The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition

Tiancheng Jin
University of Southern California
tiancheng.jin@usc.edu

Longbo Huang
Tsinghua University
longbohuang@tsinghua.edu.cn

Haipeng Luo
University of Southern California
haipengl@usc.edu

Abstract

We consider the best-of-both-worlds problem for learning an episodic Markov Decision Process through $T$ episodes, with the goal of achieving $\tilde{O}(\sqrt{T})$ regret when the losses are adversarial and simultaneously $O(\text{polylog}(T))$ regret when the losses are (almost) stochastic. Recent work by [Jin and Luo, 2020] achieves this goal when the fixed transition is known, and leaves the case of unknown transition as a major open question. In this work, we resolve this open problem by using the same Follow-the-Regularized-Leader (FTRL) framework together with a set of new techniques. Specifically, we first propose a loss-shifting trick in the FTRL analysis, which greatly simplifies the approach of [Jin and Luo, 2020] and already improves their results for the known transition case. Then, we extend this idea to the unknown transition case and develop a novel analysis which upper bounds the transition estimation error by (a fraction of) the regret itself in the stochastic setting, a key property to ensure $O(\text{polylog}(T))$ regret.

1 Introduction

We study the problem of learning finite-horizon Markov Decision Processes (MDPs) with unknown transition through $T$ episodes. In each episode, the learner starts from a fixed initial state and repeats the following for a fixed number of steps: select an available action, incur some loss, and transit to the next state according to a fixed but unknown transition function. The goal of the learner is to minimize her regret, which is the difference between her total loss and that of the optimal stationary policy in hindsight.

When the losses are stochastically generated, [Simchowitz and Jamieson, 2019, Yang et al., 2021] show that $O(\log T)$ regret is achievable (ignoring dependence on some gap-dependent quantities for simplicity). On the other hand, even when the losses are adversarially generated, [Rosenberg and Mansour, 2019a, Jin et al., 2020] show that $\tilde{O}(\sqrt{T})$ regret is achievable.\(^1\) Given that the existing algorithms for these two worlds are substantially different, Jin and Luo [2020] asked the natural question of whether one can achieve the best of both worlds, that is, enjoying (poly)logarithmic regret in the stochastic world while simultaneously ensuring some worst-case robustness in the adversarial world. Taking inspiration from the bandit literature and using the classic Follow-the-regularized-Leader (FTRL) framework with a novel regularizer, they successfully achieved this goal, albeit under a strong restriction that the transition has to be known ahead of time. Since it is highly unclear how

\(^1\)Throughout the paper, we use $\tilde{O}(\cdot)$ to hide polylogarithmic terms.
to ensure that the transition estimation error is only $O(\text{polylog}(T))$, extending their results to the unknown transition case is highly challenging and was left as a key open question.

In this work, we resolve this open question and propose the first algorithm with such a best-of-both-worlds guarantee under unknown transition. Specifically, our algorithm enjoys $\tilde{O}(\sqrt{T})$ regret always, and simultaneously $O(\log^2 T)$ regret if the losses are i.i.d. samples of a fixed distribution. More generally, our polylogarithmic regret holds under a general condition similar to that of [Jin and Luo, 2020], which requires neither independence nor identical distributions. For example, it covers the corrupted i.i.d. setting where our algorithm achieves $\tilde{O}(\sqrt{C})$ regret with $C \leq T$ being the total amount of corruption.

Techniques Our results are achieved via three new techniques. First, we propose a new loss-shifting trick for the FTRL analysis when applied to MDPs. While similar ideas have been used for the special case of multi-armed bandits (e.g., [Wei and Luo, 2018, Zimmert and Seldin, 2019, Lee et al., 2020, Zimmert and Seldin, 2021]), its extension to MDPs has eluded researchers, which is also the reason why [Jin and Luo, 2020] resorts to a different approach with a highly complex analysis involving analyzing the inverse of the non-diagonal Hessian of a complicated regularizer. Instead, inspired by the well-known performance difference lemma, we design a key shifting function in the FTRL analysis, which helps reduce the variance of the stability term and eventually leads to an adaptive bound with a certain self-bounding property known to be useful for the stochastic world. To better illustrate this idea, we use the known transition case as a warm-up example in Section 3, and show that the simple Tsallis entropy regularizer (with a diagonal Hessian) is already enough to achieve the best-of-both-worlds guarantee. This not only greatly simplifies the approach of Jin and Luo [2020] (paving the way for extension to unknown transition), but also leads to bounds with better dependence on some parameters, which on its own is a notable result already.

Our second technique is a new framework to deal with unknown transition under adversarial losses, which is important for incorporating the loss-shifting trick mentioned above. Specifically, when the transition is unknown, prior works [Rosenberg and Mansour, 2019a,b, Jin et al., 2020, Lee et al., 2020a] perform FTRL over the set of all plausible occupancy measures according to a confident set of the true transition, which can be seen as a form of optimism encouraging exploration. Since our loss-shifting trick requires a fixed transition, we propose to move the optimism from the decision set of FTRL to the losses fed to FTRL. More specifically, we perform FTRL over the empirical transition in some doubling epoch schedule, and add (negative) bonuses to the loss functions so that the algorithm is optimistic and never underestimates the quality of a policy, an idea often used in the stochastic setting (e.g., [Azar et al., 2017]). See Section 4 for the details of our algorithm.

Finally, we develop a new analysis to show that the transition estimation error of our algorithm is only polylogarithmic in $T$, overcoming the most critical obstacle in achieving best-of-both-worlds. An important aspect of our analysis is to make use of the amount of underestimation of the optimal policy, a term that is often ignored since it is nonpositive for optimistic algorithms. We do so by proposing a novel decomposition of the regret inspired by the work of Simchowitz and Jamieson [2019], and show that in the stochastic world, every term in this decomposition can be bounded by a fraction of the regret itself plus some polylogarithmic terms, which is enough to conclude the final polylogarithmic regret bound. See Section 5 for a formal summary of this idea.

Related work For earlier results in each of the two worlds, we refer the readers to the systematic surveys in [Simchowitz and Jamieson, 2019, Yang et al., 2021, Jin et al., 2020]. The work closest to ours is [Jin and Luo, 2020] which assumes known transition, and as mentioned, we strictly improve their bounds and more importantly extend their results to the unknown transition case.

Two recent works [Lykouris et al., 2021, Chen et al., 2021] also consider the corrupted stochastic setting, where both the losses and the transition function can be corrupted by a total amount of $C$. This is more general than our results since we assume a fixed transition and only allow the losses to be corrupted. On the other hand, their bounds are worse than ours when specified to our setting — [Lykouris et al., 2021] ensures a gap-dependent polylogarithmic regret bound of $O(C \log^3 T + C^2)$, while [Chen et al., 2021] achieves $O(\log^3 T + C)$ but with a potentially larger gap-dependent quantity. Therefore, neither result provides a meaningful guarantee in the adversarial world when $C = T$, while our algorithm always ensures a robustness guarantee with $\tilde{O}(\sqrt{T})$ regret. Their algorithms are also very different from ours and are not based on FTRL.
The question of achieving best-of-both-worlds guarantees for the special case of multi-armed bandits was first proposed in [Bubeck and Slivkins, 2012]. Since then, many improvements using different approaches have been established over the years [Seldin and Lugosi, 2017, Wei and Luo, 2018, Lykouris et al., 2018, Gupta et al., 2019, Zimmert et al., 2019, Zimmert and Seldin, 2021, Lee et al., 2021]. One notable and perhaps surprising approach is to use the FTRL framework, originally designed only for the adversarial settings but later found to be able to automatically adapt to the stochastic settings as long as certain regularizers are applied [Wei and Luo, 2018, Zimmert et al., 2019, Zimmert and Seldin, 2021]. Our approach falls into this category, and our regularizer design is also based on these prior works. As mentioned, however, obtaining our results requires the new loss-shifting technique as well as the novel analysis on controlling the estimation error, both of which are critical to address the extra challenges presented in MDPs.

2 Preliminaries

We consider the problem of learning an episodic MDP through $T$ episodes, where the MDP is formally defined by a tuple $(S, A, L, P, \{\ell_t\}_{t=1}^T)$ with $S$ being a finite state set, $A$ being a finite action set, $L$ being the horizon, $\ell_t : S \times A \rightarrow [0, 1]$ being the loss function of episode $t$, and $P : S \times A \times S \rightarrow [0, 1]$ being the transition function so that $P(s', a) = \text{Probability of moving to state } s' \text{ after executing action } a \text{ at state } s$.

Without loss of generality [Jin et al., 2020], the MDP is assumed to have a layer structure, that is, the state set $S$ is partitioned into $L + 1$ subsets $S_0, S_1, \ldots, S_L$ such that the state transition is only possible from one layer to the next layer (in other words, $P(s'|s, a)$ must be zero unless $s \in S_k$ and $s' \in S_{k+1}$ for some $k \in \{0, \ldots, L - 1\}$). Moreover, $S_0$ contains $s_0$ only (the initial state), and $S_L$ contains $s_L$ only (the terminal state). We use $k(s)$ to represent the layer to which state $s$ belongs.

Ahead of time, the environment decides an MDP with $P$ and $\{\ell_t\}_{t=1}^T$ unknown to the learner. The interaction proceeds through $T$ episodes. In episode $t$, the learner selects a stochastic policy $\pi_t : S \times A \rightarrow [0, 1]$ where $\pi_t(a|s)$ denotes the probability of taking action $a$ at state $s$.\footnote{Note that $\pi_t(\cdot|s_L)$ is not meaningful since no action will be taken at $s_L$. For conciseness, however, we usually define functions over $S \times A$ instead of $(S \setminus \{s_L\}) \times A$.} Starting from the initial state $s'_0 = s_0$, the learner then repeatedly selects an action $a'_k$ drawn from $\pi_t(\cdot|s'_k)$, suffers loss $\ell_t(s'_k, a'_k)$, and transits to the next state $s'_{k+1} \in S_{k+1}$ for $k = 0, \ldots, L - 1$, until reaching the terminal state $s_L$. At the end of the episode, the learner receives some feedback on the loss function $\ell_t$. In the full-information setting, the learner observes the entire loss function $\ell_t$, while in the more challenging bandit feedback setting, the learner only observes the losses of those visited state-action pairs, that is, $\ell_t(s'_0, a'_0), \ldots, \ell_t(s'_{L-1}, a'_{L-1})$.

With slight abuse of notation, we denote the expected loss of a policy $\pi$ for episode $t$ by $\ell_t(\pi) = \mathbb{E} \left[ \sum_{k=0}^{L-1} \ell_t(s'_k, a'_k) \big| P, \pi \right]$, where the trajectory $\{(s_k, a_k)\}_{k=0,\ldots,L-1}$ is the generated by executing policy $\pi$ under transition $P$. The regret of the learner against some policy $\pi$ is then defined as $\text{Reg}_T(\pi) = \mathbb{E} \sum_{t=1}^T \ell_t(\pi_t) - \ell_t(\pi)$, and we denote by $\hat{\pi}$ one of the optimal policies in hindsight such that $\text{Reg}_T(\hat{\pi}) = \max_\pi \text{Reg}_T(\pi)$.

Adversarial world versus stochastic world We consider two different setups depending on how the loss functions $\ell_1, \ldots, \ell_T$ are generated. In the adversarial world, the environment decides the loss functions arbitrarily with knowledge of the learner’s algorithm (but not her randomness). In this case, the goal is to minimize the regret against the best policy $\text{Reg}_T(\pi)\,$ with the best existing upper bound being $O(L|S|\sqrt{A|T|})$ [Rosenberg and Mansour, 2019a, Jin et al., 2020] and the best lower bound being $Ω(L\sqrt{|S||A|T|})$ [Jin et al., 2018] (for both full-information and bandit feedback).

In the stochastic world, following [Jin and Luo, 2020] (which generalizes the bandit case of [Zimmert and Seldin, 2019, 2021]), we assume that the loss functions satisfy the following condition: there exists a deterministic policy $\pi^* : S \rightarrow A$, a gap function $\Delta : S \times A \rightarrow \mathbb{R}_+$ and a constant $C > 0$ such that

$$\text{Reg}_T(\pi^*) \geq \mathbb{E} \left[ \sum_{t=1}^T \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta(s, a) \right] - C, \quad (1)$$
where \( q_\ell(s,a) \) is the probability of the learner visiting \((s,a)\) in episode \( \ell \). This general condition covers the heavily-studied i.i.d. setting where \( \ell_1, \ldots, \ell_T \) are i.i.d. samples of a fixed distribution, in which case \( C = 0, \pi^* \) is simply the optimal policy, and \( \Delta \) is the gap function with respect to the optimal \( Q \)-function. More generally, the condition also covers the corrupted i.i.d. setting with \( C \) being the total amount of corruption. We refer the readers to [Jin and Luo, 2020] for detailed explanation. In this stochastic world, our goal is to minimize regret against \( \pi^* \), that is, \( \text{Reg}_T(\pi^*) \).

With unknown transition, this general setup has not been studied before, but for specific examples which then automatically implies logarithmic regret under Eq. (1). This is because Eq. (3) admits a self-bounding property under Eq. (1) — one can bound the right-hand side of Eq. (3) as follows using [Simchowitz and Jamieson, 2019, Yang et al., 2021].

### Occupancy measure and FTRL

To solve this problem with online learning techniques, a commonly used concept is the occupancy measure. Specifically, an occupancy measure \( q \) is such that the following adaptive regret bound holds

\[
\text{Reg}_T(\hat{\pi}) \lesssim \mathbb{E} \left[ \sum_{\ell=1}^{T} \sum_{s \neq s_\ell} \sum_{a \neq \pi^*(s)} \sqrt{q_\ell(s,a) / t} \right],
\]

where \( \hat{q}_\ell \) is some estimator for \( q_\ell \) and \( \phi_\ell \) is any regularizer. The learner’s policy \( \pi_\ell \) is then defined through \( \pi^*_\ell(a|s) \propto \hat{q}_{\ell}(s,a) \). Note that we have \( \hat{q}_\ell = q_\ell^{\pi_\ell} \) but not necessarily \( \hat{q}_\ell = q_\ell \) unless \( \hat{P} = P \).

### 3 Warm-up for Known Transition: A New Loss-shifting Technique

One of the key components of our approach is a new loss-shifting technique for analyzing FTRL applied to MDPs. To illustrate the key idea in a clean manner, in this section we focus on the known transition setting with bandit feedback, the same setting studied by Jin and Luo [2020]. As we will show, our method not only improves their bounds, but also significantly simplifies the analysis, which paves the way for extending the result to the unknown transition setting studied in following sections.

First note that when \( P \) is known, one can simply take \( \hat{P} = P \) (so that \( \hat{q}_\ell = q_\ell \)) and use the standard importance-weighted estimator \( \hat{\ell}_\tau(s,a) = \ell_\tau(s,a) \Pi_\tau(s,a) / q_\tau(s,a) \) in the FTRL framework Eq. (2), where \( \Pi_\tau(s,a) = 1 \) if \( (s,a) \) is visited in episode \( \tau \), and 0 otherwise. It remains to determine the regularizer \( \phi_\ell \). While there are many choices of \( \phi_\ell \) leading to \( \sqrt{T} \)-regret in the adversarial world, obtaining logarithmic regret in the stochastic world requires some special property of the regularizer. Specifically, generalizing the idea of [Zimmert and Seldin, 2019] for multi-armed bandits, [Jin and Luo, 2020] shows that it suffices to find \( \phi_\ell \) such that the following adaptive regret bound holds

\[
\text{Reg}_T(\hat{\pi}) \lesssim \mathbb{E} \left[ \sum_{\ell=1}^{T} \sum_{s \neq s_\ell} \sum_{a \neq \pi^*(s)} \sqrt{q_\ell(s,a) / t} \right],
\]

which then automatically implies logarithmic regret under Eq. (1). This is because Eq. (3) admits a self-bounding property under Eq. (1) — one can bound the right-hand side of Eq. (3) as follows using

\[\text{Reg}_T(\hat{\pi}) \lesssim \mathbb{E} \left[ \sum_{\ell=1}^{T} \sum_{s \neq s_\ell} \sum_{a \neq \pi^*(s)} \sqrt{q_\ell(s,a) / t} \right],\]

Some works (such as [Jin and Luo, 2020]) still consider minimizing \( \text{Reg}_T(\hat{\pi}) \) as the goal in this case. More discussions are deferred to the last paragraph of Section 4.1.
Theorem 3.1. When $P$ is known, Algorithm 3 (with parameter $\gamma = 1$) ensures the optimal regret $\text{Reg}_T(\hat{\pi}) = O(\sqrt{|S||A|T})$ in the adversarial world, and simultaneously $\text{Reg}_T(\hat{\pi}^*) \leq \text{Reg}_T(\hat{\pi}) = O(U + \sqrt{|C|})$ where $U = \frac{|S||A|\log T}{2d_{\text{min}}} + L^4 \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{\log T}{\Delta(s, a)}$ in the stochastic world.

Our bound for the stochastic world is even better than [Jin and Luo, 2020] (their $U$ has an extra $|A|$ factor in the first term and an extra $L$ factor in the second term). By setting the parameter $\gamma$ differently, one can also improve $L^4$ to $L^3$, matching the best existing result from [Simchowitz and Jamieson, 2019] for the i.i.d. setting with $C = 0$ (this would worsen the adversarial bound though). Besides this improvement, we emphasize again that the most important achievement of this approach is that it significantly simplifies the analysis, making the extension to the unknown transition setting possible.
4 Main Algorithms and Results

We are now ready to introduce our main algorithms and results for the unknown transition case, with either full-information or bandit feedback. The complete pseudocode is shown in Algorithm 1, which is built with two main components: a new framework to deal with unknown transitions and adversarial losses (important for incorporating our loss-shifting technique), and special regularizers for FTRL. We explain these two components in detail below.

A new framework for unknown transitions and adversarial losses When the transition is unknown, a common practice (which we also follow) is to maintain an empirical transition along with either full-information or bandit feedback. The complete pseudocode is shown in Algorithm 1, which is built with two main components: a new framework to deal with unknown transitions and adversarial losses (important for incorporating our loss-shifting technique), and special regularizers for FTRL. We explain these two components in detail below.

To fix this, our second modification is to feed FTRL with optimistic losses constructed by adding some (negative) bonus term, an idea often used in the stochastic setting. More specifically, we subtract 
\[ \sum_{t} \hat{\ell}_{t}(s, a) \ln \left( \frac{T |S||A|}{\delta} \right) \]

from the loss for each \((s, a)\) pair, where \(B_{i}(s, a) = \min \left\{ 1, \sum_{s' \in S_{(s, a)}+1} B_{i}(s, a, s') \right\} \); see Eq. (11). In the full-information setting, this means using \(\hat{\ell}_{t}(s, a) = \ell_{t}(s, a) - L \cdot B_{i}(s, a)\). In the bandit setting, note that the importance-weighted estimator discussed in Section 3 is no longer applicable since the transition is unknown (making \(q_{t}\) also unknown), and [Jin et al., 2020] proposes to use \(\ell_{t}(s, a) \cdot u_{t}(s, a) / u_{t}(s, a)\) instead, where \(\ell_{t}(s, a)\) is again the indicator of whether \((s, a)\) is visited during episode \(t\), and \(u_{t}(s, a)\) is the so-called upper occupancy measure defined as

\[ u_{t}(s, a) = \max_{P \in \mathcal{P}_{i}(t)} q_{t}^{\pi_{t}}(s, a) \]

where \(m_{i}(s, a)\) and \(m_{i}(s, a, s')\) are the total number of visits to \((s, a)\) and \((s, a, s')\) respectively prior to epoch \(i\). The confidence set of the true transition for this epoch is then defined as

\[ \mathcal{P}_{i} = \left\{ \hat{P} : \left| \hat{P}(s' | s, a) - \hat{P}(s' | s, a) \right| \leq B_{i}(s, a, s'), \forall (s, a, s') \in S_{k} \times A \times S_{k+1}, k < L \right\} , \]

where \(B_{i}\) is Bernstein-style confidence width (taken from Jin et al. [2020]):

\[ B_{i}(s, a, s') = \min \left\{ 2 \left( \frac{\hat{P}(s' | s, a) \ln \left( \frac{T |S||A|}{\delta} \right)}{m_{i}(s, a)} + \frac{14 \ln \left( \frac{T |S||A|}{\delta} \right)}{3m_{i}(s, a)}, \right), 1 \right\} \]

for some confidence parameter \(\delta \in (0, 1)\). As [Jin et al., 2020, Lemma 2] shows, the true transition \(P\) is contained in the confidence set \(\mathcal{P}_{i}\) for all epoch \(i\) with probably at least \(1 - 4\delta\).

When dealing with adversarial losses, prior works [Rosenberg and Mansour, 2019a,b, Jin et al., 2020, Lee et al., 2020a] perform FTRL (or a similar algorithm called Online Mirror Descent) over the set of all plausible occupancy measures \(\Omega(\mathcal{P}_{i}) = \{ q \in \Omega(\hat{P}) : \hat{P} \in \mathcal{P}_{i} \}\) during epoch \(i\), which can be seen as a form of optimism and encourages exploration. This framework, however, does not allow us to apply the loss-shifting trick discussed in Section 3 — indeed, our key shifting function Eq. (6) is defined in terms of some fixed transition \(\hat{P}\), and the required invariant condition on \((q, g_{t})\) only holds for \(q \in \Omega(\hat{P})\) but not for \(q \notin \Omega(\hat{P})\).

Inspired by this observation, we propose the following new approach. First, to directly fix the issue mentioned above, for each epoch \(i\), we run a new instance of FTRL simply over \(\Omega(\hat{P}_{i})\). This is implemented by keeping track of the epoch starting time \(\ell_{i}\) and only using the cumulative loss \(\sum_{t=1}^{t-1} \hat{\ell}_{t}\) in the FTRL update (Eq. (10)). Therefore, in each epoch, we are pretending to deal with a known transition problem, making the same loss-shifting technique discussed in Section 3 applicable.

However, this removes the critical optimism in the algorithm and does not admit enough exploration. To fix this, our second modification is to feed FTRL with optimistic losses constructed by adding some (negative) bonus term, an idea often used in the stochastic setting. More specifically, we subtract \(L \cdot B_{i}(s, a)\) from the loss for each \((s, a)\) pair, where \(B_{i}(s, a) = \min \left\{ 1, \sum_{s' \in S_{(s, a)}+1} B_{i}(s, a, s') \right\} \); see Eq. (11). In the full-information setting, this means using \(\hat{\ell}_{t}(s, a) = \ell_{t}(s, a) - L \cdot B_{i}(s, a)\). In the bandit setting, note that the importance-weighted estimator discussed in Section 3 is no longer applicable since the transition is unknown (making \(q_{t}\) also unknown), and [Jin et al., 2020] proposes to use \(\ell_{t}(s, a) \cdot u_{t}(s, a) / u_{t}(s, a)\) instead, where \(\ell_{t}(s, a)\) is again the indicator of whether \((s, a)\) is visited during episode \(t\), and \(u_{t}(s, a)\) is the so-called upper occupancy measure defined as

\[ u_{t}(s, a) = \max_{P \in \mathcal{P}_{i}(t)} q_{t}^{\pi_{t}}(s, a) \]

\[ 4\text{When } m_{i}(s, a) = 0, \text{ we simply let } \hat{P}_{i}(|s, a) \text{ be an arbitrary distribution.} \]
With this new framework, it is not difficult to show world, however, we need some carefully designed regularizers discussed next.

**Algorithm 1** Best-of-both-worlds for Episodic MDPs with Unknown Transition

**Input:** confidence parameter $\delta$.

**Initialize:** epoch index $i = 1$ and epoch starting time $t_i = 1$.

**Initialize:** $\forall (s, a, s')$, set counters $m_1(s, a) = m_1(s, a, s') = m_0(s, a) = m_0(s, a, s') = 0$.

**Initialize:** empirical transition $\hat{P}_t$ and confidence width $B_1$ based on Eq. (7) and Eq. (8).

for $t = 1, \ldots, T$ do

Let $q_t$ be Eq. (13) for full-information feedback or Eq. (12) for bandit feedback, and compute

$$\hat{q}_t = \arg\min_{q \in \Omega(P)} \left\langle q, \sum_{\tau = t_i}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q).$$

(10)

Compute policy $\pi_t$ from $\hat{q}_t$ such that $\pi_t(a|s) \propto \hat{q}_t(s, a)$.

Execute policy $\pi_t$ and obtain trajectory $(s^k_{t}, a^k_t)$ for $k = 0, \ldots, L - 1$.

Construct adjusted loss estimator $\hat{\ell}_t(s, a)$ such that

$$\hat{\ell}_t(s, a) = \left\{ \begin{array}{ll} \ell_t(s, a) - L \cdot B_1(s, a), & \text{for full-information feedback,} \\ \ell_t(s, a) - L \cdot B_1(s, a), & \text{for bandit feedback,} \end{array} \right.$$  

(11)

where $B_1(s, a) = \min \{1; \sum_{s' \in S(s,a)_{t+1}} B_1(s, a, s')\}$. $\ell_t(s, a) = \|\{\exists k, (s, a) = (s^k_t, a^k_t)\}$, and $u_t$ is the upper occupancy measure defined in Eq. (9).

Increment counters: if for each $k < L$, $m_i(s^k_t, a^k_t) ≥ \max\{1, 2m_{i-1}(s^k_t, a^k_t)\}$ then $\triangleright$ entering a new epoch

Increment epoch index $i \leftarrow i + 1$ and set new epoch starting time $t_i = t + 1$.

Initialize new counters: $\forall (s, a, s')$, $m_i(s, a, s') = m_{i-1}(s, a, s')$, $m_i(s, a) = m_{i-1}(s, a)$.

Update empirical transition $\hat{P}_t$ and confidence width $B_t$ based on Eq. (7) and Eq. (8).

and can be efficiently computed via the COMP-UOB procedure of [Jin et al., 2020]. Our final adjusted loss estimator is then $\hat{\ell}_t(s, a) = \ell_t(s, a) - L \cdot B_1(s, a)$. In our analysis, we show that these adjusted loss estimators indeed make sure that we only underestimate the loss of each policy, which encourages exploration.

With this new framework, it is not difficult to show $\sqrt{T}$-regret in the adversarial world using many standard choices of the regularizer $\phi_t$ (which recovers the results of [Rosenberg and Mansour, 2019a, Jin et al., 2020] with a different approach). To further ensure polylogarithmic regret in the stochastic world, however, we need some carefully designed regularizers discussed next.

**Special regularizers for FTRL**

Due to the new structure of our algorithm which uses a fixed transition $\hat{P}_t$ during epoch $i$, the design of the regularizers is basically the same as in the known transition case. Specifically, in the bandit case, we use the same Tsallis entropy regularizer:

$$\phi_t(q) = \frac{1}{\eta_t} \sum_{s \neq s_L} \sum_{a \in A} \sqrt{q(s, a)} + \beta \sum_{s \neq s_L} \sum_{a \in A} \log \frac{1}{q(s, a)},$$

(12)

where $\eta_t = \frac{1}{\sqrt{T-t_{(i-1)}}} + 1$ and $\beta = 128L^4$. As discussed in Section 3, the small amount of log-barrier in the second part of Eq. (12) is used to stabilize the algorithm, similarly to [Jin and Luo, 2020].

In the full-information case, while we can still use Eq. (12) since the bandit setting is only more difficult, this leads to extra dependence on some parameters. Instead, we use the following Shannon entropy regularizer:

$$\phi_t(q) = \frac{1}{\eta_t} \sum_{s \neq s_L} \sum_{a \in A} q(s, a) \cdot \log q(s, a).$$

(13)

If $\sum_{b \in A} \hat{q}_t(s, b) = 0$, we let $\pi_t$ to be the uniform distribution.

We use $x \leftarrow y$ as a shorthand for the increment operation $x ← x + y$. 
Although this is a standard choice for the full-information setting, the tuning of the learning rate $\eta_t$ requires some careful thoughts. In the special case of MDPs with one layer (known as the expert problem [Freund and Schapire, 1997]), it has been shown that choosing $\eta_t$ to be of order $1/\sqrt{T}$ ensures best-of-both-worlds [Mourtada and Gaïffas, 2019, Amir et al., 2020]. However, in our general case, due to the use of the loss-shifting trick, we need to use the following data-dependent tuning (with $i$ denoting $i(t)$ for simplicity): $\eta_t = \frac{L \ln(|S||A|)}{64 L^2 \ln(|S||A|) + M_t}$ where

$$M_t = \sum_{\tau=t_i}^{t-1} \min \left\{ \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a) \hat{\ell}_t(s, a)^2, \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a) (\hat{Q}_t(s, a) - \hat{V}_t(s))^2 \right\},$$

and similar to the discussion in Section 3, $\hat{Q}_t$ and $\hat{V}_t$ are the state-action and state-value functions with respect to the transition $P_t$, the adjusted loss function $\hat{\ell}_t$, and the policy $\pi_t$, that is: $\hat{Q}_t(s, a) = \hat{\ell}_t(s, a) + \mathbb{E}_{s' \sim P_t(s, a)}[\hat{V}_t(s')]$ and $\hat{V}_t(s) = \mathbb{E}_{a \sim \pi_t(s)}[\hat{Q}_t(s, a)]$ (with $\hat{V}_t(s_L) = 0$). This particular tuning makes sure that FTRL enjoys some adaptive regret bound with a self-bounding property akin to Eq. (3), which is again the key to ensure polylogarithmic regret in the stochastic world. This concludes all the algorithm design; see Algorithm 1 again for the complete pseudocode.

