of times the i-th arm is pulled in task M_{Hi} in trajectory τ_K.

Because both M_{Lo} and M_{Hi} are two-armed Bernoulli bandits, and each arm in those MDPs has non-zero probability mass on both value 0 and 1 and the algorithms are deterministic, for arbitrary $k \geq 1$ and τ_k, $\Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_k) > 0$ if and only if $\Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_k) > 0$.

Now, we consider the following probability ratio, for arbitrary τ_k with $\Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_k) > 0$:

$$\begin{aligned}
\frac{\Pr^{M,M'}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_k)}{\Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_k)} &= \frac{\Pr_{M'}(r_{\text{Hi}} = r^{k-1}_{\text{Hi}} | a^{k-1}_{\text{Hi}})}{\Pr_{M}(r_{\text{Hi}} = r^{k-1}_{\text{Hi}} | a^{k-1}_{\text{Hi}})} \cdot \frac{\Pr^{M,M'}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_{k-1})}{\Pr^{M,M}_{\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}}}(\tau_{k-1})} \\
&\qquad (\Pr_M(\cdot) \text{ denotes the probability of event on model } M) \\
&= \prod_{k'=1}^{k} \frac{\Pr_{M'}(r_{\text{Hi}} = r^{k'}_{\text{Hi}} | a^{k'}_{\text{Hi}})}{\Pr_{M}(r_{\text{Hi}} = r^{k'}_{\text{Hi}} | a^{k'}_{\text{Hi}})} \geq (\frac{1 - \mu - \Delta}{1 - \mu + \Delta})^{N_{\text{Hi}}(2; \tau_k)}.
\end{aligned} \tag{3}$$

where for the first equality, we use the fact that the algorithms are deterministic, and the randomness of r^{k-1}_{Lo} only depends on a^{k-1}_{Lo} so it cancels out; the last inequality is because that the ratio is 1 if

$a_{\text{Hi}}^{k'} = 1$ and the ratio can be lower bounded by $(1-\mu-\Delta)/(1-\mu+\Delta)$ otherwise, Therefore, combining with Eq. (2), we have:

$$\begin{aligned}
\text{Regret}_K(M_{\text{Hi}}; M_{\text{Hi}} = M') &= \sum_{\tau_K} \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) N_{\text{Hi}}(1;\tau_K)\Delta \\
&= \sum_{\tau_K:\Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K)>0} \Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) \frac{\Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K)}{\Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K)} N_{\text{Hi}}(1;\tau_K)\Delta \\
&\geq \sum_{\tau_K:\Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K)>0} \Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) (\frac{1-\mu-\Delta}{1-\mu+\Delta})^{N_{\text{Hi}}(2;\tau_K)} N_{\text{Hi}}(1;\tau_K)\Delta
\end{aligned} \tag{4}$$

Suppose the algorithm pair $(\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}})$ can achieve constant regret C when $(M_{\text{Lo}}, M_{\text{Hi}}) = (M, M)$, i.e.

$$\mathbb{E}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}},M,M}[N_{\text{Hi}}(2;\tau_K)\Delta] \leq C, \quad \forall K \geq 1.$$

then, according to Markov inequality, for arbitrary constant $\delta \in (0,1)$ we have:

$$\Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(N_{\text{Hi}}(2;\tau_K) \leq \frac{C}{\Delta\delta}) \geq 1-\delta, \quad \forall K \geq 1. \tag{5}$$

which is equivalent to (note that τ_K is the random variable)

$$\sum_{\tau_K:N_{\text{Hi}}(2;\tau_K)\leq\frac{C}{\Delta\delta}} \Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) \geq 1-\delta.$$

Combining with Eq. (4), by choosing an arbitrary constant $\delta \in (0,1)$, for arbitrary $K \geq 1$, we have:

$$\begin{aligned}
\text{Regret}_K(M_{\text{Hi}}; M_{\text{Hi}} = M') &\geq \sum_{\tau_K:N_{\text{Hi}}(2;\tau_K)\leq\frac{C}{\Delta\delta}} \Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) (\frac{1-\mu-\Delta}{1-\mu+\Delta})^{N_{\text{Hi}}(2;\tau_K)} N_{\text{Hi}}(1;\tau_K)\Delta \\
&\geq \sum_{\tau_K:N_{\text{Hi}}(2;\tau_K)\leq\frac{C}{\Delta\delta}} \Pr^{M,M}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) (\frac{1-\mu-\Delta}{1-\mu+\Delta})^{\frac{C}{\Delta\delta}} (K - \frac{C}{\Delta\delta})\Delta \\
&\geq (1-\delta)\cdot(\frac{1-\mu-\Delta}{1-\mu+\Delta})^{\frac{C}{\Delta\delta}} (K - \frac{C}{\Delta\delta})\Delta \\
&= O(K).
\end{aligned}$$

which finishes the proof. □

Theorem 3.2. *[Transferable States are Restricted by $\Delta_{\min}$] Under Assump. B, regardless of the optimality of Alg^{Lo}, given arbitrary $\Delta_{\min}$ and arbitrary $\Delta \in [\frac{\Delta_{\min}}{2}, \Delta_{\min}]$, for each algorithm pair (Alg^{Lo}, Alg^{Hi}), it cannot simultaneously (1) achieve constant regret for the case when M_{Lo} and M_{Hi} with minimal gap $\Delta_{\min}$ are Δ-close, and (2) ensure sub-linear regret in all other cases.*

Proof. We can construct three two-armed Bernoullis bandit problem M, M' and M'' such that:

$$\begin{aligned}
&\mu_M(1) = \mu, \quad \mu_M(2) = \mu - \Delta; \\
&\mu_{M'}(1) = \mu - \Delta', \quad \mu_{M'}(2) = \mu - \Delta - \Delta'; \\
&\mu_{M''}(1) = \mu - \Delta', \quad \mu_{M''}(2) = \mu + \Delta - \Delta';
\end{aligned}$$

where Δ and μ are chosen to satisfy $0 < \mu - 2\Delta < \mu < \mu + \Delta$, and $\Delta' \in [\frac{\Delta}{2}, \Delta]$. Note that by construction, Δ is $\Delta_{\min}$. Now, consider the following learning process, the learner will be provided M as the low-tier task M_{Lo}, and the high-tier task M_{Hi} will be uniformly sampling from $\{M', M''\}$. Easy to check that, $M_{\text{Lo}} = M$, has optimal value dominance on M_{Hi} when $M_{\text{Hi}} = M'$ or $M_{\text{Hi}} = M''$. Next, we want to show that, for arbitrary algorithm pair $(\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}})$, if the learner can achieve constant regret when $M_{\text{Hi}} = M'$, it must achieve linear regret when $M_{\text{Hi}} = M''$.

The remaining proof is similar to the proof for Thm. 3.1. First of all, we have:

$$\text{Regret}_K(M_{\text{Hi}}; M_{\text{Hi}} = M') = \sum_{\tau_K : \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) > 0} \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) N_{\text{Hi}}(2;\tau_K)\Delta. \tag{6}$$

As an analogue of Eq. (3), we have:

$$\frac{\Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_k)}{\Pr^{M,M''}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_k)} \geq (\frac{\mu - \Delta - \Delta'}{\mu + \Delta - \Delta'})^{N_{\text{Hi}}(2;\tau_k)}.$$

Combining with Eq. (6), we have:

$$\text{Regret}_K(M_{\text{Hi}}; M_{\text{Hi}} = M'') \geq \sum_{\tau_K : \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) > 0} \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) (\frac{\mu - \Delta - \Delta'}{\mu + \Delta - \Delta'})^{N_{\text{Hi}}(2;\tau_K)} N_{\text{Hi}}(1;\tau_K)\Delta \tag{7}$$

Suppose the algorithm pair $(\text{Alg}^{\text{Lo}}, \text{Alg}^{\text{Hi}})$ can achieve constant regret C when $(M_{\text{Lo}}, M_{\text{Hi}}) = (M, M')$, we must have:

$$\sum_{\tau_K : N_{\text{Hi}}(2;\tau_K) \leq \frac{C}{\Delta\delta}} \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) \geq 1 - \delta.$$

By choosing an arbitrary fixed constant $\delta \in (0,1)$, for arbitrary K, we have:

$$\begin{aligned}
\text{Regret}_K(M_{\text{Hi}}; M_{\text{Hi}} = M'') \geq & \sum_{\tau_K : N_{\text{Hi}}(2;\tau_K) \leq \frac{C}{\Delta\delta}} \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) (\frac{1 - \mu - \Delta}{1 - \mu + \Delta})^{N_{\text{Hi}}(2;\tau_K)} N_{\text{Hi}}(1;\tau_K)\Delta \\
\geq & \sum_{\tau_K : N_{\text{Hi}}(2;\tau_K) \leq \frac{C}{\Delta\delta}} \Pr^{M,M'}_{\text{Alg}^{\text{Lo}},\text{Alg}^{\text{Hi}}}(\tau_K) (\frac{1 - \mu - \Delta}{1 - \mu + \Delta})^{\frac{C}{\Delta\delta}} (K - \frac{C}{\Delta\delta})\Delta \\
\geq & (1 - \delta) \cdot (\frac{\mu - \Delta - \Delta'}{\mu + \Delta - \Delta'})^{\frac{C}{\Delta\delta}} (K - \frac{C}{\Delta\delta})\Delta \\
= & O(K).
\end{aligned}$$

which finishes the proof. □

C Proofs for Tiered MAB with Single Source/Low-Tier Task

Lemma C.1 (Concentration Inequality)**.** *In Alg. 1, at each iteration k, we have:*

$$\Pr(|\mu_{Lo}(i) - \widehat{\mu}^k_{Lo}(i)| \geq \sqrt{\frac{2\alpha \log f(k)}{N^k_{Lo}(i)}}) \leq \frac{2}{f(k)^\alpha} \leq \frac{1}{8Ak^{2\alpha}}, \quad \forall i \in [A]$$

$$\Pr(|\mu_{Hi}(i) - \widehat{\mu}^k_{Hi}(i)| \geq \sqrt{\frac{2\alpha \log f(k)}{N^k_{Hi}(i)}}) \leq \frac{2}{f(k)^\alpha} \leq \frac{1}{8Ak^{2\alpha}}, \quad \forall i \in [A]$$

As a direct result, we have the following lemma:

Lemma C.2. *[Valid Under Estimation] For arbitrary $i \in [A]$, if $\mu_{Lo}(i) \leq \mu_{Hi}(i) + \varepsilon$, for arbitrary iteration k in Alg. 1, we have:*

$$\Pr(\underline{\mu}_{Lo}(i) \leq \overline{\mu}_{Hi}(i) + \varepsilon) \geq 1 - \frac{4}{f(k)^\alpha} \geq 1 - \frac{1}{4Ak^{2\alpha}}$$

Proof. According to Lem. C.1, w.p. at least $1 - \frac{4}{f(k)^\alpha}$ we have:

$$\Pr(\underline{\mu}_{\text{Lo}}(i) \leq \overline{\mu}_{\text{Hi}}(i) + \varepsilon) \geq \Pr(\{\underline{\mu}_{\text{Lo}}(i) \leq \mu_{\text{Lo}}(i)\} \cap \{\mu_{\text{Hi}}(i) + \varepsilon \leq \overline{\mu}_{\text{Hi}}(i) + \varepsilon\}) \geq 1 - \frac{4}{f(k)^\alpha} \geq 1 - \frac{1}{4Ak^{2\alpha}}.$$

□

Next, we recall two useful lemma: Lemma 4.2 and Lemma D.1 from [13].

Lemma C.3 (Property of UCB; Lem 4.2 in [13])**.** *With the choice that $f(k) = 1 + 16A^2(k+1)^2$, there exists a constant c, for arbitrary i with $\Delta_{Lo}(i) > 0$ and arbitrary $\nu \in [1, 4A]$, in UCB algorithm, we have:*

$$\Pr(N_{Lo}^k(i) \geq \frac{k}{\nu}) \leq \frac{2}{k^{2\alpha-1}}, \quad \forall k \geq \nu + c \cdot \frac{\alpha\nu}{\Delta_{Lo}^2(i)} \log(1 + \frac{\alpha A}{\Delta_{\min}}).$$

Lemma C.4 (Lemma D.1 in [13])**.** *Given an arm i, we separate all the arms into two parts depending on whether its gap is larger than $\Delta_{Lo}(i)$ and define $G_i^{lower} := \{\iota | \Delta_{Lo}(\iota) > \Delta_{Lo}(i)/2\}$ and $G_i^{upper} := \{\iota | \Delta_{Lo}(\iota) \leq \Delta_{Lo}(i)/2\}$. With the choice that $f(k) = 1 + 16A^2(k+1)^2$, there is a constant c, such that for arbitrary i with $\Delta_{Lo}(i) > 0$, for $\underline{\pi}_{Lo}^k$ in Alg 1, there exists a constant c, such that:*

$$\Pr(i = \underline{\pi}_{Lo}^k) \leq 2/k^{2\alpha} + 2A/k^{2\alpha-1}, \quad \forall k \geq k_i := 8\alpha c \Big(\sum_{\iota \in G_i^{lower}} \frac{1}{\Delta_{Lo}^2(\iota)} + \frac{4|G_i^{upper}|}{\Delta_{Lo}^2(i)} \Big) \log(1 + \frac{\alpha A}{\Delta_{\min}}) \tag{8}$$

Lemma C.5. *We denote $k_i' := 3A + c \cdot \frac{3\alpha A}{\Delta_{Lo}^2(i)} \log(1 + \frac{\alpha A}{\Delta_{\min}})$ and $\widetilde{k}_i := 3 + c \cdot \frac{3\alpha}{\Delta_{Lo}^2(i)} \log(1 + \frac{\alpha A}{\Delta_{\min}})$, where c is specified in Lem. C.3, and denote $k_{\max} := \max_{i \neq i^*} \max\{k_i,\ k_i'\}$, where k_i is defined in Lemma C.4, we have:*

$$\Pr(i^* = \underline{\pi}_{Lo}^k) = 1 - \sum_{i \neq i^*} \Pr(i = \underline{\pi}_{Lo}^k) \geq 1 - \frac{2A}{k^{2\alpha}} - \frac{2A^2}{k^{2\alpha-1}}, \quad \forall k \geq k_{\max}. \tag{9}$$

$$\Pr(N_{Lo}^k(i^*) > \frac{k}{2}) \geq \Pr(N_{Lo}^k(i^*) \geq \frac{2k}{3}) \geq 1 - \sum_{i \neq i^*} \Pr(N_{Lo}^k(i) \leq \frac{k}{3A}) \geq 1 - \frac{2A}{k^{2\alpha-1}}, \quad \forall k \geq k_{\max}. \tag{10}$$

$$\Pr(N_{Lo}^k(i) > \frac{k}{2}) \leq \Pr(N_{Lo}^k(i) \geq \frac{k}{3}) \leq \frac{2}{k^{2\alpha-1}}, \quad \forall k \geq \widetilde{k}_i. \tag{11}$$

Proof. By applying Lem. C.4 and Lem. C.3 we can obtain the results. □

Lemma C.6. *For arbitrary $K \geq A + 1$ and arbitrary $k_0 \leq K$, and $i \neq i_{Hi}^*$, we have:*

$$\begin{aligned} N_{Hi}^K(i) \leq k_0 &+ \sum_{k=k_0+1}^{K} \mathbb{I}[\{\underline{\mu}_{Lo}^k(\underline{\pi}_{Lo}^k) \leq \overline{\mu}_{Hi}^k(\underline{\pi}_{Lo}^k) + \varepsilon\} \cap \{N_{Lo}^k(\underline{\pi}_{Lo}^k) > k/2\} \cap \{i = \underline{\pi}_{Lo}^k\} \cap \{\pi_{Hi}^k = i\}] \\ &+ \sum_{k=k_0+1}^{K} \mathbb{I}[0 \geq \widehat{\mu}_{Hi}^k(i_{Hi}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{Hi}^k(i_{Hi}^*)}} - \mu_{Hi}(i_{Hi}^*)] \\ &+ \sum_{k=k_0+1}^{K} \mathbb{I}[\widetilde{\mu}_{Hi}^k(i) + \sqrt{\frac{2\alpha \log f(K)}{k}} - \mu_{Hi}(i) - \Delta_{Hi}(i) \geq 0] \end{aligned} \tag{12}$$

Proof.

$$\begin{aligned} N_{\mathrm{Hi}}^K(i) &= \sum_{k=1}^{K} \mathbb{I}[\pi_{\mathrm{Hi}}^k = i] \leq k_0 + \sum_{k_0+1}^{K} \mathbb{I}[\pi_{\mathrm{Hi}}^k = i] \\ \leq k_0 &+ \sum_{k=k_0+1}^{K} \mathbb{I}[\underbrace{\{\underline{\mu}_{\mathrm{Lo}}^k(\underline{\pi}_{\mathrm{Lo}}^k) \leq \overline{\mu}_{\mathrm{Hi}}^k(\underline{\pi}_{\mathrm{Lo}}^k) + \varepsilon\} \cap \{N_{\mathrm{Lo}}^k(\underline{\pi}_{\mathrm{Lo}}^k) > k/2\} \cap \{i = \underline{\pi}_{\mathrm{Lo}}^k\}}_{e_1} \cap \{\pi_{\mathrm{Hi}}^k = i\}] \\ &+ \sum_{k=k_0+1}^{K} \mathbb{I}[\underbrace{\{\widehat{\mu}_{\mathrm{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\mathrm{Hi}}^k(i)}} - \mu_{\mathrm{Hi}}(i) - \Delta_{\mathrm{Hi}}(i) \geq \widehat{\mu}_{\mathrm{Hi}}^k(i_{\mathrm{Hi}}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{\mathrm{Hi}}^k(i_{\mathrm{Hi}}^*)}} - \mu_{\mathrm{Hi}}(i_{\mathrm{Hi}}^*)\}}_{e_2} \cap \{\pi_{\mathrm{Hi}}^k = i\}]. \end{aligned}$$

(If $\pi_{\mathrm{Hi}}^k = i$ happens, one of e_1 and e_2 must hold)

For the second term, we have:

$$\sum_{k=k_0+1}^{K} \mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i)}} - \mu_{\text{Hi}}(i) - \Delta_{\text{Hi}}(i) \geq \widehat{\mu}_{\text{Hi}}^k(i_{\text{Hi}}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i_{\text{Hi}}^*)}} - \mu_{\text{Hi}}(i_{\text{Hi}}^*)\} \cap \{\pi_{\text{Hi}}^k = i\}]$$

$$\leq \sum_{k=k_0+1}^{K} \mathbb{I}[0 \geq \widehat{\mu}_{\text{Hi}}^k(i_{\text{Hi}}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i_{\text{Hi}}^*)}} - \mu_{\text{Hi}}(i_{\text{Hi}}^*)]$$

$$+ \sum_{k=k_0+1}^{K} \mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i)}} - \mu_{\text{Hi}}(i) - \Delta_{\text{Hi}}(i) \geq 0\} \cap \{\pi_{\text{Hi}}^k = i\}]$$

$$(\mathbb{I}[a \geq b] \leq \mathbb{I}[a \geq c] + \mathbb{I}[c \geq b]; \mathbb{I}[a \cap b] \leq \mathbb{I}[a])$$

$$\leq \sum_{k=k_0+1}^{K} \mathbb{I}[0 \geq \widehat{\mu}_{\text{Hi}}^k(i_{\text{Hi}}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i_{\text{Hi}}^*)}} - \mu_{\text{Hi}}(i_{\text{Hi}}^*)] + \sum_{k=k_0+1}^{K} \mathbb{I}[\widetilde{\mu}_{\text{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(K)}{k}} - \mu_{\text{Hi}}(i) - \Delta_{\text{Hi}}(i) \geq 0]$$

where in the last step, $\widetilde{\mu}_{\text{Hi}}^k(i)$ is defined to be the average of k random samples from reward distribution of arm i in M_{Hi}, and we replace $N_{\text{Hi}}^k(i)$ in the denominator with increasing k since the indicator function equals 1 only when $\{\pi_{\text{Hi}}^k = i\}$, which implies that $N_{\text{Hi}}^k(i)$ should increase by 1. □

Theorem 4.1. *[Tiered MAB with Single Source Tasks] Under Assump. A, B and C, by running Alg. 1 with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4}$ and $\alpha > 2$, we always have Regret$_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log K)$. Moreover, if M_{Hi} and M_{Lo} are $\frac{\widetilde{\Delta}_{\min}}{4}$-close, we have: Regret$_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log \frac{A}{\Delta_{\min}})$.*

Proof. We first study the case when M_{Hi} and M_{Lo} satisfy Def. 2.1.

Case 1: M_{Hi} and M_{Lo} are ε-close In this case, since $i_{\text{Lo}}^* = i_{\text{Hi}}^*$, we use i^* to denote the common optimal arm. As a result of Lem. C.2, we have:

$$\Pr(\underline{\mu}_{\text{Lo}}^k(i^*) \leq \overline{\mu}_{\text{Hi}}^k(i^*) + \varepsilon) \geq 1 - \frac{1}{4Ak^{2\alpha}}.$$

We consider the same $k_{\max}$ defined in Lem. C.5, as a result of Lem. C.5, for arbitrary $K \geq k_{\max} + 1$,

$$\sum_{k=k_{\max}+1}^{K} \Pr(\pi_{\text{Hi}}^k \neq i_{\text{Hi}}^*) \leq \sum_{k=k_{\max}+1}^{K} \Pr(\underline{\mu}_{\text{Lo}}^k(i^*) > \overline{\mu}_{\text{Hi}}^k(i^*) + \varepsilon) + \Pr(N_{\text{Lo}}^k(i^*) \leq \frac{k}{2}) + \Pr(i^* \neq \underline{\pi}_{\text{Lo}}^k)$$

$$\leq \sum_{k=k_{\max}+1}^{K} \frac{1}{4Ak^{2\alpha}} + \frac{2A}{k^{2\alpha}} + \frac{2A^2}{k^{2\alpha-1}} + \frac{2A}{k^{2\alpha-1}}$$

$$\leq \sum_{k=k_{\max}+1}^{\infty} \frac{8A^2}{k^{2\alpha-1}} \leq \frac{8A^2}{(2\alpha-2)k_{\max}^{2\alpha-2}}.$$

Therefore, all we need to do is to upper bound the regret up to step $k_{\max}$. In the following, we separately upper bound $\mathbb{E}[N_{\text{Hi}}^{k_{\max}}(i)]$ for $i \neq i^*$ for two cases depending on the comparison between $\Delta_{\text{Hi}}(i)$ and $\Delta_{\text{Lo}}(i)$.

