Adapting to Stochastic and Adversarial Losses in Episodic MDPs with Aggregate Bandit Feedback

Shinji Ito
The University of Tokyo and RIKEN
shinji@mist.i.u-tokyo.ac.jp

Kevin Jamieson
University of Washington
jamieson@cs.washington.edu

Haipeng Luo
University of Southern California
haipengl@usc.edu

Arnab Maiti
University of Washington
arnabm2@uw.edu

Taira Tsuchiya
The University of Tokyo and RIKEN
tsuchiya@mist.i.u-tokyo.ac.jp

Abstract

We study online learning in finite-horizon episodic Markov decision processes (MDPs) under the challenging *aggregate bandit feedback* model, where the learner observes only the cumulative loss incurred in each episode, rather than individual losses at each state-action pair. While prior work in this setting has focused exclusively on worst-case analysis, we initiate the study of *best-of-both-worlds* (BOBW) algorithms that achieve low regret in both stochastic and adversarial environments. We propose the first BOBW algorithms for episodic tabular MDPs with aggregate bandit feedback. In the case of known transitions, our algorithms achieve $O(\log T)$ regret in stochastic settings and $O(\sqrt{T})$ regret in adversarial ones. Importantly, we also establish matching lower bounds, showing the optimality of our algorithms in this setting. We further extend our approach to unknown-transition settings by incorporating confidence-based techniques. Our results rely on a combination of FTRL over occupancy measures, self-bounding techniques, and new loss estimators inspired by recent advances in online shortest path problems. Along the way, we also provide the first individual-gap-dependent lower bounds and demonstrate near-optimal BOBW algorithms for shortest path problems with bandit feedback.

1 Introduction

This paper considers online learning problems for finite-horizon episodic Markov decision processes (MDPs) with *aggregate bandit feedback* [Efroni et al., 2021, Cohen et al., 2021]. In this feedback model, the learner receives feedback on the *aggregate loss* in each episode, which is the sum of losses for all state-action pairs in the learner's trajectory of that episode, rather than individual losses for each state-action pair. The aggregate bandit feedback model naturally arises in various real-world applications where only trajectory-level outcomes are observable. For example, in personalized healthcare, a sequence of medical treatments is administered, but only the final patient outcome (e.g., recovery rate) is observed, without attributing effects to individual actions. Similarly, in application to the design of educational programs, students experience a curriculum composed of multiple learning activities, while feedback is typically available only in the form of an overall test score.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: Regret bounds for episodic MDPs with known transitions. Here π^* is an optimal policy and S^* is the set of states that can be reached by π^*. "TC" stands for computational time complexity and a checkmark (✓) indicates that an efficient implementation is possible.

Algorithm	Stochastic	Adversarial	TC
Bubeck et al. [2012]	$\sqrt{\lvert S\rvert^2\lvert A\rvert T\log\lvert A\rvert}$	$\sqrt{\lvert S\rvert^2\lvert A\rvert T\log\lvert A\rvert}$	
Lancewicki and Mansour [2025]	$\sqrt{\lvert S\rvert\lvert A\rvert LT\log\iota}$	$\sqrt{\lvert S\rvert\lvert A\rvert LT\log\iota}$	✓
Dann et al. [2023a], Ito et al. [2024]	$\frac{\lvert S\rvert^2\lvert A\rvert\log\lvert A\rvert\log T}{\Delta_{\min}}$	$\sqrt{\lvert S\rvert^2\lvert A\rvert T\log\lvert A\rvert}$	
This study (Tsallis entropy)	$\sum_{s\neq s_L,a\neq\pi^*(s)}\frac{\log T}{\Delta(s,a)}+\frac{L\lvert S\rvert\log T}{\Delta_{\min}}$	$\sqrt{\lvert S\rvert\lvert A\rvert LT}$	✓
This study (Log-barrier)	$\sum_{s\neq s_L,a\neq\pi^*(s)}\frac{\log T}{\Delta(s,a)}$	$\sqrt{\lvert S\rvert\lvert A\rvert LT\log T}$	✓
Lower bound	$\sum_{s\in S^*,a:\Delta(s,a)>0}\frac{\log T}{\Delta(s,a)}$	$\sqrt{\lvert S\rvert\lvert A\rvert LT}$	

In the study of online learning for episodic MDPs, two different models of the loss (or reward) function are commonly considered: the *stochastic model* and the *adversarial model*. In the stochastic setting, it is typically assumed that the loss function ℓ_t at each episode t is independently drawn from an unknown fixed distribution. In contrast, the adversarial model makes no such probabilistic assumptions and allows the loss function ℓ_t to vary arbitrarily over time. In various online learning/bandit problems, including those with individual loss feedback in episodic MDPs, it is well known that one can achieve instance-dependent $O(\mathrm{polylog}\, T)$-regret in the stochastic setting, and $\tilde{O}(\sqrt{T})$-regret in the adversarial setting, where T is the number of rounds/episodes and the $\tilde{O}(\cdot)$ notation hides logarithmic factors in parameters. However, to the best of our knowledge, prior work on episodic MDPs with aggregate bandit feedback has focused exclusively on the worst-case analysis (i.e., $O(\sqrt{T})$-bounds at best), and no algorithm is known to achieve instance-dependent $O(\mathrm{polylog}\, T)$-regret under the stochastic loss model with aggregate bandit feedback and unknown transitions.

This paper focuses on the design of algorithms that can effectively handle both stochastic and adversarial loss models. More specifically, we aim to develop a single algorithm that, without any prior knowledge about the nature of the environment, achieves $O(\mathrm{polylog}\, T)$-regret in stochastic settings and $\tilde{O}(\sqrt{T})$-regret in adversarial settings. Such algorithms are referred to as *best-of-both-worlds* (BOBW) algorithms [Bubeck and Slivkins, 2012]. While many prior works design separate algorithms tailored to either the stochastic or adversarial model, real-world applications often involve uncertainty about the true nature of the environment, making BOBW algorithms especially valuable in practice. Although BOBW algorithms have been developed for various settings—including episodic MDPs with individual loss feedback [Jin and Luo, 2020, Jin et al., 2021, 2023]—no such algorithm was known for episodic MDPs with aggregate bandit feedback with unknown transition, prior to this work.

1.1 Contribution

This paper presents the first BOBW algorithms for episodic tabular MDPs with aggregate bandit feedback and unknown transitions. More specifically, we consider layered MDPs with L-layers, and begin by considering the setting where the transition probability matrix P is known, designing an algorithm that achieves $\tilde{O}(\sqrt{T})$-regret in adversarial environments and $O(\log T)$-regret in stochastic environments. We then extend this approach to the more realistic and challenging setting where the transition matrix P is unknown, with the help of the techniques by Jin et al. [2021] for handling unknown transitions.

Our results are accomplished by combining some algorithmic frameworks including follow-the-regularized-leader (FTRL) over occupancy measures and self-bounding techniques [Wei and Luo, 2018, Zimmert and Seldin, 2021] with key ideas from the recent study by Maiti et al. [2025] on the online shortest path problem with bandit feedback. More precisely, our algorithm is inspired by their loss estimation method for the shortest path problem, which plays a central role in our design. By adopting their loss estimation approach, not only can we construct an estimator using only bandit feedback, but we also find that its second moment is well-controlled (see Lemma 4 and the subsequent

Table 2: Regret bounds for episodic MDPs with unknown transition. $\log \iota = O(\log(|S||A|T))$.

Algorithm	Stochastic	Adversarial	TC
Efroni et al. [2021]	$\sqrt{\lvert S\rvert^4\lvert A\rvert^3 LT\log\iota}$	–	
Cohen et al. [2021]	$\sqrt{(\lvert S\rvert\lvert A\rvert)^{O(1)}T}$	$\sqrt{(\lvert S\rvert\lvert A\rvert)^{O(1)}T}$	✓
Lancewicki and Mansour [2025]	$\sqrt{\lvert S\rvert^2\lvert A\rvert LT\log\iota}$	$\sqrt{\lvert S\rvert^2\lvert A\rvert LT\log\iota}$	✓
This study	$\frac{(\lvert S\rvert\lvert A\rvert)^{O(1)}\log^2\iota}{\Delta_{\min}}$	$\sqrt{(\lvert A\rvert+L)\lvert S\rvert^2\lvert A\rvert LT\log^2\iota}$	✓

Table 3: Comparison of regret bounds for online shortest path problems with bandit feedback. The quantity $c^* > 0$ represents the instance-dependent constant characterizing the asymptotic lower bound for linear bandits [Lattimore and Szepesvari, 2017], and it is known that $c^* \lesssim \frac{|E|}{\Delta_{\min}}$ as shown in Lee et al. [2021]. Here $E' = \bigcup_{v\in V\cup\{s\}} \partial_+ v \setminus \{\pi^*(v)\}$ and $\tilde{E} = \bigcup_{v\in V^*\cup\{s\}} \partial_+ v \setminus \{\pi^*(v)\}$, where $\partial_+ v \subset E$ is the set of outgoing edges from v, π^* is an optimal policy and V^* is the set of nodes reached by π^*. Please refer to Definition 2 for further details.

Algorithm	Stochastic	Adversarial	TC
Bubeck et al. [2012]	$\sqrt{\lvert E\rvert T\log\lvert\mathcal{P}\rvert}$	$\sqrt{\lvert E\rvert T\log\lvert\mathcal{P}\rvert}$	
Lattimore and Szepesvari [2017]	$c^*\log T$	–	
Lee et al. [2021]	$c^*\log T\log(\lvert\mathcal{P}\rvert T)$	$\sqrt{\lvert E\rvert T}\log(\lvert\mathcal{P}\rvert T)$	
Dann et al. [2023a], Ito et al. [2024]	$\frac{\lvert E\rvert\log\lvert\mathcal{P}\rvert\log T}{\Delta_{\min}}$	$\sqrt{\lvert E\rvert T\log\lvert\mathcal{P}\rvert}$	
This study (Tsallis entropy)	$\sum_{e\in E'}\frac{\log T}{\Delta(e)} + \frac{\lvert V\rvert^2\log T}{\Delta_{\min}}$	$\sqrt{\lvert E\rvert LT}$	✓
This study (Log-barrier)	$\sum_{e\in E'}\frac{\log T}{\Delta(e)}$	$\sqrt{\lvert E\rvert LT\log T}$	✓
Lower bound	$c^*\log T \gtrsim \sum_{e\in\tilde{E}}\frac{\log T}{\Delta(e)}$	$\sqrt{\lvert E\rvert T\log\lvert\mathcal{P}\rvert/\log\lvert E\rvert}$	

discussion). This property is highly beneficial for designing BOBW algorithms. Building on this insight, we propose an efficient and nearly optimal BOBW algorithm for the online shortest path problem with bandit feedback, which naturally extends to the case of MDPs with known transitions.

However, extending this estimation idea to the unknown-transition setting requires substantial care. The new estimator contains negative terms, and thus, naively replacing the occupancy measure with its upper confidence bound, as done in prior work [Jin and Luo, 2020, Jin et al., 2021], does not necessarily yield an optimistic estimator. To address this, we carefully design the estimator so that it is optimistic in expectation and its second moment remains well-controlled. This allows us not only to effectively handle aggregate bandit feedback, but also, perhaps surprisingly, to avoid the technically involved loss-shifting technique used in prior analyses [Jin et al., 2021, 2023], thereby simplifying the overall regret analysis. We refer the reader to Section 3 for a detailed discussion of these estimator constructions.

The regret upper bounds established in this work and in prior studies are summarized in Tables 1, 2, and 3. For detailed definitions of the symbols used in the tables, we refer the reader to Section 2. "TC" stands for computational time complexity and a checkmark (✓) indicates that an efficient implementation is possible. The symbol $\iota > 0$ denotes a polynomial factor in other parameters such as T, $|S|$, and $|A|$. In Table 3 for the shortest path problem, $\mathcal{P}$ denotes the set of all directed paths, and it holds that $\log|\mathcal{P}| \lesssim \min\{|V|, L\log|E|\}$, where L is the maximum number of edges in a path of the given graph. Here, $X \lesssim Y$ means $X = O(Y)$ in this paper. Similarly, $X \gtrsim Y$ means $X = \Omega(Y)$.

As shown in Tables 1 and 3, we propose computationally efficient BOBW algorithms that achieve nearly tight regret bounds for known-transition MDPs and the online shortest path problem. We note here that the corresponding lower bounds for stochastic environments are also new contributions of this paper. The adversarial lower bounds for known-transition MDPs and online

shortest paths are due to [Lancewicki and Mansour, 2025] and [Maiti et al., 2025], respectively. For unknown-transition MDPs (Table 2), we propose the first BOBW algorithm that achieves an $O\left(\sum_{s\neq s_L, a\neq\pi^*(s)} \frac{L^4|S|\log\iota+|S||A|\log^2\iota}{\Delta(s,a)} + \frac{(L^3|S|^2)(|S|+|A|)\log\iota+L|S|^2|A|\log^2\iota}{\Delta_{\min}}\right)$-regret bound in the stochastic setting and simultaneously achieves an upper bound of $\tilde{O}\left(\sqrt{(|A|+L)|S|^2|A|LT}\right)$ in the adversarial setting. Moreover, all of our BOBW algorithms exhibit robustness to corrupted stochastic environments, achieving regret bounds of the form $O(U+\sqrt{UC})$, where U is the stochastic regret and C is the corruption level. These results are established using an argument similar to the standard self-bounding technique [Zimmert and Seldin, 2021, Jin et al., 2021, 2023].

Due to differences in the problem settings, several caveats must be taken into account when making comparisons. First, while prior work on episodic MDPs assumes that the per-step loss or reward within each episode is $O(1)$, our setting assumes that the *aggregate* loss or reward over an entire episode is $O(1)$. To align the scales, we multiply the regret bounds from prior work by a factor of $1/L$. As noted in Remark 1 in the appendix, our setting is strictly more general. In addition, some prior works [Cohen et al., 2021, Lancewicki and Mansour, 2025] consider non-layered settings, and we reinterpret their bounds in terms of the layered setting by replacing $|S|L$ with $|S|$ where appropriate. Furthermore, while we evaluate the expected regret defined in Section 2, some of the prior works [Efroni et al., 2021, Lancewicki and Mansour, 2025, Lee et al., 2021] establish high-probability regret bounds.

Tables 1 and 3 include the bounds achieved by applying algorithms developed for finite-armed linear bandits [Bubeck et al., 2012, Lattimore and Szepesvari, 2017, Lee et al., 2021, Dann et al., 2023a, Ito et al., 2024], where the feature space dimension is $|S||A|$ or $|E|$, and the number of arms is calculated as $|A|^{|S|}$ or $|\mathcal{P}|$, respectively. We note that due to the exponential number of arms in this approach, it is generally unclear whether an efficient implementation is feasible. In contrast, the other algorithms listed, including our proposed methods, can be implemented efficiently using dynamic programming or convex optimization techniques. We include additional related work in Appendix A.

2 Problem setup

2.1 Episodic Markov decision process

In this paper, we consider finite-horizon Markov decision processes (MDPs) with finite actions and finite states. The model is defined by a tuple (S, A, P), where S is the finite set of states, A is the set of actions, and $P : S \times A \times S \to [0, 1]$ is the transition function that defines the probability of moving from one state to another given an action. We assume that the state space S consists of $(L+1)$ *layers*: S can be expressed as a disjoint union as $S = \bigcup_{k=0}^{L} S_k$, where $S_0 = \{s_0\}$ (initial state), $S_L = \{s_L\}$ (terminal state), $S_k \neq \emptyset$ for $k \in [L-1]$, and $S_k \cup S_{k'} = \emptyset$ for $k \neq k'$. Transitions from the k-th layer are allowed only to the $(k+1)$-th layer, i.e., for any $k \in \{0, 1, \ldots, L-1\}$, $\sum_{s'\in S_{k+1}} P(s'|s, a) = 1$ holds for all $s \in S_k$ and $a \in A$ and $P(s'|s, a) = 0$ for all $s \in S_k$, $a \in A$ and $s' \in S \setminus S_{k+1}$. Let $k(s)$ denote the index of the layer to which the state s belongs. In *known-transition* setting, we assume that the transition function P is known to the player. On the other hand, in *unknown-transition* setting, the player does not know the transition function P and it is learned through interactions with the environment.