### 4.1 Main Best-of-both-worlds Results

We now present our main best-of-both-worlds results. As mentioned, proving $\sqrt{T}$-regret in the adversarial world is relatively straightforward. However, proving polylogarithmic regret bounds for the stochastic world is much more challenging due to the transition estimation error, which is usually of order $\sqrt{T}$. Fortunately, we are able to develop a new analysis that upper bounds some transition estimation related terms by the regret itself, establishing a self-bounding property again. We defer the proof sketch to Section 5, and state the main results in the following theorems.\(^2\)

**Theorem 4.1.1.** In the full-information setting, Algorithm 1 with $\delta = \frac{1}{T^2}$ guarantees $\text{Reg}_T(\hat{\pi}) = \tilde{O} \left( \frac{(L^2 |S|^2 + L^4 |S||A| \log(|S||A|)) \log T}{\Delta_{\text{min}}} + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} L^4 |S| \log T \right)$ always, and simultaneously $\text{Reg}_T(\pi^*) = \tilde{O} \left( U + \sqrt{UC} \right)$ under Condition (1), where $U = \tilde{O} \left( \frac{(L^6 |S|^2 + L^8 |S||A| \log(|S||A|)) \log T}{\Delta_{\text{min}}} + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} (L^6 |S| + L^8 |S||A|) \log^2 T \right)$.

**Theorem 4.1.2.** In the bandit feedback setting, Algorithm 1 with $\delta = \frac{1}{T^2}$ guarantees $\text{Reg}_T(\hat{\pi}) = \tilde{O} \left( (L + \sqrt{|A|}) |S| \sqrt{|A|} T \right)$ always, and simultaneously $\text{Reg}_T(\pi^*) = \tilde{O} \left( U + \sqrt{UC} \right)$ under Condition (1), where $U = \tilde{O} \left( \frac{(L^6 |S|^2 + L^8 |S||A|) \log^2 T}{\Delta_{\text{min}}} + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} (L^6 |S| + L^8 |S||A|) \log^2 T \right)$.

While our bounds have some extra dependence on the parameters $L$, $|S|$, and $|A|$ compared to the best existing bounds in each of the two worlds, we emphasize that our algorithm is the first to be able to adapt to these two worlds simultaneously and achieve $\tilde{O}(\sqrt{T})$ and $\tilde{O}(\text{polylog}(T))$ regret respectively. In fact, with some extra twists (such as treating differently the state-action pairs that are visited often enough and those that are not), we can improve the dependence on these parameters, but we omit these details since they make the algorithms much more complicated.

Also, while [Jin and Luo, 2020] is able to obtain $\tilde{O}(\log T)$ regret for the stronger benchmark $\text{Reg}_T(\hat{\pi})$ under Condition (1) and known transition (same as our Theorem 3.1), here we only achieve so for $\text{Reg}_T(\pi^*)$ due to some technical difficulty (see Section 5). However, recall that for the most interesting i.i.d. case, one simply has $\text{Reg}_T(\pi^*) = \text{Reg}_T(\hat{\pi})$ as discussed in Section 2; even for the corrupted i.i.d. case, since $\text{Reg}_T(\hat{\pi})$ is at most $C + \text{Reg}_T(\pi^*)$, our algorithms ensure $\text{Reg}_T(\hat{\pi}) = \tilde{O}(U + C)$ (note $\sqrt{UC} \leq U + C$). Therefore, our bounds on $\text{Reg}_T(\pi^*)$ are meaningful and strong.

### 5 Analysis Sketch

In this section, we provide a proof sketch for the full-information setting (which is simpler but enough to illustrate our key ideas). The complete proofs can be found in Appendix B (full-information) and

\(^2\)For simplicity, for bounds in the stochastic world, we omit some $\tilde{O}(1)$ terms that are independent of the gap function, but they can be found in the full proof.
Appendix C (bandit). We start with the following straightforward regret decomposition:

$$\text{Reg}_T(\pi) = \mathbb{E} \left[ \sum_{t=1}^{T} V^\pi_t(s_0) - \hat{V}^\pi_t(s_0) + \sum_{t=1}^{T} \tilde{V}^\pi_t(s_0) - \hat{V}^\pi_t(s_0) + \sum_{t=1}^{T} \hat{V}^\pi_t(s_0) - V^\pi_t(s_0) \right]$$ (14)

for an arbitrary benchmark $\pi$, where $V^\pi_t$ is the state value function associated with the true transition $P$, the true loss $\ell_t$, and policy $\pi$, while $\hat{V}^\pi_t$ is the state value function associated with the empirical transition $\hat{P}_t$, the adjusted loss $\hat{\ell}_t$, and policy $\pi$. Define the corresponding state-action value functions $Q^\pi_t$ and $\hat{Q}^\pi_t$ similarly (our earlier notations $\tilde{V}_t$ and $\hat{V}_t$ are thus shorthands for $\hat{V}^\pi_t$ and $\hat{Q}^\pi_t$).

In the adversarial world, we bound each of the three terms in Eq. (14) as follows (see Proposition B.1 for details). First, $\mathbb{E} [\text{ERR}_1]$ measures the estimation error of the loss of the learner’s policy $\pi_t$, which can be bounded by $\mathcal{O}(L|S|\sqrt{AT})$ following the analysis of Jin et al. [2020]. Second, as mentioned, our adjusted losses are optimistic in the sense that it underestimates the loss of all policies (with high probability), making $\mathbb{E} [\text{ERR}_2]$ an $\mathcal{O}(1)$ term only. Finally, $\mathbb{E} [\text{ESTREG}]$ is the regret measured with $\hat{P}_t$ and $\hat{\ell}_t$, which is controlled by the FTRL procedure and of order $\mathcal{O}(L\sqrt{|S||A|T})$. Put together, this proves the $\mathcal{O}(L|S|\sqrt{|A|T})$ regret shown in Theorem 4.1.1.

In the stochastic world, we fix the benchmark $\pi = \pi^\star$. To obtain polylogarithmic regret, an important observation is that we now have to make use of the potentially negative term $\text{ERR}_2$ instead of simply bounding it by $\mathcal{O}(1)$ (in expectation). Specifically, inspired by Simchowitz and Jamieson [2019], we propose a new decomposition on $\text{ERR}_1$ and $\text{ERR}_2$ jointly as follows (see Appendix D.1): $\text{ERR}_1 + \text{ERR}_2 = \text{ERR}^\text{SUB} + \text{ERR}^\text{OPT} + \text{OCCDIFF} + \text{BIAS}$. Here,

- $\text{ERR}^\text{SUB} = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} q_t(s, a) \hat{E}^\pi_t(s, a)$ measures some estimation error contributed by the suboptimal actions, where $\hat{E}^\pi_t(s, a) = \ell_t(s, a) + \mathbb{E}_{s' \sim P(\cdot|s, a)} [\hat{V}^\pi_t(s')] - \hat{Q}^\pi_t(s, a)$ is a “surplus” function (a term taken from Simchowitz and Jamieson, 2019);
- $\text{ERR}^\text{OPT} = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^\star(s)} (q_t(s, a) - q^\star_t(s, a)) \hat{E}^\pi_t(s, a)$ measures some estimation error contributed by the optimal action, where $q^\star_t(s, a)$ is the probability of visiting a trajectory of the form $(s_0, \pi^\star(s_0)), (s_1, \pi^\star(s_1)), \ldots, (s_{k(s)} - 1, \pi^\star(s_{k(s)} - 1)), (s, a)$ when executing policy $\pi_t$;
- $\text{OCCDIFF} = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} (q_t(s, a) - \hat{q}_t(s, a)) (\hat{Q}^\pi_t(s, a) - \hat{V}^\pi_t(s))$ measures the occupancy measure difference between $q_t$ and $\hat{q}_t$;
- $\text{BIAS} = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} q_t(s, a) (\hat{V}^\pi_t(s) - V^\pi_t(s))$ measures some estimation error for $\pi^\star$, which, similar to $\text{ERR}_2$, is of order $\mathcal{O}(1)$ in expectation due to optimism.

The next step is to show that the terms $\text{ERR}^\text{SUB}, \text{ERR}^\text{OPT}, \text{OCCDIFF},$ and $\text{ESTREG}$ can all be upper bounded by some quantities that admit a certain self-bounding property similarly to the right-hand side of Eq. (3). We identify four such quantities and present them using functions $G_1, G_2, G_3,$ and $G_4$, whose definitions are deferred to Appendix D.2 due to space limit. Combining these bounds for each term, we obtain the following important lemma.

**Lemma 5.1.** With $\delta = \frac{1}{T^2}$, Algorithm 1 ensures that $\text{Reg}_T(\pi^\star)$ is at most $\mathcal{O}(L^4|S|^3|A|^2 \ln^2 T)$ plus:

$$\mathbb{E} \left[ \mathcal{O} \left( G_1(L^4|S|\ln T) + G_2(L^4|S|\ln T) + G_3(L^4 \ln T) + G_4(L^5|S||A| \ln(T \ln(|S||A|))) \right) \right].$$

Finally, as mentioned, each of the $G_1, G_2, G_3,$ and $G_4$ functions can be shown to admit the following self-bounding property, such that similarly to what we argue in Eq. (4), picking the optimal values of $\alpha$ and $\beta$ and rearranging leads to the polylogarithmic regret bound shown in Theorem 4.1.1.

**Lemma 5.2 (Self-bounding property).** Under Condition (1), we have for any $\alpha, \beta \in (0, 1),$ \n
$$\mathbb{E} [G_1(J)] \leq \alpha \cdot (\text{Reg}_T(\pi^\star) + C) + \mathcal{O} \left( \frac{1}{\alpha} \cdot \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \frac{J}{\Delta(s, a)} \right),$$
We emphasize again that the proposed joint decomposition on $E$ which simplifies the method of Jin and Luo [2020] and improves their results. Then, we design a new Theorem 4.1.2), one other key future direction is to remove the assumption that there exists a unique

$$\text{E}[G_2(J)] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + O\left(\frac{1}{\beta} \cdot \frac{L^3|J|}{\Delta_{\min}}\right),$$

$$\text{E}[G_3(J)] \leq (\alpha + \beta) \cdot (\text{Reg}_T(\pi^*) + C) + O\left(\frac{1}{\alpha} \cdot \sum_{s \neq s_\perp} \sum_{a \neq \pi^*(s)} \frac{L^2|J|}{\Delta_\perp(\pi, \pi^*)}\right) + O\left(\frac{1}{\beta} \cdot \frac{L^2|S||J|}{\Delta_{\min}}\right),$$

$$\text{E}[G_4(J)] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + O\left(\frac{1}{\beta} \cdot \frac{L^2|S||J|}{\Delta_{\min}}\right).$$

We emphasize again that the proposed joint decomposition on $E_1 + E_2$ plays a crucial role in this analysis and addresses the key challenge on how to bound the transition estimation error by something better than $\sqrt{T}$. We also point out that in this analysis, only $\text{EstReg}$ is related to the FTRL procedure, while the other three terms are purely based on our new framework to handle unknown transition. In fact, the reason that we can only derive a polylog($T$) bound on $\text{Reg}_T(\pi^*)$ but not directly on $\text{Reg}_T(\hat{\pi})$ is also due to these three terms — they can be related to the right-hand side of Condition (1) only when we use the benchmark $\pi = \pi^*$ but not when $\pi = \hat{\pi}$. This is not the case for $\text{EstReg}$, which is the reason why Jin and Luo [2020] are able to derive a bound on $\text{Reg}_T(\hat{\pi})$ directly when the transition is known. Whether this issue can be addressed is left as a future direction.

6 Conclusions

In this work, we propose an algorithm for learning episodic MDPs which achieves favorable regret guarantees simultaneously in the stochastic and adversarial worlds with unknown transition. We start from the known transition setting and propose a loss-shifting trick for FTRL applied to MDPs, which simplifies the method of Jin and Luo [2020] and improves their results. Then, we design a new framework to extend our known transition algorithm to the unknown transition case, which is critical for the application of the loss-shifting trick. Finally, we develop a novel analysis which carefully upper bounds the transition estimation error by (a fraction of) the regret itself plus a gap-dependent poly-logarithmic term in the stochastic setting, resulting in our final best-of-both-worlds result.

Besides the open questions discussed earlier (such as improving our bounds in Theorem 4.1.1 and Theorem 4.1.2), one other key future direction is to remove the assumption that there exists a unique optimal action for each state, which appears to be challenging despite the recent progress for the bandit case [Ito, 2021], since the occupancy measure computed from Eq. (10) has a very complicated structure. Another interesting direction would be to extend the sub-optimality gap function to other fine-grained gap functions, such as that of Dann et al. [2021].

Acknowledgments and Disclosure of Funding

HL is supported by NSF Award IIS-1943607 and a Google Faculty Research Award. LH is supported in part by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403. We thank Max Simchowitz for many helpful discussions, and the anonymous reviewers for their valuable feedback and suggestions.

References


## Contents

1 Introduction ................................................................. 1

2 Preliminaries ............................................................ 3

3 Warm-up for Known Transition: A New Loss-shifting Technique .... 4

4 Main Algorithms and Results ............................................ 6
   4.1 Main Best-of-both-worlds Results .......................... 8

5 Analysis Sketch .......................................................... 8

6 Conclusions ................................................................. 10

A Best of Both Worlds for MDPs with Known Transition .......... 14
   A.1 Loss-shifting Technique ..................................... 14
   A.2 Known Transition and Full-information Feedback: FTRL with Shannon Entropy . 17
   A.3 Known Transition and Bandit Feedback: FTRL with Tsallis Entropy ........ 21

B Best of Both Worlds for MDPs with Unknown Transition and Full Information 25
   B.1 Optimism of Adjusted losses and Other Lemmas .......... 25
   B.2 Proof for the Adversarial World (Proposition B.1) .... 27
   B.3 Proof for the Stochastic World (Proposition B.2) ....... 30

C Best of Both Worlds for MDPs with Unknown Transition and Bandit Feedback 34
   C.1 Auxiliary Lemmas ............................................. 35
   C.2 Proof for the Adversarial World (Proposition C.1) .... 44
   C.3 Proof for the Stochastic World (Proposition C.2) ....... 45

D General Decomposition, Self-bounding Terms, and Supplementary Lemmas 49
   D.1 General Decomposition Lemma ............................ 49
   D.2 Self-bounding Terms ........................................ 53
   D.3 Supplementary Lemmas .................................... 60
An important convention  Note that the value of $m_i(s, a)$ is changing in the algorithm. For the entire analysis, we see $m_i(s, a)$ as its initial value, which is the number of visits to $(s, a)$ from epoch 1 to epoch $i - 1$. In this sense, if we let $N$ be the total number of epochs, then $m_{N+1}(s, a)$ is naturally defined as the total number of visits to $(s, a)$ within $T$ episodes.

A Best of Both Worlds for MDPs with Known Transition

In this section, we show how to extend the loss-shifting technique to MDPs with known transition and obtain best-of-both-worlds results.

A.1 Loss-shifting Technique

First of all, we introduce a general invariant condition with a fixed transition in Lemma A.1.1

**Lemma A.1.1.** Fix the transition function $P$. For any policy $\pi$ and loss function $\hat{\ell} : S \times A \to \mathbb{R}$, define invariant function $g \in S \times A \to \mathbb{R}$ as:

$$g^{P, \pi, \hat{\ell}}(s, a) \triangleq \left( Q^{P, \pi, \hat{\ell}}(s, a) - V^{P, \pi, \hat{\ell}}(s) - \hat{\ell}(s, a) \right),$$

where $Q^{P, \pi, \hat{\ell}}$ and $V^{P, \pi, \hat{\ell}}$ are state-action value and state value functions associated with $\hat{\ell}$ and the fixed policy $\pi$. Then, it holds for any policy $\pi'$ that

$$\left\langle q^{P, \pi'}, g^{P, \pi, \hat{\ell}} \right\rangle \triangleq \sum_{s \neq s_L} \sum_{a \in A} q^{P, \pi'}(s, a) \cdot g^{P, \pi, \hat{\ell}}(s, a) = -V^{P, \pi, \hat{\ell}}(s_0)$$

where $V^{P, \pi, \hat{\ell}}(s_0)$ only depends on $\pi$ and $\hat{\ell}$ (but not $\pi'$).

**Proof.** For notational convenience, we drop the superscripts for fixed transition $P$ and loss function $\hat{\ell}$. By the standard performance difference lemma [Kakade, 2003, Theorem 5.2.1], it holds for any policy $\pi'$ that

$$V^{\pi'}(s_0) - V^{\pi}(s_0) = \sum_{s \neq s_L} \sum_{a \in A} q^{\pi'}(s, a) (Q^{\pi}(s, a) - V^{\pi}(s)).$$

On the other hand, it also holds that

$$V^{\pi'}(s_0) = \sum_{s \neq s_L} \sum_{a \in A} q^{\pi'}(s, a) \hat{\ell}(s, a).$$

Therefore, subtracting $V^{\pi'}(s_0)$ from Eq. (16) yields that

$$-V^{\pi}(s_0) = \sum_{s \neq s_L} \sum_{a \in A} q^{\pi'}(s, a) \left( Q^{\pi}(s, a) - V^{\pi}(s) - \hat{\ell}(s, a) \right)$$

which completes the proof after putting back the superscripts for $P$ and $\hat{\ell}$.

As discussed in Section 3, the invariant function $g^{P, \pi, \hat{\ell}}$ defined in Eq. (15) allows us to treat FTRL as dealing with a hypothesized loss sequence, as restated below.

**Corollary A.1.2.** Consider the selected occupancy measure $\hat{q}_t$ via FTRL with respect to a regularizer $\phi_t(\cdot)$ and loss sequence $\{\hat{\ell}_\tau\}_{\tau < t}$ (on the decision set $\Omega(P)$), then it holds that

$$\hat{q}_t = \arg\min_{q \in \Omega(P)} \left\langle q, \sum_{\tau < t} \hat{\ell}_\tau \right\rangle + \phi_t(q) = \arg\min_{q \in \Omega(P)} \left\langle q, \sum_{\tau < t} (\hat{\ell}_\tau + g_\tau) \right\rangle + \phi_t(q).$$

for any invariant function sequence $\{g_\tau\}_{\tau < t}$ which are constructed with hypothesized losses $\{\hat{\ell}_\tau\}_{\tau < t}$ and policies $\{\pi'_\tau\}_{\tau < t}$.
Proof. By Lemma A.1.1, one can verify that
\[
\left\langle q, \sum_{\tau \leq t} g_{\tau} \right\rangle = -\sum_{\tau \leq t} V^P,\pi_{t},\hat{\ell}_{t}(s_0)
\]
for any occupancy measure \( q \in \Omega(\hat{P}) \). Therefore, this term does not affect the optimization. \( \square \)

Then, we consider the “loss-shifting function” defined in Eq. (6), that is, constructing \( g_{t} \) via the loss estimator \( \hat{\ell}_{t} \) and the policy \( \pi_{t} \) selected at episode \( t \). Importantly, in the known transition setting where \( \hat{q}_{t} = q_{t} \), \( \hat{\ell}_{t} \) is inverse propensity weighted estimator, in other words, \( \hat{\ell}_{t}(s, a) = \frac{1_{t}(s, a)}{q_{t}(s, a)} \). More specifically, we have \( g_{t}(s, a) = \hat{Q}_{t}(s, a) - \hat{V}_{t}(s) - \hat{\ell}_{t}(s, a) \), where
\[
\hat{Q}_{t}(s, a) = \hat{\ell}_{t}(s, a) + \sum_{s' \in S_{k(t)+1}} \hat{P}(s'|s, a) \hat{V}_{t}(s), \quad \hat{V}_{t}(s) = \sum_{a \in A} \pi_{t}(a|s) \hat{Q}_{t}(s, a)
\]
(with \( \hat{V}_{t}(s_L) = 0 \)). Below we show several useful properties, which are key to achieve the best-of-both-worlds guarantee in the known transition setting.

**Lemma A.1.3.** With \( \hat{P} = P \) being the true transition function (therefore, \( \hat{q}_{t} = q_{t} \)), we have

- \( q_{t}(s, a)\hat{Q}_{t}(s, a) \leq L \),
- \( q_{t}(s)\hat{V}_{t}(s) \leq L \),
- \( \mathbb{E}_{t} \left[ \left( \hat{Q}_{t}(s, a) - \hat{V}_{t}(s) \right)^2 \right] \leq \frac{2L^2(1-\pi_{t}(a|s))}{q_{t}(s, a)} \),

for all state-action pairs \( (s, a) \) (where \( \mathbb{E}_{t} \) denotes the conditional expectation given everything before episode \( t \)).

Proof. Denote by \( q_{t}(s', a'|s, a) \) the probability of visiting \( (s', a') \) after taking action \( a \) at state \( s \) and following \( \pi_{t} \) afterwards. Then we have \( \hat{Q}_{t}(s, a) = \sum_{k=k(s)}^{L-1} \sum_{s' \in S_{k}} \sum_{a' \in A} q_{t}(s', a'|s, a) \hat{\ell}_{t}(s', a') \). Therefore, plugging in the definition of \( \hat{\ell}_{t}(s, a) \), we verify the following:
\[
q_{t}(s, a)\hat{Q}_{t}(s, a) = \sum_{k=k(s)}^{L-1} \sum_{s' \in S_{k}} \sum_{a' \in A} q_{t}(s', a'|s, a) q_{t}(s, a) \hat{\ell}_{t}(s', a') \leq \sum_{k=k(s)}^{L-1} \sum_{s' \in S_{k}} \sum_{a' \in A} \tau_{t}(s', a') \leq L,
\]
where the inequality is by \( q_{t}(s, a) q_{t}(s', a'|s, a) \leq q_{t}(s', a') \) and \( \tau_{t}(s', a') \in [0, 1] \). This also proves \( q_{t}(s)\hat{V}_{t}(s) \leq L \) using the definition of \( \hat{V}_{t}(s) \).

To prove the last statement, we first note that
\[
\mathbb{E}_{t} \left[ \left( \hat{Q}_{t}(s, a) - \hat{V}_{t}(s) \right)^2 \right] \leq 2\mathbb{E}_{t} \left[ \left(1 - \pi_{t}(a|s)\right)^2 \hat{Q}_{t}(s, a)^2 + \sum_{b \neq a} \pi_{t}(b|s) \hat{Q}_{t}(s, b) \right] \tag{18}
\]
by the fact \( (x - y)^2 \leq 2x^2 + 2y^2 \) for all \( x, y \in \mathbb{R} \).

For the first term in Eq. (18), we have:
\[
\mathbb{E}_{t} \left[ \hat{Q}_{t}(s, a)^2 \right] = \mathbb{E}_{t} \left[ \left( \sum_{k=k(s)}^{L-1} \sum_{s' \in S_{k}} \sum_{a' \in A} q_{t}(s', a'|s, a) q_{t}(s, a) \tau_{t}(s', a') \right)^2 \right]
\]

15
where the second line uses the Cauchy-Schwartz inequality; the third line follows from the fact 
\( E_t [s,a] \leq L \leq L = t \), the fourth line uses 
\( q_t(s',a') = 0 \) for all \( (s,a),(s',a') \in S_k \times A \) such that \( (s,a) \neq (s',a') \); and the last line follows from the fact \( q_t(s,a)q_t(s',a'|s,a) \leq q_t(s',a') \).

Repeating the similar arguments, we bound the second term as

\[
E_t \left[ \left( \sum_{b \neq a} \pi_t(b|s) \tilde{Q}_t(s,b) \right)^2 \right] = E_t \left[ \left( \sum_{k=k(s)}^{L-1} \left( \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s)q_t(s',a'|s,b) \right) \tilde{E}_t(s',a') \right)^2 \right] 
\]

\[
\leq L \cdot E_t \left[ \sum_{k=k(s)}^{L-1} \left( \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s)q_t(s',a'|s,b) \right) \tilde{E}_t(s',a') \right)^2 \right] 
\]

\[
= L \cdot \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s)q_t(s',a'|s,b) \right) \tilde{E}_t(s',a') \] \quad \text{(Cauchy-Schwartz inequality)}
\]

\[
\leq L \cdot \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s)q_t(s',a'|s,b) \right) \tilde{E}_t(s',a') \]

\[
= L \cdot \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} q_t(s,b)q_t(s',a'|s,b) \right) \tilde{E}_t(s',a') \]

\[
\leq L^2 \cdot \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s)q_t(s',a'|s,b) \right) \]

Plugging these bounds into Eq. (18) concludes the proof:

\[
E_t \left[ (\tilde{Q}_t(s,a) - \tilde{V}_t(s))^2 \right] \leq 2L^2 \left( \frac{(1 - \pi_t(a|s))^2}{q_t(s,a)} + \frac{1 - \pi_t(a|s)}{q_t(s)} \right)
\]
Appendix B.

with known transition via the FTRL framework with the Shannon entropy regularizer. Some of
Although not mentioned in the main text, in this section, we discuss a simple application of the
Algorithm 2
Best-of-both-worlds for MDPs with Known Transition and Full-information Feedback

\[ \eta \]

Theorem A.2.1. Algorithm 2 ensures that

\[ \text{Reg}_T(\tilde{\pi}) \geq \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta(s, a) - C \right]. \]  

(21)

Note that this is only weaker than Condition (1) since \( \text{Reg}_T(\tilde{\pi}) \geq \text{Reg}_T(\pi^*) \).

Then, we show that Algorithm 2 ensures a worst-case guarantee \( \text{Reg}_T(\tilde{\pi}) = \tilde{O}(L \sqrt{T}) \), and simultaneouly an adaptive regret bound which further leads to logarithmic regret under Condition (21) (Corollary A.2.2). Importantly, the worst-case regret bound matches the lower bound of learning MDPs with known transition and full-information feedback [Zimin and Neu, 2013].

Theorem A.2.1. Algorithm 2 ensures that \( \text{Reg}_T(\tilde{\pi}) \) is bounded by

\[ \mathcal{O} \left( \min \left\{ L^2 T, L^3 \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi_t(s)} q_t(s, a) \right] \right\} \ln(|S||A|) + L^2 \ln(|S||A|) \right) \]  

(22)
for any mapping \( \pi : S \to A \).

**Proof.** Due to the invariant property (that \( \langle q, g \rangle \) is independent of \( q \in \Omega(P) \)), we can apply Lemma A.2.3 with \( \ell_t \) being either \( \ell_t \) or \( \ell_t + gt \) for any \( t \) — note that the condition \( \eta_t \ell_t(s, a) \geq -1 \) is always satisfied since \( \ell_t(s, a) \in [0, 1] \) and \( \hat{Q}_t(s, a) - \hat{V}_t(s) \in [-L, L] \). Therefore, we have for any \( u \in \Omega(P) \),

\[
T \sum_{t=1} T (q_t - u, \ell_t) \leq \frac{L \ln(|S||A|)}{\eta_{T+1}} + \sum_{t=1} T \eta_t \min \left\{ \sum_{s \neq s_L, a \in A} q_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2, \sum_{s \neq s_L, a \in A} q_t(s, a) \ell_t(s, a)^2 \right\}
\]

\[
= \frac{L \ln(|S||A|)}{\eta_{T+1}} + \sum_{t=1} T \eta_t (M_t - M_{t-1}) \quad \text{(definition of } M_t) \]

\[
= \frac{L \ln(|S||A|)}{\eta_{T+1}} + \sum_{t=1} T \eta_t \sqrt{M_t + M_{t-1}} \left( \sqrt{M_t} - \sqrt{M_{t-1}} \right),
\]

\[
\leq \frac{L \ln(|S||A|)}{\eta_{T+1}} + 2 \sum_{t=1} T \eta_t \sqrt{M_{t-1} + L} \left( \sqrt{M_t} - \sqrt{M_{t-1}} \right). (M_t \leq M_{t-1} + L)
\]

Further plugging in the definition of \( \eta_t \) and taking expectation, we arrive at

\[
\text{Reg}_T(\hat{\pi}) \leq \mathbb{E} \left[ \frac{L \ln(|S||A|)}{\eta_{T+1}} + 2 \sqrt{L \ln(|S||A|)} \sum_{t=1} T \left( \sqrt{M_t} - \sqrt{M_{t-1}} \right) \right]
\]

\[
= \mathbb{E} \left[ \sqrt{L \ln(|S||A|)} (M_T + 64L^3 \ln |S||A|) + 2 \sqrt{LM_T} \ln(|S||A|) \right]
\]

\[
= O \left( \sqrt{L \mathbb{E}[M_T \ln(|S||A|)]} + L^2 \ln(|S||A|) \right).
\]

It remains to bound \( M_T \). First, we note that

\[
M_T = \sum_{t=1} T \min \left\{ \sum_{s \neq s_L, a \in A} q_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2, \sum_{s \neq s_L, a \in A} q_t(s, a) \ell_t(s, a)^2 \right\}
\]

\[
\leq \min \left\{ \sum_{t=1} T \sum_{s \neq s_L, a \in A} q_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2, \sum_{t=1} T \sum_{s \neq s_L, a \in A} q_t(s, a) \ell_t(s, a)^2 \right\}
\]

\[
\leq \min \left\{ \sum_{t=1} T \sum_{s \neq s_L, a \in A} q_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2, LT \right\}.
\]

where the second line follows from the fact \( \min \{a, b\} \leq \min \{c, d\} \leq \min \{a + c, b + d\} \), and the third line uses the property \( 0 \leq \ell_t(s, a) \leq 1 \) for all state-action pairs \( (s, a) \).

On the other hand, we have

\[
\left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2 \leq 2 \left[ (1 - \pi_t(a|s))^2 \hat{Q}_t(s, a)^2 + \left( \sum_{b \neq a} \pi_t(b|s) \hat{Q}_t(s, b) \right)^2 \right]
\]

\[
\leq 2L^2 \cdot \left[ (1 - \pi_t(a|s))^2 + (1 - \pi_t(a|s))^2 \right]
\]

\[
\leq 4L^2 (1 - \pi_t(a|s)), \quad (23)
\]
where we use the facts \((a - b)^2 \leq 2(a^2 + b^2)\) and \(0 \leq \hat{Q}_t(s, a) \leq L\) for all state-action pairs \((s, a)\).