Case 1-(a) $0 < \Delta_{\text{Hi}}(i) \leq 4\Delta_{\text{Lo}}(i)$ Recall $\widetilde{k}_i$ in Lem. C.5. In this case, since $\Delta_{\text{Lo}}^{-1}(i) \leq \Delta_{\text{Hi}}^{-1}(i)$, we have $\widetilde{k}_i = O(\frac{1}{\Delta_{\text{Hi}}^2(i)} \log \frac{A}{\Delta_{\min}})$, and by taking expectation over Eq. (12):

$$\mathbb{E}[N_{\text{Hi}}^K(i)] \leq \widetilde{k}_i + \sum_{k=\widetilde{k}_i+1}^{K} \Pr(\{N_{\text{Lo}}^k(\underline{\pi}_{\text{Lo}}^k) \geq \frac{k}{2}\}) + \sum_{k=1}^{K} \Pr(0 \geq \widehat{\mu}_{\text{Hi}}^k(i_{\text{Hi}}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i_{\text{Hi}}^*)}} - \mu_{\text{Hi}}(i_{\text{Hi}}^*))$$

$$+ \mathbb{E}[\sum_{k=1}^{K} \mathbb{I}[\widetilde{\mu}_{\text{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{k}} - \mu_{\text{Hi}}(i) - \Delta_{\text{Hi}}(i) \geq 0]]$$

$$\leq \widetilde{k}_i + \sum_{k=\widetilde{k}_i+1}^{K} \frac{2}{k^{2\alpha-1}} + \sum_{k=1}^{K} \frac{1}{8Ak^{2\alpha}} + 1 + \frac{2}{\Delta_{\text{Hi}}^2(i)}(\alpha \log f(K) + \sqrt{\pi\alpha \log f(K)} + 1)$$

(Lem. 8.2 in [17])

$$= O(\frac{1}{\Delta_{\text{Hi}}^2(i)} \log \frac{AK}{\Delta_{\min}})$$

Case 1-(b) $\Delta_{\mathbf{Hi}}(i) > 4\Delta_{\mathbf{Lo}}(i) > 0$ We introduce $\bar{k}_i := \frac{c_{\text{Hi},i}\alpha}{\Delta_{\text{Hi}}^2(i)} \log \frac{A}{\Delta_{\min}}$, where $c_{\text{Hi},i}$ is the minimal constant, such that when $k \geq \frac{c_{\text{Hi},i}\alpha}{\Delta_{\text{Hi}}^2(i)} \log \frac{\alpha A}{\Delta_{\min}}$, we always have $k \geq \frac{256\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}$. Therefore, for all $k \geq \bar{k}_i$, $N_{\text{Lo}}^k(i) > \frac{k}{2}$ implies $N_{\text{Lo}}^k(i) \geq \frac{128\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}$ and we have:

$$\begin{aligned}
&\mathbb{I}[\{\underline{\mu}_{\text{Lo}}^k(i) \leq \overline{\mu}_{\text{Hi}}^k(i) + \varepsilon\} \cap \{N_{\text{Lo}}^k(i) \geq \frac{k}{2}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
=&\mathbb{I}[\{\underline{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) + \frac{\Delta_{\text{Hi}}(i)}{4} \leq \overline{\mu}_{\text{Hi}}^k(i) \pm \mu_{\text{Hi}}(i) \pm \mu_{\text{Hi}}(i^*) \pm \mu_{\text{Lo}}(i^*) - \mu_{\text{Lo}}(i) + \varepsilon + \frac{\Delta_{\text{Hi}}(i)}{4}\} \\
&\quad \cap \{N_{\text{Lo}}^k(i) \geq \frac{128\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
\leq&\mathbb{I}[\{\underline{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) + \frac{\Delta_{\text{Hi}}(i)}{4} \leq \overline{\mu}_{\text{Hi}}^k(i) - \mu_{\text{Hi}}(i) - \Delta_{\text{Hi}}(i) + \frac{\Delta_{\text{Hi}}(i)}{4} + \frac{\Delta_{\text{Hi}}(i)}{8} + \frac{\Delta_{\text{Hi}}(i)}{8} + \frac{\Delta_{\text{Hi}}(i)}{4}\} \\
&(\Delta_{\text{Lo}}(i) \leq \tfrac{\Delta_{\text{Hi}}(i)}{4}\text{; Optimal value dominance } (\mu_{\text{Hi}}(i^*) - \mu_{\text{Lo}}(i^*) \leq \tfrac{\Delta_{\min}}{2} \leq \tfrac{\Delta_{\text{Hi}}(i)}{8}); \varepsilon < \tfrac{\Delta_{\min}}{4} \leq \tfrac{\Delta_{\text{Hi}}(i)}{8}) \\
&\quad \cap \{N_{\text{Lo}}^k(i) \geq \frac{128\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
\leq&\mathbb{I}[\{\underline{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) + \frac{\Delta_{\text{Hi}}(i)}{4} \leq \overline{\mu}_{\text{Hi}}^k(i) - \mu_{\text{Hi}}(i) - \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{N_{\text{Lo}}^k(i) \geq \frac{128\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
&\qquad (\Delta_{\text{Lo}}(i) + \tfrac{\Delta_{\min}}{2} + \varepsilon + \tfrac{\Delta_{\text{Hi}}(i)}{4} \leq \tfrac{\Delta_{\text{Hi}}(i)}{4} + \tfrac{\Delta_{\text{Hi}}(i)}{8} + \tfrac{\Delta_{\text{Hi}}(i)}{16} + \tfrac{\Delta_{\text{Hi}}(i)}{4} < \tfrac{3\Delta_{\text{Hi}}(i)}{4}) \\
\leq&\mathbb{I}[\{\underline{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) + \frac{\Delta_{\text{Hi}}(i)}{4} \leq 0\} \cap \{N_{\text{Lo}}^k(i) \geq \frac{128\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
&+ \mathbb{I}[\{0 \leq \overline{\mu}_{\text{Hi}}^k(i) - \mu_{\text{Hi}}(i) - \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
=&\mathbb{I}[\{\widehat{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) \leq \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Lo}}^k(i)}} - \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{N_{\text{Lo}}^k(i) \geq \frac{128\alpha \log f(k)}{\Delta_{\text{Hi}}^2(i)}\}] \\
&+ \mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^k(i) - \mu_{\text{Hi}}(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i)}} \geq \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}] \\
\leq&\mathbb{I}[\widehat{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) \leq -\sqrt{\frac{2\alpha \log f(k)}{N_{\text{Lo}}^k(i)}}] + \mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^k(i) - \mu_{\text{Hi}}(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i)}} \geq \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}].
\end{aligned} \tag{13}$$

By taking the expectation over both sides of Eq. (12), we have:

$$\begin{aligned}
&\mathbb{E}[N_{\text{Hi}}^K(i)] \\
\leq&\bar{k}_i + \sum_{k=\bar{k}_i+1}^{K} \Pr(\widehat{\mu}_{\text{Lo}}^k(i) - \mu_{\text{Lo}}(i) \leq -\sqrt{\frac{2\alpha \log f(k)}{N_{\text{Lo}}^k(i)}}) + \sum_{k=1}^{K} \Pr(0 \geq \widehat{\mu}_{\text{Hi}}^k(i_{\text{Hi}}^*) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i_{\text{Hi}}^*)}} - \mu_{\text{Hi}}(i_{\text{Hi}}^*)) \\
&+ \mathbb{E}[\sum_{k=\bar{k}_i+1}^{K} \mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^k(i) - \mu_{\text{Hi}}(i) + \sqrt{\frac{2\alpha \log f(k)}{N_{\text{Hi}}^k(i)}} \geq \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{i = \underline{\pi}_{\text{Lo}}^k\} \cap \{\pi_{\text{Hi}}^k = i\}]] \\
&+ \mathbb{E}[\sum_{k=1}^{K} \mathbb{I}[\widehat{\mu}_{\text{Hi}}^k(i) + \sqrt{\frac{2\alpha \log f(k)}{k}} - \mu_{\text{Hi}}(i) - \Delta_{\text{Hi}}(i) \geq 0]]
\end{aligned}$$

$$
\begin{aligned}
\leq & \bar{k}_i + \sum_{k=\bar{k}_i+1}^{K} \Pr(\widehat{\mu}^k_{\text{Lo}}(i) - \mu_{\text{Lo}}(i) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo}}(i)}}) + \sum_{k=1}^{K} \Pr(0 \geq \widehat{\mu}^k_{\text{Hi}}(i^*_{\text{Hi}}) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Hi}})}} - \mu_{\text{Hi}}(i^*_{\text{Hi}})) \\
& + 2\mathbb{E}[\sum_{k=1}^{K} \mathbb{I}[\widetilde{\mu}^k_{\text{Hi}}(i) + \sqrt{\frac{2\alpha \log f(k)}{k}} - \mu_{\text{Hi}}(i) \geq \frac{\Delta_{\text{Hi}}(i)}{4}]] \\
\leq & \bar{k}_i + \sum_{k=\bar{k}_i+1}^{K} \frac{2}{k^{2\alpha-1}} + 2 + \frac{64}{\Delta^2_{\text{Hi}}(i)}(\alpha \log f(K) + \sqrt{\pi \alpha \log f(K)} + 1) \qquad \text{(Lem. 8.2 in [17])} \\
= & O(\frac{1}{\Delta^2_{\text{Hi}}(i)} \log \frac{AK}{\Delta_{\min}}).
\end{aligned}
$$

Since $k_{\max} = \text{Poly}(A, \frac{1}{\Delta_{\min}})$, combining both cases, we have:

$$
\begin{aligned}
\text{Regret}_K(M_{\text{Hi}}) = \sum_{i \neq i^*} \Delta_{\text{Hi}}(i) \cdot \mathbb{E}[N^K_{\text{Hi}}(i)] \leq & \sum_{i \neq i^*} \Delta_{\text{Hi}}(i) \mathbb{E}[N^{k_{\max}}_{\text{Hi}}(i)] + \sum_{k=k_{\max}+1}^{K} \Pr(\pi^k_{\text{Hi}} \neq i^*_{\text{Hi}}) \\
= & O(\sum_{i \neq i^*} \frac{1}{\Delta_{\text{Hi}}(i)} \log \frac{A}{\Delta_{\min}}).
\end{aligned}
$$

Case 2: M_{Hi} and M_{Lo} are not ε-close In that case, we will use i^*_{Lo} and i^*_{Hi} to denote optimal arm in M_{Lo} and M_{Hi}, respectively, and either $i^*_{\text{Lo}} = i^*_{\text{Hi}}$ but $\mu_{\text{Lo}}(i^*_{\text{Lo}}) \geq \mu_{\text{Lo}}(i^*_{\text{Hi}}) + \varepsilon$, or $i^*_{\text{Lo}} \neq i^*_{\text{Hi}}$ and as a result of Assump. B:

$$
\mu_{\text{Hi}}(i^*_{\text{Lo}}) = \mu_{\text{Hi}}(i^*_{\text{Hi}}) - \Delta_{\text{Hi}}(i^*_{\text{Lo}}) \leq \mu_{\text{Lo}}(i^*_{\text{Lo}}) + \frac{\Delta_{\min}}{2} - \Delta_{\text{Hi}}(i^*_{\text{Lo}}) \leq \mu_{\text{Lo}}(i^*_{\text{Lo}}) - \frac{\Delta_{\text{Hi}}(i^*_{\text{Lo}})}{2}. \quad (14)
$$

Next, we first separate arms other than i^*_{Lo} and i^*_{Hi} into three cases:

Case 2-(a) $i \neq i^*_{\text{Lo}}, i \neq i^*_{\text{Hi}}$ and $0 < \Delta_{\text{Hi}}(i) \leq 4\Delta_{\text{Lo}}(i)$ The analysis is the same as Case 1-(a), and we have $\mathbb{E}[N^K_{\text{Hi}}(i)] = O(\frac{1}{\Delta^2_{\text{Hi}}(i)} \log \frac{AK}{\Delta_{\min}})$.

Case 2-(b) $i \neq i^*_{\text{Lo}}, i \neq i^*_{\text{Hi}}$ and $\Delta_{\text{Hi}}(i) > 4\Delta_{\text{Lo}}(i) > 0$ The analysis is the same as Case 1-(b), and we have $\mathbb{E}[N^K_{\text{Hi}}(i)] = O(\frac{1}{\Delta^2_{\text{Hi}}(i)} \log \frac{AK}{\Delta_{\min}})$.

Case 2-(c) Others If $i^*_{\text{Lo}} = i^*_{\text{Hi}}$, M_{Hi} suffers no regret when choosing $i = i^*_{\text{Lo}}$, and therefore:

$$
\text{Regret}_K(M_{\text{Hi}}) = \sum_{i \neq i^*_{\text{Hi}}} \Delta_{\text{Hi}}(i) \cdot \mathbb{E}[N^k_{\text{Hi}}(i)] \leq \sum_{i \neq i^*_{\text{Hi}}} \Delta_{\text{Hi}}(i) \cdot O(\frac{1}{\Delta^2_{\text{Hi}}(i)} \log \frac{AK}{\Delta_{\min}}) = O(\sum_{i \neq i^*_{\text{Hi}}} \frac{1}{\Delta_{\text{Hi}}(i)} \log \frac{AK}{\Delta_{\min}}).
$$

In the following, we study the case when $i^*_{\text{Lo}} \neq i^*_{\text{Hi}}$. For arm $i = i^*_{\text{Lo}}$, we define $k'_{\max} = \frac{c'_{\max}\alpha}{\Delta_{\text{Hi}}(i^*_{\text{Lo}})^2} \log \frac{\alpha A}{\Delta_{\min}}$, where $c'_{\max}$ is the minimal constant, such that for all $k \geq \frac{c'_{\max}\alpha}{\Delta_{\text{Hi}}(i^*_{\text{Lo}})^2} \log \frac{\alpha A}{\Delta_{\min}}$, we always have $k \geq \frac{1024\alpha \log f(k)}{\Delta_{\text{Hi}}(i^*_{\text{Lo}})^2}$. Similar to Eq. (13), we check the following event for $k \geq k'_{\max}$:

$$
\begin{aligned}
& \mathbb{I}[\{\underline{\mu}^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) \leq \overline{\mu}^k_{\text{Hi}}(\underline{\pi}^k_{\text{Lo}}) + \varepsilon\} \cap \{N^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) > k/2\} \cap \{i^*_{\text{Lo}} = \underline{\pi}^k_{\text{Lo}}\} \cap \{\pi^k_{\text{Hi}} = i^*_{\text{Lo}}\}] \\
= & \mathbb{I}[\{\widehat{\mu}^k_{\text{Lo}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}}) - \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo}}(i^*_{\text{Lo}})}} \leq \widehat{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Hi}}(i^*_{\text{Lo}}) + (\mu_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}})) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Lo}})}} + \varepsilon\} \\
& \cap \{N^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) > k/2\} \cap \{i^*_{\text{Lo}} = \underline{\pi}^k_{\text{Lo}}\} \cap \{\pi^k_{\text{Hi}} = i^*_{\text{Lo}}\}] \\
\leq & \mathbb{I}[\{\widehat{\mu}^k_{\text{Lo}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}}) - \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo}}(i^*_{\text{Lo}})}} \leq \widehat{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Hi}}(i^*_{\text{Lo}}) - \frac{\Delta_{\text{Hi}}(i^*_{\text{Lo}})}{4} + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Lo}})}}\} \\
& \qquad \text{(As a result of Eq. (14), } \mu_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}}) + \varepsilon \leq -\tfrac{\Delta_{\text{Hi}}(i^*_{\text{Lo}})}{2} + \tfrac{\Delta_{\min}}{4} \leq -\tfrac{\Delta_{\text{Hi}}(i^*_{\text{Lo}})}{4}) \\
& \cap \{N^k_{\text{Lo}}(\underline{\pi}^k_{\text{Lo}}) > k/2\} \cap \{i^*_{\text{Lo}} = \underline{\pi}^k_{\text{Lo}}\} \cap \{\pi^k_{\text{Hi}} = i^*_{\text{Lo}}\}]
\end{aligned}
$$

$$\leq \mathbb{I}[\{\widehat{\mu}^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) - \sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}})}} \leq -\frac{\Delta_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}{8}\} \cap \{N^k_{\mathrm{Lo}}(\underline{\pi}^k_{\mathrm{Lo}}) > k/2\} \cap \{i^*_{\mathrm{Lo}} = \underline{\pi}^k_{\mathrm{Lo}}\} \cap \{\pi^k_{\mathrm{Hi}} = i^*_{\mathrm{Lo}}\}]$$

$$+ \mathbb{I}[\{\widehat{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}} \geq \frac{\Delta_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}{8}\} \cap \{N^k_{\mathrm{Lo}}(\underline{\pi}^k_{\mathrm{Lo}}) > k/2\} \cap \{i^*_{\mathrm{Lo}} = \underline{\pi}^k_{\mathrm{Lo}}\} \cap \{\pi^k_{\mathrm{Hi}} = i^*_{\mathrm{Lo}}\}]$$

$$\leq \mathbb{I}[\widehat{\mu}^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}})}}] + \mathbb{I}[\{\widehat{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}} \geq \frac{\Delta_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}{8}\} \cap \{\pi^k_{\mathrm{Hi}} = i^*_{\mathrm{Lo}}\}]. \tag{15}$$

Therefore, by taking the expectation on both side of Eq. (12) and leveraging the above bound, we have:

$$\begin{aligned}
\mathbb{E}[N^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})] \leq & k'_{\max} + \mathbb{E}[\sum_{k=k'_{\max}+1}^{K} \mathbb{I}[\widehat{\mu}^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}})}}]] \\
& + \mathbb{E}[\sum_{k=k'_{\max}+1}^{K} \mathbb{I}[\{\widehat{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}} \geq \frac{\Delta_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}{8}\} \cap \{\pi^k_{\mathrm{Hi}} = i^*_{\mathrm{Lo}}\}]] \\
& + \sum_{k=1}^{K} \Pr(0 \geq \widehat{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Hi}}) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Hi}}(i^*_{\mathrm{Hi}})}} - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Hi}})) \\
& + \mathbb{E}[\sum_{k=1}^{K} \mathbb{I}[\widetilde{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) + \sqrt{\frac{2\alpha \log f(K)}{k}} - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) - \Delta_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) \geq 0]] \\
\leq & k'_{\max} + \sum_{k=k'_{\max}+1}^{K} \Pr(\widehat{\mu}^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) - \mu_{\mathrm{Lo}}(i^*_{\mathrm{Lo}}) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Lo}}(i^*_{\mathrm{Lo}})}}) \\
& + \sum_{k=1}^{K} \Pr(0 \geq \widehat{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Hi}}) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\mathrm{Hi}}(i^*_{\mathrm{Hi}})}} - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Hi}})) \\
& + 2\mathbb{E}[\sum_{k=1}^{K} \mathbb{I}[\widetilde{\mu}^k_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) + \sqrt{\frac{2\alpha \log f(K)}{k}} - \mu_{\mathrm{Hi}}(i^*_{\mathrm{Lo}}) \geq \frac{\Delta_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}{8}]] \\
\leq & k'_{\max} + 2\sum_{k=k'_{\max}+1}^{K} \frac{2}{f(k)^\alpha} + 2 \cdot (1 + \frac{128}{\Delta^2_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})}(\alpha \log f(K) + \sqrt{\pi \alpha \log f(K)} + 1)) \\
= & O(\frac{1}{\Delta^2_{\mathrm{Hi}}(i^*_{\mathrm{Lo}})} \log \frac{AK}{\Delta_{\min}})
\end{aligned}$$

As a result, for arbitrary K, we also have:

$$\mathrm{Regret}_K(M_{\mathrm{Hi}}) = \sum_{i \neq i^*_{\mathrm{Hi}}} \Delta_{\mathrm{Hi}}(i) \cdot \mathbb{E}[N^k_{\mathrm{Hi}}(i)] \leq \sum_{i \neq i^*_{\mathrm{Hi}}} \Delta_{\mathrm{Hi}}(i) \cdot O(\frac{1}{\Delta^2_{\mathrm{Hi}}(i)} \log \frac{AK}{\Delta_{\min}}) = O(\sum_{i \neq i^*_{\mathrm{Hi}}} \frac{1}{\Delta_{\mathrm{Hi}}(i)} \log \frac{AK}{\Delta_{\min}}).$$

□

D Proofs for RL Setting with Single Source/Low-Tier Task

D.1 Missing Algorithms, Conditions and Notations

Condition D.1 (Condition on $\mathrm{Alg}^{\mathrm{Lo}}$)**.** $\mathrm{Alg}^{\mathrm{Lo}}$ is an algorithm which returns deterministic policies at each iteration, and there exists C_1, C_2 only depending on S, A, H and $\Delta_{\min}$ but independent of k, such that for arbitrary $k \geq 2$, we have $\Pr(\mathcal{E}_{\mathrm{Alg}^{\mathrm{Lo}},k}) \geq 1 - \frac{1}{k^\alpha}$ for $\mathcal{E}_{\mathrm{Alg}^{\mathrm{Lo}},k}$ defined below:

$$\mathcal{E}_{\mathrm{Alg}^{\mathrm{Lo}},k} := \{\sum_{\widetilde{k}=1}^{k} V^*_{\mathrm{Lo},1}(s_1) - V^{\pi^{\widetilde{k}}_{\mathrm{Lo}}}_{\mathrm{Lo},1}(s_1) \leq C_1 + \alpha C_2 \log k\}.$$

Algorithm 6: ModelLearning

Input: Dataset D.
for $h = 1, 2, ..., H$ **do**
  for $s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$ **do**
    Use $N_h(s_h, a_h)$ and $N_h(s_h, a_h, s_{h+1})$ to denote the number of times state, action (next state) occurs in the dataset D.
    $\widehat{\mathbb{P}}_h(\cdot|s_h, a_h) \leftarrow \begin{cases} 0, & \text{if } N_h(s_h, a_h) = 0; \\ \frac{N_h(s_h, a_h, \cdot)}{N_h(s_h, a_h)}, & \text{otherwise.} \end{cases}$
  end
end
return $\{\widehat{\mathbb{P}}_1, \widehat{\mathbb{P}}_2, ..., \widehat{\mathbb{P}}_H\}$.

Remark D.2. *We consider such condition to avoid unnecessary discussion on analyzing Alg^{Lo}. Although most of the existing near-optimal algorithms fixed the confidence level before the running of algorithm, as analyzed in Appx. G in [13], one may combine those algorithms with doubling trick to realize Cond. D.1 for Alg^{Hi}, only at the cost of increase the the regret of Alg^{Lo} to $O(\log^2 K)$.*

Condition D.3 (Condition on function **Bonus** in Alg. 2)**.** Given a confidence sequence $\{\delta_k\}_{k=1}^K$ with $\delta_1, \delta_2, ..., \delta_K \in (0, 1/2)$, we define the following event at iteration $k \in [K]$ during the running of Alg. 2:

$$\mathcal{E}_{\textbf{Bonus},k} := \bigcap_{\substack{(\cdot)\in\{\text{Hi,Lo}\},\\ h\in[H],\\ s_h\in\mathcal{S}_h, a_h\in\mathcal{A}_h}} \left\{ \{H \cdot \|\widehat{\mathbb{P}}^k_{(\cdot),h}(s_h, a_h) - \mathbb{P}_{(\cdot),h}(s_h, a_h)\|_1 < b^k_{(\cdot),h}(s_h, a_h) \le B_1 \sqrt{\frac{\log(B_2/\delta_k)}{N^k_{(\cdot),h}(s_h, a_h)}} \} \right\}$$

we consider the choice of **Bonus** such that there exists such a B_1 and B_2 only depending on S, A, H but independent of δ_k, k or Δ, and $\Pr(\mathcal{E}_{\textbf{Bonus},k}) \ge 1 - \delta_k$ holds for any $k \in [K]$.[4]

For simplicity, in Cond. D.3, we directly control the l_1-norm of the error of model estimation. Our analysis framework is compatible with other bonus term for sharper analysis. We provide a simple example for the choice of B_1 and B_2 for completeness:

Example D.4. By Hoeffding's inequality and union bound, w.p. $1 - \delta$, for all s_{h+1}, s_h, a_h, we should have:

$$|\widehat{\mathbb{P}}^k_{(\cdot),h}(s_{h+1}|s_h, a_h) - \mathbb{P}_{(\cdot),h}(s_{h+1}|s_h, a_h)| \le \sqrt{\frac{1}{2N^k_{(\cdot),h}(s_h, a_h)} \log \frac{S^2 A}{\delta}}.$$

which implies:

$$H \cdot \|\widehat{\mathbb{P}}^k_{(\cdot),h}(\cdot|s_h, a_h) - \mathbb{P}_{(\cdot),h}(\cdot|s_h, a_h)\|_1 = O(SH \sqrt{\frac{\log(SA/\delta)}{N^k_{(\cdot),h}(s_h, a_h)}})$$

Therefore, one can choose $B_1 = \Theta(SH)$ and $B_2 = \Theta(SA)$.

