Online Learning for Uninformed Markov Games: Empirical Nash-Value Regret and Non-Stationarity Adaptation

Junyan Liu JUNYANL1@CS.WASHINGTON.EDU
University of Washington

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

Zihan Zhang ZIHANZ@UST.HK
Hong Kong University of Science and Technology

Lillian J. Ratliff RATLIFFL@UW.EDU
University of Washington

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study online learning in two-player uninformed Markov games, where the opponent's actions and policies are unobserved. In this setting, Tian et al. (2021) show that achieving no-external-regret is impossible without incurring an exponential dependence on the episode length H. They then turn to the weaker notion of Nash-value regret and propose a V-learning algorithm with regret $\widetilde{\mathcal{O}}(K^{2/3})$ after K episodes. However, their algorithm and guarantee do not adapt to the difficulty of the problem: even in the case where the opponent follows a fixed policy and thus $\widetilde{\mathcal{O}}(\sqrt{K})$ external regret is well-known to be achievable, their result is still the *worse* rate $\widetilde{\mathcal{O}}(K^{2/3})$ on a *weaker* metric.

In this work, we fully address both limitations. First, we introduce *empirical Nash-value regret*, a new regret notion that is strictly stronger than Nash-value regret and naturally reduces to external regret when the opponent follows a fixed policy. Moreover, under this new metric, we propose a parameter-free algorithm that achieves an $\widetilde{\mathcal{O}}\big(\min\{\sqrt{K}+(CK)^{1/3}, \sqrt{LK}\}\big)$ regret bound, where C quantifies the "variance" of the opponent's policies and L denotes the number of policy switches (both at most $\mathcal{O}(K)$). Therefore, our results not only recover the two extremes—$\widetilde{\mathcal{O}}(\sqrt{K})$ external regret when the opponent is fixed and $\widetilde{\mathcal{O}}(K^{2/3})$ Nash-value regret in the worst case—but also smoothly interpolate between these extremes by automatically adapting to the opponent's non-stationarity. We achieve so by first providing a new analysis of the epoch-based V-learning algorithm by Mao et al. (2022), establishing an $\widetilde{\mathcal{O}}(\eta C + \sqrt{K/\eta})$ regret bound, where η is the epoch incremental factor. Next, we show how to adaptively restart this algorithm with an appropriate η in response to the potential non-stationarity of the opponent, eventually achieving our final results.

Keywords: uninformed Markov games, empirical Nash-value regret, non-stationarity adaptation.

1. Introduction

Multi-agent reinforcement learning (MARL), often modeled as a Markov game (MG), provides a general framework for studying sequential decision-making problems involving multiple strategic agents whose actions jointly influence a shared environment. Recent advances have demonstrated the empirical success of MARL in domains with complex strategic interactions, including the game of Go (Silver et al., 2016, 2017), Poker (Brown and Sandholm, 2019), large-scale video games (Vinyals et al., 2019), and autonomous driving (Shalev-Shwartz et al., 2016; Zhou et al., 2021).

A substantial body of prior work studies MGs in the self-play setting, where all players follow the same learning algorithm to optimize their joint behavior, typically with the goal of minimizing

the number of episodes required to identify a good joint policy (Bai and Jin, 2020; Bai et al., 2020; Liu et al., 2021; Mao et al., 2022; Jin et al., 2024). While this formulation has led to significant algorithmic progress, it abstracts away challenges that arise in many practical multi-agent systems. In such settings, an agent may repeatedly interact with opponents whose learning rules, objectives, or update schedules are unknown and not aligned with its own. This motivates the study of learning Markov games with *arbitrary opponents* (Xie et al., 2020; Tian et al., 2021), where an agent must balance exploration of unknown environments with strategic decision-making against opponents that may follow arbitrary, potentially history-dependent policies.

Depending on the observation model, MGs with arbitrary opponents can be divided into the informed setting, where the opponent's actions are observable, and the uninformed setting, where the opponent's actions are unobserved. While near-optimal regret bounds have been established for the informed setting (Xie et al., 2020; Tian et al., 2021), learning in uninformed MGs is significantly more challenging, since the lack of access to the opponent's actions prevents explicitly learning the transition model or maintaining a table of state-action value (Q-value table). Indeed, Tian et al. (2021) show that an exponential dependence on the episode length H is unavoidable when performance is measured by the standard external regret. As a remedy, Tian et al. (2021) consider a weaker performance measure that compares the algorithm's total reward to the Nash value of the MG, known as Nash-value regret. They then show that a variant of the V-learning algorithm (Bai et al., 2020) with an appropriate choice of parameters achieves an $\widetilde{\mathcal{O}}(K^{2/3})$ Nash-value regret bound after K episodes. This remains the best result in this setting, with the current best lower bound being $\Omega(\sqrt{K})$.

However, while the tightness of their result in the worst case remains unknown, it is certainly suboptimal in special cases. For example, when the opponent follows a fixed policy, then the learner is simply facing a fixed environment, in which case $\Theta(\sqrt{K})$ *external* regret is well-known (Azar et al., 2017); on the other hand, the result of Tian et al. (2021) for this case not only is about the weaker Nash-value regret (see Remark 7), but also has a worse rate of $\widetilde{\mathcal{O}}(K^{2/3})$. Tian et al. (2021) leave it as an open question whether this limitation is fundamental or merely an artifact of the analysis, but more generally, this begs the following natural questions:

1. Is there a better regret notion that naturally interpolates between standard external regret and Nash-value regret as the opponent's non-stationarity increases?

2. Correspondingly, for such adaptive regret notions, are there efficient algorithms whose regret smoothly interpolates between $\widetilde{\mathcal{O}}(\sqrt{K})$ and $\widetilde{\mathcal{O}}(K^{2/3})$ as the opponent's non-stationarity increases?

Contributions. In this paper, we provide affirmative answers to both questions. Specifically, our main contributions are summarized as follows.

- For two-player uninformed MGs, we introduce a new regret notion termed *empirical Nash-value regret* (ENR), which is the Nash value for a game where at each state the opponent is restricted to the policies that they have played over the K episodes. Consequently, ENR not only is strictly stronger than the Nash-value regret (NR) considered in prior work, but also reduces to the standard external regret when the opponent follows a fixed policy.

- Towards achieving our second goal, we start by introducing a novel analysis of the epoch-based V-learning algorithm of Mao et al. (2022), which turns out to be more manageable than the original V-learning algorithm (Bai et al., 2020; Tian et al., 2021). Note that Mao et al. (2022) analyze this algorithm under the self-play setting, which is very different from our analysis that deals with an arbitrary opponent. Specifically, we establish an $\widetilde{\mathcal{O}}(\eta C + \sqrt{K/\eta})$ regret bound under

the new metric ENR, where $\eta \in (0, 1/H]$ is the epoch incremental factor and C is a certain non-stationarity measure that quantifies the "variance" of the opponent's policies (see Eq. (7) for its formal definition). When the opponent uses a single fixed policy, we have $C = 0$, and thus choosing $\eta = 1/H$ yields an $\widetilde{\mathcal{O}}(\sqrt{K})$ external regret bound. This result is the first to show that V-learning–type algorithms can achieve an $\widetilde{\mathcal{O}}(\sqrt{K})$ external regret bound in a stationary environment, answering the open question raised by Tian et al. (2021). More generally, setting $\eta = \min\{1/H, K^{1/3}C^{-2/3}\}$ leads to a bound of order $\widetilde{\mathcal{O}}(\sqrt{K} + (CK)^{1/3})$. The issue is of course that this tuning requires the knowledge of C.

- To address the limitation that epoch-based V-learning cannot automatically adapt to unknown non-stationarity, we further propose a meta-algorithm that repeatedly restarts epoch-based V-learning in response to potential non-stationarity in the environment. In addition, motivated by the observation that the environment is nearly stationary when the opponent switches policies infrequently (yet C can be as large as $\Omega(T)$ in this case), we further equip the meta-algorithm with a mechanism to detect such changes. As a result, the final (parameter-free) algorithm achieves a regret bound of $\widetilde{\mathcal{O}}(\min\{\sqrt{K} + (CK)^{1/3}, \sqrt{LK}\})$ for ENR, where L denotes the total number of policy switches by the opponent. This shows that our algorithm automatically adapts to both non-stationarity measures (C and L) and enjoys a bound on ENR that smoothly interpolates between $\mathcal{O}(\sqrt{K})$ and $\widetilde{\mathcal{O}}(K^{2/3})$ as C and L increase, completely resolving our second question.

Related work. Markov games, also known as stochastic games (Shapley, 1953), are a fundamental framework in MARL. Early work primarily established asymptotic convergence to Nash equilibrium under the assumption that the transition dynamics and rewards are known (Littman et al., 2001; Hu and Wellman, 2003; Hansen et al., 2013). More recently, a growing literature has developed non-asymptotic guarantees for MGs without additional structural assumptions (Wei et al., 2017; Sidford et al., 2020; Bai and Jin, 2020; Xie et al., 2020; Liu et al., 2021; Tian et al., 2021; Mao et al., 2022). In the self-play setting, where all players deploy the same (or symmetric) learning algorithms, Sidford et al. (2020); Bai and Jin (2020); Liu et al. (2021); Mao et al. (2022) study the sample complexity of computing an ϵ-approximate Nash equilibrium.

Our work is most closely related to Wei et al. (2017); Xie et al. (2020); Tian et al. (2021), which study Markov games with arbitrary opponents. In particular, Wei et al. (2017); Xie et al. (2020) consider settings in which the opponent's actions are observable, and propose algorithms that achieve $\sqrt{K}$-order regret bounds. In contrast, Tian et al. (2021) study the more challenging uninformed setting, where the opponent's actions are unobserved. They show that external regret admits a lower bound of $\Omega\big(\min\sqrt{2^H K}, K\big)$, which motivates the study of a weaker regret notion, NR. Building on Bai et al. (2020), they propose a V-learning–type algorithm that achieves $\widetilde{\mathcal{O}}(K^{2/3})$ regret under NR. Given the impossibility result for external regret in the uninformed setting, Liu et al. (2022) investigate what additional assumptions are necessary to recover $\sqrt{K}$-order external regret bounds.

In the single-agent RL literature, a line of work also studies how to adapt to the non-stationarity of a sequence of changing Markov Decision Processes (MDPs). For example, Wei and Luo (2021) studied the strong notion of dynamic regret (that competes to the best policy of each MDP) and proposed a black-box reduction approach to adapt to unknown non-stationarity. Jin et al. (2023) focused on external regret and developed an algorithm that adapts to the unknown non-stationarity of the transitions (while allowing rewards to be arbitrary). Since playing a Markov game with an arbitrary opponent can be viewed as a single agent interacting with a sequence of changing MDPs,

one can apply those algorithms to our problem. However, unlike our results, the resulting bounds are all necessarily linear in K in the worst case since the regret notions they consider are stronger.

Finally, we point out that a recent work by Appel and Kosoy (2025) obtains a $\widetilde{\mathcal{O}}(\sqrt{K})$ guarantee on NR for *turn-based* Markov games, which we emphasize does *not* translate to a same bound for our problem. This is because in their definition of NR, the Nash value is with respect to a game where at each state the min-player can make a decision based on the max-player's current action. This makes their Nash value smaller than ours and thus their regret measure weaker than those considered by Tian et al. (2021) and our work. To the best of our knowledge, whether the worst-case $\widetilde{\mathcal{O}}(K^{2/3})$ rate for our setting can be improved remains open.

2. Preliminaries

We consider an episodic two-player Markov game (MG) with finite state and action spaces.[1] Let $\Delta(X)$ denote the set of probability distributions over a finite set X. Such an MG is specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, H)$, where

- H is the number of steps in an episode;
- $S = \cup_{h\in[H+1]} S_h$ is the state space, where there is a single terminal state in S_{H+1};
- $A = \cup_{h\in[H]} A_h$ and $B = \cup_{h\in[H]} B_h$ are the action space for the max-player (learner) and the min-player (opponent) respectively;
- P is a collection of unknown transition functions $\{P_h : S_h \times A_h \times B_h \to \Delta(S_{h+1})\}_{h\in[H]}$.
- r is the collection of reward functions $\{r_h : S_h \times A_h \times B_h \to [0,1]\}_{h\in[H]}$.

In each episode $k \in [K]$, the Markov game starts from an adversarially chosen initial state $s_1^k \in \mathcal{S}_1$. At the beginning of this episode, each player commits to a policy that may depend on the entire history of the previous $k-1$ episodes. We denote our policy by $\mu^k = \{\mu_h^k\}_{h\in[H]}$ and the opponent's policy by $\nu^k = \{\nu_h^k\}_{h\in[H]}$, where $\mu_h^k : \mathcal{S}_h \to \Delta(\mathcal{A}_h)$ and $\nu_h^k : \mathcal{S}_h \to \Delta(\mathcal{B}_h)$ are policies for step h. Neither player is allowed to modify its policy within an episode. At each step $h \in [H]$, both players observe the current state s_h^k and simultaneously select actions $a_h^k \in \mathcal{A}_h$ drawn from $\mu_h^k(s_h^k)$ and $b_h^k \in \mathcal{B}_h$ drawn from $\nu_h^k(s_h^k)$, respectively. They then observe reward $r_h^k := r_h(s_h^k, a_h^k, b_h^k)$, after which the environment transitions to the next state according to $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k, b_h^k)$. Importantly, at each step $h \in [H]$, the opponent's action b_h^k and policy ν_h^k are *unobservable* to the learner. On the other hand, the opponent can observe the learner's actions and is allowed to choose policies arbitrarily and adaptively as a function of the interaction history, knowing the MG and learner's algorithm (but not their randomness) ahead of time. However, for ease of exposition, we sometimes restrict our attention to an oblivious opponent whose policies cannot depend on the learner's past actions, and defer the details for a fully adaptive opponent to the appendix.

For a policy pair (μ, ν), step $h \in [H]$, state $s \in S_h$, and actions $a \in A_h, b \in B_h$, we define the standard state value function and Q-value function as follows:

$$V_h^{\mu,\nu}(s) = \mathbb{E}_{\mu,\nu}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \mid s_h = s\right],$$

1. Here, we focus on two players only for simplicity. From the learner's perspective, an m-player general-sum MG can be written as an equivalent two-player MG by grouping the other $m-1$ players into a single opponent with a joint action. See Tian et al. (2021, Section 5) for more details.

$$Q_h^{\mu,\nu}(s,a,b) = \mathbb{E}_{\mu,\nu}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'},b_{h'}) \mid s_h = s, a_h = a, b_h = b\right],$$

where $\{(s_{h'}, a_{h'}, b_{h'})\}_{h' \geq h}$ is the random trajectory generated by following μ and ν starting from $s_h = s$ for $V_h^{\mu,\nu}(s)$ or $s_h = s, a_h = a, b_h = b$ for $Q_h^{\mu,\nu}(s,a,b)$.

For shorthand, we define operators

$$P_h V(s,a,b) = \mathbb{E}_{s' \sim P_h(\cdot|s,a,b)}[V(s')], \quad \text{and} \quad \mathbb{D}_{\mu,\nu}[Q](s) = \mathbb{E}_{a\sim\mu(\cdot|s), b\sim\nu(\cdot|s)}\left[Q(s,a,b)\right].$$

Then, we have $V_h^{\mu,\nu}(s) = \mathbb{D}_{\mu_h,\nu_h}[Q^{\mu,\nu}](s)$ and $Q_h^{\mu,\nu}(s,a,b) = (r_h + P_h V_{h+1}^{\mu,\nu})(s,a,b)$. Central to this new regret notion is the following definition of *empirical state Nash values*: for each $(h,s) \in [H] \times S_h$, recursively define

$$V_h^*(s) = \max_{\mu\in\Delta(A_h)} \min_{\nu\in\left\{\nu_h^k(s)\right\}_{k\in[K]}} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s), \tag{1}$$

with $V_{H+1}^*(s) = 0$ for all $s \in S_{H+1}$. It is "empirical" since in this definition we restrict the min-player to choose from the K policies that are used by the opponent over the K episodes. If we were to relax this restriction and allow the min-player to choose any $\nu \in \Delta(B_h)$, then $V_h^*(s)$ became the exact and standard state Nash value. With this definition, our proposed *empirical Nash-value regret* (ENR) is defined as

$$\text{ENR}_K = \sum_{k=1}^{K}\left(V_1^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right), \tag{2}$$

which compares the cumulative empirical Nash-values (at initial states) and the total expected reward of the learner over K episodes.

