Combinatorial Multivariant Multi-Armed Bandits
with Applications to Episodic Reinforcement Learning and Beyond

Xutong Liu [1 2]   Siwei Wang [3]   Jinhang Zuo [2 4]   Han Zhong [5]   Xuchuang Wang [2]   Zhiyong Wang [1]   Shuai Li [6]
Mohammad Hajiesmaili [2]   John C.S. Lui [1]   Wei Chen [3]

Abstract

We introduce a novel framework of combinatorial multi-armed bandits (CMAB) with multivariant and probabilistically triggering arms (CMAB-MT), where the outcome of each arm is a d-dimensional multivariant random variable and the feedback follows a general arm triggering process. Compared with existing CMAB works, CMAB-MT not only enhances the modeling power but also allows improved results by leveraging distinct statistical properties for multivariant random variables. For CMAB-MT, we propose a general 1-norm multivariant and triggering probability-modulated smoothness condition, and an optimistic CUCB-MT algorithm built upon this condition. Our framework can include many important problems as applications, such as episodic reinforcement learning (RL) and probabilistic maximum coverage for goods distribution, all of which meet the above smoothness condition and achieve matching or improved regret bounds compared to existing works. Through our new framework, we build the first connection between the episodic RL and CMAB literature, by offering a new angle to solve the episodic RL through the lens of CMAB, which may encourage more interactions between these two important directions.

[1]The Chinese University of Hong Kong, Hong Kong SAR, China [2]University of Massachusetts Amherst, Massachusetts, United States [3]Microsoft Research, Beijing, China [4]California Institute of Technology, California, United States [5]Peking University, Beijing, China [6]Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Shuai Li <shuaili8@sjtu.edu.cn>, Siwei Wang <siweiwang@microsoft.com>, Wei Chen <weic@microsoft.com>.

1. Introduction

The stochastic multi-armed bandit (MAB) (Robbins, 1952; Auer et al., 2002) is a classical model for sequential decision-making that has been widely studied(cf. Slivkins et al. (2019); Lattimore & Szepesvári (2020)). As a noteworthy extension of MAB, combinatorial multi-armed bandits (CMAB) have drawn considerable attention due to their rich applications in domains such as online advertising, network optimization, and healthcare systems (Gai et al., 2012; Kveton et al., 2015b; Chen et al., 2013b; 2016; Wang & Chen, 2017; Merlis & Mannor, 2019; Liu et al., 2021; Zuo & Joe-Wong, 2021; Zuo et al., 2022). In CMAB, the learning agent chooses a combinatorial action (often referred to as super arm) in each round. This combinatorial action would trigger a set of arms to be pulled simultaneously, and the outcomes of these arms are observed as feedback (typically known as semi-bandit feedback). The agent then receives a reward, which can be a general function of the pulled arms' outcomes, with the summation function being the most common example. The agent's goal is to minimize the expected *regret*, which quantifies the difference in expected cumulative rewards between always selecting the best action (i.e., the action with the highest expected reward) and following the agent's own policy. CMAB poses the challenge of balancing exploration and exploitation while dealing with an exponential number of combinatorial actions.

To model a wider range of application scenarios where the combinatorial action may probabilistically trigger arms, Chen et al. (2016) first introduce a generalization of CMAB, known as CMAB with probabilistically triggered arms (CMAB-T). This extension successfully encompasses a broader range of applications, including cascading bandits (Combes et al., 2015) and online influence maximization (OIM) (Wen et al., 2017). Subsequently, Wang & Chen (2017); Liu et al. (2022) improve the regret bounds of (Chen et al., 2016) by introducing novel triggering probability modulated (TPM) smoothness conditions and/or variance adaptive algorithms. Further elaboration on related works can be found in Appendix A.

Despite the expanded modeling capabilities and improved regret bounds, all prior CMAB-T frameworks assume that

each arm's outcome is a univariate random variable, upon which they base their smoothness conditions, algorithms, and analyses. In real-world applications, arm outcomes can be d-dimensional multivariate random variables with distinct statistical properties. One example is the indivisible goods distribution (Alkan et al., 1991; Chevaleyre et al., 2017), where each good can be distributed to one of d target users, forming a multivariant random variable. Another critical example is episodic reinforcement learning (RL) (Jaksch et al., 2010; Azar et al., 2017; Zanette & Brunskill, 2019; Neu & Pike-Burke, 2020), where in each episode the agent starts from an initial state s_1 and transits through traverses a series of H states $(s_h)_{h \in [H]}$ by taking action a_h upon each encountered state s_h. Each transition in this scenario is a multivariant random variable, with outcomes spanning the state space S. Existing CMAB-T approaches cannot effectively model these situations with appropriate smoothness conditions. Specifically, they resort to treating each multivariate arm as d separate arms, thereby neglecting the unique statistical characteristics of multivariate random variables and yielding suboptimal regret performance.

Our Contributions. We introduce a new CMAB-MT framework, which inherits the arm triggering mechanism of CMAB-T while accommodating d-dimensional multivariate random variables as arm outcomes. The key challenge lies in determining the contributions of each dimension of the arms to the overall regret and effectively leveraging the multivariate statistical characteristics. To address this challenge, we first introduce a novel 1-norm MTPM smoothness condition that assigns varying weights to different arms and dimensions, which flexibly covers existing 1-norm TPM smoothness conditions of CMAB-T and accommodates new applications such as episodic RL. Then, we construct an *action-dependent* confidence region that can incorporate problem-specific multivariant statistical properties. Leveraging this confidence region, we devise the CUCB-MT algorithm with a general joint oracle and establish the first regret bound for any CMAB-MT problem. Our new analysis combines regret decomposition techniques from the RL domain and sharp CMAB techniques to deal with arm triggering and regret amortization, which can yield matching or improved results for applications within and beyond this study.

To show the applicability of our framework, we first show that episodic RL fits into the CMAB-MT framework by mapping each transition kernel as an arm and employing the occupancy measure as the triggering probability. Leveraging this insight, we give two CUCB-MT algorithms that can achieve $\tilde{O}(\sqrt{H^4 S^2 AT})$ and $\tilde{O}(\sqrt{H^3 SAT})$ regret based on distinct smoothness conditions, with the latter matches the lower bound (Jaksch et al., 2010) up to logarithmic factors. Remarkably, our regret bound improves at least a factor of $O(\log^2 T)$ for the leading regret term compared with existing works (Zanette & Brunskill, 2019; Zhang et al.,

2023) owing to our sharp CMAB analysis. As a by-product, our framework gives a gap-dependent regret that scales with $O(\log T)$. Notably, episodic RL is widely known to be a strict generalization of the MAB and thus much harder to solve than MAB due to the state transition and long-term reward structure. Our work makes the first attempt to view episodic RL as an instance of CMAB, and offers a new angle for addressing episodic RL challenges through the lens of CMAB. Our results highlight that episodic RL is not significantly harder than CMAB-MT problems, and build a valuable connection between the RL and CMAB that may encourage more interactions between these pivotal research directions.

Furthermore, we explore another application beyond episodic RL that fits into our framework: the probabilistic maximum coverage for goods distribution (PMC-GD). For PMC-GD, we overcome the challenge of identifying a tight confidence region based on its unique statistical property and finding the efficient implementation of the joint oracle. To this end, our framework gives a regret bound that improves the best-known variance-adaptive algorithm (Merlis & Mannor, 2019; Liu et al., 2022) by a factor of $\tilde{O}(\sqrt{|V|/k})$, where $|V|$ and k are the numbers of target nodes and selected source nodes, and $V \gg k$ in most application scenarios (Chen et al., 2016; Liu et al., 2023b).

2. Combinatorial MAB with Multivariant and Probabilistically Triggering Arms

In this section, we present the combinatorial multi-armed bandit with multivariant and probabilistically triggering arms (or CMAB-MT for short), which generalizes the previous CMAB-T framework to handle d-dimensional multivariant arm outcomes. CMAB-MT covers not only existing instances of CMAB-T with univariant arms, but more importantly, the episodic RL as a new example.

Notations. We use **boldface** symbols for vectors $\boldsymbol{v} \in \mathbb{R}^d$. For matrix $\boldsymbol{v} \in \mathbb{R}^{m \times d}$, we treat $\boldsymbol{v}$ as a long column vector that sequentially stacks m sub-vectors of dimension d and $\boldsymbol{v}_i \in \mathbb{R}^d$ is the i-th sub-vector for $i \in [m]$. For function $V : [d] \to \mathbb{R}$, we use $\boldsymbol{V}$ to denote the vector $(V(x))_{x \in [d]} \in \mathbb{R}^d$. For any set S, we define probability simplex $\boldsymbol{\Delta}_S = \{\boldsymbol{p} \in [0, 1]^{|S|} : \sum_{i \in S} p(i) = 1\}$. We use $\boldsymbol{e}_i \in \mathbb{R}^d$ to denote the vector whose i-th entry is 1 and 0 elsewhere. For vector $\boldsymbol{v} \in \mathbb{R}^d$, we use $|\boldsymbol{v}|$ to denote the vector $(|v_i|)_{i \in [d]}$.

2.1. Framework Setup

Problem Instance. A CMAB-MT problem instance can be described by a tuple $([m], d, \Pi, \mathcal{D}, D_{\text{trig}}, R)$, where $[m] = \{1, 2, ..., m\}$ is the set of multivariant base arms; d is the dimension of multivariant base arm's random

outcome[1] (with bounded support $[0,1]^d$), i.e., outcome $\boldsymbol{X}_i = (X_{i,1}, ..., X_{i,d}) \in [0,1]^d$; Π is the set of eligible combinatorial actions and $\pi \in \Pi$ is a combinatorial action;[2] $\mathcal{D}$ is the set of possible distributions over the outcomes of base arms with support $[0,1]^{m \times d}$; D_{trig} is the probabilistic triggering function and R is the reward function, which shall be specified shortly after.

Learning Process. In CMAB-MT, the learning agent interacts with the unknown environment in a sequential manner as follows. First, the environment chooses a distribution $D \in \mathcal{D}$ unknown to the agent. Then, at round $t = 1, 2, ..., T$, the agent selects a combinatorial action $\pi_t \in \Pi$ and the environment draws from the unknown distribution D random outcome vectors $\boldsymbol{X}_t = (\boldsymbol{X}_{t,1}, ..., \boldsymbol{X}_{t,m}) \in [0,1]^{m \times d}$ for all m multivariate base arms. Note that the outcome $\boldsymbol{X}_t$ is assumed to be independent from outcomes generated in previous rounds, but outcomes $\boldsymbol{X}_{t,i}$ and $\boldsymbol{X}_{t,j}$ in the same round could be correlated. Let $D_{\text{trig}}(\pi, \boldsymbol{X})$ be a distribution over all possible subsets of $[m]$. When the action π_t is played on the outcome $\boldsymbol{X}_t$, base arms in a random set $\tau_t \sim D_{\text{trig}}(\pi_t, \boldsymbol{X}_t)$ are triggered, meaning that the multivariant outcomes of arms in τ_t, i.e., $(\boldsymbol{X}_{t,i})_{i \in \tau_t}$ are revealed as the feedback to the agent, and are involved in determining the reward of action π_t. Function D_{trig} is referred to as the *probabilistic triggering function*. At the end of the round t, the agent will receive a non-negative reward $R(\pi_t, \boldsymbol{X}_t, \tau_t)$, determined by $\pi_t, \boldsymbol{X}_t$ and τ_t.

Learning Objective. The goal of CMAB-MT is to accumulate as much reward as possible over T rounds, by learning distribution D or its parameters. Let vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_m) \in [0,1]^{m \times d}$, where $\boldsymbol{\mu}_i = (\mu_{i,1}, ..., \mu_{i,d}) \in [0,1]^d$ denote the mean vector of base arm i's multivariant outcome, i.e., $\boldsymbol{\mu}_i = \mathbb{E}_{\boldsymbol{X}_t \sim D}[\boldsymbol{X}_{t,i}]$. Similar to the CMAB-T framework (Wang & Chen, 2017; Liu et al., 2022; 2023a), we assume that the expected reward $\mathbb{E}[R(\pi, \boldsymbol{X}, \tau)]$ is a function of the unknown mean vector $\boldsymbol{\mu}$, where the expectation is taken over the randomness of $\boldsymbol{X} \sim D$ and $\tau \sim D_{\text{trig}}(\pi, \boldsymbol{X})$, and therefore we use $r(\pi; \boldsymbol{\mu}) \overset{\text{def}}{=} \mathbb{E}[R(\pi, \boldsymbol{X}, \tau)]$ to denote the expected reward. The performance of a learning algorithm ALG is measured by its *regret*, defined as the difference of the expected cumulative reward between always playing the best action $\pi^* \overset{\text{def}}{=} \operatorname{argmax}_{\pi \in \Pi} r(\pi; \boldsymbol{\mu})$ and that of playing actions chosen by the algorithm.

For many reward functions $r(\pi; \boldsymbol{\mu})$, it is NP-hard to compute the exact π^* even when $\boldsymbol{\mu}$ is known (Chen et al., 2013b; Wang & Chen, 2017; Liu et al., 2022), so we assume that one has access to an (α, β)-approximation

---

[1]For simplicity, we assume dimensions are the same, yet it is easy to generalize d to d_i for arm $i \in [m]$.

[2]When Π is a collection of subsets of $[m]$, we call action $\pi \in \Pi$ a super arm. Otherwise, we treat Π as a general action space, same as in (Wang & Chen, 2017).

oracle $\tilde{\mathcal{O}}$. $\tilde{\mathcal{O}}$ takes a confidence region function $\mathcal{C}$ that maps any action $\pi \in \Pi$ to possible parameters $\mathcal{C}(\pi) \subseteq [0,1]^{m \times d}$ as input, and outputs an action-parameter pair $(\tilde{\pi}, \tilde{\boldsymbol{\mu}}) = \tilde{\mathcal{O}}(\mathcal{C})$ such that $\tilde{\pi} \in \Pi$, $\tilde{\boldsymbol{\mu}} \in \mathcal{C}(\tilde{\pi})$ and $(\tilde{\pi}, \tilde{\boldsymbol{\mu}})$ is an α-approximation with probability at least β, i.e., $\Pr\left[r(\tilde{\pi}, \tilde{\boldsymbol{\mu}}) \geq \alpha \cdot \max_{\pi \in \Pi, \boldsymbol{\mu} \in \mathcal{C}(\pi)} r(\pi; \boldsymbol{\mu})\right] \geq \beta$. Formally, the T-round (α, β)-approximate regret is defined as

$$\text{Reg}(T; \alpha, \beta, \boldsymbol{\mu}) = T \cdot \alpha\beta \cdot r(\pi^*; \boldsymbol{\mu}) - \mathbb{E}\left[\sum_{t=1}^{T} r(\pi_t; \boldsymbol{\mu})\right], \tag{1}$$

where the expectation is taken over the randomness of the outcomes $\boldsymbol{X}_1, ..., \boldsymbol{X}_T$, the triggered sets $\tau_1, ..., \tau_T$, as well as the randomness of the algorithm ALG itself.

Remark 1 (CMAB-MT v.s. CMAB-T)**.** CMAB-MT is more general and reduces to CMAB-T when $d = 1$. Conversely, for any CMAB-MT instance, one can treat each multi-variant arm $i \in [m]$ as d separate arms $(i, j)_{j \in [d]}$ with unknown mean $\mu_{i,j}$ and use the CMAB-T model to learn these md arms, but in this way, one cannot enjoy some nice statistical property (e.g., concentration properties) by treating them as a whole. Take PMC-GD in Section 5 as an example, using CMAB-MT can improve the regret up to a factor of $O(\sqrt{d})$, owing to tighter concentration inequality around the mean vector $\boldsymbol{\mu}_i$ instead of using d separate concentration inequality around $\mu_{i,j}$.

Remark 2 (Computational complexity of joint oracle $\tilde{\mathcal{O}}$)**.** Different from the classical CMAB (α, β)-approximation oracle $\mathcal{O} : [0,1]^{m \times d} \rightarrow \Pi$ that takes a single parameter (e.g., mean vector $\boldsymbol{\mu}$) as input and optimizes over the action space Π, the joint oracle $\tilde{\mathcal{O}}$ can optimize over the joint action-parameter space $(\pi, \mathcal{C}(\pi))_{\pi \in \Pi}$. In the worst case, one has to compute the best reward $r^*(\pi) = \max_{\boldsymbol{\mu} \in \mathcal{C}(\pi)} r(\pi; \boldsymbol{\mu})$ for $\pi \in \Pi$, and then enumerate over $\pi \in \Pi$ to get the optimal $(\pi^*, \boldsymbol{\mu}^*)$ in time $O(|\Pi|)$. Nevertheless, the joint oracle has been used in many CMAB (Combes et al., 2015; Degenne & Perchet, 2016) and linear contextual bandits (Sec. 19.3.1 in Lattimore & Szepesvári (2020)) works. In this paper, we will show that all CMAB-MT applications considered, i.e., the episodic RL (Section 4) and PMC-GD (Section 5) have efficient implementations for the joint oracle.

2.2. Key Quantities and Conditions

In the CMAB-MT model, there are several quantities and conditions that are crucial to the subsequent study. First, we define *triggering probability* $q_i^{D, D_{\text{trig}}, \pi}$ as the probability that base arm i is triggered when the combinatorial action is π, the outcome distribution is D, and the probabilistic triggering function is D_{trig}. Since D_{trig} is always fixed in a given application context, and D often determines the triggering probability via its mean $\boldsymbol{\mu}$ in most cases, we use $q_i^{\boldsymbol{\mu}, \pi}$ to denote $q_i^{D, D_{\text{trig}}, \pi}$ for simplicity. Triggering probabilities

$q_i^{\boldsymbol{\mu},\pi}$'s are crucial for the triggering probability modulated bounded smoothness conditions to be defined below. Second, we define the batch-size $K \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \sum_{i \in [m]} q_i^{\boldsymbol{\mu},\pi}$ as the maximum expected number of arms that can be triggered. Note that this definition is much smaller than $K' \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \sum_{i \in [m]} \mathbb{I}\{q_i^{\boldsymbol{\mu},\pi} > 0\}$ originally defined in (Wang & Chen, 2017). For example, in episodic RL, this difference saves a factor of S, i.e., $K' = SH$ and $K = H$.

Owing to the nonlinearity and the combinatorial structure of the reward, it is essential to give some conditions for the reward function in order to achieve any meaningful regret bounds (Chen et al., 2013b; 2016; Wang & Chen, 2017; Degenne & Perchet, 2016; Merlis & Mannor, 2019). In this paper, we consider the smoothness condition as follows.

Condition 1 (1-norm multivariant and triggering probability modulated (MTPM) smoothness condition). *We say that a CMAB-MT problem satisfies 1-norm MTPM smoothness condition, if there exist weight vectors $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \in [0, \bar{w}]^d$ for $\tilde{\boldsymbol{\mu}} \in [0,1]^{m \times d}, \pi \in \Pi, i \in [m]$ such that, for any two distributions $\tilde{D}, D \in \mathcal{D}$ with mean $\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu} \in [0,1]^{m \times d}$, and for any action $\pi \in \Pi$, we have*

$$|r(\pi; \tilde{\boldsymbol{\mu}}) - r(\pi; \boldsymbol{\mu})| \le \sum_{i \in [m]} q_i^{\boldsymbol{\mu},\pi} \left| |\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \right|. \quad (2)$$

Furthermore, if for any two distributions $\tilde{D}, D \in \mathcal{D}$ with mean $\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu} \in [0,1]^{m \times d}$, and for any action $\pi \in \Pi$ we have

$$|r(\pi; \tilde{\boldsymbol{\mu}}) - r(\pi; \boldsymbol{\mu})| \le \sum_{i \in [m]} q_i^{\boldsymbol{\mu},\pi} \left| (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \right|, \quad (3)$$

then we say CMAB-MT problem satisfies 1-norm MTPM+ smoothness condition.

Remark 3 (Intuitions of Condition 1). The 1-norm MTPM smoothness condition aims to bound the reward difference caused by the parameter changing from $\boldsymbol{\mu}$ to $\tilde{\boldsymbol{\mu}}$. Intuitively, we use $|\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$ to characterize the parameter difference for each multivariant base arm i. Instead of directly using the 1-norm distance $\|\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i\|_1$, we use a refined weighted 1-norm where each dimension's difference, each dimension's difference $|\mu_{i,j} - \tilde{\mu}_{i,j}|$ is weighted by $w_{i,j}^{\tilde{\boldsymbol{\mu}},\pi}$ for $j \in [d]$. Then, each arm i's parameter difference is reweighted by the triggering probability $q_i^{\boldsymbol{\mu},\pi}$. Intuitively, for base arm i that is unlikely to be triggered/observed (small $q_i^{\boldsymbol{\mu},\pi}$), Condition 1 ensures that a large change in $\boldsymbol{\mu}_i$ only causes a small change (multiplied by $q_i^{\boldsymbol{\mu},\pi}$) in the reward, and thus one does not need to pay extra regret to observe such arms. Notice that $q_i^{\boldsymbol{\mu},\pi}$ and $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$ are related to $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$, respectively, to keep the balance of $\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}$, and the condition still holds if one exchanges the $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$ of $q_i^{\boldsymbol{\mu},\pi}$ and $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$. The intuition for 1-norm MTPM+ is similar to 1-norm MTPM, but 1-norm MTPM+ condition is stronger since any CMAB-MT instance satisfies 1-norm

MTPM+ condition with $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \in [0, \bar{w}]^d$ also satisfies 1-norm MTPM condition with the same $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$, owing to the fact that $\left| (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \right| \le |\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$. In Section 4, we will show episodic RL satisfies the stronger 1-norm MTPM+, which is the key to achieving minimax regret bound. On the other hand, the weaker 1-norm MTPM condition is easier to satisfy and can potentially cover more applications.