Finally, we introduce the following concentration events about the deviation of the empirical visitation frequency and its expectation:

$$\begin{aligned}\mathcal{E}_{\textbf{Con},k} := \bigcap_{\substack{h\in[H],\\ s_h\in\mathcal{S}_h,\\ a_h\in\mathcal{A}_h}} \Big\{ &\{\frac{1}{2}\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\text{Lo}}}(s_h, a_h) - \alpha \log(2SAHk) \le N^k_{\text{Lo},h}(s_h, a_h) \le e\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\text{Lo}}}(s_h, a_h) + \alpha \log(2SAHk)\} \\ &\cap \{\frac{1}{2}\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\text{Hi}}}(s_h, a_h) - \alpha \log(2SAHk) \le N^k_{\text{Hi},h}(s_h, a_h) \le e\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\text{Hi}}}(s_h, a_h) + \alpha \log(2SAHk)\} \Big\}. \end{aligned} \tag{16}$$

[4]Note that we do not require the knowledge of Δ_i's to compute $b_{k,h}$.

D.2 Some Basic Lemma

Lemma D.5 (Underestimation)**.** *Given a **Bonus** satisfying Cond. D.3, at each iteration k during the running of Alg. 2, on the events $\mathcal{E}_{\textbf{Bonus},k}$ defined in Cond. D.3, $\forall h \in [H], \forall s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$, we have:*

$$\underline{Q}^{\pi^k_{Hi}}_{Hi,h}(s_h,a_h) \leq Q^{\pi^k_{Hi}}_{Hi,h}(s_h,a_h) \leq Q^*_{Hi,h}(s_h,a_h) \tag{17}$$

$$\underline{Q}^{k}_{Lo,h}(s_h,a_h) \leq Q^{\pi^k_{Lo}}_{Lo,h}(s_h,a_h) \leq Q^*_{Lo,h}(s_h,a_h) \tag{18}$$

$$Q^*_{Lo,h}(s_h,a_h) - \underline{Q}^{k}_{Lo,h}(s_h,a_h) \leq 2\mathbb{E}_{\pi^*_{Lo}}[\sum_{h'=h}^{H} \min\{H, b^k_{Lo,h'}(s_{h'},a_{h'})\}|s_{h'},a_{h'}]. \tag{19}$$

Proof. Note that the relationship between Q^π and Q^* will always hold, and therefore, we only compare the underestimation part.

According to the initialization, we have $\underline{Q}^k_{\text{Hi},h}(s_h,a_h) = Q^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h,a_h) = Q^*_{\text{Lo},h}(s_h,a_h)$ at $h = H+1$. Under the event of $\mathcal{E}_{\textbf{Bonus},k}$, at iteration k, suppose we have the inequality Eq. (17) holds for step $h+1$ for some $h \in [H]$, then at step h, we have:

$$\begin{aligned}
&\underline{Q}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h,a_h) - Q^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h,a_h) = \widehat{\mathbb{P}}^k_{\text{Hi},h}\underline{V}^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1}(s_h,a_h) - b^k_{\text{Hi},h}(s_h,a_h) - \mathbb{P}_{\text{Hi},h}V^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1}(s_h,a_h)\\
=&(\widehat{\mathbb{P}}^k_{\text{Hi},h} - \mathbb{P}_{\text{Hi},h})\underline{V}^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1}(s_h,a_h) - b^k_{\text{Hi},h}(s_h,a_h) + \mathbb{P}_{\text{Hi},h}(\underline{V}^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1} - V^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1})(s_h,a_h)\\
\leq&\mathbb{P}_{\text{Hi},h}(\underline{Q}^k_{\text{Hi},h+1}(\cdot,\pi^k_{\text{Hi}}) - Q^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1}(\cdot,\pi^k_{\text{Hi}}))(s_h,a_h) \leq 0.
\end{aligned}$$

where the first inequality is because $(\widehat{\mathbb{P}}^k_{\text{Hi},h} - \mathbb{P}_{\text{Hi},h})\underline{V}^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1}(s_h,a_h) \leq H \cdot \|\widehat{\mathbb{P}}^k_{\text{Hi},h}(s_h,a_h) - \mathbb{P}_{\text{Hi},h}(s_h,a_h)\|_1 \leq b^k_{\text{Hi},h}(s_h,a_h)$. The proof for Eq. (18) is similar (except $\underline{Q}^k_{\text{Lo},h}$ is the greedy value $\underline{\pi}^k_{\text{Lo}}$ instead of π^k_{Lo}). Besides,

$$\begin{aligned}
&Q^*_{\text{Lo},h}(s_h,a_h) - \underline{Q}^k_{\text{Lo},h}(s_h,a_h)\\
=&\min\{H, \mathbb{P}_{\text{Lo},h}V^*_{\text{Lo},h+1}(s_h,a_h) - \widehat{\mathbb{P}}^k_{\text{Lo},h}\underline{V}^k_{\text{Lo},h+1}(s_h,a_h) + b^k_{\text{Lo},h}(s_h,a_h)\}\\
\leq&\min\{H, (\mathbb{P}_{\text{Lo},h} - \widehat{\mathbb{P}}^k_{\text{Lo},h})\underline{V}^k_{\text{Lo},h+1}(s_h,a_h) + b^k_{\text{Lo},h}(s_h,a_h)\} + \mathbb{P}_{\text{Lo},h}(V^*_{\text{Lo},h+1} - \underline{V}^k_{\text{Lo},h+1})(s_h,a_h)\\
\leq&2\min\{H, b^k_{\text{Lo},h}(s_h,a_h)\} + \mathbb{P}_{\text{Lo},h}(V^*_{\text{Lo},h+1} - \underline{Q}^k_{\text{Lo},h+1}(\cdot,\pi^*_{\text{Lo}}))(s_h,a_h)\\
\leq&\ldots \leq 2\mathbb{E}_{\pi^*_{\text{Lo}}}[\sum_{h'=h}^{H} \min\{H, b^k_{\text{Lo},h'}(s_{h'},a_{h'})\}|s_h,a_h].
\end{aligned}$$

□

Theorem D.6 (Extended from Thm. 4.7 in [13])**.** *For an arbitrary sequence of deterministic policies $\pi^1, \pi^2, ..., \pi^K$, there must exist a sequence of deterministic optimal policies $\pi^{1,*}, \pi^{2,*}, ..., \pi^{K,*}$, such that $\forall h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$:*

$$|\sum_{k=1}^{K} d^{\pi^k}(s_h,a_h) - \sum_{k=1}^{K} d^{\pi^{k,*}}(s_h,a_h)| \leq \frac{1}{\Delta_{\min}}\Big(\sum_{k=1}^{K} V^*_1(s_1) - V^{\pi^k}_1(s_1)\Big).$$

Proof. We first define the following events:

$$\mathcal{E}_{k,h,\pi} := \{\pi_{k,h}(s_h) \neq \pi_h(s_h)\}, \quad \widetilde{\mathcal{E}}_{k,h,\pi} := \mathcal{E}_{k,h,\pi} \cap \bigcap_{h'=1}^{h} \mathcal{E}^{\complement}_{k,h'-1,\pi}, \quad \bar{\mathcal{E}}_{k,\pi} := \bigcup_{h=1}^{H} \mathcal{E}_{k,h,\pi}.$$

From Thm. 4.7 in [13], we already know $\sum_{k=1}^K d^{\pi^k}(s_h,a_h) - \sum_{k=1}^K d^{\pi^{k,*}}(s_h,a_h) \geq -\frac{1}{\Delta_{\min}}\Big(\sum_{k=1}^K V^*_1(s_1) - V^{\pi^k}_1(s_1)\Big)$. Next, we start with the second step in the proof of Lem.

E.3 in [13]: by choosing $\delta_{s_h,a_h} := \mathbb{I}[S_h = s_h, A_h = a_h]$ (which equals one if the state action is (s_h, a_h) at step h and otherwise 0) as reward function, we have:

$$\begin{aligned} d^{\pi}(s_h,a_h) - d^{\pi^k}(s_h,a_h) =& V_1^{\pi}(s_1;\delta_{s_h,a_h}) - V_1^{\pi^k}(s_1;\delta_{s_h,a_h}) \\ =& \mathbb{E}_{\pi^k}[\sum_{h'=1}^{h} \mathbb{I}[\widetilde{\mathcal{E}}_{k,h',\pi}](V_{h'}^{\pi}(s_{h'};\delta_{s_h,a_h}) - V_{h'}^{\pi^k}(s_{h'};\delta_{s_h,a_h}))] \\ & (V_{h'}^{\pi} = V_{h'}^{\pi^k} = 0 \text{ for all } h' \geq h+1) \end{aligned}$$

Starting from here, we do something differently:

$$\begin{aligned} d^{\pi}(s_h,a_h) - d^{\pi^k}(s_h,a_h) \geq& \mathbb{E}_{\pi^k}[\sum_{h'=1}^{h} -\mathbb{I}[\widetilde{\mathcal{E}}_{k,h',\pi}]V_{h'}^{\pi^k}(s_{h'};\delta_{s_h,a_h})] && (V_{h'}^{\pi} \geq 0) \\ \geq& -\mathbb{E}_{\pi^k}[\sum_{h'=1}^{h} \mathbb{I}[\widetilde{\mathcal{E}}_{k,h',\pi}]] && (V_{h'}^{\pi^k} \leq 1) \\ \geq& -\mathbb{E}_{s_1,a_1,s_2,a_2\ldots,s_H,a_H\sim\pi^k}[\mathbb{I}[\bar{\mathcal{E}}_{k,\pi}]] = -\Pr(\bar{\mathcal{E}}_{k,\pi}|\pi^k) \end{aligned}$$

Therefore, combining with the results in Lem. E.3, we can conclude that:

$$d^{\pi}(s_h,a_h) - d^{\pi^k}(s_h,a_h) \geq -\Pr(\bar{\mathcal{E}}_{k,\pi}|\pi^k)$$

We define $\pi^{k,*}$ to be a policy that equals π^k on those states where π^k is optimal, and takes the optimal action when π^k is non-optimal, then we have:

$$\begin{aligned} V_1^{\pi^{k,*}}(s_1) - V_1^{\pi^k}(s_1) =& \mathbb{E}_{\pi^k}[\sum_{h=1}^{H} \mathbb{I}[\widetilde{\mathcal{E}}_{k,h,\pi^{k,*}}](V_h^{\pi^{k,*}}(s_h) - V_h^{\pi^k}(s_h))] \\ \geq& \mathbb{E}_{\pi^k}[\sum_{h=1}^{H} \mathbb{I}[\widetilde{\mathcal{E}}_{k,h,\pi^{k,*}}](V_h^{\pi^{k,*}}(s_h) - Q_h^{\pi^{k,*}}(s_h,\pi^k(s_h)))] \\ \geq& \mathbb{E}_{\pi^k}[\sum_{h=1}^{H} \mathbb{I}[\widetilde{\mathcal{E}}_{k,h,\pi^{k,*}}]\Delta_{\min}] = \Delta_{\min}\Pr(\bar{\mathcal{E}}_{k,\pi^{k,*}}|\pi^k) \\ \geq& \Delta_{\min}(d^{\pi^k}(s_h,a_h) - d^{\pi^{k,*}}(s_h,a_h)). \end{aligned}$$

Sum over all $k \in [K]$, we have

$$\sum_{k=1}^{K} d^{\pi^k}(s_h,a_h) - \sum_{k=1}^{K} d^{\pi^{k,*}}(s_h,a_h) \leq \frac{1}{\Delta_{\min}}\Big(\sum_{k=1}^{K} V_1^*(s_1) - V_1^{\pi^k}(s_1)\Big).$$

which completes the proof. □

Corollary D.7 (Unique Optimal Policy)**.** *Under Assump. A, Thm. D.6 implies that:*

$$|\sum_{k=1}^{K} d_{Lo}^{\pi_{Lo}^k}(s_h,a_h) - K d^{\pi_{Lo}^*}(s_h,a_h)| \leq \frac{1}{\Delta_{\min}}(\sum_{k=1}^{K} V_{Lo,1}^*(s_1) - V_{Lo,1}^{\pi_{Lo}^k}(s_1))\Big)$$

Lemma D.8. *Let $\mathcal{F}_i$ for $i, 1...$ be a filtration and $X_1, ...X_n$ be a sequence of Bernoulli random variables with $\Pr(X_i = 1|\mathcal{F}_{i-1}) = P_i$ with P_i being $\mathcal{F}_{i-1}$-measurable and X_i being $\mathcal{F}_i$ measurable. It holds that*

$$\Pr(\exists n: \sum_{t=1}^{n} X_t < \frac{1}{2}\sum_{t=1}^{n} P_t - W) \leq e^{-W};\ \Pr(\exists n: \sum_{t=1}^{n} X_t > e\sum_{t=1}^{n} P_t - W) \leq e^{-W}.$$

Proof. The first inequality has been proven in Lemma F.4 of [6]. Here we adopt similar techniques to prove the second one.

We first define $m_t := e^{X_t - eP_t}$, since X_t is a Bernoulli random variable with $\Pr(X_t = 1) = P_t$, we should have:

$$\mathbb{E}_{X_t}[e^{X_t - eP_t}|\mathcal{F}_{t-1}] = \frac{eP_t + (1-P_t)}{e^{eP_t}} \leq \frac{eP_t + (1-P_t)}{eP_t + 1} \leq 1$$

where in the last but two step, we use $e^x \geq x + 1$. Therefore, $M_n := \prod_{t=1}^n m_t = e^{\sum_{t=1}^n X_t - eP_t}$ is a supermartingale. By Markov inequality, we have:

$$\Pr(\sum_{t=1}^n X_t - eP_t \geq W) = \Pr(M_n \geq e^W) \leq \frac{\mathbb{E}[M_n]}{e^W} \leq e^{-W}.$$

As a result, for a fixed n, we have $\Pr(\sum_{t=1}^n X_t \geq eP_t + W) \leq e^{-W}$. After a similar discussion about stopping time as [6], we have $\Pr(\exists n : \sum_{t=1}^n X_t \geq eP_t + W) \leq e^{-W}$. □

As a direct result of Lem. D.8, we have the following result:

Lemma D.9. *For arbitrary $k \geq 1$, and arbitrary $\alpha > 2$, $\Pr(\mathcal{E}_{\textbf{Con},k}) \geq 1 - \frac{1}{k^\alpha}$.*

D.3 Analysis of Alg$^{\text{Lo}}$

Lemma D.10 (The relationship between d^*_{Lo} and N^k_{Lo})**.** *There exists a constant c_{occup} which is independent of λ, S, A, H and gap Δ, s.t., for all $k \geq k_{occup} := c_{occup} \frac{C_1 + \alpha C_2}{\lambda \Delta_{\min}} \log(\frac{\alpha C_1 C_2 SAH}{\lambda \Delta_{\min}})$, on the events of $\mathcal{E}_{Alg^{Lo},k}$ and $\mathcal{E}_{\textbf{Con},k}$, $N^k_{Lo,h}(s_h, a_h) \geq \frac{\lambda}{3}k$ implies that $d^*_{Lo}(s_h, a_h) \geq \frac{\lambda}{9} > 0$, and conversely, if $d^*_{Lo}(s_h, a_h) \geq \lambda$, we must have $N^k_{Lo,h}(s_h) \geq N^k_{Lo,h}(s_h, \pi^*_{Lo}) \geq \frac{\lambda}{3}k$.*

Proof. On the event of $\mathcal{E}_{\text{Alg}^{\text{Lo}},k}$ and $\mathcal{E}_{\textbf{Con},k}$, as a result of Cor. D.7, $N^k_{\text{Lo},h}(s_h, a_h) \geq \frac{\lambda}{3}k$ implies:

$$\frac{\lambda}{3}k \leq N^k_{\text{Lo},h}(s_h, a_h) \leq ekd^*_{\text{Lo}}(s_h, a_h) + \alpha \log(2SAHk) + \frac{1}{\Delta_{\min}}(C_1 + \alpha C_2 \log k)$$

There should exists a constant c_{occup}, such that, $\frac{3-e}{9}\lambda k \geq \alpha \log(2SAHk) + \frac{1}{\Delta_{\min}}(C_1 + \alpha C_2 \log k)$ can be satisfied for all $k \geq c_{occup} \frac{C_1 + \alpha C_2}{\lambda \Delta_{\min}} \log(\frac{\alpha C_1 C_2 SAH}{\lambda \Delta_{\min}})$, which implies that:

$$d^*_{\text{Lo}}(s_h, a_h) \geq \frac{\frac{\lambda}{3}k - \frac{3-e}{9}\lambda k}{ek} \geq \frac{\lambda}{9}.$$

On the other hand, if $d^*_{\text{Lo}}(s_h, a_h) \geq \lambda$, on the event $\mathcal{E}_{\textbf{Con},k}$ and Cor. D.7, we have:

$$\begin{aligned} N^k_{\text{Lo},h}(s_h, a_h) \geq & \frac{1}{2}kd^*_{\text{Lo}}(s_h, a_h) - \alpha \log(2SAHk) - \frac{1}{\Delta_{\min}}(C_1 + \alpha C_2 \log k) \\ \geq & \frac{\lambda}{2}k - \alpha \log(2SAHk) - \frac{1}{\Delta_{\min}}(C_1 + \alpha C_2 \log k). \end{aligned}$$

with the same c_{occup}, we have:

$$N^k_{\text{Lo},h}(s_h) \geq N^k_{\text{Lo},h}(s_h, a_h) \geq \frac{\lambda}{2}k - \frac{3-e}{9}\lambda k = \frac{3+2e}{18}\lambda k \geq \frac{\lambda}{3}k.$$

which finishes the proof. □

Lemma D.11 (Convergence Speed of PVI)**.** *There exists an absolute constant c_Ξ, such that for arbitrary fixed $\xi > 0$ and $\lambda > 0$, and for arbitrary*

$$k \geq c_\Xi \max\{\frac{\alpha B_1^2 H^2 S}{\lambda^2 \xi^2} \log(\frac{\alpha HSAB_1 B_2}{\lambda \xi}), \frac{(C_1 + \alpha C_2)SH}{\Delta_{\min} \lambda \xi} \log \frac{C_1 C_2 SAH}{\Delta_{\min} \lambda \xi}\}. \tag{20}$$

on the event $\mathcal{E}_{\textbf{Con},k}, \mathcal{E}_{\textbf{Bonus},k}$ and $\mathcal{E}_{Alg^{Lo},k}$, for arbitrary $h \in [H], s_h \in \mathcal{S}_h$ with $N^k_{Lo,h}(s_h) > \frac{\lambda}{3}$, we have

$$V^*_{Lo,h}(s_h) - \underline{V}^k_{Lo,h}(s_h) \leq \xi.$$

Proof. As a result of Lem. D.10, considering $c_\Xi \geq c_{occup}$, on the events of $\mathcal{E}_{\text{Alg}^{\text{Lo}},k}$ and $\mathcal{E}_{\textbf{Con},k}$, $N^k_{\text{Lo},h}(s_h) > \frac{\lambda}{3}$ implies $d^*_{\text{Lo}}(s_h) \geq \frac{\lambda}{9}$. According to the Lem. D.5, for arbitrary s_h with $d^{\pi^*}_{\text{Lo}}(s_h) > 0$

$$\begin{aligned}
V^*_{\text{Lo},h}(s_h) - \underline{V}^k_{\text{Lo},h}(s_h) \leq & 2\mathbb{E}_{\pi^*_{\text{Lo}},M_{\text{Lo}}}[\sum_{h'=h}^{H} \min\{b^k_{\text{Lo},h'}(s_{h'},\pi^*_{\text{Lo}}),H\}|s_h] \\
=& 2\sum_{h'=h}^{H}\sum_{s_{h'},\pi^*_{\text{Lo}}} d^{\pi^*_{\text{Lo}}}_{\text{Lo}}(s_{h'},\pi^*_{\text{Lo}}|s_h)\min\{b^k_{\text{Lo},h'}(s_{h'},\pi^*_{\text{Lo}}),H\} \\
\leq & \frac{2}{d^{\pi^*_{\text{Lo}}}_{\text{Lo}}(s_h)}\sum_{h'=h}^{H}\sum_{s_{h'},\pi^*_{\text{Lo}}} d^{\pi^*_{\text{Lo}}}_{\text{Lo}}(s_{h'},\pi^*_{\text{Lo}})\min\{b^k_{\text{Lo},h'}(s_{h'},\pi^*_{\text{Lo}}),H\} \\
\leq & \frac{18}{\lambda}\mathbb{E}_{\pi^*_{\text{Lo}},M_{\text{Lo}}}[\sum_{h'=h}^{H}\min\{b^k_{\text{Lo},h'}(s_{h'},\pi^*_{\text{Lo}}),H\}|s_1]
\end{aligned}$$

Given threshold ξ, we define the following set: $\mathcal{Y}^\xi_{\geq h} := \bigcup_{h'\geq h}\{s_h | d^*_{\text{Lo}}(s_h) > \frac{\lambda\xi}{36SH}\}$. Note that for those $s_h \in \mathcal{S}_h \setminus \mathcal{Y}^\xi_{\geq h}$, we have:

$$\sum_{h'=h}^{H}\sum_{s_{h'}\in\mathcal{S}_{h'}\setminus\mathcal{Y}^\xi_{\geq h}} \frac{2}{\lambda}\mathbb{E}_{\pi^*_{\text{Lo}},M_{\text{Lo}}}[\min\{b^k_{\text{Lo},h'}(s_{h'},\pi^*_{\text{Lo}}),H\}|s_1] \leq SH\cdot\frac{18}{\lambda}\cdot\frac{\lambda\xi}{36SH} = \frac{\xi}{2}.$$