In contrast, the Nash-value regret (denoted as NR_K) considered by Tian et al. (2021) replaces the empirical Nash values in Eq. (2) by the actual Nash values mentioned earlier. On the other hand, the standard external regret is defined as $\text{REG}_K = \max_\mu \sum_{k=1}^K \left(V_1^{\mu,\nu^k}(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right)$, comparing the total reward achieved by the best fixed policy and that of the learner. From these definitions, the following fact is straightforward.

Fact 1 *Empirical Nash-value regret is always stronger than Nash-value regret:* $\text{ENR}_K \geq \text{NR}_K$. *Moreover, when the opponent uses a fixed policy, that is,* $\nu_1 = \cdots = \nu_K$, *empirical Nash-value regret recovers external regret:* $\text{ENR}_K = \text{REG}_K$.

Consequently, the subsequent bounds we derived for ENR_K are also upper bounds on NR_K, and they also become a bound on REG_K when the opponent uses a fixed policy. More generally, ENR_K decreases as the diversity of the opponent's polices increases.

Other notations. For points $x_1, \ldots, x_n$, we use $\texttt{Conv}(x_1, \ldots, x_n)$ to denote their convex hull. For two probability distributions P, Q on the same measurable space, $\texttt{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ stands for the total variation distance. Define $\log^+(x) = \max\{\log(x), 0\}$. We also use $\mathcal{F}_h^k$ to denote the history before step h in episode k.

Algorithm 1 Epoch V-learning

Input: total episodes K, confidence $\delta \in (0,1)$, epoch incremental factor $\eta > 0$.
Initialize: $V_{H+1}^k \leftarrow 0$ for all k; $V_h^1(s) \leftarrow H-h+1, N_0(h,s) \leftarrow 1, N_1(h,s) \leftarrow 0, \mathcal{E}(h,s) \leftarrow \{1\}, \mathcal{K}_1(h,s) \leftarrow \emptyset$, bandit subroutine $\text{ADVBANDIT}_{h,s}$ for all h, s; $\pi_h^1(a|s) \leftarrow 1/|A_h|$ for all s, a, h.
for *episode* $k = 1, 2, \ldots, K$ **do**
  for *step* $h = 1, 2, \ldots, H$ **do**
    Receive state s_h^k. Set $\tau \leftarrow |\mathcal{E}(h, s_h^k)|$ to be the current epoch index for (h, s_h^k) pair.
    Take action $a_h^k \sim \pi_h^k(\cdot|s_h^k)$; observe reward r_h^k and next state s_{h+1}^k.
    Update $\mathcal{K}_\tau(h, s_h^k) \leftarrow \mathcal{K}_\tau(h, s_h^k) \cup \{k\}$ and $N_\tau(h, s_h^k) \leftarrow |\mathcal{K}_\tau(h, s_h^k)|$.
    if $N_\tau(h, s_h^k) = \lceil (1+\eta) N_{\tau-1}(h, s_h^k) \rceil$ **then** ▷ entering a new epoch
      Update $\mathcal{E}(h, s_h^k) \leftarrow \mathcal{E}(h, s_h^k) \cup \{\tau + 1\}$.
      Update $V_h^{k+1}(s_h^k) \leftarrow L_\tau(h, s_h^k)$ where $L_\tau(h,s)$ is defined in Eq. (3).
      Set $N_{\tau+1}(h, s_h^k) \leftarrow 0$, $\mathcal{K}_{\tau+1}(h, s_h^k) \leftarrow \emptyset$, and $\pi_h^{k+1}(\cdot|s_h^k) \leftarrow 1/|A_h|$.
      Restart $\text{ADVBANDIT}_{h,s}$.
    else
      Set $V_h^{k+1}(s_h^k) \leftarrow V_h^k(s_h^k)$.
      Update $\pi_h^{k+1}(\cdot|s_h^k) \leftarrow \text{ADVBANDIT}_{h,s}\big(a_h^k, \frac{1}{H}(r_h^k + V_{h+1}^k(s_{h+1}^k))\big)$.
    Set $V_h^{k+1}(s) \leftarrow V_h^k(s)$ and $\pi_h^{k+1}(\cdot|s) \leftarrow \pi_h^k(\cdot|s)$ for all $s \in S_h \backslash s_h^k$.

3. Base Algorithm: Epoch V-learning and Analysis

As mentioned, our starting point is the epoch V-learning algorithm proposed by Mao et al. (2022), which simplifies the algorithmic design of the original V-learning methods (e.g., Bai et al., 2020; Jin et al., 2024). It turns out that it also allows easier analysis for ENR_K as we show below. To introduce our novel analysis, we first review their algorithm (whose pseudocode is provided in Algorithm 1) and point out a slight change that is critical for our purpose.

Epoch schedule. At a high level, for each step-state pair $(h, s) \in [H] \times S_h$, Algorithm 1 partitions the set of all episodes where this pair is visited into epochs of geometrically increasing lengths, and within each epoch, a new adversarial bandit subroutine is used to update the policy for this (h, s) pair. More concretely, $\mathcal{E}(h, s)$ records the set of epoch indices for (h, s) pair, $\mathcal{K}_\tau(h, s)$ records the set of episodes in which (h, s) is visited during epoch τ, and $N_\tau(h, s) = |\mathcal{K}_\tau(h, s)|$ is the visitation count during that epoch. Given an epoch incremental factor $\eta > 0$, if $N_\tau(h, s) = \lceil (1+\eta) N_{\tau-1}(h, s) \rceil$, that is, the visitation count for (h, s) pair increases by a factor of $1 + \eta$, then a new epoch for this pair is created. This $(1 + \eta)$ visitation schedule provides a controlled update rule: the geometric growth keeps the number of restarts small, while the relative increase between consecutive epoch lengths is only of order η, which controls the error caused by stale value estimates. Thus, η governs the tradeoff between statistical overhead and adaptation to the opponent's non-stationarity. We point out that Mao et al. (2022) always fix η to be $1/H$, but it is important for us to dynamically tune η on the fly.

Optimistic Nash-value estimate. At the start of an epoch for (h, s), a new adversarial bandit subroutine is initialized. During the epoch, it is updated whenever state $s_h^k = s$ is visited, using the observed reward r_h^k plus a *bonus* $V_{h+1}^k(s_{h+1}^k)$, which, as we show in the analysis, serves as

an optimistic estimate of the empirical Nash value $V^*_{h+1}(s^k_{h+1})$. Note that for each (h,s), its optimistic estimate $V^k_h(s)$ stays the same within an epoch and is updated only when an epoch for (h,s) ends. Concretely, if the τ-th epoch for pair (h,s^k_h) ends at episode k, then we update $V^{k+1}_h(s^k_h) \leftarrow L_\tau(h,s^k_h)$ where $L_\tau(h,s)$ is the sum of the average "reward plus bonus" for state s during the current epoch and a confidence width $\beta_{N_\tau(h,s)}$ term, truncated to $H-h+1$:

$$L_\tau(h,s) = \min\left\{H-h+1, \frac{\sum_{j\in\mathcal{K}_\tau(h,s)}\left(r^j_h + V^j_{h+1}(s^j_{h+1})\right)}{N_\tau(h,s)} + \beta_{N_\tau(h,s)}\right\}, \tag{3}$$

for all $\tau \geq 1$ (and $L_0(h,s) = H-h+1$ for all h,s). Here, for any $n \in \mathbb{Z}_+$ and some known logarithmic factors $\iota > 0$ (that depends on the choice of adversarial bandit subroutine), β_n is a confidence width defined as

$$\beta_n = \sqrt{\iota/n} \quad \text{where} \quad \iota = \text{poly}(H, |A|, \log(K|S||A|/\delta)). \tag{4}$$

Adversarial bandit subroutine. The adversarial bandit subroutine for step-state pair (h,s) is denoted by ADVBANDIT$_{h,s}$. If (h,s) is visited in episode k during epoch τ, ADVBANDIT$_{h,s}$ updates its model with the played action a^k_h and a normalized value $\frac{1}{H}(r^k_h + V^k_{h+1}(s^k_{h+1})) \in [0,1]$. Then, ADVBANDIT$_{h,s}$ outputs a policy $\pi^{k+1}_h(\cdot|s^k_h) \in \Delta(A_h)$. In this case, the external regret of ADVBANDIT$_{h,s}$ till the n-th visit of (h,s) in epoch τ is defined as

$$\text{Reg}^\tau_{h,s}(n) := \max_{\mu\in\Delta(A_h)} \frac{1}{H}\sum_{i=1}^{n}\left(\mathbb{D}_{\mu,\nu^{t_i}_h}\left[r_h + P_h V^{t_i}_{h+1}\right](s) - \mathbb{D}_{\mu^{t_i}_h,\nu^{t_i}_h}\left[r_h + P_h V^{t_i}_{h+1}\right](s)\right), \tag{5}$$

where t_i is the episode corresponding to the i-th visit of (h,s) pair in epoch τ. As the visitation counter $N_\tau(h,s)$ may not reach $\lceil(1+\eta)N_{\tau-1}(h,s)\rceil$ when Algorithm 1 ends, we require ADVBANDIT$_{h,s}$ to satisfy a high-probability anytime regret guarantee for some function ξ, polynomial in all arguments, and confidence parameter $\delta' \in (0,1)$:

$$\mathbb{P}\left(\forall n \in \mathbb{N} : \text{Reg}^\tau_{h,s}(n) \leq \xi\left(\log(1/\delta'), n, |A_h|\right)\right) \geq 1-\delta'. \tag{6}$$

Indeed, most existing adversarial bandit algorithms under the Follow-the-Regularized-Leader (FTRL) framework with implicit exploration (IX) estimators (Neu, 2015) and doubling trick satisfy this requirement. For completeness, we explicitly spell out the following example.

Example 1 *(Luo, 2017, Theorem 2) FTRL with $1/2$-Tsallis entropy, IX estimator, and doubling trick satisfies Eq. (6) with $\xi\left(\log(1/\delta'), n, |A_h|\right) = \mathcal{O}(\sqrt{|A_h| n \log(1/\delta')})$ and $\xi\left(\log(1/\delta'), n, |A_h|\right) = \mathcal{O}(\sqrt{|A_h| n \log(|A_h|/\delta')})$ for an oblivious environment and an adaptive environment, respectively. Here, an extra $|A_h|$ in the logarithmic term for the adaptive environment is due to an additional union bound. With such a bandit subroutine, we use $\iota = \Theta(H^2|A|\log(HK|A||S|/\delta))$ in Algorithm 1.*

3.1. Main Results for Epoch V-Learning

Now, we present a new result for epoch V-learning. Note again that (Mao et al., 2022) analyze this algorithm under the self-play setting, which is rather different from the analysis herein for the new

regret metric ENR under an arbitrary opponent. To this end, we introduce a non-stationarity measure capturing the "variance" of the opponent's policies, defined as

$$C = \sum_{h=1}^{H} \sum_{k=1}^{K} \mathrm{TV}\left(\nu_h^k(s_h^k), \nu_h^*(s_h^k)\right), \tag{7}$$

where $\nu_h^*(s) \in \operatorname{argmin}_{\nu \in \mathtt{Conv}\left(\nu_h^1(s), \ldots, \nu_h^K(s)\right)} \max_{\mu \in \Delta(A_h)} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s)$ is the minimax policy for the opponent when restricted to play a mixture of empirical polices $\nu_h^1(s), \ldots, \nu_h^K(s)$. In other words, C measures the cumulative difference between the opponent's policy and a fixed minimax policy. It is clear that in the worst case, we have $C = \mathcal{O}(HK)$, while in the best case when ν_h^k stays the same over all episodes, we have $C = 0$.

We are now ready to present our main result. For simplicity, we restrict our attention to oblivious opponents, but similar results hold for adaptive opponents too; see Appendix B.

Theorem 2 *Suppose that the opponent is oblivious and $K \geq H|S|$. If we run Algorithm 1 with $\eta \in [|S|/K, 1/H]$ and the adversarial bandit subroutine is instantiated with Example 1 (so $\iota = \Theta(H^2|A|\log(HK|A||S|/\delta))$), then with probability at least $1-\delta$, the estimate holds:*

$$\mathrm{ENR}_K \leq \sum_{k=1}^{K} \left(V_1^k(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) \leq \mathcal{O}\left(\eta H^3 C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right).$$

If C is known, then setting $\eta = \min\left\{\frac{1}{H}, \left(\frac{\iota|S|K}{H^4C^2}\right)^{1/3}\right\}$ achieves $\mathrm{ENR}_K \leq \widetilde{\mathcal{O}}\big(\sqrt{H^3\iota|S|K} + \left(\iota H^5|S|KC\right)^{\frac{1}{3}}\big)$, smoothly interpolating between $\widetilde{\mathcal{O}}(\sqrt{K})$ and $\widetilde{\mathcal{O}}(K^{2/3})$. We will address the issue that the knowledge of C is required in achieving this bound in Section 4. Before that, we discuss the case when $C = 0$.

Corollary 3 (External regret under a stationary opponent) *If $\nu^1 = \cdots = \nu^K$, then running Algorithm 1 with $\eta = 1/H$ and the adversarial bandit subroutine instantiated with Example 1 ensures that with probability at least $1-\delta$, $\mathrm{REG}_K = \mathrm{ENR}_K = \widetilde{\mathcal{O}}(\sqrt{H^5|A||S|K})$.*

Note that this result also implies that one can run Algorithm 1 in the single-agent setting over a fixed MDP and obtain external regret bound $\mathrm{REG}_K = \widetilde{\mathcal{O}}(\sqrt{H^5|A||S|K})$, since this setting can be modeled as a MG with a dummy second player with only one signle action at every state. This matches the regret bound of the UCB-H algorithm (Jin et al., 2018) and is close to the optimal bound $\widetilde{\mathcal{O}}(\sqrt{H^3|A||S|K})$ (Zhang et al., 2024). To the best of our knowledge, this is first result establishing a $\sqrt{K}$-type regret bound for V-learning algorithms over a fixed MDP, thereby resolving an open problem raised by Tian et al. (2021).

3.2. Proof Sketch of Theorem 2

In this subsection, we sketch the proof of Theorem 2 and highlight the novelty of our analysis. In what follows, when we refer to a variable (e.g., $\mathcal{E}(h,s)$, $N_\tau(h,s)$, etc.), we refer to their final value after K episodes.

Similar to (Tian et al., 2021), our analysis starts with a standard optimism argument, which shows $V_h^*(s) \le V_h^k(s)$ for all k, h, s (see Lemma 11). Then, we apply optimism and write the bound in terms of $\delta_h^k := (V_h^k - V_h^{\mu^k,\nu^k})(s_h^k)$:

$$\textsc{Enr}_K \le \sum_{k=1}^K \left(V_1^k(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) = \sum_{k=1}^K \delta_1^k = \sum_{h=1}^H \sum_{k=1}^K \left(\delta_h^k - \delta_{h+1}^k\right)$$

(with $\delta_{H+1}^k = 0$ for all k). To prove the claimed bound, it suffices to show for each h,

$$\sum_{k=1}^K \delta_h^k \le \sum_{k=1}^K \delta_{h+1}^k + \mathcal{O}\big(\eta H^2 C + \sqrt{\iota|S|K\log(K)\eta^{-1}}\big). \tag{8}$$

The rest of the proof for achieving this bound is novel as far as we know. Specifically, for each h, k, we rewrite δ_h^k as

$$\delta_h^k = \underbrace{\left(V_h^k(s_h^k) - \mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^k\right](s_h^k)\right)}_{(1)_h^k} + \underbrace{\mathbb{D}_{\mu_h^k,\nu_h^k}\left[P_h(V_{h+1}^k - V_{h+1}^{\mu^k,\nu^k})\right](s_h^k)}_{(2)_h^k}.$$

Since $\sum_{k=1}^K \delta_h^k = \sum_{k=1}^K (1)_h^k + \sum_{k=1}^K (2)_h^k$, we then bound each summation.