Therefore, we have for any mapping \(\pi : S \to A\),

\[
\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2
\]

\[
\leq 4L^2 \cdot \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) (1 - \pi_t(a|s))
\]

\[
\leq 4L^2 \cdot \sum_{t=1}^{T} \sum_{s \neq s_L} \left( q_t(s) \cdot (1 - \pi_t(\pi(s)|s)) + \sum_{a \neq \pi(s)} q_t(s, a) \right)
\]

\[
= 8L^2 \cdot \sum_{t=1}^{T} \sum_{s \neq s_L} q_t(s, a),
\]

which finishes the proof. \(\Box\)

**Corollary A.2.2.** Suppose Condition (21) holds. Algorithm 2 guarantees that:

\[
\text{Reg}_T(\hat{\pi}) = O\left(U + \sqrt{CU}\right), \quad \text{where } U = \frac{L^3 \ln(|S||A|)}{\Delta_{\text{MIN}}}.
\]

**Proof.** By Theorem A.2.1, \(\text{Reg}_T(\hat{\pi})\) is bounded by

\[
\kappa \cdot \left( L^3 \ln(|S||A|) \cdot \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \right] + L^2 \ln(|S||A|) \right)
\]

where \(\kappa \geq 1\) is a universal constant, and \(\pi^*\) is the mapping specified in Condition (21).

For any \(z > 1\), \(\text{Reg}_T(\hat{\pi})\) is bounded by

\[
\kappa \cdot \left( \frac{zL^3 \ln(|S||A|)}{2\Delta_{\text{MIN}}} \cdot \left( \frac{2}{z} \cdot \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta_{\text{MIN}} \right] \right) + \kappa L^2 \ln(|S||A|) \right)
\]

\[
\leq \frac{\text{Reg}_T(\hat{\pi}) + C}{z} + \frac{zL^3 \ln(|S||A|)}{4\Delta_{\text{MIN}}} + \kappa L^2 \ln(|S||A|)
\]

\[
\leq \frac{\text{Reg}_T(\hat{\pi}) + C}{z} + z \cdot 2\kappa^2 U,
\]

where the third line uses the AM-GM inequality and Eq. (21), and the last line uses the shorthand \(U\) and the facts \(\kappa, z > 1\) and \(\Delta_{\text{MIN}} \leq 1\).

Therefore, by defining \(x = z - 1 > 0\), we can rearrange and arrive at

\[
\text{Reg}_T(\hat{\pi}) \leq \frac{C}{x - 1} + \frac{x^2}{z - 1} \cdot 2\kappa^2 U
\]

\[
= \frac{C}{x} + \frac{(x + 1)^2}{x} \cdot (2\kappa^2 U)
\]

\[
= \frac{1}{x} \cdot (C + 2\kappa^2 U) + x \cdot (2\kappa^2 U) + 4\kappa^2 U,
\]

where we replace all \(z\)’s in the second line. Picking the optimal \(x = \sqrt{\frac{C+2\kappa^2 U}{2\kappa^2 U}}\) gives

\[
\text{Reg}_T(\hat{\pi}) \leq 2\sqrt{(C + 2\kappa^2 U) \cdot (2\kappa^2 U) + 4\kappa^2 U}
\]

19
where the second line follows from the fact $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$.

**Lemma A.2.3.** Suppose $q_t = \arg \min_{q \in \Omega(P)} \left\langle q, \sum_{\tau \leq t} \hat{\ell}_\tau \right\rangle + \phi_t(q)$, where $\phi_t(q) = \frac{1}{\eta_t} \phi(q)$ for some $\eta_t > 0$, $\phi(q) = \sum_{s \neq s_L} \sum_{a \in A} q(s, a) \ln q(s, a)$, and $\eta_t \hat{\ell}_t(s, a) \geq -1$ holds for all $t$ and $(s, a)$. Then

$$\sum_{t=1}^T \left\langle q_t - u, \hat{\ell}_t \right\rangle \leq \frac{L \ln(|S||A|)}{\eta_{T+1}} + \sum_{t=1}^T \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) \hat{\ell}_t(s, a)^2,$$

holds for any $u \in \Omega(P)$.

**Proof.** Let $\Phi_t = \min_{q \in \Omega(P)} \left\langle q, \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q)$ and $D_F(u, v)$ being the Bregman divergence with convex function $F$, that is, $D_F(u, v) = F(u) - F(v) - \langle u - v, \nabla F(v) \rangle$.

Then, we have

$$\Phi_t = \left\langle q_t, \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q_t)$$

$$= \left\langle q_{t+1}, \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q_{t+1}) - \left( \left\langle q_{t+1} - q_t, \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q_{t+1}) - \phi_t(q_t) \right)$$

$$\leq \left\langle q_{t+1}, \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q_{t+1}) - \left( -\langle q_{t+1} - q_t, \nabla \phi_t(q_t) \rangle + \phi_t(q_{t+1}) - \phi_t(q_t) \right)$$

$$= \left\langle q_{t+1}, \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \right\rangle + \phi_t(q_{t+1}) - \phi_t(q_{t+1}) - \phi_t(q_t) + D_{\phi_t}(q_{t+1}, q_t)$$

$$= \Phi_{t+1} - \left\langle q_{t+1}, \hat{\ell}_t \right\rangle - (\phi_{t+1}(q_{t+1}) - \phi_t(q_{t+1})) - D_{\phi_t}(q_{t+1}, q_t),$$

where the third line follows from the first order optimality condition of $q_t$, that is, $\langle q_{t+1} - q_t, \nabla \phi_t(q_t) + \sum_{\tau = 1}^{t-1} \hat{\ell}_\tau \rangle \geq 0$.

Taking the summation over all episodes gives

$$\Phi_1 = \Phi_{T+1} - \sum_{t=1}^T \left\langle q_{t+1}, \hat{\ell}_t \right\rangle - \sum_{t=1}^T (\phi_{t+1}(q_{t+1}) - \phi_t(q_{t+1})) - \sum_{t=1}^T D_{\phi_t}(q_{t+1}, q_t).$$

Therefore, we have

$$\sum_{t=1}^T \left\langle q_t - u, \hat{\ell}_t \right\rangle$$

$$= \sum_{t=1}^T \left\langle q_t - u, \hat{\ell}_t \right\rangle + \Phi_{T+1} - \Phi_1 - \sum_{t=1}^T \left\langle q_{t+1}, \hat{\ell}_t \right\rangle - \sum_{t=1}^T (\phi_{t+1}(q_{t+1}) - \phi_t(q_{t+1})) - \sum_{t=1}^T D_{\phi_t}(q_{t+1}, q_t)$$

$$= \sum_{t=1}^T \left( \left\langle q_t - q_{t+1}, \hat{\ell}_t \right\rangle - D_{\phi_t}(q_{t+1}, q_t) \right) - \sum_{t=1}^T \left\langle u, \hat{\ell}_t \right\rangle + \Phi_{T+1} - \Phi_1 - \sum_{t=1}^T (\phi_{t+1}(q_{t+1}) - \phi_t(q_{t+1}))$$

$$\leq \sum_{t=1}^T \left( \left\langle q_t - q_{t+1}, \hat{\ell}_t \right\rangle - D_{\phi_t}(q_{t+1}, q_t) \right) + \Phi_{T+1}(u) - \phi_1(q_1) - \sum_{t=1}^T (\phi_{t+1}(q_{t+1}) - \phi_t(q_{t+1}))$$

where the last line follows from the optimality condition $\Phi_{T+1} \leq \sum_{t=1}^T \left\langle u, \hat{\ell}_t \right\rangle + \phi_{T+1}(u)$. 

20
To bound the stability term, we first consider relaxing the constraint and taking the maximum as:
\[
\langle q_t - q_{t+1}, \hat{\ell}_t \rangle - D_{\phi_t}(q_{t+1}, q_t) \leq \max_{q \in \mathbb{R}^S \times A} \langle q_t - q, \hat{\ell}_t \rangle - D_{\phi_t}(q, q_t).
\]
Denote by \( \tilde{q}_t \) the maximizer of the right hand side. Setting the gradient to zero yields the equality \( \nabla \phi_t(q_t) - \nabla \phi_t(\tilde{q}_t) = \hat{\ell}_t \). By direction calculation, one can verify that \( \tilde{q}_t(s, a) = q_t(s, a) \cdot \exp \left( -\eta_t \cdot \hat{\ell}_t(s, a) \right) \) for all state-action pairs, and the following inequality that
\[
\langle q_t - q_{t+1}, \hat{\ell}_t \rangle - D_{\phi_t}(q_{t+1}, q_t) \leq \langle q_t - \tilde{q}_t, \hat{\ell}_t \rangle - D_{\phi_t}(\tilde{q}_t, q_t)
\]
\[
= \langle q_t - \tilde{q}_t, \hat{\ell}_t \rangle - \phi_t(q_t) + \phi_t(\tilde{q}_t) - (\tilde{q}_t - q_t, \nabla \phi_t(q_t))
\]
where the second equality uses the equality \( \nabla \phi_t(q_t) - \nabla \phi_t(\tilde{q}_t) = \hat{\ell}_t \).
Moreover, the term \( D_{\phi_t}(q_t, \tilde{q}_t) \) can be bounded as:
\[
D_{\phi_t}(q_t, \tilde{q}_t) = \frac{1}{\eta_t} \sum_{s \neq s_*} \sum_{a \in A} \left( q_t(s, a) \ln \left( \frac{q_t(s, a)}{\tilde{q}_t(s, a)} \right) - q_t(s, a) + \tilde{q}_t(s, a) \right)
\]
\[
= \frac{1}{\eta_t} \sum_{s \neq s_*} \sum_{a \in A} q_t(s, a) \cdot \left( \eta_t \hat{\ell}_t(s, a) - 1 + \exp \left( -\eta_t \cdot \hat{\ell}_t(s, a) \right) \right)
\]
\[
\leq \eta_t \sum_{s \neq s_*} \sum_{a \in A} q_t(s, a) \hat{\ell}_t(s, a)^2
\]
where the last inequality follows from the facts \( y - 1 + e^{-y} \leq y^2 \) for \( y > -1 \) and \( \eta_t \cdot \hat{\ell}_t(s, a) \geq -1 \) for all state-action pairs.

On the other hand, the penalty term is at most
\[
\phi_{T+1}(u) - \phi_1(q_1) - \sum_{t=1}^{T} (\phi_{t+1}(q_{t+1}) - \phi_t(q_t+1)) \leq -\frac{\phi(q_1)}{\eta_t} - \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \phi(q_t),
\]
since \( \phi(u) \leq 0 \). Moreover, note that for any valid occupancy measure \( q \), it holds that
\[
\phi(q) = \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{a \in A} q(s, a) \geq -\sum_{k=0}^{L-1} \ln(|S_k||A|) \geq -L \ln(|S||A|).
\]
Therefore, the penalty term is bounded by
\[
-\frac{\phi(q_1)}{\eta_t} - \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \phi(q_t)
\]
\[
\leq L \ln(|S||A|) \cdot \left( \frac{1}{\eta_t} + \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \right) = \frac{L \ln(|S||A|)}{\eta_{T+1}}.
\]
Finally, combining the bounds for the stability and penalty terms finishes the proof. \( \square \)

### A.3 Known Transition and Bandit Feedback: FTRL with Tsallis Entropy

In this section, we consider the bandit feedback setting with known transition. We use the following hybrid regularizer with learning rate \( \eta_t = \gamma/\sqrt{t} \) for episode \( t \):
\[
\phi_t(q) = \frac{\phi_H(q)}{\eta_t} + \beta \sum_{s \neq s_*} \sum_{a \in A} \log \frac{1}{q(s, a)}. \tag{25}
\]
**Algorithm 3** Best-of-both-worlds for MDPs with Known Transition and Bandit Feedback

for $t = 1$ to $T$ do

compute $q_t = \arg\min_{q \in \Omega} \left\langle q, \sum_{\ell \in \ell_t} \tilde{\ell}_t \right\rangle + \phi_L(q)$ where $\phi_L(q)$ is defined in Eq. (25).

execute policy $\pi_t$ where $\pi_t(s|a) = q_t(s, a) / q_t(s)$.

observe $(s_0, a_0, \ell_0(s_0, a_0), \ldots, (s_{T-1}, a_{T-1}, \ell_T(s_{T-1}, a_{T-1}))$.

construct estimator $\hat{\ell}_t$ such that: $\forall (s, a), \hat{\ell}_t(s, a) = \frac{\ell_t(s, a)}{q_t(s, a)} \mathbb{1}\{s_k(s) = s, a_k(s) = a\}$.

end for

where $\phi_L$ is a fixed log-barrier regularizer, and $\phi_H(q)$ is the $1/2$-Tsallis entropy:

$$\phi_H(q) = -\sum_{s \neq s_L, a \neq \pi(s)} \sqrt{q(s, a)}.$$ 

We present the pseudocode of our algorithm in Algorithm 3, and show the ensured guarantees in Theorem A.3.1, which is a more detailed version of Theorem 3.1. In particular, the adaptive regret bound Eq. (26) is a strict improvement of [Jin and Luo, 2020, Theorem 1] and leads to the best-of-both-worlds guarantee as well. We emphasize that the key to achieve such a guarantee is the loss-shifting function defined in Eq. (6).

**Theorem A.3.1.** With $\beta = 64L$ and $\gamma = 1$, Algorithm 3 ensures that $\text{Reg}_T(\hat{\pi})$ is bounded by

$$\sum_{t=1}^T \mathcal{O}\left(\min_{t} \left\{ \mathbb{E} \left[ B \sum_{s \neq s_L, a \neq \pi(s)} \sqrt{q_t(s, a)} / \eta_t \right] + D \sum_{s \neq s_L, a \neq \pi(s)} q_t(s, a) + \hat{q}(s, a) / \eta_t \right\} \right)$$

for any mapping $\pi : S \rightarrow A$, where $B = L^2$ and $D = \sqrt{|S||A|}$. Therefore, the regret of Algorithm 3 is always bounded as $\text{Reg}_T(\hat{\pi}) = \mathcal{O}\left( \sqrt{|S||A|T}\right)$. Moreover, under Condition (21), $\text{Reg}_T(\hat{\pi})$ is bounded by $\mathcal{O}\left(U + \sqrt{UC}\right)$ where $U = \frac{|S|\log T}{\Delta^2} + \sum_{s \neq s_L, a \neq \pi(s)} \frac{L^2 \log T}{\Delta(s, a)} + L|S||A|\log T$.

**Proof.** By [Jin and Luo, 2020, Lemma 5], with a sufficiently large log-barrier component (in particular, $\beta = 64L$ suffices), the regret can be decomposed and bounded as:

$$\mathbb{E} \left[ \sum_{t=1}^T \left( \frac{1}{\eta_t} - \frac{1}{\eta_t-1} \right) \mathbb{E} [\phi_H(\hat{q}) - \phi_H(q_t)] \right] + 8 \sum_{t=1}^T \eta_t \mathbb{E} \left[ \left\| \hat{\ell}_t \right\|^2_{\nabla^{-2}\phi_H(q_t)} \right]$$

$$+ \mathcal{O}(L|S||A|\log T),$$

where $\hat{q}$ is the occupancy measure of an deterministic optimal policy $\hat{\pi} : S \rightarrow A$. Moreover, with the help of Corollary A.1.2, we can in fact bound $\text{Reg}_T(\hat{\pi})$ as

$$\text{Reg}_T(\hat{\pi}) \leq \sum_{t=1}^T \left( \frac{1}{\eta_t} - \frac{1}{\eta_t-1} \right) \mathbb{E} [\phi_H(\hat{q}) - \phi_H(q_t)] + \mathcal{O}(L|S||A|\log T)$$

$$+ 8 \sum_{t=1}^T \eta_t \mathbb{E} \left[ \min \left\{ \mathbb{E}_t \left[ \left\| \hat{\ell}_t \right\|^2_{\nabla^{-2}\phi_H(q_t)} \right], \mathbb{E}_t \left[ \left\| \hat{\ell}_t + g_t \right\|^2_{\nabla^{-2}\phi_H(q_t)} \right] \right\} \right].$$

where $g_t$ is the specific loss-shifting function defined in Eq. (6). This is again because adding the loss-shifting function $g_t$ does not influence the outcomes of FTRL and thus in the analysis, one can decide whether to add $g_t$ or not for episode $t$ in hindsight to establish a tighter adaptive regret bound.

Before analyzing the stability term, we point out that $\phi_H(\hat{q}) - \phi_H(q_t)$ can be bounded as

$$\phi_H(\hat{q}) - \phi_H(q_t) \leq \sum_{s \neq s_L, a \neq \pi(s)} q_t(s, a) + q_t(s, a) \sqrt{|S||L|} \sum_{s \neq s_L, a \neq \pi(s)} q_t(s, a) + \hat{q}(s, a)$$

(28)
We now bound the stability term. By direct calculation, we have
\[
x \pi \leq \min_{\pi} \left( \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)} + 2 \sqrt{\frac{|S|L}{t}} \right)
\]

We now bound the stability term. By direct calculation, we have
\[
E_t \left[ \left\| \hat{\ell}_t + g_t \right\|_{\nabla^{-2} \phi(q_t)}^2 \right] = \sum_{s \neq s_L} \sum_{a \in A} q_t(s,a)^{3/2} E_t \left[ \left( \hat{\ell}_t(s,a) + g_t(s,a) \right)^2 \right] 
\]
\[
\leq 2L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)} \cdot (1 - \pi_t(a|s)), \quad (29)
\]
where the second line applies the properties of the loss-shifting function in Lemma A.1.3.

For any mapping \( \pi : S \rightarrow A \), we can further bound Eq. (29) as
\[
2L^2 \sum_{s \neq s_L} \sum_{a \in A} \sqrt{q_t(s,a)} \cdot (1 - \pi_t(a|s)) 
\]
\[
\leq 2L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)} + 2L^2 \sum_{s \neq s_L} \sqrt{q_t(s)} \cdot \left( \sum_{a \neq \pi(s)} \pi_t(a|s) \right) 
\]
\[
\leq 4L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)},
\]
where the third line follows from the fact \( x \leq \sqrt{x} \) for \( x \in [0, 1] \).

Therefore, for any mapping \( \pi : S \rightarrow A \), the stability term is bounded by
\[
\sum_{t=1}^T \frac{8\eta_t}{\gamma} E_t \left[ \left\| \hat{\ell}_t + g_t \right\|_{\nabla^{-2} \phi(q_t)}^2 \right] \leq 32L^2 \cdot \sum_{t=1}^T \frac{\eta_t}{\gamma} \left[ \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)} \right]. \quad (30)
\]

On the other hand, without the loss-shifting function, the stability term is simultaneously bounded as
\[
\sum_{t=1}^T \frac{8\eta_t}{\gamma} E_t \left[ \left\| \hat{\ell}_t \right\|_{\nabla^{-2} \phi(q_t)}^2 \right] = \sum_{t=1}^T \frac{8\eta_t}{\gamma} E_t \left[ \sum_{s \neq s_L} \sum_{a \in A} q_t(s,a)^{3/2} \cdot \hat{\ell}_t(s,a)^2 \right] 
\]
\[
\leq \sum_{t=1}^T \frac{8\eta_t}{\gamma} E_t \left[ \sum_{s \neq s_L} \sum_{a \in A} \sqrt{q_t(s,a)} \right] \leq \sum_{t=1}^T \frac{8\eta_t}{\gamma} \sqrt{|S|L}. \quad \text{(Cauchy-Schwarz inequality)}
\]

Plugging Eq. (28) and Eq. (30) into the Eq. (27) shows that Algorithm 3 ensures the following self-bounding regret bound for \( \text{Reg}_T(\hat{\pi}) \):
\[
\frac{1}{\gamma} \sum_{t=1}^T \mathbb{E} \left[ \min \left\{ \sqrt{\frac{|S|L}{t}}, \left( \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)} \right) + 2 \sqrt{\frac{|S|L}{t}} \right\} \right] 
\]
\[
32\gamma \cdot \frac{1}{\gamma} \sum_{t=1}^T \mathbb{E} \left[ \min \left\{ \sqrt{\frac{|S|L}{t}}, L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{q_t(s,a)} \right\} \right] + O \left( L\sqrt{|S|L} \log T \right)
\]
\[
\text{for any mapping } \pi : S \rightarrow A. \text{ Picking } \gamma = 1 \text{ and using } \min \{a,b\} + \min \{c,d\} \leq \min \{a + c, b + d\} \text{ proves Eq. (26).}
\]
The (optimal) worst-case bound \( \text{Reg}_T(\hat{\pi}) = \tilde{O}(\sqrt{L|S||A|T}) \) can be obtained by using the second argument of the min operator in Eq. (26), while the logarithmic regret bound under Condition (21) is obtained by using the first argument of the min operator and the exact same reasoning as in [Jin and Luo, 2020, Appendix A.1].

We point out that with a different choice \( \gamma = \frac{1}{L} \), Algorithm 3 achieves a regret bound of \( \text{Reg}_T(\hat{\pi}) = \mathcal{O}\left(V + \sqrt{VC}\right) \) under Condition (21), where

\[
V = \frac{L^3|S|\log T}{\Delta_{\text{MIN}}} + \sum_{s \neq s_L} \sum_{a \neq \pi^*(a)} \frac{L^2 \log T}{\Delta(s,a)} + L|S||A|\log T
\]

which matches the best existing regret bound in Simchowitz and Jamieson [2019]. This choice of \( \gamma \) worsens the worst-case bound though.
B Best of Both Worlds for MDPs with Unknown Transition and Full Information

In this part, we will prove the best of both worlds results for the full-information setting. We present the bound for the adversarial world in Proposition B.1, and that for the stochastic world in Proposition B.2 (part of which is a restatement of Lemma 5.1). Together, they prove Theorem 4.1.1.

Proposition B.1. Consider the decomposition \( \text{Reg}_T(\pi) = \mathbb{E} [\text{ERR}_1 + \text{ESTReg} + \text{ERR}_2] \) stated in Eq. (14). Then, with \( \delta = \frac{1}{\sqrt{T}} \), Algorithm 1 ensures:

- \( \mathbb{E} [\text{ERR}_1] = \tilde{O} \left( L|S|^4 \sqrt{|A|^3 T} + L^3 |S|^3 |A| \right) \).
- \( \mathbb{E} [\text{ERR}_2] = \tilde{O} (1) \).
- \( \mathbb{E} [\text{ESTReg}] = \tilde{O} \left( L \sqrt{|S||A||T|} + L^2 |S|^2 |A|^2 + L^3 |S||A| \right) \).

Proposition B.2. With \( \delta = \frac{1}{\sqrt{T}} \), Algorithm 1 ensures that \( \text{Reg}_T(\pi^*) \) is bounded as

\[
\tilde{O} \left( \mathbb{E} \left[ \sum_{i} \left( G_1 \left( L^4 |S| \ln T \right) + G_2 \left( L^4 |S| \ln T \right) + G_3 \left( L^4 |S| \ln T \right) + G_4 \left( L^5 |S| |A| \ln T \ln(|S||A|) \right) \right] \right) \\
+ \tilde{O} \left( L^4 |S|^3 |A|^{2 \ln^2 T} \right),
\]

where \( G_1 \) - \( G_4 \) are defined in Definition D.2.1. Under Condition (1), this bound implies \( \text{Reg}_T(\pi^*) = \tilde{O} \left( U + \sqrt{U C} + V \right) \) where

\[
U = \frac{L^6 |S|^2 + L^5 |S||A| \ln(|S||A|)}{\Delta_{\text{MIN}}} \ln T + \frac{L^6 |S| \ln T}{\Delta(s,a)} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} L^6 |S|, \quad V = L^4 |S|^3 |A|^{2 \ln^2 T}.
\]

Before diving into the proof details, we first give formal definitions of several notations mentioned in Section 4 and Section 5 for the full-information setting. Throughout this paper, we denote by \( A \) the event that \( P \in \mathcal{P}_i \) for all \( i \), which happens with probability at least \( 1 - 4 \delta \) based on [Jin et al., 2020, Lemma 2]. We denote by \( N \) the total number of epochs, and set \( t_{N+1} = T + 1 \) for convenience (recall that \( t_i \) is the first episode for epoch \( i \)).

Then, recall the \( \hat{Q}_t^\pi \) and \( \hat{V}_t^\pi \) defined in Section 5, that is, the state-action and state value functions associated with the empirical transition \( \hat{P}_t(\cdot) \) and the adjusted loss \( \hat{\ell}_t \), formally defined as:

\[
\hat{Q}_t^\pi(s,a) = \hat{\ell}_t(s,a) + \sum_{s' \in S_k(s)+1} \hat{P}_t(s'|s,a)\hat{V}_t(s'), \quad \hat{V}_t^\pi(s) = \sum_{a \in A} \pi(a|s)\hat{Q}_t^\pi(s,a),
\]

and \( \hat{Q}_t^\pi(s_L, a) = 0 \) for all \( a \). Also recall that the notation \( \hat{Q}_t \) and \( \hat{V}_t \) used in the loss-shifting function are shorthands for \( \hat{Q}_t^\pi \) and \( \hat{V}_t^\pi \). Similarly, the true state-action and state value functions of episode \( t \) are defined as:

\[
Q_t^\pi(s,a) = \ell_t(s,a) + \sum_{s' \in S_k(s)+1} P(s'|s,a)V_t(s'), \quad V_t^\pi(s) = \sum_{a \in A} \pi(a|s)Q_t^\pi(s,a),
\]

with \( Q_t^\pi(s_L, a) = 0 \) for all \( a \). For notational convenience, we let \( \epsilon = \frac{T|S||A|}{4} \) and assume that \( \delta \in (0,1) \), and denote by \( T_k \) the set of transition tuples at layer \( k \), that is, \( T_k = \{(s, a, s') \in S_k \times A \times S_{k+1} \} \).

B.1 Optimism of Adjusted losses and Other Lemmas

First, we show that the adjusted loss \( \hat{\ell}_t \) defined in Eq. (11) ensures the optimism of the estimated state-action and state value functions as stated in Lemma B.1.1. As discussed in Section 5, this certain kind of optimism ensures that \( \mathbb{E} [\text{ERR}_2] \) is bounded by a constant with a sufficiently small confidence parameter \( \delta \).
Lemma B.1.1. Using the notations in Eq. (32) and Eq. (33) and conditioning on the event \( A \), we have
\[
\hat{Q}^*_t(s, a) \leq Q^*_t(s, a), \forall (s, a) \in S \times A, t \in [T].
\]

Proof. We prove this result via a backward induction from layer \( L \) to layer 0.

Base case: for \( s_L, \hat{Q}^*_t(s, a) = Q^*_t(s, a) = 0 \) holds always.

Induction step: Suppose \( \hat{Q}^*_t(s, a) \leq Q^*_t(s, a) \) holds for all states \( s \) with \( k(s) > h \). Then, for any state \( s \) in layer \( h \), we have
\[
\hat{Q}^*_t(s, a) = \ell_t(s, a) + \sum_{s' \in S_{h+1}} \hat{P}_i(t)(s'|s, a) \hat{V}^*_t(s') - L \cdot B_i(t, s, a)
\]
\[
\leq \ell_t(s, a) + \sum_{s' \in S_{h+1}} \hat{P}_i(t)(s'|s, a) \hat{V}^*_t(s') - L \cdot B_i(t, s, a) \quad \text{(Induction hypothesis)}
\]
\[
\leq \ell_t(s, a) + \sum_{s' \in S_{h+1}} P(s'|s, a) V^*_t(s') + \sum_{s' \in S_{h+1}} (\hat{P}_i(t)(s'|s, a) - P(s'|s, a)) V^*_t(s') - L \cdot B_i(t, s, a)
\]
\[
= Q^*_t(s, a) + \sum_{s' \in S_{h+1}} (\hat{P}_i(t)(s'|s, a) - P(s'|s, a)) V^*_t(s') - L \cdot B_i(t, s, a)
\]

where the first line follows from the definition of \( \hat{\ell}_t \).

Clearly, when \( B_i(t, s, a) = 1 \), we have
\[
\sum_{s' \in S_{h+1}} (\hat{P}_i(t)(s'|s, a) - P(s'|s, a)) V^*_t(s') - L \cdot B_i(t, s, a) \leq \sum_{s' \in S_{h+1}} \hat{P}_i(t)(s'|s, a) \cdot L = 0
\]

where the inequality follows from the fact \( 0 \leq V^*_t(s') \leq L \).

On the other hand, when \( \sum_{s' \in S_{h+1}} B_i(t, s, a, s') = B_i(t, s, a) \), we have
\[
\sum_{s' \in S_{h+1}} (\hat{P}_i(t)(s'|s, a) - P(s'|s, a)) V^*_t(s') - L \cdot B_i(t, s, a)
\]
\[
\leq \sum_{s' \in S_{h+1}} B_i(t, s, a, s') \cdot L - L \cdot B_i(t, s, a) = 0
\]

where the second line uses the definition of event \( A \).

Combining these two cases shows that \( \hat{Q}^*_t(s, a) \leq Q^*_t(s, a) \) holds for all state-action pairs \((s, a)\) at layer \( h \), finishing the induction.