In the following, we study the bonus term for $s_h \in \mathcal{Y}^\xi_{\geq h}$. According to Lem. D.9, on the event of $\mathcal{E}_{\text{Alg}^{\text{Lo}},k}$ and $\mathcal{E}_{\textbf{Con},k}$, we have:

$$N^k_{\text{Lo},h}(s_h,a_h) \geq \frac{k}{2}d^*_{\text{Lo}}(s_h,a_h) - \frac{1}{\Delta_{\min}}(C_1 + \alpha C_2\log k) - \alpha\log(2SAHk).$$

We define:

$$k_{s_h} := \arg\min_k \quad s.t. \quad kd^*_{\text{Lo}}(s_h,a_h)/4 \geq \frac{1}{\Delta_{\min}}(C_1+\alpha C_2\log k) + \alpha\log(2SAHk), \quad \forall k' \geq k$$

which implies that $k_{s_h} = c_0\cdot\frac{C_1+\alpha C_2}{\Delta_{\min}d^*_{\text{Lo}}(s_h,a_h)}\log\frac{C_1C_2SAH}{\Delta_{\min}d^*_{\text{Lo}}(s_h,a_h)} \leq c_0'\cdot\frac{(C_1+\alpha C_2)SH}{\Delta_{\min}\lambda\xi}\log\frac{C_1C_2SAH}{\Delta_{\min}\lambda\xi}$ for some absolute constants c_0 and c_0', where the second step we use $d^*_{\text{Lo}}(s_h,a_h) > \lambda\xi/36SH$. Therefore, for $k \geq k_{s_h}$, we have:

$$b^k_{\text{Lo},h}(s_h,a_h) \leq B_1\sqrt{\frac{\log(B_2/\delta_k)}{N^k_{\text{Lo},h}(s_h,a_h)}} \leq 2B_1\sqrt{\frac{\log(B_2/\delta_k)}{kd^*_{\text{Lo}}(s_h,a_h)}}.$$

which implies that:

$$\mathbb{E}_{\pi^*_{\text{Lo}},M_{\text{Lo}}}[\min\{b^k_{\text{Lo},h}(s_h,\pi^*_{\text{Lo}}),H\}|s_1] = 2d^*_{\text{Lo}}(s_h,\pi^*_{\text{Lo}})\cdot B_1\sqrt{\frac{\log(B_2/\delta_k)}{kd^*_{\text{Lo}}(s_h,\pi^*_{\text{Lo}})}} = 2B_1\sqrt{\frac{d^*_{\text{Lo}}(s_h,\pi^*_{\text{Lo}})\log(B_2/\delta_k)}{k}}.$$

Therefore,

$$\begin{aligned}
&\sum_{h'=h}^{H}\sum_{s_{h'}\in\mathcal{Y}^\xi_{\geq h}}\frac{18}{\lambda}\mathbb{E}_{\pi^*_{\text{Lo}},M_{\text{Lo}}}[\min\{b^k_{\text{Lo},h'}(s_{h'},\pi^*_{\text{Lo}}),H\}|s_1] \\
\leq & \frac{36}{\lambda}B_1\sum_{h'=h}^{H}\sum_{s_{h'}\in\mathcal{S}_{h'}\setminus\mathcal{Y}^\xi_{\geq h}}\sqrt{\frac{d^*_{\text{Lo}}(s_{h'},\pi^*_{\text{Lo}})\log(B_2/\delta_k)}{k}} \leq \frac{36}{\lambda}B_1\sqrt{\frac{\log(B_2/\delta_k)}{k}}\sum_{h'=h}^{H}\sum_{s_{h'}\in\mathcal{S}_{h'}\setminus\mathcal{Y}^\xi_{\geq h}}\sqrt{d^*_{\text{Lo}}(s_{h'},\pi^*_{\text{Lo}})} \\
\leq & \frac{36}{\lambda}B_1\sqrt{\frac{\log(B_2/\delta_k)}{k}}\sqrt{SH\sum_{h'=h}^{H}\sum_{s_{h'}\in\mathcal{S}_{h'}\setminus\mathcal{Y}^\xi_{\geq h}} d^*_{\text{Lo}}(s_{h'},\pi^*_{\text{Lo}})} \leq \frac{36}{\lambda}B_1H\sqrt{\frac{S\log(B_2/\delta_k)}{k}}.
\end{aligned}$$

Recall that $\delta_k = O(1/SAHk^\alpha)$, the RHS is less than $\frac{\xi}{2}$ when:

$$k \geq c_0'' \frac{\alpha B_1^2 H^2 S}{\lambda^2 \xi^2} \log(\frac{\alpha HSAB_1B_2}{\lambda\xi}).$$

for some constant c_0''. Therefore, by choosing $c_\Xi = \max\{c_0', c_0''\}$ we can conclude that, as long as:

$$k \geq c_\Xi \max\{\frac{\alpha B_1^2 H^2 S}{\lambda^2 \xi^2} \log(\frac{\alpha HSAB_1B_2}{\lambda\xi}), \frac{(C_1+\alpha C_2)SH}{\Delta_{\min}\lambda\xi} \log \frac{C_1C_2SAH}{\Delta_{\min}\lambda\xi}\}.$$

we have $V^*_{\text{Lo},h}(s_h) - \underline{V}^k_{\text{Lo},h}(s_h) \leq \xi$. □

D.4 Analysis of Regret on M_{Hi}

For the simplification of the notation, in the following, we will denote:

$$\zeta^k(s_h) := \mathbb{I}[\{\underline{Q}^k_{\text{Lo},h}(s_h, \underline{\pi}^k_{\text{Lo},h}) \leq \widetilde{Q}^k_{\text{Hi},h}(s_h, \underline{\pi}^k_{\text{Lo},h}) + \varepsilon\} \cap \{N^k_{\text{Lo},h}(s_h) > \frac{\lambda}{3}k\}].$$

In another word, $\zeta^k(s_h) = 1$ if and only if we will trust M_{Lo} at state s_h in Alg. 2 and therefore set π^k_{Hi} by exploiting information from M_{Lo}. Next, we define the surplus.

Definition D.12 (Definition of Surplus in Pessimistic Algorithm setting)**.** We define the surplus for pessimistic estimation in M_{Lo} and optimistic estimation in M_{Hi}:

$$\mathbf{E}^k_{\text{Hi},h}(s_h, a_h) = \widetilde{Q}^k_{\text{Hi},h}(s_h, a_h) - \mathbb{P}_{\text{Hi},h}\widetilde{V}^k_{\text{Hi},h+1}(s_h, a_h) + \mathbb{P}_{\text{Hi},h}\underline{V}^{\pi^k_{\text{Hi}}}_{\text{Hi},h+1}(s_h, a_h) - \underline{Q}^{\pi^k_{\text{Hi}}}_{\text{Hi},h}(s_h, a_h).$$

We also define:

$$\ddot{\mathbf{E}}^k_{\text{Hi},h}(s_h, a_h) := \text{Clip}\left[\mathbf{E}^k_{\text{Hi},h}(s_h, a_h)\Big| \max\{\frac{\Delta_{\min}}{4eH}, \frac{\Delta_{\text{Hi}}(s_h, a_h)}{4e}\}\right].$$

and

$$\ddot{Q}^\pi_{\text{Hi},h}(s_h, a_h) := r_{\text{Hi},h}(s_h, a_h) + \mathbb{P}_{\text{Hi},h}\ddot{V}^\pi_{\text{Hi},h+1}(s_h, a_h) + e\ddot{\mathbf{E}}^k_{\text{Hi},h}(s_h, a_h), \quad \ddot{V}^\pi_{\text{Hi},h}(s_h) := \ddot{Q}^\pi_{\text{Hi},h}(s_h, \pi).$$

We first show that $\widetilde{V}^k_{\text{Hi},h}$ will be an overestimation eventually.

Theorem 4.3. *There exists $k_{ost} = Poly(S, A, H, \lambda^{-1}, \Delta^{-1}_{\min})$, such that, for all $k \geq k_{ost}$, on some event $\mathcal{E}_k$ with $\mathbb{P}(\mathcal{E}_k) \leq 3\delta_k$, we have $Q^*_{Hi,h}(s_h, a_h) \leq \widetilde{Q}^k_{Hi,h}(s_h, a_h),\ V^*_{Hi,h}(s_h) \leq \widetilde{V}^k_{Hi,h}(s_h), \forall h \in [H],\ s_h \in \mathcal{S}_h,\ a_h \in \mathcal{A}_h$ and*

$$V^*_{Hi,1}(s_1) - V^{\pi^k_{Hi}}_{Hi,1}(s_1) \leq 2e\mathbb{E}_{\pi^k_{Hi}}\left[\sum_{h=1}^H \textit{Clip}\left[\min\{H, 4b^k_{Hi,h}(s_h, a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{Hi}(s_h, a_h)}{4e}\right]\right]. \quad (1)$$

Proof. In this theorem, $\mathcal{E}_k$ denotes the event $\mathcal{E}_{\text{Alg}^{\text{Lo}},k} \cap \mathcal{E}_{\textbf{Bonus},k} \cap \mathcal{E}_{\textbf{Con},k}$.

Part 1: Proof of Overestimation We do the proof by induction. First of all, note that the overestimation is true for horizon $h = H+1$, since all the value is zero. Next, we assume the overestimation is true for step $h+1$, we will show it holds for step h. We first show $\widetilde{Q}^k_{\text{Hi},h}$ is an overestimation:

$$\begin{aligned}
&\widetilde{Q}^k_{\text{Hi},h}(s_h, a_h) - Q^*_{\text{Hi},h}(s_h, a_h) \\
=&\min\{H - Q^*_{\text{Hi},h}(s_h, a_h), \widehat{\mathbb{P}}^k_{\text{Hi},h}\widetilde{V}^k_{\text{Hi},h+1}(s_h, a_h) + b^k_{\text{Hi},h}(s_h, a_h) - \mathbb{P}_{\text{Hi},h}V^*_{\text{Hi},h+1}(s_h, a_h)\} \\
=&\min\{H - Q^*_{\text{Hi},h}(s_h, a_h), (\widehat{\mathbb{P}}^k_{\text{Hi},h} - \mathbb{P}_{\text{Hi},h})\widetilde{V}^k_{\text{Hi},h+1}(s_h, a_h) + b^k_{\text{Hi},h}(s_h, a_h) + \mathbb{P}_{\text{Hi},h}(\widetilde{V}^k_{\text{Hi},h+1} - V^*_{\text{Hi},h+1})(s_h, a_h)\} \\
\geq&\min\{H - Q^*_{\text{Hi},h}(s_h, a_h), \mathbb{P}_{\text{Hi},h}(\widetilde{V}^k_{\text{Hi},h+1} - V^*_{\text{Hi},h+1})(s_h, a_h)\} \geq 0
\end{aligned}$$

where the first inequality is because of event $\mathcal{E}_{\textbf{Bonus},k}$. In the following, we separate to three cases:

Case 1: $\zeta^k(s_h, a_h) = 0$ In this case,

$$\widetilde{V}^k_{\text{Lo},h}(s_h) = \max_a \widetilde{Q}^k_{\text{Hi},h}(s_h, a) \geq \widetilde{Q}^k_{\text{Hi},h}(s_h, \pi^*_{\text{Hi}}) \geq Q^*_{\text{Hi},h}(s_h, \pi^*_{\text{Hi}}) = V^*_{\text{Hi},h}(s_h).$$

Case 2: $\zeta^k(s_h, a_h) = 1$ **and** $\underline{\pi}^k_{\mathbf{Lo},h}(s_h) = \pi^*_{\mathbf{Hi},h}(s_h)$ In this case, as a result of Lem. D.5, we have:

$$\widetilde{V}^k_{\mathrm{Lo},h}(s_h) = \widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Hi}}) + \frac{1}{H}(\widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Hi}}) - \underline{Q}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Hi}})) \geq \widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Hi}}) \geq Q^*_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Hi}}) = V^*_{\mathrm{Hi},h}(s_h).$$

Case 3: $\zeta^k(s_h, a_h) = 1$ **and** $\underline{\pi}^k_{\mathbf{Lo},h}(s_h) \neq \pi^*_{\mathbf{Hi},h}(s_h)$ This case is more complicated. Intuitively, we want to show that after k_{ost}, in case 3, the "uncerntainty" must be high, and therefore, adding $1/H$ of the uncerntainty interval will ensure the overestimation.

As a result of Lem. D.11, we choose k_{ost} by plugging $\xi = \frac{\Delta_{\min}}{4(H+1)}$ into Eq. (20), which yields:

$$k_{ost} := c_{ost} \cdot \max\{\alpha \frac{B_1^2 H^2 S}{\lambda^2 \Delta^2_{\min}} \log(\alpha HSAB_1B_2), \frac{(C_1 + \alpha C_2)SH}{\lambda \Delta^2_{\min}} \log \frac{C_1 C_2 SAH}{\Delta_{\min}\lambda}\}, \tag{21}$$

for some constant c_{ost}. Then, for arbitrary $k \geq k_{ost}$, on the event of $\mathcal{E}_{\mathbf{Bonus},k}, \mathcal{E}_{\mathbf{Con},k}, \mathcal{E}_{\mathrm{Alg}^{\mathrm{Lo}},k}$, case 3 implies that:

$$V^*_{\mathrm{Lo},h}(s_h) - \underline{V}^k_{\mathrm{Lo},h}(s_h) \leq \frac{\Delta_{\min}}{4(H+1)}.$$

Combining with Lem. D.5, it directly implies that $\underline{\pi}^k_{\mathrm{Lo},h}(s_h) = \pi^*_{\mathrm{Lo},h}(s_h)$. Therefore, in the following, we directly use $\pi^*_{\mathrm{Lo},h}$ to refer $\underline{\pi}^k_{\mathrm{Lo},h}(s_h)$. Then first observation is that, under Cond. B, in this case, we have:

$$\begin{aligned}
V^*_{\mathrm{Hi},h}(s_h) - \widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) \leq & V^*_{\mathrm{Hi},h}(s_h) - \underline{Q}^k_{\mathrm{Lo},h}(s_h, \pi^*_{\mathrm{Lo},h}) + \varepsilon \\
= & V^*_{\mathrm{Hi},h}(s_h) - \underline{V}^k_{\mathrm{Lo},h}(s_h) + \varepsilon \\
\leq & V^*_{\mathrm{Lo},h}(s_h) + \frac{\Delta_{\min}}{2(H+1)} - \underline{V}^k_{\mathrm{Lo},h}(s_h) + \varepsilon \\
= & V^*_{\mathrm{Lo},h}(s_h) - \underline{V}^k_{\mathrm{Lo},h}(s_h) + \frac{\Delta_{\min}}{2(H+1)} + \varepsilon \\
\leq & \frac{\Delta_{\min}}{H+1} \leq \frac{1}{H+1}(V^*_{\mathrm{Hi},h}(s_h) - Q^*_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}))
\end{aligned}$$

which implies that:

$$\begin{aligned}
V^*_{\mathrm{Hi},h}(s_h) \leq & \widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) + \frac{1}{H}(\widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) - Q^*_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) \\
\leq & \widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) + \frac{1}{H}(\widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) - \underline{Q}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Lo}}) \qquad \text{(Lem. D.5)} \\
= & \widetilde{V}^k_{\mathrm{Lo},h}(s_h)
\end{aligned}$$

which finishes the proof for overestimation.

Part 2: Proof for Eq. (1) The following proof relies on Lem. D.13 and Lem. D.14, whose proofs we just provide after finishing the proof for this theorem. The first observation is, for arbitrary policy π_{Hi},

$$\begin{aligned}
V^*_{\mathrm{Hi},h}(s_h) - V^{\pi_{\mathrm{Hi}}}_{\mathrm{Hi},h}(s_h) = & Q^*_{\mathrm{Hi},h}(s_h, \pi^*_{\mathrm{Hi}}) - Q^*_{\mathrm{Hi},h}(s_h, \pi_{\mathrm{Hi}}) \\
= & \Delta_{\mathrm{Hi}}(s_h, \pi_{\mathrm{Hi}}) + Q^*_{\mathrm{Hi},h}(s_h, \pi_{\mathrm{Hi}}) - Q^*_{\mathrm{Hi},h}(s_h, \pi_{\mathrm{Hi}}) \\
= & \Delta_{\mathrm{Hi}}(s_h, \pi_{\mathrm{Hi}}) + \mathbb{P}_{\mathrm{Hi},h}(V^*_{\mathrm{Hi},h+1} - V^{\pi_{\mathrm{Hi}}}_{\mathrm{Hi},h+1})(s_h, \pi_{\mathrm{Hi}}) \\
= & ... \\
= & \mathbb{E}_{\pi_{\mathrm{Hi}}, M_{\mathrm{Hi}}}[\sum_{h'=h}^{H} \Delta_{\mathrm{Hi}}(s_{h'}, a_{h'})|s_h].
\end{aligned} \tag{22}$$

Besides, according to the definition of $\ddot{\mathbf{E}}_{\mathrm{Hi},h}$, we have:

$$\ddot{V}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(s_h) - V^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(s_h) = e\mathbb{E}_{\pi^k_{\mathrm{Hi}}}[\sum_{h'=h}^{H} \ddot{\mathbf{E}}^k_{\mathrm{Hi}h'}(s_{h'}, a_{h'})|s_h]$$

$$\geq e\mathbb{E}_{\pi_{\text{Hi}}^k}[\sum_{h'=h}^{H}\mathbf{E}_{\text{Hi}h'}^k(s_{h'},a_{h'}) - \varepsilon_{\text{Clip}} - \frac{\Delta_{\text{Hi}}(s_{h'},a_{h'})}{4e}|s_h]$$

$$\geq e\mathbb{E}_{\pi_{\text{Hi}}^k}[\sum_{h'=h}^{H}\mathbf{E}_{\text{Hi}h'}^k(s_{h'},a_{h'}) - \frac{\Delta_{\text{Hi}}(s_{h'},a_{h'})}{4e}|s_h] - eH\cdot\frac{\Delta_{\min}}{4eH}.$$

$$\geq\widetilde{V}_{\text{Hi},h}^k(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h) - \frac{1}{4}(V_{\text{Hi},h}^*(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)) - \frac{\Delta_{\min}}{4}. \quad \text{(Eq. (22) and Lem. D.14)}$$

$$\geq\frac{3}{4}(V_{\text{Hi},h}^*(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)) - \frac{\Delta_{\min}}{4}. \quad \text{(Overestimation)}$$

If $\pi_{\text{Hi}}^k(s_h)\neq\pi_{\text{Hi}}^*(s_h)$, since $V_{\text{Hi},h}^*(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)\geq\Delta_{\text{Hi}}(s_h,\pi_{\text{Hi}}^k)$, we further have:

$$\ddot{V}_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)\geq\frac{1}{2}(V_{\text{Hi},h}^*(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)) + \frac{\Delta_{\text{Hi}}(s_h,\pi_{\text{Hi}}^k)}{4} - \frac{\Delta_{\min}}{4}\geq\frac{1}{2}(V_{\text{Hi},h}^*(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)).$$

otherwise,

$$\begin{aligned}\ddot{V}_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h) =& \ddot{\mathbf{E}}_{\text{Hi},h}(s_h,\pi_{\text{Hi}}^k) + \mathbb{E}_{\pi_{\text{Hi}}^k,M_{\text{Hi}}}[\ddot{V}_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k}(s_{h+1}) - V_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k}(s_{h+1})|s_h]\\ \geq&\mathbb{E}_{\pi_{\text{Hi}}^k,M_{\text{Hi}}}[\ddot{V}_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k}(s_{h+1}) - V_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k}(s_{h+1})|s_h].\end{aligned}$$

Therefore, we have:

$$\begin{aligned}\ddot{V}_{\text{Hi},1}^{\pi_{\text{Hi}}^k}(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^k}(s_1)\geq&\mathbb{E}_{\pi_{\text{Hi}}^k,M_{\text{Hi}}}[\sum_{h=1}^{H}\mathbb{I}[\widetilde{\mathcal{E}}_{\text{Hi},h}^k](\ddot{V}_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h))]\\ \geq&\frac{1}{2}\mathbb{E}_{\pi_{\text{Hi}}^k,M_{\text{Hi}}}[\sum_{h=1}^{H}\mathbb{I}[\widetilde{\mathcal{E}}_{\text{Hi},h}^k](V_{\text{Hi},h}^*(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h))]\\ =&\frac{1}{2}(V_{\text{Hi},h}^*(s_1) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_1)).\end{aligned}$$

Combining Lem. D.13 and the definition of $\ddot{V}_{\text{Hi},1}^{\pi_{\text{Hi}}^k}(s_1)$, we can finish the proof for Eq. (1). □

Lemma D.13 (Upper and lower bounds of the surplus)**.** *For arbitrary k, on the event of $\mathcal{E}_{\textbf{Bonus},k}$, we have:*

$$\forall h\in[H],\ \forall s_h\in\mathcal{S}_h,a_h\in\mathcal{A}_h,\quad 0\leq\mathbf{E}_{Hi,h}^k(s_h,a_h)\leq\min\{H,4b_{Hi,h}^k(s_h,a_h)\}$$

Proof. According to Alg. 2, we should have $\widetilde{Q}_{\text{Hi},h}^k,\ \widetilde{V}_{\text{Hi},h+1}^k,\ \underline{V}_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k},\ \underline{Q}_{\text{Hi},h}^{\pi_{\text{Hi}}^k}\in[0,H]$. By Lem. D.5 and Thm. 4.3, we also have $\widetilde{V}_{\text{Hi},h+1}^k(\cdot)\geq\underline{V}_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k}(\cdot)$, which implies that $\mathbf{E}_{\text{Hi},h}^k(\cdot,\cdot)\leq H$. Besides,

$$\mathbf{E}_{\text{Hi},h}^k(s_h,a_h) = (\widehat{\mathbb{P}}_{\text{Hi},h}^k - \mathbb{P}_{\text{Hi},h})\widetilde{V}_{\text{Hi},h+1}^k(s_h,a_h) + (\mathbb{P}_{\text{Hi},h} - \widehat{\mathbb{P}}_{\text{Hi},h}^k)\underline{V}_{\text{Hi},h+1}^{\pi_{\text{Hi}}^k}(s_h,a_h) + 2b_{\text{Hi},h}^k(s_h,a_h).$$

On the event of $\mathcal{E}_{\textbf{Bonus},k}$, we have $0\leq\mathbf{E}_{\text{Hi},h}^k(s_h,a_h)\leq 4b_{\text{Hi},h}^k(s_h,a_h)$, which finishes the proof. □

Lemma D.14 (Relationship between surplus and overestimation gap)**.** *Under the same condition of Thm. 4.3,*

$$\widetilde{V}_{Hi,h}^k(s_h) - V_{Hi,h}^{\pi_{Hi}^k}(s_h)\leq e\mathbb{E}_{\pi_{Hi}^k,M_{Hi}}[\sum_{h'=h}^{H}\mathbf{E}_{Hi,h}^k(s_h,\pi_{Hi}^k)|s_h].$$

Proof.

$$\widetilde{V}_{\text{Hi},h}^k(s_h) - V_{\text{Hi},h}^{\pi_{\text{Hi}}^k}(s_h)\leq\widetilde{V}_{\text{Hi},h}^k(s_h) - \underline{V}_{\text{Hi},h}^{\pi_{\text{Hi},k}}(s_h) \quad \text{(Lem. D.5)}$$

$$\leq(1+\frac{1}{H})(\widetilde{Q}_{\text{Hi},h}^k(s_h,\pi_{\text{Hi}}^k) - \underline{Q}_{\text{Hi},h}^{\pi_{\text{Hi},k}}(s_h,\pi_{\text{Hi}}^k)) \quad \text{(Update rule in Alg. 2 and } \widetilde{Q}_{\text{Hi},h}^k\geq\underline{Q}_{\text{Hi},h}^{\pi_{\text{Hi},k}})$$

$$
\begin{aligned}
=&(1+\frac{1}{H})(\widetilde{Q}^k_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}})-\mathbb{P}_{\mathrm{Hi},h}\widetilde{V}^k_{\mathrm{Hi},h+1}(s_h,\pi^k_{\mathrm{Hi}})+\mathbb{P}_{\mathrm{Hi},h}\underline{V}^k_{\mathrm{Hi},h+1}(s_h,\pi^k_{\mathrm{Hi}})-\underline{Q}^{\pi_{\mathrm{Hi},k}}_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}}))\\
&+(1+\frac{1}{H})\mathbb{P}_{\mathrm{Hi},h}(\widetilde{V}^k_{\mathrm{Hi},h+1}-\underline{V}^k_{\mathrm{Hi},h+1})(s_h,\pi^k_{\mathrm{Hi}})\\
=&(1+\frac{1}{H})\mathbf{E}^k_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}})+(1+\frac{1}{H})\mathbb{P}_{\mathrm{Hi},h}(\widetilde{V}^k_{\mathrm{Hi},h+1}-\underline{V}^k_{\mathrm{Hi},h+1})(s_h,\pi^k_{\mathrm{Hi}})\\
\leq&...\\
\leq&\mathbb{E}_{\pi^k_{\mathrm{Hi}},M_{\mathrm{Hi}}}[\sum_{h'=h}^{H}(1+\frac{1}{H})^{h'-h+1}\mathbf{E}^k_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}})]\\
\leq&e\mathbb{E}_{\pi^k_{\mathrm{Hi}},M_{\mathrm{Hi}}}[\sum_{h'=h}^{H}\mathbf{E}^k_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}})].
\end{aligned}
$$

□

In the following lemma, we show the benefits of transfer action between similar states.