In episodic MDPs, the player interacts with the environment in a sequence of episodes. Before each episode $t \in [T]$, the player selects a policy $\pi_t \in \Pi := \{\pi : (S \setminus \{s_L\}) \times A \to [0, 1] \mid \sum_{a\in A} \pi(a|s) = 1\}$ and the environment selects a loss function $\ell_t : S \times A \to [0, 1]$. Each episode $t \in [T]$ consists of a sequence of time steps. The initial state s_0^t is set to s_0 for all episodes $t \in [T]$. At each time step $k \in \{0, 1, \ldots, L-1\}$, the player chooses an action $a_k^t \in A$ according to the policy π_t, i.e., a_k^t follows the distribution $\pi_t(\cdot|s_k^t)$, and state s_{k+1}^t is sampled from the transition function $P(\cdot|s_k^t, a_k^t)$ given the current state s_k^t and action a_k^t. At the end of the episode, the player observes the *aggregate loss* feedback $c_t \in [0, 1]$ such that $\mathbf{E}\left[c_t \mid ((s_k^t, a_k^t))_{k\in[L-1]}\right] = \sum_{k=0}^{L-1} \ell_t(s_k^t, a_k^t)$, which corresponds to the sum of the losses incurred at each time step in the episode. We note that the player does not observe the individual losses $\ell_t(s_k^t, a_k^t)$ for $k \in [L-1]$. All the information revealed to the player in each episode is the state-action trajectory $((s_k^t, a_k^t))_{k\in[L-1]}$ and the aggregate loss c_t. We

assume that the loss function ℓ_t is chosen so that $\sum_{k=0}^{L-1} \ell_t(s_k^t, a_k^t) \in [0,1]$ for any possible trajectory $((s_k^t, a_k^t))_{k\in[L-1]}$. For notational convenience, we suppose that $\ell_t(s_L, a)$ is set to 0 for all $a \in A$.

For a transition P, a loss function $\ell : S \times A \to \mathbb{R}$, and a policy $\pi \in \Pi$, let $Q^{P,\pi}(s,a;\ell)$ and $V^{P,\pi}(s;\ell)$ express the values of the V- and Q- functions, i.e., we set $V^{P,\pi}(s_L) = Q^{P,\pi}(s_L, a) = 0$ and recursively define

$$Q^{P,\pi}(s,a;\ell) = \ell(s,a) + \sum_{s'\in S} P(s'|s,a)V^{P,\pi}(s';\ell), \quad V^{P,\pi}(s;\ell) \quad = \sum_{a\in A} \pi(a|s)Q^{P,\pi}(s,a;\ell)$$

for $s \neq s_L$ and $a \in A$. The Q-function and V-function satisfy the equality stated in the following lemma, serves as a fundamental property that underpins both the justification of our loss estimator and the derivation of the regret upper bound.

Lemma 1. *Suppose $\bar{\ell}$ is defined by $\bar{\ell}(s,a) = Q^{\pi}(s,a;\ell) - V^{\pi}(s;\ell)$ for some $\pi \in \Pi$ and for all $s \in S$ and $a \in A$. We then have*

$$V^{\pi'}(s;\bar{\ell}) = V^{\pi'}(s;\ell) - V^{\pi}(s;\ell), \quad Q^{\pi'}(s,a;\bar{\ell}) = Q^{\pi'}(s,a;\ell) - V^{\pi}(s;\ell), \tag{1}$$

for any $\pi' \in \Pi$, $s \in S$ and $a \in A$.

In addition, we define the *occupancy measure* $q^{P,\pi} : S \times A \to [0,1]$ by

$$q^{P,\pi}(s,a) = \Pr\left[(s_k,a_k) = (s,a) \mid ((s_k,a_k))_{k=0}^{L} \text{ is sampled according to } \pi \text{ and } P\right] \tag{2}$$

for $s \in S_k$ and $a \in A$. We also denote $q^{P,\pi}(s) = \sum_{a\in A} q^{P,\pi}(s,a) = \Pr[s_k = s]$ for $s \in S_k$, for notational convenience. Let $\mathcal{Q}^P = \left\{q^{P,\pi} \mid \pi \in \Pi\right\}$ be the set of occupation measures induced by the transition P. We note that $\mathcal{Q}^P$ is a closed convex set. From the definitions of V, Q and q, we have $\mathbf{E}[c_t \mid \pi_t, \ell_t] = V^{P,\pi_t}(s_0,\ell_t) = \left\langle \ell_t, q^{P,\pi_t}\right\rangle := \sum_{k=0}^{L-1}\sum_{s\in S_k}\sum_{a\in A} \ell_t(s,a)q^{P,\pi_t}(s,a)$. Using these concepts, we define the *regret*:

$$\mathrm{Reg}_T(\pi^*) = \mathbf{E}\left[\sum_{t=1}^{T}\left(V^{P,\pi_t}(s_0,\ell_t) - V^{P,\pi^*}(s_0,\ell_t)\right)\right] = \mathbf{E}\left[\sum_{t=1}^{T}\left\langle \ell_t, q^{P,\pi_t} - q^{P,\pi^*}\right\rangle\right], \tag{3}$$

where $\mathbf{E}[\cdot]$ denotes the expectation with respect to all the randomness of the environment and the player. We also denote $\mathrm{Reg}_T = \max_{\pi^*\in\Pi}\mathrm{Reg}_T(\pi^*)$. Hereafter, when it is clear from the context, we may omit P and π for simplicity. Additionally, for notational convenience, we denote $Q_t^{\pi}(\cdot) = Q^{P,\pi}(\cdot;\ell_t)$, $V_t^{\pi}(\cdot) = V^{P,\pi}(\cdot;\ell_t)$, $q_t = q^{P,\pi_t}$, and $q^* = q^{P,\pi^*}$.

2.2 Regime of environments

We consider several different regimes as models for the environment generating the loss function ℓ_t. In an *adversarial regime*, ℓ_t can be chosen arbitrarily by an adversary, depending on the history $((s_k^\tau, a_k^\tau))_{k\in\{0,\dots,L-1\},\tau\in[t-1]}$ so far and the policy π_t chosen by the player. In a *stochastic regime*, the loss function ℓ_t is independently and identically drawn from an unknown distribution for each episode $t \in [T]$. For a stochastic environment, let $\ell^* : S \times A \to [0,1]$ denote the expected loss function, which is defined as $\ell^*(s,a) = \mathbf{E}[\ell_t(s,a)]$ for all $s \in S$ and $a \in A$. We then have $\mathrm{Reg}_T = \mathbf{E}\left[\sum_{t=1}^{T}\langle \ell^*, q_t - q^*\rangle\right] = \mathbf{E}\left[\sum_{t=1}^{T}\sum_{s\in S\setminus\{s_L\}}\sum_{a\in A}\Delta(s,a)q_t(s,a)\right]$, where $q^* \in \arg\min_{q\in\mathcal{Q}}\langle \ell^*, q\rangle$ and $\Delta : S \times A \to [0,1]$ is defined as $\Delta(s,a) = Q^{\pi^*}(s,a;\ell^*) - V^{\pi^*}(s;\ell^*) = Q^{\pi^*}(s,a;\ell^*) - \min_{a'\in A} Q^{\pi^*}(s,a';\ell^*)$ for an optimal policy π^* that minimizes $V^{\pi}(s_0;\ell^*)$. We note that the optimal policy π^* is unique if and only if the set $\arg\min_{a\in A}\{Q^{\pi^*}(s,a;\ell^*)\} = \{a \in A \mid \Delta(s,a) = 0\}$ consists of a single action for all $s \in S$. Here, as a generalization of a stochastic regime admitting a unique optimal policy, we can define the an adversarial regime with self-bounding constraints:

Definition 1 (self-bounding regime for MDPs)**.** Let $\pi^* : S \to A$ be a deterministic policy. Suppose that $\Delta : S \times A \to [0,1]$ satisfies $\Delta(s,a) > 0$ for all $s \in S \setminus \{s_L\}$ and $a \in A \setminus \{\pi^*(s)\}$. Let $C \geq 0$. The environment is in an adversarial regime with a (π^*, Δ, C) self-bounding constraint (or, more concisely, a (π^*, Δ, C)-self-bounding regime) if it holds for any algorithm that

$$\mathrm{Reg}_T(\pi^*) \geq \mathbf{E}\left[\sum_{t=1}^{T}\sum_{s\in S\setminus\{s_L\}}\sum_{a\in A\setminus\{\pi^*(s)\}}\Delta(s,a)q_t(s,a)\right] - C. \tag{4}$$

We also denote $\Delta_{\min} = \min_{s \neq s_L, a \neq \pi^*(s)} \Delta(s, a)$. As discussed by Zimmert and Seldin [2021], this is a general regime that includes stochastic environments with adversarial corruption, where the parameter C corresponds to the total amount of corruption. For more details, see, e.g., [Zimmert and Seldin, 2021, Jin and Luo, 2020, Jin et al., 2021].

2.3 Online shortest path problem

In online shortest path problem, the player is given a directed acyclic graph (DAG) $G = (V \cup \{s, g\}, E)$, where s and g are the source and target vertices, respectively, V is the set of other vertices, and E is the set of edges. Denote $m = |E|$ and $n = |V|$. Let L denote the maximum number of edges in a s-g path in G. In each round $t \in [T]$, the environment chooses a loss function $\ell_t : E \to [0, 1]$ and the player chooses a path p_t from s to g. At the end of the round, the player can only observe the aggregate loss feedback $c_t \in [0, 1]$ such that $\mathbf{E}[c_t \mid \ell_t, p_t] = \sum_{e \in p_t} \ell_t(e)$, while the player does not observe the individual losses $\ell_t(e)$ for $e \in E$. The definition of regret, the regimes of the environment, and other related concepts are defined in a similar way to in the case of episodic MDPs. More details on the model are given in the appendix. We also note that the online shortest path problem can be interpreted as an "almost" special case of episodic MDPs with known transitions, but it is not necessarily an exact special case. For details, see Remark 2 in the appendix.

3 Core idea: construction of loss estimator with aggregate feedback

A key aspect of the proposed algorithms lies in how to estimate the loss function in a setting where only the limited aggregate loss feedback is available. In this paper, inspired by the approach of Maiti et al. [2025] to the online shortest path problem with bandit feedback, we extend the idea to the MDP setting. We begin by briefly reviewing their approach.

3.1 Review of the approach of Maiti et al. [2025] for the online shortest path problem

The online shortest path algorithm by Maiti et al. [2025] maintains an s-g flow $q_t \in \mathcal{Q} \subseteq [0, 1]^E$ of capacity 1. Note that $\mathcal{Q}$ can be interpreted as a convex hull of the set of all s-g paths. In the following, we denote $q(v) = \sum_{e \in \partial_+ v} q(e)$ for any $q \in [0, 1]^E$ and $v \in V \cup \{s\}$, where $\partial_+ v \subset E$ is the set of outgoing edges from v. From q_t, it samples a path $p_t \in \{0, 1\}^E$ in a *Markovian* way, i.e., we choose a path as follows: (i) We first initialize $p_t \in \{0, 1\}^E$ by $p_t(e) = 0$ for all $e \in E$ and set $v \leftarrow s$. (ii) While $v \neq g$, Pick $e \in \partial_+ v$ with probability $\frac{q_t(e)}{q_t(v)}$, set $p_t(e) \leftarrow 1$, and transition to e's terminal vertex e_+, i.e., set $v \leftarrow e_+$. We then have $\mathbf{E}[p_t \mid q_t] = q_t$, i.e., each edge $e \in E$ is included in the path p_t with probability $q_t(e)$.

After constructing the path p_t as described above and obtaining the aggregate loss feedback c_t such that $\mathbf{E}[c_t \mid \ell_t, p_t] = \langle \ell_t, p_t \rangle$, the loss estimator $\widehat{L}_t(p)$ for any s-g path $p = (s = v_0, e_0, v_1, e_1, \ldots, v_{L-1}, e_{L-1}, v_L = g)$ is defined as follows:[1]

$$\widehat{L}_t(p) = \sum_{k=0}^{L-1} \frac{p_t(e_k)}{q_t(e_k)} c_t - \sum_{k=1}^{L-1} \frac{p_t(v_k)}{q_t(v_k)} c_t = c_t \cdot \left(\sum_{e \in E} \frac{p(e) p_t(e)}{q_t(e)} - \sum_{v \in V} \frac{p(v) p_t(v)}{q_t(v)} \right). \tag{5}$$

We then have $\mathbf{E}\left[\widehat{L}_t(p) \mid q_t, \ell_t\right] = \langle \ell_t, p \rangle$ for any s-g path p. In fact, the conditional expectation given q_t, ℓ_t satisfies

$$\mathbf{E}\left[\frac{p_t(e_k)}{q_t(e_k)} c_t\right] = \mathbf{E}[\langle \ell_t, p_t \rangle \mid p_t(e_k) = 1] = \bar{L}_t(s \to v_k) + \ell_t(e_k) + \bar{L}_t(v_{k+1} \to g), \tag{6}$$

$$\mathbf{E}\left[\frac{p_t(v_k)}{q_t(v_k)} c_t\right] = \mathbf{E}[\langle \ell_t, p_t \rangle \mid p_t(v_k) = 1] = \bar{L}_t(s \to v_k) + \bar{L}_t(v_k \to g), \tag{7}$$

where $\bar{L}_t(v \to v')$ represents the conditional expectation of the cost of the subpath of p_t from v to v', given that p_t goes through v and v'. By combining (5), (6) and (7), we obtain $\mathbf{E}\left[\widehat{L}_t(p) \mid q_t, \ell_t\right] =$

[1] The construction method by Maiti et al. [2025] does not exactly match the one described below, as they add a uniform shift and incorporate implicit exploration Neu [2015]. These adjustments are designed to obtain high-probability regret bounds, but they are not essential in this study, which focuses on expected regret bounds.

$\sum_{k=1}^{L-1} \ell_t(e_k) = \langle \ell_t, p \rangle$. In this paper, we define the loss estimator $\widehat{\ell}_t \in \mathbb{R}^E$ by $\widehat{\ell}_t(e) = c_t \cdot \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(e_-)}{q_t(e_-)} \right)$, where $e_- \in V \cup \{s\}$ is the initial vertex of the edge $e \in E$. As we have $\left\langle \widehat{\ell}_t, p \right\rangle = \widehat{L}_t(p) - c_t$ for any s-g path p, we can use $\widehat{\ell}_t$ as an loss estimator in our FTRL framework.

3.2 Loss estimator for MDPs with known transition

Let $q_t \in \mathcal{Q}$ be the occupancy measure for the policy π_t. Suppose that the trajectory $((s_k^t, a_k^t))_{k=0}^{L-1}$ is generated according to the policy π_t and c_t is the observed aggregate loss feedback. For any $k \in \{0, 1, \ldots, L-1\}$ and for any $s \in S_k$ and $a \in A$, denote $\mathbb{I}_t(s) = \mathbb{I}[s_k^t = s]$ and $\mathbb{I}_t(s,a) = \mathbb{I}[(s_k^t, a_k^t) = (s,a)]$. Note that we then have $\mathbf{E}\left[\mathbb{I}_t(s) \mid \pi_t\right] = q_t(s)$ and $\mathbf{E}\left[\mathbb{I}_t(s,a) \mid \pi_t\right] = q_t(s,a)$. Inspired by the approach of Maiti et al. [2025], we define the loss estimator as in the following lemma:

Lemma 2. *The loss estimator $\widehat{\ell}_t : S \times A \to \mathbb{R}$ defined as $\widehat{\ell}_t(s,a) = c_t \cdot \left(\frac{\mathbb{I}_t(s,a)}{q_t(s,a)} - \frac{\mathbb{I}_t(s)}{q_t(s)} \right)$ satisfies*

$$\mathbf{E}\left[\widehat{\ell}_t(s,a) \mid \ell_t, \pi_t \right] = Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t) =: \bar{\ell}_t(s,a). \tag{8}$$

From this and Lemma 1, we have $\mathrm{Reg}_T(\pi^*) = \mathbf{E}\left[\sum_{t=1}^T \left(V^{\pi_t}(s_0; \widehat{\ell}_t) - V^{\pi^*}(s_0; \widehat{\ell}_t) \right) \right] = \mathbf{E}\left[\sum_{t=1}^T \left\langle \widehat{\ell}_t, q_t - q^* \right\rangle \right]$. Thanks to this, we can use $\widehat{\ell}_t$ instead of ℓ_t in our FTRL-based algorithms.