Remark 4 (Instances and extensions of 1-norm MTPM condition). 1-norm MTPM/MTPM+ smoothness condition reduces to B_1 bounded 1-norm TPM condition in the CMAB-T framework (Wang & Chen, 2017) when $d = 1$ and $w_i^{\tilde{\boldsymbol{\mu}},\pi} = B_1$, covering all CMAB-T problems as instances. More importantly, the 1-norm MPTM/MTPM+ smoothness condition covers the smoothness of value functions for episodic RL in Section 4.3 and Section 4.4. 1-norm MTPM/MTPM+ can also be transformed to the infinity norm or other norms, e.g., $|r(\pi; \tilde{\boldsymbol{\mu}}) - r(\pi; \boldsymbol{\mu})| \le \max_{i \in [m]} q_i^{\boldsymbol{\mu},\pi} \cdot |\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$. When $d = 1$ and $w_i^{\tilde{\boldsymbol{\mu}},\pi} = B_\infty$, ℓ_∞-norm MTPM reduces to the B_∞-bounded max-norm smoothness condition (Chen et al., 2013b).

3. Multivariant CUCB Algorithm

For CMAB-MT, we give a combinatorial upper confidence bound algorithm with a general joint oracle (CUCB-MT) in Algorithm 1. Since the CUCB-MT algorithm and its analysis are slightly different for problems satisfying the 1-norm MPTM condition in Eq. (2) and the stronger 1-norm MPTM+ condition in Eq. (3), we unify the notation and for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ use $[\boldsymbol{u} - \boldsymbol{v}]_+ \stackrel{\text{def}}{=} |\boldsymbol{u} - \boldsymbol{v}|$ for the former problems and $[\boldsymbol{u} - \boldsymbol{v}]_+ \stackrel{\text{def}}{=} (\boldsymbol{u} - \boldsymbol{v})$ for the latter problems.

CUCB-MT utilizes the principle of optimism in the face of uncertainty. In each round t, it first constructs a function $\mathcal{C}$, where $\mathcal{C}_t(\pi) \subseteq [0,1]^{m \times d}$ are action-dependent confidence region around the empirical mean $\hat{\boldsymbol{\mu}}_{t-1}$. In this work, we assume the $\mathcal{C}_t(\pi)$ for any $\pi \in \Pi$ is defined as:

$$\mathcal{C}_t(\pi) \stackrel{\text{def}}{=} \left\{ \tilde{\boldsymbol{\mu}} : \left| [\tilde{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \right| \le \phi_{t,i}, \forall i \in [m] \right\}, \quad (4)$$

where $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi}$ are weights specified by Condition 1, $\phi_{t,i}$ are confidence radius defined as $\phi_{t,i} = F_{t,i}\sqrt{\frac{1}{N_{t-1,i}}} + I_{t,i}\frac{1}{N_{t-1,i}}$, and $F_{t,i}, I_{t,i}$ are problem-specific values that will be specified for subsequent applications.

CUCB-MT then selects an optimistic action-parameter pair $(\pi_t, \tilde{\boldsymbol{\mu}}_t)$ with the help of the joint oracle $\tilde{\mathcal{O}}$. Note that the joint oracle is determined by the confidence region $\mathcal{C}_t$ and the reward function $r(\pi; \boldsymbol{\mu})$. The agent then plays the selected combinatorial action π_t, which will trigger a set of multivariant base arms τ_t whose d-dimensional outcomes

Algorithm 1 CUCB-MT: Combinatorial Upper Confidence Bound Algorithm for CMAB-MT

1: **Input:** Base arms $[m]$, dimension d, joint oracle $\tilde{\mathcal{O}}$.
2: **Initialize:** For each arm i, $N_{0,i} = 0$, $\hat{\boldsymbol{\mu}}_{0,i} = \mathbf{0}$.
3: **for** $t = 1, ..., T$ **do**
4:     Construct an action-dependent confidence region function $\mathcal{C}_t$ around $\hat{\boldsymbol{\mu}}_{t-1}$ according to Equation (4).
5:     Apply joint oracle $\tilde{\mathcal{O}}$ and get $(\pi_t, \tilde{\boldsymbol{\mu}}_t) = \tilde{\mathcal{O}}(\mathcal{C}_t)$.
6:     Play action π_t, which triggers arms $\tau_t \subseteq [m]$ with outcome $\boldsymbol{X}_{t,i}$, for $i \in \tau_t$.
7:     For $i \in \tau_t$, update $N_{t,i} = N_{t-1,i}+1$, $\hat{\boldsymbol{\mu}}_{t,i} = \hat{\boldsymbol{\mu}}_{t-1,i}+ (\boldsymbol{X}_{t,i}-\hat{\boldsymbol{\mu}}_{t-1,i})/N_{t,i}$. For $i \notin \tau_t$, keep $N_{t,i} = N_{t-1,i}$, $\hat{\boldsymbol{\mu}}_{t,i} = \hat{\boldsymbol{\mu}}_{t-1,i}$.
8: **end for**

---

are observed. Finally, CUCB-MT updates the statistics and historical information accordingly to improve future decisions. Note that though the form of $\mathcal{C}_t$ and the joint oracle are abstracted out in CUCB-MT, we will give concrete applications as examples in Section 4 and Section 5, where determining $\mathcal{C}_t$ and $\tilde{\mathcal{O}}$ serve as key ingredients of efficient algorithms and tight regret bounds.

3.1. Analysis of CUCB-MT and Its Discussion

Fix the underlying distribution $D \in \mathcal{D}$ and its mean vector $\boldsymbol{\mu} \in [0,1]^{m \times d}$ with optimal action π^*. For each action $\pi \in \Pi$, we define the (approximation) gap as $\Delta_\pi = \max\{0, \alpha r(\pi^*; \boldsymbol{\mu}) - r(\pi; \boldsymbol{\mu})\}$. For each arm $i \in [m]$, we define $\Delta_i^{\min} = \inf_{\pi \in \Pi: q_i^{\boldsymbol{\mu},\pi} > 0, \, \Delta_\pi > 0} \Delta_\pi$.

Recall that for any round t, $\hat{\boldsymbol{\mu}}_{t-1,i}$ is the empirical mean, $\mathcal{C}_t$ is the confidence region function defined in Eq. (4) with problem-specific parameters $F_{t,i}$ and $I_{t,i}$, $\tilde{\mathcal{O}}$ is the joint oracle, and $(\pi_t, \tilde{\boldsymbol{\mu}}_t) = \tilde{\mathcal{O}}(\mathcal{C}_t)$ are the pair of optimistic policy and parameter in line 5 of Algorithm 1. Define the concentration event

$$\mathcal{E}_{c,1} = \left\{ \left| [\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top \boldsymbol{w}_i^{\boldsymbol{\mu},\pi^*} \right| \leq F_{t,i}\sqrt{\frac{1}{N_{t-1,i}}} \right.$$
$$\left. + I_{t,i}\frac{1}{N_{t-1,i}}, \text{ for all } i \in [m], t \in [T] \right\}. \quad (5)$$

Let $G_{t,i}, J_{t,i}$ be another two problem-specific parameters for $i \in [m], t \in [T]$, and define the second concentration event

$$\mathcal{E}_{c,2} = \left\{ \left| [\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top (\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} - \boldsymbol{w}_i^{\boldsymbol{\mu},\pi^*}) \right| \right.$$
$$\left. \leq G_{t,i}\sqrt{\frac{1}{N_{t-1,i}}} + J_{t,i}\frac{1}{N_{t-1,i}}, \text{ for all } i \in [m], t \in [T] \right\}. \quad (6)$$

Let $\bar{F}, \bar{G}, \bar{I}, \bar{J}$ be upper bounds for problem specific parameters so that $\sum_{i \in [m]} q_i^{\boldsymbol{\mu},\pi_t} F_{t,i}^2 \leq \bar{F}$, $\sum_{i \in [m]} q_i^{\boldsymbol{\mu},\pi_t} G_{t,i}^2 \leq \bar{G}$, $I_{t,i} \leq \bar{I}$, $J_{t,i} \leq \bar{J}$, then we have the following theorem.

Theorem 1. *For a CMAB-MT problem instance* $([m], d, \Pi, \mathcal{D}, D_{trig}, R)$ *that satisfies 1-norm MTPM or MTPM+ smoothness condition (Condition 1) with weight vectors* $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}},\pi} \in [0, \bar{w}]^d$ *for* $\tilde{\boldsymbol{\mu}} \in [0,1]^{m \times d}, \pi \in \Pi, i \in [m]$, *if the oracle* $\tilde{\mathcal{O}}$ *is an* (α, β)*-approximation oracle, and concentration events* $\mathcal{E}_{c,1}, \mathcal{E}_{c,2}$ *hold with probability at least* $1 - \frac{1}{T}$, *then CUCB-MT (Algorithm 1) achieves an* (α, β)*-approximate gap-dependent regret bounded by*

$$O\left(\sum_{i \in [m]} \frac{(\bar{F} + \bar{G})}{\Delta_i^{\min}} + (\bar{I} + \bar{J})\log\left(\frac{(\bar{I} + \bar{J})K}{\Delta_i^{\min}}\right) \right), \quad (7)$$

and the gap-independent regret bounded by

$$O\left(\sqrt{m(\bar{F} + \bar{G})T} + m(\bar{I} + \bar{J})\log(KT) \right). \quad (8)$$

Remark 5. Looking at the above theorem, problem-specific parameters $F_{t,i}, I_{t,i}, G_{t,i}, J_{t,i}$ are related to concentration inequalities that hold with high probability, so they are polylogarithmic terms regarding T. For example, when arms are d-dimensional multinoulli random variables, $F_{t,i}, G_{t,i} = O(\bar{w}\sqrt{d \log T})$ and $I_{t,i} = J_{t,i} = 0$. Therefore, the leading regret is $O(\sqrt{m(\bar{F} + \bar{G})T})$. For event $\mathcal{E}_{c,1}$ in Theorem 1, we only require $\boldsymbol{\mu} \in \mathcal{C}_t(\pi^*)$ (instead of $\boldsymbol{\mu} \in \mathcal{C}_t(\pi)$ for all π), which can obtain smaller $F_{t,i}, I_{t,i}$ with tighter regret. For event $\mathcal{E}_{c,2}$ in Theorem 1, $G_{t,i}, J_{t,i}$ can be very small since $\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t}, \boldsymbol{w}_i^{\boldsymbol{\mu},\pi^*}$ can be very close to each other, e.g., if $\boldsymbol{w}_i^{\boldsymbol{\mu},\pi} = \boldsymbol{c}$ are constant vectors, then $G_{t,i} = J_{t,i} = 0$. Now looking at any CMAB-T problem following 1-norm TPM condition with B_1 mentioned in Remark 4 and arms being Bernoulli, $F_{t,i} = B_1\sqrt{1.5 \log(mT)}, G_{t,i} = I_{t,i} = J_{t,i} = 0$ and our theorem gives $\tilde{O}(B_1\sqrt{mKT})$ regret, matching the tight regret bound given by (Wang & Chen, 2017). In later sections, we will provide two representative applications that fit into the CMAB-MT framework and identify parameters $\bar{F}, \bar{G}, \bar{I}, \bar{J}$, which achieves matching or improved regret bounds compared to existing works. Due to the space limit, the detailed analysis of Theorem 1 is deferred to Appendix B.3.

4. Application to Episodic Reinforcement Learning

In this section, we first introduce the setup of episodic RL, which is modeled as a finite-horizon Markov decision process (MDP). Then we demonstrate how episodic RL can be effectively integrated into the framework of CMAB-MT and satisfy two different 1-norm MTPM smoothness conditions. For the former, we give a result matching that of the seminal work (Jaksch et al., 2010) as a warm-up case. For the latter, we achieve the minimax-optimal regret bound by leveraging a tighter confidence region function and the variance-aware analysis.

4.1. Setup of Episodic MDP and RL

We consider the finite-horizon MDP, i.e., episodic MDP, which can be described by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$. $\mathcal{S}$ is the finite state space with cardinality S. $\mathcal{A}$ is the finite action space with cardinality A. H is the number of steps for each episode. $\mathcal{P} = (\boldsymbol{p}(s, a, h))_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]}$ are transition kernels, where $\boldsymbol{p}(s, a, h) \in \boldsymbol{\Delta}_S$ and each element $p(s'|s, a, h)$ is the probability of transitioning to state s' after taking action a in state s at step h.[3] For the ease of exposition, reward distribution $\mathcal{R} = (r(s, a, h))_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]}$ are assumed to be Bernoulli random variables with mean $r(s, a, h) \in [0, 1]$, indicating the instantaneous reward collected upon taking action a in state s at step h.

In episodic RL, the agent interacts with an unknown episodic MDP environment (where $\mathcal{P}$ and $\mathcal{R}$ are unknown) in a sequence of episodes $t \in [T]$. At the beginning of episode t, the agent starts from a fixed initial state s_1 and determines a policy π_t, where $\pi_t(s, h) \in \mathcal{A}$ maps any state and step h to actions.[4] Then at the step $h = 1, ..., H$, the agent selects an action $a_{t,h} = \pi_t(s_{t,h}, h)$, receives a random reward $R_{k,h} \in [0, 1]$ with mean $r(s_{t,h}, a_{t,h}, h)$, and transits to the next state $s_{t,h+1}$ with probability $p(s_{t,h+1} \mid s_{t,h}, a_{t,h}, h)$. The trajectory $(s_{t,h}, a_{t,h})_{h \in [H]}$ and the random reward $(R_{k,h})_{h \in [H]}$ are observed as feedback to improve future policies.

Each policy π specifies a value function for every state s and step h (i.e., the expected total reward starting from state s at step h until the end of the episode), defined as $V_h^\pi(s) = \mathbb{E}[\sum_{i=h}^{H} r(s_i, a_i, i) \mid s_h = s, \pi]$, where the expectation is taken over visited state-action pairs (s_i, a_i) upon starting from state s at step h. It is easily shown that the value function satisfies the Bellman equation (with $\boldsymbol{V}_{H+1}^\pi = \boldsymbol{0}$) for any policy π:

$$V_h^\pi(s) = r(s, \pi(s, h), h) + \boldsymbol{p}(s, \pi(s, h), h)^\top \boldsymbol{V}_{h+1}^\pi. \quad (9)$$

For episodic MDP, there always exists a policy π that attains the best possible values, and we define the optimal value function $V_h^*(s) = \sup_\pi V_h^\pi(s)$. The objective of episodic RL is to minimize the regret over T episodes, which is defined as

$$\text{Reg}(T) \overset{\text{def}}{=} \sum_{t \in [T]} \left(V_1^*(s_1) - V_1^{\pi_t}(s_1) \right). \quad (10)$$

---

[3] We consider the time inhomogeneous setting where $\boldsymbol{p}(s, a, h)$'s at different steps are different.

[4] The fixed initial state can be generalized to random initial state $s_{t,1}$ by using a $H + 1$ step MDP which virtually starts from a fixed state s_0 and transits to $s_{t,1}$ with (unknown) distribution $\boldsymbol{p}_0$.

4.2. Episodic RL from the Lens of CMAB-MT

Similar to existing works, we assume transition kernels $\mathcal{P}$ are unknown while the reward distribution $\mathcal{R}$ is known.[5] For this episodic RL problem, it fits into CMAB-MT framework with tuple $([m], d, \Pi, \mathcal{D}, D_{\text{trig}}, R)$. Each transition kernel $\boldsymbol{p}(s, a, h) \in \boldsymbol{\Delta}_S$ corresponds to a base arm and there are $m = SAH$ of them. The outcome of base arm $\boldsymbol{X}_{s,a,h} \in \{0, 1\}^S$ is a multinoulli (or categorical) random variable with dimension $d = S$, i.e., a one-hot vector $\boldsymbol{X}_{s,a,h} = \boldsymbol{e}_{s'}$ indicating the state at next step $h + 1$ will be s' upon taking action a at step h. The set of feasible combinatorial actions π corresponds to the set of deterministic policies π that maps state-step pairs to actions, i.e., $\pi : \mathcal{S} \times [H] \to \mathcal{A}$. As mentioned before, we assume the reward distribution $\mathcal{R}$ is known, so the set of $\mathcal{D}$ corresponds to any feasible MDP with reward distribution $\mathcal{R}$.

Before the RL game starts, the environment draws an unknown distribution $D \in \mathcal{D}$ with transition probabilities $\boldsymbol{p} = (\boldsymbol{p}(s, a, h))_{s,a,h \in \mathcal{S} \times \mathcal{A} \times [H]}$, where $\boldsymbol{p}(s, a, h) \in \boldsymbol{\Delta}_S$. At each episode $t \in [T]$, let the outcomes of base arms be $\boldsymbol{X}_t = (\boldsymbol{X}_{t,s,a,h})_{s,a,h} \sim D$. Given the policy π_t and the starting state s_1, the triggering set $\tau_t = (s_{t,h}, a_{t,h}, h)_{h \in [H]}$ includes a cascade of H base arms starting from the state-action-step tuple $(s_1, \pi_t(s_1, 1), 1)$, and the h-th arm for $h > 1$ of this cascade is tuple $(s_{t,h}, a_{t,h}, h) = (s', \pi_t(s', h), h)$, where $s' \in \mathcal{S}$ is the index such that s'-th entry of $\boldsymbol{X}_{t,s_{t,h-1},a_{t,h-1},h-1}$ equals to 1. In this case, the triggering probability distribution $D_{\text{trig}}(\pi_t, \boldsymbol{X}_t)$ is fully determined by π_t and $\boldsymbol{X}_t$, i.e., τ_t is deterministically decided given π_t and $\boldsymbol{X}_t$. And it is easy to show that the reward function $R(\pi_t, \boldsymbol{X}_t, \tau_t) = \sum_{(s,a,h) \in \tau_t} r(s, a, h)$ and the expected reward function $r(\pi_t; \boldsymbol{p}) = \mathbb{E}[R(\pi_t, \boldsymbol{X}_t, \tau_t)] = V_1^{\pi_t}(s_1)$.

Key Quantities. The triggering probability is the occupancy measure of (s, a, h), i.e., $q_{s,a,h}^{\boldsymbol{p},\pi} = \mathbb{E}[\mathbb{I}\{s_h = s, a_h = a \mid \pi, \boldsymbol{p}\}]$, indicating the probability of visiting state-action pair (s, a) at step h when the underlying transition is $\boldsymbol{p}$ and the policy is π. And the batch-size is $K = \max_\pi \sum_{s,a,h} q_{s,a,h}^{\boldsymbol{p},\pi} = H$.

4.3. The Simple Smoothness Condition with Constant Weights Achieves Sublinear Regret

Fitting episodic RL into CMAB-MT, we can show that it satisfies the following lemma, whose proof is in Appendix C.1.

Lemma 1. *Episodic RL with unknown transition is a CMAB-MT instance, which satisfies 1-norm MTPM smoothness condition (Condition 1) with weights $\boldsymbol{w}_{s,a,h}^{\tilde{\boldsymbol{p}},\pi} = H \cdot \boldsymbol{1} \in \mathbb{R}^S$ for all $\tilde{\boldsymbol{p}}, \pi$, i.e.,* $\left| V_1^{\tilde{\boldsymbol{p}},\pi}(s_1) - V_1^{\boldsymbol{p},\pi}(s_1) \right| \le H \cdot \sum_{s,a,h} q_{s,a,h}^{\boldsymbol{p},\pi}.$

---

[5] Handling unknown reward distribution $\mathcal{R}$ is straight-forward by adding SAH arms with $d = 1$ for the SAH unknown rewards.

Algorithm 2 Extended Value Iteration Oracle in Episode t

1: **Input:** Counter $N_{t-1}(s,a,h)$, empirical transition $\hat{p}_{t-1}(s,a,h)$ for all s,a,h, and $\delta' = 1/(2T)$.
2: **Initialize:** $\phi'_t(s,a,h) = \sqrt{\frac{2S\log(SAHT/\delta')}{N_{t-1}(s,a,h)}}$, $\bar{V}_{t,H+1}(s) = 0$ for all s,a,h.
3: **for** Step $h = H, H-1, ..., 1$ **do**
4:     For all (s,a), set $\tilde{p}_t(s,a,h) = \arg\max_{p' \in \Delta_S: \|p' - \hat{p}_{t-1}(s,a,h)\|_1 \le \phi'_t(s,a,h)} {p'}^{\top} \bar{V}_{t,h+1}$
5:     For all (s,a), set $Q_t(s,a,h) = r(s,a,h) + \tilde{p}_t(s,a,h)^{\top} \bar{V}_{t,h+1}$.
6:     For all s, set $\pi_t(s,h) = \arg\max_a Q_t(s,a,h)$ and $\bar{V}_{t,h}(s) = Q_t(s, \pi_t(s,h), h)$
7: **end for**
8: **Return:** $\pi_t, \tilde{p}_t$.