Lemma D.15 (Benefits on Similar States)**.** *If s_h in M_{Hi} is ε-close to s_h in M_{Lo}, i.e. satisfying the property in Def. 2.1, and $d^*_{Lo}(s_h)\geq\lambda$, where λ is the hyper-parameter in Alg. 2, then, for arbitrary $k\geq k_{occup}:=c_{occup}\frac{C_1+\alpha C_2}{\lambda\Delta_{\min}}\log(\frac{\alpha C_1C_2SAH}{\lambda\Delta_{\min}})$, on the events of $\mathcal{E}_{Alg^{Lo},k},\mathcal{E}_{\textbf{Con},k},\mathcal{E}_{\textbf{Bonus},k}$, we have $\pi^k_{Hi}(s_h)=\pi^*_{Hi}(s_h)$.*

Proof. As a result of Lem. D.10, we have, for arbitrary s_h with $d^*_{\mathrm{Lo}}(s_h)=d^*_{\mathrm{Lo}}(s_h,\pi^*_{\mathrm{Lo}})\geq\lambda$, after $k\geq c_{occup}\frac{C_1+\alpha C_2}{\lambda\Delta_{\min}}\log(\frac{\alpha C_1C_2SAH}{\lambda\Delta_{\min}})$, we should have:

$$
N^k_{\mathrm{Lo},h}(s_h)\geq N^k_{\mathrm{Lo},h}(s_h,\pi^*_{\mathrm{Lo}})\geq\frac{\lambda}{3}k.
$$

On the events of $\mathcal{E}_{\textbf{Bonus},k}$, and the value dominance condition Def. B, we also have:

$$
\underline{Q}^k_{\mathrm{Lo},h}(s_h,\underline{\pi}^k_{\mathrm{Lo},h})\leq\widetilde{Q}^k_{\mathrm{Hi},h}(s_h,\underline{\pi}^k_{\mathrm{Lo},h})+\varepsilon.
$$

which implies that the algorithm will choose the "trust" branch and choose $\pi^k_{\mathrm{Hi},h}(s_h)=\underline{\pi}^k_{\mathrm{Lo},h}(s_h)$.

On the other hand, if there is another $a_h\neq\pi^*_{\mathrm{Lo}}(s_h)$ satisfying $N^k_{\mathrm{Lo},h}(s_h,a_h)\geq\frac{\lambda}{3}k$, by applying Lem. D.10 again, we can make a contradication and therefore, we must have

$$
N^k_{\mathrm{Lo},h}(s_h,\pi^*_{\mathrm{Lo}})\geq\frac{\lambda}{3}k>N^k_{\mathrm{Lo},h}(s_h,a_h),\quad\forall a_h\neq\pi^*_{\mathrm{Lo}}(s_h).
$$

which implies that $\pi^k_{\mathrm{Hi}}(s_h)=\pi^*_{\mathrm{Lo}}(s_h)$ in Alg. 2. Given that s_h is ε-close between M_{Hi} and M_{Lo}, we directly have $\pi^k_{\mathrm{Hi}}(s_h)=\pi^*_{\mathrm{Hi}}(s_h)$. □

Theorem D.16 (Detailed Version of Thm.4.2)**.** *Under Assump. A, B and C, Cond. D.1 for Alg^{Lo} and Cond. D.3 for **Bonus** function, by running Alg. 2 with $\varepsilon=\frac{\widetilde{\Delta}_{\min}}{4(H+1)}$, $\alpha>2$, an any $\lambda>0$, we have*

$$
\begin{aligned}
\textit{Regret}_K(M_{Hi})=&O\Big(H\cdot\max\{\alpha\frac{S^3H^4}{\lambda^2\Delta^2_{\min}}\log(\alpha SAH),\frac{(C_1+\alpha C_2)SH}{\lambda\Delta^2_{\min}}\log\frac{C_1C_2SAH}{\Delta_{\min}\lambda}\}\\
&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{C}^*_h}\frac{SH^2}{\Delta_{\min}}\log(SAH(K\wedge\frac{1}{\lambda\Delta_{\min}d^*_{Hi}(s_h)}))\\
&+SH\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}^\lambda_h}(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{Hi}(s_h,a_h)})\log(SAHK)\Big)\\
=&O\Big(SH\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}^\lambda_h}(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{Hi}(s_h,a_h)})\log(SAHK)\Big).
\end{aligned}
\tag{23}
$$

Proof. We consider $k_{start} := \max\{k_{ost}, k_{occup}\}$. We first study the regret part after $k \geq k_{start}$:

$$\begin{aligned}
&\mathbb{E}[\sum_{k=k_{start}+1}^{K} V_{\text{Hi},1}^{*}(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^k}(s_1)] \\
\leq &\sum_{k=k_{start}+1}^{K} 2e\mathbb{E}_{\pi_{\text{Hi}}^k}[\sum_{h=1}^{H} \text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},k} \cap \mathcal{E}_{\textbf{Con},k}]] \\
&+ \sum_{k=k_{start}+1}^{K} H \cdot \Pr(\mathcal{E}_{\textbf{Bonus},k}^{\complement} \cup \mathcal{E}_{\text{Alg}^{\text{Lo}},k}^{\complement} \cup \mathcal{E}_{\textbf{Con},k}^{\complement}).
\end{aligned}$$

Since the failure rate for events $\mathcal{E}_{\textbf{Bonus},k}, \mathcal{E}_{\text{Alg}^{\text{Lo}},k}, \mathcal{E}_{\textbf{Con},k}$ is only at the level of $k^{-\Theta(\alpha)}$, the second part is constant, and we mainly focus on the first term. For all state action (s_h, a_h), and for all $k \geq k_{start}$, we have:

$$\begin{aligned}
&\mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},k} \cap \mathcal{E}_{\textbf{Con},k}]] \\
=&d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)\text{Clip}\left[\min\{H, 4B_1\sqrt{\frac{\log(B_2k^{\alpha})}{N_{\text{Hi},h}^k(s_h,a_h)}}\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \\
\leq&d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)\text{Clip}\left[\min\{H, 4B_1\sqrt{\frac{\alpha\log(B_2K)}{N_{\text{Hi},h}^k(s_h,a_h)}}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \qquad (24)
\end{aligned}$$

Under the event of $\mathcal{E}_{\textbf{Con},k}$, as a result of Lem. D.6, we have:

$$N_{\text{Hi},h}^k(s_h,a_h) \geq \frac{1}{2}\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^{k'}}(s_h,a_h) - \alpha\log(2SAHk) \geq \frac{1}{2}\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^{k'}}(s_h,a_h) - \alpha\log(2SAHK)$$

We denote $\tau_{s_h,a_h}^K := \min_k \ s.t.\ \forall k' \geq k,\ \frac{1}{4}\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) \geq \alpha\log(2SAHK)$. Then we have:

$$(24) \leq d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)\text{Clip}\left[\min\{H, 8B_1\sqrt{\frac{\alpha\log(B_2K)}{\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)}}\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right], \quad \forall k \geq \tau_{s_h,a_h}^K.$$

Therefore, for arbitrary s_h, a_h, there exists an absolute constant c_{s_h,a_H}, such that:

$$\begin{aligned}
&\sum_{k=k_{start}+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]] \\
\leq &H \cdot \sum_{k=1}^{\tau_{s_h,a_h}^K} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) + \sum_{k=\tau_{s_h,a_h}^K+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4B_1\sqrt{\frac{\alpha\log(B_2K)}{N_{\text{Hi},h}^k(s_h,a_h)}}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]] \\
\leq &c_{s_h,a_h}H\log(2SAHK)) + \sum_{k=\tau_{s_h,a_h}^K+1}^{K} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)\text{Clip}\left[\min\{H, 8B_1\sqrt{\frac{\alpha\log(B_2K)}{\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)}}\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \\
\leq &c_{s_h,a_h}H\log(2SAHK) + c_{s_h,a_h} \cdot \int_{\alpha\log(2SAHK)}^{K/4} \text{Clip}\left[B_1\sqrt{\frac{\alpha\log(B_2K)}{x}}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] dx \\
= &O\left(H\log(2SAHK) + B_1(\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(B_2K)\right).
\end{aligned}$$

As a result, we can establish the following regret upper bound (note that $k_{start} = O(k_{ost})$):

$$\mathbb{E}[\sum_{k=1}^{K} V_{\text{Hi},1}^{*}(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^k}(s_1)]$$

$$
\begin{aligned}
&\leq \sum_{k=k_{start}+1}^{K} 2e\mathbb{E}_{\pi_{\text{Hi}}^k}[\sum_{h=1}^{H} \text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},k} \cap \mathcal{E}_{\textbf{Con},k}]] \\
&\quad + Hk_{start} + \sum_{k=k_{start}+1}^{K} H \cdot \Pr(\mathcal{E}^{\complement}_{\textbf{Bonus},k} \cup \mathcal{E}^{\complement}_{\text{Alg}^{\text{Lo}},k} \cup \mathcal{E}^{\complement}_{\textbf{Con},k}) \\
&= O\left(k_{start} \cdot H + SAH^2 \log(2SAHK) + B_1 \log(B_2 K) \sum_{h=1}^{H} \sum_{s_h,a_h} (\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\right).
\end{aligned} \tag{25}
$$

As introduced in maintext, because of the knowledge transfer, we may expect to achieve constant regret on some special state action pairs, and we analyze them in the following.

Type 1: $(s_h,a_h) \in \mathcal{C}_h^{\lambda,1} \cup \mathcal{C}_h^{\lambda,2}$: Constant Regret because of Low Visitation Probability As discussed in Lem. D.15, for $s_h \in \mathcal{Z}_h^{\varepsilon,\lambda}$, since $k_{start} \geq k_{occup} := c_{occup}\frac{C_1+\alpha C_2}{\lambda\Delta_{\min}}\log(\frac{\alpha C_1 C_2 SAH}{\lambda\Delta_{\min}})$, on the event of $\mathcal{E}_{\text{Alg}^{\text{Lo}},k}, \mathcal{E}_{\textbf{Bonus},k}$ and $\mathcal{E}_{\textbf{Con},k}$, we have $\forall a_h \neq \pi_{\text{Hi}}^*(s_h),\ d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) = 0$. Moreover, according to the definition of $\mathcal{C}_h^{\lambda,2}$, we also have $\forall (s_h,a_h) \in \mathcal{C}_h^{\lambda,2},\ d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) = 0$. Therefore, for all $h \in [H]$ and $(s_h,a_h) \in \mathcal{C}_h^{\lambda,1} \cup \mathcal{C}_h^{\lambda,2}$:

$$
\sum_{k=k_{start}+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},k} \cap \mathcal{E}_{\textbf{Con},k}]] = 0.
$$

Type 2: $(s_h,a_h) \in \mathcal{C}_h^*$, i.e. $d_{\textbf{Hi}}^*(s_h,a_h) = d_{\textbf{Hi}}^*(s_h) > 0$ Because of the sub-linear regret in Eq. (25), we may expect that $N_{\text{Hi},h}^k(s_h,a_h) \approx \sum_{k=1}^{K} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) \sim O(Kd_{\text{Hi}}^*(s_h,a_h))$ when K is large enough. To see this, note that $\sum_{\tilde{k}=1}^{k}(V_{\text{Hi},1}^*(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^{\tilde{k}}}(s_1)) - \mathbb{E}[\sum_{\tilde{k}=1}^{k} V_{\text{Hi},1}^*(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^{\tilde{k}}}(s_1)]$ is a martingale difference sequence with bounded difference. We define $\mathcal{E}_{\text{Alg}^{\text{Hi}},k} := \{\sum_{\tilde{k}=1}^{k} V_{\text{Hi},1}^*(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^{\tilde{k}}}(s_1) \geq H\sqrt{2\alpha k\log k} + \mathbb{E}[\sum_{\tilde{k}=1}^{k} V_{\text{Hi},1}^*(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^{\tilde{k}}}(s_1)]\}$, according to the Azuma-Hoeffding inequality, we have:

$$
\Pr(\mathcal{E}_{\text{Alg}^{\text{Hi}},k}) \leq \exp(\frac{-2\alpha H^2 k\log k}{2kH^2}) \leq \frac{1}{k^\alpha}.
$$

On the event of $\mathcal{E}_{\textbf{Bonus},k}, \mathcal{E}_{\text{Alg}^{\text{Lo}},k}, \mathcal{E}_{\textbf{Con},k}$ and $\mathcal{E}_{\text{Alg}^{\text{Hi}},k}$, as a result of Thm. D.6, we have:

$$
|\sum_{\tilde{k}=1}^{k} d^{\pi_{\text{Hi}}^{\tilde{k}}}(s_h,a_h) - kd_{\text{Hi}}^*(s_h,a_h)| \leq H\sqrt{2\alpha k\log k} + (25)
$$

To make sure $\sum_{\tilde{k}=1}^{k} d^{\pi_{\text{Hi}}^{\tilde{k}}}(s_h,a_h) \geq \frac{k}{2}d_{\text{Hi}}^*(s_h,a_h)$, we expect:

$$
\frac{k}{2}d_{\text{Hi}}^*(s_h,a_h) \geq H\sqrt{2\alpha k\log k} + (25)
$$

which can be satisfied by:

$$
k \geq \bar{\tau}_{s_h} := \bar{c}_{s_h}^* \frac{1}{(d_{\text{Hi}}^*(s_h))^2}\text{Poly}(S,A,H,\lambda^{-1},\Delta_{\min}^{-1})
$$

for some constant $\bar{c}_{s_h}^*$. As a result, for $k \geq \max\{\tau_{s_h,a_h}^K, \bar{\tau}_{s_h}\} + 1$, we should have:

$$
\begin{aligned}
&\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \\
\leq &\text{Clip}\left[\min\{H, 8B_1\sqrt{\frac{\alpha\log(kB_2)}{\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h)}}\}\Big|\frac{\Delta_{\min}}{4eH}\right] \leq \text{Clip}\left[8B_1\sqrt{\frac{2\alpha\log(kB_2)}{kd_{\text{Hi}}^*(s_h,a_h)}}\Big|\frac{\Delta_{\min}}{4eH}\right].
\end{aligned} \tag{$\Delta_{\text{Hi}}(s_h,a_h) = 0$}
$$

Note that $8B_1\sqrt{\frac{2\alpha\log(kB_2)}{kd^*_{\text{Hi}}(s_h,a_h)}} = 8B_1\sqrt{\frac{2\alpha\log(kB_2)}{kd^*_{\text{Hi}}(s_h)}} \le \frac{\Delta_{\min}}{4eH}$ can be satisfied when $k \ge \tau'_{s_h} := c'_{s_h}\frac{\alpha H^2B_1^2}{\Delta^2_{\min}d^*_{\text{Hi}}(s_h)}\log(\frac{\alpha B_1B_2H}{\Delta_{\min}d^*_{\text{Hi}}(s_h)})$ for some absolute constant c'_{s_h}. Therefore, $\min\{H, 4b^k_{\text{Hi},h}(s_h,a_h)\} \le \frac{\Delta_{\min}}{4eH}$ and the regret will not increase after $k \ge \tau^*_{s_h,a_h} := c^*_{s_h,a_h}\max\{k_{start}, \tau_{s_h,a_h}, \bar{\tau}_{s_h}, \tau'_{s_h}\} = \text{Poly}(S,A,H,\lambda^{-1},\Delta^{-1}_{\min},(d^*_{\text{Hi}}(s_h))^{-1})$, for some absolute constant $c^*_{s_h,a_h}$, which implies that,

$$\sum_{k=k_{start}+1}^{K}\mathbb{E}_{\pi^k_{\text{Hi}}}[\text{Clip}\left[\min\{H, 4b^k_{\text{Hi},h}(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH}\vee\frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]\mathbb{I}[\mathcal{E}_{\textbf{Bonus},k}\cap\mathcal{E}_{\text{Alg}^{\text{Lo}},k}\cap\mathcal{E}_{\textbf{Con},k}]]$$
$$=\sum_{k=k_{start}}^{\tau^*_{s_h,a_h}}\mathbb{E}_{\pi^k_{\text{Hi}}}[\text{Clip}\left[\min\{H, 4b^k_{\text{Hi},h}(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH}\vee\frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]\mathbb{I}[\mathcal{E}_{\textbf{Bonus},k}\cap\mathcal{E}_{\text{Alg}^{\text{Lo}},k}\cap\mathcal{E}_{\textbf{Con},k}]]$$
$$=O\left(\frac{HB_1}{\Delta_{\min}}\log(SAHB_2\min\{K,\frac{1}{\lambda\Delta_{\min}d^*_{\text{Hi}}(s_h)}\})\right).$$

where in last step, we use the fact that for $\Delta_{\text{Hi}}(s_h,a_h)=0$:

$$O\Big(H\log(2SAHK)+B_1\log(B_2K)\cdot(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\Big)=O\Big(\frac{HB_1}{\Delta_{\min}}\log(SAHB_2K)\Big).$$

As a summary, recall $k_{start}=\max\{k_{ost},k_{occup}\}$, we have:

$$\begin{aligned}\text{Regret}_K(M_{\text{Hi}}) =&O\Big(H\cdot\max\{\alpha\frac{B_1^2H^2S}{\lambda^2\Delta^2_{\min}}\log(\alpha HSAB_1B_2),\frac{(C_1+\alpha C_2)SH}{\lambda\Delta^2_{\min}}\log\frac{C_1C_2SAH}{\Delta_{\min}\lambda}\}\\&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{C}^*_h}\frac{HB_1}{\Delta_{\min}}\log(SAHB_2(K\wedge\frac{1}{\lambda\Delta_{\min}d^*_{\text{Hi}}(s_h)}))\\&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}^\lambda_h}H\log(2SAHK)+B_1(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(B_2K)\Big).\end{aligned}$$

By considering $B_1=O(SH)$ and $B_2=O(SA)$ in Example D.4, and omitting all the constant terms independent w.r.t. K, we can rewrite the above upper bound to:

$$\begin{aligned}\text{Regret}_K(M_{\text{Hi}}) =&O\Big(H\cdot\max\{\alpha\frac{S^3H^4}{\lambda^2\Delta^2_{\min}}\log(\alpha SAH),\frac{(C_1+\alpha C_2)SH}{\lambda\Delta^2_{\min}}\log\frac{C_1C_2SAH}{\Delta_{\min}\lambda}\}\\&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{C}^*_h}\frac{SH^2}{\Delta_{\min}}\log(SAH(K\wedge\frac{1}{\lambda\Delta_{\min}d^*_{\text{Hi}}(s_h)}))\\&+SH\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}^\lambda_h}(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(SAHK)\Big)\\=&O\Big(SH\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}^\lambda_h}(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(SAHK)\Big).\end{aligned}$$

□

E Proofs for Tiered MAB with Multiple Source/Low-Tier Tasks

Lemma 5.3. *[Absorbing to Similar Task] Under Assump. A, B and C, there exists a constant c^*, s.t., if there exists at least one $w^*\in[W]$ such that M_{Lo,w^*} is $\frac{\widetilde{\Delta}_{\min}}{4}$-close to M_{Hi}, by running Alg. 3 with $\varepsilon=\frac{\widetilde{\Delta}_{\min}}{4}$ and $\alpha>2$, for any $k\ge k^*:=c^*\frac{\alpha A}{\Delta^2_{\min}}\log\frac{\alpha AW}{\Delta_{\min}}$, we have $\Pr(\pi^k_{Hi}\neq i^*_{Hi})=O(\frac{A}{k^{2\alpha-2}})$.*

Proof. In the following, we denote $\mathcal{W}^* := \{w \in [W] | i^*_{\text{Lo},w} = i^*_{\text{Hi}},\ \mu_{\text{Lo}}(i^*_{\text{Hi},w^*}) \le \mu_{\text{Hi}}(i^*_{\text{Hi}}) + \frac{\widetilde{\Delta}_{\min}}{4}\}$. $\widetilde{\mathcal{W}}^* := \{w \in [W] | i^*_{\text{Lo},w} = i^*_{\text{Hi}}\}$. In another word, $\mathcal{W}^*$ includes all transferable tasks, while $\widetilde{\mathcal{W}}^*$ includes all the tasks which share the optimal action i^*_{Hi} regardless of whether the value function are close enough or not.