3.3 Loss estimator for MDPs with unknown transition

In the case of unknown transitions, when attempting to construct a loss estimator in the same manner as in Lemma 2, a key difficulty arises from the fact that the true value of q_t is not available. To address this issue, one may follow the approach of Jin et al. [2020] and compute an upper confidence bound u_t for q_t, using it as a surrogate in the estimator. However, naively replacing q_t with u_t in the definition of $\widehat{\ell}_t$ in Lemma 2 introduces yet another issue. Specifically, as $\widehat{\ell}_t$ in Lemma 2 contains a negative term (i.e., $-c_t \frac{\mathbb{I}_t(s)}{q_t(s)}$), substituting q_t with its upper bound u_t may lead to an undesirable *positive bias* in the estimator, which creates an obstacle in the regret analysis. To derive a valid regret upper bound, it is essential that the estimator is *optimistic*, i.e., its expectation must act as a lower confidence bound on $\bar{\ell}_t$ in (8). To this end, we define the following novel loss estimator:

$$\ell_t^u(s,a) = \frac{c_t \cdot \mathbb{I}_t(s,a) + (1 - \pi_t(a \mid s) - c_t) \cdot \mathbb{I}_t(s) \pi_t(a \mid s)}{u_t(s,a)} - (1 - \pi_t(a \mid s)). \tag{9}$$

We can evaluate the expectation of ℓ_t^u in a manner similar to the proof of Lemma 2, as follows:

$$\mathbf{E}\left[\ell_t^u(s,a) \mid \ell_t, \pi_t, u_t \right] = \frac{q_t(s)}{u_t(s)} \left(Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t) + 1 - \pi_t(a|s) \right) - (1 - \pi_t(a|s)). \tag{10}$$

We here have $Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t) = Q^{\pi_t}(s,a;\ell_t) - \sum_{a' \in A} \pi_t(a' \mid s) Q^{\pi_t}(s,a';\ell_t) = (1 - \pi_t(a|s)) Q^{\pi_t}(s,a;\ell_t) - \sum_{a' \neq a} \pi_t(a' \mid s) Q^{\pi_t}(s,a';\ell_t) \geq -(1 - \pi_t(a|s))$ as $Q^{\pi_t}(s,a;\ell_t) \in [0,1]$, and thus: $Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t) + 1 - \pi_t(a|s) \geq 0$. Hence, under the condition of $u_t(s) \geq q_t(s)$, the value of (10) is a lower bound on $\bar{\ell}_t(s,a) := Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t)$, i.e., $\ell_t^u(s,a)$ is an optimistic estimator of $\bar{\ell}_t(s,a)$. In addition, the gap between them is at most $\frac{u_t(s) - q_t(s)}{u_t(s)}(1 - \pi_t(a|s))$:

Lemma 3. *Under the condition of $u_t(s) \geq q_t(a)$, we have*

$$\bar{\ell}_t(s,a) - \frac{u_t(s) - q_t(s)}{u_t(s)}(1 - \pi_t(a|s)) \leq \mathbf{E}\left[\ell_t^u(s,a) \mid \ell_t, \pi_t, u_t \right] \leq \bar{\ell}_t(s,a). \tag{11}$$

3.4 Second moment of loss estimators

Our proposed algorithm, like those of Jin et al. [2020], Jin and Luo [2020], Jin et al. [2021], is based on the Follow-the-Regularized-Leader (FTRL) framework over occupancy measures. In this framework, the second moment of the loss estimator plays a crucial role. The second moment of the loss estimator introduced in this section can be bounded as follow:

Lemma 4. *Loss estimators $\widehat{\ell}_t$ in Lemma 2 and ℓ_t^u in (9) satisfy*

$$\mathbf{E}\left[\widehat{\ell}_t(s,a)^2 \mid \ell_t, \pi_t\right] \le \frac{1-\pi(a|s)}{q_t(s,a)}, \quad \mathbf{E}\left[\ell_t^u(s,a)^2 \mid \ell_t, \pi_t, u_t\right] \lesssim \frac{1-\pi(a|s)}{u_t(s,a)} \cdot \left(\frac{q_t(s)}{u_t(s)} + 1\right).$$

When applying the self-bounding technique to derive an $O(\mathrm{polylog}\, T)$ regret bound, the $(1-\pi(a|s))$ factor in this lemma plays a crucial role. In prior work [Jin et al., 2021, 2023], the original loss estimator did not exhibit this factor in its second moment, and hence the analysis relied on a carefully designed *shifting function* to apply a *loss-shifting trick* and extract the desired $(1-\pi(a|s))$ factor. However, this significantly complicated the analysis. In contrast, our regret analysis does not require the loss-shifting trick, as the self-bounding technique can be applied directly. As a result, we avoid the technically involved analysis necessitated by the loss-shifting trick in previous work.

4 Algorithm and regret bounds

4.1 Warmup: online shortest path problem

As a warm-up, let us consider the algorithm for the online shortest path problem. Following the approach of Maiti et al. [2025], we update a point q_t on the s-g unit flow polytope $\mathcal{Q}$ (i.e., the convex hull of all s-g paths) using the following FTRL framework: $q_t \in \arg\min_{q\in\mathcal{Q}} \left\{\left\langle \sum_{\tau=1}^{t-1} \widehat{\ell}_\tau, q\right\rangle + \psi_t(q)\right\}$, where $\widehat{\ell}_\tau$ is given as in Section 3.1 and $\psi_t(q)$ is a regularizer function defined as:

$$\psi_t(q) = -\frac{2}{\eta_t}\sum_{e\in E}\sqrt{q(e)} - \sum_{e\in E}\beta \ln q(e) \quad \text{with} \quad \eta_t = \frac{1}{\sqrt{t}},\ \beta = \Theta(1), \qquad \text{(Tsallis entropy)}$$

$$\psi_t(q) = -\sum_{e\in E}\frac{1}{\eta_t(e)}\ln q(e) \quad \text{with} \quad \eta_t(e) = \Big(4 + \frac{1}{\ln T}\sum_{\tau=1}^{t-1}\rho_\tau(e)\Big)^{-\frac{1}{2}}, \qquad \text{(log-barrier)}$$

where we define $\rho_t(e) = c_t^2 p_t(e)\big(1 - \frac{q_t(e)}{q_t(e_-)}\big)^2$. We then have the following regret upped bounds:

Theorem 1. *Let $p^* \in \{0,1\}^E$ be an arbitrary s-g path and $\pi^* : V \cup \{s\} \to E$ be such that $\pi^*(v) \in \partial_+ v$ for all $v \in V \cup \{s\}$ and $p^*(e) = 1 \Longrightarrow \pi^*(e_-) = e$. In the case of Tsallis entropy,*

$$\mathrm{Reg}_T(p^*) \lesssim \sum_{t=1}^T \frac{1}{\sqrt{t}}\mathbf{E}\Big[\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\{\pi^*(v)\}}\sqrt{q_t(e)} + \sum_{v\in V\setminus V^*}\sqrt{q_t(\pi^*(v))}\Big] + \sqrt{mnL} + m\log T.$$

If ψ_t is given by log-barrier regularizer, we have

$$\mathrm{Reg}_T \lesssim \mathbf{E}\Big[\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v}\sqrt{\textstyle\sum_{t=1}^T c_t^2 p_t(e)(1-\frac{q_t(e)}{q_t(v)})^2\log(T)}\Big] + m\log(T).$$

Corollary 1. *We have $\mathrm{Reg}_T \lesssim \sqrt{mL(n+T)} + n\log T$ in the Tsallis entropy case and $\mathrm{Reg}_T \lesssim \sqrt{mL(n+T)\log T}$ in the log-barrier case. Simultaneously, under the condition of $\mathrm{Reg}_T(p^*) \ge \mathbf{E}\left[\sum_{t=1}^T\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\{\pi^*(v)\}}\Delta(e)p_t(e)\right] - C$ for some $\Delta \in [0,1]^E$ and $C \ge 0$, we have $\mathrm{Reg}_T(p^*) \lesssim U + \sqrt{UC}$, where $U = \sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\{\pi^*(v)\}}\frac{\log T}{\Delta(e)} + \frac{n^2\log T}{\Delta_{\min}}$ in the Tsallis entropy case and $U = \sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\{\pi^*(v)\}}\frac{\log T}{\Delta(e)}$ for the log-barrier case.*

The tightness of the gap-dependent upper bound derived here is discussed in the appendix.

4.2 MDPs with known transition

The algorithm design for episodic MDPs with known transitions is almost identical to the case of the shortest path problem. Specifically, we apply the FTRL framework over the set of all occupancy measures as the feasible region, replace each edge $e \in E$ in the regularization functions in (Tsallis entropy) and (log-barrier) with a state-action pair $(s,a) \in S \subseteq \{s_L\} \times A$, and redefine ρ_t as $\rho_t(s,a) = c_t^2 \mathbb{I}_t(s,a)\,(1-\pi_t(a \mid s))^2$. With this setup, we obtain the following regret bound:

Corollary 2. *We have* $\mathrm{Reg}_T \lesssim \sqrt{|S||A|LT} + |S||A| \log T$ *in the Tsallis entropy case and* $\mathrm{Reg}_T \lesssim \sqrt{|S||A|LT \log T}$ *in the log-barrier case. Simultaneously, under the condition of* (4)*, we have* $\mathrm{Reg}_T(\pi^*) \lesssim U + \sqrt{UC}$*, where* $U = \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{\log T}{\Delta(s,a)} + \frac{L|S| \log T}{\Delta_{\min}}$ *in the Tsallis entropy case and* $U = \sum_{s \neq s_L} \sum_{a \neq \pi^*(s)} \frac{\log T}{\Delta(s,a)}$ *for the log-barrier case.*

The gap-dependent upper bound achieved by the log-barrier regularization in this corollary is tight. In fact, the following lower bound holds:

Theorem 2. *Consider stochastic environment in which* c_t *follows a Bernoulli distribution of parameter* $\sum_{k=1}^{L-1} \ell^*(s_t^k, a_t^k)$ *where we assume that this value is in* $[3/8, 5/8]$ *for any possible trajectories. Define* $\Delta : S \times A \to [0,1]$ *by* $\Delta(s,a) = Q^{\pi^*}(s,a;\ell^*) - V^{\pi^*}(s;\ell^*)$ *for an optimal policy* π^**. Let* S^* *be the set of all states* $s \in S \setminus \{s_L\}$ *such that* $q^{\pi^*}(s) > 0$ *for some optimal policy* π^**. Then, for any consistent algorithms, we have* $\liminf_{T \to \infty} \frac{\mathrm{Reg}_T}{\log T} \gtrsim \sum_{s \in S^*} \sum_{a \in A : \Delta(s,a) > 0} \frac{1}{\Delta(s,a)}$.

4.3 MDPs with unknown transition

Our proposed algorithm for MDPs with unknown transitions adopts an epoch-based approach, similar to prior work [Jin et al., 2021, 2020]. In each epoch i, it updates both the transition probability estimates and their corresponding confidence intervals. Based on these, we compute an upper confidence bound u_t on the occupancy measure q_t, and use it to construct the loss estimator ℓ_t^u as defined in (9). Note that $u_t(s,a)$ can be efficiently computed using the COMP-UOB procedure proposed by [Jin et al., 2020].

We then define the adjusted loss estimator as $\widehat{\ell}_t := \ell_t^u - B_i$, where B_i is a bonus term derived from the confidence width. Unlike prior work, our choice of the loss estimator ℓ_t^u allows us to avoid scaling B_i by an additional factor of L. The policy for each episode is selected by applying FTRL over the estimated occupancy measure space using $\widehat{\ell}_t$. The regularization function used here matches the Tsallis entropy regularizer from Section 4.2, except that the learning rate η_t is reset at the beginning of each epoch, and a small log-barrier term is added to stabilize updates.

A notable improvement over prior work [Jin et al., 2021] is that the second-moment bound established in Lemma 4 allows us to bypass the loss-shifting technique. As a result, our regret bounds exhibit improved dependence on the horizon L. We refer the reader to Algorithm 1 and the appendix E for full details. Our algorithm, constructed in this way, achieves the following upper bounds:

Theorem 3. *In the bandit feedback setting, Algorithm 1 with* $\delta = \frac{1}{T^3}$ *and* $\iota = \frac{|S||A|T}{\delta}$ *guarantees* $\mathrm{Reg}_T(\pi^\star) = \tilde{\mathcal{O}}\left(L|S|\sqrt{|A|T} + |S||A|\sqrt{LT} + L^2|S|^3|A|^2\right)$ *and simultaneously* $\mathrm{Reg}_T(\pi^\star) = \mathcal{O}\left(U + \sqrt{UC} + V\right)$ *under Condition* (4)*, where* $V = L^2|S|^3|A|^2 \ln^2 \iota$ *and* U *is defined as*

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

We defer the proof of the above theorem to Appendix F.

5 Conclusion

This paper initiated the study of best-of-both-worlds (BOBW) algorithms for finite-horizon episodic MDPs with aggregate bandit feedback. We proposed efficient algorithms that achieve low regret in both stochastic and adversarial settings, and established nearly tight upper and lower bounds under both known- and unknown-transition settings. Our approach is built upon FTRL over occupancy measures, combined with carefully designed loss estimators that are optimistic in expectation and admit tight second-moment bounds.

Despite these contributions, many open questions remain. A central limitation of our approach is its reliance on occupancy measure updates via FTRL, which—while grounded in convex optimization and thus computationally feasible to some extent—still requires solving a convex problem in each

Algorithm 1 BOBW algorithm for MDPs with unknown transitions and aggregate bandit feedback

Input: confidence parameter $\delta \in (0,1)$
Initialize: epoch index $i = 1$ and epoch starting time $t_i = 1$;
$\forall (s,a,s')$, set counters $m_1(s,a) = m_1(s,a,s') = m_0(s,a) = m_0(s,a,s') = 0$;
empirical transition $\bar{P}_1$ and confidence width B_1 based on Eq. (2);
for $t = 1, \dots, T$ **do**
    Let ϕ_t be the regularizer defined in Eq. (63) and compute

$$\widehat{q}_t = \operatorname*{arg\,min}_{q \in \Omega(\bar{P}_i)} \textstyle\sum_{\tau=t_i}^{t-1} \langle q, \widehat{\ell}_t \rangle + \phi_t(q),$$

    where $\widehat{\ell}_t = \ell_\tau^u - B_i$ and $B_i(s,a) = \min\{2, \sum_{s' \in S_{k(s)+1}} B_i(s,a,s')\}$.
    Compute policy π_t from $\widehat{q}_t$ such that $\pi_t(a|s) \propto \widehat{q}_t(s,a)$.
    Execute policy π_t and obtain trajectory $(s_{t,k}, a_{t,k})$ for $k = 0, \dots, L-1$.
    Construct loss estimator ℓ_t^u as defined in Eq. (9).
    Increment counters: for each $k < L$,

$$m_i(s_{t,k}, a_{t,k}, s_{t,k+1}) \leftarrow m_i(s_{t,k}, a_{t,k}, s_{t,k+1}) + 1, \quad m_i(s_{t,k}, a_{t,k}) \leftarrow m_i(s_{t,k}, a_{t,k}) + 1.$$

    if $\exists k,\ m_i(s_{t,k}, a_{t,k}) \geq \max\{1, 2m_{i-1}(s_{t,k}, a_{t,k})\}$ **then** ▷ entering a new epoch
        Increment epoch index $i \leftarrow i + 1$ and set new epoch starting time $t_i = t + 1$.
        Initialize new counters:

$$\forall (s,a,s'), m_i(s,a,s') = m_{i-1}(s,a,s'), \quad m_i(s,a) = m_{i-1}(s,a).$$

        Update empirical transition $\bar{P}_i$ and confidence width B_i based on Eq. (57) and (58).

round. Moreover, this framework does not easily extend beyond tabular MDPs to more general representations such as linear models or function approximation.

A promising direction to address these limitations is to adopt policy optimization-based methods [Shani et al., 2020, Luo et al., 2021]. In particular, a recent paper by [Lancewicki and Mansour, 2025] has shown that near-optimal and efficient adversarial regret bounds can be achieved through policy optimization. Combining this line of work with the techniques in [Dann et al., 2023a] may yield BOBW algorithms that are both computationally efficient and more broadly applicable.