Next, we analyze the estimated regret suffered within one epoch. With slightly abuse of notation, we denote by \( \text{EstReg}_t(\pi) \) the difference between the total loss suffered within epoch \( i \) and that of the fixed policy \( \pi \) with respect to the empirical transition \( \hat{P}_i \) and the adjusted losses within epoch \( i \), that is,
\[
\text{EstReg}_t(\pi) = E \left[ \sum_{t=t_i}^{t_i+1} \left( q^{\hat{P}_i, \pi} - q^{\hat{P}_i, \hat{\ell}_t} \right) \right] = E \left[ \sum_{t=t_i}^{t_i+1} \left( \hat{q}_t - q^{\hat{P}_i, \pi} , \hat{\ell}_t \right) \right].
\]

In addition, we let \( \text{EstReg}_i = \max_{\pi} \text{EstReg}_t(\pi) \) be the maximum regret suffered within epoch \( i \).

Lemma B.1.2. For full-information feedback, Algorithm 1 ensures that \( \text{EstReg}_i \) is bounded by \( O(L^3 \ln(|S||A|)) \) plus:
\[
O \left( \sqrt{E \left( \ln(|S||A|) \cdot \min \left\{ L^4 \sum_{t=t_i}^{t_i+1} \sum_{s \in S} \sum_{a \notin \pi_0(s)} \sum_{s' \in S, a \in A} \hat{q}_t(s, a), \sum_{s \in S} \sum_{s' \in S, a \in A} \sum_{t=t_i}^{t_i+1} \hat{q}_t(s, a) \hat{e}_t(s, a)^2 \right\} \right)} \right). \tag{35}
\]
Proof. The proof follows the same steps as in that of Theorem A.2.1. Due to the invariant property, the loss of episode \( t \) fed to FTRL can be seen as either \( \tilde{\ell}_t(s, a) \) or \( \tilde{Q}_t(s, a) - \tilde{V}_t(s, a) \). By the definition of \( \eta_t \), we have both \( \eta_t \tilde{\ell}_t(s, a) \geq -1 \) and \( \eta_t (\tilde{Q}_t(s, a) - \tilde{V}_t(s, a)) \geq -1 \). Therefore, we can apply Lemma A.2.3 and bound EstReg, by

\[
\mathbb{E}\left[ \frac{L \ln(|S||A|)}{\eta_{t+1}} + \sum_{t=t_i}^{t_{i+1} - 1} \eta_t \min \left\{ \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s, a) \left( \tilde{Q}_t(s, a) - \tilde{V}_t(s) \right)^2, \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s, a) \ell_t(s, a)^2 \right\} \right].
\]

The tuning of \( \eta_t \) makes sure that the above is further bounded by \( O \left( L^3 \ln(|S||A|) \right) \) plus \( \sqrt{L \ln \left( |S||A| \right)} \) multiplied with

\[
O \left( \mathbb{E} \left[ \min \left\{ \sum_{t=t_i}^{t_{i+1} - 1} \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s, a) \left( \tilde{Q}_t(s, a) - \tilde{V}_t(s) \right)^2, \sum_{t=t_i}^{t_{i+1} - 1} \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s, a) \ell_t(s, a)^2 \right\} \right] \right);
\]

see the beginning of the proof of Theorem A.2.1 for the same reasoning. Finally, it remains to bound

\[
\sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s, a) \left( \tilde{Q}_t(s, a) - \tilde{V}_t(s) \right)^2 \]

by \( 8L^4 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \tilde{q}_t(s, a) \). This is again by the same reasoning as Eq. (23) and Eq. (24), except that \( \tilde{Q}_t(s, a) \) now has a range in \([-L^2, L^2]\) which explains the extra \( L^2 \) factor.

B.2 Proof for the Adversarial World (Proposition B.1)

We analyze the regret based on the decomposition in Eq. (14) and consider bounding the terms \( \mathbb{E} [\text{ERR}_1] \), \( \mathbb{E} [\text{ERR}_2] \) and \( \mathbb{E} [\text{EstReg}] \) separately.

**ERR**

Following the similar idea of Jin et al. [2020], we decompose this term as:

\[
\text{ERR}_1 = \sum_{t=1}^{T} V_t^{\pi_t}(s_0) - V_t^{\pi_t}(s_0) = \sum_{t=1}^{T} \langle q_t, \ell_t \rangle - \langle \tilde{q}_t, \tilde{\ell}_t \rangle
\]

\[
= \sum_{t=1}^{T} \langle q_t, \ell_t \rangle - \langle \tilde{q}_t, \ell_t \rangle + L \cdot \sum_{t=1}^{T} \langle \tilde{q}_t, B_{i(t)} \rangle
\]

\[
= \sum_{t=1}^{T} \langle q_t, \ell_t \rangle - \langle \tilde{q}_t, \ell_t \rangle + L \cdot \sum_{t=1}^{T} \langle \tilde{q}_t, B_{i(t)} \rangle + L \cdot \sum_{t=1}^{T} \langle \tilde{q}_t - q_t, B_{i(t)} \rangle
\]

\[
\leq \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} |q_t(s, a) - \tilde{q}_t(s, a)| + L \cdot \sum_{t=1}^{T} \langle \tilde{q}_t - q_t, B_{i(t)} \rangle
\]

where the last line follows from the fact \( 0 \leq \ell_t(s, a) \leq 1 \). According to this decomposition, we next consider bounding the expectation of these three terms separately.

First, we focus on the second term:

\[
\mathbb{E} \left[ L \cdot \sum_{t=1}^{T} \langle q_t, B_{i(t)} \rangle \right]
\]

\[
\leq L \cdot \mathbb{E} \left[ \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{a \in A} \sum_{t=1}^{T} q_t(s, a) \left( 2 \sqrt{\frac{|S_k(s,a) + 1| \ln t}{\max \{ m_i(s,a), 1 \}}} + 14\frac{|S_k(s,a) + 1| \ln t}{3 \max \{ m_i(s,a), 1 \}} \right) \right]
\]

\[
= O \left( L \cdot \sum_{k=0}^{L-1} \left( \sqrt{|S_k||S_{k+1}| |A| T \ln t} + |S_k(s,a) + 1| |S_k||A| (2 + \ln T) \ln t \right) \right)
\]

\[
\leq O \left( L \sqrt{|A| T \ln t} \cdot \sum_{k=0}^{L-1} (|S_k| + |S_{k+1}|) + L|S||A| \ln^2 t \right)
\]

\[
\leq O \left( L \sqrt{|A| T \ln t} \cdot \sum_{k=0}^{L-1} (|S_k| + |S_{k+1}|) + L|S||A| \ln^2 t \right)
\]

\[
\leq O \left( L \sqrt{|A| T \ln t} \cdot \sum_{k=0}^{L-1} (|S_k| + |S_{k+1}|) + L|S||A| \ln^2 t \right)
\]

\[
\leq O \left( L \sqrt{|A| T \ln t} \cdot \sum_{k=0}^{L-1} (|S_k| + |S_{k+1}|) + L|S||A| \ln^2 t \right)
\]

\[
\leq O \left( L \sqrt{|A| T \ln t} \cdot \sum_{k=0}^{L-1} (|S_k| + |S_{k+1}|) + L|S||A| \ln^2 t \right)
\]
where the second line follows Lemma D.3.2, the third line follows from Lemma D.3.8 and the fourth line applies AM-GM inequality.

Then, for the first term, with the help from residual term \( r_t \) defined in Definition D.3.9, we have

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} |q_t(s, a) - \hat{q}_t(s, a)| \right]
\]

\[
\leq \mathbb{E} \left[ 4 \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} \sum_{k=0}^{k(s)-1} q_t(u,v) \sqrt{P(w|u,v) \ln t \over \max \{m_{i(t)}(u,v), 1\}} q_t(s,a|w) + \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} r_t(s,a) \right]
\]

\[
\leq \mathbb{E} \left[ 4L \cdot \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{v \in A} \sum_{u \in \mathcal{S}_k(u) + 1} q_t(u,v) \sqrt{S_k(u) \ln t \over \max \{m_{i(t)}(u,v), 1\}} + \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} r_t(s,a) \right]
\]

\[
= \mathcal{O} \left( L|S|\sqrt{|A|T \ln t + L^2|S|^2 |A|^2 \ln^2 t + \delta |S||A|T} \right)
\]

\[
= \tilde{\mathcal{O}} \left( L|S|\sqrt{|A|T + L^2|S|^2 |A|^2} \right)
\]  

(36)

where the second line uses the bound of \( |q_t(s,a) - \hat{q}_t(s,a)| \) in Lemma D.3.10; the third line follows from the fact \( \sum_{s \neq s_t} \sum_{a \in A} q_t(s,a|w) \leq L \); the forth line uses the Cauchy-Schwarz inequality; the fifth line follows the same argument in Eq. (36) and applies the expectation bound of residual terms in Lemma D.3.10; and the last line plugs in the value of \( \delta = 1/T^2 \).

For the last term, using the bound of \( |\hat{q}_t(s,a) - q_t(s,a)| \) in Lemma D.3.10, we arrive at

\[
\mathbb{E} \left[ L \cdot \sum_{t=1}^{T} \langle \hat{q}_t - q_t, B_{i(t)} \rangle \right]
\]

\[
\leq \mathbb{E} \left[ L \cdot \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} B_{i(t)}(s,a) \cdot \left( 4 \sum_{k=0}^{k(s)-1} q_t(u,v) \sqrt{P(w|u,v) \ln t \over \max \{m_{i(t)}(u,v), 1\}} q_t(s,a|w) + r_t(s,a) \right) \right]
\]

\[
\leq \mathbb{E} \left[ 4L \cdot \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} \sum_{s' \in \mathcal{S}_k(s) + 1} B_{i(t)}(s,a,s') \sum_{k=0}^{k(s)-1} q_t(u,v) \sqrt{P(w|u,v) \ln t \over \max \{m_{i(t)}(u,v), 1\}} q_t(s,a|w) \right]
\]

\[
+ \mathbb{E} \left[ L \cdot \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} r_t(s,a) \right],
\]

where the last line follows from the fact \( B_{i(t)}(s,a) \leq 1 \).

According to the definition of the residual term in Definition D.3.9, we have

\[
\sum_{s' \in \mathcal{S}_k(s) + 1} B_{i(t)}(s,a,s') \cdot \sum_{k=0}^{k(s)-1} q_t(u,v) \sqrt{P(w|u,v) \ln t \over \max \{m_{i(t)}(u,v), 1\}} q_t(s,a|w)
\]

(in particular, the second summand in the definition of \( r_t(s,a) \) is an upper bound of the right-hand side above). Therefore, we have

\[
\mathbb{E} \left[ L \cdot \sum_{t=1}^{T} \langle \hat{q}_t - q_t, B_{i(t)} \rangle \right] \text{ further bounded by}
\]

\[
\mathbb{E} \left[ (4L + L) \cdot \sum_{t=1}^{T} \sum_{s \neq s_t} \sum_{a \in A} r_t(s,a) \right] \leq \mathcal{O} \left( L^3 |S|^3 |A|^2 \ln^2 t + \delta \cdot L|S||A|T \right) = \tilde{\mathcal{O}} \left( L^3 |S|^3 |A|^2 \right)
\]

(38)
where the last inequality uses the expectation bound of residual terms in Lemma D.3.10.

Combining all bounds yields
\[
\mathbb{E}[\text{ERR}_1] = \tilde{O}\left(L|S|\sqrt{|A|T} + L^3|S|^3|A|\right).
\]

**ERR** 
According to Lemma B.1.1, Lemma D.3.5, and the fact \(|V_\pi^\pi(s)| \leq L^2\), we have
\[
\mathbb{E}[\text{ERR}_2] = \mathbb{E}\left[\sum_{t=1}^T \tilde{V}_t^\pi(s_0) - V_t^\pi(s_0)\right] \leq L^2 T \Pr[\mathcal{A}'] \leq 4L^2 T \delta = \tilde{O}(1).
\]

**ESTREG** 
By Lemma B.1.2, we have \(\mathbb{E}[\text{ESTREG}]\) bounded as
\[
\mathbb{E}\left[\sum_{t=1}^N \sum_{t'=t}^{t+1} \langle \tilde{q}_t - q_{t',\pi}, \tilde{\ell}_t \rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^N \text{EstReg}_t\right] \leq \tilde{O}\left(\mathbb{E}\left[L|S||A| \sum_{t=1}^T \sum_{s \in S} \sum_{a \in A} \tilde{q}_t(s,a) \tilde{\ell}_t(s,a)^2 + L^3\right]\right)
\]
\[
\leq \tilde{O}\left(\mathbb{E}\left[L|S||A| \sum_{t=1}^T \sum_{s \in S} \sum_{a \in A} \tilde{q}_t(s,a) \tilde{\ell}_t(s,a)^2\right] + L^3|S||A|\right)
\]
where the last line follows from the fact \(N \leq 4|S||A| (\log T + 1)\) according to Lemma D.3.12 and uses Cauchy-Schwarz inequality.

Next, we continue to bound the following key term:
\[
\mathbb{E}\left[\sum_{t=1}^T \sum_{s \in S} \sum_{a \in A} \tilde{q}_t(s,a) \tilde{\ell}_t(s,a)^2\right]
\]
\[
= \mathbb{E}\left[\sum_{t=1}^T \sum_{s \in S} \sum_{a \in A} \tilde{q}_t(s,a) \left(\ell_t(s,a) - L \cdot B_{t(i)}(s,a)\right)^2\right]
\]
\[
\leq 2 \cdot \mathbb{E}\left[\sum_{t=1}^T \sum_{s \in S} \sum_{a \in A} \tilde{q}_t(s,a) \left(\ell_t(s,a)^2 + L^2 \cdot B_{t(i)}(s,a)^2\right)\right]
\]
\[
\leq 2LT + 2L^2 \cdot \mathbb{E}\left[\sum_{t=1}^T \langle \tilde{q}_t, B_{t(i)} \rangle\right]
\]
\[
= 2LT + 2L \cdot \left(L \cdot \mathbb{E}\left[\sum_{t=1}^T \langle \tilde{q}_t - q_t, B_{t(i)} \rangle\right] + L \cdot \mathbb{E}\left[\sum_{t=1}^T \langle q_t, B_{t(i)} \rangle\right]\right),
\]
where the third line uses \((x + y)^2 \leq 2(x^2 + y^2)\) and the fourth line uses \(B_{t(i)}(s,a) \leq 1\). Moreover, in the previous analysis of the term ERR, we bound the terms in the bracket with
\[
\mathbb{E}\left[L \cdot \sum_{t=1}^T \langle q_t, B_{t(i)} \rangle\right] \leq \tilde{O}\left(L|S|\sqrt{|A|T} + L^2|S|^2|A|\right), \quad \text{(from Eq. (36))}
\]
\[
\mathbb{E}\left[L \cdot \sum_{t=1}^T \langle \tilde{q}_t - q_t, B_{t(i)} \rangle\right] \leq \tilde{O}\left(L^3|S|^3|A|^2\right). \quad \text{(from Eq. (38))}
\]

Therefore, we have
\[
\mathbb{E}\left[\sum_{t=1}^T \sum_{s \in S} \sum_{a \in A} \tilde{q}_t(s,a) \tilde{\ell}_t(s,a)^2\right] = \tilde{O}\left(LT + L|S|\sqrt{|A|T} + L^3|S|^3|A|^2\right) = \tilde{O}\left(LT + L^3|S|^3|A|^2\right),
\]
which further proves
\[
\mathbb{E}[\text{ESTREG}] = \tilde{O}\left(L\sqrt{|S||A|T} + L^2|S|^2|A|^2 + L^3|S||A|\right).
\]
B.3 Proof for the Stochastic World (Proposition B.2)

As discussed in Section 5, we decompose $\text{ERR}_1 + \text{ERR}_2$ as (see Corollary D.1.2):

$$\text{ERR}_1 + \text{ERR}_2 = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \hat{E}^\pi_t(s, a) \tag{ERRSUB}$$

$$+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a=\pi^*(s)} (q_t(s, a) - q_t(s, a)) \hat{E}^\pi_t(s, a) \tag{ERROPT}$$

$$+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} (q_t(s, a) - \hat{q}_t(s, a)) \left( \hat{Q}^\pi_t(s, a) - \hat{V}^\pi_t(s) \right) \tag{OCCDIFF}$$

$$+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \left( \hat{V}^\pi_t(s) - V^\pi_t(s) \right) \tag{BIAS}$$

where $\hat{E}^\pi_t(s, a)$ is defined as:

$$\hat{E}^\pi_t(s, a) = \ell_t(s, a) + \sum_{s' \in S_{k(t)+1}} P(s'|s, a) \hat{V}^\pi_t(s') - \hat{Q}^\pi_t(s, a).$$

Then, we proceed to bound each of the five terms: ERRSUB, ERROPT, OCCDIFF, BIAS, and ESTREG.

**ERRSUB** Conditioning on $A$, we know that

$$\hat{E}^\pi_t(s, a) = LB_t(s, a) + \sum_{s' \in S_{k(t)+1}} (P(s'|s, a) - \bar{P}_t(s'|s, a)) \hat{V}^\pi_t(s')$$

$$\leq LB_t(s, a) + L^2 \cdot \sum_{s' \in S_{k(t)+1}} B_t(s, a, s')$$

$$\leq 4L^2 \cdot \sum_{s' \in S_{k(t)+1}} \left( \sqrt{\frac{P_t(s'|s, a) \ln \ell}{\max \{m_t(s, a), 1\}}} + \frac{7 |S| \ln \ell}{3 \max \{m_t(s, a), 1\}} \right)$$

$$\leq 4L^2 \left( \sqrt{\frac{|S| \ln \ell}{\max \{m_t(s, a), 1\}}} + \frac{7 |S| \ln \ell}{3 \max \{m_t(s, a), 1\}} \right),$$

where the second line follows from the event $A$ and the fact $\left| \hat{V}^\pi_t(s) \right| \leq L^2$, and the last line applies the Cauchy-Schwarz inequality.

Therefore, under event $A$, ERRSUB can be bounded as:

$$\text{ERRSUB} \leq \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \cdot 4L^2 \left( \sqrt{\frac{|S| \ln \ell}{\max \{m_t(s, a), 1\}}} + \frac{7 |S| \ln \ell}{3 \max \{m_t(s, a), 1\}} \right)$$

$$\leq 4G_1 (L^4 |S| \ln \ell) + \frac{28 |S| L^2 \ln \ell}{3} \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) \sum_{a \neq \pi^*(s)} \frac{q_t(s, a)}{3 \max \{m_t(s, a), 1\}},$$

where the second line follows from the definition of $G_1(\cdot)$ in Definition D.2.1.

With the help of Lemma D.3.5 and the fact $|\text{ERRSUB}| \leq L^3 T$, we have

$$\mathbb{E}[\text{ERRSUB}] \leq O \left( L^3 T \delta + \mathbb{E} \left[ G_1 \left( L^4 |S| \ln \ell \right) \right] \right) + \mathbb{E} \left( \frac{28 |S| L^2 \ln \ell}{3} \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) \sum_{a \neq \pi^*(s)} \frac{q_t(s, a)}{3 \max \{m_t(s, a), 1\}} \right)$$

$$= O \left( \mathbb{E} \left[ G_1 \left( L^4 |S| \ln \ell \right) \right] \right) + L^3 |S|^2 |A| \ln^2 \ell, \tag{39}$$

where the last line uses Lemma D.3.8.
\[ \text{ErrOpt} \] By the similar arguments above, we have \text{ErrOpt} bounded by the following given event \( A \):

\[
\text{ErrOpt} \leq \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \cdot 4L^2 \left( \sqrt{\frac{|S| \ln t}{\max \{m_{i(t)}(s, a), 1\}}} + \frac{7|S| \ln t}{3 \max \{m_{i(t)}(s, a), 1\}} \right).
\]

Using the definition of \( G_2(\cdot) \) in \text{Definition D.2.1} and \text{Lemma D.3.5}, we have

\[
E[\text{ErrOpt}] \leq O \left( L^3 T \delta + E \left[ G_2 \left( L^4 |S| \ln t \right) \right] \right) + E \left[ \sum_{t=1}^{T} \frac{28|S| L^2 \ln t}{3} \sum_{s \neq s_L} \sum_{a \in A} \frac{q_t(s, a)}{3 \max \{m_{i(t)}(s, a), 1\}} \right]
\]

where the second line follows from the fact \( \hat{V}_t^\pi(s) = Q_t^\pi(s, a) \) for all state-action pairs \((s, a)\) satisfying \( a = \pi^*(s) \), and the last line uses the fact \( Q_t^\pi(s, a) - \hat{V}_t^\pi(s) \leq 2L^2 \) for all state-action pairs. With the help of the residual terms in \text{Definition D.3.9} and \text{Lemma D.3.10}, we further bound \text{OccDiff} as

\[
2L^2 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} |q_t(s, a) - \hat{q}_t(s, a)|
\]

\[
\leq 2L^2 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} r_t(s, a)
\]

\[
+ 8L^2 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \sqrt{\frac{P(u|v, w) \ln t}{\max \{m_{i(t)}(u, v), 1\}}} q_t(s, a|w)
\]

where the last line is by the definition of \( G_3(\cdot) \) in \text{Definition D.2.1}. Therefore, we conclude

\[
E[\text{OccDiff}] \leq O \left( L^4 |S|^3 |A|^2 \ln^2 t + L^2 |S||A|T \cdot \delta + G_3(L^4 \ln t) \right)
\]

\[
\text{Bias} \quad \text{Conditioning on the event } A, \text{Bias is nonpositive due to \text{Lemma B.1.1}. Then, by \text{Lemma D.3.5}, we bound the expectation of \text{Bias by}}
\]

\[
E[\text{Bias}] \leq 0 + E \left[ \mathbb{P}\left( \{A^c\} \right) \right] \cdot L^3 T = O \left( 1 \right).
\]

\[
\text{EstReg} \quad \text{By the analysis of estimated regret in \text{Lemma B.1.2}, we have } E[\text{EstReg}] \text{ bounded by (with } C_{\text{EstReg}} = L^5 |S||A| \ln T \ln \left( |S||A| \right))
\]

\[
O \left( E \left[ \sum_{k=1}^{N} \left( L^5 \ln (|S||A|) \cdot \sum_{k=t_i+1}^{t_{i+1}-1} \sum_{s \in S} \sum_{a \neq \pi^*(s)} \hat{q}_t(s, a) + L^3 \ln (|S||A|) \right) \right] \right)
\]

\[
\leq O \left( E \left[ C_{\text{EstReg}} \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \hat{q}_t(s, a) + L^3 |S||A| \ln T \ln (|S||A|) \right] \right)
\]
where the second line uses the Cauchy-Schwarz inequality and the fact $N \leq 4|S||A|$ (log $T + 1$) according to Lemma D.3.12; the third line uses the fact that $\sqrt{x} \leq \sqrt{y} + \sqrt{|x - y|}$ for $x, y > 0$; the last line uses the definition of $G_4(\cdot)$ in Definition D.2.1 and the AM-GM inequality.

Note that in the analysis of OCCDIFF (see Eq. (41)), we have already shown that

$$\sum_{t=1}^{T} \sum_{s \neq S_L, a \neq \pi^*(s)} |q(s, a) - \tilde{q}(s, a)| = O\left(L^2|S|^3|A|^2 \ln^2 \epsilon + \mathbb{E}[G_3(\ln \epsilon)]\right). \quad (44)$$

Combining everything, we have $\mathbb{E}[[\text{ESTREG}]]$ bounded by:

$$O\left(\mathbb{E}[G_4(L^5|S||A| \ln (|S||A|))) + G_5(\ln \epsilon)] + L^5|S|^3|A|^2 \ln^2 \epsilon\right). \quad (45)$$

Finally, combining everything we have shown that Algorithm 1 ensures the following regret bound for $\text{Reg}_T(\pi^*)$:

$$\mathcal{O}\left(\mathbb{E}\left[G_1\left(L^4|S| \ln \epsilon\right)\right] + \mathbb{E}\left[G_2\left(L^4|S| \ln \epsilon\right)\right] + \mathbb{E}\left[G_3\left(L^4 \ln \epsilon\right)\right] + \mathbb{E}\left[G_4\left(L^5|S||A| \ln (|S||A|) \ln T\right)\right]\right)$$

Now suppose that Condition (1) holds. For some universal constant $\kappa > 0$, $\text{Reg}_T(\pi^*)$ is bounded as

$$\text{Reg}_T(\pi^*) \leq \kappa \cdot \left(\mathbb{E}\left[G_1\left(L^4|S| \ln \epsilon\right)\right] + \mathbb{E}\left[G_2\left(L^4|S| \ln \epsilon\right)\right] + \mathbb{E}\left[G_3\left(L^4 \ln \epsilon\right)\right] + \mathbb{E}\left[G_4\left(L^5|S||A| \ln (|S||A|) \ln T\right)\right]\right) + \kappa \cdot \left(L^4|S|^3|A|^2 \ln^2 \epsilon\right).$$

For any $z > 0$, by Lemma D.2.2, Lemma D.2.3, Lemma D.2.4 and Lemma D.2.5 with $\alpha = \beta = \frac{1}{12\pi \kappa}$, we have

$$\text{Reg}_T(\pi^*) \leq \frac{\text{Reg}_T(\pi^*) + C}{z} + 12z \cdot \left(\sum_{s \neq S_L, a \neq \pi^*(s)} \frac{8\kappa^2}{\Delta(s, a)} \cdot \left(L^4|S| \ln \epsilon + L^6|S| \ln \epsilon\right)\right)$$

$$+ 12z \cdot \left(\frac{\kappa^2}{\Delta_{\min}} \cdot \left(8L^5|S| \ln \epsilon + 8L^6|S|^2 \ln \epsilon + \frac{L^5|S||A| \ln (|S||A|) \ln T}{4}\right)\right)$$

$$+ \kappa \cdot \left(L^4|S|^3|A|^2 \ln^2 \epsilon\right)$$

$$\leq \frac{\text{Reg}_T(\pi^*) + C}{z} + 288z \kappa^2 \cdot U + 2\kappa \cdot V,$$
where the last line uses the shorthands $U$ and $V$ defined in Proposition B.2.

Rearranging the terms arrive at:

$$\text{Reg}_T(\pi^*) \leq \frac{C}{z-1} + \frac{z^2}{z-1} \cdot 288\kappa^2 U + \frac{z}{z-1} \cdot 2\kappa \cdot V$$

$$= \frac{C}{x} + \frac{(x+1)^2}{x} \cdot 288\kappa^2 U + \frac{x+1}{x} \cdot 2\kappa \cdot V$$

$$= \frac{1}{x} \cdot (C + 288\kappa^2 U + 2\kappa \cdot V) + x \cdot 288\kappa^2 U + 2\kappa \cdot V + 576\kappa^2 U$$

where we replace all $z$'s by $x = z - 1 > 0$ in the second line. Finally, by selecting the optimal $x$ to balance the first two terms, we have

$$\text{Reg}_T(\pi^*) \leq 2\sqrt{(C + 288\kappa^2 U + 2\kappa \cdot V) \cdot 288\kappa^2 U + 2\kappa \cdot V + 576\kappa^2 U}$$

$$= \mathcal{O}(U \sqrt{UC} + V),$$

finishing the entire proof for Proposition B.2.
C  Best of Both Worlds for MDPs with Unknown Transition and Bandit Feedback

In this section, we prove the best of both worlds results for the bandit setting with unknown transition. We present the bound for the adversarial world in Proposition C.1, and that for the stochastic world in Proposition C.2. Together, they prove Theorem 4.1.2.

**Proposition C.1.** With $\delta = \frac{1}{T^{2}}$, Algorithm 1 ensures

$$
\text{Reg}_{\pi}(\hat{\pi}) = O \left( \left( L + \sqrt{A} \right) |S||A|T \right).
$$

**Proposition C.2.** Suppose Condition (1) holds. With $\delta = \frac{1}{T^{2}}$, Algorithm 1 ensures that $\text{Reg}_{\pi}(\pi^{*})$ is bounded by $O \left( U + \sqrt{CU} + V \right)$ where $V = L^{6}|S|^{3}|A|^{3}\ln^{2}T$ and $U$ is defined as

$$
U = \sum_{s\neq s_{L}} \sum_{a\neq \pi(s)} \left[ \frac{L^{6}|S|\ln T + L^{4}|S||A|\ln^{2}T}{\Delta(s, a)} \right] + \left[ \frac{L^{6}|S|^{2}\ln T + L^{3}|S|^{2}|A|\ln^{2}T}{\Delta_{\text{MIN}}} \right].
$$

The analysis is similar to that for the full-information setting, except that we need to handle some bias terms caused by the new loss estimators. To this end, we denote by $\hat{\ell}_{i}$ the conditional expectation of $\ell_{i}$, that is

$$
\hat{\ell}_{i}(s, a) = E_{i} \left[ \ell_{i}(s, a) \right] = \frac{q_{i}(s, a)}{u_{i}(s, a)} \cdot \ell_{i}(s, a) - L \cdot B_{i}(s, a).
$$

Then we define the following:

**Definition C.3.** For any policy $\pi$, the estimated state-action and state value functions associated with $\hat{P}_{i}(t)$ and loss function $\hat{\ell}_{i}$ are defined as:

$$
\hat{Q}_{i}^{\pi}(s, a) = \ell_{i}(s, a) + \sum_{s'\in S_{k+1}} \hat{P}_{i}(s'|s, a)\hat{V}_{i}^{\pi}(s'), \quad \forall(s, a) \in (S - \{s_{L}\}) \times A,
$$

$$
\hat{V}_{i}^{\pi}(s) = \sum_{a\in A} \pi(a|s)\hat{Q}_{i}^{\pi}(s, a), \quad \forall s \in S,
$$

$$
\hat{Q}_{i}^{s_{L}}(s_{L}, a) = 0, \quad \forall a \in A.
$$

On the other hand, the true state-action and value functions are again defined as:

$$
Q_{i}^{\pi}(s, a) = \ell_{i}(s, a) + \sum_{s'\in S_{k+1}} P(s'|s, a)V_{i}^{\pi}(s'), \quad \forall(s, a) \in (S - \{s_{L}\}) \times A,
$$

$$
V_{i}^{\pi}(s) = \sum_{a\in A} \pi(a|s)Q_{i}^{\pi}(s, a), \quad \forall s \in S,
$$

$$
Q_{i}^{s_{L}}(s_{L}, a) = 0, \quad \forall a \in A.
$$

where $P$ denotes the true transition function.