Consider the event $\mathcal{E} := \{\exists k' \in [\frac{k}{2}, k],\ s.t.\ \mathcal{W}^* \cap \mathcal{I}^{k'} = \emptyset\} \cup \{\exists k' \in [\frac{k}{2}, k],\ \exists w \in [W],\ s.t.\ \underline{\pi}^{k'}_{\text{Lo},w} \neq i^*_{\text{Hi}}\}$, note that on its complement: $\mathcal{E}^{\complement} := \{\forall k' \in [\frac{k}{2}, k],\ \mathcal{W}^* \cap \mathcal{I}^{k'} \neq \emptyset\} \cap \{\forall k' \in [\frac{k}{2}, k],\ \forall w \in [W],\ \underline{\pi}^{k'}_{\text{Lo},w} = i^*_{\text{Lo},w}\}$, if $\pi^k_{\text{Hi}} \neq i^*_{\text{Hi}}$ still happens, we must have: $\{\forall k' \in [\frac{k}{2}, k],\ \exists w \notin \widetilde{\mathcal{W}}^*,\ w \in \mathcal{I}^{k'},\ w^{k'} = w\}$. That's because, if $\mathcal{E}^{\complement}$ holds, and $\mathcal{I}^{k'} \subset \widetilde{\mathcal{W}}^*$ for some k', no matter which task in $\widetilde{\mathcal{W}}^*$ is chosen as $w^{k'}$, for all $\widetilde{k} \in [k', k]$, no matter whether $w^{\widetilde{k}}$ changes or not, the action we transfer is always i^*_{Hi} (i.e. $\pi^{\widetilde{k}}_{\text{Hi}} = i^*_{\text{Hi}}$), because of the action inheritance startegy. Therefore,

$$\begin{aligned}
\Pr(\pi^k_{\text{Hi}} \neq i^*_{\text{Hi}}) \le& \Pr(\mathcal{E}) + \Pr(\mathcal{E}^{\complement} \cap \{\pi^k_{\text{Hi}} \neq i^*_{\text{Hi}}\}) \\
\le& \Pr(\mathcal{E}) + \Pr(\mathcal{E}^{\complement} \cap \{\forall k' \in [\frac{k}{2}, k],\ \exists w \notin \widetilde{\mathcal{W}}^*,\ w \in \mathcal{I}^{k'},\ w^{k'} = w\}) \\
\le& \sum_{k'=\frac{k}{2}}^{k} \Pr(w^* \notin \mathcal{I}^{k'}) + \sum_{k'=\frac{k}{2}}^{k} \sum_{w=1}^{W} \Pr(\underline{\pi}^k_{\text{Lo},w} \neq i^*_{\text{Lo},w}) \\
& \qquad\qquad (\Pr(\mathcal{W}^* \cap \mathcal{I}^{k'} = \emptyset) \le \Pr(w^* \notin \mathcal{I}^{k'})) \\
&+ \sum_{k'=\frac{k}{2}}^{k} \sum_{w=1}^{W} \Pr(\mathcal{E}^{\complement} \cap \{\forall k' \in [\frac{k}{2}, k],\ \exists w \notin \widetilde{\mathcal{W}}^*,\ w \in \mathcal{I}^{k'},\ w^{k'} = w\}). \qquad (26)
\end{aligned}$$

For the first and second term, by considering $f(k) = 1 + 16A^2W(k+1)^2$, with a similar discussion as Lem. C.2 and Lem. C.5, we have, for arbitrary $k \ge k^{[W]}_{\max} := \max_{w \in [W]}\{k^w_{\max}\}$, where $k^w_{\max} := c(A + \frac{\alpha A}{\Delta^2_{\text{Lo}}(i)} \log(1 + \frac{\alpha AW}{\Delta_{\min}}))$ for some constant c is an analogue of $k_{\max}$ defined in Lem. C.5 specified on task t:

$$\sum_{k'=\frac{k}{2}}^{k} \Pr(w^* \notin \mathcal{I}^{k'}) + \sum_{w=1}^{W} \sum_{k'=\frac{k}{2}}^{k} \Pr(\underline{\pi}^k_{\text{Lo},w} \neq i^*_{\text{Lo},w}) \le \frac{2^{2\alpha}}{T \cdot k^{2\alpha-1}} + \frac{2 \cdot 2^{2\alpha} AW}{T \cdot k^{2\alpha-1}} + \frac{2 \cdot 2^{2\alpha-1} A^2 T}{T \cdot k^{2\alpha-2}} \le \frac{3 \cdot 2^{2\alpha-1} A^2}{k^{2\alpha-2}}. \qquad (27)$$

Therefore, we mainly focus on the second term. We denote $k'' := c'' \frac{\alpha A}{\Delta^2_{\min}} \log \frac{\alpha AW}{\Delta_{\min}}$ for some constant c'', such that for all $\widetilde{k} \ge k''$, we always have $\widetilde{k} \ge 2A \cdot \frac{512\alpha \log f(\widetilde{k})}{\Delta^2_{\min}}$. Therefore, for arbitrary $w \notin \widetilde{\mathcal{W}}^*$, and arbitrary $\widetilde{k} \ge \max\{k^{[W]}_{\max}, k''\}$ we have:

$$\begin{aligned}
&\Pr(\{w \in \mathcal{I}^{\widetilde{k}}\} \cap \{N^{\widetilde{k}}_{\text{Hi}}(i^*_{\text{Lo},w}) \ge \frac{512\alpha \log f(\widetilde{k})}{\Delta^2_{\min}}\}) \\
\le& \Pr(\{w \in \mathcal{I}^{\widetilde{k}}\} \cap \{N^{\widetilde{k}}_{\text{Hi}}(i^*_{\text{Lo},w}) \ge \frac{512\alpha \log f(\widetilde{k})}{\Delta^2_{\min}}\} \cap \{N^{\widetilde{k}}_{\text{Lo},w}(i^*_{\text{Lo},w}) > \frac{\widetilde{k}}{2}\}) + \Pr(\{N^{\widetilde{k}}_{\text{Lo},w}(i^*_{\text{Lo},w}) \le \frac{\widetilde{k}}{2}\}) \\
\le& \frac{2A}{T \cdot \widetilde{k}^{2\alpha-1}} + \Pr(\{\underline{\mu}^{t,\widetilde{k}}_{\text{Lo}}(i^*_{\text{Lo},w}) \le \overline{\mu}^{t,\widetilde{k}}_{\text{Hi}}(i^*_{\text{Lo},w}) + \varepsilon\} \cap \{N^{\widetilde{k}}_{\text{Lo},w}(i^*_{\text{Lo},w}) > \frac{\widetilde{k}}{2}\} \cap \{N^{\widetilde{k}}_{\text{Hi}}(i^*_{\text{Lo},w}) \ge \frac{512\alpha \log f(\widetilde{k})}{\Delta^2_{\min}}\})
\end{aligned}$$

For the second part, it equals:

$$\begin{aligned}
&\Pr(\{\widehat{\mu}^k_{\text{Lo}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}}) - \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo},w}(i^*_{\text{Lo}})}} \le \widehat{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Hi}}(i^*_{\text{Lo}}) + (\mu_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}})) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Lo}})}} + \varepsilon\} \\
&\quad \cap \{N^k_{\text{Lo},w}(i^*_{\text{Lo},w}) > \frac{k}{2}\} \cap \{N^k_{\text{Hi}}(i^*_{\text{Lo},w}) \ge \frac{512\alpha \log f(k)}{\Delta^2_{\min}}\}) \\
\le& \Pr(\{\widehat{\mu}^k_{\text{Lo}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}}) - \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo},w}(i^*_{\text{Lo}})}} \le \widehat{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Hi}}(i^*_{\text{Lo}}) - \frac{\Delta_{\min}}{4} + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Lo}})}}\}
\end{aligned}$$

$$\cap \{N^k_{\text{Lo},w}(i^*_{\text{Lo},w}) > \frac{k}{2}\} \cap \{N^k_{\text{Hi}}(i^*_{\text{Lo},w}) \geq \frac{512\alpha \log f(k)}{\Delta^2_{\min}}\})$$

$$\leq \Pr(\widehat{\mu}^k_{\text{Lo}}(i^*_{\text{Lo}}) - \mu_{\text{Lo}}(i^*_{\text{Lo}}) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo},w}(i^*_{\text{Lo}})}}) + \Pr(\sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Lo}})}} \leq \widehat{\mu}^k_{\text{Hi}}(i^*_{\text{Lo}}) - \mu_{\text{Hi}}(i^*_{\text{Lo}}))$$

$$(\sqrt{\tfrac{2\alpha \log f(k)}{N^k_{\text{Lo},w}(i^*_{\text{Lo}})}}, \sqrt{\tfrac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i^*_{\text{Lo}})}} \leq \tfrac{\Delta_{\min}}{16})$$

$$\leq \frac{1}{T \cdot k^{2\alpha}}.$$

Therefore, we can conclude that

$$\forall k \geq \max\{k^{[W]}_{\max}, k''\}, \quad \Pr(\{w \in \mathcal{I}^k\} \cap \{N^k_{\text{Hi}}(i^*_{\text{Lo},w}) \geq \frac{512\alpha \log f(k)}{\Delta^2_{\min}}\}) \leq \frac{4}{T \cdot k^{2\alpha}} + \frac{2A}{T \cdot k^{2\alpha-1}} \leq \frac{3A}{T \cdot k^{2\alpha-1}}.$$

Next, we are ready to upper bound the second term in Eq. (26). The key observation is that, as long as the event $\{\forall k' \in [\frac{k}{2}, k],\ \exists w \notin \widetilde{\mathcal{W}}^*,\ w \in \mathcal{I}^{k'},\ w^{k'} = w\}$ happens, no matter what the sequence $\{w^{k'}\}^k_{k'=k/2}$ is, since we only have $A-1$ sub-optimal arms, and $i^*_{\text{Lo},w^{k'}} \neq i^*_{\text{Hi}}$ for all $k' \in [k/2, k]$, there must be an arm which has been taken for at least $\frac{k}{2(A-1)}$ times from step $k/2$ to k, therefore,

$$\begin{aligned}
&\Pr(\forall k' \in [\frac{k}{2}, k],\ \exists w \notin \widetilde{\mathcal{W}}^*,\ w \in \mathcal{I}^{k'},\ w^{k'} = w) \\
\leq &\Pr(\exists k' \in [\frac{k}{2}, k],\ \exists w \notin \widetilde{\mathcal{W}}^*,\ s.t.\ N^{k'}_{\text{Hi}}(i^*_{\text{Lo},w}) = N^{k/2}_{\text{Hi}}(i^*_{\text{Lo},w}) + \frac{k}{2(A-1)} - 1,\ w \in \mathcal{I}^{k'}) \\
\leq &\sum_{k'=\frac{k}{2}}^{k} \sum_{w \notin \widetilde{\mathcal{W}}^*} \Pr(\{w \in \mathcal{I}^{k'}\} \cap \{N^{k'}_{\text{Hi}}(i^*_{\text{Lo},w}) \geq \frac{k}{2(A-1)} - 1,\ w \in \mathcal{I}^{k'}\}) \\
\leq &\sum_{k'=\frac{k}{2}}^{k} \sum_{w \notin \widetilde{\mathcal{W}}^*} \Pr(\{w \in \mathcal{I}^{k'}\} \cap \{N^{k'}_{\text{Hi}}(i^*_{\text{Lo},w}) \geq \frac{512\alpha \log f(k)}{\Delta^2_{\min}}\}) \\
\leq &\frac{3 \cdot 2^{2\alpha-1} A}{k^{2\alpha-2}}.
\end{aligned} \tag{28}$$

According to the definition of $k^{[W]}_{\max}$ and k'', there must exists a constant c^* such that $\max\{k'', k^{[W]}_{\max}\} \leq c^* \frac{\alpha A}{\Delta^2_{\min}} \log \frac{\alpha AW}{\Delta_{\min}}$. By choosing such c^*, and combining Eq. (27) and Eq. (28), we have:

$$\Pr(\pi^k_{\text{Hi}} \neq i^*_{\text{Hi}}) \leq \frac{3 \cdot 2^{2\alpha-1} A^2}{k^{2\alpha-2}} + \frac{3 \cdot 2^{2\alpha-1} A}{k^{2\alpha-2}} \leq \frac{16A}{(k/2)^{2\alpha-2}} = O(\frac{A}{k^{2\alpha-2}}).$$

□

Lemma E.1 (Extension of Lem. C.6)**.** *For arbitrary $K \geq A+1$ and arbitrary $1 \leq k_0 \leq K$, we have:*

$$\begin{aligned}
N^K_{Hi}(i) \leq &k_0 + \sum_{k=k_0+1}^{K} \mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{\underline{\mu}^k_{Lo,w^k}(\underline{\pi}^k_{Lo,w^k}) \leq \overline{\mu}^k_{Hi}(\underline{\pi}^k_{Lo,w^k}) + \varepsilon\} \\
&\qquad \cap \{N^k_{Lo,w^k}(\underline{\pi}^k_{Lo,w^k}) > k/2\} \cap \{i = \underline{\pi}^k_{Lo,w^k}\} \cap \{\pi^k_{Hi} = i\}] \\
&+ \sum_{k=k_0+1}^{K} \mathbb{I}[0 \geq \widehat{\mu}^k_{Hi}(i^*) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{Hi}(i^*)}} - \mu_{Hi}(i^*)] \\
&+ \sum_{k=k_0+1}^{K} \mathbb{I}[\widetilde{\mu}^k_{Hi}(i) + \sqrt{\frac{2\alpha \log f(K)}{k}} - \mu_{Hi}(i) - \Delta_{Hi}(i) \geq 0]
\end{aligned} \tag{29}$$

We omit the proof here since it is almost the same as Lem. C.6, except that we need to specify w^k.

Theorem 5.1. *[Tiered MAB with Multiple Source Tasks] Under Assump. A, B, and C, by running Alg. 3 with $\mathcal{M}_{Lo} = \{M_{Lo,w}\}_{w=1}^{W}$ and M_{Lo}, with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4}$ and $\alpha > 2$, we always have: $Regret_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log(WK))$. Moreover, if at least one task in $\mathcal{M}_{Lo}$ is $\frac{\widetilde{\Delta}_{\min}}{4}$-close to M_{Hi}, we further have: $Regret_K(M_{Hi}) = O(\sum_{\Delta_{Hi}(i)>0} \frac{1}{\Delta_{Hi}(i)} \log \frac{AW}{\Delta_{\min}})$.*

Proof. One key observation is that when analyzing N_{Hi}^K, we only need to analyze the second term in Eq. (29), since the others can be directly bounded.

We first study the case when there exists w^* such that M_{Hi} and M_{Lo,w^*} are ε-close. As a result of Lem. 5.3, we only need to upper bound the regret before step k^*. Similar to the proof of Thm. 4.1, we separate two cases for each k and i.

Case 1-(a) $w^k \neq$ **Null,** $0 < \Delta_{\textbf{Hi}}(i) \leq 4\Delta_{\textbf{Lo},w^k}(i)$ In the following, we will define $\widetilde{k}_{t,i} := 3 + c \cdot \frac{3\alpha}{\Delta^2_{\text{Lo},w}(i)} \log(T + \frac{\alpha AT}{\Delta_{\min}})$ (i.e. similar to the role of $\widetilde{k}_i$ in Lem. C.5 with specified task index t). Since $\Delta_{\text{Hi}}(i) \leq 4\Delta_{\text{Lo},w^k}(i)$, we define $\widetilde{k}_{[W],i} := 3 + c \cdot \frac{48\alpha}{\Delta^2_{\text{Hi}}(i)} \log(T + \frac{\alpha AT}{\Delta_{\min}})$. In this case, obviously $\widetilde{k}_{[W],i} \geq \widetilde{k}_{t,i}$. As a result of Lem. C.5,

$$\begin{aligned}
&\mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{\underline{\mu}^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) \leq \overline{\mu}^k_{\text{Hi}}(\underline{\pi}^k_{\text{Lo},w^k}) + \varepsilon\} \cap \{N^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) > k/2\} \cap \{i = \underline{\pi}^k_{\text{Lo},w^k}\} \cap \{\pi^k_{\text{Hi}} = i\}] \\
\leq &\mathbb{I}[k \leq \widetilde{k}_{[W],i}] + \mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{k > \widetilde{k}_{[W],i}\} \cap \{N^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) > k/2\}]. \qquad (30)
\end{aligned}$$

by taking expectation, we have:

$$\begin{aligned}
&\Pr(\mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{\underline{\mu}^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) \leq \overline{\mu}^k_{\text{Hi}}(\underline{\pi}^k_{\text{Lo},w^k}) + \varepsilon\} \\
&\qquad \cap \{N^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) > k/2\} \cap \{i = \underline{\pi}^k_{\text{Lo},w^k}\} \cap \{\pi^k_{\text{Hi}} = i\}]) \\
\leq &\mathbb{I}[k \leq \widetilde{k}_{[W],i}] + \frac{2}{k^{2\alpha-1}}.
\end{aligned}$$

Case 1-(b) $w^k \neq$ **Null,** $\Delta_{\textbf{Hi}}(i) > 4\Delta_{\textbf{Lo},w^k}(i) \geq 0$ We consider $\bar{k}_i := \frac{c_{\text{Hi},i}\alpha}{\Delta^2_{\text{Hi}}(i)} \log \frac{AW}{\Delta_{\min}}$, where $c_{\text{Hi},i}$ is the minimal constant, such that when $k \geq \frac{c_{\text{Hi},i}\alpha}{\Delta^2_{\text{Hi}}(i)} \log \frac{\alpha AW}{\Delta_{\min}}$, we always have $k \geq \frac{256\alpha \log f(k)}{\Delta^2_{\text{Hi}}(i)}$. With a similar discussion as Eq. (13), we have:

$$\begin{aligned}
&\mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{\underline{\mu}^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) \leq \overline{\mu}^k_{\text{Hi}}(\underline{\pi}^k_{\text{Lo},w^k}) + \varepsilon\} \cap \{N^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) > k/2\} \cap \{i = \underline{\pi}^k_{\text{Lo},w^k}\} \cap \{\pi^k_{\text{Hi}} = i\}] \\
\leq &\mathbb{I}[k < \bar{k}_i] + \mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{\{\underline{\mu}^k_{\text{Lo},w^k}(\underline{\pi}^k_{\text{Lo},w^k}) \leq \overline{\mu}^k_{\text{Hi}}(\underline{\pi}^k_{\text{Lo},w^k}) + \varepsilon\}\} \\
&\qquad \cap \{N^k_{\text{Lo},w^k}(i) \geq \frac{128\alpha \log f(k)}{\Delta^2_{\text{Hi}}(i)}\} \cap \{i = \underline{\pi}^k_{\text{Lo}}\} \cap \{\pi^k_{\text{Hi}} = i\}] \\
\leq &\mathbb{I}[k < \bar{k}_i] + \mathbb{I}[\{w^k \neq \texttt{Null}\} \cap \{\widehat{\mu}^k_{\text{Lo},w^k}(i) - \mu_{\text{Lo},w^k}(i) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo},w^k}(i)}}\}] \\
&+ \mathbb{I}[\{\widehat{\mu}^k_{\text{Hi}}(i) - \mu_{\text{Hi}}(i) + \sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Hi}}(i)}} \geq \frac{\Delta_{\text{Hi}}(i)}{4}\} \cap \{\pi^k_{\text{Hi}} = i\}]. \qquad (31)
\end{aligned}$$

We denote $k_i^* := \max\{\widetilde{k}_{[W],i}, \bar{k}_i\}$. Combining Eq. (30) and (31) with Eq. (29), for arbitrary K, we have (recall that $\widetilde{\mu}^k_{\text{Hi}}(i)$ is defined to be the average of k random samples from reward distribution of arm i in M_{Hi}):

$$\begin{aligned}
&\mathbb{E}[N^K_{\text{Hi}}(i)] \\
\leq &\sum_{k=1}^{K} \mathbb{I}[k \leq \widetilde{k}_{[W],i}] + \sum_{k=1}^{K} \frac{2}{k^{2\alpha-1}} + \sum_{k=1}^{K} \mathbb{I}[k < \bar{k}_i] \\
&+ \sum_{k=1}^{K} \Pr(\{w^k \neq \texttt{Null}\} \cap \{\widehat{\mu}^k_{\text{Lo},w^k}(i) - \mu_{\text{Lo},w^k}(i) \leq -\sqrt{\frac{2\alpha \log f(k)}{N^k_{\text{Lo},w^k}(i)}}\})
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}[\sum_{k=1}^{K}\mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^{k}(i)-\mu_{\text{Hi}}(i)+\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Hi}}^{k}(i)}}\geq\frac{\Delta_{\text{Hi}}(i)}{4}\}\cap\{\pi_{\text{Hi}}^{k}=i\}]]\\
&+\sum_{k=1}^{K}\Pr(0\geq\widehat{\mu}_{\text{Hi}}^{k}(i^{*})+\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Hi}}^{k}(i^{*})}}-\mu_{\text{Hi}}(i^{*}))\\
&+\sum_{k=1}^{K}\mathbb{E}[\mathbb{I}[\widetilde{\mu}_{\text{Hi}}^{k}(i)+\sqrt{\frac{2\alpha\log f(K)}{k}}-\mu_{\text{Hi}}(i)-\Delta_{\text{Hi}}(i)\geq 0]]\\
\leq&\widetilde{k}_{[W],i}+\sum_{k=1}^{K}\frac{2}{k^{2\alpha-1}}+\bar{k}_{i}+2\cdot\sum_{k=1}^{K}\frac{1}{8Ak^{2\alpha}}+2\sum_{k=1}^{K}\mathbb{E}[\mathbb{I}[\widetilde{\mu}_{\text{Hi}}^{k}(i)+\sqrt{\frac{2\alpha\log f(K)}{k}}-\mu_{\text{Hi}}(i)-\frac{\Delta_{\text{Hi}}(i)}{4}\geq 0]]\\
=&O(\frac{1}{\Delta_{\text{Hi}}^{2}(i)}\log WK).
\end{aligned}
\tag{32}
$$

As a result, combining with Lem. 5.3, we can conclude that:

$$
\text{Regret}_{K}(M_{\text{Hi}})=\sum_{i\neq i^{*}}\Delta_{\text{Hi}}(i)\mathbb{E}[N_{\text{Hi}}^{k^{*}}(i)]+\sum_{k=k^{*}+1}^{K}\Pr(\pi_{\text{Hi}}^{k}\neq i_{\text{Hi}}^{*})=O(\sum_{i\neq i^{*}}\frac{1}{\Delta_{\text{Hi}}(i)}\log\frac{AW}{\Delta_{\min}})
$$

Next, we study the case when there is no task among $\{M_{\text{Lo},w}\}_{w=1}^{W}$ close enough to M_{Hi}. Similarly, we also decompose into three cases.

Case 2-(a) $w^{k}\neq$ **Null,** $i\neq i_{\textbf{Lo},w^{k}}^{*}, i\neq i_{\textbf{Hi}}^{*}$ **and** $0<\Delta_{\textbf{Hi}}(i)\leq 4\Delta_{\textbf{Lo},w^{k}}(i)$ The result is the same as Eq. (30).

Case 2-(b) $w^{k}\neq$ **Null,** $i\neq i_{\textbf{Lo},w^{k}}^{*}, i\neq i_{\textbf{Hi}}^{*}$ **and** $\Delta_{\textbf{Hi}}(i)>4\Delta_{\textbf{Lo},w^{k}}(i)>0$ The result is the same as Eq. (31).