Another important direction for future research is to extend the present results beyond the online shortest path problem to other combinatorial optimization settings, or to more challenging MDP formulations such as stochastic shortest path problems [Chen et al., 2021]. Addressing these challenges may lead to a more comprehensive understanding of online learning under aggregate feedback.

Acknowledgments

Ito is supported by JSPS KAKENHI Grant Number JP25K03184. Jamieson is funded in part by NSF Award CAREER 2141511 and Microsoft Grant for Customer Experience Innovation. Luo is funded by NSF award IIS-1943607. Tsuchiya is supported by JSPS KAKENHI Grant Number JP24K23852.

References

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 42.1–42.23, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, volume 23, pages 41.1–41.14, 2012.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 111–127. PMLR, 2018.

Clément L Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2022.

Asaf Cassel, Haipeng Luo, Aviv Rosenberg, and Dmitry Sotnikov. Near-optimal regret in linear MDPs with aggregate bandit feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 5757–5791. PMLR, 2024.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. In *Advances in Neural Information Processing Systems*, volume 34, pages 3401–3412. Curran Associates, Inc., 2021.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory*, pages 1180–1215. PMLR, 2021.

Alon Cohen, Haim Kaplan, Tomer Koren, and Yishay Mansour. Online Markov decision processes with aggregate bandit feedback. In *Conference on Learning Theory*, pages 1301–1329. PMLR, 2021.

Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5503–5570. PMLR, 2023a.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. Best of both worlds policy optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 6968–7008. PMLR, 2023b.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35-8, pages 7288–7295, 2021.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 34, pages 28511–28521. Curran Associates, Inc., 2021.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pages 2552–2583. PMLR, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 28631–28643. Curran Associates, Inc., 2022.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2522–2563. PMLR, 2024.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems*, volume 33, pages 16557–16566. Curran Associates, Inc., 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. In *Advances in Neural Information Processing Systems*, volume 34, pages 20491–20502. Curran Associates, Inc., 2021.

Tiancheng Jin, Junyan Liu, Chloé Rouyer, William Chang, Chen-Yu Wei, and Haipeng Luo. No-regret online reinforcement learning with adversarial losses and transitions. In *Advances in Neural Information Processing Systems*, volume 36, pages 38520–38585. Curran Associates, Inc., 2023.

Tal Lancewicki and Yishay Mansour. Near-optimal regret using policy optimization in online MDPs with aggregate bandit feedback. *arXiv preprint arXiv:2502.04004*, 2025.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

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

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151. PMLR, 2021.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

Arnab Maiti, Zhiyuan Fan, Kevin Jamieson, Lillian J. Ratliff, and Gabriele Farina. Efficient near-optimal algorithm for online shortest paths in directed acyclic graphs with bandit feedback against adaptive adversaries, 2025.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. In *Advances in Neural Information Processing Systems*, volume 37, pages 141071–141102. Curran Associates, Inc., 2024.

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

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. PMLR, 2023a.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3285–3294. PMLR, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Theorems match the claims made in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: Limitations and future works are discussed in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Every assumption on our problem setting is clearly mentioned in the introduction.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper is theoretical in nature.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our paper is theoretical in nature.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper is theoretical in nature.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our paper is theoretical in nature.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper is theoretical in nature.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper is theoretical in nature.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper is theoretical in nature.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Contents

A Additional related work

FTRL and best-of-both-worlds algorithms In episodic tabular MDPs with adversarial losses, Jin and Luo [2020] is the first to propose a BOBW algorithm under known transitions. Jin et al. [2021] extended this to the unknown-transition setting, which is further improved and extend to the setting where the transition can vary over episodes [Jin et al., 2023]. Subsequently, policy optimization algorithms was shown to achieve BOBW guarantees with improved computational efficiency [Dann et al., 2023b]. In spite of these developments, our work is the first to consider BOBW algorithms under aggregate feedback. We build on the analysis of Jin et al. [2021]. While we may improve our bounds by using the optimistic transition technique from the recent work by Jin et al. [2023], it remains unclear whether this technique can be effectively combined with our loss estimator, which can be negative.

A key challenge in achieving BOBW is the design and analysis of the regularizer in FTRL. In this work, we consider two types of regularizers (see Section 4.1). The first one is a hybrid regularizer [Bubeck et al., 2018] that combines the Tsallis entropy with a small-coefficient log-barrier. The first BOBW algorithm based on the Tsallis entropy was initially explored by Zimmert and Seldin [2021], and the hybrid regularizers to ensure the stability of FTRL have been shown to be powerful in obtaining BOBW guarantees for complex online learning problems or for obtaining adaptive bounds [Zimmert and Seldin, 2020, Erez and Koren, 2021, Ito et al., 2022, 2024, Tsuchiya et al., 2023a,b, Masoudian et al., 2024]. The second regularizer we consider is the log-barrier regularizer with adaptive learning rates, developed in Wei and Luo [2018], Ito [2021]. As we show in this paper, although the strong regularization of the log-barrier can introduce an additional $O(\sqrt{\log T})$ multiplicative factor in adversarial settings, it ensures a strong stability of FTRL.

Episodic MDPs with aggregate feedback. Recently, MDPs with aggregate feedback have received growing attention. For example, in tabular MDPs, aggregate feedback has been studied in both the stochastic and adversarial settings (see [Efroni et al., 2021, Cohen et al., 2021, Chatterji et al., 2021]), as well as in the context of policy optimization [Lancewicki and Mansour, 2025]. Similar interest has emerged for linear MDPs as well (see [Cassel et al., 2024]). However, to the best of our knowledge, our work is the first to study best-of-both-worlds guarantees in the setting of aggregate feedback.

Remarks in comparing results

Remark 1 (On the scale of loss)**.** In existing studies on online learning for MDPs with adversarial losses, it's common to assume that $\ell_t(s,a) \in [0,1]$ for all s and a. Such setting is reduced to our

setting by scaling the losses by a factor of $1/L$, which therefore can be regarded as a special case of our setting. If the loss is given by this reduction, values of losses have the same scale of $O(1/L)$ for all layers. In contrast, in our setting, the losses may have different scales (possibly $> 1/L$) for each layer, which can be interpreted to be more general.

Remark 2 (On online shortest path problems and episodic MDPs)**.** The online shortest path problem can be interpreted as an "almost" special case of episodic MDPs with known transitions, but it is not necessarily an exact special case. Intuitively, the vertices $v \in V$ in the shortest path problem correspond to the states $s \in S$ in an MDP, and selecting one outgoing edge $e \in \partial_+ v$ from a vertex v corresponds to choosing an action $a \in A$ in the MDP. The shortest path problem however differs in several aspects: the set of vertices V does not necessarily have a hierarchical structure, the number of edges in a path from the source to the sink is not necessarily fixed, and the set $\partial_+ v$ of outgoing edges available for selection can vary depending on the vertex v. Therefore, the shortest path problem cannot always be directly interpreted as an MDP. Consequently, regret bounds for MDPs do not immediately translate to results for the shortest path problem. We therefore provide a separate discussion on the online shortest path problem. However, the overall framework for algorithm design and analysis is similar to that for episodic MDPs with known transitions.

B Auxiliary lemmas

B.1 Lemmas for FTRL

B.1.1 Stability terms for one-dimensional functions

Lemma 5. *Let $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be defined as $\phi(x) = -2\sqrt{x}$ and $D_\phi(y, x)$ be the Bregman divergence associated with ϕ, i.e.,*

$$D_\phi(y, x) = -2\sqrt{y} + 2\sqrt{x} + \frac{1}{\sqrt{x}}(y - x) = \frac{1}{\sqrt{x}}\left(\sqrt{x} - \sqrt{y}\right)^2.$$

Then, for any $x \in (0, 1)$, $\ell \in \mathbb{R}$ and $\eta > 0$ such that $\eta\sqrt{x}\ell > -1$, we have

$$\sup_{y \in [0,1]} \left\{ \ell \cdot (x - y) - \frac{1}{\eta} D_\phi(y, x) \right\} \leq \frac{\eta x^{3/2} \ell^2}{1 + \eta\ell\sqrt{x}}.$$

Proof. We have

$$\begin{aligned}
&\ell \cdot (x - y) - \frac{1}{\eta} D_\phi(y, x) = \frac{1}{\eta}\left(2\sqrt{y} - 2\sqrt{x} + \left(\frac{1}{\sqrt{x}} + \eta\ell\right)(x - y)\right) \\
&= \frac{1}{\eta}\left(-y\left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-1}\left(\left(\frac{1}{\sqrt{x}} + \eta\ell - \frac{1}{\sqrt{y}}\right)^2 - \frac{1}{y}\right) - \sqrt{x} + \eta\ell x\right) \\
&= \frac{1}{\eta}\left(-y\left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-1}\left(\frac{1}{\sqrt{x}} + \eta\ell - \frac{1}{\sqrt{y}}\right)^2 + \left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-1} - \sqrt{x} + \eta\ell x\right) \\
&\leq \frac{1}{\eta}\left(\left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-1} - \sqrt{x} + \eta\ell x\right) \\
&= \frac{\sqrt{x}}{\eta}\left(\frac{1}{1 + \eta\ell\sqrt{x}} - 1 + \eta\ell\sqrt{x}\right) = \frac{\sqrt{x}}{\eta(1 + \eta\ell\sqrt{x})}\left(\eta^2\ell^2 x\right) = \frac{\eta x^{3/2}\ell^2}{1 + \eta\ell\sqrt{x}}.
\end{aligned}$$

□

Lemma 6. *Let $\eta > 0$ and $\beta > 0$. Suppose that $\phi : \mathbb{R}_{>0} \to \mathbb{R}$ is defined as $\phi(x) = -\frac{2}{\eta}\sqrt{x} - \beta \ln(x)$. Let $D_\phi(y, x)$ be the Bregman divergence associated with ϕ. Then, for any $x \in (0, 1)$, $\ell \in \mathbb{R}$ and η such that $x\ell \geq -\frac{\beta}{2}$, we have*

$$\sup_{y \in [0,1]} \left\{ \ell \cdot (x - y) - D_\phi(y, x) \right\} \leq 6\eta x^{3/2} \ell^2. \tag{12}$$

Proof. If $\ell \geq 0$, it immediately follows from Lemma 5 that the left-hand side of (12) is bounded by $\eta x^{3/2}\ell^2$ from above. We next consider the case of $\ell < 0$. The derivative of $\ell \cdot (x-y) - D_\phi(y,x)$ in y is given as

$$g(y) := -\ell + \frac{1}{\eta\sqrt{y}} - \frac{1}{\eta\sqrt{x}} + \frac{\beta}{y} - \frac{\beta}{x}.$$

This is a monotone decreasing function and hence the maximum of $\ell \cdot (x-y) - D_\phi(y,x)$ is attained at $y^* \in \mathbb{R}_{>0}$ such that $g(y^*) = 0$. As we have

$$g\left(\frac{\beta}{\beta+\ell x}x\right) \leq -\ell + \beta \cdot \frac{\beta+\ell x}{\beta x} - \frac{\beta}{x} = 0$$

and

$$g\left(\left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-2}\right) \leq -\ell + \frac{1}{\eta}\left(\frac{1}{\sqrt{x}} + \eta\ell\right) - \frac{1}{\eta\sqrt{x}} = 0,$$

we have

$$\begin{aligned} y^* \leq \min\left\{\frac{\beta}{\beta+\ell x}x, \left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-2}\right\} &\leq \min\left\{2x, \left(\frac{1}{\sqrt{x}} + \eta\ell\right)^{-2}\right\} \\ &= x\min\left\{2, \left(1+\eta\ell\sqrt{x}\right)^{-2}\right\}, \end{aligned}$$

where the last inequality follows from the assumption of $\ell x \geq -\frac{\beta}{2}$. We hence have

$$\begin{aligned} \sup_{y\in(0,1]}\left\{\ell\cdot(x-y) - D_\phi(y,x)\right\} &\leq -\ell(y^*-x) - D_\phi(y^*,x) \leq -\ell(y^*-x) \\ &\leq -\ell x \cdot \min\left\{1, \left(1+\eta\ell\sqrt{x}\right)^{-2} - 1\right\}. \end{aligned} \tag{13}$$

If $0 < -\eta\ell\sqrt{x} \leq 1/2$, we then have $(1+\eta\ell\sqrt{x})^{-2} - 1 \leq -6\eta\ell\sqrt{x}$, which implies that the value of (13) is at most $-\ell x \cdot (-6\eta\ell\sqrt{x}) = 6\eta x^{3/2}\ell^2$. If $-\eta\ell\sqrt{x} > 1/2$, we then have the value of (13) is at most $-\ell x < -\ell x \cdot (-2\eta\ell\sqrt{x}) = 2\eta x^{3/2}\ell^2$. This completes the proof. □

Lemma 7. *Let $\phi : \mathbb{R}_{>0} \to \mathbb{R}$ be defined as $\phi(x) = -\log(x)$ and $D_\phi(y,x)$ be the Bregman divengence associated with ϕ, i.e.,*

$$D_\phi(y,x) = -\log y + \log x + \frac{1}{x}(y-x) = -\log\frac{y}{x} + \frac{y}{x} - 1.$$

Then, for any $x \in (0,1)$, $\ell \in \mathbb{R}$ and $\eta > 0$ such that $\eta x\ell > -1$, we have

$$\sup_{y\in[0,1]}\left\{\ell\cdot(x-y) - \frac{1}{\eta}D_\phi(y,x)\right\} \leq \frac{1}{\eta}\left(-\log\left(1+\eta x\ell\right) + \eta x\ell\right).$$

Consequently, if $\eta\ell x \geq -\frac{1}{2}$, we have

$$\sup_{y\in[0,1]}\left\{\ell\cdot(x-y) - \frac{1}{\eta}D_\phi(y,x)\right\} \leq \eta x^2\ell^2\left(\frac{1}{2} + \mathbb{I}[\ell<0]\cdot\eta x|\ell|\right) \leq \eta x^2\ell^2.$$

Proof. We have

$$\ell\cdot(x-y) - \frac{1}{\eta}D_\phi(y,x) = \frac{1}{\eta}\left(\log\frac{y}{x} + \left(\frac{1}{x} + \eta\ell\right)(x-y)\right).$$

For fixed x, ℓ and η, this value is maximized when $\frac{1}{y} = \frac{1}{x} + \eta\ell$. We then have

$$\ell\cdot(x-y) - \frac{1}{\eta}D_\phi(y,x) = \frac{1}{\eta}\left(-\log\frac{x}{y} + \frac{x}{y} - 1\right) = \frac{1}{\eta}\left(-\log\left(1+\eta x\ell\right) + \eta x\ell\right).$$

As we have $-\log(1+a)+a \le \frac{1}{2}a^2 + \mathbb{I}[a<0]\cdot|a|^3$ for $a \ge -1/2$, we have

$$\begin{aligned}
\ell\cdot(x-y) - \frac{1}{\eta}D_\phi(y,x) &\le \frac{1}{\eta}\left(-\log\left(1+\eta x\ell\right)+\eta x\ell\right)\\
&\le \frac{1}{\eta}\left(\frac{1}{2}(\eta x\ell)^2 + \mathbb{I}[\ell<0]\cdot|\eta x\ell|^3\right)\\
&= \eta x^2\left(\frac{1}{2}\ell^2 + \mathbb{I}[\ell<0]\cdot\eta x|\ell|^3\right)\\
&\le \eta x^2\ell^2.
\end{aligned}$$

□

Proof of Lemma 1. We can show this by backward induction in layers. For $s = s_L$, (1) is clear as both sides are equal to 0. For $s \in S_k$ with $k < L$,

$$\begin{aligned}
Q^{\pi'}(s,a;\bar{\ell}) &= \bar{\ell}(s,a) + \sum_{s'\in S_{k+1}} P(s'|s,a)V^{\pi'}(s';\bar{\ell})\\
&= Q^{\pi}(s,a;\ell) - V^{\pi}(s;\ell) + \sum_{s'\in S_{k+1}} P(s'|s,a)\left(V^{\pi'}(s';\ell) - V^{\pi}(s';\ell)\right)\\
&= \ell(s,a) + \sum_{s'\in S_{k+1}} P(s'|s,a)V^{\pi}(s';\ell) - V^{\pi}(s;\ell) + \sum_{s'\in S_{k+1}} P(s'|s,a)\left(V^{\pi'}(s';\ell) - V^{\pi}(s';\ell)\right)\\
&= \ell(s,a) - V^{\pi}(s;\ell) + \sum_{s'\in S_{k+1}} P(s'|s,a)V^{\pi'}(s';\ell)\\
&= Q^{\pi'}(s,a;\ell) - V^{\pi}(s;\ell),
\end{aligned}$$

where the second equality follows from the induction hypothesis and the definition of $\bar{\ell}$. We hence have

$$V^{\pi'}(s;\bar{\ell}) = \sum_{a\in A}\pi'(a|s)Q^{\pi'}(s,a;\bar{\ell}) = \sum_{a\in A}\pi'(a|s)\left(Q^{\pi'}(s,a;\ell) - V^{\pi}(s;\ell)\right) = V^{\pi'}(s;\ell) - V^{\pi}(s;\ell).$$

□

C Online shortest path problem with bandit feedback

In this section, we analyze our algorithm for online shortest path problem. Specifically, we prove Theorems 4 and 5, which together directly imply Corollary 1 in the main body. In addition, we prove our lower bound result in Theorem 6.