Bounding $\sum_k (2)_h^k$. Note that $\sum_k \left(\mathbb{D}_{\mu_h^k,\nu_h^k}\left[P_h(V_{h+1}^k - V_{h+1}^{\mu^k,\nu^k})\right](s_h^k) - \delta_{h+1}^k\right) \le \widetilde{\mathcal{O}}(H\sqrt{K})$ by Azuma-Hoeffding inequality for martingale difference sequence. Hence, for each $h \in [H]$, we have

$$\sum_{k=1}^K (2)_h^k \le \sum_{k=1}^K \delta_{h+1}^k + \widetilde{\mathcal{O}}(H\sqrt{K}). \tag{9}$$

Bounding $\sum_k (1)_h^k$. Considering the objective in Eq. (8) and the bound on $\sum_k (2)_h^k$ in Eq. (9), we need only to show that $\sum_k (1)_h^k \le \mathcal{O}\big(\eta H^2 C + \sqrt{\iota|S|K\log(K)\eta^{-1}}\big)$. Indeed, we have that

$$\begin{aligned}
\sum_{k=1}^K (1)_h^k &= \sum_{k=1}^K \left(V_h^k(s_h^k) - \mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^k\right](s_h^k)\right) \\
&= \sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} \sum_{k\in K_\tau(h,s)} \left(V_h^k(s) - \mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^k\right](s)\right) \\
&\le \sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_{\tau-1}(h,s) - L_\tau(h,s)\right) + \mathcal{O}\Big(\sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)}\Big),
\end{aligned}$$

where in the last step, we use the fact $V_h^k(s) = L_{\tau-1}(h,s)$ for all $k \in \mathcal{K}_\tau(h,s)$, the definition of $L_\tau(h,s)$ from Eq. (3), and a standard concentration argument applied to $\sum_k \mathbb{D}_{\mu_h^k,\nu_h^k}[r_h + P_h V_{h+1}^k](s_h^k) - \sum_{k\in\mathcal{K}_\tau(h,s)}(r_h^k + V_{h+1}^k(s_{h+1}^k))$; see Eq. (22) for more details. The upper bound of $\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)}$ depends on the number of epochs for each (h,s) pair, which is bounded according to the following lemma via a simple calculation.

Lemma 4 *For any $h \in [H]$ and $s \in S_h$, we have $|\mathcal{E}(h,s)| \le \left\lceil \frac{(1+\eta)\log(K)}{\eta} \right\rceil$.*

Then, we can show that $\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)} \le \mathcal{O}\big(\sqrt{\iota|S|K\log(K)\eta^{-1}}\big)$ by Cauchy–Schwarz inequality. It remains to bound $\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_{\tau-1}(h,s)-L_\tau(h,s)\right)$. For each $s\in S_h$, one can show that

$$
\begin{aligned}
&\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_{\tau-1}(h,s)-L_\tau(h,s)\right)\\
&= N_1(h,s)L_0(h,s) + \sum_{\tau=1}^{|\mathcal{E}(h,s)|-1} L_\tau(h,s)\left(N_{\tau+1}(h,s)-N_\tau(h,s)\right) - N_{|\mathcal{E}(h,s)|}(h,s)L_{|\mathcal{E}(h,s)|}(h,s)\\
&= N_1(h,s)\left(L_0(h,s)-V_h^*(s)\right) + \sum_{\tau=1}^{|\mathcal{E}(h,s)|-1} \left(N_{\tau+1}(h,s)-N_\tau(h,s)\right)\left(L_\tau(h,s)-V_h^*(s)\right)\\
&\quad - \sum_{s\in S_h} N_{|\mathcal{E}(h,s)|}(h,s)\left(L_{|\mathcal{E}(h,s)|}(h,s)-V_h^*(s)\right)\\
&\le N_1(h,s)\left(L_0(h,s)-V_h^*(s)\right) + \eta\sum_{\tau=1}^{|\mathcal{E}(h,s)|-1} N_\tau(h,s)\left(L_\tau(h,s)-V_h^*(s)\right) + |\mathcal{E}(h,s)|H\\
&\le \eta\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_\tau(h,s)-V_h^*(s)\right) + \mathcal{O}\left(\frac{H\log(K)}{\eta}\right), \qquad (10)
\end{aligned}
$$

where the first equality follows from the fact that for any $\{a_i\}_{i=1}^n, \{b_i\}_{i=1}^n$, we have $\sum_{i=1}^n a_i(b_{i-1}-b_i) = a_1b_0 + \sum_{i=1}^{n-1}(a_{i+1}-a_i)b_i - a_nb_n$, the first inequality uses facts $N_{\tau+1}(h,s)-N_\tau(h,s)\le \eta N_\tau(h,s)+1$ and $L_\tau(h,s)\ge V_h^*(s)$ by optimism, and the last inequality uses Lemma 4 to bound $|\mathcal{E}(h,s)|\le\mathcal{O}(\log(K)/\eta)$ for all (h,s).

Now, the following lemma is a key to handle $\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_\tau(h,s)-V_h^*(s)\right)$, which is also the most technically novel part of our proof. As discussed in (Tian et al., 2021, Remark 4), the gap between the optimistic estimate $L_\tau(h,s)$ and $V_h^*(s)$ may not diminish. Indeed, the following lemma presents an upper bound that depends on the non-stationarity C and the epoch incremental factor η.

Lemma 5 *Suppose that $\eta\in(0,1/H]$. With probability at least $1-\delta$, for any step $h\in[H]$,*

$$
\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_\tau(h,s)-V_h^*(s)\right) \le \mathcal{O}\left(H^2C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H^2|S|\log(K)}{\eta}\right).
$$

The proof of Lemma 5 relies on a recursive argument from step h to H, which requires $\eta\le 1/H$ to bypass an exponential dependence on H. We refer readers to Appendix A.4 for details.

Summing Eq. (10) for all $s\in S_h$ and then applying Lemma 5, we obtain

$$
\sum_{k=1}^K (1)_h^k \le \mathcal{O}\left(\eta H^2C + H\sqrt{\eta\iota|S|K\log(K)} + \sqrt{\iota|S|K\log(K)\eta^{-1}} + H|S|\log(K)\eta^{-1}\right). \qquad (11)
$$

The desired bound in Eq. (8) is immediate according to Eq. (11), Eq. (9). Finally, summing Eq. (8) for all $h\in[H]$, using the fact $\eta\le 1/H$ to bound $H^2\sqrt{\eta\iota|S|K\log(K)} \le H\sqrt{\iota|S|K\log(K)\eta^{-1}}$, and using $K\ge H|S|$ to bound $H^2|S|\log(K)/\eta \le H\sqrt{\iota|S|K\log(K)\eta^{-1}}$, we obtain the claimed regret bound.

Algorithm 2 Adaptive Epoch V-Learning

Input: confidence $\delta \in (0,1)$, total episode K, absolute constant $c_0 \geq 2$ specified in Eq. (39), epoch incremental scheduling $\{\eta_\ell^{(b)}\}_{b,\ell}$.

for *block* $b = 1, 2, \ldots$ **do**

  for $\ell = 1, 2 \ldots, 2^{2b} + 1$ **do**

    Create a new instance of Algorithm 1, denoted by $\mathbf{Alg}_\ell^{(b)}$, with input $\left(K, \frac{\delta}{2K^6}, \eta_\ell^{(b)}\right)$.

    Initialize $\Phi_\ell^{(b)} \leftarrow 0$, $\mathcal{T}_\ell^{(b)} \leftarrow \emptyset$, and $D \leftarrow 0$.

    while $\Phi_\ell^{(b)} \leq D$ **do**

      Let k be the current episode.

      Run $\mathbf{Alg}_\ell^{(b)}$ to collect rewards $\{r_h^k\}_{h=1}^H$ and construct optimistic estimate $V_1^k(s_1^k)$.

      Update $\mathcal{T}_\ell^{(b)} \leftarrow \mathcal{T}_\ell^{(b)} \cup \{k\}$ and $\Phi_\ell^{(b)} \leftarrow \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_1^k(s_1^k) - \sum_{h=1}^H r_h^k\right) + \sqrt{\iota \left|\mathcal{T}_\ell^{(b)}\right|}$.

      If $\ell \leq 2^{2b}$, set $D \leftarrow 3c_0 H \sqrt{\frac{\iota |S| |\mathcal{T}_\ell^{(b)}| \log(K)}{\eta_\ell^{(b)}}}$; otherwise, set $D \leftarrow 4c_0 H \sqrt{\frac{\iota |S| K \log(K)}{\eta_\ell^{(b)}}}$.

4. Adapting to Unknown Non-Stationarity: A Meta-Algorithm

As mentioned, Algorithm 1 cannot adapt to the unknown level of policy variance and requires tuning the epoch incremental factor η based on C. In fact, even setting this issue aside, a regret bound solely depends on C is not always satisfactory, because there are simple examples where the opponent is intuitively benign yet C is still $\Omega(K)$—for example, when the opponent switches their policy only once. In this section, we address both of these limitations simultaneously, leading to our final parameter-free algorithm that automatically adapts to both C and another non-stationarity measure $L := 1 + \sum_{k=1}^{K-1} \mathbb{I}\{\nu^k \neq \nu^{k+1}\}$, that is, the number of policy switches by the opponent (plus one).

4.1. Algorithm and Main Results

The proposed meta-algorithm, called Adaptive Epoch V-Learning, is shown in Algorithm 2. It uses Algorithm 1 as a base algorithm and adaptively restarts it with different epoch incremental factors in response to potential environment changes. The meta-algorithm runs in blocks $b = 1, 2, \ldots$, and each block b is further divided into a number of sub-blocks $\ell = 1, 2, \ldots, 2^{2b} + 1$. For each sub-block ℓ in block b, Algorithm 2 creates a new instance of Algorithm 1 with input $(K, \delta/(2K^6), \eta_\ell^{(b)})$. Here, if the opponent is oblivious, then one can set the epoch incremental factor $\eta_\ell^{(b)}$ as follows (for adaptive opponents, refer to Theorem 27 for the choice of $\eta_\ell^{(b)}$):

$$\eta_\ell^{(b)} = \begin{cases} \frac{1}{H}, & \ell \leq 2^{2b}, \\ \max\left\{\frac{2^{-2b}}{H}, \frac{|S|}{K}\right\}, & \ell = 2^{2b} + 1. \end{cases} \tag{12}$$

When running an instance of Algorithm 1, we monitor a (computable) upper bound $\Phi_\ell^{(b)}$ on its ENR, and when it exceeds a threshold D whose value follows the second term of the regret bound in Theorem 2, we terminate this sub-block.

High-level ideas. If, for a moment, we only focus on the last sub-block for each block and assume K is large so that $\max\left\{\frac{2^{-2b}}{H}, \frac{|S|}{K}\right\} = \frac{2^{-2b}}{H}$. Then the incremental factor schedule in Eq. (12) is simply performing a standard doubling trick on the unknown non-stationarity C, that is: maintain a guess on C, set the incremental factor accordingly, and when $\Phi_\ell^{(b)} \le D$, which we know cannot be true if the guess on C is correct, we restart the algorithm and double the guess. This would be enough to obtain an $\widetilde{\mathcal{O}}\big(\sqrt{K} + (CK)^{1/3}\big)$ regret bound without knowing C.

To further adapt to L, the algorithm additionally introduces the first 2^b sub-blocks, each serving as a test for a potential policy switch. By Corollary 3, if the opponent does not change its policy, then the condition $\Phi_\ell^{(b)} \le D$ will hold and the algorithm keeps running. Therefore, the termination of a sub-block implies that at least one policy switch has occurred, leading to an $\widetilde{\mathcal{O}}(\sqrt{LK})$ regret bound.

Finally, since each of the first 2^b sub-blocks uses $\eta_\ell^{(b)} = 1/H$ while the last sub-block uses $\eta_\ell^{(b)} = 2^{-2b}/H$ (for a large K), when all 2^b sub-blocks terminate, the regret incurred by the first 2^b sub-blocks and that incurred by the last sub-block are of the same order. This justifies taking the minimum of the two bounds and yields the refined regret bound $\widetilde{\mathcal{O}}(\min\{\sqrt{K} + (CK)^{1/3}, \sqrt{LK}\})$.

The final result is summarized in the following theorem, where when we refer to Algorithm 1, we implicitly mean running it with Example 1 as the adversarial bandit subroutine, and hence ι is again of order $\Theta(H^2|A|\log(HK|A||S|/\delta))$.

Theorem 6 *For an oblivious opponent, running Algorithm 2 with the choice of $\eta_\ell^{(b)}$ specified in Eq. (12) guarantees, with probability at least $1-\delta$*

$$\textsc{Enr}_K \le \widetilde{\mathcal{O}}\left(H^2|S| + \min\left\{\sqrt{H^3\iota|S|K} + \left(\iota H^5|S|KC\right)^{\frac{1}{3}}, \sqrt{LH^3\iota|S|K}\right\}\right).$$

Similar results for an adaptive opponent can be found in Theorem 27. Once again, we emphasize that our algorithm is completely parameter-free and its regret bound on $\textsc{Enr}_K$ smoothly interpolates between $\widetilde{\mathcal{O}}(\sqrt{K})$ and $\widetilde{\mathcal{O}}(K^{2/3})$ when the non-stationarity measures C and L increase. This means that it recovers both the $\textsc{Reg}_K = \widetilde{\mathcal{O}}(\sqrt{K})$ external regret bound for a fixed opponent as achieved by standard single-agent RL algorithms and also the $\textsc{Nr}_K = \widetilde{\mathcal{O}}(K^{2/3})$ Nash-value regret bound of Tian et al. (2021) in the worst case. Beyond the worst case, our bound can be strictly sharper when the opponent is moderately non-stationary. For example, if $C = o(K)$, then our regret is $o(K^{2/3})$, and if the opponent switches policies only logarithmically many times, then our regret scales as $\widetilde{\mathcal{O}}(\sqrt{K})$.

5. Conclusion and Open Problems

In this paper, we study online learning in uninformed Markov games, where the learner does not observe the opponent's actions. We introduced ENR, a new regret notion that is strictly stronger than NR used in Tian et al. (2021) and that reduces to standard external regret when the opponent plays a fixed policy. Under ENR, we provide a new analysis of epoch-based V-learning and show how its performance depends on a natural measure of opponent's non-stationarity. Building on this analysis, we design a parameter-free meta-algorithm that adaptively restarts epoch V-learning and achieves a regret bound that smoothly interpolates between the stationary and worst-case regimes, attaining $\widetilde{\mathcal{O}}(\min\{\sqrt{K} + (CK)^{1/3}, \sqrt{LK}\})$ regret bound. Our results also elicit compelling open questions:

- **Is the worst-case $\widetilde{\mathcal{O}}(K^{2/3})$ bound tight in uninformed MGs?** There remains a gap between the current $\widetilde{\mathcal{O}}(K^{2/3})$ upper bound and the best known $\Omega(\sqrt{K})$ lower bound. A trivial baseline is

obtained by enumerating all $|A|^{|S|H}$ deterministic policies and running an adversarial bandit algorithm over this class, which yields $\widetilde{\mathcal{O}}(H\sqrt{|A|^{|S|H}K})$ regret. Therefore, closing the gap requires either a lower bound of order $\Omega(\min\{H\sqrt{|A|^{|S|H}K}, K^{2/3}\})$ or a genuinely new algorithmic or analytical idea that improves the $\widetilde{\mathcal{O}}(K^{2/3})$ rate.

- **Can $\widetilde{\mathcal{O}}(\sqrt{LK})$ be improved under ENR?** If one only aims for a $\widetilde{\mathcal{O}}(\sqrt{LK})$ guarantee, the algorithm of Wei and Luo (2021) can be applied to obtain such a bound. However, their result is stated for dynamic regret, which is stronger than ENR. Since we work with the weaker notion ENR, it is natural to ask whether one can obtain a better dependence on L under ENR while still degrading gracefully to the worst-case $\widetilde{\mathcal{O}}(K^{2/3})$ rate when L is on the order of K.

- **A "best-of-all-worlds" guarantee across regret notions.** The setting we consider generalizes online learning in a sequence of changing MDPs, yet our guarantees are proved only for ENR. In contrast, Wei and Luo (2021) and Jin et al. (2023) study changing MDPs under different regret notions. They consider dynamic regret and transition adaptive external regret, respectively, and both notions are stronger than ENR. This raises a natural open question: can we design a single algorithm that, without knowing the regime in advance, attains our sublinear ENR guarantee in the worst case, while also matching the sharper guarantees of Wei and Luo (2021) and Jin et al. (2023) under their respective regret notions whenever the interaction reduces to their changing MDP models?

Acknowledgments

HL is supported by NSF award IIS-1943607. LJR and JL are supported in part by NSF award AF-2312775 and CPS-1844729.

References

Alexander Appel and Vanessa Kosoy. Regret bounds for robust online decision making. In *Conference on learning theory*, 2025.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 2020.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 2019.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 2013.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 2003.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*, 49(4): 2295–2322, 2024.

Tiancheng Jin, Junyan Liu, Chloé Rouyer, William Chang, Chen-Yu Wei, and Haipeng Luo. No-regret online reinforcement learning with adversarial losses and transitions. In *Advances in Neural Information Processing Systems*, 2023.

Michael L Littman et al. Friend-or-foe q-learning in general-sum games. In *ICML*, 2001.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *International Conference on Machine Learning*, pages 14036–14053. PMLR, 2022.

Haipeng Luo. Lecture note 13, Introduction to Online Learning. 2017. URL https://haipeng-luo.net/courses/CSCI699/lecture13.pdf.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28:3168–3176, 2015.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 1953.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 2017.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International Conference on Machine Learning*, 2021.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 2019.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2021.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, 2024.

Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. Smarts: An open-source scalable multi-agent rl training school for autonomous driving. In *Conference on robot learning*, pages 264–285. PMLR, 2021.

Appendix

A Regret Bound of Epoch V-learning for Oblivious Opponent 17

A.1 Notation For Epoch V-learning . . . 17

A.2 Construction of Nice Event $\mathcal{G}$ . . . 17

A.3 Proof of Theorem 2 . . . 19

A.4 Supporting Lemmas . . . 21

B Regret Bound of Epoch V-learning for Adaptive Opponent 26

B.1 Regret Bound under ENR_K . . . 26

B.2 Regret Bound under NR_K . . . 29

C Regret Bound of Adaptive Epoch V-learning for Oblivious Opponent 29

C.1 Notations. . . . 29

C.2 Construction of Nice Event . . . 30

C.3 Supporting Lemmas . . . 31

D Regret Bound of Adaptive Epoch V-learning for Adaptive Opponent 35

D.1 Supporting Lemmas . . . 36

D.2 Proof of Theorem 27 . . . 38

Appendix A. Regret Bound of Epoch V-learning for Oblivious Opponent

Remark 7 *We remark that what Tian et al. (2021) obtain for the case with a fixed opponent is subtle. Applying their result literally, one indeed only gets an $\widetilde{\mathcal{O}}(K^{2/3})$ bound on the Nash-value regret instead of the external regret. However, if one applies their result to a different MG, then it in fact does imply an $\widetilde{\mathcal{O}}(K^{2/3})$ bound on the external regret. To see this, note that since the opponent is fixed, the learner is equivalently facing a fixed MDP, which in turn can be treated as another MG with a dummy min-player with only one action in all states. It is clear that Nash-value regret for this MG is the external regret of the original MG.*

Nevertheless, we argue that this type of reasoning is cumbersome and, more importantly, does not generalize as long as the opponent deviates slightly from the ideal stationary case. On the contrary, our proposed notion of empirical Nash-value regret is much more convenient and versatile.

A.1. Notation For Epoch V-learning

For ease of reading, we here list key notations that will be used in the analysis of epoch V-learning.

Table 1: Notation Summary

Description	Description
$\pi_h^k(\cdot\|s)$	Distribution over A_h at state s in step h at episode k.
r_h^k	Reward in step h at episode k.
a_h^k	$a_h^k \sim \pi_h^k(\cdot\|s)$, action taken at step h at episode k.
$N_\tau(h,s)$	Number of visits of (h,s) pair in epoch τ.
$V_h^k(s)$	The optimistic estimation of $V_h^*(s)$.
$\mathcal{K}_\tau(h,s)$	Set of episodes that visit (h,s) in epoch τ.
$\mathcal{E}(h,s)$	Set of epochs that (h,s) has proceeded
ι	$\iota = \mathrm{poly}(H, \lvert A\rvert, \log(K\lvert S\rvert\lvert A\rvert/\delta))$.
β_n	For positive integer n, $\beta_n = \sqrt{\iota/n}$.

A.2. Construction of Nice Event $\mathcal{G}$

Lemma 8 (Restatement of Lemma 4) *For any $h \in [H]$ and $s \in S_h$, $|\mathcal{E}(h,s)| \le \left\lceil \frac{(1+\eta)\log(K)}{\eta} \right\rceil$.*

Proof Consider an arbitrary pair (h,s). We will show that the total number of epochs cannot exceed $\tau = \lceil \log(K)/\log(1+\eta) \rceil$. Suppose that (h,s) is in epoch τ. As $N_0(h,s) = 1$ for all (h,s), we have $\lceil (1+\eta) N_{\tau-1}(h,s) \rceil \ge (1+\eta) N_{\tau-1}(h,s) \ge \cdots \ge (1+\eta)^\tau N_0(h,s) = (1+\eta)^\tau$. Thus,

$$\lceil (1+\eta) N_{\tau-1}(h,s) \rceil \ge (1+\eta)^\tau \ge (1+\eta)^{\log(K)/\log(1+\eta)} = K,$$

which implies that to end epoch τ, (h,s) should be visited over K times. Using the fact that $\log(1+x) \ge \frac{x}{1+x}$ for any $x > -1$ to lower bound $\log(1+\eta) \ge \eta/(1+\eta)$, we complete the proof. ■

Given Lemma 4, we define the upper bound of epoch number as

$$M := \left\lceil \frac{(1+\eta)\log(K)}{\eta} \right\rceil . \tag{13}$$

For any $h \in [H+1]$, we define function $F_h : S_h \to [0, H]$, and define

$$\epsilon_h^k(s; F_{h+1}) := \mathbb{D}_{\mu_h^k, \nu_h^k}[r_h + P_h F_{h+1}](s) - \left(r_h^k + F_{h+1}(s_{h+1}^k)\right). \tag{14}$$

We choose ι as:

$$\iota = \max\{4c^2, 8\} H^2 |A| \log(8HKM|A||S|/\delta). \tag{15}$$

Recall that $\mathcal{F}_h^k$ is the history before step h in episode k. We define events

$$\mathcal{G}^1 := \left\{ \forall h : \left| \sum_{k=1}^K \mathbb{E}\left[\delta_h^k \mid \mathcal{F}_h^k\right] - \sum_{k=1}^K \delta_h^k \right| \le 2H\sqrt{2K \log(8H/\delta)} \right\}, \tag{16}$$

$$\mathcal{G}^2 := \left\{ \forall h, s, \tau : \left| \sum_{t \in \mathcal{K}_\tau(h,s)} \epsilon_h^t\left(s; V_{h+1}^t\right) \right| \le \frac{1}{2} N_\tau(h,s) \beta_{N_\tau(h,s)} \right\}, \tag{17}$$

$$\mathcal{G}^3 := \left\{ \forall h, s, \tau : \left| \sum_{t \in \mathcal{K}_\tau(h,s)} \epsilon_h^t\left(s; V_{h+1}^*\right) \right| \le \frac{1}{2} N_\tau(h,s) \beta_{N_\tau(h,s)} \right\}, \tag{18}$$

$$\mathcal{G}^4 := \left\{ \forall h, s, \tau : \mathrm{Reg}_{h,s}^\tau(N_\tau(h,s)) \le \frac{1}{2H} N_\tau(h,s) \beta_{N_\tau(h,s)} \right\}, \tag{19}$$

where $c > 0$ is some absolute constant and $\mathrm{Reg}_{h,s}^\tau(N_\tau(h,s))$ is defined in Eq. (5).

Definition 9 *Let $\mathcal{G}$ be the event such that $\mathcal{G} = \mathcal{G}^1 \cap \mathcal{G}^2 \cap \mathcal{G}^3 \cap \mathcal{G}^4$.*

Lemma 10 *We have $\mathbb{P}(\mathcal{G}) \ge 1 - \delta$.*

Proof For $\mathcal{G}^1$, one can see that $|\mathbb{E}\left[\delta_h^k \mid \mathcal{F}_h^k\right] - \delta_h^k| \le 2H$ for all k, h. We apply Azuma–Hoeffding for martingale difference sequence $\{\mathbb{E}\left[\delta_h^k \mid \mathcal{F}_h^k\right] - \delta_h^k\}_{k=1}^K$ with respect to any fixed step $h \in [H]$, and then use a union bound over all $h \in [H]$ to obtain $\mathbb{P}(\mathcal{G}^1) \ge 1 - \delta/4$.

For $\mathcal{G}^2$, we rewrite $\sum_{t \in \mathcal{K}_\tau(h,s)} \epsilon_h^t\left(s; V_{h+1}^t\right) = \sum_{i=1}^{N_\tau(h,s)} \epsilon_h^{t_i}\left(s; V_{h+1}^{t_i}\right)$ where t_i is the episode, corresponding to the i-th visit of (h,s) pair in epoch τ. We have that $\mathbb{E}[\epsilon_h^{t_i}\left(s; V_{h+1}^{t_i}\right) \mid \mathcal{F}_h^{t_i}] = 0$ and $|\epsilon_h^{t_i}\left(s; V_{h+1}^{t_i}\right)| \le H$. Consider fixed (h,s) and epoch τ, and we further fix a $N_\tau(h,s) \in [K]$. Applying Azuma–Hoeffding gives that with probability at least $1 - \delta/(4HKM|S|)$,

$$\left| \sum_{i=1}^{N_\tau(h,s)} \epsilon_h^{t_i}\left(s; V_{h+1}^{t_i}\right) \right| \le H\sqrt{2N_\tau(h,s) \log(8KH|S|M/\delta)} \le \frac{1}{2} N_\tau(h,s) \beta_{N_\tau(h,s)}.$$

Using a union bound over all (h,s), τ, $N_\tau(h,s)$, and following Lemma 4 that the number of epochs of any (h,s) pair is at most M, we have $\mathbb{P}(\mathcal{G}^2) \ge 1 - \delta/4$.

The argument for proving $\mathbb{P}(\mathcal{G}^3) \ge 1 - \delta/4$ is analogous to that of $\mathcal{G}^2$. One caveat here is that we can apply Azuma-Hoeffding inequality because the opponent is assumed to be oblivious.

Then, we show that $\mathbb{P}(\mathcal{G}^4) \ge 1 - \delta/4$. For any (h,s) pair, Algorithm 1 runs a new instance of FTRL with $1/2$-Tsallis, IX estimator, and doubling trick in each epoch. By (Luo, 2017, Theorem 2), running this algorithm ensures that with probability at least $1 - \delta/(4H|S|M)$, for any $N_\tau(h,s) \in \mathbb{N}$

$$\mathrm{Reg}_{h,s}^\tau(N_\tau(h,s)) \le c\sqrt{|A_h| N_\tau(h,s) \log(HM|A_h||S|/\delta)} \le \frac{1}{2H} N_\tau(h,s) \beta_{N_\tau(h,s)},$$

where $c > 0$ is some absolute constant. Using a union bound over all h, s, τ gives the claimed result.

Finally, using a union bound over $\mathcal{G}^1, \mathcal{G}^2, \mathcal{G}^3, \mathcal{G}^4$ completes the proof. ■

A.3. Proof of Theorem 2

The proof of Theorem 2 conditions on event $\mathcal{G}$ defined in Definition 9, which occurs with probability at least $1-\delta$. By the optimism in Lemma 11 and the definition $\delta_h^k = (V_h^k - V_h^{\mu^k,\nu^k})(s_h^k)$, we write

$$\textsc{Enr}_K \le \sum_{k=1}^{K} \left(V_1^k(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k) \right) = \sum_{k=1}^{K} \delta_1^k.$$

To bound δ_1^k, we analyze every δ_h^k. Consider any fixed episode k and step h. By subtracting and adding $\mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^k\right](s_h^k)$, we write

$$\begin{aligned}
\delta_h^k &= V_h^k(s_h^k) - \mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^{\mu^k,\nu^k}\right](s_h^k) \\
&= \underbrace{\left(V_h^k(s_h^k) - \mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^k\right](s_h^k)\right)}_{(1)_h^k} + \underbrace{\mathbb{D}_{\mu_h^k,\nu_h^k}\left[P_h(V_{h+1}^k - V_{h+1}^{\mu^k,\nu^k})\right](s_h^k)}_{(2)_h^k}.
\end{aligned}$$

With this decomposition, for any step $h \in [H]$,

$$\sum_{k=1}^{K} \delta_h^k = \sum_{k=1}^{K} (1)_h^k + \sum_{k=1}^{K} (2)_h^k. \tag{20}$$

Bounding $\sum_{k=1}^{K}(2)_h^k$. Recall that $\mathcal{F}_h^k$ is the history before step h in episode k. We have

$$\mathbb{E}\left[\delta_{h+1}^k \mid \mathcal{F}_{h+1}^k\right] = \mathbb{E}\left[(V_{h+1}^k - V_{h+1}^{\mu^k,\nu^k})(s_{h+1}^k) \mid \mathcal{F}_{h+1}^k\right] = \mathbb{D}_{\mu_h^k,\nu_h^k}\left[P_h(V_{h+1}^k - V_{h+1}^{\mu^k,\nu^k})\right](s_h^k).$$

Then, $(2)_h^k$ can be bounded by

$$(2)_h^k = \delta_{h+1}^k + \mathbb{D}_{\mu_h^k,\nu_h^k}\left[P_h(V_{h+1}^k - V_{h+1}^{\mu^k,\nu^k})\right](s_h^k) - \delta_{h+1}^k = \delta_{h+1}^k + \mathbb{E}\left[\delta_{h+1}^k \mid \mathcal{F}_{h+1}^k\right] - \delta_{h+1}^k.$$

We use Eq. (16) of $\mathcal{G}$ to bound $\sum_{k=1}^{K}\left(\mathbb{E}\left[\delta_{h+1}^k \mid \mathcal{F}_{h+1}^k\right] - \delta_{h+1}^k\right)$ to obtain

$$\sum_{k=1}^{K} (2)_h^k \le 2H\sqrt{2K\log(8H/\delta)} + \sum_{k=1}^{K} \delta_{h+1}^k. \tag{21}$$

Bounding $\sum_{k=1}^{K}(1)_h^k$. By the definition of $L_\tau(h,s)$ in Eq. (3), we have $V_h^k(s) = L_{\tau-1}(h,s)$ for all $k \in K_\tau(h,s)$. For any step $h \in [H]$, one can write

$$\sum_{k=1}^{K} (1)_h^k = \sum_{k=1}^{K} \left(V_h^k(s_h^k) - \mathbb{D}_{\mu_h^k,\nu_h^k}\left[r_h + P_h V_{h+1}^k\right](s_h^k)\right)$$

$$
\begin{aligned}
&= \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in K_\tau(h,s)}\left(V_h^t(s) - \mathbb{D}_{\mu_h^t,\nu_h^t}\left[r_h + P_h V_{h+1}^t\right](s)\right)\\
&\le \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in K_\tau(h,s)}\left(V_h^t(s) - \left(r_h^t + V_{h+1}^t(s_{h+1}^t)\right)\right) + \frac{1}{2}\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)}\\
&= \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in K_\tau(h,s)}\left(L_{\tau-1}(h,s) - \left(r_h^t + V_{h+1}^t(s_{h+1}^t)\right)\right) + \frac{1}{2}\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)}\\
&\le \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_{\tau-1}(h,s) - L_\tau(h,s)\right) + \frac{3}{2}\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)},
\end{aligned}
\tag{22}
$$

where the first inequality uses Eq. (17) of $\mathcal{G}$, the last equality holds since for any $t \in K_\tau(h,s)$, we have $V_h^t(s) = L_{\tau-1}(h,s)$, and the last inequality holds due to:

$$
\begin{aligned}
&\sum_{t\in K_\tau(h,s)}\left(r_h^t + V_{h+1}^t(s_{h+1}^t)\right)\\
&= N_\tau(h,s)\left(\frac{1}{N_\tau(h,s)}\sum_{t\in K_\tau(h,s)}\left(r_h^t + V_{h+1}^t(s_{h+1}^t)\right) + \beta_{N_\tau(h,s)}\right) - N_\tau(h,s)\beta_{N_\tau(h,s)}\\
&\ge N_\tau(h,s)L_\tau(h,s) - N_\tau(h,s)\beta_{N_\tau(h,s)}.
\end{aligned}
$$

We use Lemma 12 to bound the second term of Eq. (22) as:

$$
\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)} \le \mathcal{O}\left(\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right). \tag{23}
$$

