Provably Efficient Multi-Task Reinforcement Learning with Model Transfer

Chicheng Zhang [1] **Zhi Wang** [2]

Abstract

We study multi-task reinforcement learning (RL) in tabular episodic Markov decision processes (MDPs). We formulate a heterogeneous multi-player RL problem, in which a group of players concurrently face similar but not necessarily identical MDPs, with a goal of improving their collective performance through inter-player information sharing. We design and analyze a model-based algorithm, and provide gap-dependent and gap-independent upper and lower bounds that characterize the intrinsic complexity of the problem.

1. Introduction

In many real-world applications, reinforcement learning (RL) agents can be deployed as a group to complete similar tasks at the same time. For example, in healthcare robotics, robots are paired with people with dementia to perform personalized cognitive training activities by learning their preferences (Tsiakas et al., 2016; Kubota et al., 2020); in autonomous driving, a set of autonomous vehicles learn how to navigate and avoid obstacles in various environments (Liang et al., 2019). In these settings, each learning agent alone may only be able to acquire a limited amount of data, while the agents as a group have the potential to collectively learn faster through sharing knowledge among themselves. Multi-task learning (Caruana, 1997) is a practical framework that can be used to model such settings, where a set of learning agents share/transfer knowledge to improve their collective performance.

Despite many empirical successes of multi-task RL (e.g., Zhuo et al., 2019; Liu et al., 2019; Liang et al., 2019) and transfer learning for RL (e.g., Lazaric et al., 2008; Taylor et al., 2008), a theoretical understanding of when and how information sharing or knowledge transfer can provide benefits remains limited. Exceptions include (e.g., Guo & Brunskill, 2015; Brunskill & Li, 2013; D'Eramo et al., 2020; Hu et al., 2021; Pazis & Parr, 2016; Lazaric & Restelli, 2011), which study multi-task learning from parameter or representation transfer perspectives. However, these works still do not provide a completely satisfying answer: for example, in many application scenarios, the reward structures and the environment dynamics are only slightly different for each task—this is, however, not captured by representation transfer (D'Eramo et al., 2020; Hu et al., 2021) or existing works on clustering-based parameter transfer (Guo & Brunskill, 2015; Brunskill & Li, 2013). In such settings, is it possible to design provably efficient multi-task RL algorithms that have guarantees never worse than agents learning individually, while outperforming the individual agents in favorable situations?

In this work, we formulate a multi-task RL problem that is applicable to the aforementioned settings. Specifically, inspired by a recent study on multi-task multi-armed bandits (Wang et al., 2021), we formulate the ϵ-Multi-Player Episodic Reinforcement Learning (abbreviated as ϵ-MPERL) problem, in which all tasks share the same state and action spaces, and the tasks are assumed to be similar—i.e., the dissimilarities between the environments of different tasks (specifically, the reward distributions and transition dynamics associated with the players/tasks) are bounded in terms of a dissimilarity parameter $\epsilon \geq 0$. This problem not only models concurrent RL (Silver et al., 2013a; Guo & Brunskill, 2015) as a special case by taking $\epsilon = 0$, but also captures richer multi-task RL settings when ϵ is nonzero. In this work, we study regret minimization for the ϵ-MPERL problem, specifically:

1. We identify a problem complexity notion named *subpar* state-action pairs, which captures the amenability of information sharing among tasks in ϵ-MPERL problem instances. Subpar state-action pairs, intuitively, are those that are clearly suboptimal for all tasks, for which we can robustly take advantage of (possibly biased) data collected for other tasks for faster learning.
2. We design a model-based algorithm MULTI-TASK-EULER (Algorithm 1) for the ϵ-MPERL problem, which is built upon state-of-the-art algorithms for learning single-task Markov decision processes (MDPs) (Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019) as well as ideas for model transfer in RL (Taylor et al., 2008). Specifically,

[1]University of Arizona [2]University of California, San Diego. Correspondence to: Chicheng Zhang <chichengz@cs.arizona.edu>.

ICML 2021 Workshop on Reinforcement Learning Theory.

compared with a baseline algorithm that does not utilize information sharing, MULTI-TASK-EULER has gap-dependent and gap-independent regret guarantees that: (1) is never worse, i.e., it avoids negative transfer (Rosenstein et al., 2005); (2) can be much superior when there are a large number of subpar state-action pairs.

3. Complementary to the upper bounds, we also present gap-dependent and gap-independent regret lower bounds for the ϵ-MPERL problem in terms of subpar state-action pairs.

Due to space constraints, we only present gap-independent upper and lower bounds in the main text; our gap-dependent upper and lower bounds can be found at Appendix B.

2. Preliminaries

Throughout this paper, we denote by $[n] := \{1, \ldots, n\}$. For a set A, we use A^C to denote its complement. Denote by $\Delta(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$. For functions f, g, we use $f \lesssim g$ (resp. $f \gtrsim g$) to denote that there exists some constant $c > 0$, such that $f \leq cg$ (resp. $f \geq cg$), and use $f \eqsim g$ to denote $f \lesssim g$ and $f \gtrsim g$ simultaneously. Define $a \vee b := \max(a, b)$, and $a \wedge b := \min(a, b)$. We use $\mathbb{E}$ to denote the expectation operator, and use var to denote the variance operator. Throughout, we use $\tilde{O}(\cdot)$ to hide logarithmic factors.

Multi-task RL in episodic MDPs. We have a set of M layered MDPs $\left\{\mathcal{M}_p = (H, \mathcal{S}, \mathcal{A}, p_0, \mathbb{P}_p, r_p)\right\}_{p=1}^{M}$, each associated with a player $p \in [M]$. Each MDP $\mathcal{M}_p$ is regarded as a task. The MDPs share the same episode length $H \in \mathbb{N}_+$, finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and initial state distribution $p_0 \in \Delta(\mathcal{S})$. The transition probabilities $\mathbb{P}_p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and reward distributions $r_p : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ of the players are not necessarily identical. The shared state space $\mathcal{S}$ can be partitioned into disjoint subsets $(\mathcal{S}_h)_{h=1}^{H}$, where p_0 is supported on $\mathcal{S}_1$, and for every $p \in [M]$, $h \in [H]$, and every $s \in \mathcal{S}_h, a \in \mathcal{A}$, $\mathbb{P}_p(\cdot \mid s, a)$ is supported on $\mathcal{S}_{h+1}$; here, we define $\mathcal{S}_{H+1} = \{\perp\}$ so that it contains a default terminal state $\perp$ (note that $\perp \notin \mathcal{S}$). We denote by $S := |\mathcal{S}|$ the size of the state space, and $A := |\mathcal{A}|$ the size of the action space.

Interaction process. For each episode $k \in [K]$, each player $p \in [M]$ interacts with its respective MDP $\mathcal{M}_p$; specifically, player p starts with state $s_{1,p}^k \sim p_0$, and at every step $h \in [H]$, it chooses action $a_{h,p}^k$, transitions to next state $s_{h+1,p}^k \sim \mathbb{P}_p(\cdot \mid s_{h,p}^k, a_{h,p}^k)$ and receives a stochastic immediate reward $r_{h,p}^k \sim r_p(\cdot \mid s_{h,p}^k, a_{h,p}^k)$; after all players have finished their k-th episode, they can share information.

Policy and value functions. A deterministic policy π is a mapping from $\mathcal{S}$ to $\mathcal{A}$, which can be used by players to make decisions. For player p and step h, we define the value function $V_{h,p}^{\pi} : \mathcal{S}_h \to [0, H]$ and the action value function $Q_{h,p}^{\pi} : \mathcal{S}_h \times \mathcal{A} \to [0, H]$ as the expected return of player p conditioned on its being at a state at step h, and its being at a state and taking an action at step h, respectively. $V_{h,p}$ and $Q_{h,p}$'s satisfy the Bellman equation: $\forall h \in [H]$:

$$V_{h,p}^{\pi}(s) = Q_{h,p}^{\pi}(s, \pi(s)), \;\; Q_{h,p}^{\pi}(s, a) = R_p(s, a) + (\mathbb{P}_p V_{h+1,p}^{\pi})(s, a),$$

Denote by $V_{0,p}^{\pi} = \mathbb{E}_{s_1 \sim p_0}\left[V_{1,p}^{\pi}(s_1)\right]$ π's expected reward on $\mathcal{M}_p$.

For player p, we also define its optimal value function $V_{h,p}^{\star} : \mathcal{S}_h \to [0, H]$ and the optimal action value function $Q_{h,p}^{\star} : \mathcal{S}_h \times \mathcal{A} \to [0, H]$ using the Bellman optimality equation:

$$V_{h,p}^{\star}(s) = \max_{a \in \mathcal{A}} Q_{h,p}^{\star}(s, a), \;\; Q_{h,p}^{\star}(s, a) = R_p(s, a) + (\mathbb{P}_p V_{h+1,p}^{\star})(s, a), \tag{1}$$

where $V_{H+1,p}^{\star}(\perp) := 0$. Denote by $V_{0,p}^{\star} = \mathbb{E}_{s_1 \sim p_0}\left[V_{1,p}^{\star}(s_1)\right]$ the optimal expected reward on $\mathcal{M}_p$. When it is clear from context, we will drop the subscript h in $V_{h,p}^{\pi}$, $Q_{h,p}^{\pi}$, $V_{h,p}^{\star}$, $Q_{h,p}^{\star}$.

Suboptimality gap. For player p, we define the suboptimality gap of state-action pair (s, a) as $\mathrm{gap}_p(s, a) = V_p^{\star}(s) - Q_p^{\star}(s, a)$.

Performance metric. We use collective regret to measure the players' performance: suppose for each episode k, player p executes policy $\pi^k(p)$, then the collective regret of the players is defined as: $\mathrm{Reg}(K) = \sum_{p=1}^{M} \sum_{k=1}^{K} \left(V_{0,p}^{\star} - V_{0,p}^{\pi^k(p)}\right)$.

Baseline: individual EULER. A naive baseline for multi-player RL is to let each player run a separate RL algorithm without communication. For concreteness, we choose to let each player run the state of the art EULER algorithm (Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019), which enjoys minimax gap independent (Azar et al., 2017; Dann & Brunskill, 2015) and gap-dependent regret guarantees, and refer to this strategy as individual EULER. Our goal is to design multi-task RL algorithms that can achieve collective regret strictly lower than this baseline.

Notion of similarity. Throughout this paper, we will consider the following notion of similarity between MDPs in the multi-task episodic RL setting.

Definition 1. *A collection of MDPs $(\mathcal{M}_p)_{p=1}^{M}$ is said to be ϵ-dissimilar, if for all $p, q \in [M]$, and $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|R_p(s, a) - R_q(s, a)\right| \leq \epsilon, \;\; \|\mathbb{P}_p(\cdot \mid s, a) - \mathbb{P}_q(\cdot \mid s, a)\|_1 \leq \frac{\epsilon}{H}.$$

We call such $(\mathcal{M}_p)_{p=1}^M$ an ϵ-Multi-Player Episodic Reinforcement Learning (abbrev. ϵ-MPERL) problem instance.

We have the following intuitive lemma that shows the closeness of optimal value functions of different MDPs, in terms of the dissimilarity parameter ϵ:

Lemma 2. *If $(\mathcal{M}_p)_{p=1}^M$ are ϵ-dissimilar, then for every $p, q \in [M]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \le 2H\epsilon$; consequently, $\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\right| \le 4H\epsilon$.*

3. Algorithm

We now describe our main algorithm, MULTI-TASK-EULER (Algorithm 1). Our model-based algorithm is built upon recent works on episodic RL that provide algorithms with sharp instance-dependent guarantees in the single task setting (Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019). In a nutshell, for each episode k and each player p, the algorithm performs optimistic value iteration to construct high-probability upper and lower bounds for the optimal value and action value functions $V_p^\star$ and $Q_p^\star$, and uses them to guide its exploration and decision making process.

Empirical estimates of model parameters. For each player p, the construction of its value function bound estimates relies on empirical estimates on its transition probability and expected reward function. For both estimands, we use two estimators with complementary roles, which are at two different points of the bias-variance tradeoff spectrum: one estimator uses only the player's own data (termed *individual estimate*), which is unbiased but has large variance, the other estimator uses the data collected by all players (termed *aggregate estimate*), which is biased but has lower variance. We construct individual estimates of transition probability $\hat{\mathbb{P}}_p$ and reward $\hat{R}_p$, along with aggregate estimates of transition probability $\hat{\mathbb{P}}$ and reward $\hat{R}$, using the standard maximum likelihood principle, and defer their exact formulae for to Appendix A.

Constructing value function estimates via optimistic value iteration. For each player p, based on these model parameter estimates, MULTI-TASK-EULER performs optimistic value iteration to compute the value function estimates for states at all layers (lines 6 to 21). For the terminal layer $H+1$, $V_{H+1}^\star(\perp) = 0$ trivially, so nothing needs to be done. For earlier layers $h \in [H]$, MULTI-TASK-EULER iteratively builds its value function estimates in a backward fashion. At the time of estimating values for layer h, the algorithm has already obtained optimal value estimates for layer $h+1$. Based on the Bellman optimality equation (1), MULTI-TASK-EULER estimates $(Q_p^\star(s,a))_{s\in\mathcal{S}_h, a\in\mathcal{A}}$ using model parameter estimates and its estimates of $(V_p^\star(s))_{s\in\mathcal{S}_{h+1}}$, i.e., $(\overline{V}_p(s))_{s\in\mathcal{S}_{h+1}}$ and $(\underline{V}_p(s))_{s\in\mathcal{S}_{h+1}}$. (lines 8 to 15).

Specifically, MULTI-TASK-EULER constructs estimates of $(Q_p^\star(s,a))$ for all $s \in \mathcal{S}_h, a \in \mathcal{A}$ in two different ways. First, it uses the individual estimates of model of player p to construct $\text{ind-}\underline{Q}_p$ and $\text{ind-}\overline{Q}_p$, upper and lower bound estimates of $Q_p^\star$ (lines 9 and 10); this construction is reminis-

Algorithm 1 MULTI-TASK-EULER

1: **Input:** Failure probability $\delta \in (0,1)$.
2: **Initialize:** Set $V_p(\perp) = 0$ for all p in $[M]$, where $\perp$ is the only state in $\mathcal{S}_{H+1}$.
3: **for** $k = 1, 2, \ldots, K$ **do**
4:   **for** $p = 1, 2, \ldots, M$ **do**
5:     Construct optimal value estimates for player p:
6:     **for** $h = H, H-1, \ldots, 1$ **do**
7:       **for** $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ **do**
8:         Compute:
9:         $\text{ind-}\overline{Q}_p(s,a) = \hat{R}_p(s,a) + (\hat{\mathbb{P}}_p\overline{V}_p)(s,a) + \text{ind-}b_p(s,a)$;
10:         $\text{ind-}\underline{Q}_p(s,a) = \hat{R}_p(s,a) + (\hat{\mathbb{P}}_p\underline{V}_p)(s,a) - \text{ind-}b_p(s,a)$;
11:         $\text{agg-}\overline{Q}_p(s,a) = \hat{R}(s,a) + (\hat{\mathbb{P}}\overline{V}_p)(s,a) + \text{agg-}b_p(s,a)$;
12:         $\text{agg-}\underline{Q}_p(s,a) = \hat{R}(s,a) + (\hat{\mathbb{P}}\underline{V}_p)(s,a) - \text{agg-}b_p(s,a)$.
13:         Update optimal action value function upper and lower bound estimates:
14:         $\overline{Q}_p(s,a) = (H-h+1) \wedge \text{ind-}\overline{Q}_p(s,a) \wedge \text{ind-}\underline{Q}_p(s,a)$;
15:         $\underline{Q}_p(s,a) = 0 \vee \text{ind-}\underline{Q}_p(s,a) \vee \text{agg-}\underline{Q}_p(s,a)$.
16:       **end for**
17:       **for** $s \in \mathcal{S}_h$ **do**
18:         Define $\pi^k(p)(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, \overline{Q}_p(s,a)$;
19:         Update $\overline{V}_p(s) = \overline{Q}_p(s, \pi^k(p)(s))$, $\underline{V}_p(s) = \underline{Q}_p(s, \pi^k(p)(s))$.
20:       **end for**
21:     **end for**
22:     All players p interact with their respective environments, and update reward and transition estimates:
23:     **for** $p = 1, 2, \ldots, M$ **do**
24:       Player p executes policy $\pi^k(p)$ on $\mathcal{M}_p$ and obtains trajectory $(s_{h,p}^k, a_{h,p}^k, r_{h,p}^k)_{h=1}^H$;
25:       Update individual estimates of transition probability $\hat{\mathbb{P}}_p$, reward $\hat{R}_p$ and count $n_p(\cdot,\cdot)$.
26:     **end for**
27:     Update aggregate estimates of transition probability $\hat{\mathbb{P}}$, reward $\hat{R}$ and count $n(\cdot,\cdot)$.
28:   **end for**
29: **end for**

cent of EULER and STRONG-EULER (Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019), in that if we were to use $\underline{\text{ind-}Q}_p$ and $\overline{\text{ind-}Q}_p$ as our optimal action value function estimate $\overline{Q}$ and $\underline{Q}$, our algorithm becomes individual EULER. Second, it uses the aggregate estimate of model to construct $\underline{\text{agg-}Q}_p$ and $\overline{\text{agg-}Q}_p$, also upper and lower bound estimates of $Q_p^\star$ (lines 11 and 12); this construction is unique to multitask RL, and is our main algorithmic contribution.

To ensure that $\overline{\text{agg-}Q}_p$ and $\overline{\text{ind-}Q}_p$ (resp. $\underline{\text{agg-}Q}_p$ and $\underline{\text{ind-}Q}_p$) are valid upper bounds (resp. lower bounds) of $Q_p^\star$, MULTI-TASK-EULER adds bonus terms ind-$b_p(s,a)$ and ind-$b_p(s,a)$, respectively, in the optimistic value iteration process, to account for estimation error of the model estimates against the true models. Specifically, ind-$b_p(s,a)$ has the same form as the bonus term of STRONG-EULER; agg-$b_p(s,a)$ accounts the bias of $\hat{\mathbb{P}}$ and $\hat{R}$ in estimating $\mathbb{P}_p$ and R_p, taking into account the ϵ-dissimilarity assumption of the $\mathcal{M}_p$'s to ensure that $\overline{\text{ind-}Q}$ and $\underline{\text{ind-}Q}$ are indeed valid upper and lower bounds of $Q_p^\star$. The formulae of ind-$b_p(s,a)$ and agg-$b_p(s,a)$, and a more detailed motivation of their constructions can be found at Appendix A. By taking intersections of all confidence bounds of $Q_p^\star$ it has obtained, MULTI-TASK-EULER constructs its final upper and lower bound estimates for $Q_p^\star$, $\overline{Q}_p$ and $\underline{Q}_p$, respectively for $s \in \mathcal{S}_h$ (line 14 to 15).[1] Based on these, it also constructs upper and lower bound estimates for $V_p^\star$, $\overline{V}_p$ and $\underline{V}_p$, respectively for $s \in \mathcal{S}_h$ (line 19).

Executing optimistic policies. For each episode k and each player p, its optimal action-value function upper bound estimate $\overline{Q}_p$ induces a greedy policy $\pi^k(p) : s \mapsto \mathrm{argmax}_{a\in\mathcal{A}} \overline{Q}_p(s,a)$ (line 18); the player then executes this policy at this episode to collect a new trajectory and use this to update its individual model parameter estimates. After all players finish their episode k, the algorithm also updates its aggregate model parameter estimates (line 23 to 27) and continues to the next episode.

4. Performance Guarantees

We provide performance guarantees of Algorithm 1, which centers around an instance-dependent complexity measure that characterizes the amenability of information sharing. This measure generalizes the notion of subpar arms studied in multi-task multi-armed bandit learning (Wang et al., 2021), and is defined as follows:

Definition 3. *Define the set of subpar state-action pairs as* $\mathcal{I}_\epsilon := \Big\{(s,a) \in \mathcal{S}\times\mathcal{A} : \exists p \in [M], \mathrm{gap}_p(s,a) \geq 96H\epsilon\Big\}$.

[1]Similar ideas on using multiple data sources to construct confidence intervals and guide explorations have been explored by e.g. Soare et al. (2014) in multi-task linear contextual bandits.

4.1. Gap-independent upper and lower bounds

In the interest of space, we only present gap-independent upper and lower bounds here in this extended abstract, and we defer the gap-dependent upper and lower bounds on the collective regret, as well as all the proofs, to the appendices.

Theorem 4 (Gap-independent upper bound)**.** *If* $\{\mathcal{M}_p\}_{p=1}^M$ *are ϵ-dissimilar, then* MULTI-TASK-EULER *satisfies that, with probability* $1-\delta$,

$$\mathrm{Reg}(K) \leq \tilde{O}\left(M\sqrt{H^2|\mathcal{I}_\epsilon^C|K} + \sqrt{MH^2|\mathcal{I}_\epsilon|K} + MH^3S^2A\right).$$

We compare this regret upper bound with that of the baseline, individual STRONG-EULER. By summing over the regret guarantees for each individual player, we have that, with probability $1-\delta$,

$$\mathrm{Reg}(K) \leq \tilde{O}\left(M\sqrt{H^2SAK} + MH^3S^2A\right).$$

We focus on the comparison on the leading terms, i.e., the $\sqrt{K}$ terms. As $M\sqrt{H^2SAK} \approx M\sqrt{H^2|\mathcal{I}_\epsilon|\,K} + M\sqrt{H^2\big|\mathcal{I}_\epsilon^C\big|\,K}$, we see that an improvement in the collective regret bound comes from the contributions from the subpar state-action pairs: the $M\sqrt{H^2|\mathcal{I}_\epsilon|K}$ term is reduced to $\sqrt{MH^2|\mathcal{I}_\epsilon|K}$, which is an improvement by a factor of $\sqrt{M}$. Moreover, if $\big|\mathcal{I}_\epsilon^C\big| \ll SA$ and $M \gg 1$, MULTI-TASK-EULER provides a regret bound of lower order than individual EULER.

In addition to the above upper bound, we present a gap-independent lower bound that also depends on our subpar state-action pair notion.[2]

Theorem 5 (Gap-independent lower bound)**.** *For any* $A \geq 2$, $H \geq 2$, $S \geq 4H$, $K \geq SA$, $M \in \mathbb{N}$, *and* $l, l^C \in \mathbb{N}$ *such that* $l + l^C = SA$ *and* $l \leq SA - 4(S+HA)$, *there exists some ϵ such that for any algorithm* Alg, *there exists an ϵ-MPERL problem instance with S states, A actions, M players and an episode length of H such that* $\Big|\mathcal{I}_{\frac{\epsilon}{192H}}\Big| \geq l$, *and*

$$\mathbb{E}\Big[\mathrm{Reg}_{\mathrm{Alg}}(K)\Big] \geq \Omega\left(M\sqrt{H^2l^CK} + \sqrt{MH^2lK}\right).$$

Comparing the lower bound with Theorem 4, we see that the upper and lower bounds nearly match for any constant H. When H is large, the key difference between the regret bounds is that the former are in terms of $\mathcal{I}_\epsilon$, while the latter are in terms of $\mathcal{I}_{\Theta(\frac{\epsilon}{H})}$. We conjecture that our upper bounds can be improved by replacing $\mathcal{I}_\epsilon$ with $\mathcal{I}_{\Theta(\frac{\epsilon}{H})}$—our analysis

[2]Our gap-dependent and gap-independent lower bounds are inspired by regret bounds for episodic RL (Simchowitz & Jamieson, 2019; Dann & Brunskill, 2015) and multi-task bandits (Wang et al., 2021).

uses a clipping trick similar to (Simchowitz & Jamieson, 2019), which may be the reason for a suboptimal dependence on H; we leave closing this gap as an open question.

5. Related Work

Regret minimization for MDPs. Our work belongs to the literature of regret minimization for MDPs (e.g., Bartlett & Tewari, 2009; Jaksch et al., 2010; Dann & Brunskill, 2015; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Dann et al., 2019; Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019; Zhang et al., 2020; Yang et al., 2021; Xu et al., 2021). In the episodic setting, (Azar et al., 2017; Dann et al., 2019; Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019; Zhang et al., 2020) achieve minimax $\sqrt{H^2 SAK}$ regret bounds for general stationary MDPs. Furthermore, the EULER algorithm (Zanette & Brunskill, 2019) achieves adaptive problem-dependent regret guarantees when the total reward within an episode is small or when the environmental norm of the MDP is small. Simchowitz & Jamieson (2019) refine EULER, proposing STRONG-EULER that provides more fine-grained gap-dependent $O(\log K)$ regret guarantees. Yang et al. (2021); Xu et al. (2021) show that the optimistic Q-learning algorithm (Jin et al., 2018) and its variants can also achieve gap-dependent logarithmic regret guarantees. Remarkably, Xu et al. (2021) achieve a regret bound that improves over that of (Simchowitz & Jamieson, 2019), in that it replaces the dependence on the number of optimal state-action pairs with the number of non-unique state-action pairs.

Transfer and lifelong learning for RL. A considerable portion of related works concerns transfer learning for RL tasks (see Taylor & Stone, 2009; Lazaric, 2012; Zhu et al., 2020, for surveys from different angles), and many studies investigate a batch setting: given some source tasks and target tasks, transfer learning agents have access to batch data collected for the source tasks (and sometimes for the target tasks as well). In this setting, model-based approaches have been explored by Taylor et al. (e.g., 2008); theoretical guarantees for transfer of samples across tasks have been established in (e.g., Lazaric & Restelli, 2011; Tirinzoni et al., 2019). Similarly, sequential transfer has been studied under the framework of lifelong RL in (e.g., Tanaka & Yamamura, 2003; Abel et al., 2018; Garcia & Thomas, 2019; Landolfi et al., 2019)—in this setting, an agent faces a sequence of RL tasks and aims to take advantage of knowledge gained from previous tasks for better performance in future tasks; in particular, analyses on the sample complexity of transfer learning algorithms are presented in (Brunskill & Li, 2013; Liu et al., 2016) under the assumption that an upper bound on the total number of unique (and well-separated) RL tasks is known. We note that, in contrast, we study an online setting in which no prior data are available and multiple RL tasks are learned concurrently by RL agents.