C.1 Notation and problem setup

- $G = (V \cup \{s,g\}, E)$: a directed acyclic graph.
- s: Source node.
- g: Sink node.
- V: Set of vertices that are neither sources nor sinks.
- $E \subseteq (V\cup\{s\})\times(V\cup\{g\})$: set of directed edges.
- $e_-, e_+ \in V\cup\{s,g\}$: initial and terminal vertices of an edge $e\in E$, i.e., $e = (e_-, e_+)$.
- $\partial_- v, \partial_+ v \subseteq E$: sets of incoming and outgoing edges of a vertex $v \in V\cup\{s,g\}$, i.e., $\partial_- v = \{e\in E \mid e_+ = v\}$, $\partial_+ v = \{e\in E \mid e_- = v\}$.
- $n = |V|$.
- $m = |E|$.
- $\mathcal{P} \subseteq \{0,1\}^E$: set of (vector representations of) s-g paths.

- $\mathcal{Q} = \mathrm{conv}(\mathcal{P}) \subseteq [0,1]^E$: set of s-g flows of value 1, equivalently convex hull of $\mathcal{P}$. Note $\mathcal{Q} = \{q \in [0,1]^E \mid \sum_{e\in\partial_+ s} q(e) = \sum_{e\in\partial_- g} q(e) = 1, \sum_{e\in\partial_+ v} q(e) = \sum_{e\in\partial_- v} q(e) (\forall v \in V)\}$.
- $L = \max_{p\in\mathcal{P}}\{\|p\|_1\} \leq |V| + 1$: maximum length of s-g paths.
- Without loss of generality, we assume that every vertex $v \in V$ admits a path from s to g passing through v.

In each round $t \in [T]$, an environment chooses $\ell_t \in \mathbb{R}^E_{\geq 0}$ and then the player chooses $p_t \in P$, after which the player observes a feedback $c_t \in [0,1]$ such that $\mathbf{E}[c_t|\ell_t, p_t] = \langle \ell_t, p_t\rangle$. We here assume that ℓ_t satisfies $\langle \ell_t, p\rangle \leq 1$ for all $p \in P$. The performance of the player is evaluated in terms of regret defined as:

$$\mathrm{Reg}_T(p^*) = \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - p^*\rangle\right], \quad \mathrm{Reg}_T = \max_{p^*\in P} \mathrm{Reg}_T(p^*),$$

where the expectation is taken over all randomness arising from the environment and the algorithm.

C.2 Algorithm

The algorithm updates $q_t \in \mathcal{Q}$ by an FTRL approach and picks $p_t \in \mathcal{P}$ so that $\mathbf{E}[p_t|q_t] = q_t$, following the technique of [Maiti et al., 2025].

In the following, for $q \in \mathcal{Q}$ and $v \in V$, we denote

$$q(v) = \sum_{e\in\partial_- v} q(e) = \sum_{e\in\partial_+ v} q(e) \tag{14}$$

for the notational simplicity.

Let $p_v \in P$ be an s-g path that passes through v, i.e., $p_v(v) = 1$. Define $q_0 \in Q$ by

$$q_0 = \frac{1}{|V|}\sum_{v\in V} p_v. \tag{15}$$

We then have $q_0(v) \geq 1/|V| = 1/n$ for any $v \in V$.

Define q_t by

$$q_t \in \operatorname*{arg\,min}_{q\in Q}\left\{\left\langle \sum_{\tau=1}^{t-1} \widehat{\ell}_\tau, q\right\rangle + \psi_t(q)\right\}, \tag{16}$$

where $\widehat{\ell}_\tau$ is an estimator for ℓ_τ defined later. The regularizer ψ_t is given by

$$\psi_t(q) = -\frac{2}{\eta_t}\sum_{e\in E}\sqrt{q(e)} - \sum_{e\in E}\beta \ln q(e) \quad \text{with} \quad \eta_t = \frac{1}{\sqrt{t}},\ \beta = 2, \quad \text{or} \tag{17}$$

$$\psi_t(q) = -\sum_{e\in E}\frac{1}{\eta_t(e)}\ln q(e) \quad \text{with} \quad \eta_t(e) = \Big(4 + \frac{1}{\ln T}\sum_{\tau=1}^{t-1}\rho_\tau(e)\Big)^{-\frac{1}{2}}, \tag{18}$$

where $\rho_\tau \in [0,1]$ will be defined later.

Based on $q_t \in \mathcal{Q}$, we pick $p_t \in P$ in the same way as in [Maiti et al., 2025]:

- Initialize $p \in \{0,1\}^E$ by $p(e) = 0$ for all $e \in E$ and set $v \leftarrow s$.
- While $v \neq g$:
  - Pick $e \in \partial_+ v$ with probability $q_t(e)/q_t(v)$.
  - Set $p(e) \leftarrow 1$ and transition to the next node e_+, i.e., $v \leftarrow e_+$.

We then have $\mathbf{E}[p_t|q_t] = q_t$.

After outputting p_t, we get feedback $c_t \in [0,1]$ such that $\mathbf{E}[c_t|p_t, \ell_t] = \langle \ell_t, p_t \rangle$. Based on this, we define $\widehat{\ell}_t \in \mathbb{R}^E$ by

$$\widehat{\ell}_t(e) = c_t \cdot \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(e_-)}{q_t(e_-)} \right). \tag{19}$$

Note that the notation of (14) applies to $p \in P \subseteq Q$ as well, and that $p_t(v) = 1$ if and only if the path passes through the node v. Then, it holds for any $q \in Q$ that

$$\begin{aligned}
\left\langle \widehat{\ell}_t, q \right\rangle &= c_t \cdot \sum_{e \in E} \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(e_-)}{q_t(e_-)} \right) q(e) \\
&= c_t \cdot \left(\sum_{e \in E} \frac{p_t(e)}{q_t(e)} q(e) - \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v} \frac{p_t(e_-)}{q_t(e_-)} q(e) \right) && \left(E = \bigcup_{v \in V \cup \{s\}} \partial_+ v\right) \\
&= c_t \cdot \left(\sum_{e \in E} \frac{p_t(e)}{q_t(e)} q(e) - \sum_{v \in V \cup \{s\}} \frac{p_t(v)}{q_t(v)} q(v) \right) && (e \in \partial_+ v \iff e_- = v, (14)) \\
&= c_t \cdot \left(\sum_{e \in E} \frac{p_t(e)}{q_t(e)} q(e) - \sum_{v \in V} \frac{p_t(v)}{q_t(v)} q(v) - 1 \right). && (p_t(s) = q_t(s) = q(s) = 1)
\end{aligned} \tag{20}$$

We note that, an alternative definition of $\widehat{\ell}_t$ given as

$$\widehat{\ell}'_t(e) = c_t \cdot \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(e_+)}{q_t(e_+)} \right) \tag{21}$$

also satisfies (20) similarly, and hence we have

$$\left\langle \widehat{\ell}_t, q \right\rangle = \left\langle \widehat{\ell}'_t, q \right\rangle$$

for any $q \in Q$. Therefore, using $\widehat{\ell}'_t$ in (21) instead of $\widehat{\ell}_t$ in (19) does not change the behavior of the algorithm.

We now state the following that can be proved in a similar way as in [Maiti et al., 2025]:

Lemma 8. *For any $q, q' \in Q$, we have*

$$\mathbf{E}\left[\left\langle \widehat{\ell}_t, q - q' \right\rangle | q_t, \ell_t \right] = \langle \ell_t, q - q' \rangle,$$

where the expectation is taken w.r.t. p_t.

Note that the effect of the path-length does not appear here. That is, we do not need to assume that the path lengths are the same.

C.3 Regret analysis

Definition 2 (consistent policy)**.** Define $\Pi = \{\pi : V \cup \{s\} \to E \mid \pi(v) \in \partial_+ v \ (\forall v \in V \cup \{s\})\}$. Let $p^* \in P$ be an arbitrary s-g path. Let $E^* \subseteq E$ and $V^* \subseteq V$ denote the sets of edges and nodes included in p^*, i.e., $E^* = \{e \in E \mid p^*(e) = 1\}$ and $V^* = \{v \in V \mid p^*(v) = 1\}$. We say $\pi^* \in \Pi$ is *consistent* with $p^* \in P$ if and only if $\pi^*(v) \in E^*$ for all $v \in V^* \cup \{s\}$. We denotes $E' = E \setminus \mathrm{Im}(\pi^*)$.

Definition 3 (self-bounding regime for online shortest path)**.** Let $p^* \in P$ be an arbitrary s-g path and suppose that $\pi^* \in \Pi$ is consistent with p^*. Suppose that $\Delta \in [0,1]^E$ satisfies $\Delta(e) > 0$ for all $e \in E' = E \setminus \mathrm{Im}(\pi^*)$. The environment is in a (p^*, π^*, Δ, C)-self-bounding regime if it holds that

$$\mathrm{Reg}_T(p^*) \geq \mathbf{E}\left[\sum_{t=1}^{T} \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \Delta(e) p_t(e) \right] - C. \tag{22}$$

Remark 3. An example of $\Delta \in [0,1]^E$ can be constructed as follows: Assume that ℓ_t follows an identical distribution independently for all $t \in [T]$ and denote $\ell^* = \mathbf{E}[\ell_t]$. For each $u, v \in V \cup \{s, g\}$, let $\mathrm{dist}(u, v)$ denote the length of u-v shortest path w.r.t. the weight ℓ^*. (Set $\mathrm{dist}(u, v) = +\infty$ if there is no u-v path.) For each $e \in E$, define $\Delta \in [0,1]^E$ by

$$\Delta(e) = \ell^*(e) + \mathrm{dist}(e_+, g) - \mathrm{dist}(e_-, g). \tag{23}$$

Then, if $p^* \in P$ is a shortest s-g path, then (22) holds. In fact, for any s-g path $p \in P$ expressed as a sequence of $(s = v_0, e_1, v_1, e_2, v_2, \ldots, v_{h-1}, e_h, g = v_h)$, we have

$$\begin{aligned}\langle \Delta, p \rangle &= \sum_{j=1}^{h} \Delta(e_j) = \sum_{j=1}^{h} \left(\ell^*(e_j) + \mathrm{dist}(v_j, g) - \mathrm{dist}(v_{j-1}, g)\right) \\ &= \sum_{j=1}^{h} \ell^*(e_j) + \mathrm{dist}(v_h, g) - \mathrm{dist}(v_0, g) = \langle \ell^*, p \rangle - \langle \ell^*, p^* \rangle ,\end{aligned}$$

which implies $\mathrm{Reg}_T(p^*) = \mathbf{E}\left[\sum_{t=1}^{T} \langle \Delta, p_t \rangle\right]$. Suppose π^* is chosen so that $\pi^*(v) \in \arg\min_{e \in \partial_+ v}\{\Delta(e)\}$. Then, $\Delta(e)$ for all $e \in E' = E \setminus \mathrm{Im}(\pi^*)$ if and only if the shortest v-g path is unique for all $v \in V \cup \{s\}$.

Remark 4. An issue of the definition of Δ in (23) is the requirement for a strong assumption that the v-g shortest path is unique for all $v \in V \cup \{s\}$ to ensure that $\Delta(e) > 0$ for all $e \in E'$. We can relax this assumption by some alternative definitions of Δ. When using the following definition, it suffices to assume that the s-g shortest path is unique: Define $L' = \max_{p \in \mathcal{P}}\{\sum_{e \in E'} p(e)\} \le L$. For $e \in E'$ and $k \in [L']$, define $\tilde{\mathcal{P}}(e, k) \subseteq P$ and $\tilde{\Delta}(e)$ by

$$\begin{aligned}\tilde{\mathcal{P}}(e, k) &= \left\{ p \in \mathcal{P} \mid p(e) = 1, \sum_{e' \in E'} p(e') = k \right\}, \\ \tilde{\Delta}(e) &= \min_{k \in [L']} \left\{ \frac{1}{k} \inf_{p \in \tilde{\mathcal{P}}(e,k)} \{\langle \ell^*, p - p^* \rangle\} \right\} = \min_{k \in [L']} \left\{ \frac{1}{k} \inf_{p \in \tilde{P}(e,k)} \{\langle \Delta, p \rangle\} \right\}. \end{aligned} \tag{24}$$

We then have $\sum_{e \in E'} \tilde{\Delta}(e) p(e) \le \sum_{e \in E'} \Delta(e) p(e)$ for all $p \in P$, which implies that the environment is in $(p^*, \pi^*, \tilde{\Delta})$ as well. Further, as we have $p^* \notin \tilde{P}(e, k)$ for any $e \in E'$ and k, we have

$$\tilde{\Delta}(e) \ge \frac{1}{L'} \min_{p \in P \setminus \{p^*\}} \{\langle \ell^*, p - p^* \rangle\} =: \frac{1}{L'} \Delta_{\min}$$

for any $e \in E'$. Hence, this value is positive as long as the s-g shortest path is unique. We also have $\tilde{\Delta}(e) \ge \min_{e' \in E'} \Delta(e)$ for any $e \in E'$.

Let $q_0 \in \mathcal{Q}$ be such that $q_0(e) \ge 1/m$. For any $p^* \in \mathcal{Q}$ and $\varepsilon \in [0,1]$, set q^* by

$$q^* = (1 - \varepsilon)\, p^* + \varepsilon q_0. \tag{25}$$

Using Lemma 8 and standard analysis for FTRL (see, e.g., Exercise 28.12 of [Lattimore and Szepesvári, 2020]), we obtain:

$$\begin{aligned}&\mathrm{Reg}_T(p^*) = \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, p_t - p^* \rangle\right] = \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, p_t - p^* \right\rangle\right] = \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, q_t - p^* \right\rangle\right] \\ &= \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, q_t - q^* \right\rangle\right] + \varepsilon \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, q_0 - p^* \right\rangle\right] \le \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, q_t - q^* \right\rangle\right] + \varepsilon T \\ &\le \varepsilon T + \sum_{t=1}^{T} \mathbf{E}\left[\underbrace{\left\langle \widehat{\ell}_t, q_t - q_{t+1} \right\rangle - D_t(q_{t+1}, q_t)}_{=:\mathrm{stab}_t} + \underbrace{\left(\psi_t(q^*) - \psi_{t-1}(q^*) - \psi_t(q_t) + \psi_{t-1}(q_t)\right)}_{=:\mathrm{pena}_t} \right],\end{aligned} \tag{26}$$

where $D_t(\cdot, \cdot)$ represents the Bregman divergence associated with ψ_t defined by (17) or (18), and we set $\psi_t(q) = 0$ for $t = 0$ as an exception.