Then, we turn to bound the first term of Eq. (22) as:

$$
\begin{aligned}
&\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_{\tau-1}(h,s) - L_\tau(h,s)\right)\\
&\overset{(a)}{=} \sum_{s\in S_h}\left(N_1(h,s)L_0(h,s) + \sum_{\tau=1}^{|\mathcal{E}(h,s)|-1} L_\tau(h,s)\left(N_{\tau+1}(h,s) - N_\tau(h,s)\right) - N_{|\mathcal{E}(h,s)|}(h,s)L_{|\mathcal{E}(h,s)|}(h,s)\right)\\
&= \sum_{s\in S_h}\left(N_1(h,s)\left(L_0(h,s) - V_h^*(s)\right) + \sum_{\tau=1}^{|\mathcal{E}(h,s)|-1}\left(N_{\tau+1}(h,s) - N_\tau(h,s)\right)\left(L_\tau(h,s) - V_h^*(s)\right)\right)\\
&\quad - \sum_{s\in S_h} N_{|\mathcal{E}(h,s)|}(h,s)\left(L_{|\mathcal{E}(h,s)|}(h,s) - V_h^*(s)\right)\\
&\overset{(b)}{\le} \sum_{s\in S_h}\left(N_1(h,s)\left(L_0(h,s) - V_h^*(s)\right) + \eta\sum_{\tau=1}^{|\mathcal{E}(h,s)|-1} N_\tau(h,s)\left(L_\tau(h,s) - V_h^*(s)\right) + |\mathcal{E}(h,s)|H\right)\\
&\quad - \sum_{s\in S_h} N_{|\mathcal{E}(h,s)|}(h,s)\left(L_{|\mathcal{E}(h,s)|}(h,s) - V_h^*(s)\right)
\end{aligned}
$$

$$= \sum_{s\in S_h}\left(N_1(h,s)\left(L_0(h,s)-V_h^*(s)\right)+\eta\sum_{\tau\in\mathcal{E}(h,s)}N_\tau(h,s)\left(L_\tau(h,s)-V_h^*(s)\right)+|\mathcal{E}(h,s)|H\right)$$
$$-(1+\eta)\sum_{s\in S_h}N_{|\mathcal{E}(h,s)|}(h,s)\left(L_{|\mathcal{E}(h,s)|}(h,s)-V_h^*(s)\right)$$
$$\overset{(c)}{\leq} H|S|N_1(h,s)+\eta\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}N_\tau(h,s)\left(L_\tau(h,s)-V_h^*(s)\right)+\mathcal{O}\left(\frac{H|S|\log(K)}{\eta}\right)$$
$$\leq \mathcal{O}\left(\eta H^2C+H\sqrt{\eta\iota|S|K\log(K)}+\frac{H|S|\log(K)}{\eta}\right), \tag{24}$$

where step (a) follows from the fact that for any $\{a_i\}_{i=1}^n, \{b_i\}_{i=1}^n$, we have $\sum_{i=1}^n a_i(b_{i-1}-b_i) = a_1b_0+\sum_{i=1}^{n-1}(a_{i+1}-a_i)b_i-a_nb_n$, step (b) uses facts that $N_{\tau+1}(h,s)-N_\tau(h,s)\leq \eta N_\tau(h,s)+1$ and $L_\tau(h,s)\geq V_h^*(s)$ by Lemma 11, and step (c) uses again $L_\tau(h,s)\geq V_h^*(s)$ and Lemma 4 to bound $|\mathcal{E}(h,s)|\leq\mathcal{O}(\log(K)/\eta)$ for all (h,s), and the last inequality invokes Lemma 5 and uses $\eta\leq 1/H$ to bound $H^2|S|\log(K)\leq H|S|\log(K)/\eta$.

Plugging Eq. (24) and Eq. (23) into Eq. (22), we have that for any step $h\in[H]$

$$\sum_{k=1}^K(1)_h^k\leq\mathcal{O}\left(\eta H^2C+H\sqrt{\eta\iota|S|K\log(K)}+\sqrt{\frac{\iota|S|K\log(K)}{\eta}}+\frac{H|S|\log(K)}{\eta}\right). \tag{25}$$

Putting together. Plugging Eq. (25) and Eq. (21) into Eq. (20), we have that for any step $h\in[H]$

$$\sum_{k=1}^K\delta_h^k\leq\sum_{k=1}^K\delta_{h+1}^k+\mathcal{O}\left(\eta H^2C+H\sqrt{\eta\iota|S|K\log(K)}+\sqrt{\frac{\iota|S|K\log(K)}{\eta}}+\frac{H|S|\log(K)}{\eta}\right).$$

Summing over all $h\in[H]$, using the fact $\delta_{H+1}^k=0$ for any k, and rearranging, we have

$$\text{ENR}_K\leq\sum_{k=1}^K\delta_1^k\leq\mathcal{O}\left(\eta H^3C+H^2\sqrt{\eta\iota|S|K\log(K)}+H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}+\frac{H^2|S|\log(K)}{\eta}\right)$$
$$\leq\mathcal{O}\left(\eta H^3C+H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}+\frac{H^2|S|\log(K)}{\eta}\right),$$

where the second inequality uses the fact $\eta\leq 1/H$ to bound $H^2\sqrt{\eta\iota|S|K\log(K)}\leq H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}$.

Furthermore, if one assumes $K\geq H|S|$ and constrains $\eta\in\left[\frac{|S|}{K},\frac{1}{H}\right]$, then, $\frac{H^2|S|\log(K)}{\eta}\leq H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}$, which gives

$$\text{ENR}_K\leq\sum_{k=1}^K\left(V_1^k(s_1^k)-V_1^{\mu^k,\nu^k}(s_1^k)\right)\leq\mathcal{O}\left(\eta H^3C+H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right).$$

A.4. Supporting Lemmas

Lemma 11 (Optimism) *Suppose that $\mathcal{G}$ holds where $\mathcal{G}$ is defined in Definition 9. For all $h\in[H+1]$, we have $V_h^*(s)\leq V_h^k(s)$ for all $k\in[K], s\in S_h$.*

Proof Here, we use backward induction on h to show that for all h, $V_h^* \le V_h^k$ holds in entry-wise for all episode k. For the base case $h = H+1$, $V_h^* \le V_h^k$ holds for all $k \in [K]$ by the definitions that $V_{H+1}^*(s) = V_{H+1}^k(s) = 0$ for all state s and episode k. Suppose that for step $h+1$, we have $V_{h+1}^* \le V_{h+1}^k$ for all $k \in [K]$, and then we show that at step h, $V_h^* \le V_h^k$ holds for all $k \in [K]$.

Let $\tau(k,h,s)$ be the epoch that (h,s) pair lies at episode k. Consider any fixed episode $k \in [K]$, and any fixed state $s \in S_h$. Notice that if $\tau(k,h,s) = 1$, then $V_h^k(s) = H-h+1$, directly implying $V_h^*(s) \le V_h^k(s)$. In the following, we consider $\tau(k,h,s) > 1$ and write

$$
\begin{aligned}
V_h^*(s) &= \max_{\mu\in\Delta(A_h)} \min_{\nu\in\{\nu_h^1(s),\ldots,\nu_h^K(s)\}} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s) \\
&= \frac{1}{N_{\tau(k,h,s)-1}(h,s)} \max_{\mu\in\Delta(A_h)} \sum_{t\in\mathcal{K}_{\tau(k,h,s)-1}(h,s)} \min_{\nu\in\{\nu_h^1(s),\ldots,\nu_h^K(s)\}} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s) \\
&\le \frac{1}{N_{\tau(k,h,s)-1}(h,s)} \max_{\mu\in\Delta(A_h)} \sum_{t\in\mathcal{K}_{\tau(k,h,s)-1}(h,s)} \mathbb{D}_{\mu,\nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) \\
&\le \frac{1}{N_{\tau(k,h,s)-1}(h,s)} \max_{\mu\in\Delta(A_h)} \sum_{t\in\mathcal{K}_{\tau(k,h,s)-1}(h,s)} \mathbb{D}_{\mu,\nu_h^t}\left[r_h + P_h V_{h+1}^t\right](s) \\
&\le \frac{1}{N_{\tau(k,h,s)-1}(h,s)} \sum_{t\in\mathcal{K}_{\tau(k,h,s)-1}(h,s)} \mathbb{D}_{\mu_h^t,\nu_h^t}\left[r_h + P_h V_{h+1}^t\right](s) + \frac{1}{2}\beta_{N_{\tau(k,h,s)-1}(h,s)} \\
&\le \frac{1}{N_{\tau(k,h,s)-1}(h,s)} \sum_{t\in\mathcal{K}_{\tau(k,h,s)-1}(h,s)} \left(r_h^t + V_{h+1}^t(s_{h+1}^t)\right) + \beta_{N_{\tau(k,h,s)-1}(h,s)}, \qquad (26)
\end{aligned}
$$

where the second inequality uses the induction hypothesis, and the third inequality follows from Eq. (19) of event $\mathcal{G}$, and the last inequality uses Eq. (17) of event $\mathcal{G}$.

Since $V_h^*(s) \le H-h+1$, combining this and Eq. (26), we have $V_h^*(s) \le V_h^k(s)$. As this argument holds for all episodes k and states $s \in S_h$, the induction is done, and the lemma thus follows. ■

We define the variance of opponent's policy in epoch τ of (h,s) pair as:

$$
C_\tau(h,s) = \sum_{t\in K_\tau(h,s)} \mathtt{TV}\left(\nu_h^t(s), \nu_h^*(s)\right). \qquad (27)
$$

Lemma 12 *Suppose that $\eta \in (0, 1/H]$. For any step $h \in [H]$, the following holds.*

$$
\sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)} \le \mathcal{O}\left(\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right).
$$

Proof We show that

$$
\begin{aligned}
&\sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)} \\
&= \sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} \sqrt{\iota N_\tau(h,s)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{s\in S_h} \sqrt{\iota|\mathcal{E}(h,s)|\left(\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\right)} \\
&\leq \sum_{s\in S_h} \sqrt{\iota M\left(\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\right)} \\
&\leq \sqrt{\iota M|S_h|K} \\
&\leq \mathcal{O}\left(\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right),
\end{aligned}
$$

where the first inequality applies the Cauchy–Schwarz inequality, the second inequality uses Lemma 4 to bound $|\mathcal{E}(h,s)| \leq M$ where M is given in Eq. (13), the third inequality uses again the Cauchy–Schwarz inequality and bounds $\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s) \leq K$, and the last inequality follows from the definition of M and simply bounds $|S_h| \leq |S|$.

The proof is thus complete. ■

Lemma 13 *Suppose that $\mathcal{G}$ holds where $\mathcal{G}$ is defined in Definition 9. For any pair (h,s) and any epoch τ of this pair, we have*

$$
\begin{aligned}
&\sum_{t\in\mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^t(s_{h+1}^t) - V_h^*(s)\right) \\
&\leq \mathcal{O}\left(HC_\tau(h,s) + N_\tau(h,s)\beta_{N_\tau(h,s)}\right) + \sum_{t\in\mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right).
\end{aligned}
$$

Proof We show that for (h,s) pair and any epoch τ

$$
\begin{aligned}
&\sum_{t\in\mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^t(s_{h+1}^t) - V_h^*(s)\right) \\
&= \sum_{t\in\mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^*(s_{h+1}^t) - V_h^*(s)\right) + \sum_{t\in\mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right). \qquad (28)
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\sum_{t\in\mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^*(s_{h+1}^t) - V_h^*(s)\right) \\
&\leq \sum_{t\in\mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t,\nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - V_h^*(s)\right) + \mathcal{O}\left(N_\tau(h,s)\beta_{N_\tau(h,s)}\right) \\
&= \sum_{t\in\mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t,\nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - \mathbb{D}_{\mu_h^*,\nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s)\right) + \mathcal{O}\left(N_\tau(h,s)\beta_{N_\tau(h,s)}\right) \\
&\leq \sum_{t\in\mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t,\nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - \mathbb{D}_{\mu_h^t,\nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s)\right) + \mathcal{O}\left(N_\tau(h,s)\beta_{N_\tau(h,s)}\right),
\end{aligned}
$$

where the first inequality uses Eq. (18) of $\mathcal{G}$ and the last inequality follows from the following fact that for any h, s, t

$$\begin{aligned}
&\mathbb{D}_{\mu_h^*,\nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s) \\
&= \max_{\mu\in\Delta(A_h)} \min_{\nu\in\left\{\nu_h^1(s),\ldots,\nu_h^K(s)\right\}} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s) \\
&= \max_{\mu\in\Delta(A_h)} \min_{\nu\in\mathtt{Conv}\left(\nu_h^1(s),\ldots,\nu_h^K(s)\right)} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s) \\
&= \min_{\nu\in\mathtt{Conv}\left(\nu_h^1(s),\ldots,\nu_h^K(s)\right)} \max_{\mu\in\Delta(A_h)} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s) \\
&= \max_{\mu\in\Delta(A_h)} \mathbb{D}_{\mu,\nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s) \\
&\geq \mathbb{D}_{\mu_h^t,\nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s),
\end{aligned}$$

where the second equality holds since for any μ, $\mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1}^*\right](s)$ is linear in ν, the third equality follows from the minimax theorem, and the fourth equality uses the definition of ν_h^*.

Let us define

$$g_{h,s}^t(b) = \mathbb{E}_{a\sim\mu_h^t}\left[r_h(s,a,b) + \mathbb{E}_{s'\sim P_h(\cdot|s,a,b)}\left[V_{h+1}^*(s')\right]\right] \in [0,H].$$

Then, we have

$$\begin{aligned}
&\sum_{t\in\mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t,\nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - \mathbb{D}_{\mu_h^t,\nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s)\right) \\
&= \sum_{t\in\mathcal{K}_\tau(h,s)} \left(\mathbb{E}_{b\sim\nu_h^t}[g_{h,s}^t(b)] - \mathbb{E}_{b\sim\nu_h^*}[g_{h,s}^t(b)]\right) \\
&\leq 2 \sum_{t\in\mathcal{K}_\tau(h,s)} \left\|g_{h,s}^t(\cdot)\right\|_\infty \mathtt{TV}(\nu_h^t(s), \nu_h^*(s)) \\
&\leq 2H \sum_{t\in\mathcal{K}_\tau(h,s)} \mathtt{TV}(\nu_h^t(s), \nu_h^*(s)) \\
&= 2HC_\tau(h,s). \qquad (29)
\end{aligned}$$

The claimed result thus follows. ■

We prove the following lemma, which states the claim of Lemma 5 in a slightly different way, i.e., we prove the bound under a nice event $\mathcal{G}$, which satisfies $\mathbb{P}(\mathcal{G}) \geq 1-\delta$.

Lemma 14 *Suppose that $\mathcal{G}$ holds where $\mathcal{G}$ is defined in Definition 9 and $\eta \in (0, 1/H]$. For any step $h \in [H]$, we have*

$$\sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_\tau(h,s) - V_h^*(s)\right) \leq \mathcal{O}\left(H^2 C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H^2|S|\log(K)}{\eta}\right).$$

Proof For shorthand, let

$$\Gamma(h) := \sum_{s\in S_h} \sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_\tau(h,s) - V_h^*(s)\right).$$

We then write

$$\begin{aligned}
\Gamma(h) &= \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\left(L_\tau(h,s) - V_h^*(s)\right) \\
&\le \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in K_\tau(h,s)} \left(r_h^t + V_{h+1}^t(s_{h+1}^t) - V_h^*(s)\right) + \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} N_\tau(h,s)\beta_{N_\tau(h,s)} \\
&\le \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)} \mathcal{O}\left(HC_\tau(h,s) + N_\tau(h,s)\beta_{N_\tau(h,s)}\right) \\
&\quad + \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in\mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right) \\
&\le \mathcal{O}\left(HC + \sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right) + \sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in\mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right),
\end{aligned} \tag{30}$$

where the first inequality uses the definition of $L_\tau(h,s)$ in Eq. (3), the second inequality uses Lemma 13, and the last inequality uses Lemma 12.

Now, we show that

$$\begin{aligned}
&\sum_{s\in S_h}\sum_{\tau\in\mathcal{E}(h,s)}\sum_{t\in\mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right) \\
&= \sum_{s\in S_{h+1}}\sum_{\tau\in\mathcal{E}(h+1,s)}\sum_{t\in\mathcal{K}_\tau(h+1,s)} \left(V_{h+1}^t(s) - V_{h+1}^*(s)\right) \\
&= \sum_{s\in S_{h+1}}\sum_{\tau\in\mathcal{E}(h+1,s)} N_\tau(h+1,s)\left(L_{\tau-1}(h+1,s) - V_{h+1}^*(s)\right) \\
&\le \sum_{s\in S_{h+1}}\sum_{\tau\in\mathcal{E}(h+1,s)} (1+\eta)N_{\tau-1}(h+1,s)\left(L_{\tau-1}(h+1,s) - V_{h+1}^*(s)\right) + \mathcal{O}\left(\frac{H|S|\log(K)}{\eta}\right) \\
&\le (1+\eta)\sum_{s\in S_{h+1}}\sum_{\tau\in\mathcal{E}(h+1,s)} N_\tau(h+1,s)\left(L_\tau(h+1,s) - V_{h+1}^*(s)\right) + \mathcal{O}\left(\frac{H|S|\log(K)}{\eta}\right) \\
&= (1+\eta)\Gamma(h+1) + \mathcal{O}\left(\frac{H|S|\log(K)}{\eta}\right),
\end{aligned}$$

where the first inequality follows from the optimism (see Lemma 11) that $L_\tau(h,s) \ge V_h^*(s)$ for all τ, h, s, and the fact that $N_\tau(h,s) \le (1+\eta)N_{\tau-1}(h,s) + 1$, the second inequality uses $\forall(h,s)$, $N_0(h,s) = 1$ and $L_0(h,s) = H - h + 1$ to bound $(1+\eta)\sum_{s\in S_{h+1}} N_0(h+1,s)(L_0(h+1,s) - V_{h+1}^*(s)) \le 2|S|H \le \mathcal{O}(H|S|\log(K)/\eta)$.