$\|\tilde{p}(s,a,h) - p(s,a,h)\|_1$.

Confidence Region Function $\mathcal{C}_t$ and Joint Oracle $\tilde{\mathcal{O}}$. By definition, we have counter $N_t(s,a,h) = \sum_{t'=1}^{t} \mathbb{I}\{(s,a,h) \in \tau_{t'}\}$ and the empirical mean $\hat{p}_t(s,a,h) = \frac{\sum_{t'=1}^{t} \mathbb{I}\{(s,a,h) \in \tau_{t'}\} X_{t',s,a,h}}{N_t(s,a,h)}$. Based on the fact that outcomes $X_{t,s,a,h}$ are multinoulli random variables, we use the concentration for multinoulli distributions (Lemma 7), i.e., $\|p(s,a,h) - \hat{p}_{t-1}(s,a,h)\|_1 \le \sqrt{\frac{2S\log(2/\delta)}{N_{t-1}(s,a,h)}}$ with probability at least $1 - \delta$. The confidence region function defined as Eq. (4) becomes

$$\mathcal{C}_t(\pi) = \{\tilde{p} : \text{for all } (s,a,h), \tilde{p}(s,a,h) \in \Delta_S,$$
$$H \cdot \|\tilde{p}(s,a,h) - \hat{p}_{t-1}(s,a,h)\|_1 \le \phi_t(s,a,h)\}, \quad (11)$$

where $\phi_t(s,a,h) = F_{t,s,a,h}\sqrt{\frac{1}{N_{t-1}(s,a,h)}} + \frac{I_{t,s,a,h}}{N_{t-1}(s,a,h)}$, and $F_{t,s,a,h} = H\sqrt{2S\log(SAHT/\delta')}$, $I_{t,s,a,h} = 0$.

Since this region is not policy-dependent, we use $\mathcal{C}_t$ as a shortcut of $\mathcal{C}_t(\pi)$ for all $\pi \in \Pi$. Based on the confidence region $\mathcal{C}_t$, we identify the joint oracle as $(\pi_t, \tilde{p}_t) = \arg\max_{\pi \in \Pi, \tilde{p} \in \mathcal{C}_t} V_1^{\tilde{p},\pi}(s_1)$. According to Jaksch et al. (2010), this joint oracle can be implemented efficiently using extended value iteration described in Algorithm 2. Note that in line 4 in Algorithm 2, we need to solve a linear optimization problem over a convex polytope, which can be solved in $O(S^2 A)$.

Regret Bound and Discussion. Based on the above confidence region function and the joint oracle, we have

Theorem 2. *For episodic RL fitting into the CMAB-MT framework with weights in Lemma 1, CUCB-MT algorithm with the confidence region function $\mathcal{C}_t$ in Eq. (11) and the joint oracle in Algorithm 2 satisfies the requirements of Theorem 1 with parameters $\bar{F} = \tilde{O}(H^3 S), \bar{G} = \bar{I} = \bar{J} = 0$, and thus achieves a regret bounded by $\tilde{O}(\sqrt{H^4 S^2 AT})$.*

Looking at the above theorem, we achieve a regret bound

Algorithm 3 Optimistic Value Iteration Oracle in Episode t

1: **Input:** Counter $N_{t-1}(s,a,h)$, empirical transition $\hat{p}_{t-1}(s,a,h)$ for all s,a,h, and $\delta' = \frac{SAHT}{8T}$.
2: **Initialize:** Constant $L = \log\left(\frac{SAHT}{\delta'}\right)$, value function $\bar{V}_{t,H+1}(s) = \underline{V}_{t,H+1}(s) = 0$, for all s.
3: **for** $h = H, H-1, ..., 1$ **do**
4:     For all (s,a), set confidence radius
$$\phi_t(s,a,h) = 2\sqrt{\frac{\text{Var}_{s' \sim \hat{p}_{t-1}(s,a,h)}(\bar{V}_{t,h+1}(s'))L}{N_{t-1}(s,a,h)}} +$$
$$2\sqrt{\frac{\mathbb{E}_{s' \sim \hat{p}_{t-1}(s,a,h)}[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')]^2 L}{N_{t-1}(s,a,h)}} +$$
$$\frac{5HL}{N_{t-1}(s,a,h)}.$$
5:     Set $s^* = \arg\max_s \bar{V}_{t,h+1}(s)$.
6:     **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
7:       **if** $\bar{V}_{t,h+1}(s^*) < \hat{p}_{t-1}(s,a,h)^{\top} \bar{V}_{t,h+1} + \phi_t(s,a,h)$ **then**
8:         Set $\tilde{p}_t(s,a,h) = e_{s^*}$
9:       **else**
10:         Pick any $\tilde{p}_t(s,a,h) \in \Delta_S$ s.t. $\tilde{p}_t(s,a,h)^{\top} \bar{V}_{t,h+1} = \hat{p}_{t-1}(s,a,h)^{\top} \bar{V}_{t,h+1} + \phi_t(s,a,h)$.
11:       **end if**
12:       $Q_t(s,a,h) = r(s,a,h) + \tilde{p}_t(s,a,h)^{\top} \bar{V}_{t,h+1}$.
13:     **end for**
14:     For all s, set $\pi_t(s,h) = \arg\max_a Q_t(s,a,h)$.
15:     For all s, set $\bar{V}_{t,h}(s) = Q_t(s, \pi_t(s,h), h)$.
16:     For all s, set $\underline{V}_{t,h}(s) = \max\{r(s, \pi_t(s,h), h) + \hat{p}_{t-1}(s, \pi_t(s,h), h)^{\top} \underline{V}_{t,h+1} - \phi_t(s,a,h), 0\}$.
17: **end for**
18: **Return:** $\pi_t, \tilde{p}_t$.

matching the seminal work Jaksch et al. (2010), and up to a factor of $\tilde{O}(\sqrt{HS})$ compared with lower bound given by Jaksch et al. (2010), see Appendix C.2 for detailed analysis.

4.4. The Value Function Related Smoothness Condition Achieves Optimal Regret

A natural question to ask is whether we can achieve the minimax optimal regret using the CMAB-MT framework. The answer is affirmative by leveraging the RL structures for stronger 1-norm MTPM+ smoothness condition, tighter confidence region $\mathcal{C}_t$, and variance-aware analysis. We start with a stronger smoothness condition. Compared with Lemma 1, we use the future value function $V_{h+1}^{\tilde{\mu},\pi}$ instead of the constant $H \cdot \mathbf{1}$ as the weight $w_{s,a,h}^{\tilde{\mu},\pi}$, whose proof is in Appendix C.1.

Lemma 2. *Episodic RL with unknown transition is a CMAB-MT instance, which satisfies 1-norm MTPM+ smoothness (Condition 1) with weight vector $w_{s,a,h}^{\tilde{p},\pi} = V_{h+1}^{\tilde{p},\pi}$ for all $\tilde{p}, \pi$, i.e., $\left|V_1^{\tilde{p},\pi}(s_1) - V_1^{p,\pi}(s_1)\right| \le \sum_{s,a,h} q_{s,a,h}^{p,\pi} \cdot \left|[\tilde{p}(s,a,h) - p(s,a,h)]^{\top} V_{h+1}^{\tilde{p},\pi}\right|$.*

Confidence Region Function $\mathcal{C}_t$ and Joint Oracle

$\tilde{\mathcal{O}}$. Based on the above lemma, we use the following confidence region that bounds the expected future value $\tilde{\boldsymbol{p}}(s,a,h)^\top \boldsymbol{V}_{h+1}^{\tilde{\boldsymbol{p}},\pi}$ around the empirical future value $\hat{\boldsymbol{p}}(s,a,h)^\top \boldsymbol{V}_{h+1}^{\tilde{\boldsymbol{p}},\pi}$, i.e.,

$$\mathcal{C}_t(\pi) = \{\tilde{\boldsymbol{p}}: \text{ for all } (s,a,h), \tilde{\boldsymbol{p}}(s,a,h) \in \boldsymbol{\Delta}_S,$$
$$\left| [\tilde{\boldsymbol{p}}(s,a,h) - \hat{\boldsymbol{p}}_t(s,a,h)]^\top \boldsymbol{V}_{h+1}^{\tilde{\boldsymbol{p}},\pi} \right| \le \phi_t(s,a,h)\}. \quad (12)$$

where $\phi_t(s,a,h)$ is the confidence radius to be determined later on. According to Dann et al. (2017) (Lemma D.1 in particular), we can show that the exact joint oracle over $\mathcal{C}_t$, i.e., $(\pi_t, \tilde{\boldsymbol{p}}_t) = \operatorname{argmax}_{\pi,\tilde{\boldsymbol{p}} \in \mathcal{C}_t(\pi)} V_1^{\tilde{\boldsymbol{p}},\pi}(s_1)$, is optimistic value iteration with bonus $\phi_t(s,a,h)$ described in Algorithm 3. For the value of $\phi_t(s,a,h)$, we only need $\boldsymbol{p} \in \mathcal{C}_t(\pi^*)$ as specified in Theorem 1, so one possibility is to set $\phi_t(s,a,h) = \tilde{O}(\sqrt{\operatorname{Var}_{s' \sim \boldsymbol{p}(s,a,h)} (V_{h+1}^*(s'))/N_{t-1}(s,a,h)})$ according to the concentration of optimal future value (Lemma 8), saving a factor of $O(\sqrt{S})$ compared to the $\phi_t(s,a,h)$ in Section 4.3. However, since both $\boldsymbol{p}$ and $\boldsymbol{V}_{h+1}^*$ are unknown and inspired by Zanette & Brunskill (2019), we use the concentration of Lemma 11 and set $\phi_t(s,a,h)$ using optimistic $\bar{V}_{t,h}$ and pessimistic $\underline{V}_{t,h}$ as in line 4 in Algorithm 3. Mapping back to the form of Eq. (4), we have $F_{t,s,a,h} = 2\sqrt{\operatorname{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)} (\bar{V}_{t,h+1}(s')) L} + \sqrt{\mathbb{E}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)} [\bar{V}_{t,h+1}(s') - V_{h+1}^*(s')]^2 L}, I_{t,s,a,h} = 5HL$.

Regret Bound and Discussion. Based on the above tighter confidence region function and the joint oracle, we have

Theorem 3. *For episodic RL fitting into the CMAB-MT framework with weight in Lemma 2, CUCB-MT algorithm (Algorithm 1) with the confidence region function $\mathcal{C}_t$ in Eq. (12) and the joint oracle in Algorithm 3 achieves a regret bounded by $O(\sqrt{H^3 SAT \log(SAHT)} + H^3 S^2 A \log^{3/2}(SAHT))$ according to the analysis procedure of Theorem 1.*

Looking at the above regret bound, we obtain a minimax optimal worst-case regret matching the lower bound $\Omega(\sqrt{H^3 SAT})$ up to logarithmic factors. Our regret also saves at least a $O(\log^2(SAHT))$ factor for the leading $O\left(\sqrt{H^3 SAT \log(SAHT)}\right)$ term compared with $O\left(\sqrt{H^3 SAT \log^5(SAHT)}\right)$ (Zanette & Brunskill, 2019; Zhang et al., 2023) due to our tight analysis that uses sharp CMAB proof techniques, and see Section 4.4 for details. As a by-product, we give a gap-dependent bound that scales with $O(\log T)$. In the worst case, our result is at most a factor of $O(1/q^*)$ larger than Simchowitz & Jamieson (2019) that uses involved clipping techniques. However, when considering gap-independent bound, ours still improves theirs

by a factor of $O(\sqrt{\log(SAHT)})$, see Appendix D.5 for details.

5. Applications Beyond Episodic RL

In this section, we first consider the probabilistic maximum coverage problem for goods distribution (PMC-GD), which is a new variant of the PMC problem (Chen et al., 2013b; Merlis & Mannor, 2019; Liu et al., 2022). For PMC-GD, we show that CMAB-MT framework can give an improved regret bound compared with using the CMAB-T framework.

Application Setup. The PMC-GD problem is modeled by a weighted bipartite graph $G = (U, V, E, p)$, where U are the nodes to be selected, V are the nodes to be covered, and E are the edges between U and V. Each edge (u, v) in E is associated with a probability $p(u, v)$. The probability $p(u, v)$ indicates the likelihood that node u from U can cover a target node v in V. In the classical PMC problem, each selected node u' can independently cover v, and edges (u', v) are independent Bernoulli random variables with mean $p(u', v)$. In goods distribution applications (Alkan et al., 1991; Chevaleyre et al., 2017), the good (e.g., food, medicine, product, coupon) given to nodes in U is indivisible and u can only randomly distribute it to exactly one of the target users in V. Thus for PMC-GD, each selected node u' will cover one of its neighbors in V, and the edges $((u', v))_{v:(u',v) \in E}$ form a multinoulli distribution with $\sum_{v:(u',v) \in E} p(u', v) \le 1$. The coverage means the target node $v \in V$ receives such indivisible goods. The objective of the decision maker is to select at most k nodes in U to maximize the number of covered nodes in V.

For the online PMC-GD problem, we consider T rounds of repeated PMC-GD where the edge probabilities $p(u, v)$'s are unknown initially. Without loss of generality, we assume G is a complete bipartite graph. For each round $t \in [T]$, the agent selects k nodes in U as combinatorial action π_t, the feedback are k node pairs (u, v), where v receives indivisible goods from $u \in \pi_t$.

Fitting into CMAB-MT Framework. The PMC-GD problem fits into CMAB-MT framework as follows: the nodes U are the set of multivariant base arms, the unknown outcome distribution $D \in \mathcal{D}$ is the joint of $m = |U|$ multinoulli distribution with dimension $d = |V|$, the vectors $\boldsymbol{\mu}_i = \boldsymbol{p}(i, \cdot) \in \boldsymbol{\Delta}_V$ are unknown mean vectors for $i \in U$, the set of combinatorial action Π are any set of nodes $\pi \subseteq U$ with size $|\pi| \le k$. For the arm triggering in round t, the triggering set is $\tau_t = \pi_t$. Let $\boldsymbol{X}_t = \{0,1\}^{|U| \times |V|}$ be the random outcome where $X_{t,u,v} = 1$ if and only if user u sends the good to user v at time step t. The total reward is $R(\pi_t, \boldsymbol{X}_t, \tau_t) = \sum_{v \in V} \mathbb{I}\{\exists u \in \pi_t \text{ s.t. } X_{t,u,v} = 1\}$, and the expected reward $r(\pi_t; \boldsymbol{p}) = \sum_{v \in V} \left(1 - \prod_{u \in \pi_t} (1 - p(u, v))\right)$.

Key Quantities and Conditions. For the triggering proba-

bility, $q_i^{p, \pi_t} = 1$ if $i \in \pi_t$ and $q_i^{p, \pi_t} = 0$ otherwise. And the batch-size $K = k$.

Lemma 3. *PMC-GD is a CMAB-MT instance, which satisfies 1-norm MTPM smoothness (**Condition 1**) with weights $\boldsymbol{w}_u^{\tilde{p},\pi} = \mathbf{1}$, i.e., $|r(\pi; \tilde{p}) - r(\pi; p)| \leq \sum_{u \in \pi} \|\tilde{p}(u, \cdot) - p(u, \cdot)\|_1$.*

Confidence Region Function $\mathcal{C}_t$ and Joint Oracle $\tilde{\mathcal{O}}$. Since each base arm u's outcome follows from multinoulli distribution, indicating $\|p(u, \cdot) - \hat{p}_{t-1}(u, \cdot)\|_1 \leq \sqrt{\frac{2|V| \log(2/\delta)}{N_{t-1,u}}}$ with probability at least $1 - \delta$. The confidence region defined by Eq. (4) becomes

$$\mathcal{C}_t = \{\tilde{p} : \text{for any } u \in U, p(u, \cdot) \in \boldsymbol{\Delta}_V,$$
$$\|p(u, \cdot) - \hat{p}_{t-1}(u, \cdot)\|_1 \leq \phi_{t,u}\}, \quad (13)$$

and does not depend on the action π, where $\phi_{t,u} = F_{t,i}\sqrt{\frac{1}{N_{t-1,u}}} + \frac{I_{t,i}}{N_{t-1,u}}$ and $F_{t,i} = \sqrt{2|V| \log(|U||V|T/\delta')}, I_{t,i} = 0$.

The joint oracle is $(\pi_t, \tilde{p}_t) = \text{argmax}_{|\pi| \leq k, \tilde{p}_t \in \mathcal{C}_t} r(\pi; \tilde{p})$. A new challenge arises since the above joint oracle is a hard optimization problem. In particular, $\arg\max_{|\pi| \leq k} r(\pi; \tilde{p})$ itself is NP-hard given $\tilde{p}$, and now we also have to jointly optimize $\tilde{p}$ within the confidence region. To obtain an efficient oracle, our strategy is to bypass $r(\pi; \tilde{p})$ and optimize an upper bound of $r(\pi; \tilde{p})$ using Lemma 3 for all $\tilde{p}$ as the pseudo reward function for PMC-GD:

$$\bar{r}_t(\pi; \tilde{p}) = r(\pi; \hat{p}_{t-1}) + \sum_{u \in \pi} \|\tilde{p}(u, \cdot) - \hat{p}_{t-1}(u, \cdot)\|_1$$
$$(14)$$

We now optimize $(\pi_t, \tilde{p}_t) = \text{argmax}_{|\pi| \leq k, \tilde{p} \in \mathcal{C}_t} \bar{r}_t(\pi; \tilde{p})$, which is solved in Algorithm 4. Based on Eq. (13), first, we can find optimal $\tilde{p}_t$ in line 3 of Algorithm 4. Then in line 5, we can optimize $\bar{r}_t(\pi; \tilde{p}_t) = r(\pi; \hat{p}_{t-1}) + \sum_{u \in \pi} q_u$ efficiently using a greedy algorithm with $O(k|U|)$ calls to $\bar{r}_t(\pi; \tilde{p}_t)$, yielding a $(1 - 1/e, 1)$-approximation since $\bar{r}_t(\pi; \tilde{p}_t)$ is a submodular function regarding $\pi \subseteq U$. Since we use pseudo reward $\bar{r}_t(\pi; p)$, mapping back the true reward $r(\pi; p)$ will have an additional $\sum_{t \in [T]} \bar{r}_t(\pi_t; p) - r(\pi_t; p)$ term for the final regret, see Appendix E for details.

Regret Bound and Discussion. Based on the above argument, we have the following theorem.

Theorem 4. *For PMC-GD equipped with pseudo-reward in Eq. (14), CUCB-MT algorithm (Algorithm 1) with the confidence region function $\mathcal{C}_t$ in Eq. (13) and the joint oracle in Algorithm 4 satisfies the requirements of Theorem 1 with parameters $\bar{F} = \tilde{O}(k|V|)$, $\bar{G} = \bar{I} = \bar{J} = 0$, and thus achieves a $(1 - 1/e, 1)$-approximate regret bounded by $\tilde{O}(\sqrt{k|U||V|T})$.*

Algorithm 4 Efficient Joint Oracle for PMC-GD in round t

1: **Input:** Counter $N_{t-1,u}$, empirical edge probability $\hat{p}_{t-1}(u, \cdot)$ for all $u \in U$, and $\delta' = 1/(2T)$.
2: **Initialize:** $\phi_{t,u} = \sqrt{\frac{2|V| \log(|U||V|T/\delta')}{N_{t-1,u}}}$ for all $u \in U$.
3: For all $u \in U$, compute $\tilde{p}_t(u, \cdot) = \text{argmax}_{p \in \boldsymbol{\Delta}_V : \|p - \hat{p}_{t-1}(u, \cdot)\|_1 \leq \phi_{t,u}} \|p - \hat{p}_{t-1}(u, \cdot)\|_1$.
4: For all $u \in U$, set $q_u = \|\tilde{p}_t(u, \cdot) - \hat{p}_{t-1}(u, \cdot)\|_1$.
5: $\pi_t = \text{argmax}_{\pi \in \Pi} r(\pi; \hat{p}_{t-1}) + \sum_{u \in \pi} q_u$.
6: **Return:** $\pi_t, \tilde{p}_t$.

Compared with existing works, our regret improves upon CUCB-T algorithm (Wang & Chen, 2017) with regret $\tilde{O}(\sqrt{k|U||V|^2 T})$ by a factor of $\tilde{O}(\sqrt{V})$, and the recent variance-adaptive algorithms (Merlis & Mannor, 2019; Liu et al., 2022) with regret $\tilde{O}(\sqrt{|U||V|^2 T})$ by a factor of $\tilde{O}(\sqrt{|V|/k})$ when $|V| \geq k$, where in most application scenarios $|V| \gg k$ (Chen et al., 2016; Liu et al., 2023b).

6. Conclusion and Future Directions

In this work, we propose a new combinatorial multi-armed bandit framework with multivariant and probabilistically triggering arms (CMAB-MT). Through our framework, we build the first connection between episodic RL and CMAB literature, achieving matching or improved results for episodic RL and beyond. For future work, it will be interesting to study the CMAB-MT framework when considering the linear or nonlinear function approximation. One can also explore new application scenarios that can fit into the CMAB-MT framework for improved results.

Acknowledgements

Impact Statement

This paper presents a theoretical study on multi-armed bandits and reinforcement learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Alkan, A., Demange, G., and Gale, D. Fair allocation of indivisible goods and criteria of justice. *Econometrica: Journal of the Econometric Society*, pp. 1023–1039, 1991.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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.