Case 2-(c) $w^{k}\neq$ **Null,** $i=i_{\textbf{Lo},w^{k}}^{*}$ If $i_{\text{Lo},w^{k}}^{*}=i_{\text{Hi}}^{*}$, M_{Hi} suffers no regret when choosing $i_{\text{Lo},w^{k}}^{*}$. Therefore, in the following, we only study the case when $i_{\text{Lo},w^{k}}^{*}\neq i_{\text{Hi}}^{*}$. For arm i (note that $i=i_{\text{Lo},w^{k}}^{*}$ in this case), we define $k'_{\max}=\frac{c'_{\max}\alpha}{\Delta_{\text{Hi}}(i)^{2}}\log\frac{\alpha AT}{\Delta_{\min}}$, where $c'_{\max}$ is the minimal constant, such that for all $k\geq\frac{c'_{\max}\alpha}{\Delta_{\text{Hi}}(i)^{2}}\log\frac{\alpha AT}{\Delta_{\min}}$, we always have $k\geq\frac{1024\alpha\log f(k)}{\Delta_{\text{Hi}}(i)^{2}}$. With a similar discussion as Eq. (15), for the following event, we have:

$$
\begin{aligned}
&\mathbb{I}[\{w^{k}\neq\texttt{Null}\}\cap\{\underline{\mu}_{\text{Lo},w^{k}}^{k}(\underline{\pi}_{\text{Lo},w^{k}}^{k})\leq\overline{\mu}_{\text{Hi}}^{k}(\underline{\pi}_{\text{Lo},w^{k}}^{k})+\varepsilon\}\cap\{N_{\text{Lo},w^{k}}^{k}(\underline{\pi}_{\text{Lo},w^{k}}^{k})>k/2\}\cap\{i=\underline{\pi}_{\text{Lo},w^{k}}^{k}\}\cap\{\pi_{\text{Hi}}^{k}=i\}]\\
\leq&\mathbb{I}[k\leq k'_{\max}]+\mathbb{I}[\{w^{k}\neq\texttt{Null}\}\cap\widehat{\mu}_{\text{Lo},w^{k}}^{k}(i)-\mu_{\text{Lo},w^{k}}(i)\leq-\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Lo},w^{k}}^{k}(i_{\text{Lo}}^{*})}}]\\
&+\mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^{k}(i)-\mu_{\text{Hi}}(i)+\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Hi}}^{k}(i)}}\geq\frac{\Delta_{\text{Hi}}(i)}{8}\}\cap\{\pi_{\text{Hi}}^{k}=i\}].
\end{aligned}
$$

Combining all the cases above, similar to Case 1, for arbitrary $i\neq i_{\text{Hi}}^{*}$, we can conclude:

$$
\begin{aligned}
&\mathbb{E}[N_{\text{Hi}}^{K}(i)]\\
\leq&\sum_{k=1}^{K}\mathbb{I}[k\leq\widetilde{k}_{[W],i}]+\sum_{k=1}^{K}\frac{2}{k^{2\alpha-1}}+\sum_{k=1}^{K}\mathbb{I}[k<\bar{k}_{i}]+\sum_{k=1}^{K}\Pr(\{w^{k}\neq\texttt{Null}\}\cap\{\widehat{\mu}_{\text{Lo}}^{k}(i)-\mu_{\text{Lo}}(i)\leq-\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Lo},w^{k}}^{k}(i)}}\})\\
&+\mathbb{E}[\sum_{k=1}^{K}\mathbb{I}[\{\widehat{\mu}_{\text{Hi}}^{k}(i)-\mu_{\text{Hi}}(i)+\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Hi}}^{k}(i)}}\geq\frac{\Delta_{\text{Hi}}(i)}{4}\}\cap\{\pi_{\text{Hi}}^{k}=i\}]]\\
&+\sum_{k=1}^{K}\Pr(0\geq\widehat{\mu}_{\text{Hi}}^{k}(i^{*})+\sqrt{\frac{2\alpha\log f(k)}{N_{\text{Hi}}^{k}(i^{*})}}-\mu_{\text{Hi}}(i^{*}))\\
&+\sum_{k=1}^{K}\mathbb{E}[\mathbb{I}[\widetilde{\mu}_{\text{Hi}}^{k}(i)+\sqrt{\frac{2\alpha\log f(K)}{k}}-\mu_{\text{Hi}}(i)-\Delta_{\text{Hi}}(i)\geq 0]]
\end{aligned}
$$

$$+\sum_{k=1}^{K}\mathbb{I}[k\le k'_{\max}]+\sum_{k=1}^{K}\Pr(\{w^k\neq \texttt{Null}\}\cap\{\widehat{\mu}^k_{\text{Lo}}(i)-\mu_{\text{Lo}}(i)\le -\sqrt{\frac{2\alpha\log f(k)}{N^k_{\text{Lo},w^k}(i)}}\})$$
$$+\mathbb{E}[\sum_{k=1}^{K}\mathbb{I}[\{\widehat{\mu}^k_{\text{Hi}}(i)-\mu_{\text{Hi}}(i)+\sqrt{\frac{2\alpha\log f(k)}{N^k_{\text{Hi}}(i)}}\ge\frac{\Delta_{\text{Hi}}(i)}{4}\}\cap\{\pi^k_{\text{Hi}}=i\}]]$$
$$\le\widetilde{k}_{[W],i}+\sum_{k=1}^{K}\frac{2}{k^{2\alpha-1}}+\bar{k}_i+2\cdot\sum_{k=1}^{K}\frac{1}{8Ak^{2\alpha}}+3\sum_{k=1}^{K}\mathbb{E}[\mathbb{I}[\widetilde{\mu}^k_{\text{Hi}}(i)+\sqrt{\frac{2\alpha\log f(K)}{k}}-\mu_{\text{Hi}}(i)-\frac{\Delta_{\text{Hi}}(i)}{8}\ge 0]]$$
$$=O(\frac{1}{\Delta^2_{\text{Hi}}(i)}\log WK).$$

which implies:

$$\text{Regret}_K(M_{\text{Hi}})=\sum_{i\neq i^*}\Delta_{\text{Hi}}(i)\mathbb{E}[N^K_{\text{Hi}}(i)]=O(\sum_{i\neq i^*}\frac{1}{\Delta_{\text{Hi}}(i)}\log WK).$$

□

F Proofs for Tiered RL with Multiple Source/Low-Tier Tasks

We first introduce the notion of transferable states in this multi-source tasks setting. Comparing with Def. 2.2, we have an additional constraint on $d^*_{\text{Hi}}(s_h)>0$. This is because, to distinguish which source task to transfer, we require s_h in M_{Hi} to be visited frequently enough for accurate estimation, and it is only possible for those s_h on optimal trajectories given that we expect $\text{Regret}_K(M_{\text{Hi}})$ is at least near-optimal.

Definition F.1 (λ-Transferable States in TRL-MST)**.** Given any $\lambda>0$, we say s_h is λ-transferable if $d^*_{\text{Hi}}(s_h)>0$, and $\exists w\in[W]$, such that $d^*_{\text{Lo},w}(s_h)\ge\lambda$ and M_{Hi} is $\frac{\widetilde{\Delta}_{\min}}{4(H+1)}$-close to $M_{\text{Lo},w}$ on state s_h. We use $\mathcal{Z}_h^{\lambda,[W]}$ to denote the set of λ-transferable state at step $h\in[H]$.

Definition F.2 (Benefitable States in TRL-MST)**.** Similar to the single task case, we define $\mathcal{C}_h^{\lambda,[W],1}:=\{(s_h,a_h)|s_h\in\mathcal{Z}^{\lambda,[W]},a_h\neq\pi^*_{\text{Hi},h}(s_h)\}$, $\mathcal{C}_h^{\lambda,[W],2}:=\{(s_h,a_h)|\text{Block}(\{\mathcal{C}_{h'}^{\lambda,[W],1}\}_{h'=1}^{h-1},s_h)=\text{True},s_h\notin\mathcal{C}_h^{\lambda,[W],1},a_h\in\mathcal{A}_h\}$ and $\mathcal{C}_h^*:=\{(s_h,a_h)|d^*_{\text{Hi}}(s_h,a_h)>0\}$, which represents the three categories of state-action pairs with constant regret. We define $\mathcal{C}_h^{\lambda,[W]}:=\mathcal{C}_h^{\lambda,[W],1}\cup\mathcal{C}_h^{\lambda,[W],2}\cup\mathcal{C}_h^*$, which captures the benefitable state-action pairs.

Remark F.3 (Constant Regret in Entire M_{Hi})**.** *For each individual task $M_{Lo,w}$, the additional constraint $d^*_{Hi}(s_h)>0$ reduces the size of transferable states comparing with single task learning setting. However, if the tasks are diverse enough, we expect $\mathcal{C}_h^{\lambda,[W]}$ to be much larger than $\mathcal{C}_h^{\lambda}$ in Sec. 4.2 and we can achieve more benefits with only an additional cost of order $\log W$. Besides, if $\forall h,s_h$ with $d^*_{Hi}(s_h)>0$ we have $s_h\in\mathcal{Z}_h^{\lambda,[W]}$, then, $\mathcal{C}_h^{\lambda,[W]}=\mathcal{S}_h\times\mathcal{A}_h$ and we can achieve constant regret for the entire M_{Hi}.*

F.1 Additional Algorithms, Conditions and Notations

Our algorithm is provided in Alg. 7, which is extended from Alg. 2 and integrated with the "Trust till Failure" strategy introduced in bandit setting in Alg. 3. We consider the same **ModelLearning** and **Bonus** algorithm in Sec. D.1, but a different condition for Alg^{Lo} listed below:

Condition F.4 (Condition on Alg^{Lo} in MT-TRL)**.** Alg^{Lo} is an algorithm which returns deterministic policies at each iteration for each task $M_{tO,w}\in\mathcal{M}_{\text{Lo}}$, and there exists C_1,C_2 only depending on S,A,H and $\Delta_{\min}$ but independent of k, such that for arbitrary $k\ge 2$, we have $\Pr(\mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k})\ge 1-\frac{1}{k^\alpha}$ for $\mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k}$ defined below:

$$\mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k}:=\bigcap_{w\in W}\{\sum_{\widetilde{k}=1}^{k}V^*_{\text{Lo},w,1}(s_1)-V^{\widetilde{\pi}^{\widetilde{k}}_{\text{Lo}}}_{\text{Lo},w,1}(s_1)\le C_1+\alpha C_2\log Wk\}.$$

Algorithm 7: Robust Tiered RL with Multiple Low-Tier Tasks

Input: Ratio $\lambda \in (0,1)$; Bonus term computation function **Bonus**; Sequence of confidence level $(\delta_k)_{k\geq 1}$ with $\delta_k = 1/SAHWk^\alpha$; Model learning function **ModelLearning**. $\varepsilon < \underline{\Delta}_{\min}/4(H+1)$

Initialize: $D^0_{\mathrm{Lo},w} \leftarrow \{\}$ for $w \in [W]$, $D^0_{\mathrm{Hi}} \leftarrow \{\}$, set $\underline{V}^k_{\mathrm{Hi},h+1}, \underline{Q}^k_{\mathrm{Hi},h+1}, \underline{V}^k_{\mathrm{Lo},w,h+1}, \underline{Q}^k_{\mathrm{Lo},w,h+1}, \widetilde{V}^k_{\mathrm{Lo},w,h+1}, \widetilde{Q}^k_{\mathrm{Lo},w,h+1}$ to be 0 for all $k = 1, 2, ...$

for $k = 1, 2, ...$ **do**

  for $t = 1, 2, ...T$ **do**

    $\pi^k_{\mathrm{Lo},w} \leftarrow \mathrm{Alg}^{\mathrm{Lo}}(D^{k-1}_{\mathrm{Lo},w})$; collect data from $M_{\mathrm{Lo},w}$ with $\pi^k_{\mathrm{Lo},w}$; update $D^k_{\mathrm{Lo},w} \leftarrow D^{k-1}_{\mathrm{Lo},w} \cup \{\tau^k_t\}$.

    $\{\widehat{\mathbb{P}}^k_{\mathrm{Lo},w,h}\}_{h=1}^H \leftarrow$ **ModelLearning**$(D^{k-1}_{\mathrm{Lo},w})$, $\quad \{b^k_{\mathrm{Lo},h}\}_{h=1}^H \leftarrow$ **Bonus**$(D^{k-1}_{\mathrm{Lo},w},\ \delta_k)$.

    for $h = H, H-1..., 1$ **do**

      $\underline{Q}^k_{\mathrm{Lo},w,h}(\cdot,\cdot) \leftarrow \max\{0, r_{\mathrm{Lo},h}(\cdot,\cdot) + \widehat{\mathbb{P}}^k_{\mathrm{Lo},w,h}\underline{V}^k_{\mathrm{Lo},w,h+1}(\cdot,\cdot) - b^k_{\mathrm{Lo},h}(\cdot,\cdot)\}$.

      $\underline{V}^k_{\mathrm{Lo},w,h}(\cdot) = \max_a \underline{Q}^k_{\mathrm{Lo},w,h}(\cdot,a)$, $\quad \underline{\pi}^k_{\mathrm{Lo},w,h}(\cdot) \leftarrow \arg\max_a \underline{Q}^k_{\mathrm{Lo},w,h}(\cdot,a)$.

    end

  end

  $\{\widehat{\mathbb{P}}^k_{\mathrm{Hi},h}\}_{h=1}^H \leftarrow$ **ModelLearning**(D^{k-1}_{Hi}), $\quad \{b^k_{\mathrm{Hi},h}\}_{h=1}^H \leftarrow$ **Bonus**$(D^{k-1}_{\mathrm{Hi}},\ \delta_k)$.

  for $h = H, H-1..., 1$ **do**

    $\underline{Q}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(\cdot,\cdot) \leftarrow \max\{0, r^k_{\mathrm{Hi},h}(\cdot,\cdot) + \widehat{\mathbb{P}}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}\underline{V}^k_{\mathrm{Hi},h+1}(\cdot,\cdot) - b^k_{\mathrm{Hi},h}(\cdot,\cdot)\}$, $\quad \underline{V}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(\cdot) = \underline{Q}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(\cdot, \pi^k_{\mathrm{Hi}})$

    $\widetilde{Q}^k_{\mathrm{Hi},h}(\cdot,\cdot) \leftarrow \min\{H, r_{\mathrm{Hi}}(\cdot,\cdot) + \widehat{\mathbb{P}}^k_{\mathrm{Hi},h}\widetilde{V}^k_{\mathrm{Hi},h+1}(\cdot,\cdot) + b^k_{\mathrm{Hi},h}(\cdot,\cdot)\}$.

    for $s_h \in \mathcal{S}_h$ **do**

      $\mathcal{I}^k_{s_h} \leftarrow \{w \in [W] | \{\underline{V}^k_{\mathrm{Lo},w,h}(s_h) \leq \widetilde{Q}^k_{\mathrm{Hi},h}(s_h, \underline{\pi}^k_{\mathrm{Lo},w,h}) + \varepsilon\} \cap \{\max_a N^k_{\mathrm{Lo},w,h}(s_h,a) \geq \frac{\lambda}{3}k\}\}$.

      if $\mathcal{I}^k(s_h) \neq \emptyset$ **then**

        if $w^{k-1}_{s_h} \neq$ `Null` *and* $w^{k-1} \in \mathcal{I}^k_{s_h}$ **then** $w^k_{s_h} \leftarrow w^{k-1}_{s_h}$;

        else if $w^{k-1}_{s_h} \neq$ `Null` *and* $\exists w \in \mathcal{I}^k(s_h)$ *s.t.* $\pi^{k-1}_{Hi}(s_h) = \arg\max_a N_{Lo,w,h}(s_h,a)$ **then** $w^k_{s_h} \leftarrow w$;

        else $w^k_{s_h} \leftarrow \mathrm{Unif}(\mathcal{I}^k_{s_h})$. ;

        $\pi^k_{\mathrm{Hi}}(s_h) \leftarrow \arg\max_a N^k_{\mathrm{Lo},w^k_{s_h},h}(s_h,a)$.

      end

      else $w^k_{s_h} \leftarrow$ `Null`, $\pi^k_{\mathrm{Hi}}(s_h) \leftarrow \arg\max_a \widetilde{Q}^k_{\mathrm{Hi},h}(s_h,a)$;

      $\widetilde{V}^k_{\mathrm{Hi},h}(s_h) \leftarrow \min\{H, \widetilde{Q}^k_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}}) + \frac{1}{H}(\widetilde{Q}^k_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}}) - \underline{Q}^{\pi^k_{\mathrm{Hi}}}_{\mathrm{Hi},h}(s_h,\pi^k_{\mathrm{Hi}}))\}$

    end

  end

  Deploy π_{Hi} to interact with M_{Hi} and receive τ^k_{Hi}; update $D^k_{\mathrm{Hi}} \leftarrow D^{k-1}_{\mathrm{Hi}} \cup \{\tau^k_{\mathrm{Hi}}\}$

end

Next, we introduce some notations. As analogues of $\mathcal{E}_{\textbf{Bonus},k}$ and $\mathcal{E}_{\textbf{Con},k}$ in single low-tier task setting, we consider the following events:

$$\mathcal{E}_{\textbf{Bonus},[W],k} := \bigcap_{\substack{(\cdot)\in\{\mathrm{Hi},\mathrm{Lo}_1,\ldots,\mathrm{Lo}_T\},\\ h\in[H],\\ s_h\in\mathcal{S}_h,a_h\in\mathcal{A}_h}} \Big\{\{H\cdot\|\widehat{\mathbb{P}}^k_{(\cdot),h}(s_h,a_h)-\mathbb{P}_{(\cdot),h}(s_h,a_h)\|_1 < b^k_{(\cdot),h}(s_h,a_h) \le B_1\sqrt{\frac{\log(B_2/\delta_k)}{N^k_{(\cdot),h}(s_h,a_h)}}\}\Big\};$$

$$\begin{aligned}\mathcal{E}_{\textbf{Con},[W],k} := \bigcap_{\substack{h\in[H],\\ s_h\in\mathcal{S}_h,\\ a_h\in\mathcal{A}_h}} \Big\{\{\frac{1}{2}\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\mathrm{Hi}}}(s_h,a_h)-\alpha\log(2SAHWk) \le N^k_{\mathrm{Hi},h}(s_h,a_h)\\ \le e\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\mathrm{Hi}}}(s_h,a_h)+\alpha\log(2SAHWk)\}\\ \cap\Big(\bigcap_{w\in[W]}\{\frac{1}{2}\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\mathrm{Lo},w}}(s_h,a_h)-\alpha\log(2SAHWk) \le N^k_{\mathrm{Lo},w,h}(s_h,a_h)\\ \le e\sum_{\widetilde{k}=1}^{k} d^{\pi^{\widetilde{k}}_{\mathrm{Lo},w}}(s_h,a_h)+\alpha\log(2SAHWk)\}\Big)\Big\}.\end{aligned}$$

Under the choice of $\delta_k = 1/SAHWk^\alpha$, we have $\Pr(\mathcal{E}_{\textbf{Bonus},[W],k}) \ge 1-\frac{1}{k^\alpha}$. Besides, as a result of Lem. D.8, we have $\Pr(\mathcal{E}_{\textbf{Con},[W],k}) \ge 1-\frac{1}{k^\alpha}$.

F.2 Analysis

In this sub-section, we introduce the analysis for Alg. 7. We first provide several lemma and theorem for preparation, which are extended from single task setting. We will omit the detailed proofs if they are almost the same expect there are multiple source tasks and the additional W in the log factors.

Lemma F.5 (Lem. D.11 in TRL-MST Setting)**.** *There exists an absolute constant $c^{[W]}_{\Xi}$, such that for arbitrary fixed $\xi>0$, and for arbitrary*

$$k \ge c^{[W]}_{\Xi}\max\{\frac{\alpha B_1^2H^2S}{\lambda^2\xi^2}\log(\frac{\alpha HSAB_1B_2W}{\lambda\xi}), \frac{(C_1+\alpha C_2)SH}{\Delta_{\min}\lambda\xi}\log\frac{C_1C_2SAHW}{\Delta_{\min}\lambda\xi}\}$$

on the event $\mathcal{E}_{\textbf{Con},[W],k}, \mathcal{E}_{\textbf{Bonus},[W],k}$ and $\mathcal{E}_{Alg^{Lo},[W],k}$, $\forall h\in[H], s_h\in\mathcal{S}_h$ with $N^k_{Lo,h}(s_h) > \frac{\lambda}{3}$, we have

$$V^*_{Lo,h}(s_h) - \underline{V}^k_{Lo,h}(s_h) \le \xi.$$

Lemma F.6 (Lem. D.10 in TRL-MST Setting)**.** *There exists a constant $c^{[W]}_{occup}$ which is independent of λ, S, A, H and gap Δ, s.t., for all $k \ge k^{[W]}_{occup} := c^{[W]}_{occup}\frac{C_1+\alpha C_2}{\lambda\Delta_{\min}}\log(\frac{\alpha C_1C_2SAHW}{\lambda\Delta_{\min}})$, on the events of $\mathcal{E}_{Alg^{Lo},[W],k}$ and $\mathcal{E}_{\textbf{Con},[W],k}$, forall $w\in[W]$, $N^k_{Lo,w,h}(s_h,a_h)\ge\frac{\lambda}{3}k$ implies that $d^*_{Lo,w}(s_h,a_h)\ge\frac{\lambda}{9}$, and conversely, if $d^*_{Lo,w}(s_h,a_h)\ge\lambda$, we must have $N^k_{Lo,w,h}(s_h) \ge N^k_{Lo,w,h}(s_h,\pi^*_{Lo,w}) \ge \frac{\lambda}{3}k$.*

Theorem F.7. *[Thm. 4.3 in TRL-MST Setting] There exists a constant $c^{[W]}_{overest}$, such that, for arbitrary $k \ge k^{[W]}_{ost}$ with*

$$k^{[W]}_{ost} := c^{[W]}_{overest}\cdot\max\{\alpha\frac{B_1^2H^2S}{\lambda^2\Delta^2_{\min}}\log(HSAWB_1B_2), \frac{(C_1+\alpha C_2)SH}{\lambda\Delta^2_{\min}}\log\frac{C_1C_2SAHW}{\Delta_{\min}\lambda}\},$$

on the event of $\mathcal{E}_{\textbf{Con},[W],k}, \mathcal{E}_{\textbf{Bonus},[W],k}, \mathcal{E}_{Alg^{Lo},[W],k}$, we have:

$$Q^*_{Hi,h}(s_h,a_h) \le \widetilde{Q}^k_{Hi,h}(s_h,a_h), \quad V^*_{Hi,h}(s_h) \le \widetilde{V}^k_{Hi,h}(s_h), \quad \forall h\in[H],\ s_h\in\mathcal{S}_h,\ a_h\in\mathcal{A}_h.$$

and

$$V^*_{Hi,1}(s_1) - V^{\pi^k_{Hi}}_{Hi,1}(s_1) \le 2e\mathbb{E}_{\pi^k_{Hi}}[\sum_{h=1}^{H}\textit{Clip}\left[\min\{H,4b^k_{Hi,h}(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH}\vee\frac{\Delta_{Hi}(s_h,a_h)}{4e}\right]]. \quad (33)$$

Next, we first establish a $O(\log K)$-regret bound regardless of similarity between $\mathcal{M}_{\text{Lo}}$ and M_{Hi}, and use it to prove Lem. F.9, which will be further used to establish the tighter bound in Thm. 5.2.