Concurrent RL. Data sharing between multiple RL agents that learn concurrently has also been investigated. In (e.g., Kretchmar, 2002; Silver et al., 2013b; Guo & Brunskill, 2015; Dimakopoulou & Van Roy, 2018), a group of agents interact in parallel with *identical* environments. Another setting is studied in (Guo & Brunskill, 2015), in which agents solve different RL tasks (MDPs); however, similar to (Brunskill & Li, 2013; Liu et al., 2016), it is assumed that there is a finite number of unique tasks, and different tasks are well-separated, i.e., there is a minimum gap. In this work, we assume that players face similar but not necessarily identical MDPs, and we do not assume a minimum gap. Hu et al. (2021) study multi-task RL with linear function approximation with representation transfer, where it is assumed that the optimal value functions of all tasks are from a low dimensional linear subspace. Our setting and results are the most similar to (Pazis & Parr, 2016) and (Dubey & Pentland, 2021). Pazis & Parr (2016) study concurrent exploration in similar MDPs with continuous states in the PAC setting; however, their PAC guarantee does not hold for target error rate arbitrarily close to zero; in contrast, our algorithm has a fall-back guarantee, in that it always has a sublinear regret. Concurrent RL from similar *linear* MDPs has also been recently studied by Dubey & Pentland (2021): under the assumption of small heterogeneity between different MDPs (a setting very similar to ours), the provided regret guarantee involves a term that is linear in the number of episodes, whereas our algorithm in this paper always has a sublinear regret; concurrent RL under the assumption of large heterogeneity is also studied but additional contextual information is assumed to be available for the players to ensure a sublinear regret.

Multi-agent RL. In many multi-agent RL models (Zhang et al., 2019; OroojlooyJadid & Hajinezhad, 2019), a set of learning agents interact with a common environment and have shared global states; in particular, Zhang et al. (2018) study the setting with heterogeneous reward distributions, and provides convergence guarantees for two policy gradient-based algorithms. In contrast, in our setting, our learning agents interact with separate environments.

6. Acknowledgements

We thank Kamalika Chaudhuri for helpful initial discussions, and thank Akshay Krishnamurthy and Tongyi Cao for discussing the applicability of adaptive RL in metric spaces to this problem. CZ acknowledges startup funding support from the University of Arizona. ZW thanks the National Science Foundation under IIS 1915734 and CCF 1719133 for research support.

References

Abel, D., Jinnai, Y., Guo, S. Y., Konidaris, G., and Littman, M. Policy and value transfer in lifelong reinforcement learning. In *International Conference on Machine Learning*, pp. 20–29. PMLR, 2018.

Audibert, J.-Y., Munos, R., and Szepesvári, C. Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, pp. 150–165. Springer, 2007.

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

Bartlett, P., Dani, V., Hayes, T., Kakade, S., Rakhlin, A., and Tewari, A. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pp. 335–342. Omnipress, 2008.

Bartlett, P. L. and Tewari, A. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 35–42, 2009.

Brunskill, E. and Li, L. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 122–131, 2013.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2818–2826, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5717–5727, 2017.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

D'Eramo, C., Tateo, D., Bonarini, A., Restelli, M., and Peters, J. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgpv2VFvr.

Dimakopoulou, M. and Van Roy, B. Coordinated exploration in concurrent reinforcement learning. In *International Conference on Machine Learning*, pp. 1271–1279. PMLR, 2018.

Dubey, A. and Pentland, A. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Garcia, F. M. and Thomas, P. S. A meta-mdp approach to exploration for lifelong reinforcement learning. *arXiv preprint arXiv:1902.00843*, 2019.

Guo, Z. and Brunskill, E. Concurrent pac rl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. Near-optimal representation learning for linear bandits and linear rl. *arXiv preprint arXiv:2102.04132*, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4868–4878, 2018.

Kretchmar, R. M. Parallel reinforcement learning. In *The 6th World Conference on Systemics, Cybernetics, and Informatics*. Citeseer, 2002.

Kubota, A., Peterson, E. I., Rajendren, V., Kress-Gazit, H., and Riek, L. D. Jessie: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 121–130, 2020.

Landolfi, N. C., Thomas, G., and Ma, T. A model-based approach for sample-efficient multi-task reinforcement learning. *arXiv preprint arXiv:1907.04964*, 2019.

Lazaric, A. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pp. 143–173. Springer, 2012.

Lazaric, A. and Restelli, M. Transfer from multiple mdps. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Lazaric, A., Restelli, M., and Bonarini, A. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 544–551, 2008.

Liang, X., Liu, Y., Chen, T., Liu, M., and Yang, Q. Federated transfer reinforcement learning for autonomous driving. *arXiv preprint arXiv:1910.06001*, 2019.

Liu, B., Wang, L., and Liu, M. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019.

Liu, Y., Guo, Z., and Brunskill, E. Pac continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, pp. 438–446, 2016.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *COLT*, 2009.

OroojlooyJadid, A. and Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

Pazis, J. and Parr, R. Efficient pac-optimal exploration in concurrent, continuous state mdps with delayed updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, 2005.

Silver, D., Newnham, L., Barker, D., Weller, S., and McFall, J. Concurrent reinforcement learning from customer interactions. In *International conference on machine learning*, pp. 924–932. PMLR, 2013a.

Silver, D., Newnham, L., Barker, D., Weller, S., and McFall, J. Concurrent reinforcement learning from customer interactions. In *International conference on machine learning*, pp. 924–932. PMLR, 2013b.

Simchowitz, M. and Jamieson, K. Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*, 2019.

Soare, M., Alsharif, O., Lazaric, A., and Pineau, J. Multi-task linear bandits. *NIPS2014 Workshop on Transfer and Multi-task Learning : Theory meets Practice*, 2014.

Tanaka, F. and Yamamura, M. Multitask reinforcement learning on the distribution of mdps. In *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, volume 3, pp. 1108–1113. IEEE, 2003.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Taylor, M. E., Jong, N. K., and Stone, P. Transferring instances for model-based reinforcement learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 488–505. Springer, 2008.

Tirinzoni, A., Salvini, M., and Restelli, M. Transfer of samples in policy search via multiple importance sampling. In *International Conference on Machine Learning*, pp. 6264–6274. PMLR, 2019.

Tsiakas, K., Abellanoza, C., and Makedon, F. Interactive learning and adaptation for robot assisted therapy for people with dementia. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 1–4, 2016.

Wang, Z., Zhang, C., Singh, M. K., Riek, L., and Chaudhuri, K. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1531–1539. PMLR, 2021.

Xu, H., Ma, T., and Du, S. S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Basar, T. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pp. 5872–5881. PMLR, 2018.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhu, Z., Lin, K., and Zhou, J. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020.

Zhuo, H. H., Feng, W., Xu, Q., Yang, Q., and Lin, Y. Federated reinforcement learning. *arXiv preprint arXiv:1901.08277*, 2019.

A. Deferred Materials from Section 3

Empirical estimates of model parameters. At the end of episode k, for every $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, the individual and aggregate estimates of immediate reward $R(s,a)$ are defined formally as:

$$\hat{R}_p(s,a) := \frac{\sum_{l=1}^{k} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right) r_{h,p}^l}{n_p(s,a)},$$
$$\hat{R}(s,a) := \frac{\sum_{l=1}^{k} \sum_{p=1}^{M} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right) r_{h,p}^l}{n(s,a)},$$

where $n_p(s,a) := \sum_{l=1}^{k} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right)$, and $n(s,a) := \sum_{l=1}^{k} \sum_{p=1}^{M} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l) = (s,a)\right)$.

Similarly, for every $h \in [H]$ and $(s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$, we also define the individual and aggregate estimates of transition probability as:

$$\hat{\mathbb{P}}_p(s' \mid s,a) := \frac{\sum_{l=1}^{k} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l, s_{h+1,p}^l) = (s,a,s')\right)}{n_p(s,a)}.$$
$$\hat{\mathbb{P}}(s' \mid s,a) := \frac{\sum_{l=1}^{k} \sum_{p=1}^{M} \mathbf{1}\left((s_{h,p}^l, a_{h,p}^l, s_{h+1,p}^l) = (s,a,s')\right)}{n(s,a)},$$

If $n(s,a) = 0$, we define $\hat{R}(s,a) := 0$ and $\hat{\mathbb{P}}(s' \mid s,a) := \frac{1}{|\mathcal{S}_{h+1}|}$; and if $n_p(s,a) = 0$, we define $\hat{R}_p(s,a) := 0$ and $\hat{\mathbb{P}}(s' \mid s,a) := \frac{1}{|\mathcal{S}_{h+1}|}$. The counts and reward estimates can be maintained by MULTI-TASK-EULER efficiently in an incremental manner.

Additional details on constructions of value function estimates. To ensure that $\overline{\text{agg-}Q}_p$ and $\overline{\text{ind-}Q}_p$ (resp. $\underline{\text{agg-}Q}_p$ and $\underline{\text{ind-}Q}_p$) are valid upper bounds (resp. lower bounds) of $Q_p^\star$, MULTI-TASK-EULER adds bonus terms ind-$b_p(s,a)$ and agg-$b_p(s,a)$, respectively, in the optimistic value iteration process, to account for estimation error of the model estimates against the true models. Specifically, both bonus terms are composed of three parts:

$$\text{ind-}b_p(s,a) := b_{\text{rw}}(n_p(s,a), 0) + b_{\text{prob}}(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0) + b_{\text{str}}(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0),$$
$$\text{agg-}b_p(s,a) := b_{\text{rw}}(n(s,a), \epsilon) + b_{\text{prob}}(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon) + b_{\text{str}}(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon),$$

where

$$b_{\text{rw}}(n,\kappa) := 1 \;\; \wedge \;\; \kappa + \Theta\left(\sqrt{\frac{L(n)}{n}}\right),$$

$$b_{\text{prob}}(q,n,\overline{V},\underline{V},\kappa) := H \;\; \wedge \;\; 2\kappa + \Theta\left(\sqrt{\frac{\text{var}_{s'\sim q}\left[\overline{V}(s')\right] L(n)}{n}} + \sqrt{\frac{\mathbb{E}_{s'\sim q}\left[(\overline{V}(s') - \underline{V}(s'))^2\right] L(n)}{n}} + \frac{HL(n)}{n}\right),$$

$$b_{\text{str}}(q,n,\overline{V},\underline{V},\kappa) := \kappa + \Theta\left(\sqrt{\frac{S\,\mathbb{E}_{s'\sim q}\left[(\overline{V}(s') - \underline{V}(s'))^2\right] L(n)}{n}} + \frac{HSL(n)}{n}\right),$$

and $L(n) \eqsim \ln(\frac{MSAn}{\delta})$.

The three components in the bonus terms serve for different purposes:

1. The first component accounts for the uncertainty in the reward estimation: with probability $1 - O(\delta)$, $\left|\hat{R}_p(s,a) - R_p(s,a)\right| \le b_{\text{rw}}(n_p(s,a), 0)$, and $\left|\hat{R}(s,a) - R_p(s,a)\right| \le b_{\text{rw}}(n(s,a), \epsilon)$.

2. The second component accounts for the uncertainty in estimating $(\mathbb{P}_p V_p^\star)(s,a)$: with probability $1-O(\delta)$, $\left|(\hat{\mathbb{P}}_p V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| \leq b_{\mathrm{prob}}(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0)$ and $\left|(\hat{\mathbb{P}} V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| \leq b_{\mathrm{prob}}(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon)$.

3. The third component accounts for the lower order terms to ensure strong optimism (Simchowitz & Jamieson, 2019): with probability $1-O(\delta)$, $\left|(\hat{\mathbb{P}}_p - \mathbb{P}_p)(\overline{V}_p - V_p^\star)(s,a)\right| \leq b_{\mathrm{str}}(\hat{\mathbb{P}}_p(\cdot \mid s,a), n_p(s,a), \overline{V}_p, \underline{V}_p, 0)$, and $\left|(\hat{\mathbb{P}} - \mathbb{P}_p)(\overline{V}_p - V_p^\star)(s,a)\right| \leq b_{\mathrm{prob}}(\hat{\mathbb{P}}(\cdot \mid s,a), n(s,a), \overline{V}_p, \underline{V}_p, \epsilon)$.

Based on the above concentration inequalities and the definitions of bonus terms, it can be shown inductively that, with probability $1-O(\delta)$, both $\overline{\text{agg-}Q}_p$ and $\overline{\text{ind-}Q}_p$ (resp. $\underline{\text{agg-}Q}_p$ and $\underline{\text{ind-}Q}_p$) are valid upper bounds (resp. lower bounds) of $Q_p^\star$.

B. Deferred Materials from Section 4

Here, we present a pair of gap-dependent upper and lower bounds on the collective regret. To ease our presentation, we first provide two helpful lemmas.

Lemma 6. *If $(\mathcal{M}_p)_{p=1}^M$ are ϵ-dissimilar, then for every $p, q \in [M]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \leq 2H\epsilon$; consequently, $\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\right| \leq 4H\epsilon$.*

Lemma 7. *For any $(s,a) \in \mathcal{I}_\epsilon$, we have that: (1) for all $p \in [M]$, $(s,a) \notin Z_{p,\mathrm{opt}}$, where $Z_{p,\mathrm{opt}} := \left\{(s,a) : \mathrm{gap}_p(s,a) = 0\right\}$ is the set of optimal state-action pairs with respect to p; (2) for all $p, q \in [M]$, $\mathrm{gap}_p(s,a) \geq \frac{1}{2}\mathrm{gap}_q(s,a)$.*

Lemma 7 implies that, for every player p, the state-action space $\mathcal{S} \times \mathcal{A}$ can be partitioned to three disjoint sets: $\mathcal{I}_\epsilon$, $Z_{p,\mathrm{opt}}$, $(\mathcal{I}_\epsilon \cup Z_{p,\mathrm{opt}})^C$; furthermore, for any subpar (s,a), its suboptimality gaps with respect to all players are within a constant of each other.

We also define the minimum suboptimality gap of player p as $\mathrm{gap}_{p,\min} = \min_{(s,a):\mathrm{gap}_p(s,a)>0} \mathrm{gap}_p(s,a)$, and the minimum suboptimaily gap over all players as $\mathrm{gap}_{\min} = \min_{p\in[M]} \mathrm{gap}_{p,\min}$. We are now ready to present the gap-dependent upper and lower bounds.

Theorem 8 (Gap-dependent upper bound)**.** *If $\left\{\mathcal{M}_p\right\}_{p=1}^M$ are ϵ-dissimilar, then* MULTI-TASK-EULER *satisfies, with probability $1-\delta$,*

$$\mathrm{Reg}(K) \lesssim \ln(\frac{MSAK}{\delta}) \left(\sum_{p\in[M]} \left(\sum_{(s,a)\in Z_{p,\mathrm{opt}}} \frac{H^3}{\mathrm{gap}_{p,\min}} + \sum_{(s,a)\in(\mathcal{I}_\epsilon \cup Z_{p,\mathrm{opt}})^C} \frac{H^3}{\mathrm{gap}_p(s,a)} \right) + \sum_{(s,a)\in\mathcal{I}_\epsilon} \frac{H^3}{\min_p \mathrm{gap}_p(s,a)} \right) + \ln(\frac{MSAK}{\delta}) \cdot MH^3S^2A \ln\frac{MHSA}{\mathrm{gap}_{\min}}.$$

We compare this regret bound with the regret bound obtained by the individual STRONG-EULER baseline. By summing over the regret guarantees of STRONG-EULER for all players, we deduce that individual STRONG-EULER guarantees a collective regret bound of

$$\mathrm{Reg}(K) \lesssim \ln(\frac{MSAK}{\delta}) \left(\sum_{p \in [M]} \left(\sum_{(s,a) \in Z_{p,\mathrm{opt}}} \frac{H^3}{\mathrm{gap}_{p,\min}} + \sum_{(s,a) \in (\mathcal{I}_\epsilon \cup Z_{p,\mathrm{opt}})^C} \frac{H^3}{\mathrm{gap}_p(s,a)} \right) + \sum_{(s,a) \in \mathcal{I}_\epsilon} \sum_{p \in [M]} \frac{H^3}{\mathrm{gap}_p(s,a)} \right) + \ln(\frac{MSAK}{\delta}) \cdot MH^3 S^2 A \ln \frac{MHSA}{\mathrm{gap}_{\min}},$$

that holds with probability $1 - \delta$.

We again focus on comparing the leading terms, i.e., the terms that have polynomial dependences on the suboptimality gaps in the above two bounds. It can be seen that an improvement in the regret bound by MULTI-TASK-EULER comes from the contributions from the subpar state-action pairs: for each $(s,a) \in \mathcal{I}_\epsilon$, the regret bound is reduced from $\sum_{p \in [M]} \frac{H^3}{\mathrm{gap}_p(s,a)}$ to $\frac{H^3}{\min_p \mathrm{gap}_p(s,a)}$, which is an improvement by a factor of $O(\frac{1}{M})$. Recent work of Xu et al. (2021) has shown that in the single-task setting, it is possible to replace $\sum_{(s,a) \in Z_{p,\mathrm{opt}}} \frac{H^3}{\mathrm{gap}_{p,\min}}$ with a sharper problem-dependent complexity term that depends on the multiplicity of optimal state-action pairs. We leave improving the guarantee of Theorem 8 in a similar manner as an interesting open problem.

Complementary to the upper bound, we also present a gap-dependent lower bound. Before that, we first formally define the notion of sublinear regret algorithms: for any fixed ϵ, we say that an algorithm Alg is a sublinear regret algorithm for the ϵ-MPERL problem if there exists some $C > 0$ and $\alpha < 1$ such that $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \leq CK^\alpha$.

Theorem 9 (Gap-dependent lower bound). *Fix $\epsilon \geq 0$. For any $S \in \mathbb{N}$, $A \geq 2$, $H \geq 2$, $M \in \mathbb{N}$, such that $S \geq 2(H-1)$, let $S_1 = S - 2(H-1)$; and let $\left\{\Delta_{s,a,p}\right\}_{(s,a,p) \in [S_1] \times [A] \times [M]}$ be any set of values such that*

- *for every $(s,a,p) \in [S_1] \times [A] \times [M]$, $\Delta_{s,a,p} \in [0, H/48]$;*
- *for every $(s,p) \in [S_1] \times [M]$, there exists at least one action $a \in [A]$ such that $\Delta_{s,a,p} = 0$;*
- *and, for every $(s,a) \in [S_1] \times [A]$ and $p, q \in [M]$, $\left|\Delta_{s,a,p} - \Delta_{s,a,q}\right| \leq \epsilon/4$.*

There exists an ϵ-MPERL problem instance with S states, A actions, M players and an episode length of H, such that $\mathcal{S}_1 = [S_1]$, $|\mathcal{S}_h| = 2$ for all $h \geq 2$, and

$$\mathrm{gap}_p(s,a) = \Delta_{s,a,p}, \quad \forall (s,a,p) \in [S_1] \times [A] \times [M].$$

For this problem instance, any sublinear regret algorithm Alg for the ϵ-MPERL problem must have regret at least

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(\ln K \left(\sum_{p \in [M]} \sum_{\substack{(s,a) \in \mathcal{I}^C_{(\epsilon/192H)}: \\ \mathrm{gap}_p(s,a) > 0}} \frac{H^2}{\mathrm{gap}_p(s,a)} + \sum_{(s,a) \in \mathcal{I}_{(\epsilon/192H)}} \frac{H^2}{\min_p \mathrm{gap}_p(s,a)} \right)\right).$$

Again, we note that the upper and lower bounds nearly match for any constant H, and we leave closing the gap between the upper and lower bounds for large H as an open problem.

C. Proofs of Lemmas 6 and 7

C.1. Proof of Lemma 6

Lemma 6. *If $(\mathcal{M}_p)_{p=1}^M$ is ϵ-dissimilar, then for every $p, q \in [M]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left| Q^\star_p(s,a) - Q^\star_q(s,a) \right| \leq 2H\epsilon,$$

consequently, $\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\right| \leq 4H\epsilon$.

Proof. For the first claim, we prove a stronger statement by backward induction on h, namely, for every $p, q \in [M]$, every $h \in [1, H]$, and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \leq 2(H-h+1)\epsilon.$$

Base case: For $h = H+1$, we have $Q_p^\star(s,a) = 0$ for every $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, and $p \in [M]$. It follows trivially that $\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| = 0 \leq 2(H-h+1)\epsilon$.

Inductive case: Suppose by inductive hypothesis that for some $h \in [1, H]$ and, for every $(s,a) \in \mathcal{S}_{h+1} \times \mathcal{A}$ and $p, q \in [M]$, $\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \leq 2(H-h)\epsilon$.

We first prove the following auxiliary statement: for every $s \in \mathcal{S}_{h+1}$ and $p, q \in [M]$,

$$\left|V_p^\star(s) - V_q^\star(s)\right| \leq 2(H-h)\epsilon. \tag{2}$$

Let $a_p = \mathrm{argmax}_{a \in \mathcal{A}} Q_p^\star(s,a)$ and $a_q = \mathrm{argmax}_{a \in \mathcal{A}} Q_q^\star(s,a)$. The above auxiliary statement can be easily proven by contradiction: without loss of generality, suppose that $V_p^\star(s) - V_q^\star(s) = Q_p^\star(s,a_p) - Q_q^\star(s,a_q) > 2(H-h)\epsilon$. Since $Q_q^\star(s,a_p) \geq Q_p^\star(s,a_p) - 2(H-h)\epsilon$, it follows that $Q_q^\star(s,a_p) > Q_q^\star(s,a_q)$, which contradicts the fact that $a_q = \mathrm{argmax}_{a \in \mathcal{A}} Q_q^\star(s,a)$.

We now return to the inductive proof, and we show that given the inductive hypothesis, for every $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ and $p, q \in [M]$,

$$\begin{aligned}
&\left|Q_p^\star(s,a) - Q_q^\star(s,a)\right| \\
\leq &\left|R_p(s,a) - R_q(s,a)\right| + \left|\sum_{s' \in \mathcal{S}_{h+1}} \left[\mathbb{P}_p(s' \mid s,a) V_p^\star(s') - \mathbb{P}_q(s' \mid s,a) V_q^\star(s')\right]\right| \\
\leq &\epsilon + \left|\sum_{s' \in \mathcal{S}_{h+1}} \left[\mathbb{P}_p(s' \mid s,a) V_p^\star(s') - \mathbb{P}_q(s' \mid s,a) V_p^\star(s')\right]\right| + \left|\sum_{s' \in \mathcal{S}_{h+1}} \mathbb{P}_q(s' \mid s,a) \left(V_p^\star(s') - V_q^\star(s')\right)\right| \\
\leq &\epsilon + \left\|\mathbb{P}_p(\cdot \mid s,a) - \mathbb{P}_q(\cdot \mid s,a))\right\|_1 \left(\max_{s' \in \mathcal{S}_{h+1}} \left|V_p^\star(s')\right|\right) + \left\|\mathbb{P}_q(\cdot \mid s,a)\right\|_1 \left(\max_{s' \in \mathcal{S}_{h+1}} \left|V_p^\star(s') - V_q^\star(s')\right|\right) \\
\leq &\epsilon + \frac{\epsilon}{H} \cdot H + 2(H-h)\epsilon \\
= &2(H-h+1)\epsilon,
\end{aligned}$$

where the first inequality follows from Eq. (1) and the triangle inequality; the second inequality follows from Definition 1 and the triangle inequality; the third inequality follows from Hölder's inequality; and the fourth inequality uses Definition 1 and Eq. (2).

For the second claim, we note that from the first claim, we have for any p, q, s,

$$\left|V_p^\star(s) - V_q^\star(s)\right| = \left|\max_{a \in \mathcal{A}} Q_p^\star(s,a) - \max_{a \in \mathcal{A}} Q_p^\star(s,a)\right| \leq 2H\epsilon,$$

therefore, for any p, q, s, a,

$$\left|\mathrm{gap}_p(s,a) - \mathrm{gap}_q(s,a)\right| \leq \left|V_p^\star(s) - V_q^\star(s)\right| + \left|Q_p^\star(s,a) - Q_p^\star(s,a)\right| \leq 4H\epsilon. \qquad \square$$

C.2. Proof of Lemma 7

Lemma 7. *For any $(s,a) \in \mathcal{I}_\epsilon$, we have that: (1) for all $p \in [M]$, $(s,a) \notin Z_{p,\text{opt}}$, where we recall that $Z_{p,\text{opt}} = \left\{(s,a) : \text{gap}_p(s,a) = 0\right\}$ is the set of optimal state-action pairs with respect to p; (2) for all $p, q \in [M]$, $\text{gap}_p(s,a) \geq \frac{1}{2}\text{gap}_q(s,a)$.*

Proof. For any $(s,a) \in \mathcal{I}_\epsilon$, there exists some p_0 such that $\text{gap}_{p_0}(s,a) \geq 96H\epsilon$. Therefore, for every $p \in [M]$,

$$\text{gap}_p(s,a) \geq \text{gap}_{p_0}(s,a),$$

From Lemma 6 we know that $\left|\text{gap}_p(s,a) - \text{gap}_{p_0}(s,a)\right| \leq 4H\epsilon$. Therefore, for all p,

$$\text{gap}_p(s,a) \geq \text{gap}_{p_0}(s,a) - 4H\epsilon \geq 92H\epsilon > 0.$$

This proves the first item.