Chen, W., Lakshmanan, L. V., and Castillo, C. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013a.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159. PMLR, 2013b.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

Chevaleyre, Y., Endriss, U., and Maudet, N. Distributed fair allocation of indivisible goods. *Artificial Intelligence*, 242:1–22, 2017.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 1–12, 2021.

Degenne, R. and Perchet, V. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, pp. 2972–2980, 2016.

Demirel, I. and Tekin, C. Combinatorial gaussian process bandits with probabilistically triggered arms. In *International Conference on Artificial Intelligence and Statistics*, pp. 3844–3852. PMLR, 2021.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5): 1466–1478, 2012.

Huyuk, A. and Tekin, C. Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1322–1330. PMLR, 2019.

Hwang, T., Chai, K., and Oh, M.-h. Combinatorial neural bandits. In *International Conference on Machine Learning*, pp. 14203–14236. PMLR, 2023.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 06–11 Aug 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pp. 767–776. PMLR, 2015a.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. Combinatorial cascading bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1450–1458, 2015b.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*, 2015c.

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

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776, 2021.

Li, S., Wang, B., Zhang, S., and Chen, W. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pp. 1245–1253. PMLR, 2016.

Liu, X., Zuo, J., Chen, X., Chen, W., and Lui, J. C. Multi-layered network exploration via random walks: From offline optimization to online learning. In *International Conference on Machine Learning*, pp. 7057–7066. PMLR, 2021.

Liu, X., Zuo, J., Wang, S., Joe-Wong, C., Lui, J., and Chen, W. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. In *Advances in Neural Information Processing Systems*, 2022.

Liu, X., Zuo, J., Wang, S., Lui, J. C., Hajiesmaili, M., Wierman, A., and Chen, W. Contextual combinatorial bandits with probabilistically triggered arms. In *International Conference on Machine Learning*, pp. 22559–22593. PMLR, 2023a.

Liu, X., Zuo, J., Xie, H., Joe-Wong, C., and Lui, J. C. Variance-adaptive algorithm for probabilistic maximum coverage bandits with general feedback. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2023b.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, 36, 2024.

Ménard, P., Domingues, O. D., Shang, X., and Valko, M. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pp. 7609–7618. PMLR, 2021.

Merlis, N. and Mannor, S. Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. In *Conference on Learning Theory*, pp. 2465–2489. PMLR, 2019.

Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Nie, G., Nadew, Y. Y., Zhu, Y., Aggarwal, V., and Quinn, C. J. A framework for adapting offline algorithms to solve combinatorial multi-armed bandit problems with bandit feedback. In *International Conference on Machine Learning*, pp. 26166–26198. PMLR, 2023.

Nika, A., Elahi, S., and Tekin, C. Contextual combinatorial volatile multi-armed bandit with adaptive discretization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1486–1496. PMLR, 2020.

Perrault, P. When combinatorial thompson sampling meets approximation regret. *Advances in Neural Information Processing Systems*, 35:17639–17651, 2022.

Qin, L., Chen, S., and Zhu, X. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469. SIAM, 2014.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Saha, A. and Gopalan, A. Combinatorial bandits with relative feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Tsuchiya, T., Ito, S., and Honda, J. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 8117–8144. PMLR, 2023.

Wan, Z., Zhang, J., Chen, W., Sun, X., and Zhang, Z. Bandit multi-linear dr-submodular maximization and its applications on adversarial submodular bandits. In *International Conference on Machine Learning*, pp. 35491–35524. PMLR, 2023.

Wang, Q. and Chen, W. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pp. 1161–1171, 2017.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122, 2018.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Wen, Z., Kveton, B., Valko, M., and Vaswani, S. Online influence maximization under independent cascade model with semi-bandit feedback. *Advances in neural information processing systems*, 30, 2017.

Wu, T., Yang, Y., Zhong, H., Wang, L., Du, S., and Jiao, J. Nearly optimal policy optimization with stable at any time guarantee. In *International Conference on Machine Learning*, pp. 24243–24265. PMLR, 2022.

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, Z., Ji, X., and Du, S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Zhang, Z., Chen, Y., Lee, J. D., and Du, S. S. Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*, 2023.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

Zimmert, J., Luo, H., and Wei, C.-Y. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pp. 7683–7692. PMLR, 2019.

Zuo, J. and Joe-Wong, C. Combinatorial multi-armed bandits for resource allocation. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–4. IEEE, 2021.

Zuo, J., Liu, X., Joe-Wong, C., Lui, J. C., and Chen, W. Online competitive influence maximization. In *International Conference on Artificial Intelligence and Statistics*, pp. 11472–11502. PMLR, 2022.

Appendix

A. Extended Related Works

In this section, we review two lines of literature that are related to this work.

Stochastic Combiantorial Multi-Armed Bandits. There has been a vast literature on stochastic combinatorial multi-armed bandit (CMAB) (Gai et al., 2012; Kveton et al., 2015c; Combes et al., 2015; Chen et al., 2016; Wang & Chen, 2017; Merlis & Mannor, 2019; Saha & Gopalan, 2019; Liu et al., 2022). Gai et al. (2012) is the first work to consider the stochastic CMAB with semi-bandit feedback. Since then, its algorithm and regret have been improved by Kveton et al. (2015c); Combes et al. (2015); Chen et al. (2016); Merlis & Mannor (2019) in different settings. To model a broader range of applications, such as online learning to rank (Kveton et al., 2015a;b) and online influence maximization (Chen et al., 2013a; Wen et al., 2017), Chen et al. (2016) first generalizes the CMAB to CMAB with probabilistically triggered armed (CMAB-T). Later on, Wang & Chen (2017) improve the regret bound of Chen et al. (2016) by introducing a new smoothness condition called the triggering probability modulated (TPM) condition, which removes a factor of $1/q^*$ compared to Chen et al. (2016), where q^* is the minimum positive probability that any arm can be triggered. Recently, Liu et al. (2022) introduce a new variance-modulated TPM condition (TPVM) and variance-adaptive algorithms that can further remove a factor of K, where K is the number of arms that can be triggered in each round. Beyond these works, Qin et al. (2014); Li et al. (2016); Nika et al. (2020); Demirel & Tekin (2021); Liu et al. (2023a); Hwang et al. (2023) study contextual environments with linear/nonlinear base arm structures, Zimmert et al. (2019); Tsuchiya et al. (2023); Nie et al. (2023); Wan et al. (2023) consider adversarial environments, and Wang & Chen (2018); Huyuk & Tekin (2019); Perrault (2022) investigate Thompson sampling algorithms for both CMAB and CMAB-T settings. However, all the above works assume the outcome of each arm is a uni-variant sub-Gaussian random variable. In this work, we consider a different setting where arms' outcomes are multivariant random variables, and propose a new CMAB-MT framework that can cover new applications, e.g. episodic RL, and give matching/improved regrets by leveraging the statistical properties of the multivariant random variables.

Episodic Reinforcement Learning. In recent years, there has been an emerging number of works that study provably efficient RL for regret minimization (c.f. (Agarwal et al., 2019)). For episodic RL, the seminal work (Jaksch et al., 2010) proposes the UCRL2 algorithm that adds optimistic bonuses on transition probabilities and achieves a regret bound of $\tilde{O}(\sqrt{H^4 S^2 AT})$, matching the lower bound $\Omega(\sqrt{H^3 SAT})$ given by the same work up to a factor of $\tilde{O}(\sqrt{HS})$. Later on, Azar et al. (2017) build confidence region directly for value functions rather than transition probabilities and provide a minimax-optimal regret of $\tilde{O}(\sqrt{H^3 SAT})$. Their result is then improved by Zanette & Brunskill (2019) who proposes an algorithm based on both optimistic and pessimistic values for the bonus design and achieves tighter problem-dependent regret bounds. After this, various works (Li et al., 2021; Zhang et al., 2021; Ménard et al., 2021; Wu et al., 2022; Zhang et al., 2023) refine the lower-order terms of regret. In addition, many studies (Jiang et al., 2017; Sun et al., 2019; Jin et al., 2020; 2021; Du et al., 2021; Zhong et al., 2022; Liu et al., 2024; Foster et al., 2021) extend beyond tabular RL and explore function approximation, although their regret bounds become suboptimal when applied to the tabular setting. The above works all focus on giving gap-independent regret bound that scales with $\tilde{O}(\sqrt{T})$. There are also other works (Simchowitz & Jamieson, 2019; Dann et al., 2021) that focus on studying gap-dependent regret bound that scales with $O(\log T)$ via clipping techniques. To the best of our knowledge, we are the first to solve the episodic RL problem by modeling it as a CMAB-MT instance. From this perspective, we propose new algorithms and analysis that achieves the minimax-optimal leading regret with improved logarithmic factors for the leading regret term. Our approach also gives gap-dependent regret bounds "for free" without using the involved clipping techniques, which matches the Simchowitz & Jamieson (2019) up to a factor of $1/q^*$ in the worst case, where q^* is the minimum positive occupancy measure of any state-action pair for any policy. To this end, we build an important connection between the RL and CMAB literature, which may encourage more interactions between these two important directions.

B. Analysis for CMAB-MT Framework

B.1. Definitions

Definition 1 ((Approximation) Gap). *Fix a distribution $D \in \mathcal{D}$ and its mean vector $\boldsymbol{\mu} \in [0,1]^{m \times d}$, for each action $\pi \in \Pi$, we define the (approximation) gap as $\Delta_\pi = \max\{0, \alpha r(\pi^*; \boldsymbol{\mu}) - r(\pi; \boldsymbol{\mu})\}$. For each arm $i \in [m]$, we define $\Delta_i^{\min} = \inf_{\pi \in \Pi: q_i^{\boldsymbol{\mu}, \pi} > 0, \Delta_\pi > 0} \Delta_\pi$, $\Delta_i^{\max} = \sup_{\pi \in \Pi: q_i^{\boldsymbol{\mu}, \pi} > 0, \Delta_\pi > 0} \Delta_\pi$. As a convention, if there is no action $\pi \in \Pi$ such that $q_i^{\boldsymbol{\mu}, \pi} > 0$ and $\Delta_\pi > 0$, then $\Delta_i^{\min} = +\infty, \Delta_i^{\max} = 0$. We define $\Delta_{\min} = \min_{i \in [m]} \Delta_i^{\min}$ and $\Delta_{\max} = \max_{i \in [m]} \Delta_i^{\max}$.*

Definition 2 (Event-Filtered Regret). *For any series of events $(\mathcal{E}_t)_{t \in [T]}$ indexed by round number t, we define the*

$Reg_{\alpha,\boldsymbol{\mu}}^{ALG}(T,(\mathcal{E}_t)_{t\in[T]})$ *as the regret filtered by events* $(\mathcal{E}_t)_{t\in[T]}$, *or the regret is only counted in* t *if* $\mathcal{E}$ *happens in* t. *Formally,*

$$Reg_{\alpha,\boldsymbol{\mu}}^{ALG}(T,(\mathcal{E}_t)_{t\in[T]}) \overset{def}{=} \mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}(\mathcal{E}_t)(\alpha\cdot r(\pi^*;\boldsymbol{\mu})-r(\pi_t;\boldsymbol{\mu}))\right]. \tag{15}$$

For simplicity, we will omit $ALG, \alpha, \boldsymbol{\mu}, t\in[T]$ *and rewrite* $Reg_{\alpha,\boldsymbol{\mu}}^{ALG}(T,(\mathcal{E}_t)_{t\in[T]})$ *as* $Reg(T,\mathcal{E}_t)$ *when contexts are clear.*

B.2. Bounds for event-filtered regrets

Lemma 4 (Decomposition of the filtered regret). *Let* $K\in\mathbb{N}_+$. *For all* $t\geq 1$, *consider the event*

$$\mathcal{E}_t = \left\{\Delta_{\pi_t} \leq \sum_{k\in[T]}R_{t,k}\right\} \tag{16}$$

and K *decomposed events*

$$\mathcal{E}_{t,k}' = \{\Delta_{\pi_t}\leq KR_{t,k}\} \tag{17}$$

for some $R_{t,k}\geq 0$. *Then, we have*

$$Reg(T,\mathcal{E}_t) \leq \sum_{k\in[T]}Reg(T,\mathcal{E}_{t,k}') \tag{18}$$

.

Proof. It suffices to prove that $\mathbb{I}\{\mathcal{E}_t\}\Delta_{\pi_t} \leq \sum_{k\in[T]}\mathbb{I}\{\mathcal{E}_{t,k}'\}\Delta_{\pi_t}$ for each round t. If $\mathcal{E}_t$ does not hold, we are done. If $\mathcal{E}_t$ holds, there exists $k'\in[T]$ such that $\Delta_{\pi_t}\leq KR_{t,k'}$, so at least one of the K decomposed events holds and $1\leq\sum_{k\in[T]}\mathbb{I}\{\Delta_{\pi_t}\leq KR_{t,k}\}$, which gives $\mathbb{I}\{\mathcal{E}_t\}\Delta_{\pi_t}\leq\sum_{k\in[T]}\mathbb{I}\{\Delta_{\pi_t}\leq KR_{t,k}\}\Delta_{\pi_t}$. ∎

Lemma 5 (Null counters). *For all* $i\in[m]$, *if there exists constants* $K_i\in\mathbb{R}^+$, *consider the event*

$$\mathcal{E}_t = \left\{\Delta_{\pi_t}\leq\sum_{i\in[m]:N_{t-1,i}=0}q_i^{\boldsymbol{\mu},\pi_t}K_i\right\}. \tag{19}$$

Then, the event filtered regret $Reg(T,\mathcal{E}_t)\leq\sum_{i\in[m]}K_i$.

Proof. Let $\mathcal{F}_{t-1}=((\pi_1,\tau_1,(\boldsymbol{X}_{t,i})_{i\in\tau_1}),...,(\pi_{t-1},\tau_{t-1},(\boldsymbol{X}_{t,i})_{i\in\tau_{t-1}}),\pi_t)$ be all historical information before t plus the action at t, where τ_t is the triggered arm set at round t. By the definition of the triggering probability $q_i^{\boldsymbol{\mu},\pi_t}$, we have

$$Reg(T,\mathcal{E}_t) = \mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\{\mathcal{E}_t\}\Delta_{\pi_t}\right] \tag{20}$$

$$\leq \mathbb{E}\left[\sum_{t\in[T]}\mathbb{E}\left[\sum_{i\in[m]}K_i\mathbb{I}\{N_{t-1,i}=0,i\in\tau_t\}\mid\mathcal{F}_{t-1}\right]\right] \tag{21}$$

$$\leq \sum_{i\in[m]}K_i. \tag{22}$$

The last inequality holds since the counter $N_{t-1,i}$ will be added by one if $i\in\tau_t$, indicating the event can only occur for at most one round, giving the upper bound $\sum_{i\in[m]}K_i$. ∎

B.3. Proof of Theorem 1

By Condition 1, we have the weight vector is bounded by $0 \leq w_{i,j}^{\tilde{\boldsymbol{\mu}},\pi} \leq \bar{w}$ for all $i \in [m], j \in [d], \tilde{\boldsymbol{\mu}} \in [0,1]^{m \times d}, \pi \in \Pi$. Define the event $\mathcal{E}_{s,t}$ as the event joint oracle successfully yields an α-approximation in round $t \in [T]$, i.e., $\mathcal{E}_{s,t} = \left\{ r(\pi_t, \tilde{\boldsymbol{\mu}}_t) \geq \alpha \cdot \max_{\pi \in \Pi, \boldsymbol{\mu} \in \mathcal{C}(\pi)} r(\pi; \boldsymbol{\mu}) \right\}$. Recall that $\mathcal{F}_{t-1} = ((\pi_1, \tau_1, (\boldsymbol{X}_{t,i})_{i \in \tau_1}), ..., (\pi_{t-1}, \tau_{t-1}, (\boldsymbol{X}_{t,i})_{i \in \tau_{t-1}}), \pi_t)$ is all historical information before t plus the action at t.

We bound the regret under the event $(\mathcal{E}_{s,t})_{t \in [T]}$, the concentration event $\mathcal{E}_{c,1}$ in Eq. (5) and the concentration event $\mathcal{E}_{c,2}$ in Eq. (6) as follows.

Step 1: Regret Decomposition.

First, recall that $[\boldsymbol{u} - \boldsymbol{v}]_+ \overset{\text{def}}{=} |\boldsymbol{u} - \boldsymbol{v}|$ for any problem that satisfies the 1-norm MPTM smoothness condition and $[\boldsymbol{u} - \boldsymbol{v}]_+ \overset{\text{def}}{=} (\boldsymbol{u} - \boldsymbol{v})$ for problem that satisfies the 1-norm MPTM+ smoothness condition. We can decompose the round t instantaneous regret as follows.

$$\Delta_{\pi_t} = \alpha \cdot r(\pi^*; \boldsymbol{\mu}) - r(\pi_t; \boldsymbol{\mu}) \tag{23}$$

$$\overset{(a)}{\leq} r(\pi_t; \tilde{\boldsymbol{\mu}}_t) - r(\pi_t; \boldsymbol{\mu}) \tag{24}$$

$$\overset{(b)}{=} \sum_{i \in [m]} q_i^{\boldsymbol{\mu},\pi_t} \left| [\tilde{\boldsymbol{\mu}}_{t,i} - \boldsymbol{\mu}_i]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} \right| \tag{25}$$

$$\leq \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} \left| [\tilde{\boldsymbol{\mu}}_{t,i} - \boldsymbol{\mu}_i]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} \right| + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{26}$$

$$= \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} \left| [\tilde{\boldsymbol{\mu}}_{t,i} - \hat{\boldsymbol{\mu}}_{t-1,i} + \hat{\boldsymbol{\mu}}_{t-1,i} - \boldsymbol{\mu}_i]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} \right| + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{27}$$

$$\overset{(c1)}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} \left| [\tilde{\boldsymbol{\mu}}_{t,i} - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} \right| + q_i^{\boldsymbol{\mu},\pi_t} \left| [\hat{\boldsymbol{\mu}}_{t-1,i} - \boldsymbol{\mu}_i]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} \right| + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{28}$$

$$\overset{(c2)}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} \left| [\tilde{\boldsymbol{\mu}}_{t,i} - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top \boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} \right| + q_i^{\boldsymbol{\mu},\pi_t} \left| [\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top \boldsymbol{w}_i^{\boldsymbol{\mu},\pi^*} \right|$$
$$+ q_i^{\boldsymbol{\mu},\pi_t} \left| [\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t-1,i}]_+^\top (\boldsymbol{w}_i^{\tilde{\boldsymbol{\mu}}_t,\pi_t} - \boldsymbol{w}_i^{\boldsymbol{\mu},\pi^*}) \right| + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{29}$$

$$\overset{(d)}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} (2F_{t,i} + G_{t,i}) \sqrt{\frac{1}{N_{t-1,i}}} + \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} (2I_{t,i} + J_{t,i}) \frac{1}{N_{t-1,i}} + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{30}$$

$$\overset{(e)}{\leq} 2 \left(\sqrt{\sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} F_{t,i}^2} + \sqrt{\sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu},\pi_t} G_{t,i}^2} \right) \sqrt{\sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu},\pi_t}}{N_{t-1,i}}}$$
$$+ 2(\bar{I} + \bar{J}) \sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu},\pi_t}}{N_{t-1,i}} + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{31}$$

$$\overset{(f)}{\leq} 2 \left(\sqrt{\bar{F}} + \sqrt{\bar{G}} \right) \sqrt{\sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu},\pi_t}}{N_{t-1,i}}} + 2(\bar{I} + \bar{J}) \sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu},\pi_t}}{N_{t-1,i}} + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu},\pi_t} \bar{w} d \tag{32}$$

where inequality (a) is due to $\boldsymbol{\mu} \in \mathcal{C}_t(\pi^*)$ and $r(\pi_t, \tilde{\boldsymbol{\mu}}_t) \geq \alpha \cdot \max_{\pi \in \Pi, \boldsymbol{\mu} \in \mathcal{C}(\pi)} r(\pi; \boldsymbol{\mu})$ under $\mathcal{E}_s$, inequality (b) is due to the 1-norm MTPM/MTPM+ smoothness condition (Condition 1), inequality (c1) and (c2) are by triangle inequality for both $[\boldsymbol{u} - \boldsymbol{v}]_+ = |\boldsymbol{u} - \boldsymbol{v}|$ and $[\boldsymbol{u} - \boldsymbol{v}]_+ = (\boldsymbol{u} - \boldsymbol{v})$ cases, inequality (d) is due to the confidence region Eq. (4), the event $\mathcal{E}_{c,1}$ in Eq. (5) and $\mathcal{E}_{c,2}$ in Eq. (6), inequality (e) is by Cauchy-Schwarz inequality and the definition of $\bar{I}, \bar{J}$, and inequality (f) is due to the definition of $\bar{F}, \bar{G}$.