By plugging the above into Eq. (30) and using the facts that $\Gamma(H+1) = 0$, we have

$$\begin{aligned}
\Gamma(h) &\le (1+\eta)\Gamma(h+1) + \mathcal{O}\left(HC + \sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H|S|\log(K)}{\eta}\right) \\
&\le \left(1 + \frac{1}{H}\right)\Gamma(h+1) + \mathcal{O}\left(HC + \sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H|S|\log(K)}{\eta}\right)
\end{aligned}$$

$$\leq \sum_{h'=h}^{H} \left(1+\frac{1}{H}\right)^{H-h'} \mathcal{O}\left(HC + \sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H|S|\log(K)}{\eta}\right)$$
$$\leq \mathcal{O}\left(H^2C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H^2|S|\log(K)}{\eta}\right),$$

where the second inequality bounds $\eta \leq 1/H$ and the last inequality unrolls from $h+1$ to $H+1$. ■

Appendix B. Regret Bound of Epoch V-learning for Adaptive Opponent

B.1. Regret Bound under $\textsc{Enr}_K$

Theorem 15 *Suppose that the opponent is adaptive and $K \geq H|S|^{3/2}$. If we run Algorithm 1 with $\eta \in [|S|/K, 1/H]$ and the adversarial bandit subroutine is instantiated by Example 1 (so $\iota = \Theta(H^2|A|\log(HK|A||S|/\delta))$), then with probability at least $1-\delta$,*

$$\textsc{Enr}_K \leq \sum_{k=1}^{K}\left(V_1^k(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) \leq \mathcal{O}\left(\eta H^3C + H^2|S|\sqrt{\eta\iota K\log(K)} + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right).$$

Further, if we constrain $\eta \in \left[|S|/K, 1/(H\sqrt{|S|})\right]$, then $\textsc{Enr}_K \leq \mathcal{O}\left(\eta H^3C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right)$.

For the adaptive opponent, the regret bound suffers an additional term $\widetilde{\mathcal{O}}(H^2|S|\sqrt{\eta\iota K})$. Unlike Theorem 6, which only suffers a $\sqrt{|S|}$ dependence, this term has a linear dependence on $|S|$, and thus choosing $\eta \leq 1/H$ cannot suppress it by $\widetilde{\mathcal{O}}\big(H\sqrt{\frac{\iota|S|K}{\eta}}\big)$. Compared to the oblivious opponent case, the worse dependence on $|S|$ arises because $V_h^*(s)$ depends on the entire sequence of $(\nu_h^1(s), \ldots, \nu_h^K(s))$, which prevents us from directly applying martingale-based concentration bounds to control $\left|\sum_{k\in\mathcal{K}_\tau(h,s)}\left(V_h^*(s_{h+1}^k) - \mathbb{D}_{\mu_h^k,\nu_h^k}[P_hV_{h+1}^*](s)\right)\right|$. To address this issue, we discretize each $V_h^*(s)$, and applying a union bound over all discretized points yields the extra dependence on $|S|$.

The proof of Theorem 15 is largely based on that of Theorem 2, and the difference comes from discretizing $V_h^*(s)$ to handle the adaptive adversary. Recall from Eq. (18) that $\mathcal{G}^3$ is the only place relying on the assumption of oblivious adversary. As $V_h^*(s)$ depends on entire sequence of opponent's polices, if the opponent selects policies adaptively, $V_h^*(s)$ is not predictable given the history. As a result, one cannot simply follow Lemma 10 to apply Azuma-Hoeffding's inequality. One way to handling the future-dependent issue of $V_h^*(s)$ is to discretize $V_h^*(s)$, and then apply a union bound over all of them.

For any $\Delta > 0$ and step $h \in [H+1]$, we define a finite class of functions that $\mathcal{V}_{H+1,\Delta} = \{f : S_{H+1} \to \{0\}\}$ and for any step $h \in [H]$

$$\mathcal{V}_{h,\Delta} := \left\{f : S_h \to \left\{0, \Delta, \ldots, \left\lfloor\frac{H-h+1}{\Delta}\right\rfloor\Delta, H-h+1\right\}\right\}. \tag{31}$$

With the definition of $\mathcal{V}_{h,\Delta}$, we then define event $\widetilde{\mathcal{G}}^3$ as:

$$\widetilde{\mathcal{G}}^3 := \left\{\forall h, s, \tau : \max_{f\in\mathcal{V}_{h+1,\Delta}}\left|\sum_{t\in\mathcal{K}_\tau(h,s)}\epsilon_h^t(s;f)\right| \leq \frac{1}{2}\sqrt{|S|}N_\tau(h,s)\beta_{N_\tau(h,s)}\right\}, \tag{32}$$

Definition 16 *Let $\widetilde{\mathcal{G}}$ be the event such that $\mathcal{G} = \mathcal{G}^1 \cap \mathcal{G}^2 \cap \widetilde{\mathcal{G}}^3 \cap \mathcal{G}^4$.*

Lemma 17 *We have $\mathbb{P}(\widetilde{\mathcal{G}}) \geq 1 - \delta$.*

Proof As Lemma 10 has shown $\mathbb{P}(\mathcal{G}^1 \cap \mathcal{G}^2 \cap \mathcal{G}^4) \geq 1 - 3\delta/4$, it suffices to show $\mathbb{P}(\widetilde{\mathcal{G}}^3) \geq 1 - \delta/4$. The argument for proving $\mathbb{P}(\widetilde{\mathcal{G}}^3) \geq 1 - \delta/4$ is analogous to that of $\mathcal{G}^2$, and the only difference is to apply an additional union bound over all $f \in \mathcal{V}_{h+1,\Delta}$. ■

Then, we show how to modify lemmas that are affected by the discretization step.

Lemma 18 (Counterpart of Lemma 13) *Suppose that $\widetilde{\mathcal{G}}$ holds where $\widetilde{\mathcal{G}}$ is defined in Definition 16 and $\Delta = 1/K$. For any pair (h, s) and any epoch τ of this pair, we have*

$$\begin{aligned}
&\sum_{t \in \mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^t(s_{h+1}^t) - V_h^*(s)\right) \\
&\leq \mathcal{O}\left(HC_\tau(h,s) + \Delta N_\tau(h,s) + \sqrt{|S|} N_\tau(h,s) \beta_{N_\tau(h,s)}\right) + \sum_{t \in \mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right).
\end{aligned}$$

Proof We show that for (h, s) pair and any epoch τ

$$\begin{aligned}
&\sum_{t \in \mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^t(s_{h+1}^t) - V_h^*(s)\right) \\
&= \sum_{t \in \mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^*(s_{h+1}^t) - V_h^*(s)\right) + \sum_{t \in \mathcal{K}_\tau(h,s)} \left(V_{h+1}^t(s_{h+1}^t) - V_{h+1}^*(s_{h+1}^t)\right). \qquad (33)
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\sum_{t \in \mathcal{K}_\tau(h,s)} \left(r_h^t + V_{h+1}^*(s_{h+1}^t) - V_h^*(s)\right) \\
&\leq \sum_{t \in \mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t, \nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - V_h^*(s)\right) + 2\Delta N_\tau(h,s) + \mathcal{O}\left(\sqrt{|S|} N_\tau(h,s) \beta_{N_\tau(h,s)}\right) \\
&= \sum_{t \in \mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t, \nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - \mathbb{D}_{\mu_h^*, \nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s)\right) \\
&\quad + 2\Delta N_\tau(h,s) + \mathcal{O}\left(\sqrt{|S|} N_\tau(h,s) \beta_{N_\tau(h,s)}\right) \\
&\leq \sum_{t \in \mathcal{K}_\tau(h,s)} \left(\mathbb{D}_{\mu_h^t, \nu_h^t}\left[r_h + P_h V_{h+1}^*\right](s) - \mathbb{D}_{\mu_h^t, \nu_h^*}\left[r_h + P_h V_{h+1}^*\right](s)\right) \\
&\quad + 2\Delta N_\tau(h,s) + \mathcal{O}\left(\sqrt{|S|} N_\tau(h,s) \beta_{N_\tau(h,s)}\right) \\
&\leq \mathcal{O}\left(HC_\tau(h,s) + \Delta N_\tau(h,s) + \sqrt{|S|} N_\tau(h,s) \beta_{N_\tau(h,s)}\right),
\end{aligned}$$

where the first inequality first rounds each $V_{h+1}^*(s_{h+1}^t)$ to a function $f \in \mathcal{V}_{h+1,\Delta}$ such that $\left\|f - V_{h+1}^*\right\|_\infty \leq \Delta$, then uses Eq. (32) of $\widetilde{\mathcal{G}}$ together with the fact that for any step h and $\Delta = 1/K$, we have $\log(|\mathcal{V}_{h,\Delta}|) \leq \mathcal{O}(|S| \log(H/\Delta))$, and finally rounding each f back to $V_{h+1}^*(s_{h+1}^t)$. Rounding twice incurs extra $2\Delta N_\tau(h,s)$ term; the last inequality uses Eq. (29).

The claimed result thus follows. ■

Lemma 19 (Counterpart of Lemma 14) *Suppose that $\widetilde{\mathcal{G}}$ holds where $\widetilde{\mathcal{G}}$, $\eta \in (0, 1/H]$, and $\Delta = 1/K$. For any step $h \in [H]$, we have*

$$\sum_{s \in S_h} \sum_{\tau \in \mathcal{E}(h,s)} N_\tau(h,s) \left(L_\tau(h,s) - V_h^*(s)\right)$$
$$\leq \mathcal{O}\left(H^2 C + HK\Delta + H|S|\sqrt{\frac{\iota K \log(K)}{\eta}} + \frac{H^2|S|\log(K)}{\eta}\right).$$

Proof The proof follows the same argument of Lemma 5 to use Lemma 13. Then, the claimed bound is immediate. ■

Proof [Proof of Theorem 15.] This proof mostly follows the argument of Theorem 2 and diverges from Eq. (24). Specifically, we apply Lemma 19 in Eq. (24) to get

$$\sum_{s \in S_h} \sum_{\tau \in \mathcal{E}(h,s)} N_\tau(h,s) \left(L_{\tau-1}(h,s) - L_\tau(h,s)\right)$$
$$\leq \mathcal{O}\left(\eta H^2 C + \eta HK\Delta + H|S|\sqrt{\eta\iota K \log(K)} + \frac{H|S|\log(K)}{\eta}\right)$$
$$\leq \mathcal{O}\left(\eta H^2 C + H|S|\sqrt{\eta\iota K \log(K)} + \frac{H|S|\log(K)}{\eta}\right), \tag{34}$$

where the second inequality uses $\Delta = 1/K$ and $\eta \leq 1/H$ to bound $\eta HK\Delta$ by a constant.

Plugging Eq. (34) and Eq. (23) into Eq. (22), we have that for any step $h \in [H]$

$$\sum_{k=1}^{K} (1)_h^k \leq \mathcal{O}\left(\eta H^2 C + H|S|\sqrt{\eta\iota K \log(K)} + \sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H|S|\log(K)}{\eta}\right). \tag{35}$$

Putting together. Plugging Eq. (35) and Eq. (21) into Eq. (20), we have that for any step $h \in [H]$

$$\sum_{k=1}^{K} \delta_h^k \leq \sum_{k=1}^{K} \delta_{h+1}^k + \mathcal{O}\left(\eta H^2 C + H|S|\sqrt{\eta\iota K \log(K)} + \sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H|S|\log(K)}{\eta}\right).$$

Summing over all $h \in [H]$, using the fact $\delta_{H+1}^k = 0$ for any k, and rearranging, we have

$$\text{ENR}_K \leq \sum_{k=1}^{K} \delta_1^k \leq \mathcal{O}\left(\eta H^3 C + H^2|S|\sqrt{\eta\iota K \log(K)} + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H^2|S|\log(K)}{\eta}\right)$$
$$\leq \mathcal{O}\left(\eta H^3 C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}} + \frac{H^2|S|\log(K)}{\eta}\right),$$

where the second inequality uses the choice $\eta \leq 1/(H\sqrt{|S|})$ to bound $H^2|S|\sqrt{\eta\iota K\log(K)} \leq H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}$.

Furthermore, if one assumes $K \geq H|S|^{3/2}$ and constrains $\eta \in \left[\frac{|S|}{K}, \frac{1}{H\sqrt{|S|}}\right]$, then,

$$\text{ENR}_K \leq \sum_{k=1}^{K} \left(V_1^k(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) \leq \mathcal{O}\left(\eta H^3 C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\right).$$

The proof is thus complete. ■

B.2. Regret Bound under NR_K

In fact, if one evaluates the regret using NR_K, as in (Tian et al., 2021), then the discretization step can be avoided, and the same analysis as in Theorem 6 applies, yielding an identical regret bound. The result of Algorithm 1 under NR_K is presented in the following.

Theorem 20 *Suppose that the opponent is adaptive and $K \geq H|S|$. If we run Algorithm 1 with $\eta \in [|S|/K, 1/H]$ and the adversarial bandit subroutine is instantiated by Example 1 (so $\iota = \Theta(H^2|A|\log(HK|A||S|/\delta))$), then with probability at least $1-\delta$, $\text{NR}_K \leq \mathcal{O}\Big(\eta H^3 C + H\sqrt{\frac{\iota|S|K\log(K)}{\eta}}\Big)$.*

Appendix C. Regret Bound of Adaptive Epoch V-learning for Oblivious Opponent

Before proving our main results, we first introduce some notations.

C.1. Notations.

Given any interval $I \subseteq [K]$, we recursively define the locally empirical state Nash values: for each $(h,s) \in [H] \times S_h$,

$$V_{h,I}^*(s) = \max_{\mu \in \Delta(A_h)} \min_{\nu \in \left\{\nu_h^k(s)\right\}_{k \in I}} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1,I}^*\right](s), \tag{36}$$

with $V_{H+1,I}^*(s) = 0$ for all $s \in S_{H+1}$. Based on this, we further define for any interval I

$$\text{ENR}_\ell^{(b)}(I) = \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_{1,I}^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right). \tag{37}$$

If one chooses $I = [K]$, then $\text{ENR}_\ell^{(b)}([K])$ is the local regret in sub-block ℓ in block b. Note that since each sub-block runs a new instance of epoch V-learning algorithm, we can even directly work on $\text{ENR}_\ell^{(b)}(\mathcal{T}_\ell^{(b)})$, a stronger regret metric than $\text{ENR}_\ell^{(b)}([K])$, and obtain a regret bound in the same order. Indeed, from the definition of $V_{h,I}^*(s)$ in Eq. (36), one can see that $V_{1,[K]}^*(s_1^k) \leq V_{1,\mathcal{T}_\ell^{(b)}}^*(s_1^k)$, and thus

$$\text{ENR}_\ell^{(b)}([K]) = \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_{1,[K]}^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) \leq \text{ENR}_\ell^{(b)}(\mathcal{T}_\ell^{(b)}).$$

For each step h, we define the minimax policy for the opponent when restricted to play a mixture of empirical polices $\{\nu_h^k(s)\}_{k\in I}$ in the interval $I \subseteq [K]$

$$\nu_{h,I}^*(s) \in \operatorname*{argmin}_{\nu\in \mathtt{Conv}\left(\{\nu_h^k(s)\}_{k\in I}\right)} \max_{\mu\in\Delta(A_h)} \mathbb{D}_{\mu,\nu}\left[r_h + P_h V_{h+1,I}^*\right](s).$$

For any contiguous interval $I \subseteq [K]$, we define

$$\mathcal{S}(I) := \{[a,b] : 1 \le a \le b \le K, [a,b] \subseteq [K], I \subseteq J\}.$$

Let us define

- B as the total number of blocks when K episodes end.
- Z_b as the total number of sub-blocks in block b when K episodes end.
- $L^{(b)} := 1 + \sum_{\ell=1}^{Z_b} \sum_{k\in\mathcal{T}_\ell^{(b)}\setminus\{K\}} \mathbb{I}\{\nu^k \neq \nu^{k+1}\}$ be the number of policy switches by the opponent (plus one).