For the second item, for all $p, q \in [M]$,

$$\frac{\text{gap}_p(s,a)}{\text{gap}_q(s,a)} = \frac{\text{gap}_q(s,a) - 4H\epsilon}{\text{gap}_q(s,a)} \geq 1 - \frac{4H\epsilon}{\text{gap}_q(s,a)} \geq 1 - \frac{4}{92} \geq \frac{1}{2}.$$ □

D. Additional Definitions Used in the Proofs

In this section, we define a few useful notations that will be used in our proofs. For state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, player $p \in [M]$, episode $k \in [K]$:

1. Define $n^k(s,a)$ (resp. $n_p^k(s,a)$, $\hat{\mathbb{P}}^k$, $\hat{\mathbb{P}}_p^k$, $\hat{R}^k$, $\hat{R}_p^k$) to be the value of $n(s,a)$ (resp. $n_p(s,a)$, $\hat{\mathbb{P}}$, $\hat{\mathbb{P}}_p$, $\hat{R}$, $\hat{R}_p$) at the *beginning* of episode k of MULTI-TASK-EULER.

2. Denote by $\overline{Q}_p^k$ (resp. $\underline{Q}_p^k, \overline{V}_p^k, \underline{V}_p^k$, ind-$b_p^k(s,a)$, agg-$b_p^k(s,a)$) the values of $\overline{Q}_p$ (resp. $\underline{Q}_p, \overline{V}_p, \underline{V}_p$, ind-$b_p(s,a)$, agg-$b_p(s,a)$) right after MULTI-TASK-EULER finishes its optimistic value iteration (line 21) at episode k.

3. Define the *surplus* (Simchowitz & Jamieson, 2019) (also known as the Bellman error) of (s,a) at episode k and player p as:
$$E_p^k(s,a) := \overline{Q}_p^k(s,a) - R_p(s,a) - (\mathbb{P}_p\overline{V}_p^k)(s,a).$$

4. Define $w_p^k(s,a) := \frac{n_p^k(s,a)}{n^k(s,a)}$ be the proportion of player p on (s,a) at the beginning of episode k; this induces (s,a)'s *mixture expected reward*:
$$\bar{R}^k(s,a) := \sum_{q=1}^M w_q^k(s,a)R_q(s,a),$$
and *mixture transition probability*:
$$\bar{\mathbb{P}}^k(\cdot \mid s,a) := \sum_{q=1}^M w_q^k(s,a)\mathbb{P}_q(\cdot \mid s,a).$$

5. Define $\rho_p^k(s,a) := \mathbb{P}((s_h,a_h) = (s,a) \mid \pi^k(p), \mathcal{M}_p)$ to be the occupancy measure of $\pi^k(p)$ over $\mathcal{M}_p$ on (s,a), where $h \in [H]$ is the layer that s is in (so that $s \in \mathcal{S}_h$). It can be seen that ρ_p^k, when restricted to $\mathcal{S}_h \times \mathcal{A}$, is a probability distribution on it.
Define $\rho^k(s,a) := \sum_{p=1}^M \rho_p^k(s,a)$. It can be seen that $\bar{\rho}^k(s,a) \in [0,M]$. Define $\bar{n}_p^k(s,a) := \sum_{j=1}^k \rho_p^j(s,a)$, and $\bar{n}^k(s,a) := \sum_{j=1}^k \rho^j(s,a)$.[3]

[3]These are the cumulative occupancy measures up to episode k, inclusively; this is in contrast with the definition of $n^k(s,a)$ and $n_p^k(s,a)$, which do not count the trajectories observed at episode k.

6. Define $N^k(s) := \sum_{a \in \mathcal{A}} n^k(s,a)$ and $N_p^k(s) := \sum_{a \in \mathcal{A}} n_p^k(s,a)$ to be the total number of encounters of state s by all players, and by player p only, respectively, at the beginning of episode k.

7. Define $N_1 := 84M \ln(\frac{SAK}{\delta})$, and $N_2 := 84 \ln(\frac{MSAK}{\delta})$; define $\tau(s,a) := \min\left\{k : \bar{n}^k(s,a) \geq N_1\right\}$, and $\tau_p(s,a) := \min\left\{k : \bar{n}_p^k(s,a) \geq N_2\right\}$; With high probability, so long as $k \geq \tau(s,a)$ (resp. $k \geq \tau_p(s,a)$), $n_p^k(s,a)$ and $\bar{n}_p^k(s,a)$ (resp. $n^k(s,a)$ and $\bar{n}^k(s,a)$) are within a constant factor of each other; see Lemma 12.

8. Define $\check{\text{gap}}_p(s,a) := \frac{\text{gap}_p(s,a)}{4H} \vee \frac{\text{gap}_{p,\min}}{4H}$, where $\text{gap}_p(s,a)$ and $\text{gap}_{p,\min}$ are defined in Section 2.

Define $\text{Reg}(K,p) := \sum_{k=1}^K \left(V_{0,p}^\star - V_{0,p}^{\pi^k(p)}\right)$ as player p's contribution to the collective regret; in this notation, $\text{Reg}(K) = \sum_{p=1}^M \text{Reg}(K,p)$.

Define the clipping function $\text{clip}(\alpha, \Delta) := \alpha \mathbf{1}(\alpha \geq \Delta)$.

We also adopt the following conventions in our proofs:

1. As ϵ-dissimilarity with $\epsilon > 2H$ does not impose any constraints on $\left\{\mathcal{M}_p\right\}_{p=1}^M$, throughout the proof, we only focus on the regime that $\epsilon \leq 2H$.

2. We will use $\pi^k(p)$ and π_p^k interchangeably. To avoid notational clutter, we will also sometimes slightly abuse notation, using $V_{p,h}^{\pi^k}$, $V_p^{\pi^k}$ to denote $V_{p,h}^{\pi^k(p)}$, $V_p^{\pi^k(p)}$ respectively.

E. Proof of the Upper Bounds

This section establishes the regret guarantees of MULTI-TASK-EULER (Theorems 4 and 8). The proof follows a similar outline as STRONG-EULER (Simchowitz & Jamieson, 2019)'s analysis, with important modifications tailored to the multitask setting. The proof has the following roadmap:

1. Subsection E.1 defines a clean event E that we show happens with probability $1-\delta$. When E happens, the observed samples are typical enough so that standard concentration inequalities apply. This will serve as the basis of our subsequent arguments.

2. Subsection E.2 shows that when E happens, the value function upper and lower bounds are valid; furthermore, MULTI-TASK-EULER enjoys strong optimism (Simchowitz & Jamieson, 2019), in that all players' surpluses are always nonnegative for all state-action pairs at all time steps.

3. Subsection E.3 establishes a distribution-dependent upper bound on MULTI-TASK-EULER's surpluses when E happens, which is key to our regret theorems. In comparison with STRONG-EULER (Simchowitz & Jamieson, 2019) in the single task setting, MULTI-TASK-EULER exploits inter-task similarity, so that its surpluses on state-action pair (s,a) for player p are further controlled by a new term that depends on the dissimilarity parameter ϵ, along with $n^k(s,a)$, the total visitation counts of (s,a) *by all players*.

4. Subsection E.4 uses the strong optimism property and the surplus bounds established in the previous two subsections to conclude our final gap-independent and gap-dependent regret guarantees, via the clipping lemma (Simchowitz & Jamieson, 2019) (see also Lemma 21).

5. Finally, Subsection E.5 collects miscellaneous technical lemmas used in the proofs.

E.1. A clean event

Below we will define a "clean" event E in which all concentration bounds used in the analysis hold, which we will show happens with high probability. Specifically, we will define $E = E_{\text{ind}} \cap E_{\text{agg}} \cap E_{\text{sample}}$, where $E_{\text{ind}}, E_{\text{agg}}, E_{\text{sample}}$ are defined respectively below.

In subsequent definitions of events, we will abbreviate $\forall k \in [K], h \in [H], p \in [M], s \in \mathcal{S}_h, a \in \mathcal{A}, s' \in \mathcal{S}_{h+1}$ as $\forall k, h, p, s, a, s'$. Also, recall that $L(n) \eqsim \ln(\frac{MSAn}{\delta})$.

Define event E_{ind} as:

$$E_{\text{ind}} = E_{\text{ind,rw}} \cap E_{\text{ind,val}} \cap E_{\text{ind,prob}} \cap E_{\text{ind,var}}, \tag{3}$$

$$E_{\text{ind,rw}} = \left\{ \forall k, h, p, s, a \,.\, \left| \hat{R}_p^k(s,a) - R_p(s,a) \right| \le \sqrt{\frac{L(n^k(s,a))}{2n^k(s,a)}} \right\} \tag{4}$$

$$E_{\text{ind,val}} = \left\{ \forall k, h, p, s, a \,.\, \left| (\hat{\mathbb{P}}_p^k V_p^\star - \mathbb{P}_p V_p^\star)(s,a) \right| \le 4\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{2HL(n_p^k(s,a))}{n_p^k(s,a)} \right\} \tag{5}$$

$$E_{\text{ind,prob}} = \left\{ \forall k, h, p, s, a, s' \,.\, \left| (\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(s' \mid s,a) \right| \le 4\sqrt{\frac{L(n_p^k(s,a)) \cdot \mathbb{P}_p(s' \mid s,a)}{n_p^k(s,a)}} + \frac{2L(n_p^k(s,a))}{n_p^k(s,a)} \right\} \tag{6}$$

$$\begin{aligned} E_{\text{ind,var}} = \Bigg\{ \forall k, h, p, s, a \,.\, & \left| \frac{1}{n_p^k(s,a)} \sum_{i=1}^{n_p^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right| \\ & \le 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{2H^2 L(n_p^k(s,a))}{n_p^k(s,a)} \Bigg\}, \end{aligned} \tag{7}$$

where in Equation (7), s_i' denotes the next state player p transitions to, for the i-th time it experiences (s,a). E_{ind} captures the concentration behavior of each player's individual model estimates.

Lemma 10. $\mathbb{P}(E_{\text{ind}}) \ge 1 - \frac{\delta}{3}$.

Proof. The proof follows a similar reasoning as the proof of (e.g., Simchowitz & Jamieson, 2019, Proposition F.9) using Freedman's Inequality. We would like to show that each of $E_{\text{ind,rw}}, E_{\text{ind,val}}, E_{\text{ind,prob}}, E_{\text{ind,var}}$ happens with probability $1 - \frac{\delta}{12}$, which would give the lemma statement by a union bound. For brevity, we only show that $\mathbb{P}(E_{\text{ind,var}}) \ge 1 - \frac{\delta}{12}$, and the other probability statements follow from a similar reasoning.

Fix $h \in [H]$, $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, and $p \in [M]$. We will show

$$\begin{aligned} \mathbb{P}\Bigg(\exists k \in [K] \,.\, & \left| \frac{1}{n_p^k(s,a)} \sum_{i=1}^{n_p^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right| \\ & \ge 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{2H^2 L(n_p^k(s,a))}{n_p^k(s,a)} \Bigg) \le \frac{\delta}{12MSA}. \end{aligned} \tag{8}$$

For every $j \in \mathbb{N}_+$, define stopping time k_j as the j-th episode when (s,a) is experienced by player p, if such episode exists; otherwise, k_j is defined as ∞. it suffices to show that

$$\begin{aligned} \mathbb{P}\Bigg(\exists j \in \mathbb{N}_+ \,.\, k_j < \infty \wedge \frac{1}{j} & \left| \sum_{i=1}^{j} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right| \\ & \ge 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(j)}{j}} + \frac{2H^2 L(j)}{j} \Bigg) \le \frac{\delta}{12MSA}. \end{aligned} \tag{9}$$

Define $\mathcal{G}_j$ as the σ-algebra generated by all observations up to time step k_j. We have that $\{\mathcal{G}_j\}_{j=0}^\infty$ is a filtration. It can be seen that the sequence $\left\{ X_j := (V_p^\star(s_j') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \right\}_{j=1}^\infty$

is a martingale difference sequence adapted to $\{\mathcal{G}_j\}_{j=0}^{\infty}$; in addition, for every j, $\left|X_j\right| \leq H^2$, and $\mathbb{E}\left[X_j^2 \mid \mathcal{G}_{j-1}\right] \leq \mathbb{E}\left[(V_p^\star(s_j') - (\mathbb{P}_p V_p^\star)(s,a))^4 \mid \mathcal{G}_{j-1}\right] \leq H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]$. This implies that for any $\lambda \geq 0$, $\left\{Y_j(\lambda) = \exp\left(\lambda \frac{1}{H^2}(\sum_{i=1}^j X_i) - \left((e^\lambda - \lambda - 1)\frac{j}{H^2}\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right)\right)\right\}_{j=0}^{\infty}$ is a nonnegative supermartingale (Freedman, 1975), and by optional sampling theorem, $\mathbb{E}\left[Y_j(\lambda)\mathbf{1}(k_j < \infty)\right] \leq \mathbb{E}\left[Y_0(\lambda)\right] = 1$. As a result, for any fixed thresholds $a, v \geq 0$ (see Freedman, 1975, Theorem 1.6),

$$\mathbb{P}\left(\sum_{i=1}^j X_i \geq a \wedge \sum_{i=1}^j H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] \leq v \wedge k_j < \infty\right) \leq \exp\left(-\frac{a^2}{2v + 2aH^2/3}\right)$$

Now, by the doubling argument of (Bartlett et al., 2008, Lemma 2) (observe that $\sum_{i=1}^j \mathbb{E}\left[X_i^2 \mid \mathcal{G}_{i-1}\right] \in [0, H^4 j]$), we have that for all $j \in \mathbb{N}_+$:

$$\mathbb{P}\left(k_j < \infty \wedge \left|\frac{1}{j}\sum_{i=1}^j (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right| \right.$$
$$\left. \geq 4\sqrt{\frac{H^2 \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] L(j)}{j}} + \frac{2H^2 L(j)}{j}\right) \leq \ln(4j) \cdot \frac{\delta}{48 j^2 MSA}.$$

A union bound over all $j \in \mathbb{N}_+$ yields Equation (9). □

Define event E_{agg} as:

$$E_{\text{agg}} = E_{\text{agg,rw}} \cap E_{\text{agg,val}} \cap E_{\text{agg,prob}} \cap E_{\text{agg,var}}, \tag{10}$$

$$E_{\text{agg,rw}} = \left\{\forall k, h, p, s, a \,.\, \left|\hat{R}^k(s,a) - \bar{R}^k(s,a)\right| \leq \sqrt{\frac{L(n^k(s,a))}{2n^k(s,a)}}\right\} \tag{11}$$

$$E_{\text{agg,val}} = \left\{\forall k, h, p, s, a \,.\, \left|(\mathbb{P}^k V_p^\star - \bar{\mathbb{P}}^k V_p^\star)(s,a)\right|\right. \tag{12}$$

$$\left. \leq 4\sqrt{\frac{\left(\sum_{q=1}^M w_q^k(s,a) \mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{n^k(s,a)}} + \frac{2HL(n^k(s,a))}{n^k(s,a)}\right\} \tag{13}$$

$$E_{\text{agg,prob}} = \left\{\forall k, h, p, s, a, s' \,.\, \left|(\mathbb{P}^k - \bar{\mathbb{P}}^k)(s' \mid s,a)\right| \leq 4\sqrt{\frac{\bar{\mathbb{P}}^k(s' \mid s,a) \cdot L(n^k(s,a))}{n^k(s,a)}} + \frac{2L(n^k(s,a))}{n^k(s,a)}\right\} \tag{14}$$

$$E_{\text{agg,var}} = \left\{\forall k, h, p, s, a \,.\, \left|\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_{p_i} V_q^\star)(s,a))^2 - \sum_{q=1}^M w_q^k(s,a) \mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right|\right. \tag{15}$$

$$\left. \leq 4\sqrt{\frac{H^2\left(\sum_{q=1}^M w_q^k(s,a) \mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{n^k(s,a)}} + \frac{2H^2 L(n^k(s,a))}{n^k(s,a)}\right\},$$

where in Equation (15), s_i' denotes the next state for the i-th time some player experiences (s,a). E_{agg} captures the concentration behavior of the aggregate model estimates.

Lemma 11. $\mathbb{P}(E_{\text{agg}}) \geq 1 - \frac{\delta}{3}$.

Proof. The proof follows a similar reasoning as the proof of (e.g., Simchowitz & Jamieson, 2019, Proposition F.9) using Freedman's Inequality. We would like to show that each of $E_{\text{agg,rw}}, E_{\text{agg,val}}, E_{\text{agg,prob}}, E_{\text{agg,var}}$ happen with probability $1-\frac{\delta}{12}$, which would give the lemma statement by a union bound. For brevity, we show that $\mathbb{P}(E_{\text{agg,var}}) \geq 1-\frac{\delta}{12}$, and the other probability statements follow from a similar reasoning.

Fix $h \in [H]$, $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ and $p \in [M]$; denote by p_i the identity of the player when (s,a) is experienced for the i-th time for some player. It suffices to show that

$$\mathbb{P}\left(\exists k \in [K]. \left|\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} \left((V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star]\right)\right| \geq 4\sqrt{\frac{H^2\left(\sum_{i=1}^{n^k(s,a)} \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{(n^k(s,a))^2}} + \frac{2H^2 L(n^k(s,a))}{n^k(s,a)}\right) \leq \frac{\delta}{12MSA}, \tag{16}$$

because $\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] = \sum_{q=1}^M w_q^k(s,a) \mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]$.

For every $j \in \mathbb{N}_+$, define stopping time k_j as follows: it is the index of the j-th micro-episode when (s,a) is experienced by some player, if such micro-episode exists; and k_j is defined to be ∞ otherwise. With this notation, it suffices to show:

$$\mathbb{P}\left(\exists j \in \mathbb{N}_+ . k_j < \infty \wedge \left|\frac{1}{j} \sum_{i=1}^{j} \left((V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star]\right)\right| \geq 4\sqrt{\frac{H^2\left(\sum_{i=1}^{j} \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star]\right) L(j)}{j^2}} + \frac{2H^2 L(j)}{j}\right) \leq \frac{\delta}{12MSA}, \tag{17}$$

Define $\mathcal{G}_j$ as the σ-algebra generated by all observations up to micro-episode k_j. We have that $\{\mathcal{G}_j\}_{j=0}^\infty$ is a filtration. It can be seen that $\left\{X_j := (V_p^\star(s_j') - (\mathbb{P}_{p_j} V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_{p_j}(\cdot|s,a)}[V_p^\star]\right\}_{j=1}^\infty$ is a martingale difference sequence adapted to $\{\mathcal{G}_j\}_{j=0}^\infty$; in addition, for every j, $|X_j| \leq H^2$, and $\mathbb{E}\left[X_j^2 \mid \mathcal{G}_{j-1}\right] \leq \mathbb{E}\left[(V_p^\star(s_j') - (\mathbb{P}_{p_j} V_p^\star)(s,a))^4 \mid \mathcal{G}_{j-1}\right] \leq H^2 \mathrm{var}_{\mathbb{P}_{p_j}(\cdot|s,a)}[V_p^\star]$. Using the same reasoning as in the proof of Lemma 10 (and observing that $\sum_{i=1}^j \mathbb{E}\left[X_i^2 \mid \mathcal{G}_{i-1}\right] \in [0, H^4 j]$), we have that for all $j \in \mathbb{N}_+$:

$$\mathbb{P}\left(k_j < \infty \wedge \left|\frac{1}{j} \sum_{i=1}^{j} \left((V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right)\right| \geq 4\sqrt{\frac{H\sum_{i=1}^{j} \mathrm{var}_{\mathbb{P}_{p_i}(\cdot|s,a)}[V_p^\star] L(j)}{j^2}} + \frac{2H^2 L(j)}{j}\right) \leq \ln(4j) \cdot \frac{\delta}{48 j^2 MSA}.$$

A union bound over all $j \in \mathbb{N}_+$ implies that Equation (17) holds. $\square$

Define

$$
\begin{aligned}
E_{\text{sample}} &= E_{\text{ind,sample}} \cap E_{\text{agg,sample}}, \\
E_{\text{agg,sample}} &= \left\{ \forall s, a, k \centerdot \bar{n}^k(s,a) \geq N_1 \implies n^k(s,a) \geq \frac{1}{2}\bar{n}^k(s,a) \right\}, \\
E_{\text{ind,sample}} &= \left\{ \forall s, a, k, p \centerdot \bar{n}_p^k(s,a) \geq N_2 \implies n_p^k(s,a) \geq \frac{1}{2}\bar{n}_p^k(s,a) \right\},
\end{aligned}
$$

where we recall from Section D that $N_1 \approx M \ln(\frac{SAK}{\delta})$, and $N_2 \approx \ln(\frac{MSAK}{\delta})$.

Lemma 12. $\mathbb{P}(E_{\text{sample}}) \geq 1 - \frac{\delta}{3}$.

Proof. We first show $\mathbb{P}(E_{\text{agg,sample}}) \geq 1 - \frac{\delta}{6}$. Specifically, fix $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, define random variable $X_k = \sum_{p=1}^M \left(\mathbf{1}\left((s_{h,p}^k, a_{h,p}^k \right) = (s,a)) - \rho_p^k(s,a) \right)$. Also, define $\mathcal{G}_k$ as the σ-algebra generated by all observations up to episode k. It can be readily seen that $\{X_k\}_{k=1}^K$ is a martingale difference sequence adapted to filtration $\{\mathcal{G}_k\}_{k=0}^K$. Freedman's inequality (specifically, Lemma 2 of (Bartlett et al., 2008)) implies that for every fixed k, with probability $1 - \frac{\delta}{6K}$,

$$
\left| n^k(s,a) - \bar{n}^{k-1}(s,a) \right| \leq 4\sqrt{\bar{n}^{k-1}(s,a) \cdot M \ln\left(\frac{6SAK^2}{\delta}\right)} + 4M \ln\left(\frac{6SAK^2}{\delta}\right), \tag{18}
$$

If Equation (18) happens, then by AM-GM inequality that $\sqrt{\bar{n}^{k-1}(s,a) \cdot M \ln\left(\frac{6SAK^2}{\delta}\right)} \leq \frac{1}{4}\bar{n}^{k-1}(s,a) + 16M \ln\left(\frac{6SAK^2}{\delta}\right)$, we have

$$
\bar{n}^{k-1}(s,a) - n^k(s,a) \leq \frac{1}{4}\bar{n}^{k-1}(s,a) + 20M \ln\left(\frac{6SAK^2}{\delta}\right),
$$

implying that

$$
n^k(s,a) \geq \frac{3}{4}\bar{n}^{k-1}(s,a) - 20M \ln\left(\frac{6SAK^2}{\delta}\right).
$$

Additionally, as $\bar{n}^{k-1}(s,a) \geq \bar{n}^k(s,a) - M$ always holds, we have

$$
n^k(s,a) \geq \frac{3}{4}\bar{n}^k(s,a) - 21M \ln\left(\frac{6SAK^2}{\delta}\right).
$$

In summary, for any fixed k, with probability $1 - \frac{\delta}{6K}$, if $\bar{n}^k(s,a) \geq N_1 := 84M \ln\left(\frac{6SAK^2}{\delta}\right)$,

$$
n^k(s,a) \geq \frac{1}{2}\bar{n}^k(s,a).
$$

Taking a union bound over all $k \in [K]$, we have $\mathbb{P}(E_{\text{agg,sample}}) \geq 1 - \frac{\delta}{6}$.

It follows similarly that $\mathbb{P}(E_{\text{ind,sample}}) \geq 1 - \frac{\delta}{6}$; the only difference in the proof is that, we need to take an extra union bound over all $p \in [M]$ - hence an additional factor M within $\ln(\cdot)$ in the definition of N_2. The lemma statement follows from a union bound over these two statements. $\square$

Lemma 13. $\mathbb{P}(E) \geq 1 - \delta$.

Proof. Follows from Lemmas 10, 11, and 12, along with a union bound. $\square$

E.2. Validity of value function bounds

In this section, we show that if the clean event E happens, then for all k and p, the value function estimates $\overline{Q}_p^k$, $\underline{Q}_p^k$, $\overline{V}_p^k$, $\underline{V}_p^k$ are valid upper and lower bounds of the optimal value functions $Q_p^\star$, $V_p^\star$ (Lemma 16). As a by-product, we also give a general bound on the surplus (Lemma 15) which will be refined and used in the subsequent regret bound calculations. Before going into the proof of the above two lemmas, we need a technical lemma below (Lemma 14) that gives necessary concentration results which motivate the bonus constructions; its proof can be found at Section E.2.1.

Lemma 14. *Fix $p \in [M]$. Suppose E happens, and suppose that for episode k and step h, we have that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \le V^\star(s') \le \overline{V}_p^k(s')$. Then, for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$:*

1.

$$\left|\hat{R}_p^k(s,a) - R_p(s,a)\right| \le b_{\mathrm{rw}}\left(n_p^k(s,a), 0\right), \tag{19}$$

$$\left|\hat{R}^k(s,a) - R_p(s,a)\right| \le b_{\mathrm{rw}}\left(n^k(s,a), \epsilon\right). \tag{20}$$

2.

$$\left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(V_p^\star)(s,a)\right| \le b_{\mathrm{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right), \tag{21}$$

$$\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_p^\star)(s,a)\right| \le b_{\mathrm{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right). \tag{22}$$

3. For any $V_1, V_2 : \mathcal{S}_{h+1} \to \mathbb{R}$ such that $\overline{V}_p^k \le V_1 \le V_2 \le \underline{V}_p^k$,

$$\left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \le b_{\mathrm{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right), \tag{23}$$

$$\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \le b_{\mathrm{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right). \tag{24}$$

Lemma 15. *If event E happens, and suppose that for episode k and step h, we have that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \le V_p^\star(s') \le \overline{V}_p^k(s')$. Then, for $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,*

$$\overline{Q}_p^k(s,a) - \left(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)\right) \in \left[0, (H-h+1) \wedge 2\textit{ind-}b_p^k(s,a) \wedge 2\textit{agg-}b_p^k(s,a)\right], \tag{25}$$

and

$$\left(R_p(s,a) + (\mathbb{P}_p \underline{V}_p^k)(s,a)\right) - \underline{Q}_p^k(s,a) \in \left[0, (H-h+1) \wedge 2\textit{ind-}b_p^k(s,a) \wedge 2\textit{agg-}b_p^k(s,a)\right], \tag{26}$$

where we recall that

$$\textit{ind-}b_p^k(s,a) = b_{\mathrm{rw}}\left(n_p^k(s,a), 0\right) + b_{\mathrm{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right) + b_{\mathrm{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right),$$

$$\textit{agg-}b_p^k(s,a) = b_{\mathrm{rw}}\left(n^k(s,a), \epsilon\right) + b_{\mathrm{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right) + b_{\mathrm{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right).$$

Proof. We only show Equation (25) for brevity; Equation (26) follows from an exact symmetrical reasoning.