Let $c_1 = 3 \times 2(\sqrt{\bar{F}} + \sqrt{\bar{G}})$, $c_2 = 3 \times 2(\bar{I} + \bar{J})$, $c_3 = 3 \times \bar{w} d$. Now we define the main event $\mathcal{E}_t$ and its three decomposed

events $\mathcal{E}'_{t,1}, \mathcal{E}'_{t,2}, \mathcal{E}'_{t,3}$ as follows:

$$\mathcal{E}_t = \left\{ \Delta_{\pi_t} \leq 2(\sqrt{F} + \sqrt{G}) \sqrt{\sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu}, \pi_t}}{N_{t-1,i}}} + 2(\bar{I} + \bar{J}) \sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu}, \pi_t}}{N_{t-1,i}} + \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu}, \pi_t} \bar{w} d \right\}, \tag{33}$$

$$\mathcal{E}'_{t,1} = \left\{ \Delta_{\pi_t} \leq c_1 \sqrt{\sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu}, \pi_t}}{N_{t-1,i}}} \right\}, \mathcal{E}'_{t,2} = \left\{ \Delta_{\pi_t} \leq c_2 \sum_{i \in [m]: N_{t-1,i} > 0} \frac{q_i^{\boldsymbol{\mu}, \pi_t}}{N_{t-1,i}} \right\}, \tag{34}$$

$$\mathcal{E}'_{t,3} = \left\{ \Delta_{\pi_t} \leq \sum_{i \in [m]: N_{t-1,i} = 0} q_i^{\boldsymbol{\mu}, \pi_t} c_3 \right\}. \tag{35}$$

By Lemma 4, we have

$$\text{Reg}(T, \mathcal{E}_t) \leq \sum_{i \in [3]} \text{Reg}(T, \mathcal{E}'_{t,i}) \tag{36}$$

Step 2: Bound the Reg$(T, \mathcal{E}'_{t,1})$ term

Let $\mathbb{E}_t = [\cdot \mid \mathcal{F}_{t-1}]$. Suppose $\mathcal{E}'_{t,1}$ holds, we use the reverse amortization trick as follows:

$$\Delta_{\pi_t} \overset{(a)}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} \frac{c_1^2 q_i^{\boldsymbol{\mu}, \pi_t} \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} \tag{37}$$

$$\overset{(b)}{\leq} -\Delta_{\pi_t} + 2 \sum_{i \in [m]: N_{t-1,i} > 0} \frac{c_1^2 q_i^{\boldsymbol{\mu}, \pi_t} \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} \tag{38}$$

$$\leq -\frac{\sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t} \Delta_{\pi_t}}{\sum_{i \in [m]} q_i^{\boldsymbol{\mu}, \pi_t}} + 2 \sum_{i \in [m]: N_{t-1,i} > 0} \frac{c_1^2 q_i^{\boldsymbol{\mu}, \pi_t} \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} \tag{39}$$

$$\overset{(c)}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t} \left(\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K} \right) \tag{40}$$

where inequality (a) follows from event $\mathcal{E}'_{t,1}$, inequality (b) is due to the reverse amortization trick that multiplies two to both sides of inequality (a) and rearranges the terms, and inequality (c) is due to definition of $K \overset{\text{def}}{=} \max_{\pi \in \Pi} \sum_{i \in [m]} q_i^{\boldsymbol{\mu}, \pi}$.

Then we use the triggering probability equivalence trick (TPE) in (Liu et al., 2023a) to deal with the triggering probability $q_i^{\boldsymbol{\mu}, \pi_t}$ as follows:

$$\mathbb{E}_t[\Delta_{\pi_t}] \overset{(a)}{\leq} \mathbb{E}_t \left[\sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t} \left(\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K} \right) \right] \tag{41}$$

$$\overset{(b)}{=} \mathbb{E}_t \left[\sum_{i \in \tau_t: N_{t-1,i} > 0} \left(\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K} \right) \right] \tag{42}$$

where inequality (a) follows from Equation (40), inequality (b) follows from TPE trick to replace $q_i^{\boldsymbol{\mu}, \pi_t} = \mathbb{E}_t[\mathbb{I}\{i \in \tau_t\}]$.

Now we claim that

$$\mathbb{E}_t[\Delta_{\pi_t}] \overset{(a)}{\leq} \mathbb{E}_t \left[\sum_{i \in \tau_t: N_{t-1,i} > 0} \kappa_i(N_{t-1,i}) \right], \tag{43}$$

16

where we define $L_{i,1} = \frac{c_1^2}{(\Delta_i^{\min})^2}$, $L_{i,2} = \frac{2c_1^2 K}{(\Delta_i^{\min})^2}$, and

$$
\kappa_i(\ell) = \begin{cases} 2\sqrt{\frac{c_1^2}{\ell}}, & \text{if } 1 \le \ell \le L_{i,1}, \\ \frac{2c_1^2}{\Delta_i^{\min}} \frac{1}{\ell}, & \text{if } L_{i,1} < \ell \le L_{i,2}, \\ 0, & \text{if } \ell > L_{i,2}, \end{cases}
\tag{44}
$$

We now show inequality (a) because of the following argument.

Case 1: If there exists $i' \in \tau_t$ with $1 \le N_{t-1,i'} \le \frac{c_1^2}{\Delta_{\pi_t}^2}$.

We have $N_{t-1,i'} \le \frac{c_1^2}{\Delta_{\pi_t}^2} \le L_{i',1}$, thus $\sum_{i \in \tau_t : N_{t-1,i} > 0} \kappa_i(N_{t-1,i}) \ge \kappa_{i'}(N_{t-1,i'}) \ge 2\sqrt{\frac{c_1^2}{N_{t-1,i'}}} = 2\Delta_{\pi_t}$, then inequality (a) holds.

Case 2: For any arm $i \in \tau_t$ with $N_{t-1,i} > 0$, they satisfy $N_{t-1,i} \ge \frac{c_1^2}{\Delta_{\pi_t}^2}$.

If $N_{t-1,i} \le L_{i,1}$, then $\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K} \le \frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} = 2\sqrt{\frac{c_1^2}{\Delta_{\pi_t}^2 \cdot N_{t-1,i}}}\sqrt{\frac{c_1^2}{N_{t-1,i}}} \le 2\sqrt{\frac{c_1^2}{N_{t-1,i}}} = \kappa_i(N_{t-1,i})$; Else if $L_{i,1} < N_{t-1,i} \le L_{i,2}$, then $\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K} \le \frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} \le \frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_i^{\min}} = \kappa_i(N_{t-1,i})$; Else if $N_{t-1,i} > L_{i,2}$, $\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K} \le 0 = \kappa_i(N_{t-1,i})$. Therefore, we have

$$
\mathbb{E}_t[\Delta_{\pi_t}] \le \mathbb{E}_t\left[\sum_{i \in \tau_t : N_{t-1,i} > 0}\left(\frac{2c_1^2 \frac{1}{N_{t-1,i}}}{\Delta_{\pi_t}} - \frac{\Delta_{\pi_t}}{K}\right)\right] \le \mathbb{E}\left[\sum_{i \in \tau_t : N_{t-1,i} > 0} \kappa_i(N_{t-1,i})\right].
\tag{45}
$$

Combining the above two cases proves inequality (a).

Now we have

$$
\text{Reg}(T, \mathcal{E}'_{t,1}) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\pi_t}\right]
\tag{46}
$$

$$
\overset{(a)}{\le} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t\left[\sum_{i \in \tau_t : N_{t-1,i} > 0} \kappa_i(N_{t-1,i})\right]\right]
\tag{47}
$$

$$
\overset{(b)}{=} \mathbb{E}\left[\sum_{t \in [T]} \sum_{i \in \tau_t : N_{t-1,i} > 0} \kappa_i(N_{t-1,i})\right]
\tag{48}
$$

$$
\overset{(c)}{=} \mathbb{E}\left[\sum_{i \in [m]} \sum_{s=1}^{N_{T-1,i}} \kappa_i(s)\right]
\tag{49}
$$

$$
\le \sum_{i \in [m]} \sum_{s=1}^{L_{i,1}} 2\sqrt{\frac{c_1^2}{s}} + \sum_{i \in [m]} \sum_{s=L_{i,1}+1}^{L_{i,2}} \frac{2c_1^2}{\Delta_i^{\min}} \frac{1}{s}
\tag{50}
$$

$$
\le \sum_{i \in [m]} \int_{s=0}^{L_{i,1}} 2\sqrt{\frac{c_1^2}{s}} \cdot ds + \sum_{i \in [m]} \int_{s=L_{i,1}}^{L_{i,2}} \frac{2c_1^2}{\Delta_i^{\min}} \frac{1}{s} \cdot ds
\tag{51}
$$

$$
\le \sum_{i \in [m]} \frac{2c_1^2}{\Delta_i^{\min}} (3 + \log K),
\tag{52}
$$

where (a) follows from Equation (43), (b) follows from the tower rule, (c) follows from that $N_{t-1,i}$ is increased by 1 if and only if $i \in \tau_t$.

Step 3: Bound the $\text{Reg}(T, \mathcal{E}'_{t,2})$ term

Let $\mathbb{E}_t = [\cdot \mid \mathcal{F}_{t-1}]$. Suppose $\mathcal{E}'_{t,2}$ holds, we use the reverse amortization trick as follows:

$$
\Delta_{\pi_t} \overset{(a)}{\leq} \sum_{i \in [m]: N_{t-1,i} > 0} c_2 q_i^{\boldsymbol{\mu}, \pi_t} \frac{1}{N_{t-1,i}}
$$

$$
\overset{(b)}{\leq} -\Delta_{\pi_t} + 2 \sum_{i \in [m]: N_{t-1,i} > 0} c_2 q_i^{\boldsymbol{\mu}, \pi_t} \frac{1}{N_{t-1,i}}
$$

$$
= -\frac{\sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t} \Delta_{\pi_t}}{\sum_{i \in [m]: N_{t-1,i} > 0} q_i^{\boldsymbol{\mu}, \pi_t}} + 2 \sum_{i \in [m]: N_{t-1,i} > 0} c_2 q_i^{\boldsymbol{\mu}, \pi_t} \frac{1}{N_{t-1,i}} \tag{53}
$$

$$
\overset{(c)}{\leq} \sum_{\substack{i \in [m]: N_{t-1,i} > 0}} q_i^{\boldsymbol{\mu}, \pi_t} \left(-\frac{\Delta_{\pi_t}}{K} + 2c_2 \frac{1}{N_{t-1,i}} \right), \tag{54}
$$

where inequality (a) follows from event $E_{t,2}$, inequality (b) is due to the reverse amortization trick that multiplies two to both sides of inequality (a) and rearranges the terms, inequality (c) is due to definition of $K \overset{\text{def}}{=} \max_{\pi \in \Pi} \sum_{i \in [m]} q_i^{\boldsymbol{\mu}, \pi}$.

It follows that

$$
\mathbb{E}_t[\Delta_{\pi_t}] \overset{(a)}{\leq} \mathbb{E}_t \left[\sum_{i \in [m]} q_i^{\boldsymbol{\mu}, \pi_t} \left(-\frac{\Delta_{\pi_t}}{K} + 2c_2 \frac{1}{N_{t-1,i}} \right) \right]
$$

$$
\overset{(b)}{=} \mathbb{E}_t \left[\sum_{i \in \tau_t: N_{t-1,i} > 0} \left(-\frac{\Delta_{\pi_t}}{K} + 2c_2 \frac{1}{N_{t-1,i}} \right) \right]
$$

$$
\overset{(c)}{\leq} \mathbb{E}_t \left[\sum_{i \in \tau_t: N_{t-1,i} > 0} \kappa_i(N_{t-1,i}) \right] \tag{55}
$$

where the following regret allocation function follows from

$$
\kappa_i(\ell) = \begin{cases} \frac{2c_2}{\ell}, & \text{if } 1 \leq \ell \leq L_i \\ 0, & \text{if } \ell > L_i + 1, \end{cases} \tag{56}
$$

where $L_i = \frac{2c_2 K}{\Delta_i^{\min}}$. And inequality (a) follows from Equation (54), (b) is due to the TPE to replace $q_i^{\boldsymbol{\mu}, \pi_t} = \mathbb{E}_t[\mathbb{I}\{i \in \tau_t\}]$, (c) follows from the fact that if $N_{t-1,i} > \frac{2c_2 K}{\Delta_i^{\min}}$, then $-\frac{\Delta_{\pi_t}}{K} + 2c_2 \frac{1}{N_{t-1,i}} \leq -\frac{\Delta_{\pi_t}}{K} + \frac{\Delta_i^{\min}}{K} \leq 0$; Else, $-\frac{\Delta_{\pi_t}}{K} + 2c_2 \frac{1}{N_{t-1,i}} \leq 2c_2 \frac{1}{N_{t-1,i}}$.

$$
Reg(T, E_{t,2}) = \mathbb{E} \left[\sum_{t=1}^{T} \Delta_{\pi_t} \right] \tag{57}
$$

$$
\overset{(a)}{\leq} \mathbb{E} \left[\sum_{t \in [T]} \mathbb{E}_t \left[\sum_{i \in \tau_t: N_{t-1,i} > 0} \kappa_i(N_{t-1,i}) \right] \right] \tag{58}
$$

$$
\overset{(b)}{=} \mathbb{E} \left[\sum_{t \in [T]} \sum_{i \in \tau_t: N_{t-1,i} > 0} \kappa_i(N_{t-1,i}) \right] \tag{59}
$$

$$
\overset{(c)}{=} \mathbb{E} \left[\sum_{i \in [m]} \sum_{s=1}^{N_{T-1,i}} \kappa_i(s) \right] \tag{60}
$$

$$\leq \sum_{i \in [m]} \sum_{\ell=1}^{L_i} \frac{2c_2}{\ell} \tag{61}$$

$$\leq \sum_{i \in [m]} 2c_2 \left(1 + \int_{s=1}^{L_i} \frac{1}{s} \cdot ds\right) \tag{62}$$

$$= \sum_{i \in [m]} 2c_2 \left(1 + \log\left(\frac{2c_2 K}{\Delta_i^{\min}}\right)\right), \tag{63}$$

where (a) follows from Equation (55), (b) follows from the tower rule, (c) follows from that $N_{t-1,i}$ is increased by 1 if and only if $i \in \tau_t$.

Step 4: Bound the $\text{Reg}(T, \mathcal{E}'_{t,3})$ term

By Lemma 5, we have

$$\text{Reg}(T, \mathcal{E}'_{t,3}) \leq c_3 m. \tag{64}$$

Step 5: Putting everything together

Plugging Eq. (52), Eq. (63), Eq. (64) into Eq. (36), we have

$$\text{Reg}(T, \mathcal{E}_t) \leq \sum_{i \in [m]} \frac{2c_1^2}{\Delta_i^{\min}}(3 + \log K) + \sum_{i \in [m]} 2c_2\left(1 + \log\left(\frac{2c_2 K}{\Delta_i^{\min}}\right)\right) + c_3 m \tag{65}$$

$$\leq \sum_{i \in [m]} \frac{144(\bar{F} + \bar{G})}{\Delta_i^{\min}}(3 + \log K) + \sum_{i \in [m]} 12(\bar{I} + \bar{J})\left(1 + \log\left(\frac{12(\bar{I} + \bar{J})K}{\Delta_i^{\min}}\right)\right) + 3m\bar{w}d \tag{66}$$

$$= O\left(\sum_{i \in [m]} \frac{(\bar{F} + \bar{G})}{\Delta_i^{\min}} + 2(\bar{I} + \bar{J})\log\left(\frac{2(\bar{I} + \bar{J})K}{\Delta_i^{\min}}\right)\right) \tag{67}$$

Let $\text{Reg}(T, \{\}) \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t \in [T]}(\alpha \cdot r(\pi^*; \boldsymbol{\mu}) - r(\pi_t; \boldsymbol{\mu}))\right]$ be the regret event without any filter events. Now consider the regret caused by the failure event $(\neg\mathcal{E}_{s,t})_{t \in [T]}, \neg\mathcal{E}_{c,1}, \neg\mathcal{E}_{c,2}$, we have

$$\text{Reg}(T, \{\}) \leq \text{Reg}(T, \mathcal{E}_{s,t} \cap \mathcal{E}_{c,1} \cap \mathcal{E}_{c,2}) + \text{Reg}(T, \neg\mathcal{E}_{s,t}) + \text{Reg}(T, \neg\mathcal{E}_{c,1}) + \text{Reg}(T, \neg\mathcal{E}_{c,2}) \tag{68}$$

$$\overset{(a)}{\leq} \text{Reg}(T, \mathcal{E}_t) + (1 - \beta)T\Delta_{\max} + 2\Delta_{\max} \tag{69}$$

where inequality (a) is due to $\mathcal{E}_{s,t} \cap \mathcal{E}_{c,1} \cap \mathcal{E}_{c,2}$ implies $\mathcal{E}_t$ for the first term and the fact that $\Delta_{\pi_t} \leq \Delta_{\max}, \Pr[\neg\mathcal{E}_{s,t}] \leq \beta, \Pr[\neg\mathcal{E}_{c,1}] \leq 1/T, \Pr[\neg\mathcal{E}_{c,2}] \leq 1/T$ for the rest of the terms.

Therefore we can derive that the regret is upper bounded by

$$\text{Reg}(T; \alpha, \beta, \boldsymbol{\mu}) = \text{Reg}(T, \{\}) - (1 - \beta)T\alpha r(\pi^*; \boldsymbol{\mu}) \tag{70}$$

$$\leq \text{Reg}(T, \{\}) - (1 - \beta)T\Delta_{\max} \tag{71}$$

$$\leq \text{Reg}(T, \mathcal{E}_t) + 2\Delta_{\max} \tag{72}$$

$$= O\left(\sum_{i \in [m]} \frac{(\bar{F} + \bar{G})}{\Delta_i^{\min}} + 2(\bar{I} + \bar{J})\log\left(\frac{2(\bar{I} + \bar{J})K}{\Delta_i^{\min}}\right)\right) \tag{73}$$

For the gap-independent regret take $\Delta = \sqrt{144m(\bar{F} + \bar{G})/T}$. On one hand, $\text{Reg}(T, \mathbb{I}\{\Delta_{\pi_t} < \Delta\} \cap \mathcal{E}_t) \leq T\Delta$; On the other hand, $\text{Reg}(T, \mathbb{I}\{\Delta_{\pi_t} \geq \Delta\} \cap \mathcal{E}_t) \leq \sum_{i \in [m]} \frac{144(\bar{F} + \bar{G})K}{\Delta}(3 + \log K) + \sum_{i \in [m]} 12(\bar{I} + \bar{J})\left(1 + \log\left(\frac{12(\bar{I} + \bar{J})K}{\Delta}\right)\right) + 3m\bar{w}d$ according to Eq. (67). Therefore, $\text{Reg}(T; \alpha, \beta, \boldsymbol{\mu}) \leq \text{Reg}(T, \mathbb{I}\{\Delta_{\pi_t} < \Delta\} \cap \mathcal{E}_t) + \text{Reg}(T, \mathbb{I}\{\Delta_{\pi_t} \geq \Delta\} \cap \mathcal{E}_t) + 2\Delta_{\max} \leq O\left(\sqrt{(\bar{F} + \bar{G})mT} + m(\bar{I} + \bar{J})\log(KT)\right)$ using the similar proof of Eq. (73).

C. Analysis for Episodic RL with Sublinear Regret in Section 4.3

C.1. Proof of Lemma 1 and Lemma 2

We use $q_{(s',i)}^{\boldsymbol{p},\pi}(s,a,h)$ to denote the triggering probability $q_{s,a,h}^{\boldsymbol{p},\pi}$ when the policy is π, the transition is $\boldsymbol{p}$, and starting from initial state s' at step i. For notational simplicity, we use $q_{(s',i)}(s,a,h)$ to denote $q_{(s',i)}^{\boldsymbol{p},\pi}(s,a,h)$ and $\tilde{q}_{(s',i)}(s,a,h)$ to denote $q_{(s',i)}^{\tilde{\boldsymbol{p}},\pi}(s,a,h)$. Similarly, we use $q(s,a,h)$ to denote $q_{(s_1,1)}^{\boldsymbol{p},\pi}(s,a,h)$ and $\tilde{q}(s,a,h)$ to denote $q_{(s_1,1)}^{\tilde{\boldsymbol{p}},\pi}(s,a,h)$ when the starting state is fixed from s_1 at the initial step.