As each sub-block runs a new instance of epoch V-learning algorithm, we define another form of non-stationary measure:

$$\widetilde{C} = \sum_{b=1}^{B}\sum_{\ell=1}^{Z_b} \widetilde{C}_\ell^{(b)}, \quad \widetilde{C}_\ell^{(b)} = \min_{I\in\mathcal{S}(\mathcal{T}_\ell^{(b)})} \widetilde{C}_\ell^{(b)}(I), \quad \widetilde{C}_\ell^{(b)}(I) := \sum_{h=1}^{H}\sum_{k\in\mathcal{T}_\ell^{(b)}} \mathtt{TV}\left(\nu_h^k(s_h^k), \nu_{h,I}^*(s_h^k)\right).$$

Here, $\widetilde{C}$ sums the non-stationarity over all sub-blocks. As $[K] \in \mathcal{S}(I)$ for any I, we have $\widetilde{C} \le C$ by the definition.

C.2. Construction of Nice Event

Let $c_0 \ge 2$ be some absolute constant. We define

$$\mathcal{G}_5 := \left\{\forall \text{ contiguous interval } I \subseteq [K] : \left|\sum_{k\in I}\left(\sum_{h=1}^{H} r_h^k - V_1^{\mu^k,\nu^k}(s_1^k)\right)\right| \le \sqrt{\iota|I|}\right\}, \tag{38}$$

$$\mathcal{G}_6 := \left\{\forall b,\ell : \textsc{Enr}_\ell^{(b)}([K]) \le \sum_{k\in\mathcal{T}_\ell^{(b)}} \delta_1^k \le c_0\left(\eta_\ell^{(b)} H^3 \widetilde{C}_\ell^{(b)} + H\sqrt{\frac{\iota|S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right)\right\}. \tag{39}$$

Definition 21 *Let $\overline{\mathcal{G}}$ be the nice event $\overline{\mathcal{G}} = \mathcal{G}_5 \cap \mathcal{G}_6$.*

Lemma 22 *Suppose that $K \ge 16H^2|S|$. We have $\mathbb{P}(\overline{\mathcal{G}}) \ge 1-\delta$.*

Proof We first prove $\mathbb{P}(\mathcal{G}_5) \ge 1-\delta/2$. Consider any fixed contiguous interval $I \subseteq [K]$, containing consecutive rounds. Since $\left|\left(\sum_{h=1}^{H} r_h^k - V_1^{\mu^k,\nu^k}(s_1^k)\right)\right| \le H$ for any episode k, applying Azuma-Hoeffding inequality for the interval I ensures with probability at least $1-\delta/(2K^2)$,

$$\left|\sum_{k\in I}\left(\sum_{h=1}^{H} r_h^k - V_1^{\mu^k,\nu^k}(s_1^k)\right)\right| \le H\sqrt{2|I|\log\left(4K^2/\delta\right)} \le \sqrt{\iota|I|}.$$

As $I \subseteq [K]$ is a set of consecutive rounds, there are at most $K(K+1)/2 \leq K^2$ such intervals. Using a union bound over all possible intervals gives that $\mathbb{P}(\mathcal{G}_5) \geq 1 - \delta/2$.

We then show $\mathbb{P}(\mathcal{G}_6) \geq 1 - \delta/2$. Consider any fixed block b and any fixed sub-block ℓ in block b. We fix a contiguous interval $\mathcal{T}_\ell^{(b)} \subseteq [K]$, containing consecutive episodes, and further fix an interval $I \in \mathcal{S}(\mathcal{T}_\ell^{(b)})$ (i.e., $\mathcal{T}_\ell^{(b)} \subseteq I$). Since Algorithm 2 runs a new instance of Algorithm 1 with input total episode K, confidence $\delta/(2K^6)$, and epoch incremental factor $\eta_\ell^{(b)} \in [|S|/K, 1/H]$, Theorem 2 ensures that with probability at least $1 - \delta/(2K^6)$,

$$\begin{aligned}\textsc{Enr}_\ell^{(b)}(I) &= \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_{1,I}^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) \\ &\leq \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_1^k(s_1^k) - V_1^{u^k,v^k}(s_1^k)\right) \leq c_0 \left(\eta_\ell^{(b)} H^3 \widetilde{C}_\ell^{(b)}(I) + H\sqrt{\frac{\iota |S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right),\end{aligned}$$

where $c_0 > 0$ is some absolute constant, and we further constrain $c_0 \geq 2$. We here highlight that when applying Theorem 2, one just replaces each $V_h^*(s)$ by $V_{h,I}^*(s)$ to obtain the claimed bound since $\mathcal{T}_\ell^{(b)} \subseteq I$ ensures $V_{h,I}^*(s)$ to preserve the optimism.

Based on the fact that $V_{1,I}^*(s_1^k) \leq V_{1,I'}^*(s_1^k)$ for $I' \subseteq I$, and $[K] \in \mathcal{S}(\mathcal{T}_\ell^{(b)})$, we have

$$\textsc{Enr}_\ell^{(b)}([K]) = \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_{1,[K]}^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right) = \min_{I \in \mathcal{S}(\mathcal{T}_\ell^{(b)})} \textsc{Enr}(I). \tag{40}$$

Since $|\mathcal{S}(\mathcal{T}_\ell^{(b)})| \leq K^2$ for any $\mathcal{T}_\ell^{(b)} \subseteq [K]$, using a union bound over all intervals in $\mathcal{S}(\mathcal{T}_\ell^{(b)})$, we have that with probability at least $1 - \delta(2K^4)$

$$\begin{aligned}\textsc{Enr}_\ell^{(b)}([K]) &\leq \min_{I \in \mathcal{S}(\mathcal{T}_\ell^{(b)})} c_0 \left(\eta_\ell^{(b)} H^3 \widetilde{C}_\ell^{(b)}(I) + H\sqrt{\frac{\iota |S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right) \\ &= c_0 \left(\eta_\ell^{(b)} H^3 \widetilde{C}_\ell^{(b)} + H\sqrt{\frac{\iota |S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right).\end{aligned}$$

As the total number of such interval $\mathcal{T}_\ell^{(b)}$ is also bounded by K^2, and the number of blocks and sub-blocks are bounded by K, applying a union bound over all $b, \ell, T_\ell^{(b)}$ yields $\mathbb{P}(\mathcal{G}_6) \geq 1 - \delta/2$.

Finally, the claimed result follows by using a union bound over $\mathcal{G}_5, \mathcal{G}_6$. ■

C.3. Supporting Lemmas

Lemma 23 *Suppose that $K \geq 16H^2|S|$ and $\overline{\mathcal{G}}$ holds where $\overline{\mathcal{G}}$ is given in Definition 21. There exists an absolute constant $c_0 \geq 2$ given in Eq. (39) such that for any block b, and any sub-block ℓ in block b enjoys the following:*

$$\Phi_\ell^{(b)} \leq 2c_0 \left(\eta_\ell^{(b)} H^3 \widetilde{C}_\ell^{(b)} + H\sqrt{\frac{\iota |S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right).$$

Proof For any block b and any sub-block ℓ in this block, one can show

$$\begin{aligned}\Phi_\ell^{(b)} &= \sum_{k\in\mathcal{T}_\ell^{(b)}} \left(V_1^k(s_1^k) - \sum_{h=1}^H r_h^k\right) + \sqrt{\iota\left|\mathcal{T}_\ell^{(b)}\right|} \\ &\le \sum_{k\in\mathcal{T}_\ell^{(b)}} \left(V_1^k(s_1^k) - V_1^{u^k,v^k}(s_1^k)\right) + 2\sqrt{\iota|\mathcal{T}_\ell^{(b)}|} \\ &\le c_0\left(\eta_\ell^{(b)}H^3\widetilde{C}_\ell^{(b)} + H\sqrt{\frac{\iota|S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right) + 2\sqrt{\iota|\mathcal{T}_\ell^{(b)}|} \\ &\le 2c_0\left(\eta_\ell^{(b)}H^3\widetilde{C}_\ell^{(b)} + H\sqrt{\frac{\iota|S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}}\right),\end{aligned}$$

where the first inequality uses $\mathcal{G}_5$ of $\overline{\mathcal{G}}$ and the definition of ι, and the third inequality follows from $\mathcal{G}_6$ of $\overline{\mathcal{G}}$. ■

Lemma 24 *Suppose that $\overline{\mathcal{G}}$ holds and $K \ge 16H^2|S|$. We have*

$$B \le 1 + \log_4^+\left(4\left(\frac{H(1+\widetilde{C})^2}{\iota|S|K\log(K)}\right)^{1/3}\right),$$

where $\log_4^+(x) = \max\{\log_4(x), 0\}$. Moreover, $\eta_{2^{2b}+1}^{(b)} = \frac{2^{-2b}}{H}$ for all $b \le B$.

Proof If $B = 1$, then the claimed bound on B holds trivially. Then, we consider the case $B > 1$. It suffices to show that there exists a block such that the last sub-block termination condition will not be met. From Lemma 23, there exists an absolute constant $c_0 \ge 2$ such that for any block b, the last sub-block $\ell = 2^{2b} + 1$ enjoys the following

$$\Phi_\ell^{(b)} \le 2c_0\left(\eta_\ell^{(b)}H^3(1+\widetilde{C}) + H\sqrt{\frac{\iota|S|K\log(K)}{\eta_\ell^{(b)}}}\right),$$

where the inequality simply bounds $|\mathcal{T}_\ell^{(b)}| \le K$ and $\widetilde{C}_\ell^{(b)} \le 1 + \widetilde{C}$.

For shorthand, we use $\Lambda_b = \eta_{2^{2b}+1}^{(b)}$ to denote the epoch incremental factor for the last sub-block of block b. Since Λ_b is non-increasing w.r.t. b and $\widetilde{C} \le HK$, there exist at least one blocks $b \le \log_4\left(\frac{K}{|S|H}\right)$ such that $\Lambda_b H^3(1+\widetilde{C}) \le H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_b}}$. Then, the existence can be verified by using the assumption $K \ge 16H^2|S|$ to show that for $b = \lfloor\log_4\left(\frac{K}{|S|H}\right)\rfloor$, we have

$$\Lambda_b H^3(1+\widetilde{C}) = \frac{H^2(1+\widetilde{C})}{2^{2b}} \le \frac{2KH^3}{2^{2b}} \le 8|S|H^4 \le \frac{1}{2}KH^2 \le H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_b}},$$

where the first equality follows from the fact that $\Lambda_b = \frac{2^{-2b}}{H}$ since $b < \log_4(K/(|S|H))$ implies $2^{-2b}/H \geq |S|/K$, the first inequality bounds $1 + \widetilde{C} \leq 2KH$, the second inequality bounds $b = \lfloor \log_4\left(\frac{K}{|S|H}\right) \rfloor \geq \log_4\left(\frac{K}{|S|H}\right) - 1$, the third inequality uses the assumption $K \geq 16H^2|S|$, and the last inequality follows from facts that $\iota \geq H^2$ and $\Lambda_b \leq H^{-1}4^{-\log_4\left(\frac{K}{|S|H}\right)+1} = 4|S|/K$.

Furthermore, $B > 1$ also implies that there exist at least one blocks $b \leq \log_4\left(\frac{K}{|S|H}\right)$ such that $\Lambda_b H^3(1+\widetilde{C}) > H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_b}}$. Thus, there should exist a block $\hat{b} \leq \log_4\left(\frac{K}{|S|H}\right)$ such that

$$\Lambda_{\hat{b}} H^3(1+\widetilde{C}) \leq H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}} \quad \text{and} \quad \Lambda_{\hat{b}-1} H^3(1+\widetilde{C}) > H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}-1}}}. \tag{41}$$

In such a block $\hat{b}$, the last sub-block $\ell = 2^{2\hat{b}} + 1$ satisfies

$$\Phi_\ell^{(\hat{b})} \leq 2c_0\left(\Lambda_{\hat{b}} H^3 \widetilde{C} + H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}}\right) \leq 4c_0 H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}},$$

where the first inequality uses Lemma 23, and the second inequality uses $\Lambda_{\hat{b}} H^3(1+\widetilde{C}) \leq H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}}$ given in Eq. (43). Thus, the block termination condition will never be met in block $\hat{b}$. Moreover, $\hat{b} \leq \log_4\left(\frac{K}{|S|H}\right)$ implies that for all $b \leq \hat{b}$, $\Lambda_b = \frac{2^{-2b}}{H} \geq \frac{|S|}{K}$. Notice that $\Lambda_{\hat{b}-1} H^3(1+\widetilde{C}) > H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}-1}}}$ gives

$$\frac{2^{-2(\hat{b}-1)}}{H} = \Lambda_{\hat{b}-1} \geq \left(\frac{\iota|S|K\log(K)}{(1+\widetilde{C})^2 H^4}\right)^{1/3} \Longrightarrow \hat{b} \leq \log_4\left(4\left(\frac{H(1+\widetilde{C})^2}{\iota|S|K\log(K)}\right)^{1/3}\right).$$

Combining two cases of $B = 1$ and $B > 1$, we obtain the claimed bound on B.

Finally, as $\Lambda_b = \frac{2^{-2b}}{H}$ for all $b \leq \log_4(\frac{K}{|S|H})$ and we know the upper bound of B, it suffices to show that $\left(\frac{4^3 H(1+\widetilde{C})^2}{\iota|S|K\log(K)}\right)^{1/3} \leq K/(|S|H)$ to conclude the proof. One can easily verify that

$$\frac{4^3 H(1+\widetilde{C})^2}{\iota|S|K\log(K)} \leq \frac{4^4 H^3 K}{\iota|S|} \leq \frac{4^4 HK}{|S|} = \frac{4^4 HK^3}{|S|K^2} \leq \frac{K^3}{|S|^3 H^3},$$

where the first inequality bounds $(1+\widetilde{C})^2 \leq (2HK)^2$ and $\log(K) \geq 1$, the second inequality bounds $\iota \geq H^2$, and the last inequality uses the assumption that $K \geq 16H^2|S|$.

The proof is thus complete. ■

Lemma 25 *Suppose that $\overline{\mathcal{G}}$ holds and $K \geq 16H^2|S|$. For any block b and any sub-block $\ell \leq 2^{2b}$ of block b, if $v^k = v^{k+1}$ for all $k \in \mathcal{T}_\ell^{(b)}$, then*

$$\Phi_\ell^{(b)} \leq 2c_0\sqrt{H^3\iota|S|\left|\mathcal{T}_\ell^{(b)}\right|\log(K)},$$

where $c_0 \geq 2$ is an absolute constant given in Eq. (39).

Proof Recall that Algorithm 2 runs a new instance of Algorithm 1 in each sub-block. If the opponent uses a fixed policy in a sub-block, then, $\widetilde{C}_\ell^{(b)} = 0$. Hence, Lemma 23 gives that $\Phi_\ell^{(b)} \le 2c_0 H\sqrt{\frac{\iota|S||\mathcal{T}_\ell^{(b)}|\log(K)}{\eta_\ell^{(b)}}} = 2c_0\sqrt{H^3\iota|S|\left|\mathcal{T}_\ell^{(b)}\right|\log(K)}$ where the equality holds since for any block b, $\eta_\ell^{(b)} = H^{-1}$ for all sub-blocks $\ell \le 2^{2b}$. ■

Lemma 26 *Suppose that $\overline{\mathcal{G}}$ holds and $K \ge 16H^2|S|$. For each block b, we have*

$$\sum_{\ell=1}^{Z_b}\Phi_\ell^{(b)} \le \mathcal{O}\big(\sqrt{\min\{2^{2b}, L^{(b)}\}H^3\iota|S|K\log(K)}\big).$$

Proof Consider any block b. We then consider the following two cases.

Case 1: $Z_b \le 2^{2b}$. In this case, when a sub-block ℓ in block b ends, we have $\Phi_\ell^{(b)} \le \mathcal{O}\big(H\sqrt{\frac{\iota|S||\mathcal{T}_\ell|\log(K)}{\eta_\ell^{(b)}}}\big)$. Since $Z_b \le 2^{2b}$ implies that $\eta_\ell^{(b)} = 1/H$ for all $\ell \le Z_b \le 2^{2b}$, we have

$$\sum_{\ell=1}^{Z_b}\Phi_\ell^{(b)} \le \mathcal{O}\left(\sum_{\ell=1}^{Z_b}\sqrt{H^3\iota|S||\mathcal{T}_\ell|\log(K)}\right) \le \mathcal{O}\left(\sqrt{Z_b H^3\iota|S|K\log(K)}\right),$$

where the last inequality uses Cauchy–Schwarz inequality and $\sum_{\ell=1}^{Z_b}|\mathcal{T}_\ell| \le K$.