Besides the definition of event $A$, we also define $A_{i}$ to be the event $P \in A_{i}$. Importantly, the value of $E\{A_{i}\}$ is only based on observations prior to epoch $i$. For notational convenience, we again let $\ell = \frac{T|S||A|}{\delta}$ and assume $\delta \in (0, 1)$.

Similarly to the full-information setting, we decompose the regret against policy $\pi$, $\text{Reg}(\pi) = E \left[ \sum_{t=1}^{T} V_{t}^{\pi}(s_{0}) - V_{t}^{\pi}(s_{0}) \right]$, as

$$
E \left[ \sum_{t=1}^{T} V_{t}^{\pi}(s_{0}) - \hat{V}_{t}^{\pi}(s_{0}) \right] + E \left[ \sum_{t=1}^{T} \hat{V}_{t}^{\pi}(s_{0}) - \hat{V}_{t}^{\pi}(s_{0}) \right] + E \left[ \sum_{t=1}^{T} \hat{V}_{t}^{\pi}(s_{0}) - V_{t}^{\pi}(s_{0}) \right].
$$

Note that, the second term is exactly

$$
E \left[ \text{ESTREG} \right] = E \left[ \sum_{t=1}^{T} \langle q_{t}, \pi_{t} - q_{t}^{\hat{P}_{i}}, \hat{\ell}_{i} \rangle \right] = E \left[ \sum_{t=1}^{T} \langle q_{t}, \pi_{t} - q_{t}^{\hat{P}_{i}}, \hat{\ell}_{i} \rangle \right],
$$

which is controlled by the FTRL process.

34
C.1 Auxiliary Lemmas

First, we show the following optimism lemma.

**Lemma C.1.1.** With the notations defined in Eq. (47) and Eq. (48), the following holds conditioning on event $\mathcal{A}$:

$$\bar{Q}^π_t(s, a) \leq Q^π_t(s, a), \forall (s, a) \in S \times A, t \in [T].$$

Specifically, we have

$$\langle q^{\bar{P}_i(\cdot, \pi), \ell_i} \rangle = \bar{V}^π_t(s_0) \leq V^π_t(s_0) = \langle q^{P, \pi}, \ell_i \rangle.$$

**Proof.** We prove this result via a backward induction from layer $L$ to layer 0.

**Base case:** for $s_L$, $\bar{Q}^π_t(s, a) = Q^π_t(s, a) = 0$ holds always.

**Induction step:** Suppose $\bar{Q}^π_t(s, a) \leq Q^π_t(s, a)$ holds for all states $s$ with $k(s) > h$. Then, for any state $s$ with $k(s) = h$, we have

$$\bar{Q}^π_t(s, a) = \frac{q_t(s, a)}{u_t(s, a)} \cdot \ell_t(s, a) + \sum_{s' \in S_{k(s) + 1}} \bar{P}_i(t)(s'|s, a)\bar{V}^π_t(s') - L \cdot B_i(t)(s, a) \quad \text{(Eq. (46))}$$

$$\leq \frac{q_t(s, a)}{u_t(s, a)} \cdot \ell_t(s, a) + \sum_{s' \in S_{k(s) + 1}} \bar{P}_i(t)(s'|s, a)V^π_t(s') - L \cdot B_i(t)(s, a) \quad \text{(induction hypothesis)}$$

$$\leq \frac{q_t(s, a)}{u_t(s, a)} \cdot \ell_t(s, a) + \sum_{s' \in S_{k(s) + 1}} P(s'|s, a)V^π_t(s')$$

$$+ \sum_{s' \in S_{k(s) + 1}} \left( \bar{P}_i(t)(s'|s, a) - P(s'|s, a) \right)V^π_t(s') - L \cdot B_i(t)(s, a)$$

$$\leq \frac{q_t(s, a)}{u_t(s, a)} \cdot \ell_t(s, a) + \sum_{s' \in S_{k(s) + 1}} P(s'|s, a)V^π_t(s')$$

$$\leq \ell_t(s, a) + \sum_{s' \in S_{k(s) + 1}} P(s'|s, a)V^π_t(s') = Q^π_t(s, a),$$

where the forth step follows from the same arguments in Lemma B.1.1, and the last step holds since under event $\mathcal{A}$, we have $q_t(s, a) \leq u_t(s, a)$ by the definition of $u_t$. This finishes the induction.

Next, we provide a sequence of boundedness results, useful for regret analysis.

**Lemma C.1.2 (Lower Bound of Upper Occupancy Bound).** Algorithm 1 ensures $u_t(s) \geq \frac{1}{|S|}$ for all $t$ and $s$.

**Proof.** We prove by constructing a special transition function $\tilde{P}_i(t)$ within the confidence set $\mathcal{P}_i(t)$, which ensures $q^{P_i(t), \pi_t}(s) \geq \frac{1}{|S|}$ for all state-action pairs. Specifically, let $\tilde{P}_i(t)$ be such that

$$\tilde{P}_i(t)(s'|s, a) = \frac{1}{t} \cdot \frac{1}{|S_{k(s) + 1}|} + \frac{t - 1}{t} \cdot \bar{P}_i(t)(s'|s, a), \quad \forall (s, a, s') \in T_k, k < L.$$

Clearly, $\tilde{P}_i(t)(\cdot|s, a)$ is a valid transition distribution over $S_{k(s) + 1}$ for all state-action pairs. Then, we prove that $\tilde{P}_i(t) \in \mathcal{P}_i$ by

$$\left| \tilde{P}_i(t)(s'|s, a) - \bar{P}_i(t)(s'|s, a) \right| = \frac{1}{t} \cdot \left| \bar{P}_i(t)(s'|s, a) - \frac{1}{|S_{k(s) + 1}|} \right| \leq \frac{1}{t} \leq \frac{14 \ln \left( \frac{T|S||A|}{\delta} \right)}{3 \max \{m_{i(t)}(s, a), 1\}}$$

where the last inequality follows from the fact that $m_{i(t)}(s, a) \leq t$. 

35
Then, for any state \( s \neq s_0 \), we have by the definition of occupancy measures
\[
q^{\hat{P}_{(i)}\pi_t}(s) = \sum_{s' \in S_{k(i)} - 1} \sum_{a' \in A} q^{\hat{P}_{(i)}\pi_t}(s', a') \cdot \hat{P}_{(i)}(s|s', a') \\
\geq \sum_{s' \in S_{k(i)} - 1} \sum_{a' \in A} q^{\hat{P}_{(i)}\pi_t}(s', a') \cdot \frac{1}{|S_{k(i)}|} t \\
= \frac{1}{|S_{k(i)}|} t \geq \frac{1}{|S||t|}
\]
Clearly, for \( s_0 \) it holds that \( q^{\hat{P}_{(i)}\pi_t}(s_0) = 1 \geq 1/|S||t| \), which finishes the proof.

\[\square\]

**Corollary C.1.3.** Algorithm 1 ensures that, the adjusted loss \( \hat{\ell}_i \) defined in Eq. (11) for bandit-feedback is bounded as:
\[
|\hat{\ell}_i(s, a)| \leq L + \frac{I_i(s, a)}{q_i(s, a)} \cdot |S||t|
\]
Also, we have
\[
E \left[ \frac{I_i(s, a)}{q_i(s, a)} \cdot A_i(t) \right] = E \left[ \frac{I_i(s, a)}{q_i(s, a)} \cdot A^c_i(t) \right] = 1.
\]

\[\text{Proof.}\] By Lemma C.1.2, we have
\[
|\hat{\ell}_i(s, a)| \leq \frac{I_i(s, a)}{u_i(s) \cdot \pi_i(a|s)} + L \leq \frac{I_i(s, a)}{q_i(s, a) \cdot \pi_i(a|s)} \cdot |S||t| + L = L + \frac{I_i(s, a)}{q_i(s, a)} \cdot |S||t|
\]
where the first inequality follows from \( B_i(s, a) \leq 1 \) and \( \ell_i(s, a) \leq 1 \), and the second inequality uses Lemma C.1.2 and the fact \( q_i(s) \leq 1 \).

For the second statement, we have
\[
E \left[ \frac{I_i(s, a)}{q_i(s, a)} \cdot A_i(t) \right] = E \left[ \frac{I_i(s, a)}{q_i(s, a)} \cdot A^c_i(t) \right] = 1,
\]
By the same arguments we can prove \( E \left[ \frac{I_i(s, a)}{q_i(s, a)} \cdot A^c_i(t) \right] = 1 \) as well.

\[\square\]

**Lemma C.1.4.** Algorithm 1 ensures that, the expected adjusted loss \( \bar{\ell}_i \) defined in Eq. (46) is bounded as:
\[
|\bar{\ell}_i(s, a)| \leq L + |S| \cdot t \leq 2|S| \cdot t, \quad \forall (s, a) \in S \times A, t \in [T].
\]

\[\text{Proof.}\] By Eq. (46), we know that
\[
|\bar{\ell}_i(s, a)| = \frac{q_i(s, a)}{u_i(s)} \cdot \ell_i(s, a) - L \cdot B_i(t)(s, a) \leq \frac{q_i(s)}{u_i(s)} + L \leq L + |S| \cdot t
\]
where the last inequality follows from Lemma C.1.2. Combining with the fact \(|S| \geq L\) finishes the proof.

\[\square\]

**Corollary C.1.5.** Algorithm 1 ensures that, the estimated state-action value functions defined in Eq. (47) are bounded as:
\[
|\bar{Q}_i^G(s, a)| \leq 2L|S||t|, \quad \forall (s, a) \in S \times A, t \in [T].
\]

\[\text{Proof.}\] This is directly by Lemma C.1.4 and the definition of \( \bar{Q}_i^G(s, a) \).

\[\square\]

Next, we analyze the estimated regret in each epoch. Reloading the notation from the full-information setting, we define
\[
\text{EstReg}_i(\pi) = E \left[ \sum_{t=i}^{\ell_i+1} \left( q^{\hat{P}_{(i)}\pi_t} - q^{\hat{P}_{(i)}\pi_t} \cdot \hat{\ell}_i \right) \right] = E \left[ \sum_{t=i}^{\ell_i+1} \left( \hat{q}_t - q^{\hat{P}_{(i)}\pi_t} \cdot \hat{\ell}_i \right) \right].
\]
Lemma C.1.6. With $\beta = 128L^4$, for any epoch $i$, Algorithm 1 ensures

$$
\text{EstReg}_i(\pi) \leq \mathcal{O} \left( \mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1}-1} \eta_t \cdot \left( \sqrt{L|S||A|} + L^2 \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s,a) \cdot B_t(t)(s,a)^2 \right) \right] \right) \tag{50}
$$

for any policy $\pi$, and simultaneously

$$
\text{EstReg}_i(\pi) \leq \mathcal{O} \left( \mathbb{E} \left[ \sqrt{L|S|} \sum_{t=t_i}^{t_{i+1}-1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \tilde{q}_t(s,a) \right] \right) + \mathcal{O} \left( L^2 \mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1}-1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{\tilde{q}_t(s,a)} \right] \right) + \mathcal{O} \left( L^4|A| \cdot \mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1}-1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \tilde{q}_t(s,a) \cdot B_t(t)(s,a)^2 \right] \right) \tag{51}
$$

for any deterministic policy $\pi : S \rightarrow A$.

Proof. The proof is largely based on that of Theorem A.3.1, but with some careful adjustments based on whether $A_t$ holds or not. Let $q = q^\hat{P}, \pi$ be the occupancy measure we want to compete against. When $A_t$ does not hold, we first derive the following naive bound on $\sum_{t=t_i}^{t_{i+1}-1} \left\langle \tilde{q}_t - q, \hat{\ell}_t \right\rangle$:

$$
\sum_{t=t_i}^{t_{i+1}-1} \left\langle \tilde{q}_t - q, \hat{\ell}_t \right\rangle \leq \sum_{t=t_i}^{t_{i+1}-1} \sum_{s \neq s_L} \sum_{a \in A} (\tilde{q}_t(s,a) + q(s,a)) \cdot \left\langle \hat{\ell}_t(s,a) \right\rangle 
$$

$$
\leq \sum_{t=t_i}^{t_{i+1}-1} \sum_{s \neq s_L} \sum_{a \in A} (\tilde{q}_t(s,a) + q(s,a)) \cdot \left( L + \frac{1}{\eta_t(s,a)} \cdot L_t(s,a) \cdot |S| \cdot t \right) 
$$

Therefore, we have the conditional expectation $\mathbb{E} \left[ \sum_{t=t_i}^{t_{i+1}-1} \left\langle \tilde{q}_t - q, \hat{\ell}_t \right\rangle \middle| A_t^i \right]$ bounded by

$$
\mathbb{E} \left[ 2L^2 \cdot (t_{i+1} - t_i) + |S| \cdot t_{i+1} \sum_{t=t_i}^{t_{i+1}-1} \sum_{s \neq s_L} \sum_{a \in A} (\tilde{q}_t(s,a) + q(s,a)) \cdot \frac{L_t(s,a)}{q_t(s,a)} \middle| A_t^i \right] 
$$

$$
\leq \mathbb{E} \left[ (2L^2 + 2L|S|T) \cdot (t_{i+1} - t_i) \middle| A_t^i \right] \tag{Corollary C.1.3}
$$

Next, we condition on event $A_t$. In this case, by the same argument as [Jin and Luo, 2020, Lemma 5] and also our loss-shifting technique, Algorithm 1 with $\beta = 128L^4$ ensures that $\sum_{t=t_i}^{t_{i+1}-1} \left\langle \tilde{q}_t - q, \hat{\ell}_t \right\rangle$ is bounded by

$$
\mathcal{O} \left( L^4|S||A| \log T \right) + \sum_{t=t_i+1}^{t_{i+1}-1} \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (\phi_H(q) - \phi_H(\tilde{q}_t)) 
$$

$$
+ 8 \sum_{t=t_i}^{t_{i+1}-1} \eta_t \min \left\{ \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s,a)^{3/2} \left( \tilde{Q}_t(s,a) - \hat{V}_t(s) \right)^2, \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}_t(s,a)^{3/2} \hat{\ell}_t(s,a)^2 \right\} \tag{52}
$$
where $\phi_H(q) = -\sum_{s \neq s_L} \sum_{a \in A} \sqrt{q(s, a)}$, and $\hat{Q}_t$ and $\hat{V}_t$ are state-action and state value functions associated with the loss estimator $\hat{\ell}_t$ and the empirical transition $\hat{P}(t)$:

$$\hat{Q}_t(s, a) = \hat{\ell}_t(s, a) + \sum_{s' \in S_{\hat{P}(t)}+1} \hat{P}(t)(s'|s, a)\hat{V}_t(s'), \quad \hat{V}_t(s) = \sum_{a \in A} \pi_t(a|s)\hat{Q}_t(s, a).$$

Below, we discuss how to proceed from here to prove Eq. (50) and Eq. (51) respectively.

**Proving Eq. (50)** In this case, we take the second argument of the min operator from Eq. (52) and bound $\phi_H(q) - \phi_H(\hat{q}_t) \leq \sum_{s \neq s_L} \sum_{a \in A} \sqrt{q_t(s, a)}$ trivially by $\sqrt{L|S||A|}$ using Cauchy-Schwarz inequality, leading to

$$\sum_{t=i}^{t_{i+1} - 1} \langle \hat{q}_t - q_t, \hat{\ell}_t \rangle$$

$$\leq O \left( L|S||A| \log T + \sqrt{L|S||A|} \cdot \sum_{t=i}^{t_{i+1} - 1} \eta_t + 8 \sum_{t=i}^{t_{i+1} - 1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a)^{3/2} \hat{\ell}_t(s, a)^2 \right)$$

$$\leq O \left( L|S||A| \log T + 2\sqrt{L|S||A|} \cdot \sum_{t=i}^{t_{i+1} - 1} \eta_t + 16 \sum_{t=i}^{t_{i+1} - 1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a)^{3/2} \cdot I_t(s, a) \cdot u_t(s, a)^2 \right)$$

$$+ 16L^2 \sum_{t=i}^{t_{i+1} - 1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a)^{3/2} \cdot B_t(s, a)^2$$

where the second step follows from the definition of $\hat{\ell}_t$ in Eq. (11) and the last step follows from the fact $\hat{q}_t(s, a) \leq u_t(s, a)$ and $\hat{q}_t(s, a) \leq u_t(s, a)$ since $\hat{P}_t, P_t \in P_t$ according to event $A_t$.

Therefore, by Lemma D.3.6 we have for any policy $\pi$ that,

$$\mathbb{E} \left[ \text{EstReg}_1(\pi) \right] \leq \mathbb{E} \left[ 2\sqrt{L|S||A|} \cdot \sum_{t=i}^{t_{i+1} - 1} \eta_t + 16 \sum_{t=i}^{t_{i+1} - 1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} \sqrt{\hat{q}_t(s, a)} \cdot I_t(s, a) \right]$$

$$+ \mathbb{E} \left[ 16L^2 \sum_{t=i}^{t_{i+1} - 1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a) \cdot B_t(s, a)^2 \right]$$

$$+ O \left( L^4|S||A| \log T + \delta \cdot \mathbb{E} \left[ L|S|T(t_{i+1} - t_i) \right] \right)$$

$$\leq O \left( \mathbb{E} \left[ \sqrt{L|S||A|} \cdot \sum_{t=i}^{t_{i+1} - 1} \eta_t \right] + \mathbb{E} \left[ L^2 \sum_{t=i}^{t_{i+1} - 1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s, a) \cdot B_t(s, a)^2 \right] \right)$$

$$+ O \left( L^4|S||A| \log T + \delta \cdot \mathbb{E} \left[ L|S|T(t_{i+1} - t_i) \right] \right)$$

where the second step takes the conditional expectation of $I_t(s, a)$ and applies the Cauchy-Schwarz inequality to get $\sum_{s \neq s_L} \sum_{a \in A} \sqrt{\hat{q}_t(s, a)} \leq \sqrt{L|S||A|}$. This finishes the proof of Eq. (50).

**Proving Eq. (51)** In this case, recall that $\pi$ is a deterministic policy, so that

$$\phi_H(q) - \phi_H(\hat{q}_t) = \sum_{s \neq s_L} \sqrt{\hat{q}_t(s)} \left( \sum_{a \in A} \sqrt{\pi_t(a|s)} - 1 \right) + \sum_{s \neq s_L} \left( \sqrt{\hat{q}_t(s)} - \sqrt{q(s)} \right).$$
Therefore, taking the first argument of the min operator from Eq. (52) and using \( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \leq \eta_t \) again, we arrive at

\[
\sum_{t=t_i}^{t_i+1} \left( \hat{q}_t - q_t, \hat{\ell}_t \right) \leq \sqrt{L|S|} \sum_{t=t_i}^{t_i+1} \eta_t \cdot \sqrt{\sum_{s \neq s_L, a \neq \pi(s)} \sqrt{q_t(s,a)}} + \sum_{t=t_i}^{t_i+1} \eta_t \cdot \sum_{s \neq s_L, a \neq \pi(s)} \sqrt{q_t(s,a)} + 8 \sum_{t=t_i}^{t_i+1} \eta_t \cdot \sum_{s \neq s_L, a \in A} \sqrt{q_t(s,a)}^{3/2} \left( \hat{Q}_t(s,a) - \hat{V}_t(s) \right)^2 + O \left( L^4 |S||A| \log T \right).
\]

Finally, we apply Lemma C.1.7 to bound the term \( \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s,a)^{3/2} \left( \hat{Q}_t(s,a) - \hat{V}_t(s) \right)^2 \), and use Lemma D.3.6 again to take expectation and arrive at Eq. (51) (with the help of Eq. (54)).

**Lemma C.1.7.** Under event \( \mathcal{A} \), we have for any \( t \),

\[
\sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s,a)^{3/2} \left( \hat{Q}_t(s,a) - \hat{V}_t(s) \right)^2 \leq 4L^4 |A| \sum_{s' \neq s_L} \sum_{a' \in A} \hat{q}_t(s',a') \cdot B_{\mathcal{I}(t)}(s',a')^2 + \sum_{s \neq s_L} \sum_{a \in A} \sqrt{q_t(s,a)} \cdot (O_t(s,a) + W_t(s,a))
\]

where

\[
O_t(s,a) = 4L \cdot (1 - \pi_t(a|s)) \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s',a'|s,a) \frac{I_t(s',a')}{{q}_t(s',a')},
\]

\[
W_t(s,a) = 4L \cdot \sum_{b \neq a} \pi_t(b|s) \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s',a'|s,b) \frac{I_t(s',a')}{{q}_t(s',a')}.
\]

and \( \hat{q}_t(s',a'|s,a) \) is the probability of visiting \( (s', a') \) starting from \( (s, a) \) under \( \pi_t \) and \( \bar{P}(t) \). Moreover, we have

\[
\mathbb{E}_t \left[ \sum_{s \neq s_L} \sum_{a \in A} \sqrt{q_t(s,a)} \cdot (O_t(s,a) + W_t(s,a)) \right] \leq 16L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{\hat{q}_t(s,a)},
\]

for any mapping \( \pi : S \to A \).

**Proof.** First, \( \left( \hat{Q}_t(s,a) - \hat{V}_t(s) \right)^2 \) is bounded by

\[
\left( \hat{Q}_t(s,a) - \hat{V}_t(s) \right)^2 = \left( 1 - \pi_t(a|s) \right) \hat{Q}_t(s,a) - \left( \sum_{b \neq a} \pi_t(b|s) \hat{Q}_t(s,b) \right) \right)^2 \leq 2 \left( 1 - \pi_t(a|s) \right)^2 \hat{Q}_t(s,a)^2 + 2 \left( \sum_{b \neq a} \pi_t(b|s) \hat{Q}_t(s,b) \right)^2.
\]
Following the same idea of Lemma A.1.3, the first term can be bounded as
\[(1 - \pi_t(a|s))^2 \overline{Q}_t(s, a)^2\]
\[= (1 - \pi_t(a|s))^2 \left( \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a) \hat{\ell}_t(s', a') \right)^2\]
\[\leq 2 (1 - \pi_t(a|s))^2 \left( \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a) \frac{I_t(s', a')}{u_t(s', a')} \cdot \hat{\ell}_t(s', a') \right)^2\]
\[+ 2 (1 - \pi_t(a|s))^2 \left( \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a) \cdot L \cdot B_{i(t)}(s', a') \right)^2\]
\[\leq 2L \cdot (1 - \pi_t(a|s))^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a)^2 \cdot \frac{I_t(s', a')}{u_t(s', a')}^2\]
\[+ 2L^3 (1 - \pi_t(a|s))^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a) \cdot B_{i(t)}(s', a')^2 \] (55)

where the equality follows from the definition of $\hat{Q}_t$; the first inequality uses the fact $(x+y)^2 \leq 2(x^2 + y^2)$; the second inequality applies the Cauchy-Schwarz inequality with the facts $I_t(s, a)I_t(s', a') = 0$ for $(s, a) \neq (s', a')$ and $\sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a) \leq L$.

By the same arguments, the second term is bounded as
\[
\left( \sum_{b \neq a} \pi_t(b|s) \hat{Q}_t(s, b) \right)^2
\]
\[= \left( \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s) \hat{q}_t(s', a'|s, b) \hat{\ell}_t(s, a) \right)^2\]
\[\leq 2L \cdot \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s', a'|s, b) \right)^2 \cdot \frac{I_t(s', a')}{u_t(s', a')}^2\]
\[+ 2L^3 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s', a'|s, b) \right) \cdot B_{i(t)}(s', a')^2, \] (56)

where in the last step we use $\sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \pi_t(b|s) \cdot \hat{q}_t(s', a'|s, b) \leq L$ (after applying Cauchy-Schwarz).

Combining Eq. (55) and Eq. (56), we show that $\hat{q}_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2$ can be bounded as
\[
\hat{q}_t(s, a) \left( \hat{Q}_t(s, a) - \hat{V}_t(s) \right)^2
\]
\[\leq 4L \cdot \hat{q}_t(s, a) \left( 1 - \pi_t(a|s) \right)^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a)^2 \cdot \frac{I_t(s', a')}{u_t(s', a')}^2\]
\[+ 4L \cdot \hat{q}_t(s, a) \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s', a'|s, b) \right)^2 \cdot \frac{I_t(s', a')}{u_t(s', a')}^2\]
\[+ 4L^3 \hat{q}_t(s, a) \left( 1 - \pi_t(a|s) \right)^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s', a'|s, a) \cdot B_{i(t)}(s', a')^2\]
Moreover, we have the summation of the first two terms bounded as

\[
4L \cdot \bar{q}_t(s,a) \left(1 - \pi_t(a|s)\right)^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left(\sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s',a'|s,b)\right)^2 \frac{\|s',a'\|}{u_t(s',a')^2}
\]

\[
+ 4L \cdot \bar{q}_t(s,a) \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left(\sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s',a'|s,b)\right)^2 \frac{\|s',a'\|}{u_t(s',a')^2} \cdot \frac{\|s',a'\|}{\hat{q}_t(s',a')} \cdot \frac{\sum_{b \neq a} \bar{q}_t(s,b) \cdot \hat{q}_t(s',a'|s,b)}{u_t(s',a')} \cdot \left(\sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s',a'|s,b)\right) \frac{\|s',a'\|}{\hat{q}_t(s',a')}
\]

\[
\leq 4L \cdot \left(1 - \pi_t(a|s)\right)^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \bar{q}_t(s,a) \bar{q}_t(s',a'|s,a) \frac{\|s',a'\|}{u_t(s',a')} \cdot \frac{\|s',a'\|}{\hat{q}_t(s',a')} \cdot \frac{\sum_{b \neq a} \bar{q}_t(s,b) \cdot \hat{q}_t(s',a'|s,b)}{u_t(s',a')} \cdot \left(\sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s',a'|s,b)\right) \frac{\|s',a'\|}{\hat{q}_t(s',a')}
\]

\[
\leq \text{O}_t(s,a) + W_t(s,a)
\]

where we use \( \bar{q}_t(s',a') \leq u_t(s',a') \) due to event \( \mathcal{A}_i \) in the first step and \( \sum_{a \in A} \bar{q}_t(s,a) \bar{q}_t(s',a'|s,a) \leq \bar{q}_t(s',a') \leq u_t(s',a') \) in the second step to bound the fractions by 1.

On the other hand, the summation of the other two terms is bounded as

\[
4L^3 \bar{q}_t(s,a) \left(1 - \pi_t(a|s)\right)^2 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s',a'|s,a) \cdot B_{i(t)}(s',a')^2
\]

\[
+ 4L^3 \bar{q}_t(s,a) \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left(\sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s',a'|s,b)\right)^2 \cdot B_{i(t)}(s',a')^2
\]

\[
\leq 4L^3 \bar{q}_t(s) \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \left(\sum_{b \neq a} \hat{q}_t(s',a'|s,a) \pi_t(a|s) + \sum_{b \neq a} \pi_t(b|s) \cdot \hat{q}_t(s',a'|s,b)\right) \cdot B_{i(t)}(s',a')^2
\]

\[
= 4L^3 \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s',a'|s) \hat{q}_t(s) \cdot B_{i(t)}(s',a')^2.
\]

Note that, taking the summation of the last bound over all state-action pairs yields

\[
4L^3 \sum_{s \neq s_L} \sum_{a \neq a_L} \sum_{k=k(s)}^{L-1} \sum_{s' \in S_k} \sum_{a' \in A} \hat{q}_t(s',a'|s) \hat{q}_t(s) \cdot B_{i(t)}(s',a')^2
\]

\[
= 4L^3 |A| \sum_{s' \neq s_L} \sum_{a' \in A} \left(\sum_{k(s')}^{L-1} \sum_{s \in S_k} \hat{q}_t(s',a'|s) \hat{q}_t(s)\right) \cdot B_{i(t)}(s',a')^2
\]

\[
\leq 4L^4 |A| \sum_{s' \neq s_L} \sum_{a' \in A} \hat{q}_t(s',a') \cdot B_{i(t)}(s',a')^2.
\]

Therefore, combining everything, we have shown:

\[
\sum_{s \neq s_L, a \neq \pi(s)} \hat{q}_t(s,a)^3 \left(\hat{Q}_t(s,a) - \hat{V}_t(s)\right)^2 \leq 4L^4 |A| \sum_{s' \neq s_L} \sum_{a' \in A} \hat{q}_t(s',a') \cdot B_{i(t)}(s',a')^2 + \sum_{s \neq s_L, a \neq \pi(s)} \sqrt{\hat{q}_t(s,a) \cdot (O_t(s,a) + W_t(s,a))},
\]
proving the first statement of the lemma.

To prove the second statement, we first show

\[
E_t [O_t(s, a) + W_t(s, a)] = 4L (1 - \pi_t(a|s)) \cdot \sum_{k=k(s)} \sum_{s' \in S_k} \sum_{a' \in A} \tilde{q}(s', a'|s, a) \\
+ 4L \cdot \sum_{k=k(s)} \sum_{s' \in S_k} \sum_{a' \in A} \left( \sum_{b \neq a} \pi_t(b|s) \cdot \tilde{q}(s', a'|s, b) \right) \\
= 4L (1 - \pi_t(a|s)) \sum_{k=k(s)} 1 + 4L \cdot \sum_{k=k(s)} (1 - \pi_t(a|s)) \\
\leq 8L^2 (1 - \pi_t(a|s)),
\]

and therefore

\[
E_t \left[ \sum_{s \neq s_L} \sum_{a \in A} \sqrt{\tilde{q}(s, a) \cdot (O_t(s, a) + W_t(s, a))} \right] \\
\leq 8L^2 \sum_{s \neq s_L} \sum_{a \in A} \sqrt{\tilde{q}(s, a)(1 - \pi_t(a|s))} \\
\leq 8L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{\tilde{q}(s, a)} + 8L^2 \sum_{s \neq s_L} \sqrt{\tilde{q}(s)} (1 - \pi_t(\pi(s)|s)) \\
\leq 16L^2 \sum_{s \neq s_L} \sum_{a \neq \pi(s)} \sqrt{\tilde{q}(s, a)},
\]

which proves Eq. (54).