Lemma 6 (Smoothness of the Triggering Probability and the Value Function). *For any triggering probability q and $\tilde{q}$ given by the same policy π but different transition $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$, respectively, we have*

$$\tilde{q}(s,a,h) - q(s,a,h) = \sum_{(s',a'),s''} \sum_{i=1}^{h-1} q(s',a',i)(\tilde{p}(s''|s',a',i) - p(s''|s',a',i))\tilde{q}_{(s'',i+1)}(s,a,h). \tag{74}$$

And we have

$$\left| V_1^{\tilde{\boldsymbol{p}},\pi}(s_1) - V_1^{\boldsymbol{p},\pi}(s_1) \right| \leq \sum_{s,a,h} q_{s,a,h}^{\boldsymbol{p},\pi} \left| [\tilde{\boldsymbol{p}}(s,a,h) - \boldsymbol{p}(s,a,h)]^\top V_{h+1}^{\tilde{\boldsymbol{p}},\pi} \right| \tag{75}$$

$$\leq H \sum_{s,a,h} q_{s,a,h}^{\boldsymbol{p},\pi} \|\tilde{\boldsymbol{p}}(s,a,h) - \boldsymbol{p}(s,a,h)\|_1 \tag{76}$$

Proof. We first prove Eq. (74) by induction on h. When $h = 1$, then $\tilde{q}(s,a,h) = q(s,a,h) = \mathbb{I}\{\pi(s,h) = a\}\mathbb{I}\{s = s_1\}$, Eq. (74) holds. For the induction step $h > 1$:

$$\tilde{q}(s,a,h) - q(s,a,h) \stackrel{(a)}{=} \mathbb{I}\{\pi(s,h) = a\}\left[\sum_{s',a'} \tilde{q}(s',a',h-1)\tilde{p}(s|s',a',h-1) - q(s',a',h-1)p(s|s',a',h-1) \right] \tag{77}$$

$$= \underbrace{\mathbb{I}\{\pi(s,h) = a\} \sum_{s',a'} \tilde{p}(s|s',a',h-1)(\tilde{q}(s',a',h-1) - q(s',a',h-1))}_{\text{Term 1}}$$

$$+ \underbrace{\mathbb{I}\{\pi(s,h) = a\} \sum_{s',a'} q(s',a',h-1)(\tilde{p}(s|s',a',h-1) - p(s|s',a',h-1))}_{\text{Term 2}} \tag{78}$$

where inequality (a) is due to $q(s,a,h) = \sum_{s',a'} \mathbb{I}\{\pi(s,h) = a\}q(s',a',h-1)p(s|s',a',h-1)$. Then we bound Term 1 and Term 2 as follows.

$$\text{Term 1} \stackrel{(a)}{=} \mathbb{I}\{\pi(s,h) = a\} \sum_{s',a'} \tilde{p}(s|s',a',h-1)$$

$$\cdot \left(\sum_{(s'',a''),s'''} \sum_{i=1}^{h-2} q(s'',a'',i)(\tilde{p}(s'''|s'',a'',i) - p(s'''|s'',a'',i))\tilde{q}_{(s''',i+1)}(s',a',h-1) \right) \tag{79}$$

$$= \sum_{(s'',a''),s'''} \sum_{i=1}^{h-2} q(s'',a'',i)(\tilde{p}(s'''|s'',a'',i) - p(s'''|s'',a'',i))$$

$$\cdot \sum_{s',a'} \tilde{p}(s|s',a',h-1)\mathbb{I}\{\pi(s,h) = a\}\tilde{q}_{(s''',i+1)}(s',a',h-1) \tag{80}$$

$$\stackrel{(b)}{=} \sum_{(s'',a''),s'''} \sum_{i=1}^{h-2} q(s'',a'',i)(\tilde{p}(s'''|s'',a'',i) - p(s'''|s'',a'',i))\tilde{q}_{(s''',i+1)}(s,a,h) \tag{81}$$

$$(82)$$

where equality (a) is due to the induction hypothesis, equality (b) is due to $\tilde{q}_{(s''',i+1)}(s,a,h) = \sum_{s',a'} \tilde{p}(s|s',a',h-1)\mathbb{I}\{\pi(s,h) = a\}\tilde{q}_{(s''',i+1)}(s',a',h-1)$.

$$\text{Term 2} = \sum_{(s',a'),s''} \mathbb{I}\{\pi(s'',h) = a\}\mathbb{I}\{s'' = s\}q(s',a',h-1)(\tilde{p}(s''|s',a',i) - p(s''|s',a',i)) \tag{83}$$

$$\stackrel{(a)}{=} \sum_{(s',a'),s''} q(s',a',h-1)(\tilde{p}(s''|s',a',i) - p(s''|s',a',i))\tilde{q}_{(s'',h)}(s,a,h) \tag{84}$$

where equality (a) is due to $\tilde{q}_{(s'',h)}(s,a,h) = \mathbb{I}\{\pi(s'',h) = a\}\mathbb{I}\{s'' = s\}$.

Plugging in Term 1 (with changing variables s''', s'', a'' to s'', s', a') and Term 2 proves the Eq. (74).

Now we prove the smoothness for the value function for the second part of Lemma 6.

$$\left|V_1^{\tilde{p},\pi}(s_1) - V_1^{p,\pi}(s_1)\right| = \left|\sum_{s,a,h}(\tilde{q}(s,a,h) - q(s,a,h))r(s,a,h)\right| \tag{85}$$

$$\stackrel{(a)}{=} \left|\sum_{s,a,h}\sum_{(s',a'),s''}\sum_{i=1}^{h-1} q(s',a',i)(\tilde{p}(s''|s',a',i) - p(s''|s',a',i))\tilde{q}_{(s'',i+1)}(s,a,h)r(s,a,h)\right| \tag{86}$$

$$= \left|\sum_{i=1}^{H}\sum_{(s',a'),s''} q(s',a',i)(\tilde{p}(s''|s',a',i) - p(s''|s',a',i)) \sum_{s,a,h>i}\tilde{q}_{(s'',i+1)}(s,a,h)r(s,a,h)\right|$$
$$\tag{87}$$

$$= \left|\sum_{i=1}^{H}\sum_{(s',a'),s''} q(s',a',i)(\tilde{p}(s''|s',a',i) - p(s''|s',a',i))V_{h+1}^{\tilde{p},\pi}(s'')\right| \tag{88}$$

$$\leq \sum_{s,a,h} q_{s,a,h}^{p,\pi}\left|[\tilde{p}(s,a,h) - p(s,a,h)]^\top V_{h+1}^{\tilde{p},\pi}\right| \tag{89}$$

$$\stackrel{(b)}{\leq} H\sum_{s,a,h} q_{s,a,h}^{p,\pi}\|\tilde{p}(s,a,h) - p(s,a,h)\|_1 \tag{90}$$

where equality (a) is due to the first part of Lemma 6 we just proved, inequality (b) is due to $V_{h+1}^{\tilde{p},\pi}(s) \leq H$. ∎

C.2. Proof of Theorem 2

We suppose the concentration event $\mathcal{E}_{\text{tran1}}$ as defined in Eq. (102) holds with probability $\delta' = 1/(2T)$. Let $L = \log(SAHT)$. Since $\boldsymbol{\mu} \in \mathcal{C}_t$ as defined in Eq. (11) due to the event $\mathcal{E}_{\text{tran1}}$, we follow the same regret decomposition and derivation as in Step 1 of Appendix B.3, and identify $F_{t,s,a,h} = 2H\sqrt{SL}, I_{t,s,a,h} = 0$ due to the definition of $\phi_t(s,a,h)$ in Eq. (11), and $G_{t,s,a,h} = J_{t,s,a,h} = 0$ since $\boldsymbol{w}_{s,a,h}^{p,\pi} = H \cdot \boldsymbol{1}$ are constants. Now we can derive that $\bar{F} = \sum_{s,a,h} q_{s,a,h}^{p,\pi} F_{t,s,a,h}^2 = 4H^3SL, \bar{G} = \bar{I} = \bar{J} = 0, \bar{w} = H, d = S$, we have

$$\Delta_{\pi_t} \leq 2\left(\sqrt{4H^3SL}\right)\sqrt{\sum_{i\in[m]:N_{t-1,i}>0}\frac{q_i^{p,\pi_t}}{N_{t-1,i}} + \sum_{i\in[m]:N_{t-1,i}=0}q_i^{p,\pi_t}SH} \tag{91}$$

Following step 2,3 in Appendix B.3, we have gap-dependent regret

$$Reg(T) = O\left(\sum_{s,a,h}\frac{H^3S\log(SAHT)}{\Delta_{s,a,h}^{\min}}\right) \tag{92}$$

and gap-independent regret

$$Reg(T) = O\left(\sqrt{H^4 S^2 A T \log(SAHT)}\right).$$ (93)

D. Analysis that Achieve Minimax Optimal Regret for Episodic RL in Section 4.4

D.1. Concentration Inequalities

Lemma 7 (Concentration of the Transition).

$$\Pr\left[\|\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h)\|_1 \le \sqrt{\frac{2S \log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}}, \text{ for any } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], t \in [T]\right] \ge 1 - 2\delta'$$ (94)

and

$$\Pr\left[|\hat{p}_{t-1}(s'|s,a,h) - p(s'|s,a,h)| \le \sqrt{\frac{p(s'|s,a,h)(1 - p(s'|s,a,h)) \log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}} + \frac{\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)},\right.$$

$$\left.\text{for any } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], t \in [T]\right] \ge 1 - 2\delta'$$ (95)

Proof. Using the (Weissman et al., 2003) for the first part, and using the Bernstein inequality for the second part, and taking the union bound over $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$ and the counter $N_{t-1}(s,a,h) \in [T]$, we obtain the lemma. ∎

Lemma 8 (Concentration of the Optimal Future Value).

$$\Pr\left[\left|(\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}_{h+1}^*\right| \le 2\sqrt{\frac{\text{Var}_{s' \sim \boldsymbol{p}(s,a,h)}\left(V_{h+1}^*(s')\right) \log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}} + \frac{H \log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)},\right.$$

$$\left.\text{for any } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], t \in [T]\right] \ge 1 - 2\delta'$$ (96)

Proof. Using the Bernstein inequality and taking the union bound over $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$ and the counter $N_{t-1}(s,a,h) \in [T]$, we obtain the lemma. ∎

Lemma 9 (Concentration of the Variance).

$$\Pr\left[\left|\sqrt{\text{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right)} - \sqrt{\text{Var}_{s' \sim \boldsymbol{p}(s,a,h)}\left(V_{h+1}^*(s')\right)}\right| \le \sqrt{\mathbb{E}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - V_{h+1}^*(s')\right]^2}\right.$$

$$\left. + 2H\sqrt{\frac{\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}}, \text{ for any } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], t \in [T]\right] \ge 1 - 2\delta'$$ (97)

Proof. According to Proposition 2 (i.e., Eq. (53)) in (Zanette & Brunskill, 2019), we have

$$\Pr\left[\left|\sqrt{\text{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(V_{h+1}^*(s')\right)} - \sqrt{\text{Var}_{s' \sim \boldsymbol{p}(s,a,h)}\left(V_{h+1}^*(s')\right)}\right| \le 2H\sqrt{\frac{\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}},\right.$$

$$\left.\text{for any } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], t \in [T]\right] \ge 1 - 2\delta'$$ (98)

Then we have

$$\left|\sqrt{\text{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right)} - \sqrt{\text{Var}_{s' \sim \boldsymbol{p}(s,a,h)}\left(V_{h+1}^*(s')\right)}\right|$$ (99)

$$\le \left|\sqrt{\text{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right)} - \sqrt{\text{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(V_{h+1}^*(s')\right)}\right|$$

$$+ \left|\sqrt{\text{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(V_{h+1}^*(s')\right)} - \sqrt{\text{Var}_{s' \sim \boldsymbol{p}(s,a,h)}\left(V_{h+1}^*(s')\right)}\right|$$ (100)

$$\overset{(a)}{\leq} \sqrt{\mathbb{E}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - V^*_{h+1}(s')\right]^2} + 2H\sqrt{\frac{\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}} \tag{101}$$

where inequality (a) is due to Eq. (48)-(52) in (Zanette & Brunskill, 2019) and Eq. (98) that holds with probability at least $1 - 2\delta'$. ∎

Based on the concentration lemmas we use, we define the following events.

$$\mathcal{E}_{\text{tran1}} \overset{\text{def}}{=} \left[\|\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h)\|_1 \leq \sqrt{\frac{2S\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}}, \text{ for any } (s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H], t\in[T]\right] \tag{102}$$

$$\mathcal{E}_{\text{tran2}} \overset{\text{def}}{=} \left[|\hat{p}_{t-1}(s'|s,a,h) - p(s'|s,a,h)| \leq \sqrt{\frac{p(s'|s,a,h)(1 - p(s'|s,a,h))\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}} + \frac{\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}\right] \tag{103}$$

$$\mathcal{E}_{\text{future}} \overset{\text{def}}{=} \left[\left|(\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}^*_{h+1}\right| \leq 2\sqrt{\frac{\text{Var}_{s'\sim\boldsymbol{p}(s,a,h)}\left(V^*_{h+1}(s')\right)\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}} + \frac{H\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)},\right.$$

$$\left.\text{for any } (s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H], t\in[T]\right] \tag{104}$$

$$\mathcal{E}_{\text{var}} \overset{\text{def}}{=} \left[\left|\sqrt{\text{Var}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right)} - \sqrt{\text{Var}_{s'\sim\boldsymbol{p}(s,a,h)}\left(V^*_{h+1}(s')\right)}\right| \leq \sqrt{\mathbb{E}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - V^*_{h+1}(s')\right]^2}\right.$$

$$\left. + 2H\sqrt{\frac{\log\left(\frac{SAHT}{\delta'}\right)}{N_{t-1}(s,a,h)}}, \text{ for any } (s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H], t\in[T]\right] \tag{105}$$

$$\tag{106}$$

$$\mathcal{E} = \mathcal{E}_{\text{tran1}} \cap \mathcal{E}_{\text{tran2}} \cap \mathcal{E}_{\text{future}} \cap \mathcal{E}_{\text{var}} \tag{107}$$

Lemma 10 (High Probability Event). *Let $\delta = \delta'/8$, then*

$$\Pr[\mathcal{E}] \geq 1 - 8\delta' = 1 - \delta. \tag{108}$$

Proof. We can obtain this lemma by Lemmas 7 to 9. ∎

Lemma 11 (Concentration of the Optimal Future Value Regarding Known Statistics). *Let $L = \log\left(\frac{8SAHT}{\delta}\right)$. Let*
$\phi_t(s,a,h) = 2\sqrt{\frac{\text{Var}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right)L}{N_{t-1}(s,a,h)}} + 2\sqrt{\frac{\mathbb{E}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}} + \frac{5HL}{N_{t-1}(s,a,h)}.$ *With probability at least $1 - \delta$, we have*

$$\left|(\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}^*_{h+1}\right| \leq \phi_t(s,a,h) \tag{109}$$

Proof. Under $\mathcal{E}$, we can obtain the lemma by applying Lemma 8, Lemma 9, and Lemma 12 ∎

D.2. Optimism and Pessimism

Let $L = \log\left(\frac{SAHT}{\delta'}\right)$. Let $\phi_t(s,a,h) = 2\sqrt{\frac{\text{Var}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right)L}{N_{t-1}(s,a,h)}} + 2\sqrt{\frac{\mathbb{E}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}} + \frac{5HL}{N_{t-1}(s,a,h)}.$

Lemma 12. *If concentration event $\mathcal{E}$ holds, then it holds that*

$$\underline{V}_{t,h}(s) \leq V^*_h(s) \leq \bar{V}_{t,h}(s) \tag{110}$$

for all $s\in\mathcal{S}, h\in[H], t\in[T]$.

Proof. We prove this lemma by induction. Since it holds that $\underline{V}_{t,H+1}(s) = V^*_{H+1}(s) = \bar{V}_{t,H+1}(s) = 0$, so it suffices to prove that if $\underline{V}_{t,h+1}(s) \leq V^*_{h+1}(s) \leq \bar{V}_{t,h+1}(s)$, then $\underline{V}_{t,h}(s) \leq V^*_h(s) \leq \bar{V}_{t,h}(s)$.

We first prove the optimistic part, i.e., $V^*_h(s) \leq \bar{V}_{t,h}(s)$. If

$$r(s, \pi_t(s,h), h) + \hat{\boldsymbol{p}}_{t-1}(s, \pi_t(s,h), h)^\top \bar{\boldsymbol{V}}_{t,h+1} + \phi_t(s, \pi_t(s,h), h) \geq H - h, \tag{111}$$

then we are done. If the above does not hold, we have the following,

$$\bar{V}_{t,h}(s) = r(s, \pi_t(s,h), h) + \hat{\boldsymbol{p}}_{t-1}(s, \pi_t(s,h), h)^\top \bar{\boldsymbol{V}}_{t,h+1} + \phi_t(s, \pi_t(s,h), h) \tag{112}$$

$$\overset{(a)}{\geq} r(s, \pi^*(s,h), h) + \hat{\boldsymbol{p}}_{t-1}(s, \pi^*(s,h), h)^\top \bar{\boldsymbol{V}}_{t,h+1} + \phi_t(s, \pi^*(s,h), h) \tag{113}$$

$$\overset{(b)}{\geq} r(s, \pi^*(s,h), h) + \hat{\boldsymbol{p}}_{t-1}(s, \pi^*(s,h), h)^\top \boldsymbol{V}^*_{h+1} + \phi_t(s, \pi^*(s,h), h) \tag{114}$$

$$\overset{(c)}{\geq} r(s, \pi^*(s,h), h) + \hat{\boldsymbol{p}}_{t-1}(s, \pi^*(s,h), h)^\top \boldsymbol{V}^*_{h+1} + 2\sqrt{\frac{\mathrm{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s, \pi^*(s,h), h)}\left(\bar{V}_{t,h+1}(s')\right) L}{N_{t-1}(s, \pi^*(s,h), h)}}$$

$$+ 2\sqrt{\frac{\mathbb{E}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s, \pi^*(s,h), h)}\left[\bar{V}_{t,h+1}(s') - V^*_{h+1}(s')\right]^2 L}{N_{t-1}(s, \pi^*(s,h), h)}} + \frac{5HL}{N_{t-1}(s, \pi^*(s,h), h)} \tag{115}$$

$$\overset{(d)}{\geq} r(s, \pi^*(s,h), h) + \hat{\boldsymbol{p}}_{t-1}(s, \pi^*(s,h), h)^\top \boldsymbol{V}^*_{h+1} + 2\sqrt{\frac{\mathrm{Var}_{s' \sim \boldsymbol{p}(s, \pi^*(s,h), h)}\left(V^*_{h+1}(s')\right) L}{N_{t-1}(s, \pi^*(s,h), h)}} + \frac{HL}{N_{t-1}(s, \pi^*(s,h), h)} \tag{116}$$

$$\overset{(e)}{\geq} r(s, \pi^*(s,h), h) + \boldsymbol{p}(s, \pi^*(s,h), h)^\top \boldsymbol{V}^*_{h+1} \tag{117}$$

$$= V^*_h(s), \tag{118}$$

where inequality (a) is due to taking the maximization over the actions in the optimistic MDP, inequality (b) is due to the inductive hypothesis $V^*_{h+1}(s) \leq \bar{V}_{t,h+1}(s)$, inequality (c) is due to the inductive hypothesis $\underline{V}_{t,h+1}(s) \leq V^*_{h+1}(s)$, inequality (d) is due to Lemma 9, and inequality (e) is due to Lemma 8.