C.3.1 Analysis for Tsallis-entropy case

Theorem 4 (First part of Theorem 1)**.** *Let $p^* \in P$ be an arbitrary s-g path and suppose that $\pi^* \in \Pi$ is consistent with p^*. Then the proposed algorithm with the Tsallis-entropy regularizer* (17) *achieves:*

$$\mathrm{Reg}_T(p^*) \lesssim \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbf{E}\left[\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \sqrt{q_t(e)} + \sum_{v \in V \setminus V^*} \sqrt{q_t(\pi^*(v))} \right] + m \log T \quad (27)$$

$$\leq \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbf{E}\left[\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \sqrt{q_t(e)} + n \sqrt{\sum_{v \in V^* \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} q_t(e)} \right] + m \log T \quad (28)$$

Corollary 3 (First part of Corollary 1)**.** *In the adversarial regime, the proposed algorithm with the Tsallis-entropy regularizer* (17) *achieves* $\mathrm{Reg}_T = O(\sqrt{mLT} + m \log T)$*. Further, if the environment is in a (p^*, π^*, Δ, C)-self-bounding regime given in Definition 3, we then have* $\mathrm{Reg}_T(p^*) \lesssim U + \sqrt{UC} + m \log T$*, where we define*

$$U = \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \frac{\log T}{\Delta(e)} + \frac{n^2 \log T}{\Delta^*}, \quad (29)$$

$$\Delta^* = \min \left\{ \Delta(e) \mid \exists v \in V^* \cup \{s\}, e \in \partial_+ v \setminus \{\pi^*(v)\} \right\}. \quad (30)$$

In the following, we provide a proof of Theorem 4.

Lemma 9. *When we use the Tsallis-entropy regularizer* (17)*, stability terms are bounded as*

$$\mathbf{E}[\mathrm{stab}_t] \lesssim \mathbf{E}\left[\eta_t \sum_{v \in V \cup \{s\}} \left(\sum_{e \in \partial_+ v} \sqrt{q_t(e)} - \sqrt{q_t(v)} \right) \right] \lesssim \mathbf{E}\left[\eta_t \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(e)\}} \sqrt{q_t(e)} \right]. \quad (31)$$

Proof. To bound the stability term, we can apply Lemma 6 with $\ell = \widehat{\ell}_t(e)$, $x = q_t(e)$, $\eta = \eta_t$, and $\beta = 2$. In fact, we can verify that $-\widehat{\ell}_t(e) q_t(e) \leq \frac{q_t(e)}{q_t(e_-)} \leq 1 \leq \frac{\beta}{2}$. Hence, from Lemma 6, we have

$$\begin{aligned}
&\left\langle \widehat{\ell}_t, q_t - q_{t+1} \right\rangle - \frac{1}{\eta_t} D(q_{t+1}, q_t) \lesssim \eta_t \sum_{e \in E} (q_t(e))^{3/2} \left(\widehat{\ell}_t(e) \right)^2 \\
&\lesssim \eta_t \sum_{e \in E} (q_t(e))^{3/2} \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(e_-)}{q_t(e_-)} \right)^2 \\
&= \eta_t \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v} (q_t(e))^{3/2} \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(e_-)}{q_t(e_-)} \right)^2 && (E = \textstyle\bigcup_{v \in V \cup \{s\}} \partial_+ v) \\
&= \eta_t \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v} (q_t(e))^{3/2} \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(v)}{q_t(v)} \right)^2 && (e \in \partial_+ v \iff e_- = v) \\
&= \eta_t \sum_{v \in V \cup \{s\}} \frac{p_t(v)}{\sqrt{q_t(v)}} \sum_{e \in \partial_+ v} \left(\frac{q_t(e)}{q_t(v)} \right)^{3/2} \left(\frac{p_t(e) q_t(v)}{q_t(e)} - 1 \right)^2,
\end{aligned}$$

where we used the fact that $p_t(v) = 0$ implies $p_t(e) = 0$ for $e \in \partial_+ v$. Taking the conditional expectation w.r.t. p_t given q_t, we obtain:

$$\begin{aligned}
&\mathbf{E}\left[\frac{p_t(v)}{\sqrt{q_t(v)}} \sum_{e\in\partial_+ v} \left(\frac{q_t(e)}{q_t(v)}\right)^{3/2} \left(\frac{p_t(e)q_t(v)}{q_t(e)} - 1\right)^2\right] \\
&= \sqrt{q_t(v)} \cdot \mathbf{E}\left[\sum_{e\in\partial_+ v} \left(\frac{q_t(e)}{q_t(v)}\right)^{3/2} \left(\frac{p_t(e)q_t(v)}{q_t(e)} - 1\right)^2 | p_t(v) = 1\right] \\
&= \sqrt{q_t(v)} \sum_{e\in\partial_+ v} \left(\frac{q_t(e)}{q_t(v)}\right)^{3/2} \left(\frac{q_t(e)}{q_t(v)} \cdot \left(\frac{q_t(v)}{q_t(e)} - 1\right)^2 + \left(1 - \frac{q_t(e)}{q_t(v)}\right) \cdot 1\right) \\
&= \sqrt{q_t(v)} \sum_{e\in\partial_+ v} \left(\frac{q_t(e)}{q_t(v)}\right)^{1/2} \left(1 - \frac{q_t(e)}{q_t(v)}\right) = \sum_{e\in\partial_+ v} \sqrt{q_t(e)} \left(1 - \frac{q_t(e)}{q_t(v)}\right).
\end{aligned}$$

Further, as we have $(1 - x) \le 2(1 - \sqrt{x})$ for $x \in [0, 1]$, we have

$$\begin{aligned}
&\sum_{e\in\partial_+ v} \sqrt{q_t(e)} \left(1 - \frac{q_t(e)}{q_t(v)}\right) \le 2 \sum_{e\in\partial_+ v} \sqrt{q_t(e)} \left(1 - \sqrt{\frac{q_t(e)}{q_t(v)}}\right) = 2 \left(\sum_{e\in\partial_+ v} \sqrt{q_t(e)} - \sum_{e\in\partial_+ v} \frac{q_t(e)}{\sqrt{q_t(v)}}\right) \\
&= 2 \left(\sum_{e\in\partial_+ v} \sqrt{q_t(e)} - \frac{q_t(v)}{\sqrt{q_t(v)}}\right) = 2 \left(\sum_{e\in\partial_+ v} \sqrt{q_t(e)} - \sqrt{q_t(v)}\right).
\end{aligned}$$

By combining the above inequalities,

$$\mathbf{E}\left[\left\langle \widehat{\ell}_t, \tilde{q}_t - \tilde{q}_{t+1} \right\rangle - \frac{1}{\eta_t} D(\tilde{q}_{t+1}, \tilde{q}_t) | q_t\right] \lesssim \eta_t \sum_{v\in V\cup\{s\}} \left(\sum_{e\in\partial_+ v} \sqrt{q_t(e)} - \sqrt{q_t(v)}\right).$$

The second inequality in (31) follows from the fact that $q_t(e) \le q_t(v)$ for any $e \in \partial_+ v$. □

Lemma 10. *Let p^* be a path consisting of $V^* \in V$ and $E^* \subseteq E$. Suppose that q^* is given by* (25)*. For $t \ge 2$, if q^* is a path consisting of $V^* \in V$ and $E^* \subseteq E$, penalty terms are bounded as*

$$\begin{aligned}
\mathrm{pena}_t &\lesssim \eta_t \left(\sum_{e\in E} \sqrt{q(e)} - |E| + m\varepsilon\right) \le \eta_t \sum_{e\in E\setminus E^*} \sqrt{q_t(e)} + m\varepsilon \\
&= \eta_t \left(\sum_{v\in V\cup\{s\}} \sum_{e\in\partial_+ v\setminus\{\pi^*(v)\}} \sqrt{q_t(e)} + \sum_{v\in V\setminus V^*} \sqrt{q_t(\pi^*(v))}\right) + m\varepsilon \qquad (32)
\end{aligned}$$

In the case of $t = 1$, the bound includes an $O(m \log(m/\varepsilon))$ term in addition to the above.

Proof. Suppose that $t \ge 2$. We note it follows from $\eta_t = 1/\sqrt{t}$ that $\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} = \Theta(\eta_t)$. We hence have

$$\begin{aligned}
\mathrm{pena}_t &= 2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \sum_{e\in E} \left(\sqrt{q_t(e)} - \sqrt{q^*(e)}\right) \lesssim \eta_t \sum_{e\in E} \left(\sqrt{q_t(e)} - \sqrt{q^*(e)}\right) \\
&\lesssim \eta_t \sum_{e\in E} \left(\sqrt{q_t(e)} - \sqrt{p^*(e)} + \varepsilon\right) = \eta_t \left(\sum_{e\in E} \sqrt{q_t(e)} - |E^*| + m\varepsilon\right) \\
&\le \eta_t \sum_{e\in E\setminus E^*} \sqrt{q_t(e)} + m\varepsilon.
\end{aligned}$$

The equality in (32) follows from $E \setminus E^* = (\bigcup_{v\in V\cup\{s\}} (\partial_+ v \setminus \pi^*(v))) \cup (V \setminus V^*)$. In the case of $t = 1$, the penalty term includes an additional term of

$$\beta \sum_{e\in E} (\ln(q_t(e)) - \ln(q^*(e))) \lesssim \sum_{e\in E} \ln \frac{1}{q^*(e)} \lesssim m \log\left(\frac{m}{\varepsilon}\right),$$

which completes the proof. □

Lemmas 9 and 10 combined with (26) immediately lead to (27). Given this, the next step to (28) follows from the following lemma:

Lemma 11. *For any $v \in V \setminus V^*$ and any $q \in \mathcal{Q}$, we have*

$$q(v) \leq \sum_{v' \in V^* \cup \{s\}} \sum_{e \in \partial_+ v' \setminus \pi^*(v')} q(e). \tag{33}$$

Proof. As $\mathcal{Q}$ is a convex hull of $\mathcal{P}$ and both sides of (33) are linear in q, it suffices to show (33) for $q \in \mathcal{P}$. Then, for any $q \in \mathcal{P}$, if RHS of (33) is positive, then it is at least 1, and hence (33) holds. Therefore, it suffices to show that RHS $= 0 \implies$ LHS $= 0$ for $q \in \mathcal{P}$. Suppose q corresponds to the sequence of $(s = v_0, e_1, v_1, \ldots, e_h, v_h = g)$. If RHS $= 0$, then q consists of $V^* = \{v_1^*, \ldots, v_{h^*}^* = g\}$ and $E^* = \{\pi^*(v_j^*)\}_{j=0,1,\ldots,h^*}$. In fact, we can show this in induction in j, under the condition that $\sum_{e \in \partial_+ v' \setminus \{\pi^*(v')\}} q(e) = 0$ for all $v' \in V^* \cup \{s\}$: $q(v_j^*) = 1 \implies \sum_{e \in \partial_+ v_j^*} q(e) = 1 \implies q(\pi^*(v_j^*)) = 1 \implies q(v_{j+1}^*) = 1$. □

Proof of Theorem 4. We choose $\varepsilon = \frac{1}{T} \in [0, 1]$ in the definition of q^* in (25). Then, Lemmas 9 and 10 combined with (26) lead to (27). The other inequality (28) follows from

$$\sum_{v \in V \setminus V^*} \sqrt{q_t(\pi^*(v))} \leq \sum_{v \in V \setminus V^*} \sqrt{q_t(v)} \leq n \sqrt{\sum_{v \in V^* \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} q_t(e)},$$

where we used Lemma 11 in the second inequality. □

Proof of Corollary 3. As we have $\sum_{e \in E} q(e) \leq L$ and $\sum_{v \in V} q(v) \leq L$ for any $q \in \mathcal{Q}$, from the Cauchy-Schwarz inequality, we have

$$\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \sqrt{q_t(e)} \leq \sum_{e \in E} \sqrt{q_t(e)} \leq \sqrt{|E| \sum_{e \in E} q_t(e)} \leq \sqrt{mL}$$

and

$$\sum_{v \in V \setminus V^*} \sqrt{q_t(v)} \leq \sum_{v \in V} \sqrt{q_t(v)} \leq \sqrt{|V| \sum_{v \in V} q_t(v)} \leq \sqrt{nL}.$$

Combining Theorem 4 with this and $\sum_{t=1}^T \frac{1}{\sqrt{t}} \lesssim \sqrt{T}$, we obtain $\mathrm{Reg}_T \lesssim \sqrt{mLT} + m \log T$. By using the Cauchy-Schwarz inequality, we obtain the following:

$$\begin{aligned}
&\sum_{t=1}^T \frac{1}{\sqrt{t}} \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \sqrt{q_t(e)} \\
&= \sum_{t=1}^T \frac{1}{\sqrt{t}} \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \frac{1}{\sqrt{\Delta(e)}} \sqrt{\Delta(e) q_t(e)} \\
&\leq \sum_{t=1}^T \frac{1}{\sqrt{t}} \sqrt{\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \frac{1}{\Delta(e)}} \sqrt{\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \Delta(e) q_t(e)} \\
&\lesssim \sqrt{\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \frac{1}{\Delta(e)}} \sqrt{\sum_{t=1}^T \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \Delta(e) q_t(e) \log T}.
\end{aligned}$$

Similarly, we have have

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{n}{\sqrt{t}} \sqrt{\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} q_t(e)} \\
&\lesssim n \sqrt{\frac{1}{\Delta^*}} \sqrt{\sum_{t=1}^{T} \Delta^* \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} q_t(e) \log T} \\
&\leq n \sqrt{\frac{1}{\Delta^*}} \sqrt{\sum_{t=1}^{T} \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \Delta(e) q_t(e) \log T}.
\end{aligned}$$

Hence, from Theorem 4, Jensen's inequality, and the assumption of self-bounding regime in Definition 3, we have

$$\begin{aligned}
\mathrm{Reg}_T(p^*) &\lesssim \sqrt{U \cdot \mathbf{E}\left[\sum_{t=1}^{T} \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \Delta(e) q_t(e)\right]} + m \log T \\
&\leq \sqrt{U \cdot (\mathrm{Reg}_T(p^*) + C)} + m \log T \\
&\lesssim \sqrt{U \cdot \mathrm{Reg}_T(p^*)} + \sqrt{UC} + m \log T,
\end{aligned}$$

which implies $\mathrm{Reg}_T(p^*) \lesssim U + \sqrt{UC} + m \log T$. □

C.3.2 Analysis for log-barrier case

In the case of the log-barrier regularizer, we have the following regret bounds:

Theorem 5 (Second part of Theorem 1)**.** *The proposed algorithm with the log-barrier regularizer* (18) *achieves:*

$$\mathrm{Reg}_T \lesssim \mathbf{E}\left[\sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v} \sqrt{\sum_{t=1}^{T} c_t^2 p_t(v) \left(p_t(e) - \frac{q_t(e)}{q_t(v)}\right)^2 \log(T)}\right] + m \log(T).$$

Corollary 4 (Second part of Corollary 1)**.** *In the adversarial regime, the proposed algorithm with the log-barrier regularizer* (18) *achieves*

$$\mathrm{Reg}_T \lesssim \mathbf{E}\left[\sqrt{mL \sum_{t=1}^{T} c_t^2 \log T}\right] \lesssim \sqrt{mL \log T \cdot \min_{p^* \in \mathcal{P}} \mathbf{E}\left[\sum_{t=1}^{T} \langle \ell_t, p^* \rangle\right]} \leq \sqrt{mLT}. \tag{34}$$

$\mathrm{Reg}_T(p^*) \lesssim U + \sqrt{UC} + m \log T$*, where we define*

$$U = \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v \setminus \{\pi^*(v)\}} \frac{\log T}{\Delta(e)}. \tag{35}$$

To show these, we use the following lemmas:

Lemma 12. *When we use the log-barrier regularizer* (18)*, we have*

$$\mathrm{stab}_t \lesssim c_t^2 \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v} \eta_t(e) p_t(v) \left(p_t(e) - \frac{q_t(e)}{q_t(v)}\right)^2.$$

Proof. As we have $-\eta_t q_t(e)\widehat{\ell}_t(e) \leq \eta_t \frac{q_t(e)}{q_t(e_-)} \leq \eta_t \leq 1/2$, we can use Lemma 7 to bound the stability terms:

$$\begin{aligned}
\mathrm{stab}_t &\lesssim \sum_e \eta_t(e)(q_t(e))^2(\widehat{\ell}_t(e))^2 \\
&= c_t^2 \sum_{v\in V\cup\{s\}} \sum_{e\in\partial_+ v} \eta_t(e)(q_t(e))^2 \left(\frac{p_t(e)}{q_t(e)} - \frac{p_t(v)}{q_t(v)}\right)^2 \\
&= c_t^2 \sum_{v\in V\cup\{s\}} \sum_{e\in\partial_+ v} \eta_t(e) \left(p_t(e) - \frac{q_t(e)p_t(v)}{q_t(v)}\right)^2 \\
&= c_t^2 \sum_{v\in V\cup\{s\}} \sum_{e\in\partial_+ v} \eta_t(e)p_t(v) \left(p_t(e) - \frac{q_t(e)}{q_t(v)}\right)^2,
\end{aligned}$$

where we used the fact that $q(v) = 0$ implies $q(e) = 0$ for $e \in \partial_+(v)$ and for any $q \in Q$. □

Define $\rho_t(e)$ by

$$\rho_t(e) = c_t^2 p_t(e_-) \left(p_t(e) - \frac{q_t(e)}{q_t(e_-)}\right)^2.$$

Define q^* by (25) with $\varepsilon = m/T$. Then, from (26) and Lemma 12, we have

$$\begin{aligned}
\mathrm{Reg}_T(p^*) \leq m + \mathrm{Reg}_T(q^*) &\lesssim \mathbf{E}\left[\sum_{t=1}^T \sum_{e\in E} \eta_t(e)\rho_t(e) + \sum_{e\in E} \frac{1}{\eta_T(e)} \log\frac{1}{q^*(e)}\right] + m \\
&\leq \mathbf{E}\left[\sum_{e\in E}\left(\sum_{t=1}^T \eta_t(e)\rho_t(e) + \frac{\log(T)}{\eta_T(e)}\right)\right] + m \\
&\lesssim \mathbf{E}\left[\sum_{e\in E}\sqrt{\sum_{t=1}^T \rho_t(e)\log(mT)}\right] + m\log(T),
\end{aligned}$$

where the last inequality follows from the setting of $\eta_t(e)$:

$$\eta_t(e) = \left(4 + \frac{1}{\log(T)}\sum_{\tau=1}^{t-1}\rho_\tau(e)\right)^{-1/2}.$$

This completes the proof of Theorem 5.