Recall that $\overline{Q}_p^k(s,a) = \min\left(\overline{\text{ind-}Q}_p^k(s,a), \overline{\text{agg-}Q}_p^k(s,a), H\right)$. We compare each term in the $\min(\cdot)$ operator with $(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a))$:

- For $\overline{\text{ind-}Q}_p^k(s,a)$, using Lemma 14 and our assumption on $\overline{V}_p^k$ and $\underline{V}_p^k$ on $\mathcal{S}_{h+1}$, we have:

$$\begin{aligned}
&\overline{\text{ind-}Q}_p^k(s,a) - \left(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)\right) \\
=& (\hat{R}_p^k - R_p)(s,a) + b_{\text{rw}}\left(n_p^k(s,a), 0\right) \\
&+ ((\hat{\mathbb{P}}_p^k - \mathbb{P}_p)V_p^\star)(s,a) + b_{\text{prob}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right) \\
&+ (\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(\overline{V}_p^k - V_p^\star))(s,a) + b_{\text{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n_p^k(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right) \\
\in& [0, 2\text{ind-}b_p^k(s,a)].
\end{aligned}$$

- For $\overline{\text{agg-}Q}_p^k(s,a)$, using Lemma 14 and our assumptions on $\overline{V}_p^k$ and $\underline{V}_p^k$ over $\mathcal{S}_{h+1}$, we have:

$$\begin{aligned}
&\overline{\text{agg-}Q}_p^k(s,a) - \left(R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)\right) \\
=& (\hat{R}_p^k - R_p)(s,a) + b_{\text{rw}}\left(n^k(s,a), \epsilon\right) \\
&+ ((\hat{\mathbb{P}}^k - \mathbb{P}_p)V_p^\star)(s,a) + b_{\text{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right) \\
&+ ((\hat{\mathbb{P}}^k - \mathbb{P}_p)(\overline{V}_p^k - V_p^\star))(s,a) + b_{\text{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right) \\
\in& [0, 2\text{agg-}b_p^k(s,a)],
\end{aligned}$$

- For $H-h+1$, we have:

$$(H-h+1) - (R_p(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)) \in [0, H-h+1],$$

 where we use the observation that $R(s,a) \in [0,1]$, and $(\mathbb{P}_p \overline{V}_p^k)(s,a) \in [0, H-h]$, and their sum is in $[0,H]$.

Combining the above three establishes that

$$\overline{Q}_p^k(s,a) - (R(s,a) + (\mathbb{P}_p \overline{V}_p^k)(s,a)) \in \left[0, (H-h+1) \wedge 2\text{ind-}b_p^k(s,a) \wedge 2\text{agg-}b_p^k(s,a)\right]. \qquad \square$$

Lemma 16. *Under event E, for every $k \in [K]$, and every $p \in [M]$, and for every $h \in [H]$, For all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,*

$$\underline{Q}_p^k(s,a) \le Q_p^{\pi^k}(s,a) \le Q_p^\star(s,a) \le \overline{Q}_p^k(s,a), \tag{27}$$

and

$$\underline{V}_p^k(s) \le V_p^{\pi^k}(s) \le V_p^\star(s) \le \overline{V}_p^k(s), \tag{28}$$

Proof. The proof of this lemma extends (Simchowitz & Jamieson, 2019), Proposition F.1 to our multitask setting.

For every k and p, we show the above holds for all layers $h \in [H]$ and every $(s,a) \in \mathcal{S}_h \times \mathcal{A}$; to this end, we do backward induction on layer h.

Base case: For layer $h = H+1$, we have $\underline{V}_p^k(\perp) = V_p^{\pi^k}(\perp) = V_p^\star(\perp) = \overline{V}_p^k(\perp) = 0$.

Inductive case: By our inductive hypothesis, for layer $h+1$ and every $s \in \mathcal{S}_{h+1}$,

$$\underline{V}_p^k(s) \le V_p^{\pi^k}(s) \le V_p^\star(s) \le \overline{V}_p^k(s).$$

We will show that Equations (27) and (28) holds holds for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$.

We first show Equation (27). First, $Q_p^{\pi^k}(s,a) \le Q_p^\star(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ is trivial.

To show $Q^\star_p(s,a) \le \overline{Q}^k_p(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, by Lemma 15 and inductive hypothesis, we have:

$$Q^\star_p(s,a) = R_p(s,a) + (\mathbb{P}_p V^\star_p)(s,a) \le R_p(s,a) + (\mathbb{P}_p \overline{V}^k_p)(s,a) \le \overline{Q}^k_p(s,a).$$

Likewise, we show $Q^{\pi^k}_p(s,a) \ge \underline{Q}^k_p(s,a)$ for all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, using Lemma 15 and inductive hypothesis:

$$Q^{\pi^k}_p(s,a) = R_p(s,a) + (\mathbb{P}_p V^{\pi^k}_p)(s,a) \ge R_p(s,a) + (\mathbb{P}_p \overline{V}^k_p)(s,a) \ge \underline{Q}^k_p(s,a).$$

This completes the proof of Equation (27) for layer h.

We now show Equation (28) for layer h. Again $V^{\pi^k}_p(s) \le V^\star_p(s)$ for all $s \in \mathcal{S}_h$ is trivial.

To show $V^\star_p(s) \le \overline{V}^k_p(s)$ for all $s \in \mathcal{S}_h$, observe that

$$V^\star_p(s) = \max_{a \in \mathcal{A}} Q^\star_p(s,a) \le \max_{a \in \mathcal{A}} \overline{Q}^k_p(s,a) = \overline{V}^k_p(s).$$

To show $V^{\pi^k}_p(s) \ge \underline{V}^k_p(s)$ for all $s \in \mathcal{S}_h$, observe that

$$V^{\pi^k}_p(s) = Q^{\pi^k}_p(s, \pi^k(p)(s)) \ge \underline{Q}^k_p(s, \pi^k(p)(s)) = \underline{V}^k_p(s).$$

This completes the induction. $\square$

E.2.1. Proof of Lemma 14

Proof of Lemma 14. Equations (19), (21), and (23) essentially follow the same reasoning as in (Simchowitz & Jamieson, 2019); we still include their proofs for completeness. Equations (20), (22), and (24) are new, and require a more involved analysis. Our proof also relies on a technical lemma, namely Lemma 17; we defer its statement and proof to the end of this subsection.

1. Equation (19) follows directly from the definition of $E_{\text{ind,rw}}$. Equation (20) follows from the definition of $E_{\text{agg,rw}}$, and the fact that $\left|\bar{R}^k(s,a) - R_p(s,a)\right| \le \epsilon$.

2. We have:

$$\begin{aligned}
&\left|(\hat{\mathbb{P}}^k_p V^\star - \mathbb{P}_p V^\star_p)(s,a)\right| \\
\le& O\left(\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V^\star] L(n^k_p(s,a))}{n^k_p(s,a)}} + \frac{HL(n^k_p(s,a))}{n^k_p(s,a)}\right) \\
\le& O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k_p(\cdot|s,a)}[V^\star] L(n^k_p(s,a))}{n^k_p(s,a)}} + \frac{HL(n^k_p(s,a))}{n^k_p(s,a)}\right) \\
\le& O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k_p(\cdot|s,a)}[\overline{V}^k_p]\, L(n^k_p(s,a))}{n^k_p(s,a)}} + \sqrt{\frac{\|V^\star_p - \overline{V}^k_p\|^2_{\hat{\mathbb{P}}^k_p(\cdot|s,a)}\, L(n^k_p(s,a))}{n^k_p(s,a)}} + \frac{HL(n^k_p(s,a))}{n^k_p(s,a)}\right) \\
\le& O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k_p(\cdot|s,a)}[\overline{V}^k_p]\, L(n^k_p(s,a))}{n^k_p(s,a)}} + \sqrt{\frac{\|\overline{V}^k_p - \underline{V}^k_p\|^2_{\hat{\mathbb{P}}^k_p(\cdot|s,a)}\, L(n^k_p(s,a))}{n^k_p(s,a)}} + \frac{HL(n^k_p(s,a))}{n^k_p(s,a)}\right) \\
\le& b_{\text{prob}}\left(\hat{\mathbb{P}}^k_p(\cdot \mid s,a), n^k_p(s,a), \overline{V}^k_p, \underline{V}^k_p, 0\right),
\end{aligned}$$

where the first inequality is from the definition of $E_{\text{ind,val}}$; the second inequality is from Equation (29) of Lemma 17; the third inequality is from Lemma 24; the fourth inequality is from our assumption that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}^k_p(s') \le$

$V^\star(s') \le \overline{V}_p^k(s')$, and thus $\left|(V_p^\star - \underline{V}_p^k)(s')\right| \le \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all s' in the support of $\hat{\mathbb{P}}_p^k(\cdot \mid s,a)$. This proves Equation (21).

We prove Equation (22) as follows:

$$\begin{aligned}
&\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_p^\star)(s,a)\right| \\
\le& \epsilon + \left|(\hat{\mathbb{P}}^k - \bar{\mathbb{P}}_p^k)(V_p^\star)(s,a)\right| \\
\le& \epsilon + O\left(\sqrt{\frac{\left(\sum_{q=1}^M w_q^k(s,a)\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]\right) L(n^k(s,a))}{n^k(s,a)}} + \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\le& \epsilon + O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[V_p^\star]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}\cdot \epsilon H} + \frac{HL(n^k(s,a))}{n^k(s,a)} + \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\le& 2\epsilon + O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[\overline{V}_p^k]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - V_p^\star\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\le& 2\epsilon + O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[\overline{V}_p^k]\ L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\ L(n^k(s,a))}{n^k(s,a)}} + \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\le& b_{\mathrm{prob}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n^k(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right),
\end{aligned}$$

where the first inequality is from the observation that $\|\bar{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_p(\cdot \mid s,a)\|_1 \le \frac{\epsilon}{H}$ and Lemma 25; the second inequality is from the definition of $E_{\mathrm{agg,val}}$; the third inequality is from Equation (30) of Lemma 17; the fourth inequality is from Lemma 24; the fifth inequality is from our assumption that for all $s' \in \mathcal{S}_{h+1}$, $\underline{V}_p^k(s') \le V^\star(s') \le \overline{V}_p^k(s')$, and thus $\left|(V_p^\star - \underline{V}_p^k)(s')\right| \le \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all s' in the support of $\hat{\mathbb{P}}^k(\cdot \mid s,a)$. This proves Equation (21).

3. We prove Equation (23) as follows:

$$\begin{aligned}
&\left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \\
\le& \sum_{s' \in \mathcal{S}_{h+1}} \left|(\hat{\mathbb{P}}_p^k - \mathbb{P}_p)(s' \mid s,a)\right| \cdot (V_2 - V_1)(s') \\
\le& O\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n_p^k(s,a)) \cdot \mathbb{P}_p^k(s' \mid s,a)}{n_p^k(s,a)}} + \frac{L(n_p^k(s,a))}{n_p^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\le& O\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n_p^k(s,a)) \cdot \hat{\mathbb{P}}_p^k(s' \mid s,a)}{n_p^k(s,a)}} + \frac{L(n_p^k(s,a))}{n_p^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\le& O\left(\sum_{s' \in \mathcal{S}_{h+1}} \sqrt{\hat{\mathbb{P}}_p^k(s' \mid s,a)}(\overline{V}_p^k - \underline{V}_p^k)(s') \cdot \sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}} + \sum_{s' \in \mathcal{S}_{h+1}} \frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\le& O\left(\sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\ L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right) \\
\le& b_{\mathrm{str}}\left(\hat{\mathbb{P}}_p^k(\cdot \mid s,a), n(s,a), \overline{V}_p^k, \underline{V}_p^k, 0\right),
\end{aligned}$$

where the first inequality is from the elementary fact that $\left|\sum_{i=1}^n a_i\right| \leq \sum_{i=1}^n |a_i|$; the second inequality is from the definition of $E_{\text{ind,prob}}$; the third inequality is from the definition of $E_{\text{ind,prob}}$ and Lemma 26; the fourth inequality is by algebra and $0 \leq (V_2 - V_1)(s') \leq \min(H, (\overline{V}_p^k - \underline{V}_p^k)(s'))$ for all $s' \in \mathcal{S}_{h+1}$; the last inequality is by Cauchy-Schwarz;

We now deduce Equation (24):

$$
\begin{aligned}
&\left|(\hat{\mathbb{P}}^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| \\
\leq& \left|(\bar{\mathbb{P}}^k - \mathbb{P}_p)(V_2 - V_1)(s,a)\right| + \left|(\hat{\mathbb{P}}^k - \bar{\mathbb{P}}^k)(V_2 - V_1)(s,a)\right| \\
\leq& \epsilon + \sum_{s' \in \mathcal{S}_{h+1}} \left|(\hat{\mathbb{P}}^k - \bar{\mathbb{P}}^k)(s' \mid s,a)\right| \cdot (V_2 - V_1)(s') \\
\leq& \epsilon + O\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n^k(s,a)) \cdot \bar{\mathbb{P}}^k(s' \mid s,a)}{n^k(s,a)}} + \frac{L(n^k(s,a))}{n^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\leq& \epsilon + O\left(\sum_{s' \in \mathcal{S}_{h+1}} \left(\sqrt{\frac{L(n^k(s,a)) \cdot \hat{\mathbb{P}}^k(s' \mid s,a)}{n^k(s,a)}} + \frac{L(n^k(s,a))}{n^k(s,a)}\right) \cdot (V_2 - V_1)(s')\right) \\
\leq& \epsilon + O\left(\sum_{s' \in \mathcal{S}_{h+1}} \sqrt{\hat{\mathbb{P}}^k(s' \mid s,a)}(\overline{V}_p^k - \underline{V}_p^k)(s') \cdot \sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}} + \sum_{s' \in \mathcal{S}_{h+1}} \frac{HL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq& \epsilon + O\left(\sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\, L(n^k(s,a))}{n^k(s,a)}} + \frac{SHL(n^k(s,a))}{n^k(s,a)}\right) \\
\leq& b_{\text{str}}\left(\hat{\mathbb{P}}^k(\cdot \mid s,a), n(s,a), \overline{V}_p^k, \underline{V}_p^k, \epsilon\right),
\end{aligned}
$$

where the first inequality is triangle inequality; the second inequality is from the elementary fact that $\left|\sum_{i=1}^n a_i\right| \leq \sum_{i=1}^n |a_i|$, along with $\|\bar{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_p(\cdot \mid s,a)\|_1 \leq \frac{\epsilon}{H}$ and Lemma 25; the third inequality is from the definition of $E_{\text{ind,prob}}$; the fourth inequality is from the definition of $E_{\text{ind,prob}}$ and Lemma 26; the fifth inequality is by algebra and $0 \leq (V_2 - V_1)(s') \leq \min(H, (\overline{V}_p^k - \underline{V}_p^k)(s'))$ for all $s' \in \mathcal{S}_{h+1}$; the last inequality is by Cauchy-Schwarz. □

Lemma 14 relies on the following technical lemma on the concentrations of the conditional variances. Specifically, Equation (29) is well-known (see, e.g., Audibert et al., 2007; Maurer & Pontil, 2009); Equations (30) and (31) are new, and allows for heterogeneous data aggregation in the multi-task RL setting. We still include the proof of Equation (29) here, as it helps illustrate our ideas for proving the two new inequalities.

Lemma 17. *If event E happens, then for any s, a, k, p, we have:*

1.
$$
\left|\sqrt{\operatorname{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\operatorname{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right| \lesssim H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}, \tag{29}
$$

2.
$$
\left|\sqrt{\operatorname{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\sum_{q=1}^M w_q^k(s,a)\operatorname{var}_{\mathbb{P}_q(\cdot|s,a)}\left[V_p^\star\right]}\right| \lesssim \sqrt{H\epsilon} + 2H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}, \tag{30}
$$

and
$$
\left|\sqrt{\operatorname{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\operatorname{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right| \lesssim \sqrt{H\epsilon} + H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}, \tag{31}
$$

Proof. 1. By the definition of E, we have

$$\left|\frac{1}{n_p^k(s,a)}\sum_{i=1}^{n_p^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2-\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right|\lesssim\sqrt{\frac{H^2\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{H^2L(n_p^k(s,a))}{n_p^k(s,a)};$$

this, when combined with Lemma 26, implies that

$$\left|\sqrt{\frac{1}{n_p^k(s,a)}\sum_{i=1}^{n_p^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]}\right|\le H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}. \tag{32}$$

Now, observe that

$$\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]=\frac{1}{n_p^k(s,a)}\sum_{i=1}^{n_p^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2-((\hat{\mathbb{P}}_p^kV_p^\star)(s,a)-(\mathbb{P}_pV_p^\star)(s,a))^2.$$

Recall that by the definition of event E, we have

$$\left|(\hat{\mathbb{P}}_p^kV_p^\star)(s,a)-(\mathbb{P}_pV_p^\star)(s,a)\right|\le H\wedge\left(\sqrt{\frac{H^2L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{HL(n_p^k(s,a))}{n_p^k(s,a)}\right)\le 2H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}},$$

where the second inequality uses Lemma 27. Using the elementary fact that $|A-B|\le C\Rightarrow\sqrt{A}\le\sqrt{B}+\sqrt{C}$, we get that

$$\begin{aligned}&\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_pV_p^\star)(s,a))^2}\right|\\&\le\left|(\hat{\mathbb{P}}_p^kV_p^\star)(s,a)-(\mathbb{P}_pV_p^\star)(s,a)\right|\lesssim H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}.\end{aligned} \tag{33}$$

Combining Equations (32) and (33), using algebra, we get

$$\left|\sqrt{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right]}-\sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]}\right|\lesssim H\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}},$$

establishing Equation (29).

2. We first show Equation (30). By the definition of E, we have

$$\begin{aligned}&\left|\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_{p_i}V_p^\star)(s,a))^2-\sum_{p=1}^{M}w_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right|\\&\lesssim\sqrt{\frac{H^2\left(\sum_{p=1}^{M}w_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]\right)L(n^k(s,a))}{n^k(s,a)}}+\frac{H^2L(n^k(s,a))}{n^k(s,a)},\end{aligned}$$

this, combined with Lemma 26, implies that

$$\left|\sqrt{\frac{1}{n^k(s,a)}\sum_{i=1}^{n^k(s,a)}(V_p^\star(s_i')-(\mathbb{P}_{p_i}V_p^\star)(s,a))^2}-\sqrt{\sum_{p=1}^{M}w_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star]}\right|\lesssim H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}. \tag{34}$$

For the first term on the left hand side, observe that for each i, $|(\mathbb{P}_{p_i} V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)| \leq H \frac{\epsilon}{H} = \epsilon$, we therefore have $\left|(V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2 - (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2\right| \leq 2H\epsilon$ by $2H$-Lipschitzness of function $f(x) = x^2$ on $[-H, H]$. By averaging over all i's and taking square root, we have

$$\left| \sqrt{\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_{p_i} V_p^\star)(s,a))^2} - \sqrt{\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2} \right| \lesssim \sqrt{H\epsilon}. \tag{35}$$

Furthermore,

$$\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right] = \frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2 - ((\hat{\mathbb{P}}^k V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a))^2,$$

and

$$\left|(\hat{\mathbb{P}}^k V_p^\star)(s,a) - (\mathbb{P}_p V_p^\star)(s,a)\right| \lesssim \epsilon + H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}$$

Together with our assumption that $\epsilon \leq 2H$ (which implies that $\epsilon \lesssim \sqrt{H\epsilon}$), this gives

$$\left| \sqrt{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\frac{1}{n^k(s,a)} \sum_{i=1}^{n^k(s,a)} (V_p^\star(s_i') - (\mathbb{P}_p V_p^\star)(s,a))^2} \right| \lesssim \sqrt{H\epsilon} + H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}. \tag{36}$$

Equation (30) is a direct consequence of Equations (34), (35) and (36) along with algebra.

We now show Equation (31) using Equation (30). By Lemma 25, for every q, $\left|\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}\left[V_p^\star\right] - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]\right| \leq 3H^2 \cdot \frac{\epsilon}{H} = 3H\epsilon$. Therefore, $\left|\sum_{q=1}^M w_q^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^\star] - \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]\right| \leq 3H^2 \cdot \frac{\epsilon}{H} = 3H\epsilon$, and

$$\left| \sqrt{\sum_{q=1}^M w_q^k(s,a)\mathrm{var}_{\mathbb{P}_q(\cdot|s,a)}[V_p^\star]} - \sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]} \right| \lesssim \sqrt{H\epsilon}$$

This, together with Equation (30), implies

$$\left| \sqrt{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right]} - \sqrt{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right]} \right| \lesssim \sqrt{H\epsilon} + H\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}},$$

establishing Equation (31). □

E.3. Simplifying the surplus bounds

In this section, we show a distribution-dependent bound on the surplus terms, namely Lemma 20, which is key to establishing our regret bound. It can be seen as an extension of Proposition B.4 of (Simchowitz & Jamieson, 2019) to our multitask setting using the MULTI-TASK-EULER algorithm, under the ϵ-dissimilarity assumption. Before we present Lemma 20 (Section E.3.1), we first show and prove two auxiliary lemmas, Lemma 18 and Lemma 19.

Lemma 18 (Bounds on $\overline{V}_p^k - \underline{V}_p^k$, generalization of (Simchowitz & Jamieson, 2019), Lemma F.8)**.** *If E happens, then for all $p \in [M]$, $k \in [K]$, $h \in [H+1]$ and $s \in \mathcal{S}_h$,*

$$(\overline{V}_p^k - \underline{V}_p^k)(s) \leq 4\mathbb{E}\left[\sum_{t=h}^H \left(H \wedge \textit{ind-}b_p^k(s_t,a_t) \wedge \textit{agg-}b_p^k(s_t,a_t)\right) \mid s_h = s, \pi^k(p), \mathcal{M}_p\right]; \tag{37}$$

consequently,

$$(\overline{V}_p^k - \underline{V}_p^k)(s) \lesssim H \sum_{t=h}^{H} \mathbb{E}\left[\left(1 \wedge \sqrt{\frac{SL(n_p^k(s_t, a_t))}{n_p^k(s_t, a_t)}}\right) \mid s_h = s, \pi^k(p), \mathcal{M}_p\right]. \tag{38}$$

Proof. First, Lemmas 16 and 15 together imply that if E holds, Equations (25) and (26) holds for all p, k, s, a. Under this premise, we show Equation (37) by backward induction.

Base case: for $h = H + 1$, we have that LHS is $(\overline{V}_p^k - \underline{V}_p^k)(\perp) = 0$ which is equal to the RHS.

inductive case: Suppose Equation (37) holds for all $s \in \mathcal{S}_{h+1}$. Now consider $s \in \mathcal{S}_h$. By the definitions of $\overline{V}_p^k$ and $\underline{V}_p^k$,

$$\begin{aligned}
&(\overline{V}_p^k - \underline{V}_p^k)(s) \\
=&\overline{Q}_p^k(s, \pi_p^k(s)) - \underline{Q}_p^k(s, \pi_p^k(s)) \\
\leq&(\mathbb{P}_p(\overline{V}_p^k - \underline{V}_p^k))(s, \pi_p^k(s)) + 4(H \wedge \text{ind-}b_p^k(s, \pi_p^k(s)) \wedge \text{agg-}b_p^k(s, \pi_p^k(s))) \\
=&\mathbb{E}\left[4 \min(H, \text{ind-}b_p^k(s, a), \text{agg-}b_p^k(s, a)) + (\overline{V}_p^k - \underline{V}_p^k)(s_{h+1}) \mid s_h = s, \pi_p^k, \mathcal{M}_p\right] \\
\leq&\mathbb{E}\left[4(H \wedge \text{ind-}b_p^k(s, a) \wedge \text{agg-}b_p^k(s, a)) + \mathbb{E}\left[\sum_{t=h+1}^{H}\left(H \wedge 2\text{ind-}b_p^k(s_t, a_t) \wedge 2\text{agg-}b_p^k(s_t, a_t)\right) \mid s_{h+1}\right] \mid s_h = s, \pi_p^k, \mathcal{M}_p\right] \\
\leq&4\mathbb{E}\left[\sum_{t=h}^{H}\left(H \wedge \text{ind-}b_p^k(s_t, a_t) \wedge \text{agg-}b_p^k(s_t, a_t)\right) \mid s_h = s, \pi_p^k, \mathcal{M}_p\right],
\end{aligned}$$

where the first inequality is from Equations (25) and (26) for (s, a) and player p at episode k, and the second inequality is from the inductive hypothesis; the third inequality is by algebra. This completes the induction.

We now show Equation (38). By the definition of $\text{ind-}b_p^k(s, a)$ and algebra,

$$\begin{aligned}
&\text{ind-}b_p^k(s, a) \\
\lesssim&\sqrt{\frac{\text{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right] \; L(n_p^k(s, a))}{n_p^k(s, a)}} + \sqrt{\frac{L(n_p^k(s, a))}{n_p^k(s, a)}} + \sqrt{\frac{S\|\overline{V}_p^k - \underline{V}_p^k\|_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}^2 \; L(n_p^k(s, a))}{n_p^k(s, a)}} + \frac{HSL(n_p^k(s, a))}{n_p^k(s, a)} \\
\lesssim&H\sqrt{\frac{SL(n_p^k(s_t, a_t))}{n_p^k(s_t, a_t)}} + \frac{HSL(n_p^k(s_t, a_t))}{n_p^k(s_t, a_t)},
\end{aligned}$$

where the second inequality uses $\text{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right] \leq H^2$ and $\|\overline{V}_p^k - \underline{V}_p^k\|_{\hat{\mathbb{P}}_p^k}^2 \leq H^2$.