Theorem F.8 (Regret bound for general cases)**.**

$$\textit{Regret}_K(M_{Hi}) = O\left(k_{start}^{[W]} \cdot H + SAH^2 \log(2SAHK) + B_1 \log(B_2 K) \sum_{h=1}^{H} \sum_{s_h, a_h} \left(\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{Hi}(s_h, a_h)}\right)\right). \tag{34}$$

Proof. Consider $k_{start}^{[W]} := \max\{k_{ost}^{[W]}, k_{occup}^{[W]}\}$. Similar to the proof of Thm. 4.2, we can conduct identical techniques and provide the following bounds for arbitrary s_h, a_h:

$$\begin{aligned}
&\sum_{k=k_{start}^{[W]}+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\middle|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]] \\
\leq & H \cdot \sum_{k=1}^{\tau_{s_h,a_h}^K} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) + \sum_{k=\tau_{s_h,a_h}^K+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4B_1\sqrt{\frac{\alpha\log(KB_2)}{N_{\text{Hi},h}^k(s_h,a_h)}}\}\middle|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]] \\
= & O\left(H\log(2SAHWK) + B_1(\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(B_2WK)\right).
\end{aligned}$$

where $\tau_{s_h,a_h}^K := \min_k \ s.t.\ \forall k' \geq k,\ \frac{1}{4}\sum_{k'=1}^{k-1} d_{\text{Hi}}^{\pi_{\text{Hi}}^k}(s_h,a_h) \geq \alpha\log(2SAHWK)$. Then, we can conclude:

$$\begin{aligned}
&\mathbb{E}[\sum_{k=1}^{K} V_{\text{Hi},1}^*(s_1) - V_{\text{Hi},1}^{\pi_{\text{Hi}}^k}(s_1)] \\
\leq & \sum_{k=k_{start}^{[W]}+1}^{K} 2e\mathbb{E}_{\pi_{\text{Hi}}^k}[\sum_{h=1}^{H} \text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\middle|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},[W],k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k} \cap \mathcal{E}_{\textbf{Con},[W],k}]] \\
& + \sum_{k=k_{start}^{[W]}+1}^{K} H \cdot \Pr(\mathcal{E}_{\textbf{Bonus},[W],k}^{\complement} \cup \mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k}^{\complement} \cup \mathcal{E}_{\textbf{Con},[W],k}^{\complement}) + H \cdot k_{start}^{[W]} \\
= & O\left(k_{ost}^{[W]} \cdot H + SAH^2\log(2SAHWK) + B_1 \sum_{h=1}^{H} \sum_{s_h,a_h} (\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(B_2WK)\right).
\end{aligned}$$

□

In the following, we try to show an extension of Lem. 5.3 in RL setting. To establish result, we require some techniques to overcome the difficulty raised by state transition. We first establish sub-linear regret bound for Alg. 7 based on the multi-task version of Eq. (1), and use it we can show that $\sum_{k'=k/2}^{k} d_{\text{Hi}}^*(s_h,a_h) > 0$ when k is large enough, which imply that $\pi_{\text{Lo},w^{\widetilde{k}}}^*(s_h) = \pi_{\text{Hi}}^*(s_h)$ for some $\widetilde{k} \in [\frac{k}{2}, k]$ with high probability. Under good events, by our task selection strategy, we can expect if $s_h \in \mathcal{Z}_h^{\lambda,[W]}$ and $\pi_{\text{Lo},w^{\widetilde{k}}}^*(s_h) = \pi_{\text{Hi}}^*(s_h)$, from $\widetilde{k}$ to k, either the trusted task does not change, or it will hand over to another one recommending the same action, which is exactly $\pi_{\text{Hi}}^*(s_h)$.

Lemma F.9. *[Absorb to Similar Task] Under Assump. A, B, C for all $s_h \in \mathcal{Z}_h^{\lambda,[W]}$, by running Alg. 7 in Appx. F.1 with $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4(H+1)}$, $\alpha > 2$ and arbitrary $\lambda > 0$, there exists $\iota_{s_h}^* = \textit{Poly}(SAH, \lambda^{-1}, \Delta_{\min}^{-1}, 1/d_{Hi}^*(s_h), \log W)$, such that, $\forall k \geq \iota_{s_h}^*$, $\Pr(\pi_{Hi}^k(s_h) \neq \pi_{Hi}^*(s_h)) = O(\frac{1}{k^{\alpha-1}})$.*

Proof. Similar to MAB setting, we define $\mathcal{W}_{s_h}^* := \{t \in [W] | \pi_{\text{Lo},w}^*(s_h) = \pi_{\text{Hi}}^*(s_h), V_{\text{Lo},w,h}(s_h) \leq V_{\text{Hi},w,h}(s_h) + \frac{\widetilde{\Delta}_{\min}}{4(H+1)}, d_{\text{Lo},w}^*(s_h) > \lambda\}$ and $\widetilde{\mathcal{W}}_{s_h}^* := \{t \in [W] | \pi_{\text{Lo},w}^*(s_h) = \pi_{\text{Hi}}^*(s_h), d_{\text{Lo},w}^*(s_h) >$

$\lambda\}$. The key observation is that, when $k \geq k_{start}$:

$$\Pr(\pi^k_{\text{Hi}}(s_h) = \pi^*_{\text{Hi}}(s_h)) \geq \Pr(\underbrace{\{\forall k' \in [\frac{k}{2}, k],\ \mathcal{W}^*_{s_h} \subset \mathcal{I}^{k'}_{s_h}\}}_{\mathcal{E}_{k,1}} \cap \underbrace{\{\exists k' \in [\frac{k}{2}, k],\ s.t.\ d^{\pi^{k'}_{\text{Hi}}}_{\text{Hi}}(s_h, \pi^k_{\text{Hi}}) > 0\}}_{\mathcal{E}_{k,2}}$$
$$\cap \underbrace{\{\forall k' \in [\frac{k}{2}, k],\ \forall t \in \mathcal{I}^{k'}_{s_h},\ \underline{\pi}^{k'}_{\text{Lo},w}(s_h) = \pi^*_{\text{Lo},w^{k'}}(s_h)\}}_{\mathcal{E}_{k,3}}).$$

That's because if $\mathcal{E}_{k,2}$ holds at some k', then, we can only have the following two cases: (1) $w^{k'}_{s_h} = w^*_{s_h}$: because of $\mathcal{E}_{k,1}$, we must have $w^k_{s_h} = w^*_{s_h}$, which implies $\pi^k_{\text{Hi}}(s_h) = \pi^*_{\text{Hi}}(s_h)$ as a result of $\mathcal{E}_{k,3}$; (2) $w^{k'}_{s_h} \neq w^*_{s_h}$: because of $\mathcal{E}_{k,1}$ and $\mathcal{E}_{k,3}$, $w^{\widetilde{k}}$ can only transfer inside $\widetilde{\mathcal{W}}^*_{s_h}$ for $\widetilde{k} \in [k', k]$, and therefore we still have $\pi^k_{\text{Hi}}(s_h) = \pi^*_{\text{Hi}}(s_h)$. In the following, we will provide a upper bound for $\Pr(\mathcal{E}^{\complement}_{k,1}), \Pr(\mathcal{E}^{\complement}_{k,2}), \Pr(\mathcal{E}^{\complement}_{k,3})$.

On the event of $\mathcal{E}_{\textbf{Bonus},[W],k}$, with a similar analysis as Lem. D.5 on $\underline{Q}^k_{\text{Lo},w,h}$, we have:

$$\Pr(\mathcal{E}^{\complement}_{k,1}) \leq \sum_{k'=\frac{k}{2}}^{k} \Pr(w^*_{s_h} \notin \mathcal{I}^{k'}_{s_h}) \leq \sum_{k'=\frac{k}{2}}^{k} \Pr(\mathcal{E}^{\complement}_{\textbf{Bonus},[W],k}) \leq \frac{2}{(k/2)^{\alpha-1}}.$$

Besides, for arbitrary s_h with $d^*_{\text{Hi}}(s_h) > 0$, and arbitrary k, by Azuma-Hoeffding inequality and Thm. D.6, with probability at least $1 - \frac{1}{k^\alpha}$, we have:

$$\sum_{k'=k/2}^{k} d^{\pi^k_{\text{Hi}}}_{\text{Hi}}(s_h, \pi^*_{\text{Hi}}) \geq \sum_{k'=k/2}^{k} d^*_{\text{Hi}}(s_h, \pi^*_{\text{Hi}}) - H\sqrt{2\alpha k \log k} - \mathbb{E}[\sum_{k=1}^{K} V^*_{\text{Hi},1}(s_1) - V^{\pi^k_{\text{Hi}}}_{\text{Hi},1}(s_1)]$$
$$= \frac{k}{2} d^*_{\text{Hi}}(s_h) - H\sqrt{2\alpha k \log k} - \mathbb{E}[\sum_{k=1}^{K} V^*_{\text{Hi},1}(s_1) - V^{\pi^k_{\text{Hi}}}_{\text{Hi},1}(s_1)].$$

As a result of Thm. F.8, $H\sqrt{2\alpha k \log k} + \mathbb{E}[\sum_{k=1}^{K} V^*_{\text{Hi},1}(s_1) - V^{\pi^k_{\text{Hi}}}_{\text{Hi},1}(s_1)]$ will be sub-linear w.r.t. k, so there exists $\iota_{s_h} := c_{s_h}\text{Poly}(S, A, H, 1/\lambda, 1/\Delta_{\min}, 1/d^*_{\text{Hi}}(s_h))$ for some absolute constant c_{s_h}, such that for arbitrary $k \geq \iota_{s_h}$, $\sum_{k'=k/2}^{k} d^{\pi^k_{\text{Hi}}}_{\text{Hi}}(s_h, \pi^*_{\text{Hi}}) > 0$, which implies:

$$\forall k \geq \iota_{s_h},\ \Pr(\mathcal{E}^{\complement}_{k,2}) = \Pr(\sum_{k'=k/2}^{k} d^{\pi^k_{\text{Hi}}}_{\text{Hi}}(s_h, \pi^*_{\text{Hi}}) = 0) \leq \frac{1}{k^\alpha}.$$

Moreover, as a result of Lem. F.6, when $k \geq k^{[W]}_{occup} := c^{[W]}_{occup} \frac{C_1 + \alpha C_2}{\lambda \Delta_{\min}} \log(\frac{\alpha C_1 C_2 SAHW}{\lambda \Delta_{\min}})$, on the event of $\mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k}$ and $\mathcal{E}_{\textbf{Con},[W],k}$, if $t \in \mathcal{I}^k_{s_h}$, we must have $\max_a N_{\text{Lo},w,h}(s_h, a) > \lambda k$, which implies $\pi^*_{\text{Lo},w}(s_h) = \arg\max N^k_{\text{Lo},w,h}(s_h, a)$. Therefore, $\forall k \geq 2 \cdot k^{[W]}_{occup}$:

$$\Pr(\mathcal{E}^{\complement}_{k,3}) \leq \sum_{k'=k/2}^{k} \sum_{t \in [W]} \Pr(\{t \in \mathcal{I}^{k'}_{s_h}\} \cap \{\pi^*_{\text{Lo},w}(s_h) \neq \arg\max_a N^{k'}_{\text{Lo},w,h}(s_h, a)\})$$
$$\leq \sum_{k'=k/2}^{k} \Pr(\mathcal{E}^{\complement}_{\text{Alg}^{\text{Lo}},[W],k}) + \Pr(\mathcal{E}^{\complement}_{\textbf{Con},[W],k}) \leq \frac{4}{(k/2)^{\alpha-1}}.$$

Therefore, for arbitrary $k \geq \iota^*_{s_h} := \max\{k^{[W]}_{start}, \iota_{s_h}, k^{[W]}_{occup}\} = \text{Poly}(S, A, H, 1/\lambda, 1/\Delta_{\min}, 1/d^*_{\text{Hi}}(s_h), \log W)$, we have:

$$\Pr(\pi^k_{\text{Hi}}(s_h) = \pi^*_{\text{Hi}}(s_h)) \geq 1 - \frac{2}{(k/2)^{\alpha-1}} - \frac{2}{(k/2)^{\alpha-1}} - \frac{4}{(k/2)^{\alpha-1}} \geq 1 - \frac{8}{(k/2)^{\alpha-1}}.$$

which finishes the proof. □

Theorem F.10 (Detailed Version of Thm. 5.2)**.** *Under Assump. A, B and C, Cond. D.3 and F.4, by running Alg. 7 in Appx. F.1, with* $\varepsilon = \frac{\widetilde{\Delta}_{\min}}{4(H+1)}$*,* $\alpha > 2$ *and arbitrary* $\lambda > 0$*, we have*[5]*:*

$$
\begin{aligned}
&\textit{Regret}_K(M_{\textit{Hi}}) \\
=&O\Big(H \cdot \max\{\alpha \frac{B_1^2 H^2 S}{\lambda^2 \Delta_{\min}^2} \log(HSAWB_1B_2), \frac{(C_1+\alpha C_2)SH}{\lambda \Delta_{\min}^2} \log \frac{C_1 C_2 SAHW}{\Delta_{\min}\lambda}\} \\
&+ \sum_{h=1}^{H} \sum_{(s_h,a_h)\in \mathcal{C}_h^{\lambda,[W],1}} H \log(2SAHW(K \wedge \frac{1}{\Delta_{\min}\lambda d^*_{\textit{Hi}}(s_h)})) + \frac{B_1}{\Delta_{\textit{Hi}}(s_h,a_h)} \log(SAHB_2W(K \wedge \frac{1}{\Delta_{\min}\lambda d^*_{\textit{Hi}}(s_h)})) \\
&+ \sum_{h=1}^{H} \sum_{(s_h,a_h)\in \mathcal{C}_h^{\lambda,[W],2}} H \log(2SAHW(K \wedge \frac{1}{\Delta_{\min}\lambda d^*_{\textit{Hi},\min}})) \\
&\qquad\qquad + B_1(\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\textit{Hi}}(s_h,a_h)}) \log(SAHB_2W(K \wedge \frac{1}{\Delta_{\min}\lambda d^*_{\textit{Hi},\min}})) \\
&+ \sum_{h=1}^{H} \sum_{(s_h,a_h)\in \mathcal{C}_h^{*}} \frac{HB_1}{\Delta_{\min}} \log(SAHB_2W \min\{K, \frac{1}{\lambda \Delta_{\min} d^*_{\textit{Hi}}(s_h)}\}) \\
&+ \sum_{h=1}^{H} \sum_{(s_h,a_h)\in \mathcal{S}_h\times\mathcal{A}_h\setminus(\mathcal{C}_h^{\lambda,[W],1}\cup\mathcal{C}_h^{\lambda,[W],2})} H \log(2SAHWK) + B_1(\frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\textit{Hi}}(s_h,a_h)}) \log(B_2WK)\Big) \qquad (35) \\
=&O(SH \sum_{h=1}^{H} \sum_{(s_h,a_h)\in \mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}_h^{\lambda,[W]}} \frac{H}{\Delta_{\min}} \wedge \frac{1}{\Delta_{\textit{Hi}}(s_h,a_h)} \log(SAHWK)).
\end{aligned}
$$

Proof. Consider the same $k_{start}^{[W]} = \max\{k_{ost}^{[W]}, k_{occup}^{[W]}\}$. Similar to single task setting, we study the following two types of states.

Type 1: $(s_h,a_h) \in \mathcal{C}_h^{\lambda,[W],1} \cup \mathcal{C}_h^{\lambda,[W],2}$**: Constant Regret because of Low Visitation Probability**
For each $(s_h,a_h) \in \mathcal{C}_h^{\lambda,[W],1}$, by leveraging Lem. F.9, we have:

$$
\begin{aligned}
&\sum_{k=k_{start}^{[W]}+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},[W],k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k} \cap \mathcal{E}_{\textbf{Con},[W],k}]] \\
=&\sum_{k=k_{start}^{[W]}+1}^{\iota_{s_h}^*} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},[W],k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k} \cap \mathcal{E}_{\textbf{Con},[W],k}]] \\
=&O\Big(H\log(2SAHW(K \wedge \frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi}}(s_h)})) + \frac{B_1}{\Delta_{\text{Hi}}(s_h,a_h)} \log(SAHB_2W(K \wedge \frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi}}(s_h)}))\Big).
\end{aligned}
$$

Note that different from single task setting, the convergence speed of $\pi_{\text{Hi}}^k(s_h)$ to $\pi_{\text{Hi}}^*(s_h)$ for $s_h \in \mathcal{Z}^{\lambda,[W]}$ depends on $d_{\text{Hi}}^*(s_h)$. Therefore, for $(s_h,a_h) \in \mathcal{C}_h^{\lambda,[W],2}$, we can guarantee $d^{\pi_{\text{Hi}}^k(s_h,a_h)}$ decays to zero only after the convergence of all its ancester states in $\mathcal{Z}^{\lambda,[W]}$, and:

$$
\begin{aligned}
&\sum_{k=k_{start}^{[W]}+1}^{K} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},[W],k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k} \cap \mathcal{E}_{\textbf{Con},[W],k}]] \\
\le&\sum_{k=k_{start}^{[W]}+1}^{k_{s_h,a_h}} \mathbb{E}_{\pi_{\text{Hi}}^k}[\text{Clip}\left[\min\{H, 4b_{\text{Hi},h}^k(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH} \vee \frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right] \mathbb{I}[\mathcal{E}_{\textbf{Bonus},[W],k} \cap \mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k} \cap \mathcal{E}_{\textbf{Con},[W],k}]]
\end{aligned}
$$

[5]We only keep non-constant terms and defer the detailed result to Eq. (35).

$$=O\Big(H\log(2SAHW(K\wedge\frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi,min}}}))+B_1(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(SAHB_2W(K\wedge\frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi,min}}}))\Big).$$

where $k_{s_h,a_h}:=\max_{h'\in[h-1],s_{h'}\in\mathcal{C}^{\lambda,[W],1}_{h'}}\iota^*_{s_{h'}}$ and $d^*_{\text{Hi,min}}:=\min_{s_h:d^*_{\text{Hi}}(s_h)>0}d^*_{\text{Hi}}(s_h)$ is the minimal probability to reach state by optimal policy in Hi.

Type 2: $(s_h,a_h)\in\mathcal{C}^*_h$**, i.e.** $d^*_{\textbf{Hi}}(s_h,a_h)=d^*_{\textbf{Hi}}(s_h)>0$ Similar to the discussion of Type 2 in the proof of Thm. 4.2, when $k\geq\tau^*_{s_h}$ for some $\tau^*_{s_h}=\text{Poly}(S,A,H,\lambda^{-1},\Delta^{-1}_{\min},(d^*_{\text{Hi}}(s_h))^{-1},\log W)$, we have $\sum_{k'=1}^{k-1}d^{\pi^k_{\text{Hi}}}_{\text{Hi}}(s_h,a_h)=O(kd^*_{\text{Hi}}(s_h))$, and therefore,

$$\sum_{k=k^{[W]}_{start}+1}^{K}\mathbb{E}_{\pi^k_{\text{Hi}}}[\text{Clip}\left[\min\{H,4b^k_{\text{Hi},h}(s_h,a_h)\}\Big|\frac{\Delta_{\min}}{4eH}\vee\frac{\Delta_{\text{Hi}}(s_h,a_h)}{4e}\right]|\mathcal{E}_{\textbf{Bonus},[W],k},\mathcal{E}_{\text{Alg}^{\text{Lo}},[W],k},\mathcal{E}_{\textbf{Con},[W],k}]$$

$$=O\left(\frac{HB_1}{\Delta_{\min}}\log(SAHB_2W\min\{K,\frac{1}{\lambda\Delta_{\min}d^*_{\text{Hi}}(s_h)}\})\right).$$

Therefore, we can conclude that:

$$\begin{aligned}
&\text{Regret}_K(M_{\text{Hi}})\\
=&O\Big(H\cdot\max\{\alpha\frac{B_1^2H^2S}{\lambda^2\Delta^2_{\min}}\log(HSAWB_1B_2),\frac{(C_1+\alpha C_2)SH}{\lambda\Delta^2_{\min}}\log\frac{C_1C_2SAHW}{\Delta_{\min}\lambda}\}\\
&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{C}^{\lambda,[W],1}_h}H\log(2SAHW(K\wedge\frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi}}(s_h)}))+\frac{B_1}{\Delta_{\text{Hi}}(s_h,a_h)}\log(SAHB_2W(K\wedge\frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi}}(s_h)}))\\
&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{C}^{\lambda,[W],2}_h}H\log(2SAHW(K\wedge\frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi,min}}}))\\
&\qquad+B_1(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(SAHB_2W(K\wedge\frac{1}{\Delta_{\min}\lambda d^*_{\text{Hi,min}}}))\\
&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{C}^*_h}\frac{HB_1}{\Delta_{\min}}\log(SAHB_2W\min\{K,\frac{1}{\lambda\Delta_{\min}d^*_{\text{Hi}}(s_h)}\})\\
&+\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus(\mathcal{C}^{\lambda,[W],1}_h\cup\mathcal{C}^{\lambda,[W],2}_h)}H\log(2SAHWK)+B_1(\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)})\log(B_2WK)\Big).
\end{aligned}$$

We consider the similar Hoeffding bound for the choice of B_1 and B_2 as Example D.4, and by plugging $B_1=\Theta(SH)$ and $B_2=\Theta(SA)$ into the above equation and omitting all the terms independent with K, we have:

$$\text{Regret}_K(M_{\text{Hi}})=O(SH\sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}_h\times\mathcal{A}_h\setminus\mathcal{C}^{\lambda,[W]}_h}\frac{H}{\Delta_{\min}}\wedge\frac{1}{\Delta_{\text{Hi}}(s_h,a_h)}\log(SAHWK)).$$

□

G Missing Details for Experiments

Construction of Source and Target Tasks We first randomly construct the transition function of the high-tier task M_{Hi} (i.e. $\mathbb{P}_{\text{Hi}}$ are randomly sampled and normalized to make sure their validity). Then, similarly, we randomly construct the reward function of M_{Hi} and shift the reward function to ensure M_{Hi} has unique optimal policy and $\Delta_{\text{min,Hi}}=0.1$.

Next, we construct the source tasks by randomly permute the transition matrix. In another word, for any s_h, we randomly permute a_1,a_2,a_3 to a'_1,a'_2,a'_3 and assign $\mathbb{P}_{h,\text{Lo}}(\cdot|s_h,a'_i)\leftarrow\mathbb{P}_{h,\text{Hi}}(\cdot|s_h,a_i)$ for $i\in[3]$. In this way, the Optimal Value Dominance (OVD) condition is ensured, and we can expect some of s_h are transferable when $\pi^*_{\text{Lo}}(s_h)=\pi^*_{\text{Hi}}(s_h)$. When the number of source tasks $W>1$, we repeat the above process and construct W different source tasks.