Note that both Eq. (50) and Eq. (51) contain a term related to \( \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}(s, a) \cdot B_{i(t)}(s, a)^2 \).

Below, we show that when summed over \( t \), this is only logarithmic in \( T \).

**Lemma C.18.** Algorithm 1 ensures the following:

\[
E \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}(s, a) \cdot B_{i(t)}(s, a)^2 \right] = O \left( L^2 |S|^3 |A|^2 \ln^2 t + |S||A|T \cdot \delta \right). \tag{57}
\]

**Proof.** By Lemma D.3.2, we know that

\[
B_{i}(s, a) \leq \left( 2 \left[ \frac{|S_k(s)| + 1}{\ln t} \max \{m_i(s, a), 1\} \right] + \frac{14 |S_k(s)| + 1}{3 \max \{m_i(s, a), 1\}} \right)^2 \\
\leq O \left( \left[ \frac{|S_k(s)| + 1}{\ln t} \max \{m_i(s, a), 1\} \right]^2 \right).
\]

Then, we have

\[
E \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \tilde{q}(s, a) \cdot B_{i(t)}(s, a)^2 \right] \\
= E \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} (\tilde{q}(s, a) - q_t(s, a)) \cdot B_{i(t)}(s, a)^2 \right] + E \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) \cdot B_{i(t)}(s, a)^2 \right] \\
\leq E \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} r_t(s, a) \right]
\]

42
\[ + \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_{t}(s, a) \cdot B_{i(t)}(s, a) \cdot P(w|u, v) \ln \left( \frac{T|S||A|}{\delta} \right) \max \left\{ m_i(t)(u, v), 1 \right\} q_{t}(s, a|w) \right] \]

\[ + \mathcal{O} \left( \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_{t}(s, a) \cdot \left( |S_{k(s)+1}| \ln \frac{t}{\max \{ m_i(s, a), 1 \}} + \frac{|S_{k(s)+1}|^2 \ln^2 t}{\max \{ m_i(s, a), 1 \}^2} \right) \right] \right) \]

\[ \leq \mathcal{O} \left( \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} r_t(s, a) \right] \right) \]

\[ + \mathcal{O} \left( \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_{t}(s, a) \cdot \left( |S_{k(s)+1}| \ln \frac{t}{\max \{ m_i(s, a), 1 \}} + \frac{|S_{k(s)+1}|^2 \ln^2 t}{\max \{ m_i(s, a), 1 \}^2} \right) \right] \right) \]

where the first inequality uses Lemma D.3.10 and \( B_{i(t)}(s, a) \in [0, 1] \), and the last inequality follows from the observation that, the second term in the previous line is bounded by \( \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} r_t(s, a) \) according to the definition of residual terms in Definition D.3.9.

Finally, applying Lemma D.3.10 and Lemma D.3.8, we have

\[ \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_{t}(s, a) \cdot B_{i(t)}(s, a)^2 \right] \]

\[ = \mathcal{O} \left( L^2 |S|^3 |A|^2 \ln^2 t + |S||A| T \cdot \delta \right) + \mathcal{O} \left( \sum_{k=0}^{L-1} (|S_{k+1}| |S_k| |A| \ln T \ln t + |S_{k(s)+1}|^2 |S_k| |A| \ln^2 t) \right) \]

\[ = \mathcal{O} \left( L^2 |S|^3 |A|^2 \ln^2 t + |S||A| T \cdot \delta \right), \]

which completes the proof. \( \square \)

Finally, we provide a lemma regarding the learning rates.

**Lemma C.1.9** (Learning Rates). According to the design of the learning rate \( \eta_t = \frac{1}{\sqrt{t-t_{i(t)}+1}} \), the following inequalities hold:

\[ \sum_{t=1}^{T} \eta_t^2 \leq \mathcal{O} \left( |S||A| \log^2 T \right), \] (58)

\[ \sum_{t=1}^{T} \eta_t \leq \mathcal{O} \left( \sqrt{|S||A|T \log T} \right) \] (59)

**Proof.** By direct calculation, we have

\[ \sum_{t=t_i}^{t_{i+1}-1} \eta_t^2 = \sum_{n=1}^{t_{i+1}-t_i} \frac{1}{n} \leq 2 \int_1^{t_{i+1}-t_i+1} \frac{1}{x} dx = 2 \log (t_{i+1} - t_i + 1) \leq \mathcal{O} \left( \log T \right). \]

Combining the inequality with the fact that the total number of epochs \( N \) is at most \( 4|S||A| (\log T + 1) \) (Lemma D.3.12) finishes the proof of Eq. (58). Following the similar idea, we have

\[ \sum_{t=t_i}^{t_{i+1}-1} \eta_t = \sum_{n=1}^{t_{i+1}-t_i} \frac{1}{\sqrt{n}} \leq \int_0^{t_{i+1}-t_i} \frac{1}{\sqrt{x}} dx \leq 2 \sqrt{t_{i+1} - t_i}. \]

Taking the summation over \( N \) epochs and applying the Cauchy-Schwarz inequality yields Eq. (59). \( \square \)
C.2 Proof for the Adversarial World (Proposition C.1)

Recall the regret decomposition in Eq. (49):

\[
\mathbb{E} \left[ \sum_{t=1}^{T} V_t^\pi(s_0) - \tilde{V}_t^\pi(s_0) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \tilde{V}_t^\pi(s_0) - \tilde{V}_t^\pi(s_0) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \tilde{V}_t^\pi(s_0) - V_t^\pi(s_0) \right].
\]

We bound each of them separately below.

**ERR**

Similarly to the proof for the full-information feedback setting, we have

\[
\text{ERR}_1 = \sum_{t=1}^{T} \langle q_t, \ell_t \rangle - \langle \tilde{q}_t, \ell_t \rangle
\]

\[
= \sum_{t=1}^{T} \left( \sum_{s \neq s_L} \sum_{a \in A} \ell_t(s,a) \tilde{q}_t(s,a) \frac{u_t(s,a)}{u_t(s,a)} \cdot (u_t(s,a) - q_t(s,a)) + \sum_{s \neq s_L} (q_t - \tilde{q}_t, \ell_t) + L \cdot \sum_{s \neq s_L} \langle \tilde{q}_t, B_i(t) \rangle \right)
\]

where the last two terms have been shown to be at most \( \tilde{O} \left( L|S|\sqrt{|A|T} + L^3|S|^3|A| \right) \) according to the analysis of \( \text{ERR}_1 \) in Appendix B.2 (see Eq. (36), Eq. (37) and Eq. (38)).

Then, we bound the first term as

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \ell_t(s,a) \tilde{q}_t(s,a) \frac{u_t(s,a)}{u_t(s,a)} \cdot (u_t(s,a) - q_t(s,a)) \right] \leq \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} |u_t(s,a) - q_t(s,a)| \right]
\]

\[
\leq \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} |u_t(s,a) - q_t(s,a)| \right] \quad (\tilde{q}_t(s,a) \leq u_t(s,a))
\]

\[
\leq \mathbb{E} \left[ 4 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} r_t(s,a) + 16 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \sum_{k(s)-1}^{k(s)} \sum_{(u,v,w) \in T_k} q_t(u,v) \sqrt{\frac{P(w|u,v) \ln \ell}{\max \{m_i(t)(u,v),1\}}} q_t(s,a|w) \right]
\]

\[
= \mathcal{O} \left( L^2|S|^3|A|^2 \ln^2 \ell + |S||A|T \cdot \delta \right) + 4L \cdot \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(u,v) \sqrt{\frac{|S_{k(u)+1}| \ln \ell}{\max \{m_i(t)(u,v),1\}}} \right]
\]

\[
\leq \mathcal{O} \left( L^2|S|^3|A|^2 \ln^2 \ell + |S||A|T \cdot \delta + L \cdot \sum_{k=0}^{L-1} |S_k| \cdot |S_{k+1}| |A|T \ln \ell \right)
\]

\[
= \mathcal{O} \left( L|S|\sqrt{|A|T} + L^3|S|^3|A|^2 \ln^2 \ell + |S||A|T \cdot \delta \right).
\]

Combining the bounds together, we have \( \mathbb{E} [\text{ERR}_1] \) bounded by:

\[
\mathbb{E} [\text{ERR}_1] = \tilde{O} \left( L|S|\sqrt{|A|T} + L^3|S|^3|A|^2 \right).
\]

**ERR**

Following the same idea of bounding \( \text{ERR}_2 \), by Lemma C.1.1 and Lemma D.3.5, we have the expectation of \( \text{ERR}_2 \) bounded as

\[
\mathbb{E} [\text{ERR}_2] \leq \delta \cdot 3L|S|T^2 + 0 = \mathcal{O} \left( L|S|T^2 \cdot \delta \right) = \mathcal{O}(1).
\]

**EST**

According to Eq. (50) of Lemma C.1.6, we have

\[
\text{EstReg}(\hat{\pi}) = \mathbb{E} \left[ \sum_{t=1}^{T} \langle \tilde{q}_t - q_t^{p_t(t),\hat{\pi}}, \ell_t \rangle \right] = \mathbb{E} \left[ \sum_{i=1}^{N} \text{EstReg}_i(\hat{\pi}) \right]
\]
\begin{align*}
\leq & \mathcal{O} \left( \mathbb{E} \left[ \sum_{t=1}^{N} \sum_{b \neq b_L} \eta_t \sqrt{|S||A|} + \mathbb{E} \left[ L^4 \cdot \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a) \cdot B_{i(t)}(s, a)^2 \right] \right] \right) \\
& + \mathcal{O} \left( L^4 |S|^3 |A|^2 \ln^2 T + \delta L |S| T^2 \right) \\
\leq & \tilde{\mathcal{O}} \left( \mathbb{E} \left[ \sum_{t=1}^{T} \eta_t \cdot \sqrt{|S||A|} \right] + L^4 |S|^3 |A|^2 \ln^2 t \right) \quad \text{(Lemma C.1.8)} \\
\leq & \tilde{\mathcal{O}} \left( |S||A| \sqrt{LT} + L^4 |S|^3 |A|^2 \right). \quad \text{(Eq. (59))}
\end{align*}

Finally, we combine the bounds of \( \text{ERR}_1 \), \( \text{ERR}_2 \) and \( \text{ESTREG} \) as:

\[
\text{Reg}_T(\pi) = \tilde{\mathcal{O}} \left( |S| \sqrt{|A| T} + |S||A| \sqrt{LT} + L^4 |S|^3 |A|^2 \right),
\]

finishing the proof.

### C.3 Proof for the Stochastic World (Proposition C.2)

Similarly to the proof of Proposition B.2, we decompose \( \text{ERR}_1 \) and \( \text{ERR}_2 \) jointly into four terms \( \text{ERR}_{\text{SUB}}, \text{ERR}_{\text{OPT}}, \text{OCCDIFF} \) and \( \text{BIAS} \):

\[
\text{ERR}_1 + \text{ERR}_2 = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \tilde{E}_t^{\pi^*}(s, a) \quad \text{(ERRSUB)}
\]

\[
+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \tilde{E}_t^{\pi^*}(s, a) \quad \text{(ERROPT)}
\]

\[
+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} (q_t(s, a) - \tilde{q}_t(s, a)) \left( \tilde{Q}_t^{\pi^*}(s, a) - \tilde{V}_t^{\pi^*}(s) \right) \quad \text{(OCCDIFF)}
\]

\[
+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t^*(s, a) \left( \tilde{V}_t^{\pi^*}(s) - V_t^{\pi^*}(s) \right) \quad \text{(BIAS)}
\]

where \( \tilde{E}_t^{\pi^*} \) is defined as

\[
\tilde{E}_t^{\pi^*}(s, a) = \ell_t(s, a) + \sum_{s' \in S_k(s)+1} P(s'|s, a) \tilde{V}_t^{\pi^*}(s') - \tilde{Q}_t^{\pi^*}(s, a).
\]

By the exact same reasoning as in the full-information setting (Appendix B.3), we have \( \mathbb{E}[\text{OCCDIFF}] = \mathcal{O} \left( L^4 |S|^3 |A|^2 \ln^2 t + \mathbb{E} [\mathbb{G}_3(L^4 \ln t)] \right) \) and \( \mathbb{E}[\text{BIAS}] = \mathcal{O}(1) \), but the first two terms \( \text{ERR}_{\text{SUB}} \) and \( \text{ERR}_{\text{OPT}} \) are slightly different. To see this, note that under event \( \mathcal{A} \), we have

\[
\tilde{E}_t^{\pi^*}(s, a) = \ell_t(s, a) - \tilde{\ell}_t(s, a) + \sum_{s' \in S_k(s)+1} \left( P(s'|s, a) - \tilde{P}_{i(t)}(s'|s, a) \right) \tilde{V}_t^{\pi^*}(s')
\]

\[
= \ell_t(s, a) \left( 1 - \frac{q_t(s, a)}{u_t(s, a)} \right) + L \cdot B_{i(t)}(s, a) + \sum_{s' \in S_k(s)+1} \left( P(s'|s, a) - \tilde{P}_{i(t)}(s'|s, a) \right) \tilde{V}_t^{\pi^*}(s')
\]

\[
\leq \frac{u_t(s, a) - q_t(s, a)}{q_t(s, a)} + 2L^2 \cdot B_{i(t)}(s, a)
\]

where the last line applies the definition of event \( \mathcal{A} \) and the fact \( q_t(s, a) \leq u_t(s, a) \) given this event. Importantly, the second term has been studied and bounded in the proof of Proposition B.2 already, so we only need to focus on the first term. Before doing so, note that the range of \( \tilde{E}_t^{\pi^*} \) is \( \mathcal{O} \left( L^4 |S| t \right) \) based on Corollary C.1.5, and thus the range of \( \text{ERR}_{\text{SUB}} \) and \( \text{ERR}_{\text{OPT}} \) is \( \mathcal{O} \left( L^4 |S| T^2 \right) \). Therefore, we only need to add a term \( \mathcal{O} \left( \delta \cdot L^2 |S| T^2 \right) \) to address the event \( \mathcal{A}^c \).
Extra term in ERRSUB According to previous analysis, the extra term in ERRSUB is
\[
\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \cdot \frac{u_t(s, a) - q_t(s, a)}{q_t(s, a)} \leq 4 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} |u_t(s, a) - q_t(s, a)|
\]
\[
\leq 4 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} r_t(s, a) + 16 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_h} q_t(u,v) \sqrt{P(w|u,v) \ln \left( \frac{T|S||A|}{s} \right)} \max \left\{ m_{i(t)}(u,v), 1 \right\} q_t(s,a|w) \tag{Corollary D.3.11}
\]
\[
= 4 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} r_t(s, a) + 16G_3(\ln \ell) \tag{Definition D.2.1}
\]
\[
= 16G_3(\ln \ell) + O\left( L^2S^3A^2 \ln^2 \ell \right). \tag{Lemma D.3.10}
\]

Finally, using Lemma D.3.6 and the bound on ERRSUB for the full-information setting, we have
\[
E[ERRSUB] = O\left( G_3(\ln \ell) + G_1(L^4|S| \ln \ell) + L^2|S|^3|A|^2 \ln^2 \ell \right).
\]

Extra term in ERROPT Similarly, we consider the extra term in ERROPT:
\[
\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} (q_t(s, a) - \hat{q}_t(s, a)) \cdot \frac{u_t(s, a) - \hat{q}_t(s, a)}{\hat{q}_t(s, a)} \leq 4 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} q_t(s, a) - \hat{q}_t(s, a) \cdot r_t(s, a)
\]
\[
+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} q_t(s, a) - \hat{q}_t(s, a) \cdot \left( 16 \sum_{u,v,w} q_t(u,v) \sqrt{P(w|u,v) \ln \left( \frac{T|S||A|}{s} \right)} \max \left\{ m_{i(t)}(u,v), 1 \right\} q_t(s,a|w) \right) \tag{Corollary D.3.11}
\]
\[
= 4 \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} r_t(s, a) + 16G_6(\ln \ell) \tag{Definition D.2.1}
\]
\[
= 16G_6(\ln \ell) + O\left( L^2S^3A^2 \ln^2 \ell \right). \tag{Lemma D.3.10}
\]

Again, considering the term that appears in the full-information setting already, we have
\[
E[ERROPT] = O\left( G_6(\ln \ell) + G_2(L^4|S| \ln \ell) + L^2|S|^3|A|^2 \ln^2 \ell \right).
\]

It remains to bound ESTREG with terms that enjoy self-bounding properties.

Term ESTREG According to Eq. (51) in Lemma C.1.6, taking the summation of all the epochs, we have the following bound for $E[\text{ESTREG}]$:
\[
O\left( E\left[ \sqrt{|S|L} \sum_{t=1}^{N} \sum_{i=t_i}^{t_i+1-1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \hat{q}_t(s,a) \right] + L^2 \cdot E\left[ \sum_{t=1}^{N} \sum_{i=t_i}^{t_i+1-1} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \sqrt{\hat{q}_t(s,a)} \right] \right)
\]
\[
+ O\left( E\left[ L^4|A| \sum_{i=1}^{N} \sum_{t=t_i}^{t_i+1-1} \sum_{s \neq s_L} \sum_{a \in A} \hat{q}_t(s,a) \cdot B_{i(t)}(s,a)^2 \right] \right)
\]
\[
+ O\left( \delta \cdot E\left[ L|S|T \sum_{i=1}^{N} (t_{i+1} - t_i) + L^4|S|^3|A|^2 \ln^2 \ell \right] \right)
\]
\[
= O\left( E\left[ \sqrt{|S|L} \sum_{t=1}^{T} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \hat{q}_t(s,a) \right] \right) + O\left( L^2 \cdot E\left[ \sum_{t=1}^{T} \eta_t \cdot \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \sqrt{\hat{q}_t(s,a)} \right] \right)
\]

46
+ \mathcal{O}(L^6|S|^3|A|^3 \ln^2 \epsilon)

where the last line applies Lemma C.1.8.

Then, for the first term, we have

\[
\mathbb{E} \left[ \sqrt{|S|L} \sum_{t=1}^{T} \eta_t \cdot \sqrt{\sum_{s \neq s_L, a \neq \pi^*(s)} \hat{q}_t(s, a)} \right] \\
\leq \mathbb{E} \left[ \sqrt{|S|L} \cdot \sum_{t=1}^{T} \eta_t^2 \cdot \sum_{s \neq s_L, a \neq \pi^*(s)} \hat{q}_t(s, a) \right] \\
\leq \mathbb{E} \left[ 4L|S|^2|A| \log^2 T \cdot \sum_{t=1}^{T} \sum_{s \neq s_L, a \neq \pi^*(s)} \hat{q}_t(s, a) \right]
\]

where the second line follows from the Cauchy-Schwarz inequality, and the third line applies Eq. (58).

Then, we separate the term into two parts:

\[
\mathbb{E} \left[ \sqrt{4L|S|^2|A| \log^2 T} \cdot \sum_{t=1}^{T} \sum_{s \neq s_L, a \neq \pi^*(s)} q_t(s, a) \right] \\
+ \mathbb{E} \left[ \sqrt{4L|S|^2|A| \log^2 T} \cdot \sum_{t=1}^{T} \sum_{s \neq s_L, a \neq \pi^*(s)} |\hat{q}_t(s, a) - q_t(s, a)| \right] \\
\leq \mathbb{E} \left[ 2 \cdot G_4(L|S|^2|A| \log^2 T) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L, a \neq \pi^*(s)} |\hat{q}_t(s, a) - q_t(s, a)| \right] + 4|S|^2|A|L \log^2 T
\]

where second line follows from the fact \( \sqrt{x+y} \leq x + y \) for \( x, y \geq 0 \). Note that, the second term above can be bounded by \( \mathcal{O}(G_3(\ln \epsilon) + L^2|S|^3|A|^2 \ln^2 \epsilon) \) just as in the full-information setting (see Eq. (44)). Therefore, we have finished bounding the first term:

\[
\mathbb{E} \left[ \sqrt{|S|L} \sum_{t=1}^{T} \eta_t \cdot \sqrt{\sum_{s \neq s_L, a \neq \pi^*(s)} \hat{q}_t(s, a)} \right] \\
= \mathcal{O} \left( \mathbb{E} \left[ G_4(L|S|^2|A| \log^2 T) + G_3(\ln \epsilon) \right] + L^2|S|^3|A|^2 \ln^2 \epsilon \right).
\]

On the other hand, the second term can be bounded similarly:

\[
L^2 \cdot \mathbb{E} \left[ \sum_{t=1}^{T} \eta_t \cdot \sum_{s \neq s_L, a \neq \pi^*(s)} \sqrt{\hat{q}_t(s, a)} \right] \\
\leq L^2 \cdot \mathbb{E} \left[ \sum_{s \neq s_L, a \neq \pi^*(s)} \sum_{t=1}^{T} \eta_t^2 \cdot \sum_{t=1}^{T} \hat{q}_t(s, a) \right] \\
\leq L^2 \sqrt{4|S||A| \log^2 T} \cdot \mathbb{E} \left[ \sum_{s \neq s_L, a \neq \pi^*(s)} \sum_{t=1}^{T} \hat{q}_t(s, a) \right] \\
\leq \mathbb{E} \left[ 2 \cdot G_5(L^4|S||A| \log^2 T) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L, a \neq \pi^*(s)} |\hat{q}_t(s, a) - q_t(s, a)| \right] + L^4|S||A| \log^2 T \\
= \mathcal{O} \left( \mathbb{E} \left[ G_5(L^4|S||A| \log^2 T) + G_3(\ln \epsilon) \right] + L^2|S|^3|A|^2 \ln^2 \epsilon \right).
\]
So we have the final bound on $\mathbb{E}[\text{EstReg}]:$

$$\mathbb{E}[\text{EstReg}] = O\left(\mathbb{E}\left[\mathcal{G}_4\left(L^4|S|^2|A|\log^2 T\right) + \mathcal{G}_5\left(L^4|S||A|\log^2 T\right) + \mathcal{G}_3\left(\ln \iota\right)\right] + L^6|S|^3|A|^3\ln^2 \iota\right)$$

Finally, by combining the bounds of each term, we finally have

$$\text{Reg}_T(\pi^\star) \leq \kappa \left(\mathbb{E}\left[\mathcal{G}_1\left(L^4|S|\ln T\right) + \mathcal{G}_2\left(L^4|S|\ln T\right) + \mathcal{G}_3\left(L^4\ln T\right)\right]\right)$$

$$+ \kappa \left(\mathbb{E}\left[\mathcal{G}_4\left(L|S|^2|A|\log^2 T\right) + \mathcal{G}_5\left(L^4|S||A|\log^2 T\right) + \mathcal{G}_6\left(\ln T\right)\right]\right)$$

$$+ \kappa \left(L^6|S|^3|A|^3\ln^2 \iota\right).$$

When Condition (1) holds, we apply similar self-bounding arguments to obtain a logarithmic regret bound. Specifically, for some universal constant $\kappa > 0$, we have

$$\text{Reg}_T(\pi^\star) \leq \frac{1}{z} \cdot \left(\text{Reg}_T(\pi^\star) + C\right)$$

$$+ \frac{1}{z} \cdot O\left(\sum_{s \neq s^L} \sum_{a \neq \pi^\star(s)} \frac{\kappa^2}{\Delta(s,a)} \cdot \left(L^4|S|\ln T + L^6|S|\ln T + L^4|S||A|\log^2 T\right)\right)$$

$$+ \frac{1}{z} \cdot O\left(\frac{\kappa^2}{\Delta_{\text{min}}} \cdot \left(L^5|S|^2\ln T + L^6|S|^2\ln T + L^3|S|^2|A|\ln T + L|S|^2|A|\log^2 T\right)\right)$$

$$+ \kappa \cdot \left(L^6|S|^3|A|^3\ln^2 \iota\right)$$

$$\leq \frac{1}{z} \cdot \left(\text{Reg}_T(\pi^\star) + C\right) + \kappa \cdot \left(L^6|S|^3|A|^3\ln^2 \iota\right)$$

$$+ \frac{1}{z} \cdot O\left(\sum_{s \neq s^L} \sum_{a \neq \pi^\star(s)} \frac{L^6|S|\ln T + L^4|S||A|\log^2 T}{\Delta(s,a)} + \frac{L^6|S|^2\ln T + L^3|S|^2|A|\log^2 T}{\Delta_{\text{min}}}\right)$$

$$\leq \frac{1}{z} \cdot \left(\text{Reg}_T(\pi^\star) + C\right) + z \cdot \kappa' U + \kappa \cdot V,$$

where $\kappa'$ is a universal constant hidden in the $O(\cdot)$ notation, and $U$ and $V$ are defined in Proposition C.2. The last step is to rearrange and pick the optimal $z$, which is almost identical to that in the proof of Proposition B.2 and finally shows $\text{Reg}_T(\pi^\star) = O\left(U + \sqrt{UC} + V\right)$. This completes the entire proof.
D  General Decomposition, Self-bounding Terms, and Supplementary Lemmas

In this section, we provide details of our two key techniques: a general decomposition and self-bounding terms, as well as a set of supplementary Lemmas used throughout the analysis.

D.1  General Decomposition Lemma

In this section, we consider measuring the performance difference between a policy \( \pi \) and a mapping (deterministic policy) \( \pi^* \), that is, \( V^\pi(s_0) - V^{\pi^*}(s_0) \) where \( Q \) and \( V \) are the state-action and state value functions associated with some transition \( P \) and some loss function \( \ell \), that is,

\[
Q^\pi(s,a) = \ell(s,a) + \sum_{s' \in S_{k(s)+1}} P(s'|s,a) V^\pi(s'), \quad V^\pi(s) = \sum_{a \in A} \pi(a|s) Q^\pi(s,a),
\]

for all state-action pairs (with \( V^\pi(s_L) = 0 \)). Moreover, for some estimated transition \( \hat{P} \) and estimated loss function \( \hat{\ell} \), define similarly \( \hat{Q} \) and \( \hat{V} \) as the corresponding state-action and state value functions:

\[
\hat{Q}^\pi(s,a) = \hat{\ell}(s,a) + \sum_{s' \in S_{k(s)+1}} \hat{P}(s'|s,a) \hat{V}^\pi(s'), \quad \hat{V}^\pi(s) = \sum_{a \in A} \pi(a|s) \hat{Q}^\pi(s,a),
\]

for all state-action pairs (with \( \hat{V}^\pi(s_L) = 0 \)).

Again, we denote by \( q_{\pi}^*(s,a) \) the probability of visiting a trajectory of the form \((s_0, \pi^*(s_0)), (s_1, \pi^*(s_1)), \ldots, (s_{k(s)-1}, \pi^*(s_{k(s)-1})), (s,a)\) when executing policy \( \pi \). In other words, \( q_{\pi}^* \) can be formally defined as

\[
q_{\pi}^*(s,a) = \begin{cases} 
\pi(a|s), \\
\pi(a|s) \cdot \left( \sum_{s' \in S_{k(s)-1}} q_{\pi}^*(s', \pi^*(s)) P(s'|s', \pi^*(s)) \right), 
\end{cases} 
\]

\( s = s_0 \), otherwise.

Note that our earlier notation \( q_{\pi_1}^* \) is thus a shorthand for \( q_{\pi_1}^* \). With slight abuse of notations, we define \( q_{\pi}^*(s) = \sum_{a \in A} q_{\pi}^*(s,a) \).

Now, we present a general decomposition for \( V^\pi(s_0) - V^{\pi^*}(s_0) \).