As a consequence, using Lemma 27,

$$\begin{aligned}
H \wedge \text{ind-}b_p^k(s_t, a_t) \wedge \text{agg-}b_p^k(s_t, a_t) \lesssim& H \wedge \left(H\sqrt{\frac{SL(n_p^k(s_t, a_t))}{n_p^k(s_t, a_t)}} + \frac{HSL(n_p^k(s_t, a_t))}{n_p^k(s_t, a_t)}\right) \\
\lesssim& H\left(1 \wedge \sqrt{\frac{SL(n_p^k(s_t, a_t))}{n_p^k(s_t, a_t)}}\right).
\end{aligned}$$

□

Lemma 19. *If E happens, we have the following statements holding for all p, k, s, a:*

1. For two terms that appear in ind-$b_p^k(s,a)$, they are bounded respectively as:

$$\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} \lesssim \|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{H^2 S L(n_p^k(s,a))}{n_p^k(s,a)} \tag{39}$$

$$\begin{aligned}\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \lesssim & \sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \\ & + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{H\sqrt{S}L(n_p^k(s,a))}{n_p^k(s,a)}\end{aligned} \tag{40}$$

2. For two terms that appear in agg-$b_p^k(s,a)$, they are bounded respectively as:

$$\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} \lesssim 2\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{H^2 S L(n_p^k(s,a))}{n_p^k(s,a)} + H\epsilon \tag{41}$$

$$\begin{aligned}\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n^k(s,a))}{n^k(s,a)}} \lesssim & \sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n^k(s,a))}{n^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n^k(s,a))}{n^k(s,a)}} \\ & + \frac{H\sqrt{S}L(n^k(s,a))}{n^k(s,a)} + \sqrt{\frac{H\epsilon L(n^k(s,a))}{n^k(s,a)}}\end{aligned} \tag{42}$$

Proof. First, Lemmas 16 and 15 together imply that if E happens, the value function upper and lower bounds are valid. Conditioned on E happening, we prove the two items respectively.

1. For Equation (39), using the definition of $E_{\mathrm{ind,prob}}$ and AM-GM inequality, when E happens, we have for all p, k, s, a, s',

$$\hat{\mathbb{P}}_p^k(s' \mid s,a) \lesssim \mathbb{P}_p(s' \mid s,a) + \frac{L(n_p^k(s,a))}{n_p^k(s,a)}. \tag{43}$$

This implies that

$$\begin{aligned}&\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} \\ =& \sum_{s' \in \mathcal{S}_{h+1}} \hat{\mathbb{P}}_p^k(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 \\ \lesssim& \sum_{s' \in \mathcal{S}_{h+1}} \mathbb{P}_p^k(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 + \sum_{s' \in \mathcal{S}_{h+1}} \frac{L(n_p^k(s,a))}{n_p^k(s,a)} \cdot H^2 \\ \lesssim& \|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{SH^2 L(n_p^k(s,a))}{n_p^k(s,a)},\end{aligned}$$

where the first inequality is from Equation (43), and the fact that $\overline{V}_p^k(s') - \underline{V}_p^k(s') \in [0, H]$ for any $s' \in \mathcal{S}_{h+1}$; the second inequality is by algebra.

For Equation (40), we have:

$$
\begin{aligned}
&\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim &\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim &\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)} \\
\lesssim &\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)}
\end{aligned}
$$

where the first inequality is from Lemma 24 and the observation that when E happens, $\left|(\overline{V}_p^k - V_p^\star)(s')\right| \leq \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all $s' \in \mathcal{S}_{h+1}$; the second inequality is from Equation (29) of Lemma 17 and Equation (39); the third inequality again uses Lemma 24 and the observation that when E happens, $\left|(V_p^\star - V_p^{\pi^k})(s')\right| \leq \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for all $s' \in \mathcal{S}_{h+1}$.

2. For Equation (41), using the definition of $E_{\mathrm{agg,prob}}$ and AM-GM inequality, when E happens, we have for all p, k, s, a, s',

$$
\hat{\mathbb{P}}^k(s' \mid s,a) \lesssim \bar{\mathbb{P}}^k(s' \mid s,a) + \frac{L(n^k(s,a))}{n^k(s,a)}. \tag{44}
$$

This implies that

$$
\begin{aligned}
&\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} \\
=&\sum_{s' \in \mathcal{S}_{h+1}} \hat{\mathbb{P}}^k(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 \\
\lesssim&2\sum_{s' \in \mathcal{S}_{h+1}} \bar{\mathbb{P}}^k(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 + \sum_{s' \in \mathcal{S}_{h+1}} \frac{L(n_p^k(s,a))}{n_p^k(s,a)} \cdot H^2 \\
\lesssim&2\sum_{s' \in \mathcal{S}_{h+1}} \mathbb{P}_p(s' \mid s,a)(\overline{V}_p^k(s') - \underline{V}_p^k(s'))^2 + \epsilon H + \frac{SH^2 L(n_p^k(s,a))}{n_p^k(s,a)} \\
\lesssim&\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} + \frac{SH^2 L(n_p^k(s,a))}{n_p^k(s,a)} + \epsilon H,
\end{aligned}
$$

where the first inequality is from Equation (44) and the fact that $\overline{V}_p^k(s') - \underline{V}_p^k(s') \in [0, H]$ for any $s' \in \mathcal{S}_{h+1}$; the second inequality is from the observation that $\|\mathbb{P}_p(\cdot \mid s,a) - \bar{\mathbb{P}}^k(\cdot \mid s,a)\|_1 \leq \frac{\epsilon}{H}$; the third inequality is by algebra.

For Equation (42), we have:

$$
\begin{aligned}
&\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[\overline{V}_p^k\right] L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim&\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim&\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^\star\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)} + \sqrt{\frac{H\epsilon L(n_p^k(s,a))}{n_p^k(s,a)}} \\
\lesssim&\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}\left[V_p^{\pi^k}\right] L(n_p^k(s,a))}{n_p^k(s,a)}} + \sqrt{\frac{\|\overline{V}_p^k - \underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)} L(n_p^k(s,a))}{n_p^k(s,a)}} + \frac{\sqrt{S}HL(n_p^k(s,a))}{n_p^k(s,a)} + \sqrt{\frac{H\epsilon L(n_p^k(s,a))}{n_p^k(s,a)}},
\end{aligned}
$$

where the first inequality is from Lemma 24 and the observation that when E happens, $\left|(\overline{V}_p^k - V_p^\star)(s')\right| \leq \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for $s' \in \mathcal{S}_{h+1}$; the second inequality uses Equation (31) of Lemma 17 and Equation (41); the third inequality is from Lemma 24 and the observation that when E happens, $\left|(\overline{V}_p^\star - V_p^{\pi^k})(s')\right| \leq \left|(\overline{V}_p^k - \underline{V}_p^k)(s')\right|$ for $s' \in \mathcal{S}_{h+1}$. □

E.3.1. Distribution-dependent bound on the surplus terms

Lemma 20 (Surplus bound)**.** *If E happens, then for all p, k, s, a:*

$$
E_p^k(s,a) \lesssim B_p^{k,\mathrm{lead}}(s,a) + \mathbb{E}\left[\sum_{t=h}^{H} B_p^{k,\mathrm{fut}}(s_t,a_t) \mid (s_h,a_h) = (s,a), \pi^k(p), \mathcal{M}_p\right],
$$

where

$$
\begin{aligned}
B_p^{k,\mathrm{lead}}(s,a) &= H \wedge \left(5\epsilon + O\left(\sqrt{\frac{\left(1 + \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) L(n^k(s,a))}{n^k(s,a)}}\right)\right) \\
&\quad \wedge O\left(\sqrt{\frac{\left(1 + \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) L(n_p^k(s,a))}{n_p^k(s,a)}}\right), \\
B_p^{k,\mathrm{fut}}(s,a) &= H^3 \wedge O\left(\frac{H^3 SL(n_p^k(s,a))}{n_p^k(s,a)}\right).
\end{aligned}
$$

Proof of Lemma 20. First, Lemmas 16 and 15 together imply that if E holds, for all p,k,s,a, $E_p^k(s,a) \leq 2\left(H \wedge \text{ind-}b_p^k(s,a) \wedge \text{agg-}b_p^k(s,a)\right)$. We now bound $\text{ind-}b_p^k(s,a)$ and $\text{agg-}b_p^k(s,a)$ respectively.

Bounding ind-$b_p^k(s,a)$: We have

$$
\begin{aligned}
&\text{ind-}b_p^k(s,a)\\
=&O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}[\overline{V}_p^k]\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\hat{\mathbb{P}}_p^k(\cdot|s,a)}\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)\\
\leq&O\left(\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)\\
\leq&O\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)\\
\leq&O\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\;L(n_p^k(s,a))}{n_p^k(s,a)}}+\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}+\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)}\right)
\end{aligned}
$$

where the first inequality is by expanding the definition of ind-$b_p^k(s,a)$ and algebra; the second inequality is from Equations (40) and Equation (39) of Lemma 19, along with algebra; the third inequality is by the basic fact that $\sqrt{A}+\sqrt{B}\lesssim\sqrt{A+B}$; the fourth inequality is by AM-GM inequality.

Bounding agg-$b_p^k(s,a)$: We have:

$$
\begin{aligned}
&\text{agg-}b_p^k(s,a)\\
\lesssim&4\epsilon+O\left(\sqrt{\frac{\mathrm{var}_{\hat{\mathbb{P}}^k(\cdot|s,a)}[\overline{V}_p^k]\;L(n^k(s,a))}{n^k(s,a)}}+\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}+\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\hat{\mathbb{P}}^k(\cdot|s,a)}\;L(n^k(s,a))}{n^k(s,a)}}+\frac{SHL(n^k(s,a))}{n^k(s,a)}\right)\\
\lesssim&5\epsilon+O\left(\sqrt{\frac{\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\;L(n^k(s,a))}{n^k(s,a)}}+\sqrt{\frac{L(n^k(s,a))}{n^k(s,a)}}+\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\;L(n^k(s,a))}{n^k(s,a)}}+\frac{SHL(n^k(s,a))}{n^k(s,a)}\right)\\
\lesssim&5\epsilon+O\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\;L(n^k(s,a))}{n^k(s,a)}}+\sqrt{\frac{S\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}\;L(n^k(s,a))}{n^k(s,a)}}+\frac{SHL(n^k(s,a))}{n^k(s,a)}\right)\\
\leq&5\epsilon+O\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\;L(n^k(s,a))}{n^k(s,a)}}+\|\overline{V}_p^k-\underline{V}_p^k\|^2_{\mathbb{P}_p(\cdot|s,a)}+\frac{SHL(n^k(s,a))}{n^k(s,a)}\right)
\end{aligned}
$$

where the first inequality is by expanding the definition of agg-$b_p^k(s,a)$ and algebra; the second inequality is from Equations (42) and Equation (41) of Lemma 19, along with the observation that $\sqrt{\frac{S\epsilon HL(n^k(s,a))}{n^k(s,a)}}\leq\frac{SHL(n^k(s,a))}{n^k(s,a)}+\epsilon$ by AM-GM inequality; the third inequality is by the basic fact that $\sqrt{A}+\sqrt{B}\lesssim\sqrt{A+B}$; the fourth inequality is from AM-GM inequality.

Combining the above upper bounds, and using the observation that $\frac{L(n^k(s,a))}{n^k(s,a)} \leq \frac{L(n_p^k(s,a))}{n_p^k(s,a)}$, we get

$$
\begin{aligned}
&\text{ind-}b_p^k(s,a) \wedge \text{agg-}b_p^k(s,a) \wedge H \\
\leq& O\left(\sqrt{\frac{\left(1+\operatorname{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n_p^k(s,a))}{n_p^k(s,a)}}\right) \wedge \left(5\epsilon + O\left(\sqrt{\frac{\left(1+\operatorname{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)\ L(n^k(s,a))}{n^k(s,a)}}\right)\right) \wedge H \\
&+ O\left(\|\overline{V}_p^k - \underline{V}_p^k\|_{\mathbb{P}_p(\cdot|s,a)}^2 + \left(\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)} \wedge H\right)\right) \\
\leq& B^{k,\text{lead}}(s,a) + O\left(\|\overline{V}_p^k - \underline{V}_p^k\|_{\mathbb{P}_p(\cdot|s,a)}^2 + \left(\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)} \wedge H\right)\right).
\end{aligned}
$$

We now show that

$$
\|\overline{V}_p^k - \underline{V}_p^k\|_{\mathbb{P}_p(\cdot|s,a)}^2 + \left(\frac{SHL(n_p^k(s,a))}{n_p^k(s,a)} \wedge H\right) \lesssim \mathbb{E}\left[\sum_{t=h}^{H} B^{k,\text{fut}}(s_t,a_t) \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right], \tag{45}
$$

which will conclude the proof. To this end, we simplify the left hand side of Equation (45) using Lemma 18:

$$
\begin{aligned}
&\|\overline{V}_p^k - \underline{V}_p^k\|_{\mathbb{P}_p(\cdot|s,a)}^2 + \left(\frac{SHL(n^k(s,a))}{n^k(s,a)} \wedge H\right) \\
\lesssim& \mathbb{E}\left[\left(H\sum_{t=h+1}^{H} \mathbb{E}\left[\left(1 \wedge \sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}\right) \mid s_{h+1}\right]\right)^2 \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right] + \left(\frac{SHL(n^k(s,a))}{n^k(s,a)} \wedge H\right) \\
\lesssim& H^3\mathbb{E}\left[\sum_{t=h+1}^{H} \mathbb{E}\left[\left(1 \wedge \sqrt{\frac{SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)}}\right)^2 \mid s_{h+1}\right] \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right] + \left(\frac{SHL(n^k(s,a))}{n^k(s,a)} \wedge H\right) \\
\lesssim& \mathbb{E}\left[\sum_{t=h}^{H} H^3 \wedge \frac{H^3 SL(n_p^k(s_t,a_t))}{n_p^k(s_t,a_t)} \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right] \\
\lesssim& \mathbb{E}\left[\sum_{t=h}^{H} B^{k,\text{fut}}(s_t,a_t) \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right],
\end{aligned}
$$

where the first inequality is from Equation (38) of Lemma 18; the second inequality is by Cauchy-Schwarz and $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ for any random variable X; the third inequality is by the law of total expectation and algebra. $\square$

E.4. Concluding the regret bounds

In this section, we present the proofs of Theorems 4 and 8.

To bound the collective regret of MULTI-TASK-EULER, we first recall the following general result from (Simchowitz & Jamieson, 2019), which is useful to establish instance-dependent regret guarantees for episodic RL.

Lemma 21 (Clipping lemma, (Simchowitz & Jamieson, 2019), Lemma B.6)**.** *Fix player $p \in [M]$; suppose for each episode k, it follows $\pi^k(p)$, the greedy policy with respect to $\overline{Q}_p^k$. In addition, there exists some event E and a collection of functions $\left\{B_p^{k,\text{lead}}, B_p^{k,\text{fut}}\right\}_{k=1}^K \subset (\mathcal{S}\times\mathcal{A}\to\mathbb{R})$, such that if E happens, then for all $k \in [K]$, $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, the* *surplus of* $\overline{Q}_p^k$ *satisfies that*

$$0 \leq E_p^k(s,a) \lesssim B_p^{k,\mathrm{lead}}(s,a) + \mathbb{E}\left[\sum_{t=h}^{H} B_p^{k,\mathrm{fut}}(s_t,a_t) \mid (s_h,a_h)=(s,a), \pi^k(p), \mathcal{M}_p\right],$$

then, on E*:*

$$\mathrm{Reg}(K,p) \lesssim \sum_{s,a}\sum_{k} \rho_p^k(s,a)\,\mathrm{clip}\left(B_p^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right) + \sum_{s,a}\sum_{k} \rho_p^k(s,a)\,\mathrm{clip}\left(B_p^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right),$$

here, recall that $\mathrm{clip}(\alpha,\Delta) = \alpha \mathbf{1}(\alpha \geq \Delta)$*, and* $\breve{\mathrm{gap}}_p(s,a) = \frac{\mathrm{gap}_p(s,a)}{4H} \vee \frac{\mathrm{gap}_{p,\min}}{4H}$.

Remark 22. *Our presentation of the clipping lemma is slightly different than the original one (Simchowitz & Jamieson, 2019, Lemma B.6), in that:*

1. *We consider layered MDPs, while Simchowitz & Jamieson (2019) consider general stationary MDPs where one state may be experienced at multiple different steps in* $[H]$*. Specifically, in a layered MDP, the occupancy distributions* $\omega_{k,h}$ *defined in (Simchowitz & Jamieson, 2019) is only supported over* $\mathcal{S}_h \times \mathcal{A}$*. As a result, in the presentation here, we no longer need to sum over* h *– this is already captured in the sum over all* s *across all layers.*

2. *Our presentation here is in the multitask RL context, which is with respect to a a player* $p \in [M]$*, its corresponding MDP* $\mathcal{M}_p$*, and its policies used throughout the process* $\left\{\pi^k(p)\right\}_{k=1}^K$*. As a result, all quantities have* p *as subscripts.*

We are now ready to prove Theorems 4 and 8, MULTI-TASK-EULER's main regret theorems.

E.4.1. PROOF OF THEOREM 4

Proof of Theorem 4. From Lemma 21 and Lemma 20, we have that when E happens,

$$\begin{aligned}\mathrm{Reg}(K) &= \sum_{p=1}^{M} \mathrm{Reg}(K,p) \\ &\leq \underbrace{\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right)}_{(A)} + \underbrace{\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right)}_{(B)},\end{aligned} \tag{46}$$

We bound each term separately. We can directly use Lemma 23 to bound term (B) as:

$$\sum_{s,a}\sum_{k,p} \rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \lesssim MH^3S^2A\left(\ln\left(\frac{MSAK}{\delta}\right)\right)^2. \tag{47}$$

For term (A), we will group the sum by $(s,a) \in \mathcal{I}_\epsilon$ and $(s,a) \notin \mathcal{I}_\epsilon$ separately.

Case 1: $(s,a) \in \mathcal{I}_\epsilon$**.** In this case, we have that for all p, $\breve{\mathrm{gap}}_p(s,a) = \frac{\mathrm{gap}_p(s,a)}{4H} \geq 24\epsilon$. We simplify the corresponding term as follows:

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right) \\
\leq &\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(H \wedge \left(5\epsilon + O\left(\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\right), \frac{\min_p \mathrm{gap}_p(s,a)}{4H}\right) \\
\leq &\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \left(H \wedge \operatorname{clip}\left(5\epsilon + O\left(\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right), \frac{\min_p \mathrm{gap}_p(s,a)}{4H}\right)\right) \\
\lesssim &\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k,p} \rho_p^k(s,a) \left(H \wedge \sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)
\end{aligned}$$

where the first inequality is from the definition of $B^{k,\mathrm{lead}}$; the second inequality is from the basic fact that $\operatorname{clip}(A \wedge B, C) \leq A \wedge \operatorname{clip}(B,C)$; the third inequality uses Lemma 28 with $a_1 = 5\epsilon$, $a_2 = \sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}$, and $\Delta = \frac{\min_p \mathrm{gap}_p(s,a)}{4H}$, along with the observation that $\operatorname{clip}(5\epsilon, \frac{\min_p \mathrm{gap}_p(s,a)}{16H}) = 0$, since for all $(s,a) \in \mathcal{I}_\epsilon$ and all $p \in [M]$, $\mathrm{gap}_p(s,a) \geq 96\epsilon H$.

We now decompose the inner sum over k, $\sum_{k=1}^K$, to $\sum_{k=1}^{\tau(s,a)-1}$ and $\sum_{k=\tau(s,a)}^K$. The first part is bounded by:

$$\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=1}^{\tau_p(s,a)-1} \sum_{p=1}^M \rho_p^k(s,a) \left(H \wedge \sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right) \leq \sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=1}^{\tau_p(s,a)-1} \sum_{p=1}^M \rho_p^k(s,a) H \leq SAHN_1,$$

which is $\lesssim MHSA \ln\left(\frac{SAK}{\delta}\right)$.

For the second part,

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=\tau(s,a)}^K \sum_{p=1}^M \rho_p^k(s,a) \left(H \wedge \sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right) \\
\lesssim &\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=\tau(s,a)}^K \sum_{p=1}^M \rho_p^k(s,a) \sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}} \\
\lesssim &\sqrt{\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=\tau(s,a)}^K \sum_{p=1}^M \rho_p^k(s,a) \cdot \frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}} \cdot \sqrt{\sum_{(s,a)\in\mathcal{I}_\epsilon} \sum_{k=1}^K \sum_{p=1}^M \rho_p^k(s,a) \left(1 + \mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)},
\end{aligned}$$

where the first inequality is by dropping the "$H\wedge$" operator; the second inequality is by Cauchy-Schwarz.

We bound each factor as follows: for the first factor,

$$\begin{aligned}\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)} &= \sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\rho^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}\\ &\le L(MK)\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\frac{\rho^k(s,a)}{\bar{n}^k(s,a)}\\ &\le \sum_{(s,a)\in\mathcal{I}_\epsilon}L(MK)\cdot\int_1^{\bar{n}^K(s,a)}\frac{1}{u}du\\ &\le |\mathcal{I}_\epsilon|\,L(MK)^2 \lesssim |\mathcal{I}_\epsilon|\left(\ln\left(\frac{MSAK}{\delta}\right)\right)^2,\end{aligned}$$

where the first inequality is because L is monotonically increasing, and $\bar{n}^k(s,a)\le MK$; the second inequality is from the observation that $\rho^k(s,a)\in[0,M]$, $\bar{n}^k(s,a)\ge 2M$, and $u\mapsto\frac{1}{u}$ is monotonically decreasing; the last two inequalities are by algebra.

For the second factor,

$$\begin{aligned}\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=1}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right) &\lesssim MKH+\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\rho_p^k(s,a)\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\\ &\lesssim MKH+\sum_{p=1}^{M}\sum_{k=1}^{K}\mathrm{Var}\left[\sum_{h=1}^{H}r_{h,p}^k\mid\pi^k(p)\right]\\ &\lesssim MKH^2.\end{aligned}\tag{48}$$

where the first inequality is by the fact that ρ_p^k are probability distributions over every layer $h\in[H]$; the last two inequalities are by a law of total variance identity (see, e.g., Azar et al., 2017, Equation (26)). To summarize, the second part is at most

$$\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k=\tau(s,a)}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\lesssim\sqrt{MKH^2|\mathcal{I}_\epsilon|}\ln\left(\frac{MSAK}{\delta}\right).$$

Combining the bounds for the first and the second parts, we have:

$$\sum_{(s,a)\in\mathcal{I}_\epsilon}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a),\breve{\mathrm{gap}}_p(s,a)\right)\lesssim\left(\sqrt{MKH^2|\mathcal{I}_\epsilon|}+MHSA\right)\ln\left(\frac{MSAK}{\delta}\right).$$

Case 2: $(s,a)\notin\mathcal{I}_\epsilon$**.** We simplify the corresponding term as follows:

$$\begin{aligned}&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a),\breve{\mathrm{gap}}_p(s,a)\right)\\ \lesssim&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(H\wedge\left(\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(n_p^k(s,a))}{n_p^k(s,a)}}\right),\frac{\breve{\mathrm{gap}}_p(s,a)}{4H}\right)\\ \lesssim&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k,p}\left(H\wedge\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(n_p^k(s,a))}{n_p^k(s,a)}}\right)\end{aligned}$$

For each p and (s,a), we now decompose the inner sum over k, $\sum_{k=1}^{K}$, to $\sum_{k=1}^{\tau_p(s,a)-1}$ and $\sum_{k=\tau_p(s,a)}^{K}$. The first part is bounded by:

$$\begin{aligned}\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=1}^{\tau_p(s,a)-1}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n_p^k(s,a))}{n_p^k(s,a)}}\right)\leq&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=1}^{\tau_p(s,a)-1}\rho_p^k(s,a)H\\ \leq& MHSAN_2,\end{aligned}$$

which is $\lesssim MHSA\ln\left(\frac{MSAK}{\delta}\right)$.

For the second part,

$$\begin{aligned}&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(n_p^k(s,a))}{n_p^k(s,a)}}\right)\\ \lesssim&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(\bar{n}_p^k(s,a))}{\bar{n}_p^k(s,a)}}\\ \leq&\sqrt{\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\cdot\frac{L(\bar{n}_p^k(s,a))}{\bar{n}_p^k(s,a)}}\cdot\sqrt{\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k=1}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)}\end{aligned}$$

We bound each factor as follows: for the first factor,

$$\begin{aligned}\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\cdot\frac{L(\bar{n}^k(s,a))}{\bar{n}^k(s,a)}\leq& L(K)\cdot\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\frac{\rho_p^k(s,a)}{\bar{n}^k(s,a)}\\ \leq& L(K)\cdot\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\int_1^{\bar{n}_p^K(s,a)}\frac{1}{u}du\\ \leq&\left|\mathcal{I}_\epsilon^C\right|ML(K)^2\leq\left|\mathcal{I}_\epsilon^C\right|M\left(\ln\left(\frac{MSAK}{\delta}\right)\right)^2.\end{aligned}$$

where the first inequality is because L is monotonically increasing, and $\bar{n}_p^k(s,a)\leq K$; the second inequality is from the observation that $\rho^k(s,a)\in[0,1]$, $\bar{n}^k(s,a)\geq 2$, and $u\mapsto\frac{1}{u}$ is monotonically decreasing; the last two inequalities are by algebra.