Proof of Corollary 4. We can show (34) by using the Cauchy-Schwarz inequality, Jensen's inequality, and the fact that $\sum_e p_t(e) \le L$. The other one can be shown via the following:

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v}\sqrt{\sum_{t=1}^T c_t^2 p_t(v)\left(p_t(e)-\frac{q_t(e)}{q_t(v)}\right)^2}\right]\\
&\le \sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v}\sqrt{\mathbf{E}\left[\sum_{t=1}^T p_t(v)\left(p_t(e)-\frac{q_t(e)}{q_t(v)}\right)^2\right]}\\
&\le \sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v}\sqrt{\mathbf{E}\left[\sum_{t=1}^T q_t(e)\left(1-\frac{q_t(e)}{q_t(v)}\right)\right]}\\
&\le \sum_{v\in V\cup\{s\}}\left(\sum_{e\in\partial_+ v\setminus\pi^*(v)}\sqrt{\mathbf{E}\left[\sum_{t=1}^T q_t(e)\left(1-\frac{q_t(e)}{q_t(v)}\right)\right]}+\sqrt{\sum_{t=1}^T\mathbf{E}\left[q_t(\pi^*(v))\left(1-\frac{q_t(\pi^*(v))}{q_t(v)}\right)\right]}\right)\\
&\le \sum_{v\in V\cup\{s\}}\left(\sum_{e\in\partial_+ v\setminus\pi^*(v)}\sqrt{\mathbf{E}\left[\sum_{t=1}^T q_t(e)\right]}+\sqrt{\mathbf{E}\left[\sum_{t=1}^T\left(q_t(v)-q_t(\pi^*(v))\right)\right]}\right)\\
&\le 2\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\pi^*(v)}\sqrt{\mathbf{E}\left[\sum_{t=1}^T q_t(e)\right]}.
\end{aligned}
$$

Hence, by using the condition of the self-bounding constraint in Definition 3, we obtain

$$
\begin{aligned}
\mathrm{Reg}_T(p^*) &\lesssim \sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\pi^*(v)}\sqrt{\mathbf{E}\left[\sum_{t=1}^T q_t(e)\log T\right]}+m\log T\\
&= \sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\pi^*(v)}\sqrt{\frac{\log T}{\Delta(e)}}\sqrt{\mathbf{E}\left[\Delta(e)\sum_{t=1}^T q_t(e)\right]}+m\log T\\
&\le \sqrt{\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\pi^*(v)}\frac{\log T}{\Delta(e)}}\sqrt{\mathbf{E}\left[\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\pi^*(v)}\Delta(e)\sum_{t=1}^T q_t(e)\right]}+m\log T\\
&\le \sqrt{\sum_{v\in V\cup\{s\}}\sum_{e\in\partial_+ v\setminus\pi^*(v)}\frac{\log T}{\Delta(e)}}\sqrt{\mathrm{Reg}_T(p^*)+C}+m\log T\\
&\le \sqrt{U\mathrm{Reg}_T(p^*)}+\sqrt{UC}+m\log T.
\end{aligned}
$$

This implies that $\mathrm{Reg}_T(p^*) \lesssim U+\sqrt{UC}+m\log T$. □

C.4 Lower bound for stochastic environments

As our problem is a special case of the linear bandit problem, we can apply the lower bound given by [Lattimore and Szepesvari, 2017, Corollary 2], which is characterized by the following optimization problem:

$$
\inf_{\alpha\in[0,\infty]^{\mathcal{P}}}\sum_{p\in\mathcal{P}^-}\alpha(p)\Delta(p)\quad\text{subject to}\quad \|p\|^2_{H(\alpha)^{-1}}\le\frac{\Delta(p)^2}{2}\quad(\forall p\in P^-),
$$

where $\mathcal{P}^- = \mathcal{P}\setminus\arg\min_{p\in\mathcal{P}}\langle\ell^*,p\rangle$, $\Delta(p)=\langle\ell^*,p\rangle-\min_{p^*\in\mathcal{P}}\langle\ell^*,p^*\rangle$, and

$$
H(\alpha)=\sum_{p\in\mathcal{P}}\alpha(p)pp^\top.
$$

However, deriving an explicit expression for the optimal value of this optimization problem is not straightforward. In fact, we have not yet identified such an expression. Instead, in what follows, we present a lower bound by exploiting the specific structure inherent to this problem.

Consider stochastic environments specified by ℓ^* such that $\langle \ell^*, p \rangle \in [3/8, 5/8]$ for all $p \in \mathcal{P}$. We suppose that $c_t \in \{0, 1\}$ follows a Bernoulli distribution of parameter $\langle \ell, p_t \rangle$. Suppose that Δ is defined as (23). We then have $\langle \ell^*, p \rangle - \min_{p^* \in \mathcal{P}} \langle \ell^*, p^* \rangle = \langle \Delta, p \rangle$. Denote

$$\bar{\Delta}(e) = \min_{p \in \mathcal{P}: p(e)=1} \{\langle \Delta, p \rangle\}. \tag{36}$$

Theorem 6. *Consider an arbitrary consistent algorithm, i.e., assume that there exists $\varepsilon \in (0, 1)$ such that $\mathrm{Reg}_T \leq MT^{1-\varepsilon}$ holds for any instances, where $M > 0$ is a parameter independent of T. We then have*

$$\liminf_{T \to \infty} \frac{\mathrm{Reg}_T}{\log T} \gtrsim \varepsilon \sum_{e \in E: \bar{\Delta}(e) > 0} \frac{\Delta(e)}{\bar{\Delta}(e)^2}. \tag{37}$$

Remark 5. If Δ is given by (23), we have $\bar{\Delta}(e) = \Delta(e)$ for $e \in \partial_+ v$ where $v \in V^* := \{v \in V \mid \mathrm{dist}(s, v) + \mathrm{dist}(v, g) = \mathrm{dist}(s, g)\}$. In fact, for such an edge e, we have

$$\bar{\Delta}(e) = \mathrm{dist}(s, e_-) + \ell(e) + \mathrm{dist}(e_+, g) - \mathrm{dist}(s, g) = \ell^*(e) + \mathrm{dist}(e_+, g) - \mathrm{dist}(e_-, g) = \Delta(e).$$

We hence have

$$\sum_{e \in E: \bar{\Delta}(e) > 0} \frac{\Delta(e)}{\bar{\Delta}(e)^2} = \sum_{v \in V \cup \{s\}} \sum_{e \in \partial_+ v: \bar{\Delta}(e) > 0} \frac{\Delta(e)}{\bar{\Delta}(e)^2} \geq \sum_{v \in V^* \cup \{s\}} \sum_{e \in \partial_+ v: \bar{\Delta}(e) > 0} \frac{1}{\Delta(e)}.$$

Proof of Theorem 6. Let $p^* \in \arg\min_{p \in \mathcal{P}} \{\langle \ell^*, p \rangle\}$. As we have

$$\mathrm{Reg}_T = \mathbf{E}\left[\sum_{t=1}^T \langle \ell^*, p_t - p^* \rangle\right] = \mathbf{E}\left[\sum_{t=1}^T \langle \ell^*, p_t - p^* \rangle\right] = \sum_{e \in E} \Delta(e) \mathbf{E}\left[\sum_{t=1}^T p_t(e)\right],$$

we have

$$\liminf_{T \to \infty} \frac{\mathrm{Reg}_T}{\log T} = \sum_{e \in E} \Delta(e) \liminf_{T \to \infty} \frac{1}{\log T} \mathbf{E}\left[\sum_{t=1}^T p_t(e)\right]. \tag{38}$$

In the following, we evaluate the value of $\liminf_{T \to \infty} \frac{1}{\log T} \mathbf{E}\left[\sum_{t=1}^T p_t(e)\right]$ for any fixed $\tilde{e} \in E$ such that $\delta := \bar{\Delta}(\tilde{e}) > 0$. Consider a modified environment given by $\tilde{\ell}$ such that $\tilde{\ell}(\tilde{e}) = \ell^*(\tilde{e}) - 2\delta$ and $\tilde{\ell}(e) = \ell^*(e)$ for $e \neq \tilde{e}$. Then, as we have $p^*(\tilde{e}) = 0$, it holds for any $p \in \mathcal{P}$ that

$$\begin{aligned} \left\langle \tilde{\ell}, p - p^* \right\rangle &= \langle \ell^*, p - p^* \rangle - 2\delta(p(\tilde{e}) - p^*(\tilde{e})) \\ &= \langle \Delta, p \rangle - 2\delta(p(\tilde{e}) - p^*(\tilde{e})) \\ &\geq \delta p(\tilde{e}) - 2\delta(p(\tilde{e}) - p^*(\tilde{e})) = -\delta p(\tilde{e}). \end{aligned} \tag{39}$$

where the inequality follows from the definition of $\delta = \bar{\Delta}(\tilde{e})$ given in (36). Further, as there exists $p \in \mathcal{P}$ such that $p(e) = 1$ and $\langle \Delta, p \rangle = \delta$, we have

$$\min_{p \in \mathcal{P}} \left\langle \tilde{\ell}, p - p^* \right\rangle \leq \delta - 2\delta = -\delta. \tag{40}$$

Hence, from (39) and (40), the regret for the environment given by $\tilde{\ell}$ satisfies

$$\tilde{\mathrm{Reg}}_T = \max_{p \in \mathcal{P}} \tilde{\mathbf{E}}\left[\sum_{t=1}^T \left\langle \tilde{\ell}, p_t - p \right\rangle\right] \geq \delta \cdot \tilde{\mathbf{E}}\left[\sum_{t=1}^T (1 - p_t(\tilde{e}))\right], \tag{41}$$

where $\tilde{\mathbf{E}}[\cdot]$ represents the expected value when feedback is generated by an environment associated with $\tilde{\ell}$. On the other hand, the regret for the environment given by ℓ^* satisfies

$$\mathrm{Reg}_T = \mathbf{E}\left[\sum_{t=1}^T \langle \Delta, p_t \rangle\right] \geq \delta \cdot \mathbf{E}\left[\sum_{t=1}^T p_t(\tilde{e})\right]. \tag{42}$$

Let TV denote the total variation distance between trajectories $((p_t, c_t))_{t=1}^T$ for environments with ℓ^* and $\tilde{\ell}$. Then, as we have $\frac{1}{T}\sum_{t=1}^T p_t(\tilde{e}) \in [0,1]$, we have

$$\left| \tilde{\mathbf{E}}\left[\frac{1}{T}\sum_{t=1}^T p_t(\tilde{e})\right] - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^T p_t(\tilde{e})\right] \right| \leq \mathrm{TV}\,.$$

We hence have

$$\begin{aligned}
1 - \mathrm{TV} &\leq 1 - \tilde{\mathbf{E}}\left[\frac{1}{T}\sum_{t=1}^T p_t(\tilde{e})\right] + \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^T p_t(\tilde{e})\right] \\
&= \tilde{\mathbf{E}}\left[\frac{1}{T}\sum_{t=1}^T (1 - p_t(\tilde{e}))\right] + \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^T p_t(\tilde{e})\right] \\
&\leq \frac{1}{\delta T}\left(\tilde{\mathrm{Reg}}_T + \mathrm{Reg}_T\right),
\end{aligned}$$

where the last inequality follows from (41) and (42). Here, from the Bretagnolle-Huber inequality (e.g., [Canonne [2022], Corollary 4]) and the chain rule of the KL divergence, we have

$$\begin{aligned}
1 - \mathrm{TV} &\geq \frac{1}{2}\exp\left(\sum_{t=1}^T \mathbf{E}\left[D_{\mathrm{KL}}\left(\langle \ell^*, p_t\rangle \,||\langle \tilde{\ell}, p_t\rangle\right)\right]\right) \\
&= \frac{1}{2}\exp\left(\sum_{t=1}^T \mathbf{E}\left[p_t(\tilde{e})\, D_{\mathrm{KL}}\left(\langle \ell^*, p_t\rangle \,||\langle \ell^*, p_t\rangle - 2\delta\right)\right]\right) \\
&\geq \frac{1}{2}\exp\left(-5\delta^2 \cdot \mathbf{E}\left[\sum_{t=1}^T p_t(\tilde{e})\right]\right).
\end{aligned}$$

Combining the above inequalities, we obtain

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^T p_t(\tilde{e})\right] &\gtrsim \frac{1}{\delta^2}\log\frac{1}{2(1-\mathrm{TV})} \geq \frac{1}{\delta^2}\log\frac{\delta T}{2(\mathrm{Reg}_T + \tilde{\mathrm{Reg}}_T)} \\
&\geq \frac{1}{\delta^2}\log\frac{\delta T}{4MT^{1-\varepsilon}} = \frac{1}{\delta^2}\left(\varepsilon \log T + \log\left(\frac{\delta}{4M}\right)\right)
\end{aligned}$$

Consequently, we have

$$\liminf_{T\to\infty} \frac{1}{\log T}\mathbf{E}\left[\sum_{t=1}^T p_t(\tilde{e})\right] \gtrsim \frac{\varepsilon}{\delta^2} = \frac{\varepsilon}{\Delta(\tilde{e})^2}$$

for any $\tilde{e} \in E$ such that $\Delta(\tilde{e}) > 0$. By combining this with (38), we obtain (37). □

D MDPs with known transition

In this section, we analyze our algorithm for episodic MDPs with known transitions. Specifically, we prove Theorems 7 and 8, which together directly imply Corollary 2 in the main body. In addition, we provide our lower bound result in Theorem 9.

D.1 Algorithm

The algorithm's construction is almost identical to that of the shortest path case. The unbiased loss estimator is defined in Lemma 2.