**Lemma D.1.1.** *(General Performance Decomposition)* For any policies \( \pi \) and \( u \), and a mapping (deterministic policy) \( \pi^*: S \rightarrow A \), we have

\[
V^\pi(s_0) - V^{\pi^*}(s_0) = \sum_{s \in S_{k(s)+1}} \sum_{a \neq \pi^*(s)} q(s,a) \hat{E}_u^u(s,a) \quad \text{(Error of Sub-opt actions)}
\]

\[+ \sum_{s \in S_{k(s)+1}} \sum_{a \neq \pi^*(s)} (q(s,a) - q_{\pi}^*(s,a)) \hat{E}_u^u(s,a) \quad \text{(Error of Opt actions)}
\]

\[+ \sum_{s \in S_{k(s)+1}} \sum_{a = \pi^*(s)} q(s,a) \left( \hat{Q}^u(s,a) - \hat{V}^u(s) \right) \quad \text{(Policy Difference)}
\]

\[- \sum_{s \in S_{k(s)+1}} \sum_{a = \pi^*(s)} q_{\pi}^*(s,a) \left( \hat{Q}^u(s,a) - \hat{V}^u(s) \right) \quad \text{(Estimation Bias 1)}
\]

\[+ \sum_{s \in S_{k(s)+1}} \sum_{a \neq \pi^*(s)} q_{\pi}^*(s,a) \left( \hat{V}^u(s) - V^{\pi^*}(s) \right) \quad \text{(Estimation Bias 2)}
\]

where \( q = q_{\pi_1}^* \) is the occupancy measure associated with transition \( P \) and policy \( \pi \), and \( \hat{E}_u^u \) is a surplus function with:

\[
\hat{E}_u^u(s,a) = \ell(s,a) + \sum_{s' \in S_{k(s)+1}} P(s'|s,a) \hat{V}^\pi(s') - \hat{Q}^\pi(s,a).
\]

Moreover, selecting the surrogate policy \( u \) as the mapping \( \pi^* \) yields **Corollary D.1.2.**, which is the key decomposition lemma used in our analysis.
Corollary D.1.2. Consider an arbitrary policy sequence \( \{\pi_t\}_{t=1}^T \), an arbitrary estimated transition sequence \( \{\tilde{P}_t\}_{t=1}^T \), and an arbitrary estimated loss sequence \( \{\ell_t\}_{t=1}^T \). Then, we have

\[
\sum_{t=1}^T \left( V^{\pi_t}(s_0) - \tilde{V}^{\pi_t}(s_0) \right) + \left( \sum_{t=1}^T \tilde{V}^{\pi^*}_t(s_0) - V^{\pi^*}(s_0) \right)
\]

\[
= \sum_{t=1}^T \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \tilde{E}^{\pi^*}_t(s, a) \quad \text{(Error of Sub-opt actions)}
\]

\[
+ \sum_{t=1}^T \sum_{s \neq s_L} \sum_{a = \pi^*(s)} \left( q_t(s, a) - q^*_t(s, a) \right) \tilde{E}^{\pi^*}_t(s, a) \quad \text{(Error of Opt actions)}
\]

\[
+ \sum_{t=1}^T \sum_{s \neq s_L} \sum_{a \in A} \left( q^*_t(s, a) - \tilde{q}_t(s, a) \right) \left( \tilde{Q}^{\pi^*}_t(s, a) - \tilde{V}^{\pi^*}_t(s) \right) \quad \text{(Occupancy Difference)}
\]

\[
+ \sum_{t=1}^T \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q^*_t(s, a) \left( \tilde{V}^{\pi^*}_t(s) - V^{\pi^*}(s) \right), \quad \text{(Estimation Bias)}
\]

where \( \tilde{q}_t = q^{P, \pi_t} \), \( q_t = q^{P, \pi_t} \), \( q^*_t = q^*_t \), \( \tilde{Q}^{\pi^*}_t \) and \( \tilde{V}^{\pi^*}_t \) are the state-action and state value functions associated with \( \pi_t, \ell_t, P_t, \) and \( \tilde{E}^{\pi^*}_t \) is the surplus function defined as:

\[
\tilde{E}^{\pi^*}_t(s, a) = \ell(s, a) + \sum_{s' \in S_{k(s)+1}} P(s'|s, a) \tilde{V}^{\pi^*}_t(s') - \tilde{Q}^{\pi^*}_t(s, a).
\]

Proof. (Proof of Lemma D.1.1) By direct calculation, for all states \( s \), we have

\[
V^{\pi}(s) - \tilde{V}^{\pi}(s) = \sum_{a \in A} \pi(a|s) \left( Q^{\pi}(s, a) - \tilde{Q}^{u}(s, a) \right) + \sum_{a \in A} \pi(a|s) \left( \tilde{Q}^{u}(s, a) - \tilde{V}^{u}(s) \right)
\]

\[
= \sum_{a \in A} \pi(a|s) \sum_{s' \in S_{k(s)+1}} P(s'|s, a) \left( V^{\pi}(s') - \tilde{V}^{u}(s') \right)
\]

\[
+ \sum_{a \in A} \pi(a|s) \left( \ell(s, a) + \sum_{s' \in S_{k(s)+1}} P(s'|s, a) \tilde{V}^{u}(s') - \tilde{Q}^{u}(s, a) \right)
\]

\[
+ \sum_{a \in A} \pi(a|s) \left( \tilde{Q}^{u}(s, a) - \tilde{V}^{u}(s) \right).
\]

By repeatedly expanding \( V^{\pi}(s') - \tilde{V}^{u}(s') \) in the same way, we conclude

\[
V^{\pi}(s_0) - \tilde{V}^{u}(s_0) = \sum_{s \neq s_L} q(s, a) \tilde{E}^{u}(s, a) + \sum_{s \neq s_L} \sum_{a \in A} q(s, a) \left( \tilde{Q}^{u}(s, a) - \tilde{V}^{u}(s) \right).
\]

On the other hand, we also have for all states \( s \):

\[
V^{\pi}(s) - \tilde{V}^{u}(s)
\]

\[
= \sum_{a \in A} \pi(a|s) \left( Q^*(s, a) - \tilde{V}^{u}(s) \right) + \sum_{a \neq \pi^*(s)} \pi(a|s) \left( Q^*(s, a) - \tilde{V}^{u}(s) \right)
\]

\[
= \sum_{a = \pi^*(s)} \pi(a|s) \sum_{s' \in S_{k(s)+1}} P(s'|s, a) \left( V^{\pi}(s') - \tilde{V}^{u}(s') \right)
\]

\[
+ \sum_{a \neq \pi^*(s)} \pi(a|s) \sum_{s' \in S_{k(s)+1}} P(s'|s, a) \left( V^{\pi}(s') - \tilde{V}^{u}(s') \right).
\]
we obtain

\[
+ \sum_{a=\pi^*(s)} \pi(a|s) \left( \ell(s, a) + \sum_{s' \in S_{k+1}} P(s'|s, a) \hat{V}^u(s') - \hat{Q}^u(s, a) \right) \\
+ \sum_{a=\pi^*(s)} \pi(a|s) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right) \\
+ \sum_{a \neq \pi^*(s)} \pi(a|s) \left( Q^\pi(s, a) - \hat{V}^u(s) \right).
\]

Using Lemma D.1.3 (which repeatedly expands \( V^\pi(s') - \hat{V}^u(s') \) in the same way) with

\[
C(s) = \sum_{a=\pi^*(s)} \pi(a|s) \hat{E}^u(s, a) + \sum_{a=\pi^*(s)} \pi(a|s) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right)
\]

we obtain

\[
V^\pi(s_0) - \hat{V}^u(s_0) = \sum_{s \neq s_L} q_\pi^*(s, a) C(s)
\]

\[
= \sum_{s \neq s_L} \sum_{a=\pi^*(s)} q_\pi^*(s, a) \hat{E}^u(s, a)
\]

\[
+ \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_\pi^*(s, a) \left( Q^\pi(s, a) - \hat{V}^u(s) \right) \\
+ \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_\pi^*(s, a) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right).
\]

Combining Eq. (60) and Eq. (61), we have the following equality:

\[
\sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_\pi^*(s, a) \left( Q^\pi(s, a) - \hat{V}^u(s) \right)
\]

\[
= \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q(s, a) \hat{E}^u(s, a)
\]

\[
+ \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q(s, a) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right)
\]

\[
- \sum_{s \neq s_L} \sum_{a=\pi^*(s)} q_\pi^*(s, a) \hat{E}^u(s, a)
\]

\[
- \sum_{s \neq s_L} \sum_{a=\pi^*(s)} q_\pi^*(s, a) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right)
\]

\[
= \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q(s, a) \hat{E}^u(s, a) \\
+ \sum_{s \neq s_L} \sum_{a=\pi^*(s)} (q(s, a) - q_\pi^*(s, a)) \hat{E}^u(s, a) \\
+ \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q(s, a) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right) \\
- \sum_{s \neq s_L} \sum_{a=\pi^*(s)} q_\pi^*(s, a) \left( \hat{Q}^u(s, a) - \hat{V}^u(s) \right) \\
\text{(Error of Sub-opt actions)} \quad (62)
\]

\[
\text{(Error of Opt actions)} \quad (63)
\]

\[
\text{(Policy Difference)} \\
\text{(Estimation Bias 1)}.
\]

Next, we consider the following:

\[
V^\pi(s) - V^{\pi^*}(s)
\]
Therefore, by taking the summation over $t$ the first estimation bias term is simply 0 according to the standard performance difference lemma [Kakade, 2003, Theorem 5.2.1]. Second, (Proof of Corollary D.1.2) By applying Lemma D.1.1 with Proof.

$V$ By Lemma D.1.3 (which again repeatedly expands $V^\pi(s') - V^\pi^*(s)$ in the same way), we obtain

$$V^\pi(s_0) - V^\pi^*(s_0) = \sum_{s \notin S_L} \sum_{a \notin \pi^*(s)} q^*_\pi(s, a) \left( Q^\pi(s, a) - V^\pi(s) \right)$$

finishing the proof. 

Proof. (Proof of Corollary D.1.2) By applying Lemma D.1.1 with $u = \pi^*$, we know that $V^\pi_t(s_0) - V^\pi^*_t(s_0)$ equals to

$$\sum_{s \notin S_L} \sum_{a \notin \pi^*(s)} q^*_t(s, a) E_t^{\pi^*}(s, a)$$

$$\sum_{s \notin S_L} \sum_{a = \pi^*(s)} (q_t(s, a) - q^*_t(s, a)) E_t^{\pi^*}(s, a)$$

$$\sum_{s \notin S_L} \sum_{a = \pi^*(s)} Q_t^\pi(s, a) - \hat{V}_t^\pi(s)$$

$$\sum_{s \notin S_L} \sum_{a \in A} (q_t(s, a) - \hat{q}_t(s, a)) \left( \hat{Q}_t^\pi(s, a) - \hat{V}_t^\pi(s) \right)$$

$$- \sum_{s \notin S_L} \sum_{a \in A} \hat{q}_t(s, a) \left( \hat{Q}_t^\pi(s, a) - \hat{V}_t^\pi(s) \right)$$

Now observe the following two facts. First, the third term above is in fact equal to $\hat{V}_t^\pi(s_0) - \hat{V}_t^\pi^*(s_0)$ according to the standard performance difference difference lemma [Kakade, 2003, Theorem 5.2.1]. Second, the first estimation bias term is simply 0 since $Q_t^\pi(s, a) = \hat{V}_t^\pi(s)$ when $a = \pi^*(s)$.

Therefore, by taking the summation over $t$, we obtain

$$\text{ERR1} + \text{ERR2} = \sum_{t=1}^T \left( V_t^\pi(s_0) - V_t^\pi^*(s_0) \right) - \left( \hat{V}_t^\pi(s_0) - \hat{V}_t^\pi^*(s_0) \right)$$
\[
\begin{align*}
&= \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \hat{E}_t^\pi(s, a) \\
&\quad + \sum_{s \neq s_L} \sum_{a = \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \hat{E}_t^\pi(s, a) \\
&\quad + \sum_{s \neq s_L} \sum_{a \in A} (q_t(s, a) - \hat{q}_t(s, a)) \left( \hat{Q}_t^\pi(s, a) - \hat{V}_t^\pi(s) \right) \\
&\quad + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t^*(s, a) \left( \hat{V}_t^\pi(s) - V_t^\pi(s) \right)
\end{align*}
\]
which finishes the proof. \qed

**Lemma D.1.3.** For any functions \( F : S \to \mathbb{R} \) and \( C : S \to \mathbb{R} \) satisfying the following condition:

\[
F(s) = \sum_{a = \pi^*(s)} \pi(a|s) \sum_{s' \in S_{h(s)+1}} P(s'|s, a) F(s') + C(s)
\]
and \( F(s_L) = 0 \), we have

\[
F(s_0) = \sum_{s \neq s_L} q_t^*(s) C(s).
\]

**Proof.** By definition and direct calculation, we have \( F(s_0) \) equal to

\[
\begin{align*}
&= \sum_{a = \pi^*(s_0)} q(s_0, a) \sum_{s' \in S_1} P(s'|s_0, a) F(s') + C(s) \quad (q(s_0) = 1) \\
&= \sum_{s_1 \in S_1} q_t^*(s_1) F(s_1) + q_t^*(s_0) C(s) \\
&= \sum_{s_1 \in S_1} q_t^*(s_1) \left( \sum_{a = \pi^*(s)} \pi(a|s) \sum_{s' \in S_2} P(s'|s, a) F(s') \right) + \sum_{k=0}^{L-1} \sum_{s \in S_k} q_t^*(s) C(s) \quad \text{(definition of } q_t^*(s)\text{)} \\
&= \sum_{s_2 \in S_2} q_t^*(s_2) F(s_2) + \sum_{k=0}^{L-1} \sum_{s \in S_k} q_t^*(s) C(s) \quad \text{(repeatedly expanding)} \\
&= \sum_{s_L \in S_L} q_t^*(s_L) F(s_L) + \sum_{k=0}^{L-1} \sum_{s \in S_k} q_t^*(s) C(s) \quad (F(s_L) = 0)
\end{align*}
\]
which completes the proof. \qed

**D.2 Self-bounding Terms**

In this section, we summarize all the self-bounding terms we use in the proofs for the unknown transition settings.
Definition D.2.1 (Self-bounding Terms). For some mapping $\pi^*: S \to A$, define the following:

\[
G_1(J) = \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \sqrt{\frac{J}{\max\{m_t(s, a), 1\}}},
\]

\[
G_2(J) = \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \sqrt{\frac{J}{\max\{m_t(s, a), 1\}}},
\]

\[
G_3(J) = \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \sqrt{\frac{P(w|u, v) \cdot J}{\max\{m_t(u, v), 1\}}} q_t(s, a|w),
\]

\[
G_4(J) = \sqrt{J \cdot \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} q_t(s, a)},
\]

\[
G_5(J) = \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} \frac{1}{J} \sum_{t=1}^{T} q_t(s, a),
\]

\[
G_6(J) = \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a = \pi^*(s)} q_t(s, a) - q_t^*(s, a) \frac{1}{q_t(s, a)} \left( \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \sqrt{\frac{P(w|u, v) \cdot J}{\max\{m_t(u, v), 1\}}} q_t(s, a|w) \right).
\]

In the next six lemmas, we show that each of these six functions enjoys a certain self-bounding property under Condition (1) so that they are small whenever the regret of the learner is small. In all these lemmas, the policy $\pi^*$ is used in $G_1$–$G_6$ coincides with the $\pi^*$ in Condition (1). Also note that Lemma 5.2 is simply a collection of the first four lemmas.

**Lemma D.2.2.** Suppose Condition (1) holds. Then we have for any $\alpha \in \mathbb{R}_+$,

\[
E[G_1(J)] \leq \alpha \cdot (\text{Reg}_T(\pi^*) + C) + \frac{1}{\alpha} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} \frac{8J}{\Delta(s, a)}.
\]

**Proof.** Under the condition, for any $\alpha \in \mathbb{R}_+$, we have

\[
G_1(J) = \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \left( \sqrt{\frac{J}{\max\{m_t(s, a), 1\}}} - \alpha \Delta(s, a) \right) + \alpha \sum_{t=1}^{T} \sum_{s \notin S_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta(s, a)
\]

where the expectation of the last term is bounded by $\alpha \cdot (\text{Reg}_T(\pi^*) + C)$. It thus remains to bound the first term. To this end, for a fixed state-action pair $(s, a)$, we define $N_{s,a}$ as the last epoch where the term in the bracket is still positive, so that:

\[
m_{N_{s,a}+1}(s, a) \leq \frac{2J}{\alpha^2 \Delta(s, a)^2}
\]

due to the doubling epoch schedule. Then we have

\[
E\left[ \sum_{t=1}^{T} q_t(s, a) \left( \sqrt{\frac{J}{\max\{m_t(s, a), 1\}}} - \alpha \Delta(s, a) \right) \right]
\]

\[
= E\left[ \sum_{i=1}^{N} (m_{i+1}(s, a) - m_i(s, a)) \left( \sqrt{\frac{J}{\max\{m_i(s, a), 1\}}} - \alpha \Delta(s, a) \right) \right]
\]

\[
\leq E\left[ \sum_{i=1}^{N} (m_{i+1}(s, a) - m_i(s, a)) \left( \sqrt{\frac{J}{\max\{m_i(s, a), 1\}}} - \alpha \Delta(s, a) \right) \right]
\]

\[
\leq E\left[ 2 \int_0^{m_{N_{s,a}+1}(s, a)} \sqrt{\frac{J}{x}} dx \right] \leq E\left[ 2 \int_0^{m_{N_{s,a}+1}(s, a)} \sqrt{\frac{J}{x}} dx \right]
\]

54
where the expectation of the last term is bounded by $\beta$ (deferred to the end of this subsection). It thus remains to bound the first term. To this end, for a fixed state-action pair $(s, a)$ satisfying $a \neq \pi^*(s)$, we have

$$E[G_2(J)] \leq \alpha \cdot (\text{Reg}_T(\pi^*) + C) + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{8J}{\alpha \Delta(s, a)}.$$  

\[ \square \]

**Lemma D.2.3.** Suppose Condition (1) holds. Then we have for any $\beta \in \mathbb{R}_+$,

$$E[G_2(J)] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + \frac{8 |S| L J}{\Delta_{\text{MIN}}}.$$  

**Proof.** Clearly, under the condition, for any $\beta \in \mathbb{R}_+$, we have

$$G_2(J) = \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \left( \sqrt{\frac{J}{\max \{m_{i_t}(s, a), 1\}}} - \beta \cdot \frac{\Delta_{\text{MIN}}}{L} \right),$$

$$+ \beta \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \cdot \frac{\Delta_{\text{MIN}}}{L},$$

where the expectation of the last term is bounded by $\beta \cdot (\text{Reg}_T(\pi^*) + C)$ according to Lemma D.2.8 (deferred to the end of this subsection). It thus remains to bound the first term. To this end, for a fixed state-action pair $(s, a)$, we similarly define $N_{s, a}$ as the last epoch where the term in the bracket is still positive, so that:

$$m_{N_{s, a}+1}(s, a) \leq \frac{2JL^2}{\beta^2 \Delta_{\text{MIN}}}.$$  

due to the doubling epoch schedule. Then, we have

$$E \left[ \sum_{t=1}^{T} (q_t(s, a) - q_t^*(s, a)) \left( \sqrt{\frac{J}{\max \{m_{i_t}(s, a), 1\}}} - \beta \cdot \frac{\Delta_{\text{MIN}}}{L} \right) \right]$$

$$\leq E \left[ \sum_{i=1}^{N_{s, a}} (m_{i+1}(s, a) - m_i(s, a)) \left( \sqrt{\frac{J}{\max \{m_i(s, a), 1\}}} - \beta \cdot \frac{\Delta_{\text{MIN}}}{L} \right) \right]$$

$$\leq E \left[ 2 \int_{0}^{m_{N_{s, a}+1}(s, a)} \sqrt{\frac{J}{x}} \: dx \right] \leq E \left[ 2 \int_{0}^{\frac{2JL^2}{\beta^2 \Delta_{\text{MIN}}}} \sqrt{\frac{J}{x}} \: dx \right]$$

$$\leq 4 \cdot \sqrt{J} \cdot \sqrt{\frac{2JL^2}{\beta^2 \Delta_{\text{MIN}}} \leq \frac{8LJ}{\beta \Delta_{\text{MIN}}}.}$$

Taking the summation over all state-action pairs satisfying $a = \pi^*(s)$, we have

$$E[G_2(J)] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + \sum_{s \neq s_L} \sum_{a = \pi^*(s)} \frac{8LJ}{\beta \Delta_{\text{MIN}}}$$

$$= \beta \cdot (\text{Reg}_T(\pi^*) + C) + \frac{8 |S| L J}{\beta \Delta_{\text{MIN}}}.$$  

\[ \square \]

**Lemma D.2.4.** Suppose Condition (1) holds. Then we have for any $\alpha, \beta \in \mathbb{R}_+$,

$$E[G_3(J)] \leq (\alpha + \beta) \cdot (\text{Reg}_T(\pi^*) + C) + \frac{1}{\alpha} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} 8L^2 |S| J \Delta(s, a) + \frac{1}{\beta} \cdot \frac{8L^2 |S|^2 J}{\Delta_{\text{MIN}}}.$$  

55
where the expectation of the last term is bounded by \( \text{Reg}_{T} \).

The second step separates the optimal and sub-optimal state-action pairs, and the inequality follows from the fact \( \sum_{s \neq s_L} \sum_{a \in A} q_t(s, a|w) \leq L \) and the Cauchy-Schwarz inequality. Note that, the first term is simply \( G(T)(L^2|S|) \) and can be applied using Lemma D.2.2.

To bound the last term, we first observe the following:

\[
\sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{v \neq \pi^*(s)} q_t(u, v) \left( \sum_{w \in S_{k+1}} \frac{P(w|u, v) \cdot J}{\max \{m_i(t)(s, a), 1\}} \right) \leq \sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{v \neq \pi^*(s)} q_t(s, a) \Delta_{\text{MIN}}
\]

where the expectation of the last term is bounded by \( \text{Reg}_{T}(\pi^*) + C \) under Condition (1).

Let \( \text{clip}[x] = \max \{x, 0\} \) be the clipping function that removes the negative value. By adding and subtracting \( \beta \) times the term above, we have

\[
\sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{v \neq \pi^*(s)} q_t(u, v) \left( \sum_{w \in S_{k+1}} \sqrt{\frac{P(w|u, v) \cdot J}{\max \{m_i(t)(s, a), 1\}}} \right) \leq \sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{v \neq \pi^*(s)} q_t(s, a) \Delta_{\text{MIN}} \leq \sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{v \neq \pi^*(s)} q_t(s, a) \Delta_{\text{MIN}}
\]

where the expectation of the last term is bounded by \( \text{Reg}_{T}(\pi^*) + C \) under Condition (1).
where the last line follows from the facts $x \leq \text{clip}[x]$ and $\sum_{s \neq s_L} \sum_{a \in A} q_t(s, a|w) \leq L$.

Fix a tuple $N_{u,v,w}$ where $v = \pi^*(u)$, we similarly define $N_{u,v,w}$ as the last epoch where the argument of clip$(\cdot)$ is still positive, so that:

$$m_{N_{u,v,w}+1}(s, a) \leq \frac{2JL^2}{P(w|u, v)\beta^2\Delta_{\min}^2}$$

due to the doubling epoch schedule. Then, we have

$$\mathbb{E} \left[ \sum_{t=1}^{T} q_t(u, v) \text{clip} \left[ \frac{P(w|u, v) \cdot J}{\max \{ m_i(t), 1 \}} - \frac{\Delta_{\min} P(w|u, v)}{L} \right] \right] \leq \mathbb{E} \left[ \sum_{i=1}^{N_{u,v,w}} (m_{i+1}(u, v) - m_i(u, v)) \text{clip} \left[ \frac{P(w|u, v) \cdot J}{\max \{ m_i(t), 1 \}} - \frac{\Delta_{\min} P(w|u, v)}{L} \right] \right] \leq \mathbb{E} \left[ \sum_{i=1}^{m_{N_{u,v,w}+1}(s, a)} \int_0^1 \frac{P(u, v) \cdot J}{x} dx \right] \leq \mathbb{E} \left[ \int_0^1 \frac{P(w|u, v) \cdot J}{x} dx \right] \leq \frac{2JL^2}{\beta^2\Delta_{\min}^2} \leq \frac{8LJ}{\beta\Delta_{\min}}.
$$

Taking the summation over all transition tuple $(u, v, w)$ satisfying $v = \pi^*(s)$ and adding $\mathbb{E} \left[ G_1(L^2|S|J) \right]$, we have

$$\mathbb{E} \left[ G_3(J) \right] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + \mathbb{E} \left[ G_1(L^2|S|J) \right] + L \sum_{k=0}^{L-1} \sum_{s \in S_k} \sum_{v = \pi^*(u)} \sum_{w \in S_{k+1}} \frac{8LJ}{\beta\Delta_{\min}} \leq (\alpha + \beta) \cdot \text{Reg}_T(\pi^*) + C + \frac{1}{\alpha} \cdot \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{8L^2|J|}{\Delta(s, a)} + \frac{1}{\beta} \cdot \frac{8L^2|S|^2J}{\Delta_{\min}}.$$

where the last line follows from the fact $\sum_{k=0}^{L-1} |S_k| |S_{k+1}| \leq |S|^2$.

**Lemma D.2.5.** Suppose Condition (1) holds. Then we have for any $\beta \in \mathbb{R}_+$,

$$\mathbb{E} \left[ G_4(J) \right] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + \frac{1}{\beta} \cdot \frac{J}{4\Delta_{\min}}.$$

**Proof.** By the fact that $2\sqrt{xy} \leq x + y$ for all $x, y \geq 0$, with Condition (1), we have

$$\mathbb{E} \left[ G_4(J) \right] = \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta_{\min} \cdot \frac{J}{2\beta\Delta_{\min}} \right] \leq \beta \cdot \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta_{\min} \right] + \frac{J}{4\beta\Delta_{\min}} \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + \frac{J}{4\beta\Delta_{\min}}.$$

**Lemma D.2.6.** Suppose Condition (1) holds. Then we have for any $\alpha \in \mathbb{R}_+$,

$$\mathbb{E} \left[ G_5(J) \right] \leq \alpha \cdot (\text{Reg}_T(\pi^*) + C) + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{J}{4\alpha\Delta(s, a)}.$$
Proof. By the fact that $2\sqrt{xy} \leq x + y$ for all $x, y \geq 0$, with Condition (1), we have

$$
\mathbb{E}[G_4(J)] = \mathbb{E} \left[ \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \sum_{t=1}^{T} q_t(s, a) \Delta(s, a) \cdot \frac{J}{2\alpha \Delta(s, a)} \right]
$$

$$
\leq \alpha \cdot \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) \Delta(s, a) \right] + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{J}{4\alpha \Delta(s, a)}
$$

$$
\leq \alpha \cdot (\text{Reg}_T(\pi^*) + C) + \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{J}{4\alpha \Delta(s, a)}.
$$

Lemma D.2.7. Suppose Condition (1) holds. Then we have for any $\beta \in \mathbb{R}_+$.

$$
\mathbb{E}[G_6(J)] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + \frac{1}{\beta} \cdot \frac{8L^3 |S|^2 |A| \cdot J}{\Delta_{\min}}.
$$

Proof. By adding and subtracting terms, we have $G_6(J)$ equals to

$$
\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s, a) - q_t^*(s, a) \frac{J}{2\alpha \Delta(s, a)}
$$

$$
\left( k(s) - 1 \right) \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \frac{P(w|u, v) \ln \left( \frac{T|S||A|}{s} \right)}{\max \left\{ m_{t(i)}(u, v), 1 \right\}} q_t(s, a|w) - \beta q_t(s, a) \cdot \frac{\Delta_{\min}}{L}
$$

$$
+ \beta \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} (q_t(s, a) - q_t^*(s, a)) \Delta_{\min}
$$

where the expectation of the last term is bounded by $\beta \cdot (\text{Reg}_T(\pi^*) + C)$ according to Lemma D.2.8.

To bound the first term, we observe that

$$
\sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \frac{P(w|u, v) \ln \left( \frac{T|S||A|}{s} \right)}{\max \left\{ m_{t(i)}(u, v), 1 \right\}} q_t(s, a|w)
$$

$$
= \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \frac{P(w|u, v) \ln \left( \frac{T|S||A|}{s} \right)}{\max \left\{ m_{t(i)}(u, v), 1 \right\}} q_t(s, a|w)
$$

$$
- \beta \cdot \frac{\Delta_{\min}}{L^2} \left( \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) P(w|u, v) q_t(s, a|w) \right)
$$

$$
= \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \left( \frac{P(w|u, v) \ln \left( \frac{T|S||A|}{s} \right)}{\max \left\{ m_{t(i)}(u, v), 1 \right\}} \cdot P(w|u, v) \cdot \beta \cdot \frac{\Delta_{\min}}{L^2} \right) \cdot q_t(s, a|w)
$$

$$
\leq \sum_{k=0}^{k(s)-1} \sum_{(u, v, w) \in T_k} q_t(u, v) \text{clip} \left[ \frac{P(w|u, v) \ln \left( \frac{T|S||A|}{s} \right)}{\max \left\{ m_{t(i)}(u, v), 1 \right\}} \cdot P(w|u, v) \cdot \beta \cdot \frac{\Delta_{\min}}{L^2} \right] q_t(s, a|w)
$$

where the first equality uses $\sum_{(u, v, w) \in T_k} q_t(u, v) P(w|u, v) q_t(s, a|w) = q_t(s, a)$ for all layer $k = 0, \ldots, k(s) - 1$. (Recall clip[$x$] = max{$x, 0$}.)
Therefore, with Condition (1), we bound the $E[\mathcal{G}_0(J)]$ by

$$
E \left[ \sum_{t=1}^{T} \sum_{s \neq \overline{S}_L} \sum_{a = \pi^*(s)} \frac{q_t(s,a) - q_t^*(s,a)}{q_t(s,a)} \left( \sum_{u,v,w} q_t(u,v)h_t(u,v,w)q_t(s,a|w) \right) + \beta \cdot (\text{Reg}_T(\pi^*) + C) \right]
$$

where the second line applies the fact $\frac{q_t(s,a) - q_t^*(s,a)}{q_t(s,a)} \leq 1$, and the third line changes summation order and uses the fact that $\sum_{s \neq \overline{S}_L} \sum_{a \in A} q_t(s,a|w) \leq L$.

Finally, following the similar idea of handong $\sum_{t=1}^{T} q_t(u,v)h_t(u,v,w)$ as in Lemma D.2.4, we have

$$
E \left[ \sum_{t=1}^{T} q_t(u,v)h_t(u,v,w) \right] \leq \frac{8L^2J}{\beta \Delta_{\text{MIN}}},
$$

By taking the summation over all transition triples, we have

$$
E[\mathcal{G}_0(J)] \leq \beta \cdot (\text{Reg}_T(\pi^*) + C) + L \cdot \sum_{k=0}^{L-1} \sum_{(u,v,w) \in T_k} \frac{1}{\beta} \cdot \frac{8L^2 \cdot J}{\Delta_{\text{MIN}}}
$$

where the last line follows from the fact that $\sum_{k=0}^{L} |S_k| \leq |S|^2$.