The second factor is again bounded by (48). Therefore, the second part of the sum is at most

$$\begin{aligned}&\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{p=1}^{M}\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\left(H\wedge\sqrt{\frac{\left(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}]\right)L(n_p^k(s,a))}{n_p^k(s,a)}}\right)\\ \leq&\left(M\sqrt{KH^2\left|\mathcal{I}_\epsilon^C\right|}+MHSA\right)\ln\left(\frac{MSAK}{\delta}\right).\end{aligned}$$

Combining the bounds for the first and the second parts, we have:

$$\sum_{(s,a)\notin\mathcal{I}_\epsilon}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a),\breve{\mathrm{gap}}_p(s,a)\right)\lesssim\left(M\sqrt{KH^2\left|\mathcal{I}_\epsilon^C\right|}+MHSA\right)\ln\left(\frac{MSAK}{\delta}\right).$$

Now, combining the bounds for cases 1 and 2, we have that

$$(A)\leq\left(\sqrt{MKH^2|\mathcal{I}_\epsilon|}+M\sqrt{KH^2\left|\mathcal{I}_\epsilon^C\right|}+MHSA\right)\cdot\ln\left(\frac{MSAK}{\delta}\right).\tag{49}$$

In conclusion, by the regret decomposition Equation (46), and Equations (49) and (47), we have:

$$\mathrm{Reg}(K) \le \left(\sqrt{MH^2|\mathcal{I}_\epsilon|\,K} + M\sqrt{H^2\left|\mathcal{I}_\epsilon^C\right|K} + MH^3S^2A\ln\left(\frac{MSAK}{\delta}\right)\right)\ln\left(\frac{MSAK}{\delta}\right). \qquad \square$$

E.4.2. Proof of Theorem 8

Proof of Theorem 8. From Lemma 21, we have that when E happens,

$$\begin{aligned}
\mathrm{Reg}(K) &= \sum_{p=1}^{M} \mathrm{Reg}(K,p) \\
&\le \underbrace{\sum_{s,a}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right)}_{(A)} + \underbrace{\sum_{s,a}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right)}_{(B)},
\end{aligned}$$

We focus on each term separately. We directly use Lemma 23 to bound term (B) as:

$$\sum_{s,a}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \lesssim MH^3S^2A\ln\left(\frac{MSAK}{\delta}\right)\cdot\ln\frac{MHSA}{\mathrm{gap}_{\min}}. \tag{50}$$

For the (s,a)-th term in term (A), we will consider the cases of $(s,a)\in\mathcal{I}_\epsilon$ and $(s,a)\notin\mathcal{I}_\epsilon$ separately.

Case 1: $(s,a)\in\mathcal{I}_\epsilon$**.** In this case, we have that for all p, $\breve{\mathrm{gap}}_p(s,a) = \frac{\mathrm{gap}_p(s,a)}{4H} \ge 24\epsilon$. We simplify the corresponding term as follows:

$$\begin{aligned}
&\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\mathrm{gap}}_p(s,a)\right) \\
\le& \sum_{k=1}^{K}\sum_{p=1}^{M}\rho_p^k(s,a)\,\mathrm{clip}\left(H\wedge\left(5\epsilon + O\left(\sqrt{\frac{(1+\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}])L(n^k(s,a))}{n^k(s,a)}}\right)\right), \frac{\min_p \mathrm{gap}_p(s,a)}{4H}\right) \\
\le& \sum_{k=1}^{K}\rho^k(s,a)\,\mathrm{clip}\left(H\wedge\left(5\epsilon + O\left(\sqrt{\frac{H^2L(n^k(s,a))}{n^k(s,a)}}\right)\right), \frac{\min_p \mathrm{gap}_p(s,a)}{4H}\right) \\
\le& \sum_{k=1}^{k}\rho^k(s,a)\left(H\wedge\mathrm{clip}\left(5\epsilon + O\left(\sqrt{\frac{H^2L(n^k(s,a))}{n^k(s,a)}}\right), \frac{\min_p \mathrm{gap}_p(s,a)}{4H}\right)\right) \\
\lesssim& \sum_{k=1}^{K}\rho^k(s,a)\left(H\wedge\mathrm{clip}\left(\sqrt{\frac{H^2L(n^k(s,a))}{n^k(s,a)}}, \frac{\min_p \mathrm{gap}_p(s,a)}{16H}\right)\right)
\end{aligned}$$

where the first inequality is by the definition of $B^{k,\mathrm{lead}}$; the second inequality is from that $\mathrm{var}_{\mathbb{P}_p(\cdot|s,a)}[V_p^{\pi^k}] \le H^2$; the third inequality is from that $\mathrm{clip}(A\wedge B, C) \le A\wedge\mathrm{clip}(B,C)$; the third inequality uses Lemma 28 with $a_1 = 5\epsilon$, $a_2 = \sqrt{\frac{H^2L(n^k(s,a))}{n^k(s,a)}}$, and $\Delta = \frac{\min_p \mathrm{gap}_p(s,a)}{4H}$, along with the observation that $\mathrm{clip}(5\epsilon, \frac{\min_p \mathrm{gap}_p(s,a)}{16H}) = 0$, since for all $(s,a)\in\mathcal{I}_\epsilon$ and all $p\in[M]$, $\mathrm{gap}_p(s,a)\ge 96\epsilon H$.

We now decompose the inner sum over k, $\sum_{k=1}^{K}$, to $\sum_{k=1}^{\tau(s,a)-1}$ and $\sum_{k=\tau(s,a)}^{K}$. The first part's contribution is at most $N_1 \cdot H \lesssim MH \ln\left(\frac{SAK}{\delta}\right)$. For the second part, its contribution is at most:

$$\begin{aligned}
&\sum_{k=\tau(s,a)}^{K} \rho^k(s,a) \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}, \frac{\min_p \operatorname{gap}_p(s,a)}{16H} \right) \right) \\
\lesssim & MH + \int_1^{\bar{n}^K(s,a)} \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(u)}{u}}, \frac{\min_p \operatorname{gap}_p(s,a)}{16H} \right) \right) du \\
\lesssim & MH + \frac{H^3}{\min_p \operatorname{gap}_p(s,a)} \ln\left(\frac{MSAK}{\delta}\right)
\end{aligned}$$

where the second inequality is from Lemma 29 with $f_{\max} = H$, $C = H^2$, $\Delta = \frac{\min_p \operatorname{gap}_p(s,a)}{16H}$, $N = MSA$, $\xi = \delta$, $\Gamma = 1$, $n = \bar{n}^K(s,a) \leq K$. In summary, for all $(s,a) \in \mathcal{I}_\epsilon$,

$$\sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\operatorname{gap}}_p(s,a)\right) \leq \left(MH + \frac{H^3}{\min_p \operatorname{gap}_p(s,a)} \right) \ln\left(\frac{MSAK}{\delta}\right).$$

Case 2: $(s,a) \notin \mathcal{I}_\epsilon$**.** In this case, for each $p \in [M]$, we simplify the corresponding term as follows:

$$\begin{aligned}
&\sum_{k} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\operatorname{gap}}_p(s,a)\right) \\
\lesssim & \sum_{k=1}^{K} \rho_p^k(s,a) \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(n_p^k(s,a))}{n_p^k(s,a)}}, \frac{\breve{\operatorname{gap}}_p(s,a)}{16H} \right) \right)
\end{aligned}$$

We now decompose the inner sum over k, $\sum_{k=1}^{K}$, to $\sum_{k=1}^{\tau_p(s,a)-1}$ and $\sum_{k=\tau_p(s,a)}^{K}$. The first part's contribution is at most $N_2 \cdot H \lesssim H \ln\left(\frac{MSAK}{\delta}\right)$.

For the second part, its contribution is at most:

$$\begin{aligned}
&\sum_{k=\tau_p(s,a)}^{K} \rho_p^k(s,a) \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(n^k(s,a))}{n^k(s,a)}}, \frac{\breve{\operatorname{gap}}_p(s,a)}{16H} \right) \right) \\
\lesssim & H + \int_1^{\bar{n}_p^K(s,a)} \left(H \wedge \operatorname{clip}\left(\sqrt{\frac{H^2 L(u)}{u}}, \frac{\breve{\operatorname{gap}}_p(s,a)}{16H} \right) \right) du \\
\lesssim & H + \frac{H^3}{\breve{\operatorname{gap}}_p(s,a)} \ln\left(\frac{MSAK}{\delta}\right)
\end{aligned}$$

where the second inequality is from Lemma 29 with $f_{\max} = H$, $C = H^2$, $\Delta = \frac{\breve{\operatorname{gap}}_p(s,a)}{16H}$, $N = MSA$, $\xi = \delta$, $\Gamma = 1$, $n = \bar{n}_p^K(s,a) \leq K$. In summary, for any $(s,a) \in \mathcal{I}_\epsilon^C$ and $p \in [M]$,

$$\sum_{k} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\operatorname{gap}}_p(s,a)\right) \lesssim (H + \frac{H^3}{\breve{\operatorname{gap}}_p(s,a)}) \ln\left(\frac{MSAK}{\delta}\right),$$

summing over p, we get:

$$\sum_{k,p} \rho_p^k(s,a) \operatorname{clip}\left(B^{k,\mathrm{lead}}(s,a), \breve{\operatorname{gap}}_p(s,a)\right) \lesssim \left(MH + \sum_{p=1}^{M} \frac{H^3}{\breve{\operatorname{gap}}_p(s,a)} \right) \ln\left(\frac{MSAK}{\delta}\right),$$

In summary, combining the regret bounds of cases 1 and 2 for term (A), along with Equation (50) for term (B), and observe that $\breve{\mathrm{gap}}_p(s,a) = \mathrm{gap}_{p,\min}$ if $(s,a) \in Z_{p,\mathrm{opt}}$, and $\breve{\mathrm{gap}}_p(s,a) = \mathrm{gap}_p(s,a)$ otherwise, we have that on event E, MULTI-TASK-EULER satisfies:

$$\mathrm{Reg}(K) \lesssim \ln\left(\frac{MSAK}{\delta}\right)\left(\sum_{p\in[M]}\left(\sum_{(s,a)\in Z_{p,\mathrm{opt}}}\frac{H^3}{\mathrm{gap}_{p,\min}} + \sum_{(s,a)\in(\mathcal{I}_\epsilon\cup Z_{p,\mathrm{opt}})^C}\frac{H^3}{\mathrm{gap}_p(s,a)}\right) + \sum_{(s,a)\in\mathcal{I}_\epsilon}\frac{H^3}{\min_p \mathrm{gap}_p(s,a)}\right) + \ln\left(\frac{MSAK}{\delta}\right)\cdot MS^2AH^3\ln\frac{MHSA}{\mathrm{gap}_{\min}}. \qquad \square$$

Lemma 23 (Bounding the lower order terms)**.** *If E happens, then*

$$\sum_{s,a}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \lesssim MH^3S^2A\ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right)\wedge\ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right).$$

Proof. We expand the left hand side using the definition of $B^{k,\mathrm{fut}}$, and the fact that $\mathrm{gap}_{p,\min} \geq \mathrm{gap}_{\min}$:

$$\sum_{k=1}^{K}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{fut}}(s,a), \frac{\mathrm{gap}_{p,\min}}{8SAH^2}\right) \tag{51}$$

$$\lesssim \sum_{k=1}^{K}\rho_p^k(s,a)\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(n_p^k(s,a))}{n_p^k(s,a)}, \frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right) \tag{52}$$

We now decompose the sum $\sum_{k=1}^{K}$ to $\sum_{k=1}^{\tau_p(s,a)-1}$ and $\sum_{k=\tau_p(s,a)}^{K}$. The first part can be bounded by

$$\sum_{k=1}^{\tau_p(s,a)-1}\rho_p^k(s,a)\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(n_p^k(s,a))}{n_p^k(s,a)}, \frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right) \leq \sum_{k=1}^{\tau_p(s,a)-1}H^3\rho_p^k(s,a) \leq H^3N_2,$$

which is at most $O\left(H^3\cdot\ln\left(\frac{MSAK}{\delta}\right)\right)$. For the second part, it can be bounded by:

$$\begin{aligned}
&\sum_{k=\tau_p(s,a)}^{K}\rho_p^k(s,a)\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(n_p^k(s,a))}{n_p^k(s,a)}, \frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right)\\
\leq& H^3\cdot 1 + \int_1^{\bar{n}_p^K(s,a)}\left(H^3\wedge\mathrm{clip}\left(\frac{H^3SL(u)}{u}, \frac{\mathrm{gap}_{\min}}{8SAH^2}\right)\right)du\\
\lesssim& H^3 + H^3\ln\left(\frac{MSA}{\delta}\right) + H^3S\ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right)\wedge\ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right),
\end{aligned}$$

where the second inequality is from Lemma 29 with $f_{\max} = H^3$, $C = H^3S$, $\Delta = \frac{\mathrm{gap}_{\min}}{8SAH^2}$, $N = MSA$, $\xi = \delta$, $\Gamma = 1$, $n = \bar{n}_p^K(s,a) \leq K$. In summary,

$$\sum_k\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \frac{\mathrm{gap}_{\min}}{8SAH^2}\right) \lesssim H^3S\ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right)\wedge\ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right)$$

Summing over $s\in\mathcal{S}$, $a\in\mathcal{A}$, and $p\in[M]$, we get

$$\sum_{s,a}\sum_{k,p}\rho_p^k(s,a)\,\mathrm{clip}\left(B^{k,\mathrm{lead}}(s,a), \frac{\mathrm{gap}_{\min}}{8SAH^2}\right) \lesssim MH^3S^2A\ln\left(\frac{MSAK}{\delta}\right)\left(\ln\left(\frac{MSAK}{\delta}\right)\wedge\ln\left(\frac{MHSA}{\mathrm{gap}_{\min}}\right)\right).$$

$\square$

E.5. Miscellaneous lemmas

This subsection collects a few miscellaneous lemmas used throughout the upper bound proofs.

Lemma 24 ((Simchowitz & Jamieson, 2019), Lemma F.5)**.** *For random variables X and Y, $\left|\sqrt{\mathrm{var}[X]} - \sqrt{\mathrm{var}[Y]}\right| \leq \sqrt{\mathbb{E}\left[(X-Y)^2\right]}$.*

Lemma 25. *Suppose distributions P and Q are supported over $[0, B]$, and $\|P - Q\|_1 \leq \epsilon \leq 2$. Then:*

$$\left|\mathbb{E}_{X\sim P}[X] - \mathbb{E}_{X\sim Q}[X]\right| \leq B\epsilon,$$

$$\left|\mathrm{var}_{X\sim P}[X] - \mathrm{var}_{X\sim Q}[X]\right| \leq 3B^2\epsilon.$$

Proof. First,

$$\left|\mathbb{E}_{X\sim P}[X] - \mathbb{E}_{X\sim Q}[X]\right| = \left|\int_0^B x(p_X(x) - q_X(x))dx\right| \leq \int_0^B |x|\left|p_X(x) - q_X(x)\right| dx \leq B\|P - Q\|_1 \leq B\epsilon.$$

Second, observe that

$$\left|\mathbb{E}_{X\sim P}[X^2] - \mathbb{E}_{X\sim Q}[X^2]\right| \leq B^2\epsilon.$$

Meanwhile,

$$\left|(\mathbb{E}_{X\sim P}[X])^2 - (\mathbb{E}_{X\sim Q}[X])^2\right| \leq \left|\mathbb{E}_{X\sim P}[X] - \mathbb{E}_{X\sim Q}[X]\right| \cdot \left|\mathbb{E}_{X\sim P}[X] + \mathbb{E}_{X\sim Q}[X]\right| \leq 2B \cdot B\epsilon = 2B^2\epsilon.$$

Combining the above, we have

$$\left|\mathrm{var}_{X\sim P}[X] - \mathrm{var}_{X\sim Q}[X]\right| \leq 3B^2\epsilon.$$

□

Lemma 26. *For $A, B, C, D, E, F \geq 0$:*

1. *If $|A - B| \leq \sqrt{BC} + C$, then we have $\left|\sqrt{A} - \sqrt{B}\right| \leq 2\sqrt{C}$.*
2. *If $D \leq E + F\sqrt{D}$, then $\sqrt{D} \leq \sqrt{E} + F$.*

Proof. 1. First, $A - B \leq |A - B| \leq \sqrt{BC} + C$; this implies that $A \leq B + 2\sqrt{BC} + C$, and therefore $\sqrt{A} \leq \sqrt{B} + \sqrt{C}$. On the other hand, $B \leq A + C + \sqrt{BC}$; therefore, applying item 1 with $D = B$, $E = A + C$, and $F = \sqrt{C}$, we have $\sqrt{B} \leq \sqrt{A + C} + \sqrt{C} \leq \sqrt{A} + 2\sqrt{C}$.

2. The roots of $x^2 - Fx - E = 0$ are $\frac{F\pm\sqrt{F^2+4E}}{2}$, and therefore D must satisfy $\sqrt{D} \leq \frac{F+\sqrt{F^2+4E}}{2} \leq \frac{F+F+2\sqrt{E}}{2} = F + \sqrt{E}$.

□

Lemma 27. *For $a \geq 0$, $1 \wedge (a + \sqrt{a}) \leq 1 \wedge 2\sqrt{a}$.*

Proof. We consider the cases of $a \geq 1$ and $a < 1$ respectively. If $a \geq 1$, LHS $= 1 =$ RHS. Otherwise, $a \leq 1$; in this case, LHS $= 1 \wedge (a + \sqrt{a}) \leq 1 \wedge (\sqrt{a} + \sqrt{a}) =$ RHS. □

Lemma 28 (Special case of (Simchowitz & Jamieson, 2019), Lemma B.5)**.** *For $a_1, a_2, \Delta \geq 0$, $\mathrm{clip}(a_1 + a_2, \Delta) \leq 2\,\mathrm{clip}(a_1, \Delta/4) + 2\,\mathrm{clip}(a_2, \Delta/4)$.*

Lemma 29 (Integral calculation, (Simchowitz & Jamieson, 2019), Lemma B.9)**.** *Let $f(u) \leq \min(f_{\max}, \mathrm{clip}(g(u), \Delta))$, where $\Delta \in [0, \Gamma]$, and $g(u)$ is nonincreasing. Let $N \geq 1$ and $\xi \in (0, \frac{1}{2})$. Then:*

1. *If* $g(u) \lesssim \sqrt{\frac{C \log \frac{Nu}{\xi}}{u}}$ *for some* $C > 0$ *such that* $\ln C \lesssim \ln N$*, then*

$$\int_{\Gamma}^{n} f(u/4)du \lesssim \sqrt{Cn \ln \frac{Nn}{\xi}} \wedge \frac{C}{\Delta} \ln\left(\frac{Nn}{\xi}\right).$$

2. *If* $g(u) \lesssim \frac{C \ln \frac{Nu}{\xi}}{u}$ *for some* $C > 0$ *such that* $\ln C \lesssim \ln N$*, then*

$$\int_{\Gamma}^{n} f(u/4)du \lesssim f_{\max} \ln \frac{N}{\xi} + C \ln \frac{Nn}{\xi} \cdot \left(\ln \frac{Nn}{\xi} \wedge \ln \frac{N\Gamma}{\Delta}\right).$$

F. Proof of the Lower Bounds

F.1. Auxiliary Lemmas

Lemma 30 (Regret decomposition (Section H.2 of (Simchowitz & Jamieson, 2019)))**.** *For any MPERL problem instance and any algorithm, we have*

$$\mathbb{E}\left[\mathrm{Reg}(K)\right] \geq \sum_{p=1}^{M} \sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbb{E}\left[n_p^{K+1}(s,a)\right] \mathrm{gap}_p(s,a), \tag{53}$$

where we recall that $n_p^{K+1}(s,a)$ *is the number of visits of* (s,a) *by player* p *at the beginning of the* $(K+1)$*-th episode (after the first* K *episodes). Furthermore, for any* $(s,a) \in \mathcal{S}_1 \times \mathcal{A}$*, we have*

$$\sum_{p=1}^{M} \mathbb{E}\left[n_p^{K+1}(s,a)\right] \mathrm{gap}_p(s,a) \geq \mathbb{E}\left[n^{K+1}(s,a)\right] \left(\min_{p\in[M]} \mathrm{gap}_p(s,a)\right), \tag{54}$$

where we recall that $n^{K+1}(s,a) = \sum_{p=1}^{M} n_p^{K+1}(s,a)$.

Proof. Eq. (54) follows straightforwardly from the fact that for every $(s,a,p) \in \mathcal{S}_1 \times \mathcal{A} \times [M]$, $\min_{p'\in[M]} \mathrm{gap}_{p'}(s,a) \leq \mathrm{gap}_p(s,a)$.

We now prove Eq. (53). Let π_p^k denote $\pi^k(p)$. We have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}(K)\right] &= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} p_0(s_{1,p}^k = s)\left(V_p^\star(s) - V_p^{\pi_p^k}(s)\right)\right] \\
&\geq \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} p_0(s_{1,p}^k = s)\left(V_p^\star(s) - Q_p^\star(s,\pi_p^k(s))\right)\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} p_0(s)\mathrm{gap}_p(s,\pi_p^k(s))\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}_1} \mathbf{1}\left(s_{1,p}^k = s\right)\mathrm{gap}_p(s,\pi_p^k(s))\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{M}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbf{1}\left(s_{1,p}^k, \pi_p^k(s) = (s,a)\right)\mathrm{gap}_p(s,a)\right] \\
&= \sum_{p=1}^{M}\sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbb{E}\left[n_p^K(s,a)\right]\mathrm{gap}_p(s,a)
\end{aligned} \tag{55}$$

where the first equality is from the definition of collective regret; the first inequality is from the simple fact that $V_p^\pi(s) = Q_p^\pi(s, \pi(s)) \leq Q_p^\star(s, \pi(s))$ for any policy π; the second equality is from the definition of suboptimality gaps; and the third equality is from the basic observation that $s_{1,p}^k \sim p_0$. $\square$

Lemma 31 (Divergence decomposition (Xu et al., 2021))**.** *For two MPERL problem instances, $\mathfrak{M}$ and $\mathfrak{M}'$, which only differ in the transition probabilities $\left\{\mathbb{P}_p(\cdot \mid s, a)\right\}_{p\in[M],(s,a)\in\mathcal{S}\times\mathcal{A}}$, and for a fixed algorithm, let $\mathbb{P}_{\mathfrak{M}}$ and $\mathbb{P}_{\mathfrak{M}'}$ be the probability measures on the outcomes of running the algorithm on $\mathfrak{M}$ and $\mathfrak{M}'$, respectively. Then,*

$$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) = \sum_{p=1}^{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}_{\mathfrak{M}}\left[n_p^{K+1}(s,a)\right] \mathrm{KL}\left(\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s,a), \mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s,a)\right),$$

where $\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s,a)$ and $\mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s,a)$ are the transition probabilities of the problem instance $\mathfrak{M}$ and $\mathfrak{M}'$, respectively.

Lemma 32 (Bretagnolle-Huber inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions on the same measurable space, and A be an event. Then,*

$$\mathbb{P}(A) + \mathbb{Q}(A^C) \geq \frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}, \mathbb{Q})\right).$$

Lemma 33 (e.g., Lemma 25 of (Wang et al., 2021))**.** *For any $x, y \in [\frac{1}{4}, \frac{3}{4}]$, $\mathrm{KL}\left(\mathrm{Ber}(x), \mathrm{Ber}(y)\right) \leq 3(x-y)^2$.*

Lemma 34. *Let X be a Binomial random variable and $X \sim Bin(n,p)$, where $n \geq \frac{1}{p}$. Then,*

$$\mathbb{E}\left[X^{\frac{3}{2}}\right] \leq 2(np)^{\frac{3}{2}}.$$

Proof. Let $Y = X^2$, and $f(Y) = Y^{\frac{3}{4}}$. We have $\mathbb{E}[Y] = \mathbb{E}\left[X^2\right] = \mathrm{var}[X] + \mathbb{E}[X]^2 = (np)^2 + np(1-p) \leq (np)^2 + np \leq 2(np)^2$, where the last inequality follows from the assumption that $n \geq \frac{1}{p}$. By Jensen's inequality, we have $\mathbb{E}\left[X^{\frac{3}{2}}\right] = \mathbb{E}\left[f(Y)\right] \leq f\left(\mathbb{E}[Y]\right) \leq \left(2n^2p^2\right)^{\frac{3}{4}} \leq 2(np)^{\frac{3}{2}}$. $\square$

F.2. Gap independent lower bounds

Theorem 35 (Restatement of Theorem 5)**.** *For any $A \geq 2$, $H \geq 2$, $S \geq 4H$, $K \geq SA$, $M \in \mathbb{N}$, and $l, l^C \in \mathbb{N}$ such that $l + l^C = SA$ and $l \leq SA - 4(S + HA)$, there exists some ϵ such that for any algorithm* Alg*, there exists an ϵ-MPERL problem instance with S states, A actions, M players and an episode length of H such that $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \geq l$, and*

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(M\sqrt{H^2 l^C K} + \sqrt{MH^2 lK}\right).$$

Proof. The construction and techniques in this proof are inspired by (Wang et al., 2021, Section E.1) and (Simchowitz & Jamieson, 2019).

Fix any algorithm Alg; we consider two cases:

1. $l > Ml^C$;
2. $Ml^C \geq l$.

Case 1: $l > Ml^C$**.** Let $S_1 = S - 2(H-1)$, and $b = \lceil \frac{l}{S_1} \rceil \geq 1$. Let $\Delta = \sqrt{\frac{l+1}{384MK}}$, and let $\epsilon = \frac{1}{2}H\Delta$. We note that under the assumption that $K \geq SA$, and the observation that $l \leq SA$, we have $\Delta \leq \frac{1}{4}$. We define $(b+1)^{S_1}$ ϵ-MPERL problem instances, each indexed by an element in $[b+1]^{S_1}$. It suffices to show that, on at least one of the problem instances, $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(\sqrt{MH^2 lK}\right)$.