Case 2: $Z_b = 2^{2b}+1$. By the termination condition of sub-block $Z_b = 2^{2b}+1$, we have

$$\Phi_{Z_b}^{(b)} \le \mathcal{O}\left(H\sqrt{\frac{\iota|S|K\log(K)}{\eta_{Z_b}^{(b)}}}\right) = \mathcal{O}\left(\sqrt{2^{2b}H^3\iota|S|K\log(K)}\right) \le \mathcal{O}\left(\sqrt{Z_b H^3\iota|S|K\log(K)}\right).$$

where the equality uses Lemma 24 that $\eta_{Z_b}^{(b)} = 2^{-2b}/H$ and the last inequality uses $2^{2b} \le Z_b$.

Then, one can show that

$$\sum_{\ell=1}^{Z_b}\Phi_\ell^{(b)} = \sum_{\ell=1}^{Z_b-1}\Phi_\ell^{(b)} + \Phi_{Z_b}^{(b)} \le \mathcal{O}\left(\sqrt{Z_b H^3\iota|S|K\log(K)}\right),$$

where the first summation can be bounded via the same argument used in Case 1.

Putting together. In both cases, we have $\sum_{\ell=1}^{Z_b}\Phi_\ell^{(b)} \le \mathcal{O}\left(\sqrt{Z_b H^3\iota|S|K\log(K)}\right)$. If the opponent always uses a fixed policy in a sub-block, then Lemma 25 implies that this sub-block never ends. In other words, if a sub-block ends, then opponent switches the policy at least once. Thus, $Z_b \le L^{(b)}$. Combining this and the fact that $Z_b \le \mathcal{O}(2^{2b})$, we obtain the desired bound. ■

Proof [Proof of Theorem 6.] Assume that $K \ge 16H^2|S|$. For every block b, we have $\textsc{Enr}_\ell^{(b)}([K]) \le \Phi_\ell^{(b)}$ since

$$\textsc{Enr}_\ell^{(b)}([K]) \le \sum_{k\in\mathcal{T}_\ell^{(b)}}\left(V_1^k(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right)$$

$$\leq \sum_{k \in \mathcal{T}_\ell^{(b)}} \left(V_1^k(s_1^k) - \sum_{h=1}^{H} r_h^k \right) + \sqrt{\iota \left| \mathcal{T}_\ell^{(b)} \right|} = \Phi_\ell^{(b)}, \tag{42}$$

where the first inequality uses the optimism of epoch-V-ol, and the second inequality uses $\mathcal{G}_5$ of $\overline{\mathcal{G}}$.

Recall that Z_b is the total number of sub-blocks in block b, and B is the total number of blocks. The regret can be written as

$$\begin{aligned}
&\textsc{Enr}_K \\
&= \sum_{b=1}^{B} \sum_{\ell=1}^{Z_b} \textsc{Enr}_\ell^{(b)}([K]) \\
&\leq \sum_{b=1}^{B} \sum_{\ell=1}^{Z_b} \Phi_\ell^{(b)} \\
&\leq \mathcal{O}\left(\sum_{b=1}^{B} \sqrt{\min\left\{2^{2b}, L^{(b)}\right\} H^3 \iota |S| K \log(K)} \right) \\
&\leq \mathcal{O}\left(\min\left\{ \sum_{b=1}^{B} 2^b, \sum_{b=1}^{B} \sqrt{L^{(b)}} \right\} \sqrt{H^3 \iota |S| K \log(K)} \right) \\
&\leq \mathcal{O}\left(\min\left\{ 2^B, \sum_{b=1}^{B} \sqrt{L^{(b)}} \right\} \sqrt{H^3 \iota |S| K \log(K)} \right) \\
&\leq \mathcal{O}\left(\min\left\{ 2^B, \sqrt{(L + \log(K)\log(K))} \right\} \sqrt{H^3 \iota |S| K \log(K)} \right) \\
&\leq \mathcal{O}\left(\min\left\{ \sqrt{H^3 \iota |S| K \log(K)} + \left(H^5 \iota |S| C K \log(K)\right)^{\frac{1}{3}}, \sqrt{(L + \log(K) H^3 \iota |S| K \log^2(K)} \right\} \right),
\end{aligned}$$

where the first inequality uses Eq. (42) to bound $\sum_{b=1}^{B} \sum_{\ell=1}^{Z_b} \textsc{Enr}_\ell^{(b)}([K]) \leq \sum_{b=1}^{B} \sum_{\ell=1}^{Z_b} \Phi_\ell^{(b)}$, the second inequality follows from Lemma 26, the fifth inequality uses Cauchy–Schwarz inequality and the fact that $B \leq \mathcal{O}(\log(K))$ and $\sum_{b=1}^{B} L^{(b)} \leq \mathcal{O}(B + L) \leq \mathcal{O}(\log(K) + L)$, and the last inequality uses Lemma 24 to bound B together with the fact that $\widetilde{C} \leq C$.

Finally, if $K \leq 16 H^2 |S|$, one can bound regret by $\mathcal{O}(H^2 |S|)$. The claimed bound thus follows. ■

Appendix D. Regret Bound of Adaptive Epoch V-learning for Adaptive Opponent

Theorem 27 *For adaptive adversary, running Algorithm 2 by choosing $\eta_\ell^{(b)}$ as: $\eta_\ell^{(b)} = 1/(H\sqrt{|S|})$ for all $\ell \leq 2^{2b}$ and $\eta_\ell^{(b)} = \max\left\{2^{-2b}/(H\sqrt{|S|}), |S|/K\right\}$ otherwise, with probability at least $1 - \delta$*

$$\textsc{Enr}_K \leq \widetilde{\mathcal{O}}\left(H^2 |S|^{3/2} + \min\left\{ |S|^{3/4} \sqrt{H^3 \iota K} + \left(\iota H^5 |S| K C\right)^{\frac{1}{3}}, |S|^{3/4} \sqrt{L H^3 \iota K} \right\} \right).$$

Further, running Algorithm 2 with the choice of $\eta_\ell^{(b)}$ specified in Eq. (12) guarantees, with probability at least $1 - \delta$

$$\textsc{Nr}_K \leq \widetilde{\mathcal{O}}\left(H^2 |S| + \min\left\{ \sqrt{H^3 \iota |S| K} + \left(\iota H^5 |S| K C\right)^{\frac{1}{3}}, \sqrt{L H^3 \iota |S| K} \right\} \right).$$

D.1. Supporting Lemmas

Nice event $\overline{\mathcal{G}}$. We use the nice event $\overline{\mathcal{G}}$ defined in Definition 21 for the following analysis. For $\mathcal{G}_6$ given in Eq. (39), one can again show $\mathbb{P}(\mathcal{G}_6) \geq 1 - \delta/2$ by using Theorem 27.

Lemma 28 *Suppose that $\overline{\mathcal{G}}$ holds and $K \geq 16H^2|S|^{3/2}$. We have*

$$B \leq 1 + \log_4^+ \left(4 \left(\frac{H(1+\widetilde{C})^2}{\iota |S|^{5/2} K \log(K)} \right)^{1/3} \right),$$

where $\log_4^+(x) = \max\{\log_4(x), 0\}$. Moreover, $\eta^{(b)}_{2^{2b}+1} = \frac{2^{-2b}}{H\sqrt{|S|}}$ for all $b \leq B$.

Proof This proof follows a similar argument of Lemma 24 with minor modifications for a different schedule of $\eta^{(b)}_{2^{2b}+1}$ and a different constraint $K \geq 16H^2|S|^{3/2}$. If $B = 1$, then the claimed bound on B holds trivially. Then, we consider the case $B > 1$. It suffices to show that there exists a block such that the last sub-block termination condition will not be met. From Lemma 23, there exists an absolute constant $c_0 \geq 2$ such that for any block b, the last sub-block $\ell = 2^{2b} + 1$ enjoys the following

$$\Phi^{(b)}_\ell \leq 2c_0 \left(\eta^{(b)}_\ell H^3 (1 + \widetilde{C}) + H \sqrt{\frac{\iota |S| K \log(K)}{\eta^{(b)}_\ell}} \right).$$

For shorthand, we use $\Lambda_b = \eta^{(b)}_{2^{2b}+1}$ to denote the epoch incremental factor for the last sub-block of block b. Since Λ_b is non-increasing w.r.t. b and $\widetilde{C} \leq HK$, there exist at least one blocks $b \leq \log_4\left(\frac{K}{|S|^{3/2}H}\right)$ such that $\Lambda_b H^3(1+\widetilde{C}) \leq H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_b}}$. Then, the existence can be verified by using the assumption $K \geq 16H^2|S|^{3/2}$ to show that for $b = \lfloor \log_4\left(\frac{K}{|S|^{3/2}H}\right) \rfloor$, we have

$$\Lambda_b H^3(1+\widetilde{C}) = \frac{H^2(1+\widetilde{C})}{2^{2b}\sqrt{|S|}} \leq \frac{2KH^3}{2^{2b}\sqrt{|S|}} \leq 8|S|H^4 \leq \frac{1}{2}KH^2 \leq H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_b}},$$

where the first equality follows from the fact that $\Lambda_b = \frac{2^{-2b}}{H\sqrt{|S|}}$ since $b < \log_4(K/(|S|^{3/2}H))$ implies $2^{-2b}/(H\sqrt{|S|}) \geq |S|/K$, the first inequality bounds $1 + \widetilde{C} \leq 2KH$, the second inequality bounds $b = \lfloor \log_4\left(\frac{K}{|S|^{3/2}H}\right) \rfloor \geq \log_4\left(\frac{K}{|S|^{3/2}H}\right) - 1$, the third inequality uses the assumption $K \geq 16H^2|S|^{3/2}$, and the last inequality follows from facts that $\iota \geq H^2$ and $\Lambda_b \leq \frac{1}{H\sqrt{|S|}} 4^{-\log_4\left(\frac{K}{|S|^{3/2}H}\right)+1} = 4|S|/K$.

Furthermore, $B > 1$ also implies that there exist at least one blocks $b \leq \log_4\left(\frac{K}{|S|^{3/2}H}\right)$ such that $\Lambda_b H^3(1+\widetilde{C}) > H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_b}}$. Thus, there should exist a block $\hat{b} \leq \log_4\left(\frac{K}{|S|^{3/2}H}\right)$ such that

$$\Lambda_{\hat{b}} H^3(1+\widetilde{C}) \leq H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}} \quad \text{and} \quad \Lambda_{\hat{b}-1} H^3(1+\widetilde{C}) > H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}-1}}}. \tag{43}$$

In such a block $\hat{b}$, the last sub-block $\ell = 2^{2\hat{b}} + 1$ satisfies

$$\Phi_\ell^{(\hat{b})} \le 2c_0 \left(\Lambda_{\hat{b}} H^3 \widetilde{C} + H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}} \right) \le 4c_0 H \sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}},$$

where the first inequality uses Lemma 23, and the second inequality uses $\Lambda_{\hat{b}} H^3(1+\widetilde{C}) \le H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}}}}$ given in Eq. (43). Thus, the block termination condition will never be met in block $\hat{b}$. Moreover, $\hat{b} \le \log_4\left(\frac{K}{|S|^{3/2}H}\right)$ implies that for all $b \le \hat{b}$, $\Lambda_b = \frac{2^{-2b}}{H\sqrt{|S|}} \ge \frac{|S|}{K}$. Notice that $\Lambda_{\hat{b}-1} H^3(1+\widetilde{C}) > H\sqrt{\frac{\iota|S|K\log(K)}{\Lambda_{\hat{b}-1}}}$ gives

$$\frac{2^{-2(\hat{b}-1)}}{H\sqrt{|S|}} = \Lambda_{\hat{b}-1} \ge \left(\frac{\iota|S|K\log(K)}{(1+\widetilde{C})^2 H^4}\right)^{1/3} \Longrightarrow \hat{b} \le \log_4\left(4\left(\frac{H(1+\widetilde{C})^2}{\iota|S|^{5/2}K\log(K)}\right)^{1/3}\right).$$

Combining two cases of $B = 1$ and $B > 1$, we obtain the claimed bound on B.

Finally, as $\Lambda_b = \frac{2^{-2b}}{H\sqrt{|S|}}$ for all $b \le \log_4(\frac{K}{|S|^{3/2}H})$ and we know the upper bound of B, it suffices to show that $\left(\frac{4^3 H(1+\widetilde{C})^2}{\iota|S|^{5/2}K\log(K)}\right)^{1/3} \le K/(|S|^{3/2}H)$ to conclude the proof. One can easily verify that

$$\frac{4^3 H(1+\widetilde{C})^2}{\iota|S|^{5/2}K\log(K)} \le \frac{4^4 H^3 K}{\iota|S|^{5/2}} \le \frac{4^4 HK}{|S|^{5/2}} = \frac{4^4 HK^3}{|S|^{5/2}K^2} \le \frac{K^3}{|S|^{27/8}H^3},$$

where the first inequality bounds $(1+\widetilde{C})^2 \le (2HK)^2$ and $\log(K) \ge 1$, the second inequality bounds $\iota \ge H^2$, and the last inequality uses the assumption that $K \ge 16H^2|S|^{3/2}$.

The proof is thus complete. ■

Lemma 29 *Suppose that $\overline{\mathcal{G}}$ holds and $K \ge 16H^2|S|^{3/2}$. For any block b and any sub-block $\ell \le 2^{2b}$ of block b, if $\nu^k = \nu^{k+1}$ for all $k \in \mathcal{T}_\ell^{(b)}$, then*

$$\Phi_\ell^{(b)} \le 2c_0|S|^{3/4}\sqrt{H^3\iota\left|\mathcal{T}_\ell^{(b)}\right|\log(K)},$$

where $c_0 \ge 2$ is an absolute constant given in Eq. (39).

Proof This proof follows from a similar argument of Lemma 25 with a different choice of $\eta_\ell^{(b)}$. ■

Lemma 30 *Suppose that $\overline{\mathcal{G}}$ holds and $K \ge 16H^2|S|^{3/2}$. For each block b, we have*

$$\sum_{\ell=1}^{Z_b} \Phi_\ell^{(b)} \le \mathcal{O}\big(|S|^{3/4}\sqrt{\min\{2^{2b}, L^{(b)}\}H^3\iota K\log(K)}\big).$$

Proof This proof follows from a similar argument of Lemma 26 with a different choice of $\eta_\ell^{(b)}$. ■

D.2. Proof of Theorem 27

Assume that $K \geq 16H^2|S|^{3/2}$. Recall that Z_b is the total number of sub-blocks in block b. The regret can be written as

$$
\begin{aligned}
&\textsc{Enr}_K \\
&= \sum_{b=1}^{B}\sum_{\ell=1}^{Z_b} R_\ell^{(b)}([K]) \\
&\leq \sum_{b=1}^{B}\sum_{\ell=1}^{Z_b} \Phi_\ell^{(b)} \\
&\leq \mathcal{O}\left(\sum_{b=1}^{B} |S|^{3/4}\sqrt{\min\left\{2^{2b}, L^{(b)}\right\} H^3\iota K\log(K)}\right) \\
&\leq \mathcal{O}\left(\min\left\{\sum_{b=1}^{B} 2^b, \sum_{b=1}^{B}\sqrt{L^{(b)}}\right\}|S|^{3/4}\sqrt{H^3\iota K\log(K)}\right) \\
&\leq \mathcal{O}\left(\min\left\{2^B, \sum_{b=1}^{B}\sqrt{L^{(b)}}\right\}|S|^{3/4}\sqrt{H^3\iota K\log(K)}\right) \\
&\leq \mathcal{O}\left(\min\left\{2^B, \sqrt{(L+\log(K))\log(K)}\right\}|S|^{3/4}\sqrt{H^3\iota K\log(K)}\right) \\
&\leq \mathcal{O}\left(\min\left\{|S|^{\frac{3}{4}}\sqrt{H^3\iota K\log(K)} + \left(H^5\iota|S|CK\log(K)\right)^{\frac{1}{3}}, |S|^{\frac{3}{4}}\sqrt{(L+\log(K))H^3\iota K\log^2(K)}\right\}\right),
\end{aligned}
$$

where the first inequality uses Eq. (42) to bound $\sum_{b=1}^{B}\sum_{\ell=1}^{Z_b}\textsc{Enr}_\ell^{(b)}([K]) \leq \sum_{b=1}^{B}\sum_{\ell=1}^{Z_b}\Phi_\ell^{(b)}$, the second inequality follows from Lemma 30, the fifth inequality uses Cauchy–Schwarz inequality and the fact that $B \leq \mathcal{O}(\log(K))$ and $\sum_{b=1}^{B} L^{(b)} \leq \mathcal{O}(B+L) \leq \mathcal{O}(\log(K)+L)$, and the last inequality uses Lemma 28 to bound B together with the fact that $\widetilde{C} \leq C$.

Finally, if $K \leq 16H^2|S|^{3/2}$, one can bound regret by $\mathcal{O}(H^2|S|^{3/2})$. The claimed bound thus follows.