Next we prove the pessimistic part. Let let $a = \pi_t(s,h)$. Similarly, if

$$r(s, a, h) + \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \underline{\boldsymbol{V}}_{t,h+1} - \phi_t(s, a, h) \leq 0, \tag{119}$$

we are done. If the above inequality does not hold, we have

$$\underline{V}_{t,h}(s) = r(s, a, h) + \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \underline{\boldsymbol{V}}_{t,h+1} - \phi_t(s, a, h) \tag{120}$$

$$\overset{(a)}{\leq} r(s, a, h) + \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \boldsymbol{V}^*_{h+1} - 2\sqrt{\frac{\mathrm{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s, a, h)}\left(\bar{V}_{t,h+1}(s')\right) L}{N_{t-1}(s, a, h)}}$$

$$- 2\sqrt{\frac{\mathbb{E}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s, a, h)}\left[\bar{V}_{t,h+1}(s') - V^*_{h+1}(s')\right]^2 L}{N_{t-1}(s, a, h)}} - \frac{5HL}{N_{t-1}(s, a, h)} \tag{121}$$

$$\overset{(b)}{\leq} r(s, a, h) + \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \boldsymbol{V}^*_{h+1} - 2\sqrt{\frac{\mathrm{Var}_{s' \sim \boldsymbol{p}(s, a, h)}\left(V^*_{h+1}(s')\right) L}{N_{t-1}(s, a, h)}} - \frac{HL}{N_{t-1}(s, a, h)} \tag{122}$$

$$\overset{(c)}{\leq} r(s, a, h) + \boldsymbol{p}(s, a, h)^\top \boldsymbol{V}^*_{h+1} \tag{123}$$

$$\leq r(s, \pi^*(s,h), h) + \boldsymbol{p}(s, \pi^*(s,h), h)^\top \boldsymbol{V}^*_{h+1} \tag{124}$$

$$= V^*_h(s), \tag{125}$$

where inequality (a) is due to the inductive hypothesis $V^*_{h+1}(s) \geq \underline{V}_{t,h+1}(s)$, inequality (b) is due to Lemma 9, inequality (c) is due to Lemma 8. $\blacksquare$

Lemma 13 (Difference between optimism and pessimism). *If the concentration event $\mathcal{E}$ holds, then we have*

$$\bar{V}_{t,h}(s) - \underline{V}_{t,h}(s) \leq \sum_{i=h}^H \mathbb{E}_{(s_i, a_i) \sim \pi_t}\left[\min\left\{\frac{20HL\sqrt{S}}{\sqrt{N_{t-1}(s_i, a_i, i)}}, H\right\} \mid s_h = s, \pi_t\right] \tag{126}$$

Proof. Let $a = \pi_t(s, h)$ be the action chosen by our algorithm at t-th episode,

$$\bar{V}_{t,h}(s) \leq r(s, a, h) + \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \bar{\boldsymbol{V}}_{t,h+1} + \phi_t(s, a, h) \tag{127}$$

$$\underline{V}_{t,h}(s) \geq r(s, a, h) + \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \underline{\boldsymbol{V}}_{t,h+1} - \phi_t(s, a, h) \tag{128}$$

Then,

$$\bar{V}_{t,h}(s) - \underline{V}_{t,h}(s) \tag{129}$$

$$\leq \hat{\boldsymbol{p}}_{t-1}(s, a, h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\right) + 2\phi_t(s, a, h) \tag{130}$$

$$= \boldsymbol{p}(s, a, h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\right) + (\hat{\boldsymbol{p}}_{t-1}(s, a, h) - \boldsymbol{p}(s, a, h))^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\right)$$
$$+ 4\sqrt{\frac{\mathrm{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right) L}{N_{t-1}(s, a, h)}} + 4\sqrt{\frac{\mathbb{E}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s, a, h)}} + \frac{10HL}{N_{t-1}(s, a, h)} \tag{131}$$

$$\leq \boldsymbol{p}(s, a, h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\right) + \|\hat{\boldsymbol{p}}_{t-1}(s, a, h) - \boldsymbol{p}(s, a, h)\|_1 \|\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\|_\infty$$
$$+ 4\sqrt{\frac{H^2 L}{N_{t-1}(s, a, h)}} + 4\sqrt{\frac{H^2 L}{N_{t-1}(s, a, h)}} + \frac{10HL}{N_{t-1}(s, a, h)} \tag{132}$$

$$\overset{(a)}{\leq} \boldsymbol{p}(s, a, h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\right) + H\sqrt{\frac{2SL}{N_{t-1}(s, a, h)}} + 18HL\sqrt{\frac{1}{N_{t-1}(s, a, h)}} \tag{133}$$

$$\leq \boldsymbol{p}(s, a, h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1}\right) + \frac{20HL\sqrt{S}}{\sqrt{N_{t-1}(s, a, h)}} \tag{134}$$

$$= \mathbb{E}_{s' \sim \boldsymbol{p}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right] + \frac{20HL\sqrt{S}}{\sqrt{N_{t-1}(s, a, h)}} \tag{135}$$

$$\overset{(b)}{\leq} \sum_{i=h}^{H} \mathbb{E}_{(s_i, a_i) \sim \pi_t}\left[\min\left\{\frac{20HL\sqrt{S}}{\sqrt{N_{t-1}(s_i, a_i, i)}}, H\right\} \mid s_h = s, \pi_t\right] \tag{136}$$

where inequality (a) is due to Lemma 7, and inequality (b) is due to $a = \pi_t(s, h)$ and recursively apply the same operation on $\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')$ till step H when coupled with the fact that $\bar{V}_{t,h}(s) - \underline{V}_{t,h}(s) \leq H$. ∎

Lemma 14 (Cumulative difference between optimism and pessimism). *If the concentration event $\mathcal{E}$ holds, then*

$$\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \boldsymbol{p}(s, a, h)^\top (\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})^2 \leq \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \frac{400 H^4 L^2 S}{N_{t-1}(s, a, h)} \tag{137}$$

Proof. For any $t \in [T], h \in [H], s \in \mathcal{S}$, let $w_{t,h}(s)$ denote the probability that state s is visited at step h in episode t. We can bound

$$\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \boldsymbol{p}(s, a, h)^\top (\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})^2 \tag{138}$$

$$= \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sum_{s'} p(s'|s, a, h)(\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s'))^2 \tag{139}$$

$$= \sum_{h,s',s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} p(s'|s, a, h)(\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s'))^2 \tag{140}$$

$$= \sum_{h,s'} w_{t,h+1}(s')(\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s'))^2 \tag{141}$$

$$= \sum_{h=1}^{H} \mathbb{E}_{s_{h+1} \sim \pi_t}(\bar{V}_{t,h+1}(s_{h+1}) - \underline{V}_{t,h+1}(s_{h+1}))^2 \tag{142}$$

$$\leq \sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_t}(\bar{V}_{t,h}(s_h) - \underline{V}_{t,h}(s_h))^2 \tag{143}$$

$$\overset{(a)}{\leq} \sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_t} \left(\sum_{i=h}^{H} \mathbb{E}_{(s_i, a_i) \sim \pi_t} \left[\frac{20HL\sqrt{S}}{\sqrt{N_{t-1}(s_i, a_i, h)}} \mid s_h, \pi_t \right] \right)^2 \tag{144}$$

$$\overset{(b)}{\leq} H \sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_t} \sum_{i=h}^{H} \left(\mathbb{E}_{(s_i, a_i) \sim \pi_t} \left[\frac{20HL\sqrt{S}}{\sqrt{N_{t-1}(s_i, a_i, h)}} \mid s_h, \pi_t \right] \right)^2 \tag{145}$$

$$\overset{(c)}{\leq} H \sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_t} \sum_{i=h}^{H} \mathbb{E}_{(s_i, a_i) \sim \pi_t} \left[\frac{400H^2L^2S}{N_{t-1}(s_i, a_i, h)} \mid s_h, \pi_t \right] \tag{146}$$

$$\leq H \sum_{h=1}^{H} \sum_{i=h}^{H} \mathbb{E}_{(s_i, a_i) \sim \pi_t} \left[\frac{400H^2L^2S}{N_{t-1}(s_i, a_i, h)} \right] \tag{147}$$

$$\leq H^2 \sum_{h=1}^{H} \mathbb{E}_{(s_h, a_h) \sim \pi_t} \left[\frac{400H^2L^2S}{N_{t-1}(s_h, a_h, h)} \right] \tag{148}$$

$$= \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \frac{400H^4L^2S}{N_{t-1}(s, a, h)} \tag{149}$$

where inequality (a) is due to Lemma 13, inequality (b) is due to Cauchy-Schwarz inequality, inequality (c) is due to Jensen's inequality. ∎

D.3. Variance Inequalities

Lemma 15 (Cumulative difference of the variance). *If the concentration event $\mathcal{E}$ holds, then it holds for all $t \in [T]$ that*

$$\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sqrt{\frac{\mathrm{Var}_{s' \sim \boldsymbol{p}(s,a,h)} \left(V_{h+1}^*(s') \right) L}{N_{t-1}(s, a, h)}} - \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sqrt{\frac{\mathrm{Var}_{s' \sim \boldsymbol{p}(s,a,h)} \left(V_{h+1}^{\pi_t}(s') \right) L}{N_{t-1}(s, a, h)}} \leq \sqrt{H^2 L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p}, \pi_t} \cdot \Delta_{\pi_t}}{N_{t-1}(s, a, h)}} \tag{150}$$

Proof. For any $t \in [T], h \in [H], s \in \mathcal{S}$, let $w_{t,h}(s)$ denote the probability that state s is visited at step h in episode t.

$$\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sqrt{\frac{\mathrm{Var}_{s' \sim \boldsymbol{p}(s,a,h)} \left(V_{h+1}^*(s') \right) L}{N_{t-1}(s, a, h)}} - \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sqrt{\frac{\mathrm{Var}_{s' \sim \boldsymbol{p}(s,a,h)} \left(V_{h+1}^{\pi_t}(s') \right) L}{N_{t-1}(s, a, h)}} \tag{151}$$

$$\overset{(a)}{\leq} \sqrt{L} \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sqrt{\frac{\mathbb{E}_{s' \sim \boldsymbol{p}(s,a,h)} \left[\left(V_{h+1}^*(s') - V_{h+1}^{\pi_t}(s') \right)^2 \right]}{N_{t-1}(s, a, h)}} \tag{152}$$

$$\leq \sqrt{L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p}, \pi_t}}{N_{t-1}(s, a, h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \mathbb{E}_{s' \sim \boldsymbol{p}(s,a,h)} \left[\left(V_{h+1}^*(s') - V_{h+1}^{\pi_t}(s') \right)^2 \right]} \tag{153}$$

$$= \sqrt{L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p}, \pi_t}}{N_{t-1}(s, a, h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sum_{s'} p(s'|s, a, h) \left(V_{h+1}^*(s') - V_{h+1}^{\pi_t}(s') \right)^2} \tag{154}$$

$$\leq \sqrt{HL} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p}, \pi_t}}{N_{t-1}(s, a, h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p}, \pi_t} \sum_{s'} p(s'|s, a, h) \left(V_{h+1}^*(s') - V_{h+1}^{\pi_t}(s') \right)} \tag{155}$$

$$= \sqrt{HL} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p}, \pi_t}}{N_{t-1}(s, a, h)}} \sqrt{\sum_{h,s'} w_{t,h+1}(s') \left(V_{h+1}^*(s') - V_{h+1}^{\pi_t}(s') \right)} \tag{156}$$

$$\overset{(b)}{\leq} \sqrt{HL} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p}, \pi_t}}{N_{t-1}(s, a, h)}} \sqrt{\sum_{h} \left(V_1^*(s_1) - V_1^{\pi_t}(s_1) \right)} \tag{157}$$

$$= \sqrt{H^2 L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \sqrt{(V_1^*(s_1) - V_1^{\pi_t}(s_1))} \tag{158}$$

$$= \sqrt{H^2 L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}} \tag{159}$$

where inequality (a) is due to Eq. (48)-(52) in (Zanette & Brunskill, 2019) and inequality (b) is due to Lemma 17 in (Zanette & Brunskill, 2019) ∎

Lemma 16 (Law of total variance). *For any policy π, then it holds that*

$$\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi} \mathrm{Var}_{s' \sim \boldsymbol{p}(s,a,h)} \left(V_{h+1}^\pi(s') \right) \le H^2 \tag{160}$$

Proof. Let $E_\pi[\cdot|s_1]$ is taken over the trajectories following policy π starting from state s_1.

$$\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi} \mathrm{Var}_{s' \sim \boldsymbol{p}(s,a,h)} \left(V_{h+1}^\pi(s') \right) \tag{161}$$

$$= \sum_{h=1}^H \mathbb{E}_\pi \left[\mathrm{Var}_\pi \left(V_{h+1}^\pi(s_{h+1}) | s_h \right) | s_1 \right] \tag{162}$$

$$= \mathbb{E}_\pi \left[\sum_{h=1}^H \mathrm{Var}_\pi \left(V_{h+1}^\pi(s_{h+1}) | s_h \right) | s_1 \right] \tag{163}$$

$$\stackrel{(a)}{=} \mathbb{E}_\pi \left[\left(\sum_{h=1}^H r(s_h, \pi(s_h, h), h) - V_1^\pi(s_1) \right)^2 | s_1 \right] \tag{164}$$

$$\le H^2, \tag{165}$$

where (a) is due to Lemma 15 in (Zanette & Brunskill, 2019). ∎

D.4. Proof of Theorem 3

Now we prove the regret upper bound for Theorem 3. The proof of Theorem 3 is slightly different from that of Theorem 1 but follows the similar analysis ideas/steps and we show the details as follows.

We suppose the concentration event $\mathcal{E}$ holds. According to our joint oracle Algorithm 3, we have $\bar{\boldsymbol{V}}_{t,h} = \boldsymbol{V}_{t,h}^{\tilde{\boldsymbol{p}}_t,\pi_t}$, where $(\pi_t, \tilde{\boldsymbol{p}}_t) = \mathrm{argmax}_{\pi \in \Pi, \tilde{\boldsymbol{p}} \in \mathcal{C}_t(\pi)} V_1^{\tilde{\boldsymbol{p}},\pi}(s_1)$. Let $L = \log\left(\frac{SAHT}{\delta'}\right)$ and $\phi_t(s,a,h) = 2\sqrt{\frac{\mathrm{Var}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right) L}{N_{t-1}(s,a,h)}} + 2\sqrt{\frac{\mathbb{E}_{s' \sim \hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}} + \frac{5HL}{N_{t-1}(s,a,h)}$.

Step 1: Regret decomposition. We use a similar argument for the regret decomposition (Step 1) in Appendix B.3.

$$\Delta_{\pi_t} \stackrel{\mathrm{def}}{=} \left(V_1^{\boldsymbol{p},\pi^*}(s_1) - V_1^{\boldsymbol{p},\pi_t}(s_1) \right) \tag{166}$$

$$\stackrel{(a)}{\le} \left(V_{t,1}^{\tilde{\boldsymbol{p}}_t,\pi_t}(s_1) - V_1^{\boldsymbol{p},\pi_t}(s_1) \right) \tag{167}$$

$$\stackrel{(b)}{\le} \sum_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}} q_{s,a,h}^{\boldsymbol{p},\pi_t} \left| (\tilde{\boldsymbol{p}}_t(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}_{t,h+1}^{\tilde{\boldsymbol{p}}_t,\pi_t} \right| \tag{168}$$

$$\stackrel{(c)}{=} \underbrace{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \left| (\tilde{\boldsymbol{p}}_t(s,a,h) - \hat{\boldsymbol{p}}_{t-1}(s,a,h))^\top \boldsymbol{V}_{t,h+1}^{\tilde{\boldsymbol{p}}_t,\pi_t} \right|}_{\text{Term (I): Optimistic Future Value Regret}} \tag{169}$$

27

$$+ \sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \underbrace{\left| (\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}^*_{h+1} \right|}_{\text{Term (II): Optimal Future Value Regret}} \tag{170}$$

$$+ \sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \underbrace{\left| (\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \left(\boldsymbol{V}^{\tilde{\boldsymbol{p}}_t,\pi_t}_{t,h+1} - \boldsymbol{V}^*_{h+1} \right) \right|}_{\text{Term (III): Lower-Order Regret}}. \tag{171}$$

where inequality (a)-(c) following the same reasoning of (a)-(c) in the Step 1 of Appendix B.3. In what follows identify the $F_{t,s,a,h}, G_{t,s,a,h}, I_{t,s,a,h}, J_{t,s,a,h}$ and their upper bounds $\bar{F}, \bar{G}, \bar{I}, \bar{J}$ as in Theorem 1.

Step 2: Bound the optimistic future value regret - Term (I)

First, we can identify $F_{t,s,a,h} = 2\sqrt{\text{Var}_{s'\sim\hat{p}_{t-1}(s,a,h)}\left(\bar{V}_{t,h+1}(s')\right) L} + 2\sqrt{\mathbb{E}_{s'\sim\hat{p}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}$, $I_{t,s,a,h} = 5HL$ from Eq. (174). Slightly different from inequality (f) in Step 1 of Appendix B.3, $\sum_{s,a,h} q^{\boldsymbol{p},\pi}_{s,a,h} F^2_{t,s,a,h}$ will produce some lower-order terms $F' = O(H^4 SL^3 \sum_{s,a,h} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)})$ so that $\sqrt{F'}\sqrt{\sum_{s,a,h} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)}}$ can be merged into $\bar{I}$. Concretely, we have

$$\text{Term (I)} = \sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \left| (\tilde{\boldsymbol{p}}_t(s,a,h) - \hat{\boldsymbol{p}}_{t-1}(s,a,h))^\top \boldsymbol{V}^{\tilde{\boldsymbol{p}}_t,\pi_t}_{t,h+1} \right| \tag{172}$$

$$\overset{(a)}{\le} \sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \cdot \phi_t(s,a,h) \tag{173}$$

$$\overset{(b)}{\le} \sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \left(2\sqrt{\frac{\text{Var}_{s'\sim p(s,a,h)}\left(V^*_{h+1}(s')\right) L}{N_{t-1}(s,a,h)}} + 4\sqrt{\frac{\mathbb{E}_{s'\sim\hat{p}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}} \right.$$
$$\left. + \frac{9HL}{N_{t-1}(s,a,h)} \right) \tag{174}$$

$$\overset{(c)}{\le} 2\sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \sqrt{\frac{\text{Var}_{s'\sim p(s,a,h)}\left(V^{\pi_t}_{h+1}(s')\right) L}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2 L}\sqrt{\sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}} + 9HL \sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)}$$

$$+ 4\sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \sqrt{\frac{\mathbb{E}_{s'\sim\hat{p}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}} \tag{175}$$

$$\le 2\sqrt{L}\sqrt{\sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h}\text{Var}_{s'\sim p(s,a,h)}\left(V^{\pi_t}_{h+1}(s')\right)}\sqrt{\sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2 L}\sqrt{\sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}}$$

$$+ 9HL \sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)} + \underbrace{4\sum_{h,s,a} q^{\boldsymbol{p},\pi_t}_{s,a,h} \sqrt{\frac{\mathbb{E}_{s'\sim\hat{p}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}}}_{\text{Term (I.a)}} \tag{176}$$

$$\overset{(d)}{\le} 2\sqrt{H^2 L}\sqrt{\sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2 L}\sqrt{\sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}}$$

$$+ 9HL \sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)} + 100\sqrt{H^4 SL^3} \sum_{h,s,a} \frac{q^{\boldsymbol{p},\pi_t}_{s,a,h}}{N_{t-1}(s,a,h)} \tag{177}$$

where inequality (a) is due to the definition of confidence region function $\mathcal{C}_t(\pi)$ in Eq. (12), inequality (b) is due to Lemma 9, inequality (c) is due to Lemma 15, inequality (d) is due to Lemma 16 and the Term (I.a) bounded as follows. Before we prove the Term (I.a), we can see from Eq. (177) (and compared with Eq. (32)) that we equivalently have $\bar{F} = 4H^2 L$ and the additionally produced second, third, and fourth term in Eq. (177) can be merged together as the $\bar{I}$ term. For Term (I.a) we

have,

$$\text{Term (I.a)} = \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \sqrt{\frac{\mathbb{E}_{s'\sim\hat{\boldsymbol{p}}_{t-1}(s,a,h)}\left[\bar{V}_{t,h+1}(s') - \underline{V}_{t,h+1}(s')\right]^2 L}{N_{t-1}(s,a,h)}} \tag{178}$$

$$\leq \sqrt{L}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\hat{\boldsymbol{p}}_{t-1}(s,a,h)^\top(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})^2} \tag{179}$$

$$\leq \sqrt{L}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \left(\sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\boldsymbol{p}(s,a,h)^\top(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})^2}\right.$$
$$\left.+ \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}|\boldsymbol{p}(s,a,h) - \hat{\boldsymbol{p}}_{t-1}(s,a,h)|^\top(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})^2}\right) \tag{180}$$

$$\leq \sqrt{L}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \left(\sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\boldsymbol{p}(s,a,h)^\top(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})^2}\right.$$
$$\left.+ \sqrt{H}\sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}|\boldsymbol{p}(s,a,h) - \hat{\boldsymbol{p}}_{t-1}(s,a,h)|^\top(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1})}\right) \tag{181}$$

$$\overset{(a)}{\leq} \sqrt{L}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \left(\sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\frac{400H^4L^2S}{N_{t-1}(s,a,h)}} + \sqrt{H}\sqrt{21\sqrt{H^4S^2L^3}\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}}\right) \tag{182}$$

$$\leq 20\sqrt{H^4SL^3}\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} + 5\sqrt{H^3SL^{2.5}}\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{183}$$

$$\leq 25\sqrt{H^4SL^3}\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}, \tag{184}$$

where inequality (a) is due to Lemma 14 and the Term (III) bounded by Eq. (199). Intuitively, term

Step 3: Bound the optimal future value regret - Term (II)

Step 3 can be proved using Step 2 above since $\left[(\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}_{h+1}^*\right] \leq \phi_t(s,a,h)$ due to Lemma 11. But we can have a tighter bound as follows:

$$\text{Term (II)} = \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\left|(\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \boldsymbol{V}_{h+1}^*\right| \tag{185}$$

$$\overset{(a)}{\leq} 2\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\sqrt{\frac{\text{Var}_{s'\sim\boldsymbol{p}(s,a,h)}\left(V_{h+1}^*(s')\right)L}{N_{t-1}(s,a,h)}} + \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\frac{HL}{N_{t-1}(s,a,h)} \tag{186}$$

$$\overset{(b)}{\leq} 2\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t}\sqrt{\frac{\text{Var}_{s'\sim\boldsymbol{p}(s,a,h)}\left(V_{h+1}^{\pi_t}(s')\right)L}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2L}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}\cdot\Delta_{\pi_t}}{N_{t-1}(s,a,h)}} + HL\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{187}$$

$$\leq 2\sqrt{L}\sqrt{\sum_{h,s,a} p_{s,a,h}^{\boldsymbol{p},\pi_t}\text{Var}_{s'\sim\boldsymbol{p}(s,a,h)}\left(V_{h+1}^{\pi_t}(s')\right)}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2L}\sqrt{\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}\cdot\Delta_{\pi_t}}{N_{t-1}(s,a,h)}}$$
$$+ HL\sum_{h,s,a}\frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{188}$$

$$\overset{(c)}{\le} 2\sqrt{H^2L}\sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2L}\sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}} + HL \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{189}$$

where inequality (a) is due to Lemma 8, inequality (b) is due to the Lemma 15, and inequality (c) is due to Lemma 16.