Construction. For $\mathbf{a} = (a_1, \ldots, a_{S_1}) \in [b+1]^{S_1}$, we define the following ϵ-MPERL problem instance, $\mathfrak{M}(\mathbf{a}) = \left\{\mathcal{M}_p\right\}_{p=1}^{M}$, with S states, A actions, and an episode length of H, such that for each $p \in [M]$, $\mathcal{M}_p$ is constructed as follows:

- $\mathcal{S}_1 = [S_1]$, and p_0 is a uniform distribution over the states in $\mathcal{S}_1$.
- For $h \in [2, H]$, $\mathcal{S}_h = \{S_1 + 2h - 3, S_1 + 2h - 2\}$.
- $\mathcal{A} = [A]$.
- For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution $r_p(s, a)$ is a Bernoulli distribution, $\mathrm{Ber}(R_p(s, a))$, and we will specify $R_p(s, a)$ subsequently.
- For each state $s \in [S_1]$,

$$\mathbb{P}_p(S_1 + 1 \mid s, a) = \begin{cases} \frac{1}{2} + \Delta, & \text{if } a = a_s; \\ \frac{1}{2}, & \text{if } a \in [b+1] \setminus a_s; \\ 0, & \text{if } a \notin [b+1]; \end{cases}$$

  and for each $a \in \mathcal{A}$, $\mathbb{P}_p(S_1 + 2 \mid s, a) = 1 - \mathbb{P}_p(S_1 + 1 \mid s, a)$, and $R_p(s, a) = 0$.
- For $h \in [2, H]$, and $a \in \mathcal{A}$, let
  - $\mathbb{P}_p\left(S_1 + 2h - 1 \mid S_1 + 2h - 3, a\right) = 1$, $\mathbb{P}_p\left(S_1 + 2h \mid S_1 + 2h - 3, a\right) = 0$, and $R_p(S_1 + 2h - 3, a) = 1$.
  - $\mathbb{P}_p\left(S_1 + 2h \mid S_1 + 2h - 2, a\right) = 0$, $\mathbb{P}_p\left(S_1 + 2h - 1 \mid S_1 + 2h - 2, a\right) = 1$, and $R_p(S_1 + 2h - 2, a) = 0$.

It can be easily verified that $\mathfrak{M}(\mathbf{a}) = \left\{\mathcal{M}_p\right\}_{p=1}^{M}$ is a 0-MPERL problem instance, and hence an ϵ-MPERL problem instance—the reward distributions and the transition probabilities are the same for all players, i.e., for every $p, q \in [M]$, and every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\left|R_p(s, a) - R_q(s, a)\right| = 0 \le \epsilon, \quad \left|\mathbb{P}_p(\cdot \mid s, a) - \mathbb{P}_q(\cdot \mid s, a)\right| = 0 \le \frac{\epsilon}{H}.$$

Suboptimality gaps. We now calculate the suboptimality gaps of the state-action pairs in the above MDPs. For each $p \in [M]$ and each $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathrm{gap}_p(s, a) = V_p^\star(s) - Q_p^\star(s, a) = \max_{a'} Q_p^\star(s, a') - Q_p^\star(s, a).$$

In $\mathfrak{M}(\mathbf{a})$, it can be easily observed that for every $p \in [M]$, and every $(s, a) \in \left(\mathcal{S} \setminus \mathcal{S}_1\right) \times \mathcal{A}$, $\mathrm{gap}_p(s, a) = 0$. Now, for every $p \in [M]$, $(s, a) \in \mathcal{S}_1 \times \mathcal{A}$, we have

$$\mathrm{gap}_p(s, a) = \max_{a'} Q_p^\star(s, a') - Q_p^\star(s, a) = (H - 1)\left(\max_{a'} \mathbb{P}_p(S_1 + 1 \mid s, a') - \mathbb{P}_p(S_1 + 1 \mid s, a)\right).$$

It follows that, for every $p \in [M]$ and every state $s \in [S_1]$,

$$\mathrm{gap}_p(s, a) = \begin{cases} 0, & \text{if } a = a_s; \\ (H-1)\Delta, & \text{if } a \in [b+1] \setminus a_s; \\ (H-1)\left(\frac{1}{2} + \Delta\right), & \text{if } a \notin [b+1]. \end{cases}$$

Subpar state-action pairs. It can be verified that in $\mathfrak{M}(\mathbf{a})$, $\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \ge l$. Specifically, since $(H-1)\Delta = (H-1)\frac{2\epsilon}{H} \ge \epsilon \ge \frac{\epsilon}{2} = 96H\frac{\epsilon}{192H}$, we have that the number of subpar state-action pairs is at least $S_1 b = S_1 \lceil \frac{l}{S_1} \rceil \ge l$.

It suffices to prove that

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([b+1]^{S_1}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \ge \frac{1}{640}\sqrt{MH^2 lK},$$

where we recall that $\mathbf{a} = (a_1, \dots, a_{S_1})$; furthermore, it suffices to show that, for any $s' \in [S_1]$,

$$\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N^{K+1}(s') - n^{K+1}(s', a_{s'})\right] \geq \frac{MK}{4S_1}, \tag{56}$$

where $N^{K+1}(s') = \sum_{a\in\mathcal{A}} n^{K+1}(s', a)$; this is because it follows from Eq. (56) that

$$\begin{aligned}
\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] &\geq \sum_{s'\in\mathcal{S}_1}(H-1)\frac{\Delta}{4}\cdot\mathbb{E}_{\mathbf{a}\sim\mathrm{Unif}\left([b+1]^{S_1}\right)}\mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N^{K+1}(s') - n^{K+1}(s', a_{s'})\right] \\
&\geq \sum_{s'\in\mathcal{S}_1}\frac{H}{2}\cdot\frac{\Delta}{4}\cdot\frac{MK}{4S_1} \\
&\geq \frac{1}{640}\sqrt{MH^2lK},
\end{aligned}$$

where the first inequality uses Lemma 30 (the regret decomposition lemma).

Without loss of generality, let us choose $s' = 1$. To prove Eq. (56), we use a standard technique and define a set of helper problem instances. Specifically, for any $(a_2, a_3, \dots, a_{S_1}) \in [b+1]^{S_1-1}$, we define a problem instance $\mathfrak{M}(0, a_2, \dots, a_{s_1})$ such that it agrees with $\mathfrak{M}(a_1, a_2, \dots, a_{s_1})$ on everything but $\mathbb{P}_p(\cdot \mid 1, a_1)$'s, i.e., in $\mathfrak{M}(0, a_2, \dots, a_{s_1})$, for every $p \in [M]$,

$$\mathbb{P}_p(S_1 + 1 \mid 1, a_1) = \frac{1}{2}.$$

Now, for each $(j, a_2, \dots, a_{s_1}) \in \left([0] \cup [b+1]\right) \times [b+1]^{S_1-1}$, let $\mathbb{P}_{j,a_2,\dots,a_{S_1}}$ denote the probability measure on the outcomes of running Alg on the problem instance $\mathfrak{M}(j, a_2, \dots, a_{s_1})$. Further, for each $j \in \{0\} \cup [b+1]$, we define

$$\mathbb{P}_j = \frac{1}{(b+1)^{S_1-1}}\sum_{a_2,\dots,a_{S_1}\in[b+1]^{S_1-1}}\mathbb{P}_{j,a_2,\dots,a_{S_1}};$$

and we use $\mathbb{E}_j$ to denote the expectation with respect to $\mathbb{P}_j$.

In subsequent calculations, for any index $m \in \left([0] \cup [b+1]\right) \times [b+1]^{S_1-1}$, we also denote by $\mathbb{P}_m\left(\cdot \mid N^{K+1}(1)\right)$ and $\mathbb{E}_m\left[\cdot \mid N^{K+1}(1)\right]$ the probability and expectation, respectively, conditional on a realization of $N^{K+1}(1)$ under $\mathbb{P}_m$. Observe that, for any $j \in \{0\} \cup [b+1]$,

$$\begin{aligned}
\mathbb{P}_j(\cdot \mid N^{K+1}(1)) &= \frac{\mathbb{P}_j(\cdot, N^{K+1}(1))}{\mathbb{P}_j(N^{K+1}(1))} \\
&= \frac{\frac{1}{(b+1)^{S_1-1}}\sum_{a_2,\dots,a_{S_1}\in[b+1]^{S_1-1}}\mathbb{P}_{j,a_2,\dots,a_{S_1}}(\cdot, N^{K+1}(1))}{\mathbb{P}_j(N^{K+1}(1))} \\
&= \frac{1}{(b+1)^{S_1-1}}\sum_{a_2,\dots,a_{S_1}\in[b+1]^{S_1-1}}\frac{\mathbb{P}_{j,a_2,\dots,a_{S_1}}(\cdot, N^{K+1}(1))}{\mathbb{P}_{j,a_2,\dots,a_{S_1}}(N^{K+1}(1))} \\
&= \frac{1}{(b+1)^{S_1-1}}\sum_{a_2,\dots,a_{S_1}\in[b+1]^{S_1-1}}\mathbb{P}_{j,a_2,\dots,a_{S_1}}(\cdot \mid N^{K+1}(1)),
\end{aligned} \tag{57}$$

where the first equality is from the definition of conditional probability; the second equality is from the definition of $\mathbb{P}_j$; the third equality uses the fact that $\mathbb{P}_j(N^{K+1}(1)) = \mathbb{P}_{j,a_2,\dots,a_{S_1}}(N^{K+1}(1))$ for any $a_2, \dots, a_{S_1}$, which is true because $N^{K+1}(1)$ is independent of $a_2, \dots, a_{S_1}$ conditional on j; and the last equality, again, is from the definition of conditional probability.

We have, for each $j \in [b+1]$,

$$
\begin{aligned}
&\mathbb{E}_j\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right] - \mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right] \\
\leq & N^{K+1}(1)\left\|\mathbb{P}_j\left(\cdot \mid N^{K+1}(1)\right) - \mathbb{P}_0\left(\cdot \mid N^{K+1}(1)\right)\right\|_1 \\
\leq & N^{K+1}(1) \cdot \frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\dots,a_{S_1} \in [b+1]^{S_1-1}} \left\|\mathbb{P}_{j,a_2,\dots,a_{S_1}}\left(\cdot \mid N^{K+1}(1)\right) - \mathbb{P}_{0,a_2,\dots,a_{S_1}}\left(\cdot \mid N^{K+1}(1)\right)\right\|_1 \\
\leq & N^{K+1}(1) \cdot \frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\dots,a_{S_1} \in [b+1]^{S_1-1}} \sqrt{2\,\mathrm{KL}\left(\mathrm{Ber}(\frac{1}{2}+\Delta), \mathrm{Ber}(\frac{1}{2})\right)\mathbb{E}_{0,a_2,\dots,a_{S_1}}\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]} \\
\leq & N^{K+1}(1) \cdot \frac{1}{(b+1)^{S_1-1}} \sum_{a_2,\dots,a_{S_1} \in [b+1]^{S_1-1}} \sqrt{6\Delta^2 \mathbb{E}_{0,a_2,\dots,a_{S_1}}\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]} \\
\leq & N^{K+1}(1)\sqrt{(6)\frac{l+1}{384MK} \cdot \mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]} \\
= & \frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK} \cdot \mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]}.
\end{aligned} \tag{58}
$$

where the first inequality is based on Lemma 25 and the fact that, conditional on $N^{K+1}(1)$, $n^{K+1}(1,j)$ has distribution supported on $[0, N^{K+1}(1)]$; the second inequality follows from Equation (57) and the triangle inequality; the third inequality uses Pinsker's inequality and Lemma 31 (the divergence decomposition lemma); the fourth inequality uses Lemma 33 and the fact that $\Delta \leq \frac{1}{4}$; and the last inequality follows from Jensen's inequality.

Since $N^{K+1}(1)$ has the same distribution under both $\mathbb{P}_0$ and any $\mathbb{P}_j$ (which is $\mathrm{Bin}(K, \frac{1}{S_1})$), taking expectation with respect to $N^{K+1}(1)$, we have that, for any $j \in [b+1]$,

$$
\mathbb{E}_j\left[n^{K+1}(1,j)\right] - \mathbb{E}_0\left[n^{K+1}(1,j)\right] \leq \mathbb{E}_0\left[\frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK} \cdot \mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]}\right].
$$

In subsequent derivations, we can now avoid bounding the conditional expectation. Specifically, we have

$$
\begin{aligned}
&\frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_j\left[n^{K+1}(1,j)\right] \\
\leq & \frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_0\left[n^{K+1}(1,j)\right] + \frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_0\left[\frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK} \cdot \mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]}\right] \\
\leq & \frac{1}{b+1}\mathbb{E}_0\left[\sum_{j\in[b+1]} n^{K+1}(1,j)\right] + \mathbb{E}_0\left[\frac{1}{8}N^{K+1}(1)\sqrt{\frac{l+1}{MK} \cdot \frac{1}{b+1}\sum_{j\in[b+1]}\mathbb{E}_0\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right]}\right] \\
\leq & \frac{1}{b+1}\mathbb{E}_0\left[N^{K+1}(1)\right] + \mathbb{E}_0\left[\frac{1}{8}\sqrt{\frac{l+1}{MK} \cdot \frac{1}{b+1}}\left(N^{K+1}(1)\right)^{\frac{3}{2}}\right] \\
\leq & \frac{1}{b+1}\mathbb{E}_0\left[N^{K+1}(1)\right] + \frac{1}{8}\sqrt{\frac{S_1}{MK}} \cdot \mathbb{E}_0\left[\left(N^{K+1}(1)\right)^{\frac{3}{2}}\right],
\end{aligned} \tag{59}
$$

where the first inequality follows from Eq. (58) and algebra; the second inequality uses linearity of expectation and Jensen's inequality; the third inequality uses the facts that $\sum_{j\in[b+1]} n^{K+1}(1,j) \leq N^{K+1}(1)$ and, for every $z \in [0] \cup [b+1]$,

$$
\sum_{j\in[b+1]}\mathbb{E}_z\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right] \leq \sum_{j\in\mathcal{A}}\mathbb{E}_z\left[n^{K+1}(1,j) \mid N^{K+1}(1)\right] = N^{K+1}(1);
$$

and the last inequality uses the linearity of expectation and the construction that $b = \lceil \frac{l}{S_1} \rceil$, which implies that $l \le bS_1$ and therefore $l + 1 \le bS_1 + 1 \le bS_1 + S_1 = (b+1)S_1$.

It follows from Equation (59) that

$$\frac{1}{b+1} \sum_{j \in [b+1]} \mathbb{E}_j \left[n^{K+1}(1,j) \right] \le \frac{1}{b+1} \cdot \frac{MK}{S_1} + \frac{1}{8}\sqrt{\frac{S_1}{MK}} \cdot \mathbb{E}_0 \left[\left(N^{K+1}(1) \right)^{\frac{3}{2}} \right]$$
$$\le \frac{MK}{2S_1} + \frac{1}{4}\sqrt{\frac{S_1}{MK}\left(\frac{MK}{S_1}\right)^3}$$
$$\le \frac{3MK}{4S_1},$$

where the second inequality uses the fact that $\frac{1}{b+1} \le \frac{1}{2}$ and Lemma 34 under the assumption that $K \ge S_1$.

It then follows that

$$\frac{1}{b+1} \sum_{j \in [b+1]} \mathbb{E}_j \left[N^{K+1}(1) - n^{K+1}(1,j) \right] \ge \frac{1}{b+1} \sum_{j \in [b+1]} \mathbb{E}_j \left[N^{K+1}(1) \right] - \frac{3MK}{4S_1} = \frac{MK}{4S_1},$$

and we have

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([b+1]^{S_1}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})} \left[N^{K+1}(1) - n^{K+1}(1, a_1) \right] \ge \frac{MK}{4S_1}.$$

Case 2: $Ml^C \ge l$. Again, let $S_1 = S - 2(H-1)$. Let $u = \lceil \frac{l}{S_1} \rceil$ and $v = A - u = A - \lceil \frac{l}{S_1} \rceil$. Furthermore, let $\Delta = \sqrt{\frac{vS_1}{384K}}$, and $\epsilon = 2H\Delta$. We note that under the assumption that $K \ge SA$ and the fact that $vS_1 \le SA$, we have $\Delta \le \frac{1}{4}$. We will define $v^{S_1 \times M}$ ϵ-MPERL problem instances, each indexed by an element in $[v]^{S_1 \times M}$. It suffices to show that, on at least one of the instances, $\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \ge \Omega\left(M\sqrt{H^2 l^C K}\right)$.

Facts about v. There are two helpful facts about v that can be easily verified:

- $vS_1 \ge \frac{1}{2} l^C$. This is true because, by definition, $vS_1 \ge S_1 A - l - S_1 = S_1 A - (SA - l^C) - S_1 = l^C - (SA - S_1 A) - S_1 = l^C - \left(2(H-1)A + S_1\right)$; since, by assumption, $l \le SA - 4(S + HA)$, we have $l^C \ge 4(HA+S) \ge 2\left(2(H-1)A + S_1\right)$; it then follows that $vS_1 \ge l^C - \left(2(H-1)A + S_1\right) \ge \frac{1}{2} l^C$.
- $v \ge 2$. This is true because, as shown above, $vS_1 \ge \frac{1}{2} l^C$ and $l^C \ge 4(HA+S)$, which imply that $v \ge \frac{2(HA+S)}{S_1} \ge \frac{2S_1}{S_1} = 2$.

Construction. For $\mathbf{a} = (a_{1,1}, \ldots, a_{1,M}, a_{2,1}, \ldots, a_{S_1,M}) \in [v]^{S_1 \times M}$, we define the following ϵ-MPERL problem instance, $\mathfrak{M}(\mathbf{a}) = \left\{\mathcal{M}_p\right\}_{p=1}^M$, with S states, A actions, and an episode length of H, such that for each $p \in [M]$, $\mathcal{M}_p$ is constructed in the same way as it is for case 1, except for the transition probabilities of $(s,a) \in \mathcal{S}_1 \times \mathcal{A}$:

- For each state $s \in [S_1]$,

$$\mathbb{P}_p(S_1 + 1 \mid s, a) = \begin{cases} \frac{1}{2} + \Delta, & \text{if } a = a_{s,p}; \\ \frac{1}{2}, & \text{if } a \in [v] \setminus a_{s,p}; \\ 0, & \text{if } a \notin [v]; \end{cases}$$

  and for each $a \in \mathcal{A}$, $\mathbb{P}_p(S_1 + 2 \mid s, a) = 1 - \mathbb{P}_p(S_1 + 1 \mid s, a)$, and $R_p(s,a) = 0$.

We now verify that $\mathfrak{M}(\mathbf{a})$ is an ϵ-MPMAB problem instance. It can be easily observed that the reward distributions are the same for all players, i.e., for every $p, q \in [M]$ and every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\left|R_p(s,a) - R_q(s,a)\right| = 0 \le \epsilon.$$

Regarding the transition probabilities, for every $(s,a) \in \Big((\mathcal{S}_1 \times (\mathcal{A} \setminus [v]))\Big) \cup \Big((\mathcal{S} \setminus \mathcal{S}_1) \times \mathcal{A}\Big)$, we observe that the transition probabilities are the same for all players. Furthermore, for every $p, q \in [M]$ and every $(s,a) \in \mathcal{S}_1 \times [v]$,

$$\left\|\mathbb{P}_p\left(\cdot \mid s,a\right) - \mathbb{P}_q\left(\cdot \mid s,a\right)\right\|_1 \leq 2\Delta = \frac{\epsilon}{H}.$$

Therefore, $\mathfrak{M}(\mathbf{a})$ is an ϵ-MPMAB problem instance.

Suboptimality gaps. Similar to the arguments in Case 1, it can be shown that for every $p \in [M]$, and every $(s,a) \in (\mathcal{S} \setminus \mathcal{S}_1) \times \mathcal{A}$, $\mathrm{gap}_p(s,a) = 0$. And, for every $p \in [M]$, and every $s \in \mathcal{S}_1$,

$$\mathrm{gap}_p(s,a) = \begin{cases} 0, & \text{if } a = a_{s,p}; \\ (H-1)\Delta, & \text{if } a \in [v] \setminus a_{s,p}; \\ (H-1)\left(\frac{1}{2} + \Delta\right), & \text{if } a \notin [v]. \end{cases}$$

Subpar state-action pairs. Based on the above construction, for every $(s,a) \in \mathcal{S}_1 \times (\mathcal{A} \setminus [v])$ and every $p \in [M]$, $\mathrm{gap}_p(s,a) = (H-1)\left(\frac{1}{2} + \Delta\right) \geq 3(H-1)\Delta = \frac{3(H-1)}{2H}\epsilon \geq \frac{3}{4}\epsilon \geq 96H\left(\frac{\epsilon}{192H}\right)$, where the first inequality uses the fact that $\Delta \leq \frac{1}{4}$. Therefore, there are at least $(A-v)S_1 = uS_1 \geq l$ state-action pairs in $\mathcal{I}_{\frac{\epsilon}{192H}}$, i.e.,$\left|\mathcal{I}_{\frac{\epsilon}{192H}}\right| \geq l$.

Now, it suffices to prove that

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{1}{240} M\sqrt{H^2 l^C K},$$

where we recall that $\mathbf{a} = (a_{1,1}, \dots, a_{1,M}, a_{2,1}, \dots, a_{S_1,M})$. It suffices to show, for any $s' \in [S_1]$ and any $p' \in [M]$,

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N_{p'}^{K+1}(s') - n_{p'}^{K+1}(s', a_{s'})\right] \geq \frac{K}{4S_1}, \tag{60}$$

where $N_{p'}^{K+1}(s') = \sum_{a \in \mathcal{A}} n_{p'}^{K+1}(s', a)$. To see this, by Lemma 30, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] &\geq \sum_{p=1}^{M} \sum_{s' \in \mathcal{S}_1} (H-1)\Delta \cdot \mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})}\left[N_p^{K+1}(s') - n_p^{K+1}(s', a_{s'})\right] \\
&\geq \frac{H-1}{4} MK\sqrt{\frac{vS_1}{384K}} \\
&\geq \frac{1}{160} M\sqrt{H^2 (vS_1) K} \\
&\geq \frac{1}{240} M\sqrt{H^2 l^C K},
\end{aligned}$$

where the last inequality uses the fact that $vS_1 \geq \frac{1}{2} l^C$.

Without loss of generality, let us choose $s' = 1$ and $p' = 1$. Similar to case 1, we define a set of helper problem instances: for any $(a_{1,2}, \dots, a_{S_1,M}) \in [v]^{S_1 \times M - 1}$, we define a problem instance $\mathfrak{M}(0, a_{1,2}, \dots, a_{S_1,M})$ such that it agrees with $\mathfrak{M}(a_{1,1}, a_{1,2}, \dots, a_{S_1,M})$ on everything but $\mathbb{P}_1(\cdot \mid 1, a_1)$, namely, in $\mathfrak{M}(0, a_{1,2}, \dots, a_{S_1,M})$, $\mathbb{P}_1(S_1 + 1 \mid 1, a_1) = \frac{1}{2}$.

For each $(j, a_{1,2}, \dots, a_{S_1,M}) \in \big([0] \cup [v]\big) \times [v]^{S_1 \times M - 1}$, let $\mathbb{P}_{j, a_{1,2}, \dots, a_{S_1,M}}$ denote the probability measure on the outcomes of running Alg on the problem instance $\mathfrak{M}(j, a_{1,2}, \dots, a_{S_1,M})$. Further, for each $j \in \{0\} \cup [v]$, we define

$$\mathbb{P}_j = \frac{1}{v^{S_1 \times M - 1}} \sum_{a_{1,2}, \dots, a_{S_1,M} \in [v]^{S_1 \times M - 1}} \mathbb{P}_{j, a_{1,2}, \dots, a_{S_1,M}};$$

and we use $\mathbb{E}_j$ to denote the expectation with respect to $\mathbb{P}_j$. In subsequent calculations, for any $m \in \big([0] \cup [v]\big) \times [v]^{S_1 \times M - 1}$, we also denote by $\mathbb{P}_m\left(\cdot \mid N_1^{K+1}(1)\right)$ and $\mathbb{E}_m\left[\cdot \mid N_1^{K+1}(1)\right]$ the probability and expectation conditional on a realization of $N_1^{K+1}(1)$ under $\mathbb{P}_m$. Similar to case 1, it can be shown that, for any $j \in \{0\} \cup [v]$,

$$\mathbb{P}_j(\cdot \mid N^{K+1}(1)) = \frac{1}{v^{S_1 \times M-1}} \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \mathbb{P}_{j,a_{1,2},\ldots,a_{S_1,M}} \left(\cdot \mid N^{K+1}(1)\right). \tag{61}$$

Now, for each $j \in [v]$, we have

$$\begin{aligned}
&\mathbb{E}_j \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right] - \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right] \\
\leq & N_1^{K+1}(1) \left\|\mathbb{P}_j \left(\cdot \mid N_1^{K+1}(1)\right) - \mathbb{P}_0 \left(\cdot \mid N_1^{K+1}(1)\right)\right\|_1 \\
\leq & N_1^{K+1}(1) \cdot \frac{1}{v^{S_1 \times M-1}} \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \left\|\mathbb{P}_{j,a_{1,2},\ldots,a_{S_1,M}} \left(\cdot \mid N_1^{K+1}(1)\right) - \mathbb{P}_{0,a_{1,2},\ldots,a_{S_1,M}} \left(\cdot \mid N_1^{K+1}(1)\right)\right\|_1 \\
\leq & N_1^{K+1}(1) \cdot \frac{1}{v^{S_1 \times M-1}} \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \sqrt{2\,\mathrm{KL}\left(\mathrm{Ber}(\frac{1}{2}+\Delta), \mathrm{Ber}(\frac{1}{2})\right) \mathbb{E}_{0,a_2,\ldots,a_{S_1}} \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]} \\
\leq & N_1^{K+1}(1) \cdot \frac{1}{v^{S_1 \times M-1}} \sum_{a_{1,2},\ldots,a_{S_1,M} \in [v]^{S_1 \times M-1}} \sqrt{6\Delta^2 \mathbb{E}_{0,a_2,\ldots,a_{S_1}} \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]} \\
\leq & N_1^{K+1}(1) \cdot \sqrt{\frac{6vS_1}{384MK} \cdot \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]} \\
= & \frac{1}{8} N_1^{K+1}(1) \sqrt{\frac{vS_1}{MK} \cdot \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]}.
\end{aligned} \tag{62}$$

where the first inequality is based on Lemma 25 and the fact that, conditional on $N_1^{K+1}(1)$, $n_1^{K+1}(1,j)$ has distribution supported on $[0, N_1^{K+1}(1)]$; the second inequality follows from Equation (61) and the triangle inequality; the third inequality uses Pinsker's inequality and Lemma 31 (the divergence decomposition lemma); the fourth inequality uses Lemma 33 and the fact that $\Delta \leq \frac{1}{4}$; and the last inequality follows from Jensen's inequality.