**Lemma D.2.8.** Under Condition (1), we have

$$
E \left[ \sum_{t=1}^{T} \sum_{s \neq \overline{S}_L} \sum_{a = \pi^*(s)} (q_t(s,a) - q_t^*(s,a)) \Delta_{\text{MIN}} \right] \leq L \cdot E[\text{Reg}_T(\pi^*) + C].
$$

**Proof.** For each $k$, we proceed as

$$
\sum_{s \in S_k} \sum_{a = \pi^*(s)} (q_t(s,a) - q_t^*(s,a)) \\
\leq 1 - \sum_{s \in S_k} \sum_{a = \pi^*(s)} q_t^*(s,a) \quad (\sum_{s \in S_k} \sum_{a \in A} q_t(s,a) = 1)
$$

$$
= 1 - \sum_{s \in S_k} \sum_{a = \pi^*(s)} \pi_t(a|s) \Pr \left[ \{s_k = s\} \bigcap \left( \bigcap_{\tau=0}^{k-1} \{a_\tau = \pi^*(s_\tau)\} \right) \bigg| P, \pi_t \right] \quad \text{(definition of } q_t^*)
$$

$$
= 1 - \Pr \left[ \left( \bigcap_{\tau=0}^{k} \{a_\tau = \pi^*(s_\tau)\} \right) \bigg| P, \pi_t \right]
$$

$$
= \Pr \left[ \left( \bigcup_{\tau=0}^{k} \{a_\tau \neq \pi^*(s_\tau)\} \right) \bigg| P, \pi_t \right] \quad \text{(De Morgan’s laws)}
$$

$$
\leq \sum_{\tau=0}^{k} \Pr \left[ a_\tau \neq \pi^*(s_\tau) \big| P, \pi_t \right] \quad \text{(union bound)}
$$

59
where the last line follows from Condition (1).

\[ \sum_{t=1}^{T} \sum_{\tau \neq s_L} \sum_{a \neq \pi^*(s)} (q_t(s,a) - q^*_t(s,a)) \Delta_{MIN} \]

\[ \leq L \cdot \sum_{t=1}^{T} \sum_{\tau \neq s_L} \sum_{a \neq \pi^*(s)} q_t(s,a) \cdot \Delta(s,a) \]

\[ \leq L \cdot \mathbb{E} [\text{Reg}_T(\pi^*) + C] \]

where the last line follows from Condition (1).

**D.3 Supplementary Lemmas**

**Lemma D.3.1. (Occupancy Measure Difference)** For any policy \( \pi \) and transition functions \( P_1 \) and \( P_2 \), with \( q_1 = q_1^{P_1,\pi} \) and \( q_2 = q_2^{P_2,\pi} \) we have for all \( s \),

\[
q_1(s) - q_2(s) = \sum_{k=0}^{k(s)-1} \sum_{u \in S_k} \sum_{v \in A} \sum_{w \in S_{k+1}} q_1(u,v) [P'_1(w|u,v) - P'_2(w|u,v)] q_2(s|w) \\
= \sum_{k=0}^{k(s)-1} \sum_{u \in S_k} \sum_{v \in A} \sum_{w \in S_{k+1}} q_2(u,v) [P'_1(w|u,v) - P'_2(w|u,v)] q_1(s|w)
\]

where the conditional occupancy measure \( q_1(s'|s) \) (similarly for \( q_2(s'|s) \)) is defined recursively as

\[
q_1(s'|s) = \begin{cases} 
0, & k(s') < k(s) \text{ or } (k(s') = k(s) \text{ and } s' \neq s) \\
1, & k(s') = k(s) \text{ and } s' = s \\
\sum_{u \in S_{k(s')-1}} q_1(u|s) \left( \sum_{v \in A} \pi(v|u) P(s'|u,v) \right), & k(s') > k(s)
\end{cases}
\]

which is the conditional probability of visiting state \( s' \) from \( s \) under \( \pi \) and transition \( P_1 \).

**Proof.** Fix a state \( s \). We proceed as:

\[
q_1(s) - q_2(s) = \sum_{s' \in S_{k(s)-1}} \sum_{a' \in A} (q_1(s'|s, a') P_1(s'|s, a') - q_2(s'|s, a') P_2(s'|s, a')) \\
= \sum_{s' \in S_{k(s)-1}} \sum_{a' \in A} (q_1(s') - q_2(s')) P_1(s|s', a') \pi(a'|s') \\
+ \sum_{s' \in S_{k(s)-1}} \sum_{a' \in A} q_2(s'|s, a') (P_1(s|s', a') - P_2(s|s', a'))
\]

where the second step follows by subtracting and adding \( q_2(s', a') P_1(s|s', a') \). Note that, \( \sum_{a' \in A} \pi(a'|s') P_1(s|s', a') \) is exactly the conditional probability of transiting to state \( s \) from state \( s' \) with transition \( P_1 \). Therefore, we have \( \sum_{a' \in A} \pi(a'|s') P_1(s|s', a') = q_1(s|s') \) according to Eq. (66), and further expand \( q_1(s) - q_2(s) \) as:

\[
\sum_{s' \in S_{k(s)-1}} \sum_{a' \in A} (q_1(s') - q_2(s')) P_1(s|s', a') \pi(a'|s') \\
+ \sum_{s' \in S_{k(s)-1}} \sum_{a' \in A} q_2(s', a') (P_1(s|s', a') - P_2(s|s', a'))
\]

\[ = \sum_{s' \in S_{k(s)-1}} q_1(s|s') (q_1(s') - q_2(s'))
\]
where the second line follows from the fact that \( q_1(s) = 1 \).
Therefore, we can recursively expand \( q_1(s) - q_2(s) \) as:

\[
q_1(s) - q_2(s) = \sum_{s' \in S_{k(s)}^{-1}} (q_1(s') - q_2(s')) q_1(s'|s') \\
+ \sum_{s' \in S_{k(s)}^{-1}} \sum_{a \in A} q_2(s', a') [P_1(s|s', a') - P_2(s|s', a')] q_1(s|s)
\]

By applying the same technique, we also have

\[
\text{Lemma D.3.2. The following holds:}
\]

\[
B_i(s, a) \leq 2 \sqrt{\frac{|S_{k(s)}+1| \ln \left( \frac{T|S||A|}{\delta} \right)}{\max \{m_i(s, a), 1\}}} + 14|S_{k(s)}+1| \ln \left( \frac{T|S||A|}{\delta} \right) \\
+ \frac{14|S_{k(s)}+1| \ln \left( \frac{T|S||A|}{\delta} \right)}{3 \max \{m_i(s, a), 1\}}.
\]

\text{Proof.} \] By the definition of \( B_i(s, a) \), we have

\[
B_i(s, a) = \sum_{s' \in S_{k(s)}+1} B_i(s, a, s')
\]
We often use the following two lemmas to deal with the small-probability event $A$ where the third line applies the fact that $\max\{m_i(s,a), 1\}$. The last line follows from the Cauchy-Schwarz inequality.

**Lemma D.3.3.** Conditioning on event $A$, we have

$$B_i(s, a, s') \leq 4 \sqrt{\frac{\sum_{s' \in S_{k(s)+1}} P(s'|s, a) \ln \left( \frac{T[S|A]}{2 \delta} \right)}{\max\{m_i(s,a), 1\}}} + \frac{40 \ln \left( \frac{T[S|A]}{\delta} \right)}{3 \max\{m_i(s,a), 1\}}. \quad (67)$$

**Proof.** By direct calculation based on Eq. (8) and the condition of event $A$, we have

$$B_i(s, a, s') \leq 2 \sqrt{\frac{\sum_{s' \in S_{k(s)+1}} P(s'|s, a) \ln \left( \frac{T[S|A]}{\delta} \right)}{\max\{m_i(s,a), 1\}}} \leq 2 \sqrt{\frac{(P(s'|s, a) + B_i(s, a, s')) \ln \left( \frac{T[S|A]}{\delta} \right)}{\max\{m_i(s,a), 1\}}} + \frac{14 \ln \left( \frac{T[S|A]}{\delta} \right)}{3 \max\{m_i(s,a), 1\}}$$

where the third line applies the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, and the last line follows from the fact $2\sqrt{xy} \leq x + y$ for $x, y > 0$.

Rearranging the terms yields that

$$B_i(s, a, s') \leq 4 \sqrt{\frac{P(s'|s, a) \ln \left( \frac{T[S|A]}{\delta} \right)}{\max\{m_i(s,a), 1\}}} + \frac{40 \ln \left( \frac{T[S|A]}{\delta} \right)}{3 \max\{m_i(s,a), 1\}}.$$

Combining with the fact $B_i(s, a, s') \leq 1$, we have the following tighter bound of confidence width.

**Corollary D.3.4.** Conditioning on event $A$, we have

$$B_i(s, a, s') \leq \min \left\{ \frac{4 \sqrt{\frac{P(s'|s, a) \ln \left( \frac{T[S|A]}{\delta} \right)}{\max\{m_i(s,a), 1\}}} + \frac{40 \ln \left( \frac{T[S|A]}{\delta} \right)}{3 \max\{m_i(s,a), 1\}}}{1}, 1 \right\} \leq \min \left\{ \frac{4 \sqrt{\frac{P(s'|s, a) \ln \left( \frac{T[S|A]}{\delta} \right)}{\max\{m_i(s,a), 1\}}}}{1}, 1 \right\}.$$

We often use the following two lemmas to deal with the small-probability event $A^c$ when taking expectation.

**Lemma D.3.5.** Suppose that a random variable $X$ satisfies the following conditions:
Then, we have

$$
\mathbb{E} [X] \leq C \cdot \Pr [E^c] + \mathbb{E} [Y].
$$

**Proof.** By writing the random variable $X$ as $X \cdot \mathbb{I} \{E\} + X \cdot \mathbb{I} \{E^c\}$, and noting

$$
X \cdot \mathbb{I} \{E\} \leq Y \cdot \mathbb{I} \{E\} \leq Y, \quad X \cdot \mathbb{I} \{E^c\} \leq C \cdot \mathbb{I} \{E^c\},
$$

we prove the statement after taking the expectations. \qed

**Lemma D.3.6.** Suppose that a random variable $X$ satisfies the following conditions:

- Conditioning on event $E$, $X < Y$ where $Y > 0$ is another random variable;
- $X < C$ holds where $C$ is another random variable which ensures $\mathbb{E} [C|E^c] \leq D$ for some fixed $D \in \mathbb{R}^+$. Then, we have

$$
\mathbb{E} [X] \leq D \cdot \Pr [E^c] + \mathbb{E} [Y].
$$

**Proof.** By writing the random variable $X$ as $X \cdot \mathbb{I} \{E\} + X \cdot \mathbb{I} \{E^c\}$, and noting

$$
X \cdot \mathbb{I} \{E\} \leq Y \cdot \mathbb{I} \{E\} \leq Y, \quad X \cdot \mathbb{I} \{E^c\} \leq C \cdot \mathbb{I} \{E^c\}, \quad \mathbb{E} [C \cdot \mathbb{I} \{E^c\}] \leq \mathbb{E} [C|E^c],
$$

we prove the statement after taking the expectations. \qed

**Lemma D.3.7.** (Jin et al., 2020, Lemma 10) With probability at least $1 - 2\delta$, we have for all $k = 0, \ldots, L - 1$,

$$
\sum_{t=1}^{T} \sum_{s \in S_k} \frac{q_t(s, a)}{\max \{1, m_{i(t)}(s, a)\}} = \mathcal{O} \left( |S_k| \|A| \ln T + \ln(L/\delta) \right)
$$

and

$$
\sum_{t=1}^{T} \sum_{s \in S_k} \frac{q_t(s, a)}{\max \{1, m_{i(t)}(s, a)\}} = \mathcal{O} \left( \sqrt{|S_k| \|A| T} + |S_k| \|A| \ln T + \ln(L/\delta) \right).
$$

Simultaneously, for all $k < h$, we have

$$
\sum_{t=1}^{T} \sum_{(u,v,w) \in T_k} \sum_{(x,y,z) \in T_h} q_t(u, v) \sqrt{\frac{P(w|u, v)}{\max \{1, m_{i(t)}(u, v)\}}} \cdot q_t(x, y) \sqrt{\frac{P(z|x, y)}{\max \{1, m_{i(t)}(x, y)\}}} \leq \mathcal{O} \left( |A| \ln T + \ln \left(\frac{T}{\delta}\right) \right) \cdot \sqrt{|S_k| |S_{k+1}| |S_h| |S_{h+1}|}.
$$

**Proof.** Eq. (68) and Eq. (69) are from Jin et al. [2020]. For Eq. (70), by direct calculation we have

$$
\sum_{t=1}^{T} \sum_{(u,v,w) \in T_k} \sum_{(x,y,z) \in T_h} q_t(u, v) \sqrt{\frac{P(w|u, v)}{\max \{1, m_{i(t)}(u, v)\}}} \cdot q_t(x, y) \sqrt{\frac{P(z|x, y)}{\max \{1, m_{i(t)}(x, y)\}}} \leq \mathcal{O} \left( |A| \ln T + \ln \left(\frac{T}{\delta}\right) \right) \cdot \sqrt{|S_k| |S_{k+1}| |S_h| |S_{h+1}|}.
$$

63
where the last inequality follows from the Cauchy-Schwarz inequality.

\[
E \left[ \sum_{t=1}^{T} \frac{q_t(s,a)}{\max\{1, m_i(t)(s,a)\}} \right] = O \left( |S_k| \ln T + |S_k||A| \right) \tag{71}
\]

and

\[
E \left[ \sum_{t=1}^{T} \frac{q_t(s,a)}{\max\{1, m_i(t)(s,a)\}} \right] = O \left( \sqrt{|S_k||A|T} + |S_k||A| \right). \tag{72}
\]

**Lemma D.3.8.** For all \( k = 0, \ldots, L - 1 \), we have

\[
E \left[ \sum_{t=1}^{T} \frac{q_t(s,a)}{\max\{1, m_i(t)(s,a)\}} \right] = O \left( |S_k||A| \ln T + |S_k||A| \right)
\]

where the second line follows from the indicator and occupancy measure \( q_t \), and the last line applies the fact \( m_{i+1}(s,a) \leq 2m_i(s,a) \) when \( m_i(s,a) \geq 1 \). Taking the summation over all state-action pairs at layer \( k \) finishes the proof of Eq. (71).

Similarly, we have

\[
E \left[ \sum_{t=1}^{T} \frac{q_t(s,a)}{\max\{1, m_i(t)(s,a)\}} \right] = O \left( |S_k||A| \ln T + |S_k||A| \right)
\]

where \( m_{N+1}(s,a) \) is the total number of visiting state-action pair \( (s,a) \). Taking the summation over all state-action pairs of layer \( k \) yields that

\[
E \left[ \sum_{s \in S_k} \sum_{a \in A} \sum_{t=1}^{T} \frac{q_t(s,a)}{\max\{1, m_i(t)(s,a)\}} \right] \leq \sum_{s \in S_k} \sum_{a \in A} 2 \left( 2 \sqrt{m_{N+1}(s,a)} + 1 \right) \leq 2 \left( 2 \sqrt{|S_k||A|T} + |S_k||A| \right)
\]

where the last inequality follows from the Cauchy-Schwarz inequality.
Definition D.3.9. (Residual Term) We define the residual term \( r_t(s, a) \) as

\[
r_t(s, a) = \frac{40}{3} \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v, w) \cdot \frac{P(w|u, v) \ln \left( \frac{T|S|A}{\delta} \right)}{\max \left\{ m_{i(t)}(u, v), 1 \right\}} \cdot q_t(s, a|w)
\]

(73)

for all state-action pair \((s, a) \in S \times A\) and all episodes \(t \in [T]\).

Lemma D.3.10. The following hold:

\[
|q_t(s, a) - \tilde{q}_t(s, a)| \leq r_t(s, a) + 4 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) \sqrt{\frac{P(w|u, v) \ln \left( \frac{T|S|A}{\delta} \right)}{\max \left\{ m_{i(t)}(u, v), 1 \right\}}} q_t(s, a|w)
\]

and

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} r_t(s, a) \right] = O \left( L^2|S|^3|A|^2 \ln^2 \left( \frac{T|S|A}{\delta} \right) + |S||A|^T \right).
\]

Proof. For simplicity, we let \( \epsilon = \frac{T|S|A}{\delta} \) and assume \( \delta \in (0, 1) \). According to the Lemma D.3.1, conditioning on event \( \mathcal{A} \), we have

\[
|q_t(s, a) - \tilde{q}_t(s, a)| = \left| \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) \left( P(w|u, v) - \tilde{P}_{i(t)}(w|u, v) \right) \tilde{q}_t(s, a|w) \right|
\]

\[
\leq \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) \left| P(w|u, v) - \tilde{P}_{i(t)}(w|u, v) \right| \tilde{q}_t(s, a|w)
\]

Moreover, we apply Lemma D.3.1 again to conditional occupancy measure and obtain

\[
|q_t(s, a|w) - \tilde{q}_t(s, a|w)| \leq \sum_{h=k(w)}^{k(s)-1} \sum_{(x,y,z) \in T_h} q_t(x, y|w) B_{i(t)}(x, y, z) \tilde{q}_t(s, a|z)
\]

where the second line applies the fact \( \tilde{q}_t(s, a|z) \leq 1 \).

Combining these inequalities yields (under the event \( \mathcal{A} \))

\[
|q_t(s, a) - \tilde{q}_t(s, a)| \leq \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) B_{i(t)}(u, v, w) q_t(s, a|w)
\]

\[
+ \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) B_{i(t)}(u, v, w) \left( \sum_{h=k(w)}^{k(s)-1} \sum_{(x,y,z) \in T_h} q_t(x, y|w) B_{i(t)}(x, y, z) \right)
\]

\[
\leq 4 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) \sqrt{\frac{P(w|u, v) \ln \left( \frac{T|S|A}{\delta} \right)}{\max \left\{ m_{i(t)}(u, v), 1 \right\}}} q_t(s, a|w)
\]
\[
+ \frac{40}{3} \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) \cdot \frac{P(w|u,v) \ln \ell}{\max \{ m_{i(t)}(u,v), 1 \}} \cdot q_t(s,a|w)
\]
\[
+ \sum_{k=0}^{k(s)-1} \sum_{h=k+1}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) B_{i(t)}(u,v,w) q_t(x,y|w) B_{i(t)}(x,y,z)
\]

where the second line follows from Lemma D.3.3.

On the other hand, \( |q_t(s,a) - \hat{q}_t(s,a)| \leq 1 \) always holds. Combining the bounds of these two cases finishes the first statement.

Recall the definition of the residual terms, we decompose the following into three terms \( \text{SUM}_1, \text{SUM}_2 \) and \( \text{SUM}_3 \):

\[
\begin{align*}
\mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} r_t(s,a) \right] &= \frac{40}{3} \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) \cdot \frac{P(w|u,v) \ln \ell}{\max \{ m_{i(t)}(u,v), 1 \}} \cdot q_t(s,a|w) \right] \\
&+ \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \mathbb{I}\{ A^c \} \right] \\
&+ \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \sum_{k=0}^{k(s)-1} \sum_{h=k+1}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) B_{i(t)}(u,v,w) q_t(x,y|w) B_{i(t)}(x,y,z) \right].
\end{align*}
\]

Then, we show that these terms are all logarithmic in \( T \).

**SUM_1**  By direct calculation, we have

\[
\text{SUM}_1 = \frac{40}{3} \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(u,v) \cdot \frac{P(w|u,v) \ln \ell}{\max \{ m_{i(t)}(u,v), 1 \}} \cdot \sum_{s \neq s_L} \sum_{a \in A} P(w|u,v) q_t(s,a|w) \right]
= \frac{40}{3} \ln \ell \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} q_t(u,v) \cdot \frac{\ln \ell}{\max \{ m_{i(t)}(u,v), 1 \}} \cdot \sum_{s \neq s_L} \sum_{a \in A} P(w|u,v) q_t(s,a|w) \right]
\leq \frac{40L}{3} \ln \ell \left( \sum_{k=0}^{L-1} |S_k||A| (\ln T + 1) \right) = O(L|S||A| \ln^2 \ell) \quad (74)
\]

where the first line follows from the property of occupancy measures, and the last line applies Eq. (71) of Lemma D.3.8.

**SUM_2**  According to the definition of event \( A \), we have

\[
\text{SUM}_2 = \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \mathbb{I}\{ A^c \} \right] = |S||A|T \cdot \mathbb{E} [\mathbb{I}\{ A^c \}] = |S||A|T \cdot \delta. \quad (75)
\]
where the third line follows from the AM-GM inequality. Taking the expectation with
where the second inequality follows from Lemma D.3.3 and Corollary D.3.4.
Then we consider bounding these three different terms with the help of previous analysis. According
On the other hand, for the second term, we have
\[
\left| 1 + 40 \sum_{t=1}^{T} \sum_{k=k_0}^{k(s)-1} \sum_{h=k+1}^{k(s)-1} \sum_{(u,v) \in T_h} q_t(u,v) \right| S_h \leq \frac{16|S||A| \ln \ell}{3} \sum_{t=1}^{T} \sum_{k=k_h}^{k} \sum_{(u,v) \in T_h} q_t(u,v) \sqrt{P(w|u,v) \ln \ell} \min \left\{ \frac{P(z|x,y) \ln \ell}{\max \left\{ m_i(t)(x,y), 1 \right\}}, 1 \right\},
\]
where the second inequality follows from Lemma D.3.3 and Corollary D.3.4.

Then we consider bounding these three different terms with the help of previous analysis. According
to Eq. (70) of Lemma D.3.7, the first term is bounded with probability at least \( 1 - 2\epsilon' \):
\[
16|S||A| \ln \ell \cdot \mathcal{O} \left( |A| \ln T + \ln (L/\delta') \sum_{k=k_h}^{k} \sum_{(u,v) \in T_h} q_t(u,v) \right) \leq 16|S||A| \ln \ell \cdot \mathcal{O} \left( |A| \ln T + \ln (L/\delta') \sum_{k=k_h}^{k} \sum_{(u,v) \in T_h} q_t(u,v) \right) \leq \mathcal{O} \left( |A| \ln T + \ln (L/\delta') \right) L|S|^3|A| \ln \ell,
\]
where the third line follows from the AM-GM inequality. Taking the expectation with \( \delta' = \frac{\ell}{16} \), we
have the expectation of the first term bounded by \( \mathcal{O} \left( L|S|^3|A|^2 \ln^2 \ell \right) \) using Lemma D.3.5.

On the other hand, for the second term, we have
\[
\frac{160|S||A|}{3} \sum_{t=1}^{T} \sum_{k=k_h}^{k} \sum_{(u,v) \in T_h} q_t(u,v) \sqrt{P(w|u,v) \ln \ell} \min \left\{ \frac{P(z|x,y) \ln \ell}{\max \left\{ m_i(t)(x,y), 1 \right\}}, 1 \right\},
\]
\[
\left| 1 + \frac{160|S||A|}{3} \sum_{t=1}^{T} \sum_{k=k_h}^{k} \sum_{(u,v) \in T_h} q_t(u,v) \sqrt{P(w|u,v) \ln \ell} \min \left\{ \frac{P(z|x,y) \ln \ell}{\max \left\{ m_i(t)(x,y), 1 \right\}}, 1 \right\},
\]
\[
\left| 1 + \frac{80|S||A|}{3} \sum_{t=1}^{T} \sum_{k=k_h}^{k} \sum_{(u,v) \in T_h} q_t(u,v) \sqrt{P(w|u,v) \ln \ell} \min \left\{ \frac{P(z|x,y) \ln \ell}{\max \left\{ m_i(t)(x,y), 1 \right\}}, 1 \right\},
\]
\[
\left| 1 + \frac{80L|S|^2|A|}{3} \ln \ell \sum_{t=1}^{T} \sum_{u \in S} \sum_{v \in A} \left( \frac{q_t(x,y)}{\max \left\{ m_i(t)(x,y), 1 \right\}} \right),
\]
\[
\left| 1 + \frac{80L|S|^2|A|}{3} \ln \ell \sum_{t=1}^{T} \sum_{u \in S} \sum_{v \in A} \left( \frac{q_t(x,y)}{\max \left\{ m_i(t)(x,y), 1 \right\}} \right),
\]
67
where the expectation of the final term is bounded $O(LS^3|A|^2 \ln^2 \epsilon)$ with the help from Lemma D.3.8. Similarly, we have the expectation of the third term bounded by $O(LS^3|A|^2 \ln^2 \epsilon)$ following the same idea.

Therefore, we have $S_{\epsilon}$ bounded as

$$S_{\epsilon} = O(LS^3|A|^2 \ln^2 \epsilon + LS^3|A|^2 \ln^2 \epsilon + |S||\mathcal{A}|T \cdot \delta)$$

where $|S||\mathcal{A}|T \cdot \delta$ comes from the range of $S_{\epsilon}$ and the probability of event $\mathcal{A}^c$.

Combining the bounds of $S_{\epsilon}$, $S_{\epsilon}$, and $S_{\epsilon}$ stated in Eq. (74), Eq. (75) and Eq. (76) finishes the proof.

**Corollary D.3.11.** The following holds:

$$|q_t(s,a) - u_t(s,a)| \leq 4r_t(s,a) + 16 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) \left( \frac{P(w|u,v) \ln \left( \frac{T|S||A|}{\delta} \right)}{\max \left\{ m_{i(t)}(u,v), 1 \right\}} \right) q_t(s,a|w).$$

where $q_t$ is the true occupancy measure of episode $t$, and $u_t$ is the upper occupancy bound of episode $t$ associated with confidence set $\mathcal{P}_{\epsilon}(t)$ and policy $\pi_t$.

**Proof.** Fix the state-action pair $(s,a)$ and episode $t$. Let $\hat{P}$ be the transition in $\mathcal{P}_{\epsilon}(t)$ that realizes the maximum in the definition of $u_t(s,a)$, and $\bar{q}_t = q_{\pi_t}$ be the associated occupancy measure. Therefore, we have $\bar{q}_t(s,a) = u_t(s,a)$.

Conditioning on event $\mathcal{A}$, we have

$$|q_t(s,a) - \bar{q}_t(s,a)| = \left| \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) \left( P(w|u,v) - \hat{P}(w|u,v) \right) \bar{q}_t(s,a|w) \right|$$

$$\leq \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v) \left| P(w|u,v) - \hat{P}(w|u,v) \right| \bar{q}_t(s,a|w)$$

$$\leq 2 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v)B_{i(t)}(u,v,w)\bar{q}_t(s,a|w).$$

Moreover, we apply Lemma D.3.1 to terms $\bar{q}_t(s,a|w)$ and obtain

$$|q_t(s,a|w) - \bar{q}_t(s,a|w)| \leq 2 \sum_{h=k(w)}^{k(s)-1} \sum_{(x,y,z) \in T_h} q_t(x,y|w)B_{i(t)}(x,y,z)\bar{q}_t(s,a|z)$$

$$\leq 2 \sum_{h=k(w)}^{k(s)-1} \sum_{(x,y,z) \in T_h} q_t(x,y|w)B_{i(t)}(x,y,z)$$

where the second line uses $\bar{q}_t(s,a|z) \leq 1$.

Combining these inequalities yields (under the event $\mathcal{A}$)

$$|q_t(s,a) - \bar{q}_t(s,a)|$$

$$\leq 4 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u,v)B_{i(t)}(u,v,w)q_t(s,a|w).$$

68
\[+ 4 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v)B_{i(t)}(u, v, w) \left( \sum_{h=k(w)}^{k(s)-1} \sum_{(x,y,z) \in T_h} q_t(x, y|w)B_{i(t)}(x, y, z) \right)\]
\[\leq 16 \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) \sqrt{P(w|u, v) \ln \frac{\ln t}{\max \{m_i(t)(u, v), 1\}}} q_t(s, a|w)\]
\[+ \frac{160}{3} \sum_{k=0}^{k(s)-1} \sum_{(u,v,w) \in T_k} q_t(u, v) \cdot \frac{P(w|u, v) \ln \frac{\ln t}{\max \{m_i(t)(u, v), 1\}}}{\max \{m_i(t)(u, v), 1\}} \cdot q_t(s, a|w)\]
\[+ 4 \sum_{k=0}^{k(s)-1} \sum_{h=k+1}^{k(s)-1} \sum_{(x,y,z) \in T_h} q_t(u, v)B_{i(t)}(u, v, w)q_t(x, y|w)B_{i(t)}(x, y, z)\]

where the second line follows from Lemma D.3.3.

On the other hand, \(|q_t(s, a) - \tilde{q}_t(s, a)| \leq 1\) holds always. Combining the bounds of these two cases finishes the proof. \(\square\)

**Lemma D.3.12.** Algorithm 1 ensures \(N \leq 4|S||A| (\log T + 1)\) where \(N\) is the number of epochs.

**Proof.** For a fixed state-action pair \((s, a)\), let the \(i_1 \leq i_2 \leq \ldots \leq i_k\) denotes the epochs that triggered by this state-action pair, that is
\[
\{i_1, i_2, \ldots, i_k\} = \{i : i \in 1, \ldots, N, m_i(s, a) \geq \max \{1, 2 \cdot m_{i-1}(s, a)\}\}.
\]

Clearly, it holds that
\[1 = m_{i_1}(s, a), \text{ and } m_{i_k}(s, a) \geq 2m_{i_{k-1}}(s, a) \tau \in 2, \ldots, k\]

which indicates that \(m_{i_k}(s, a) \geq 2^{k-1}\). Combining with the fact that \(m_{i_k}(s, a) \leq T\), we have
\[k = \{|i_1, i_2, \ldots, i_k| \leq 4 \log T + 4\}.
\]

Taking the summation over all state-action pairs finishes the proof. \(\square\)