Step 4: Bound the lower-order regret - Term (III)

For Term (III), we can identify $G_{t,s,a,h} = \sum_{s'} \left(\sqrt{p(s'|s,a,h)(1-p(s'|s,a,h))L} \right) \left(\bar{V}_{t,h+1}(s') - V_{h+1}^*(s') \right)$, $J_{t,s,a,h} = HSL$ as in Eq. (194). Then we can show that $\sum_{s,a,h} q_{s,a,h}^{\boldsymbol{p},\pi_t} G_{t,s,a,h}^2$ will produce lower order terms $G' = O(H^4 S^2 L^3 \sum_{s,a,h} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)})$ so that $\sqrt{G'}\sqrt{\sum_{s,a,h} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}}$ can be merged into $\bar{J}$. Therefore it is equivalent to have $\bar{G} = 0$, $\bar{J} = 21\sqrt{H^4 S^2 L^3}$ as follows. Concretely, we have

$$\text{Term (III)} = \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \left| (\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \left(\boldsymbol{V}_{t,h+1}^{\tilde{\boldsymbol{p}}_t,\pi_t} - \boldsymbol{V}_{h+1}^* \right) \right| \tag{190}$$

$$= \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \left| (\hat{\boldsymbol{p}}_{t-1}(s,a,h) - \boldsymbol{p}(s,a,h))^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \boldsymbol{V}_{h+1}^* \right) \right| \tag{191}$$

$$\le \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \sum_{s'} |\hat{p}_{t-1}(s'|s,a,h) - p(s'|s,a,h)| \left(\bar{V}_{t,h+1}(s') - V_{h+1}^*(s') \right) \tag{192}$$

$$\overset{(a)}{\le} \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \sum_{s'} \left(\sqrt{\frac{p(s'|s,a,h)(1-p(s'|s,a,h))L}{N_{t-1}(s,a,h)}} + \frac{L}{N_{t-1}(s,a,h)} \right) \left(\bar{V}_{t,h+1}(s') - V_{h+1}^*(s') \right) \tag{193}$$

$$\overset{(b)}{\le} \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \sum_{s'} \sqrt{\frac{p(s'|s,a,h)(1-p(s'|s,a,h))L}{N_{t-1}(s,a,h)}} \left(\bar{V}_{t,h+1}(s') - V_{h+1}^*(s') \right) + \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \frac{HSL}{N_{t-1}(s,a,h)} \tag{194}$$

$$\overset{(c)}{\le} \sqrt{SL} \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \sqrt{\frac{\sum_{s'} p(s'|s,a,h) \left(\bar{V}_{t,h+1}(s') - V_{h+1}^*(s') \right)^2}{N_{t-1}(s,a,h)}} + \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \frac{HSL}{N_{t-1}(s,a,h)} \tag{195}$$

$$\le \sqrt{SL} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \boldsymbol{p}(s,a,h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \boldsymbol{V}_{h+1}^* \right)^2} + \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \frac{HSL}{N_{t-1}(s,a,h)} \tag{196}$$

$$\le \sqrt{SL} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \boldsymbol{p}(s,a,h)^\top \left(\bar{\boldsymbol{V}}_{t,h+1} - \underline{\boldsymbol{V}}_{t,h+1} \right)^2} + \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \frac{HSL}{N_{t-1}(s,a,h)} \tag{197}$$

$$\overset{(d)}{\le} \sqrt{SL} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \sqrt{\sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \frac{400 H^4 L^2 S}{N_{t-1}(s,a,h)}} + \sum_{h,s,a} q_{s,a,h}^{\boldsymbol{p},\pi_t} \frac{HSL}{N_{t-1}(s,a,h)} \tag{198}$$

$$\le 21\sqrt{H^4 S^2 L^3} \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{199}$$

where inequality (a) is due to Lemma 7, inequality (b) is due to $\bar{V}_{t,h+1}(s') - V_{h+1}^*(s') \le H$, inequality (c) is due to Cauchy-Schwarz inequality, and inequality (d) is due to Lemma 14.

Step 5: Putting all together and using CMAB-MT techniques in Appendix B

Using Eq. (177), Eq. (189) and Eq. (199), we have

$$\Delta_{\pi_t} \le 2\sqrt{H^2L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2L} \sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}} + 9HL \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}$$

$$+ 100\sqrt{H^4SL^3} \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} + 2\sqrt{H^2L}\sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} + 2\sqrt{H^2L}\sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}}$$

$$+ HL \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} + 21\sqrt{H^4S^2L^3} \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{200}$$

$$\leq 4\sqrt{H^2L}\sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} + 4\sqrt{H^2L}\sqrt{\sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t} \cdot \Delta_{\pi_t}}{N_{t-1}(s,a,h)}} + 131\sqrt{H^4S^2L^3} \sum_{h,s,a} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{201}$$

Let $c_1 = 4 \times 4\sqrt{H^2L}, c_2 = 4 \times 4\sqrt{H^2L}, c_3 = 4 \times 131\sqrt{H^4S^2L^3}, c_4 = 4 \times 2H$. We define the four decomposed events as follows.

$$\mathcal{E}'_{t,1} = \left\{ \Delta_{\pi_t} \leq c_1 \sqrt{\sum_{s,a,h:N_{t-1}(s,a,h)>0} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)}} \right\}, \mathcal{E}'_{t,2} = \left\{ \Delta_{\pi_t} \leq c_2 \sqrt{\sum_{s,a,h:N_{t-1}(s,a,h)>0} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}\Delta_{\pi_t}}{N_{t-1}(s,a,h)}} \right\}, \tag{202}$$

$$\mathcal{E}'_{t,3} = \left\{ \Delta_{\pi_t} \leq c_3 \sum_{s,a,h:N_{t-1}(s,a,h)>0} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \right\}, \mathcal{E}'_{t,3} = \left\{ \Delta_{\pi_t} \leq c_4 \sum_{s,a,h:N_{t-1}(s,a,h)=0} q_{s,a,h}^{\boldsymbol{p},\pi_t} \cdot 2H \right\}. \tag{203}$$

By Lemma 4 and Eq. (201), we have

$$\text{Reg}(T,\mathcal{E}) \leq \sum_{i=1}^4 \text{Reg}(T,\mathcal{E}'_{t,i}) \tag{204}$$

Regarding $\text{Reg}(T,\mathcal{E}'_{t,i})$ for $i = 1,3,4$, we can apply similar analysis to that of steps 2,3,4 in Appendix B.3 respectively. For $\text{Reg}(T,\mathcal{E}'_{t,2})$, if $\mathcal{E}'_{t,2}$ holds, then we have

$$\Delta_{\pi_t} \leq c_2^2 \sum_{s,a,h:N_{t-1}(s,a,h)>0} \frac{q_{s,a,h}^{\boldsymbol{p},\pi_t}}{N_{t-1}(s,a,h)} \tag{205}$$

which can be bounded exactly the same way as $\text{Reg}(T,\mathcal{E}'_{t,3})$.

Using a similar analysis to Appendix B.3, we have

$$\text{Reg}(T,\mathcal{E}_t) \leq \sum_{i\in[m]} \frac{2c_1^2}{\Delta_i^{\min}}(3 + \log K) + \sum_{s,a,h} 2(c_2^2 + c_3)\left(1 + \log\left(\frac{2(c_2^2 + c_3)K}{\Delta_i^{\min}}\right)\right) + c_4 m \tag{206}$$

$$= \sum_{s,a,h} \frac{512H^2L}{\Delta_{s,a,h}^{\min}}(3 + \log H) + \sum_{s,a,h} 1560H^2SL^{1.5}\left(1 + \log\left(\frac{1560H^3SL^{1.5}}{\Delta_{s,a,h}^{\min}}\right)\right) + 8SAH^2 \tag{207}$$

$$= O\left(\sum_{s,a,h} \frac{H^2L}{\Delta_{s,a,h}^{\min}} + \sum_{s,a,h} H^2SL^{1.5}\log\left(\frac{1}{\Delta_{s,a,h}^{\min}}\right)\right) \tag{208}$$

Similar to the analysis of Appendix B.3, the gap-independent regret bound is $\tilde{O}(\sqrt{H^3SAT} + H^3S^2A)$ when considering the inhomogeneous episodic RL setting.

D.5. Discussion about Gap-Dependent Regret Bound

In this section, we discuss the tightness of our gap-dependent bound in Eq. (206). Since we use a different definition of the gap, it is not directly comparable to the existing works such as (Simchowitz & Jamieson, 2019). Here we specify

the value of $\Delta_{s,a,h}^{\min} = \min_{\pi \in \Pi : q_{s,a,h}^{p,\pi} > 0, V_1^*(s_1) - V_1^\pi(s_1) > 0}(V_1^*(s_1) - V_1^\pi(s_1))$, where we will omit the underlying transition probabilities p for V and Q functions. Similar to Simchowitz & Jamieson (2019), we first divide (s, a, h) into two parts: $\mathcal{Z}_{sub} = \{(s, a, h) : \pi^*(s, h) \neq a\}$ and $\mathcal{Z}_{opt} = \mathcal{S} \times \mathcal{A} \times [H] - \mathcal{Z}_{sub}$. We use $\text{gap}(s, a, h) = V_h^*(s) - Q_h^*(s, a)$ to denote the state-dependent suboptimality gap, and $\text{gap}_{\min} = \min_{s,a,h}\{\text{gap}(s, a, h) : \text{gap}(s, a, h) > 0\}$ the minimum gap. Let $q^* = \min_{\pi,(s,a,h)}\{q_{s,a,h}^{p,\pi} : q_{s,a,h}^{p,\pi} > 0\}$ be the minimum occupancy measure for any policy π and state-action-step pair (s, a, h).

We use the following performance difference lemma for episodic MDP as follows, which is slightly different from Lemma 1.16 for infinite horizon discounted MDP in Agarwal et al. (2019):

Lemma 17 (Performance difference lemma for episodic MDP). *For any MDP with transition kernel p and for any two policies π and π', the difference of their value function starting from the initial state s_1 can be bounded by*

$$V_1^\pi(s_1) - V_1^{\pi'}(s_1) = \sum_{s,a,h} q_{s,a,h}^{p,\pi'} [V_h^\pi(s) - Q_h^\pi(s, a)] \tag{209}$$

Proof. Let $q_{s,h}^{p,\pi}$ be the probability of visiting state s at step h following policy π.

$$V_1^\pi(s_1) - V_1^{\pi'}(s_1) = V_1^\pi(s_1) - \sum_{s,a,h} q_{s,a,h}^{p,\pi'} r(s, a, h) \tag{210}$$

$$\overset{(a)}{=} V_1^\pi(s_1) - \sum_{s,a,h} q_{s,a,h}^{p,\pi'} \left[Q_h^\pi(s, a) - \sum_{s'} p_h(s' \mid s, a, h) V_{h+1}^\pi(s') \right] \tag{211}$$

$$= V_1^\pi(s_1) + \sum_{s,a,h} \sum_{s'} q_{s,a,h}^{p,\pi'} p_h(s' \mid s, a, h) V_{h+1}^\pi(s') - \sum_{s,a,h} q_{s,a,h}^{p,\pi'} Q_h^\pi(s, a) \tag{212}$$

$$\overset{(b)}{=} V_1^\pi(s_1) + \sum_{s',h} q_{s',h+1}^{p,\pi'} V_{h+1}^\pi(s') - \sum_{s,a,h} q_{s,a,h}^{p,\pi'} Q_h^\pi(s, a) \tag{213}$$

$$\overset{(c)}{=} \sum_{s',h} q_{s',h}^{p,\pi'} V_h^\pi(s') - \sum_{s,a,h} q_{s,a,h}^{p,\pi'} Q_h^\pi(s, a) \tag{214}$$

$$= \sum_{s,a,h} q_{s,a,h}^{p,\pi'} [V_h^\pi(s) - Q_h^\pi(s, a)] \tag{215}$$

■

where equality (a) is due to the Bellman equation $Q_h^\pi(s, a) = r(s, a, h) + \sum_{s'} p(s' \mid s, a, h) V_{h+1}^\pi(s')$, equality (b) is due to $\sum_{s,a} q_{s,a,h}^{p,\pi} p(s' \mid s, a, h) = q_{s',h+1}^{p,\pi}$, and equality (c) is due to $q_{s_1,1}^{p,\pi} = 1$ and $q_{s,1}^{p,\pi} = 0$ for $s \neq s_1$.

Thus $V_1^*(s_1) - V_1^\pi(s_1) = \sum_{s,a,h} q_{s,a,h}^{p,\pi} \text{gap}(s, a, h)$, and for $(s, a, h) \in \mathcal{Z}_{sub}$, we have $\Delta_{s,a,h}^{\min} = \min_{\pi \in \Pi : q_{s,a,h}^{p,\pi} > 0, V_1^*(s_1) - V_1^\pi(s_1) > 0}(V_1^*(s_1) - V_1^\pi(s_1)) \geq q^* \cdot \text{gap}(s, a, h)$. For $(s, a, h) \in \mathcal{Z}_{opt}$, we have $\Delta_{s,a,h}^{\min} \geq \text{gap}_{\min}$ since in the worst case, π allocates all the triggering probability $q_{s,a,h}^{p,\pi}$ to the (s, a, h) that attains $\text{gap}_{\min}$. Now based on Eq. (206) and the above reasoning, we have

$$\text{Reg}(T) \leq O\left(\sum_{(s,a,h) \in \mathcal{Z}_{sub}} \frac{H^2}{q^* \cdot \text{gap}(s, a, h)} \log(SAHT) + \frac{H^2 |\mathcal{Z}_{opt}|}{\text{gap}_{\min}} \log(SAHT) \right.$$

$$\left. + H^3 S^2 A \log^{1.5}(SAHT) \log\left(\frac{1}{\text{gap}_{\min}} \log^{1.5}(SAHT) \right) \right) \tag{216}$$

which matches the regret bound of (Simchowitz & Jamieson, 2019) up to a factor of $1/q^*$.

E. Analysis for PMC-GD in Section 5

E.1. Proof of Lemma 3

$$|r(\pi; \tilde{\boldsymbol{p}}) - r(\pi; \boldsymbol{p})| = \left| \sum_{v \in V} \left[\left(1 - \prod_{u \in \pi} (1 - \tilde{p}(u, v)) \right) - \left(1 - \prod_{u \in \pi} (1 - p(u, v)) \right) \right] \right| \tag{217}$$

$$= \left| \sum_{v \in V} \left[\prod_{u \in \pi} (1 - p(u, v)) - \prod_{u \in \pi} (1 - \tilde{p}(u, v)) \right] \right| \tag{218}$$

$$\leq \sum_{v \in V} \left| \prod_{u \in \pi} (1 - p(u, v)) - \prod_{u \in \pi} (1 - \tilde{p}(u, v)) \right| \tag{219}$$

$$\overset{(a)}{\leq} \sum_{u \in \pi} \| \tilde{\boldsymbol{p}}(u, \cdot) - \boldsymbol{p}(u, \cdot) \|_1 \tag{220}$$

where inequality (a) is due to the fact that let $(a_1, ..., a_{|\pi|}) \overset{\text{def}}{=} (1 - p(u, v))_{u \in \pi}$, $(b_1, ..., b_{|\pi|}) \overset{\text{def}}{=} (1 - \tilde{p}(u, v))_{u \in \pi}$, $\left| \prod_{i=1}^{|\pi|} a_i - \prod_{i=1}^{|\pi|} b_i \right| = \left| \sum_{i=1}^{|\pi|} \prod_{j=1}^{i-1} a_j \cdot (a_i - b_i) \cdot \prod_{k=i+1}^{|\pi|} b_k \right| \leq \sum_{i=1}^{|\pi|} |a_i - b_i|$.

For the pseudo-reward function $\bar{r}_t(\pi; \tilde{\boldsymbol{p}}) = r(\pi; \hat{\boldsymbol{p}}_{t-1}) + \sum_{u \in \pi} \| \tilde{\boldsymbol{p}}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot) \|_1$, we also have for all $\pi, \boldsymbol{p}, \tilde{\boldsymbol{p}}$, it holds

$$|\bar{r}_t(\pi; \tilde{\boldsymbol{p}}) - \bar{r}_t(\pi; \boldsymbol{p})| = \left| \sum_{u \in \pi} \| \tilde{\boldsymbol{p}}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot) \|_1 - \| \boldsymbol{p}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot) \|_1 \right| \tag{221}$$

$$\leq \sum_{u \in \pi} \left| \| \tilde{\boldsymbol{p}}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot) \|_1 - \| \boldsymbol{p}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot) \|_1 \right| \tag{222}$$

$$\overset{(a)}{\leq} \sum_{u \in \pi} \| \tilde{\boldsymbol{p}}(u, \cdot) - \boldsymbol{p}(u, \cdot) \|_1 \tag{223}$$

where inequality (a) is due to $- \| \boldsymbol{x} - \boldsymbol{y} \|_1 \leq \| \boldsymbol{x} \|_1 - \| \boldsymbol{y} \|_1 \leq \| \boldsymbol{x} - \boldsymbol{y} \|_1$ by triangle inequality.

E.2. Proof of Theorem 4

We define the concentration event as:

$$\mathcal{E} \overset{\text{def}}{=} \left[\| \hat{\boldsymbol{p}}_{t-1}(u, \cdot) - \boldsymbol{p}(u, \cdot) \|_1 \leq \sqrt{\frac{2|V| \log \left(\frac{|U||V|T}{\delta'} \right)}{N_{t-1, u}}}, \text{ for any } u \in U, t \in [T] \right] \tag{224}$$

Suppose the concentration event $\mathcal{E}$ holds with probability $\delta' = 1/(2T)$ as in Lemma 7. Let $\alpha = 1 - 1/e$ and $L = \log(|U||V|T)$. Also we can initialize each counter by $N_{t_0, u} = 1$ using $t_0 = |U|$ rounds which pays an extra $O(k|U|)$ regret. Now we have

$$\Delta_{\pi_t} = \alpha \cdot r(\pi^*; \boldsymbol{p}) - r(\pi_t; \boldsymbol{p}) \tag{225}$$

$$\overset{(a)}{\leq} \alpha \cdot \bar{r}_t(\pi^*; \boldsymbol{p}) - r(\pi_t; \boldsymbol{p}) \tag{226}$$

$$\overset{(b)}{\leq} \alpha \bar{r}_t(\pi^*; \tilde{\boldsymbol{p}}_t) - r(\pi_t; \boldsymbol{p}) \tag{227}$$

$$\overset{(c)}{\leq} \bar{r}_t(\pi_t; \tilde{\boldsymbol{p}}_t) - r(\pi_t; \boldsymbol{p}) \tag{228}$$

$$= \bar{r}_t(\pi_t; \tilde{\boldsymbol{p}}_t) - \bar{r}_t(\pi_t; \boldsymbol{p}) + \bar{r}_t(\pi_t; \boldsymbol{p}) - r(\pi_t; \boldsymbol{p}) \tag{229}$$

$$\overset{(d)}{\leq} \sum_{u \in \pi_t} \| \tilde{\boldsymbol{p}}_t(u, \cdot) - \boldsymbol{p}(u, \cdot) \|_1 + \bar{r}_t(\pi_t; \boldsymbol{p}) - r(\pi_t; \boldsymbol{p}) \tag{230}$$

$$\overset{(e)}{\leq} 4 \sum_{u \in \pi_t} \sqrt{\frac{|V|L}{N_{t-1,u}}} + \underbrace{\bar{r}_t(\pi_t; \boldsymbol{p}) - r(\pi_t; \boldsymbol{p})}_{\text{Additional Term}} \tag{231}$$

$$\overset{(f)}{\leq} 8 \sum_{u \in \pi_t} \sqrt{\frac{|V|L}{N_{t-1,u}}}, \tag{232}$$

where inequality (a) is due to $\bar{r}_t(\pi; \boldsymbol{p}) \geq r(\pi; \boldsymbol{p})$ for any $\pi, \boldsymbol{p}$ by Lemma 3, inequality (b) is due to the definition of $\tilde{\boldsymbol{p}}_t$ in Algorithm 4, inequality (c) is due to π_t is a $(1 - 1/e, 1)$-approximate solution to the problem $\operatorname{argmax}_{|\pi| \leq k} \bar{r}_t(\pi; \tilde{\boldsymbol{p}}_t)$, inequality (d) is by Eq. (223), inequality (e) is due to Eq. (224) and Eq. (13), and inequality (f) is due to the following inequality to deal with additional regret term brought by pseudo-reward $\bar{r}_t(\pi; \boldsymbol{p})$.

$$\text{Additional Term} = \bar{r}_t(\pi_t; \boldsymbol{p}) - r(\pi_t; \boldsymbol{p}) \tag{233}$$

$$= \sum_{u \in \pi_t} \|\boldsymbol{p}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot)\|_1 + r(\pi_t; \hat{\boldsymbol{p}}_{t-1}) - r(\pi_t; \boldsymbol{p}) \tag{234}$$

$$\overset{(a)}{\leq} 2 \sum_{u \in \pi_t} \|\boldsymbol{p}(u, \cdot) - \hat{\boldsymbol{p}}_{t-1}(u, \cdot)\|_1 \tag{235}$$

$$\overset{(b)}{\leq} 4 \sum_{u \in \pi_t} \sqrt{\frac{|V|L}{N_{t-1,u}}} \tag{236}$$

$$\tag{237}$$

where inequality (a) is due to Lemma 3, and inequality (b) is due to event $\mathcal{E}$.

Compared with Eq. (30), it is equivalent to have $F_{t,u} = 8\sqrt{|V|L}, G_{t,u} = I_{t,u} = J_{t,u} = 0$ and following step 2 in Appendix B.3 where $\bar{F} = 64k|V|L, \bar{G} = \bar{I} = \bar{J} = 0$, we have gap-dependent regret

$$Reg(T) = O\left(\sum_{u \in U} \frac{k|V| \log(|U||V|T)}{\Delta_u^{\min}}\right) \tag{238}$$

and gap-independent regret

$$Reg(T) = O\left(\sqrt{k|V||U|T \log(|U||V|T)}\right). \tag{239}$$