Using arguments similar to the ones shown for case 1, we have that

$$\begin{aligned}
&\frac{1}{v} \sum_{j \in [v]} \mathbb{E}_j \left[n_1^{K+1}(1,j)\right] \\
\leq & \frac{1}{v} \mathbb{E}_0 \left[n_1^{K+1}(1,j)\right] + \mathbb{E}_0 \left[\frac{1}{8} N_1^{K+1}(1) \sqrt{\frac{vS_1}{K} \cdot \frac{1}{v} \sum_{j \in [v]} \mathbb{E}_0 \left[n_1^{K+1}(1,j) \mid N_1^{K+1}(1)\right]}\right] \\
\leq & \frac{1}{v} \mathbb{E}_0 \left[N^{K+1}(1)\right] + \frac{1}{8} \sqrt{\frac{S_1}{K}} \cdot \mathbb{E}_0 \left[\left(N_1^{K+1}(1)\right)^{\frac{3}{2}}\right] \\
\leq & \frac{1}{v} \cdot \frac{K}{S_1} + \frac{1}{4} \sqrt{\frac{S_1}{K} \left(\frac{K}{S_1}\right)^3} \\
\leq & \frac{3K}{4S_1},
\end{aligned}$$

where the second to last inequality is from Lemma 34 under the assumption that $K \geq S_1$, and the last inequality uses the fact that $v \geq 2$.

It then follows that

$$\frac{1}{v} \sum_{j \in [v]} \mathbb{E}_j \left[N_1^{K+1}(1) - n_1^{K+1}(1,j)\right] \geq \frac{1}{v} \sum_{j \in [v]} \mathbb{E}_j \left[N_1^{K+1}(1)\right] - \frac{K}{4S_1} = \frac{K}{4S_1},$$

and we thereby have shown that

$$\mathbb{E}_{\mathbf{a} \sim \mathrm{Unif}\left([v]^{S_1 \times M}\right)} \mathbb{E}_{\mathfrak{M}(\mathbf{a})} \left[N_1^{K+1}(1) - n_1^{K+1}(1,a_1)\right] \geq \frac{K}{4S_1}.$$

□

F.3. Gap dependent lower bound

Theorem 36 (Restatement of Theorem 9)**.** *Fix $\epsilon \geq 0$. For any $S \in \mathbb{N}$, $A \geq 2$, $H \geq 2$, $M \in \mathbb{N}$, such that $S \geq 2(H-1)$, let $S_1 = S - 2(H-1)$; and let $\left\{\Delta_{s,a,p}\right\}_{(s,a,p)\in[S_1]\times[A]\times[M]}$ be any set of values such that*

- *for every $(s,a,p) \in [S_1] \times [A] \times [M]$, $\Delta_{s,a,p} \in [0, H/48]$;*
- *for every $(s,p) \in [S_1] \times [M]$, there exists at least one action $a \in [A]$ such that $\Delta_{s,a,p} = 0$;*
- *and, for every $(s,a) \in [S_1] \times [A]$ and $p, q \in [M]$, $\left|\Delta_{s,a,p} - \Delta_{s,a,q}\right| \leq \epsilon/4$.*

There exists an ϵ-MPERL problem instance with S states, A actions, M players and an episode length of H, such that $\mathcal{S}_1 = [S_1]$, $|\mathcal{S}_h| = 2$ for all $h \geq 2$, and

$$\mathrm{gap}_p(s,a) = \Delta_{s,a,p}, \quad \forall (s,a,p) \in [S_1] \times [A] \times [M].$$

For this problem instance, any sublinear regret algorithm Alg *for the ϵ-MPERL problem must have regret at least*

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \Omega\left(\ln K\left(\sum_{p\in[M]} \sum_{\substack{(s,a)\in\mathcal{I}^C_{(\epsilon/192H)}: \\ \mathrm{gap}_p(s,a)>0}} \frac{H^2}{\mathrm{gap}_p(s,a)} + \sum_{(s,a)\in\mathcal{I}_{(\epsilon/192H)}} \frac{H^2}{\min_p \mathrm{gap}_p(s,a)}\right)\right).$$

Proof. The construction and techniques in this proof are inspired by (Simchowitz & Jamieson, 2019) and (Wang et al., 2021).

Proof outline. We will construct an ϵ-MPERL problem instance, $\mathfrak{M}$, and show that, for any sublinear regret algorithm and sufficiently large K, the following two claims are true:

1. for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that for all p, $\mathrm{gap}_p(s,a) > 0$,

$$\mathbb{E}_{\mathfrak{M}}\left[n^K(s,a)\right] \geq \Omega\left(\frac{H^2}{\left(\min_p \mathrm{gap}_p(s,a)\right)^2} \ln K\right); \tag{63}$$

2. for any $(s,a) \in \mathcal{I}^C_{\frac{\epsilon}{192H}}$ and any $p \in [M]$ such that $\mathrm{gap}_p(s,a) > 0$,

$$\mathbb{E}_{\mathfrak{M}}\left[n^K_p(s,a)\right] \geq \Omega\left(\frac{H^2}{\left(\mathrm{gap}_p(s,a)\right)^2} \ln K\right). \tag{64}$$

The rest then follows from Lemma 30 (the regret decomposition lemma).

Construction of $\mathfrak{M}$. Given any set of values $\left\{\Delta_{s,a,p}\right\}_{(s,a,p)\in[S_1]\times[A]\times[M]}$ that satisfies the assumptions in the theorem statement, we can construct a collection of MDPs $\left\{\mathcal{M}_p\right\}_{p=1}^M$, such that for each $p \in [M]$, $\mathcal{M}_p$ is as follows, and $\mathfrak{M} = \left\{\mathcal{M}_p\right\}_{p=1}^M$ is an ϵ-MPERL problem instance:

- $\mathcal{S}_1 = [S_1]$, and p_0 is a uniform distribution over the states in $\mathcal{S}_1$.
- For $h \in [2, H]$, $\mathcal{S}_h = \{S_1 + 2h - 3, S_1 + 2h - 2\}$.

- $\mathcal{A} = [A]$.
- For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution $r_p(s, a)$ is a Bernoulli distribution, $\mathrm{Ber}(R_p(s, a))$, and we specify $R_p(s, a)$ subsequently.
- For every $(s, a) \in \mathcal{S}_1 \times [A]$, set $\bar{\Delta}^p_{s,a} = \frac{\Delta_{s,a,p}}{H-1}$. Then, let

  $$\mathbb{P}_p\left(S_1 + 1 \mid s, a\right) = \frac{1}{2} - \bar{\Delta}^p_{s,a}, \quad \mathbb{P}_p\left(S_1 + 2 \mid s, a\right) = \frac{1}{2} + \bar{\Delta}^p_{s,a},$$

  and $R_p(s, a) = 0$. Since $\Delta_{s,a,p} \in [0, H/48]$, $\bar{\Delta}^p_{s,a} \le \frac{H}{48(H-1)} \le \frac{1}{24}$, where the last inequality follows from the assumption that $H \ge 2$. Therefore, $\mathbb{P}_p\left(S_1 + 1 \mid s, a\right) \in [0, 1]$, and $\mathbb{P}_p\left(S_1 + 2 \mid s, a\right) \in [0, 1]$.
- For $h \in [2, H]$, and $a \in [A]$, let
  - $\mathbb{P}_p\left(S_1 + 2h - 1 \mid S_1 + 2h - 3, a\right) = 1$, $\mathbb{P}_p\left(S_1 + 2h \mid S_1 + 2h - 3, a\right) = 0$, and $R_p(S_1 + 2h - 3, a) = 1$.
  - $\mathbb{P}_p\left(S_1 + 2h \mid S_1 + 2h - 2, a\right) = 0$, $\mathbb{P}_p\left(S_1 + 2h - 1 \mid S_1 + 2h - 2, a\right) = 1$, and $R_p(S_1 + 2h - 2, a) = 0$.

By the assumption that for every $(s, p) \in [S_1] \times [M]$, there exists at least one action $a \in [A]$ such that $\Delta_{s,a,p} = 0$, we have that there is at least one action a such that $\bar{\Delta}^p_{s,a} = 0$. We verify that for every $(s, a, p) \in [S_1] \times [A] \times [M]$,

$$\begin{aligned} \mathrm{gap}_p(s, a) &= V^\star_p(s) - Q^\star_p(s, a) \\ &= \max_{a'} Q^\star_p(s, a') - Q^\star_p(s, a) \\ &= (H - 1)\bar{\Delta}^p_{s,a} \\ &= \Delta_{s,a,p}. \end{aligned}$$

We now verify that the above MPERL problem instance $\mathfrak{M} = \left\{\mathcal{M}_p\right\}_{p=1}^M$ is an ϵ-MPERL problem instance:

1. The reward distributions are the same for all players, namely, for all p, q,

   $$\left|R_p(s, a) - R_q(s, a)\right| = 0 \le \epsilon, \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

2. Further, by the assumption that for every $(s, a) \in [S_1] \times [A]$ and $p, q \in [M]$, $\left|\Delta_{s,a,p} - \Delta_{s,a,q}\right| \le \epsilon/4$, we have that

   $$\left|\bar{\Delta}^p_{s,a} - \bar{\Delta}^q_{s,a}\right| = \frac{\left|\Delta_{s,a,p} - \Delta_{s,a,q}\right|}{H - 1} \le \frac{\epsilon}{4(H - 1)} \le \frac{\epsilon}{2H}.$$

   It then follows that

   $$\|\mathbb{P}_p\left(\cdot \mid s, a\right) - \mathbb{P}_q\left(\cdot \mid s, a\right)\|_1 = 2\left|\bar{\Delta}^p_{s,a} - \bar{\Delta}^q_{s,a}\right| \le \frac{\epsilon}{H}.$$

   Meanwhile, for every $(s, a) \in \left(\mathcal{S} \setminus \mathcal{S}_1\right) \times \mathcal{A}$

   $$\|\mathbb{P}_p\left(\cdot \mid s, a\right) - \mathbb{P}_q\left(\cdot \mid s, a\right)\|_1 = 0 \le \frac{\epsilon}{H}.$$

   In summary, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

   $$\|\mathbb{P}_p\left(\cdot \mid s, a\right) - \mathbb{P}_q\left(\cdot \mid s, a\right)\|_1 \le \frac{\epsilon}{H}.$$

We are now ready to prove the two claims.

1. **Proving claim 1 (Equation (63)):**

   Fix any $(s_0, a_0) \in [S_1] \times [A]$ such that $\bar{\Delta}^{\min}_{s_0,a_0} = \min_p \bar{\Delta}^{p}_{s_0,a_0} > 0$. It can be easily observed that $\mathrm{gap}_p(s_0,a_0) > 0$ for all p. Define $p_0 = \mathrm{argmin}_p \bar{\Delta}^{p}_{s_0,a_0}$. We can construct a new problem instance, $\mathfrak{M}'$, which agrees with $\mathfrak{M}$, except that

   $$\forall p \in [M], \mathbb{P}_p\left(S_1+1 \mid s_0,a_0\right) = \frac{1}{2} - \bar{\Delta}^{p}_{s_0,a_0} + 2\bar{\Delta}^{\min}_{s_0,a_0}, \mathbb{P}_p\left(S_1+2 \mid s_0,a_0\right) = \frac{1}{2} + \bar{\Delta}^{p}_{s_0,a_0} - 2\bar{\Delta}^{\min}_{s_0,a_0}.$$

   $\mathfrak{M}'$ is an ϵ-MPERL problem instance. To see this, we note that the only change is in $\mathbb{P}_p\left(\cdot \mid s_0,a_0\right)$ for all $p \in [M]$. In this new instance, it is still true that for every $p, q \in [M]$,

   $$\|\mathbb{P}_p\left(\cdot \mid s_0,a_0\right) - \mathbb{P}_q\left(\cdot \mid s_0,a_0\right)\|_1 = 2\left|\bar{\Delta}^{p}_{s_0,a_0} - \bar{\Delta}^{q}_{s_0,a_0}\right| \le \frac{\epsilon}{H}.$$

   Fix any sublinear regret algorithm Alg for the ϵ-MPERL problem. By Lemma 31 (the divergence decomposition lemma), we have

   $$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) = \sum_{p=1}^{M} \mathbb{E}_{\mathfrak{M}}\left[n_p^K(s_0,a_0)\right] \mathrm{KL}\left(\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s_0,a_0), \mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s_0,a_0)\right),$$

   where $\mathbb{P}_{\mathfrak{M}}$ and $\mathbb{P}_{\mathfrak{M}'}$ are the probability measures on the outcomes of running Alg on $\mathfrak{M}$ and $\mathfrak{M}'$, respectively; $\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s_0,a_0)$, $\mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s_0,a_0)$ are the transition probabilities for (s_0,a_0) and player p in $\mathfrak{M}$ and $\mathfrak{M}'$, respectively.

   We observe that, for any $p \in [M]$,

   $$\begin{aligned}
   &\mathrm{KL}\left(\mathbb{P}_p^{\mathfrak{M}}(\cdot \mid s_0,a_0), \mathbb{P}_p^{\mathfrak{M}'}(\cdot \mid s_0,a_0)\right) \\
   =&\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2} - \bar{\Delta}^{p}_{s_0,a_0}\right), \mathrm{Ber}\left(\frac{1}{2} - \bar{\Delta}^{p}_{s_0,a_0} + 2\bar{\Delta}^{\min}_{s_0,a_0}\right)\right) \\
   \le& 12(\bar{\Delta}^{\min}_{s_0,a_0})^2,
   \end{aligned}$$

   where the last inequality follows from Lemma 33 and the assumption that $\Delta_{s,a,p} \le \frac{H}{48}$.

   In addition, $\sum_{p=1}^{M} \mathbb{E}_{\mathfrak{M}}\left[n_p^K(s_0,a_0)\right] = \mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right]$. It then follows that

   $$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) \le 12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{\min}_{s_0,a_0})^2. \tag{65}$$

   Now, in the original ϵ-MPERL problem instance, $\mathfrak{M}$, by Equation (53) and Markov's Inequality, we have

   $$\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \ge \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}}\left(n_{p_0}^K(s_0,a_0) \ge \frac{K}{4S_1}\right);$$

   where we note that $\bar{\Delta}^{p_0}_{s_0,a_0} = \bar{\Delta}^{\min}_{s_0,a_0}$. In $\mathfrak{M}'$, the new ϵ-MPERL problem instance, we have

   $$\begin{aligned}
   \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \ge& \left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{E}_{\mathfrak{M}'}\left[\sum_{a \neq a_0} n_{p_0}(s_0,a)\right] \\
   =&\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{E}_{\mathfrak{M}'}\left[N_{p_0}^K(s_0) - n_{p_0}(s_0,a_0)\right] \\
   \ge&\frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}'}\left(N_{p_0}^K(s_0) - n_{p_0}(s_0,a_0) \ge \frac{K}{4S_1}\right) \\
   \ge&\frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}'}\left(N_{p_0}^K(s_0) \ge \frac{K}{2S_1}, n_{p_0}(s_0,a_0) \le \frac{K}{4S_1}\right) \\
   \ge&\frac{K}{4S_1}\left((H-1)\bar{\Delta}^{\min}_{s_0,a_0}\right)\left(\mathbb{P}_{\mathfrak{M}'}\left(n_{p_0}(s_0,a_0) \le \frac{K}{4S_1}\right) - \exp(-\frac{K}{8S_1})\right),
   \end{aligned}$$

where the first inequality is by Equation (53); the second inequality is by Markov's Inequality; the third inequality is by simple algebra; and the last inequality is by Chernoff bound that $\mathbb{P}_{\mathfrak{M}'}\left(N_{p_0}^K(s_0) < \frac{K}{2S_1}\right) \leq \exp(-\frac{K}{8S_1})$, and $\mathbb{P}(A \cap B) \geq \mathbb{P}(B) - \mathbb{P}(A^C)$ for events A, B.

It then follows that

$$
\begin{aligned}
&\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] + \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\
=&\frac{K}{2}\left((H-1)\bar{\Delta}_{s_0,a_0}^{\min}\right)\left(\mathbb{P}_{\mathfrak{M}}\left(n_{p_0}^K(s_0,a_0) \geq \frac{K}{2}\right) + \mathbb{P}_{\mathfrak{M}'}\left(n_{p_0}^K(s_0,a_0) < \frac{K}{2}\right) - \exp(-\frac{K}{8S_1})\right) \\
\geq&\frac{K}{2}\left((H-1)\bar{\Delta}_{s_0,a_0}^{\min}\right)\left(\frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'})\right) - \exp(-\frac{K}{8S_1})\right) \\
\geq&\frac{K}{2}\left((H-1)\bar{\Delta}_{s_0,a_0}^{\min}\right)\left(\frac{1}{2}\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}_{s_0,a_0}^{\min})^2\right) - \exp(-\frac{K}{8S_1})\right),
\end{aligned}
$$

where the first inequality follows from Lemma 32 (the Bretagnolle–Huber inequality), and the second inequality follows from Eq. (65). Observe that $\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right] \leq \frac{K}{S_1}$; in addition, by our assumption that $\Delta_{s,a,p} \leq \frac{H}{48}$ for every (s,a,p), we have $\bar{\Delta}_{s_0,a_0}^{\min} \leq \frac{1}{24}$. These together implies that $\frac{1}{4}\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}_{s_0,a_0}^{\min})^2\right) \geq \exp(-\frac{K}{8S_1})$. Therefore, we have

$$
\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] + \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{K}{2}\left((H-1)\bar{\Delta}_{s_0,a_0}^{\min}\right)\cdot\frac{1}{4}\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}_{s_0,a_0}^{\min})^2\right).
$$

Now, under the assumption that Alg is a sublinear regret algorithm, we have

$$
\frac{K}{8}\left((H-1)\bar{\Delta}_{s_0,a_0}^{\min}\right)\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}_{s_0,a_0}^{\min})^2\right) \leq 2CK^{\alpha}.
$$

It follows that

$$
\begin{aligned}
\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right] &\geq \frac{1}{12\left(\bar{\Delta}_{s_0,a_0}^{\min}\right)^2}\ln\left(\frac{(H-1)\bar{\Delta}_{s_0,a_0}^{\min}K^{1-\alpha}}{16C}\right) \\
&= \frac{(H-1)^2}{12\left(\min_p \mathrm{gap}_p(s_0,a_0)\right)^2}\ln\left(\frac{\min_p \mathrm{gap}_p(s_0,a_0)K^{1-\alpha}}{16C}\right) \\
&\geq \frac{H^2}{24\left(\min_p \mathrm{gap}_p(s_0,a_0)\right)^2}\ln\left(\frac{\min_p \mathrm{gap}_p(s_0,a_0)K^{1-\alpha}}{16C}\right).
\end{aligned}
$$

We then have

$$
\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right] \geq \Omega\left(\frac{H^2}{\left(\min_p \mathrm{gap}_p(s_0,a_0)\right)^2}\ln K\right).
$$

2. **Proving Claim 2 (Equation (64)):**

Fix any $(s_0,a_0) \in \mathcal{I}_{\frac{\epsilon}{192H}}^C$ and $p_0 \in [M]$ such that $\bar{\Delta}_{(s_0,a_0)}^{p_0} > 0$, which means that $\mathrm{gap}_{p_0}(s_0,a_0) > 0$. We have that for all $p \in [M]$,

$$
\bar{\Delta}_{s_0,a_0}^p = \frac{\Delta_{s_0,a_0}^p}{H-1} = \frac{\mathrm{gap}_p(s_0,a_0)}{H-1} \leq \frac{24H(\epsilon/(192H))}{(H-1)} \leq \frac{\epsilon}{8(H-1)} \leq \frac{\epsilon}{4H}. \tag{66}
$$

We can construct a new problem instance, $\mathfrak{M}'$, which agrees with $\mathfrak{M}$ except that

$$\mathbb{P}_{p_0}\left(S_1+1 \mid s_0,a_0\right) = \frac{1}{2} - \bar{\Delta}^{p_0}_{s_0,a_0} + 2\bar{\Delta}^{p_0}_{s_0,a_0} = \frac{1}{2} + \bar{\Delta}^{p_0}_{s_0,a_0},$$
$$\mathbb{P}_{p_0}\left(S_1+2 \mid s_0,a_0\right) = \frac{1}{2} + \bar{\Delta}^{p_0}_{s_0,a_0} - 2\bar{\Delta}^{p_0}_{s_0,a_0} = \frac{1}{2} - \bar{\Delta}^{p_0}_{s_0,a_0}.$$

$\mathfrak{M}'$ is an ϵ-MPERL problem instance. To see this, we note that the only change is in $\mathbb{P}_{p_0}\left(\cdot \mid s_0,a_0\right)$. In this new instance, it is still true that for any $q \neq p_0$,

$$\|\mathbb{P}_{p_0}(\cdot \mid s_0,a_0) - \mathbb{P}_q(\cdot \mid s_0,a_0)\|_1 \leq 2\left|\bar{\Delta}^{p_0}_{s_0,a_0} + \bar{\Delta}^{q}_{s_0,a_0}\right| \leq \frac{\epsilon}{H}.$$

where the last inequality uses Equation (66) that $\bar{\Delta}^{p}_{s_0,a_0} \leq \frac{\epsilon}{4H}$ for every $p \in [M]$.

Fix any sublinear regret algorithm Alg. By Lemma 31 (the divergence decomposition lemma), we have

$$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) = \mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] \mathrm{KL}\left(\mathbb{P}^{\mathfrak{M}}_{p_0}(\cdot \mid s_0,a_0), \mathbb{P}^{\mathfrak{M}'}_{p_0}(\cdot \mid s_0,a_0)\right).$$

Using a similar reasoning as before, we can show that

$$\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'}) \leq 12\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right](\bar{\Delta}^{p_0}_{s_0,a_0})^2. \tag{67}$$

Similar to case 1, we have the following argument. In the original ϵ-MPERL problem instance, $\mathfrak{M}$, we have $\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\mathbb{P}_{\mathfrak{M}}\left(n^K_{p_0}(s_0,a_0) \geq \frac{K}{4S_1}\right)$; and in $\mathfrak{M}'$, the new ϵ-MPERL problem instance, we have $\mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq \frac{K}{4S_1}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\left(\mathbb{P}_{\mathfrak{M}'}\left(n^K_{p_0}(s_0,a_0) < \frac{K}{4S_1}\right) - \exp(-\frac{K}{8S_1})\right)$.

It then follows that

$$\begin{aligned}
&\mathbb{E}_{\mathfrak{M}}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] + \mathbb{E}_{\mathfrak{M}'}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \\
\geq& \frac{K}{2}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\left(\frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathfrak{M}}, \mathbb{P}_{\mathfrak{M}'})\right) - \exp(-\frac{K}{8S_1})\right) \\
\geq& \frac{K}{8}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K(s_0,a_0)\right](\bar{\Delta}^{p_0}_{s_0,a_0})^2\right).
\end{aligned}$$

Now, under the assumption that Alg is a sublinear regret algorithm, we have

$$\frac{K}{8}\left((H-1)\bar{\Delta}^{p_0}_{s_0,a_0}\right)\exp\left(-12\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right](\bar{\Delta}^{p_0}_{s_0,a_0})^2\right) \leq 2CK^{\alpha}.$$

It follows that

$$\begin{aligned}
\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] &\geq \frac{1}{12\left(\bar{\Delta}^{p_0}_{s_0,a_0}\right)^2}\ln\left(\frac{(H-1)\bar{\Delta}^{p_0}_{s_0,a_0}K^{1-\alpha}}{16C}\right) \\
&\geq \frac{H^2}{24\left(\mathrm{gap}_{p_0}(s_0,a_0)\right)^2}\ln\left(\frac{\mathrm{gap}_{p_0}(s_0,a_0)K^{1-\alpha}}{16C}\right).
\end{aligned}$$

We then have that

$$\mathbb{E}_{\mathfrak{M}}\left[n^K_{p_0}(s_0,a_0)\right] \geq \Omega\left(\frac{H^2}{\left(\mathrm{gap}_{p_0}(s_0,a_0)\right)^2}\ln K\right).$$

Combing the two claims: We note that in $\mathfrak{M}$, for any $(s,a,p) \in \left(\mathcal{S} \setminus \mathcal{S}_1\right) \times \mathcal{A} \times [M]$, $\mathrm{gap}_p(s,a) = 0$. It then follows from Lemma 30 (the regret decomposition lemma) and the fact that for any $(s,a,p) \in \mathcal{I}_{\epsilon/192H} \times [M]$, $\mathrm{gap}_p(s,a) > 0$, that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_{\mathrm{Alg}}(K)\right] \geq & \sum_{p=1}^{M} \sum_{(s,a)\in\mathcal{S}_1\times\mathcal{A}} \mathbb{E}\left[n_p^K(s,a)\right] \mathrm{gap}_p(s,a) \\
\geq & \Omega\left(\ln K \left(\sum_{p\in[M]} \sum_{\substack{(s,a)\in\mathcal{I}^C_{\epsilon/192H}: \\ \mathrm{gap}_p(s,a)>0}} \frac{H^2}{\mathrm{gap}_p(s,a)} + \sum_{(s,a)\in\mathcal{I}_{\epsilon/192H}} \frac{H^2}{\min_p \mathrm{gap}_p(s,a)}\right)\right).
\end{aligned}$$

□