Proof of Lemma 2. For notational simplicity, we omit the conditioning in expectations throughout this proof. Fix an arbitrary $s \in S_k$ and a. We then have

$$\begin{aligned}
&\mathbf{E}\left[c_t \cdot \frac{\mathbb{I}_t(s,a)}{q_t(s,a)}\right] = \mathbf{E}\left[\sum_{i=0}^{L-1} \ell(s_i^t, a_i^t) \mid \mathbb{I}_t(s,a) = 1\right] = \mathbf{E}\left[\sum_{i=0}^{L-1} \ell(s_i^t, a_i^t) \mid (s_k^t, a_k^t) = (s,a)\right] \\
&= \mathbf{E}\left[\sum_{i=0}^{k-1} \ell(s_i^t, a_i^t) \mid s_k^t = s\right] + \mathbf{E}\left[\sum_{i=k}^{L-1} \ell(s_i^t, a_i^t) \mid s_k^t = s, a_k^t = a\right] \\
&= \mathbf{E}\left[\sum_{i=0}^{k-1} \ell(s_i^t, a_i^t) \mid s_k^t = s\right] + Q^{\pi_t}(s,a;\ell_t).
\end{aligned}$$

Similarly, for any fixed $s \in S_k$, we have

$$\mathbf{E}\left[c_t \cdot \frac{\mathbb{I}_t(s)}{q_t(s)}\right] = \mathbf{E}\left[\sum_{i=0}^{L-1} \ell(s_i^t, a_i^t) \mid s_k^t = s\right] = \mathbf{E}\left[\sum_{i=0}^{k-1} \ell(s_i^t, a_i^t) \mid s_k^t = s\right] + V^{\pi_t}(s;\ell_t).$$

By combining the above two equalities, we obtain

$$\mathbf{E}\left[\widehat{\ell}_t(s,a)\right] = \mathbf{E}\left[c_t \cdot \frac{\mathbb{I}_t(s,a)}{q_t(s,a)}\right] - \mathbf{E}\left[c_t \cdot \frac{\mathbb{I}_t(s)}{q_t(s)}\right] = Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t).$$

□

In the following, we denote

$$\bar{\ell}_t := \mathbf{E}\left[\widehat{\ell}_t \mid \pi_t, \ell_t\right] = Q^{\pi_t}(s,a;\ell_t) - V^{\pi_t}(s;\ell_t) \in [-1,1]^{S \times A}.$$

By combining Lemmas 1 and 2, we obtain the following expression of the regret:

$$\begin{aligned}
\mathrm{Reg}_T(\pi^*) &= \mathbf{E}\left[\sum_{t=1}^{T} \left(V^{\pi_t}(s_0;\ell_t) - V^{\pi^*}(s_0;\ell_t)\right)\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T} \left(V^{\pi_t}(s_0;\bar{\ell}_t) - V^{\pi^*}(s_0;\bar{\ell}_t)\right)\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \bar{\ell}_t, q_t - q^* \right\rangle\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, q_t - q^* \right\rangle\right], \qquad (43)
\end{aligned}$$

where we denote $q^* = q^{\pi^*}$. To upper bound the value of $\sum_{t=1}^{T} \left\langle \widehat{\ell}_t, q_t - q^* \right\rangle$, we choose $q_t \in \mathcal{Q}$ by using the following FTRL approach similarly to the case of online shortest path problem:

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{Q}} \left\{ \left\langle \sum_{\tau=1}^{t-1} \widehat{\ell}_\tau, q \right\rangle + \psi_t(q) \right\},$$

where

$$\psi_t(q) = -\frac{2}{\eta_t} \sum_{s \neq s_L, a \in A} \sqrt{q(s,a)} - \sum_{s \neq s_L, a \in A} \beta \ln q(s,a) \quad \text{with} \quad \eta_t = \frac{1}{\sqrt{t}},\ \beta = 2, \quad \text{or} \qquad (44)$$

$$\psi_t(q) = -\sum_{s \neq s_L, a \in A} \frac{1}{\eta_t(s,a)} \ln q(s,a) \quad \text{with} \quad \eta_t(s,a) = \left(4 + \frac{1}{\ln T} \sum_{\tau=1}^{t-1} \rho_\tau(s,a)\right)^{-\frac{1}{2}}, \qquad (45)$$

where $\rho_\tau \in [0,1]$ will be defined later.

D.2 Regret analysis

In our regret analysis, we use the following lemma:

Lemma 13 (First part of Lemma 4)**.** *If $\widehat{\ell}_t$ is given by as in Lemma 2, the expectation of $\widehat{\ell}(s,a)^2$ taken w.r.t. the randomness of p satisfies*

$$\mathbf{E}\left[\widehat{\ell}_t(s,a)^2\right] \leq \frac{1-\pi_t(a|s)}{q_t(s,a)}.$$

Proof. For notational simplicity, we omit the subscript t throughout this proof. From the definition of $\widehat{\ell}$, we have

$$\begin{aligned}
\mathbf{E}\left[\widehat{\ell}(s,a)^2\right] &\leq q(s)\left(\pi(a|s)\left(\frac{1}{q(s,a)}-\frac{1}{q(s)}\right)^2+(1-\pi(a|s))\frac{1}{q(s)^2}\right)\\
&= \frac{q(s)}{q(s,a)^2}\left(\pi(a|s)(1-\pi(a|s))^2+(1-\pi(a|s))\pi(a|s)^2\right)\\
&= \frac{q(s)}{q(s,a)^2}\pi(a|s)(1-\pi(a|s)) = \frac{1-\pi(a|s)}{q(s,a)}.
\end{aligned}$$

□

We also use q_0 and $\tilde{q}^*$, which are defined as follows in the analysis: For all $s \in S$ and $a \in A$, let $q_{s,a} \in \arg\max_{q' \in \mathcal{Q}} q'(s,a)$. Define $q_0 \in \mathcal{Q}$ by $q_0 = \frac{1}{|S||A|}\sum_{s,a} q_{s,a}$. For $q^* = q^{\pi^*} \in \mathcal{Q}$ and $\varepsilon \in [0,1]$, define $\tilde{q}^*$ by

$$\tilde{q}^* = (1-\varepsilon)q^* + \varepsilon q_0. \tag{46}$$

Then, it holds for any $q \in \mathcal{Q}$, $s \in S$ and $a \in A$ that

$$\frac{q(s,a)}{\tilde{q}^*(s,a)} \leq \frac{1}{\varepsilon}\frac{q(s,a)}{q_0(s,a)} \leq \frac{|S||A|}{\varepsilon}\frac{q(s,a)}{q_{s,a}(s,a)} \leq \frac{|S||A|}{\varepsilon}. \tag{47}$$

From (43), by a similar way to (26), we can show that

$$\mathrm{Reg}_T(\pi^*) \leq \varepsilon T + \sum_{t=1}^{T}\mathbf{E}\left[\underbrace{\left\langle \widehat{\ell}_t, q_t - q_{t+1}\right\rangle - D_t(q_{t+1},q_t)}_{=:\mathrm{stab}_t} + \underbrace{(\psi_t(\tilde{q}^*) - \psi_{t-1}(\tilde{q}^*) - \psi_t(q_t) + \psi_{t-1}(q_t))}_{=:\mathrm{pena}_t}\right]. \tag{48}$$

D.2.1 Analysis for Tsallis-entropy case

Theorem 7. *For any deterministic policy $\pi^* \in \Pi$, the proposed algorithm with the Tsallis-entropy regularizer* (44) *achieves:*

$$\begin{aligned}
&\mathrm{Reg}_T(\pi^*)\\
&\lesssim \sum_{t=1}^{T}\frac{1}{\sqrt{t}}\mathbf{E}\left[\sum_{s\neq s_L}\sum_{a\neq\pi^*(s)}\sqrt{q_t(s,a)} + \sqrt{L|S||A|\sum_{s\neq s_L}\sum_{a\neq\pi^*(s)}(q_t(s,a)+q^*(s,a))}\right] + |S||A|\log T
\end{aligned}$$

In the proof of this theorem, we can bound the stability term by using the following lemma:

Lemma 14. *When we use the Tsallis-entropy regularizer* (17)*, stability terms are bounded as*

$$\mathbf{E}[\mathrm{stab}_t] \lesssim \mathbf{E}\left[\eta_t\sum_{s\neq s_L}\left(\sum_{a\neq\pi^*(s)}\sqrt{q_t(s,a)} - \sqrt{q_t(s)}\right)\right] \lesssim \mathbf{E}\left[\eta_t\sum_{s\neq s_L}\sum_{a\neq\pi^*(s)}\sqrt{q_t(s,a)}\right]. \tag{49}$$

This can be shown in a similar way as Lemma 9, by using Lemmas 6 and 13. Furthermore, the penalty term can be bounded in the same manner as done by Jin and Luo [2020], using their Lemma 6 with $\alpha = 0$. By combining these results, Theorem 7 can be established in the same way as Theorem 4.

D.2.2 Analysis for log-barrier case

In the case of the log-barrier regularizer, we have the following regret bounds:

Theorem 8. *For any deterministic policy π^*, the proposed algorithm with the log-barrier regularizer* (45) *achieves:*

$$\mathrm{Reg}_T \lesssim \mathbf{E}\left[\sum_{s\neq s_L}\sum_{a\neq\pi^*(s)}\sqrt{\sum_{t=1}^{T} c_t^2 \mathbb{I}_t(s)\left(\mathbb{I}_t(s,a)-\pi_t(a|s)\right)^2 \log(T)}\right] + |S||A|\log(T). \tag{50}$$

To show these, we use the following lemmas:

Lemma 15. *When we use the log-barrier regularizer* (45)*, we have*

$$\mathrm{stab}_t \lesssim c_t^2 \sum_{s\neq s_L}\sum_{a\in A} \eta_t(s,a)\mathbb{I}_t(s)\left(\mathbb{I}_t(s,a)-\pi_t(a|s)\right)^2.$$

Proof. We have

$$\begin{aligned}\mathrm{stab}_t &= \left\langle \widehat{\ell}_t, q_t - q_{t+1}\right\rangle - D_t(q_{t+1}, q_t)\\ &= \sum_{s,a}\left(\widehat{\ell}_t(s,a)(q_t(s,a)-q_{t+1}(s,a)) - \frac{1}{\eta_t(s,a)} D_\phi(q_{t+1}(s,a), q_t(s,a))\right),\end{aligned}$$

where D_ϕ is the Bregman divergence associated with $\phi(x) = -\ln(x)$. As we have

$$-\eta_t(s,a)q_t(s,a)\widehat{\ell}_t(s,a) \le \eta_t(s,a)q_t(s,a)\frac{c_t}{q_t(s)} \le \eta_t(s,a) \le \frac{1}{2},$$

we can apply Lemma 7 to obtain the following:

$$\mathrm{stab}_t \le \sum_{s,a}\eta_t(s,a)q_t(s,a)^2\widehat{\ell}_t(s,a)^2 = \sum_{s,a}\eta_t(s,a)c_t^2\mathbb{I}_t(s)\left(\mathbb{I}_t(s,a)-\pi_t(a|s)\right)^2.$$

□

Define $\rho_t(s,a)$ by

$$\rho_t(s,a) = c_t^2\mathbb{I}_t(s)\left(\mathbb{I}_t(s,a) - \frac{q_t(s,a)}{q_t(s)}\right)^2.$$

Then, Theorem 8 can be established in the same way as Theorem 5.

Lastly, results in Corollary 2 follows from Theorems 7 and 8 by an argument similar to that used for Corollaries 3 and 4.

D.3 Lower bound for stochastic MDPs

Consider stochastic environment in which c_t follows a Bernoulli distribution of parameter $\langle \ell^*, p_t\rangle$, where we assume that $\ell^* : S\times A \to [0,1]$ satisfies $\langle \ell^*, p\rangle \in [3/8, 5/8]$ for any possible trajectories p and $\ell^*(s_L, a) = 0$ for all $a \in A$. Define $\Delta : S\times A \to [0,1]$ by optimal Q function values:

$$\begin{aligned}Q^*(s_L,a) &= 0 && (a\in A),\\ Q^*(s,a) &= \ell^*(s,a) + \sum_{s'\in S} P(s'|s,a)\cdot V^*(s'), && (s\in S\setminus\{s_L\}, a\in A),\\ V^*(s) &= \min_{a'\in A} Q^*(s,a') && (s\in S),\\ \Delta(s,a) &= Q^*(s,a) - \min_{a'\in A} Q^*(s,a') = Q^*(s,a) - V^*(s) && (s\in S, a\in A).\end{aligned}$$

We then have $\min_{a\in A}\Delta(s,a)=0$ for all $s\in S$ and $\|\Delta\|_1\le 1/4$. Also, we have

$$\begin{aligned}
\langle \ell^*, q\rangle &= \sum_{k=0}^{L-1}\sum_{s\in S_k}\sum_{a\in A}\ell^*(s,a)q(s,a)\\
&= \sum_{k=0}^{L-1}\sum_{s\in S_k}\sum_{a\in A}\left(Q^*(s,a)-\sum_{s'\in S_{k+1}}P(s'|s,a)\cdot V^*(s')\right)q(s,a)\\
&= \sum_{k=0}^{L-1}\left(\sum_{s\in S_k}\sum_{a\in A}Q^*(s,a)q(s,a)-\sum_{s'\in S_{k+1}}q(s')\cdot V^*(s')\right)\\
&= \sum_{k=0}^{L-1}\left(\sum_{s\in S_k}\sum_{a\in A}\left(Q^*(s,a)-V^*(s)\right)q(s,a)\right)+V^*(s_0)-V^*(s_L)\\
&= \langle \Delta, q\rangle + \min_{q^*\in\mathcal{Q}}\langle \ell^*, q^*\rangle
\end{aligned}$$

as $V^*(s_L)=\ell^*(s_L,a)=0$ for all $a\in A$. Denote

$$\begin{aligned}
\mathcal{Q}^* &= \arg\min_{q\in\mathcal{Q}}\{\langle \ell^*, q\rangle\} = \arg\min_{q\in\mathcal{Q}}\{\langle \Delta, q\rangle\}\\
&= \{q\in\mathcal{Q}\mid \langle\Delta,q\rangle=0\} = \{q\in\mathcal{Q}\mid \Delta(s,a)>0\Longrightarrow q(s,a)=0\}.
\end{aligned}$$

We also define

$$S^*=\{s\in S\setminus\{s_L\}\mid \exists q\in\mathcal{Q}^*, q(s)>0\}.$$

Then, for any consistent algorithms, we have

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}_T}{\log T}\gtrsim\sum_{s\in S^*}\sum_{a\in A:\Delta(s,a)>0}\frac{1}{\Delta(s,a)}.$$

Theorem 9. *Consider an arbitrary consistent algorithm, i.e., assume that there exists $\varepsilon\in(0,1)$ such that $\mathrm{Reg}_T\le MT^{1-\varepsilon}$ holds for all instances, where $M>0$ is a parameter independent of T. Then for any MDP with $\ell^*: S\times A\to[0,1]$ satisfying $\langle\ell^*,p\rangle\in[3/8,5/8]$ for any possible trajectories p and $\ell^*(s_L,a)=0$ for all $a\in A$, we have*

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}_T}{\log T}\gtrsim\varepsilon\sum_{s\in S^*}\sum_{a\in A:\Delta(s,a)>0}\frac{1}{\Delta(s,a)}. \tag{51}$$

Proof. Let $q^*\in\mathcal{Q}^*$. As we have

$$\mathrm{Reg}_T=\mathbf{E}\left[\sum_{t=1}^T\langle\ell^*,p_t-q^*\rangle\right]=\mathbf{E}\left[\sum_{t=1}^T\langle\ell^*,q_t-q^*\rangle\right]=\sum_{s,a}\Delta(s,a)\mathbf{E}\left[\sum_{t=1}^T q_t(s,a)\right],$$

we have

$$\liminf_{T\to\infty}\frac{\mathrm{Reg}_T}{\log T}=\sum_{s,a}\Delta(s,a)\liminf_{T\to\infty}\frac{1}{\log T}\mathbf{E}\left[\sum_{t=1}^T q_t(s,a)\right]. \tag{52}$$

In the following, we evaluate the value of $\liminf_{T\to\infty}\frac{1}{\log T}\mathbf{E}\left[\sum_{t=1}^T q_t(s,a)\right]$ for any fixed $\tilde{s}\in S^*$ and $\tilde{a}\in A$ such that $\delta:=\Delta(\tilde{s},\tilde{a})>0$. Let $q^*\in\arg\max_{q\in\mathcal{Q}^*}\{q(\tilde{s})\}$ and denote $\bar{q}=\max_{q\in\mathcal{Q}^*}\{q(\tilde{s})\}=q^*(\tilde{s})>0$. Then, from Lemma 16 below, there exists $c\in(0,\delta]$ such that

$$\langle\Delta,q\rangle\ge c\max\{0,q(\tilde{s})-\bar{q}\}+\delta q(\tilde{s},\tilde{a}).$$

Consider a modified environment given by $\tilde{\ell}$ such that $\tilde{\ell}(\tilde{s},\tilde{a})=\ell^*(\tilde{s},\tilde{a})-\delta-c/2$ and $\tilde{\ell}(s,a)=\ell^*(s,a)$ for $(s,a)\neq(\tilde{s},\tilde{a})$. Then, as we have $q^*(\tilde{s},\tilde{a})=0$, it holds for any $q\in\mathcal{Q}$ that

$$\begin{aligned}
\left\langle\tilde{\ell},q-q^*\right\rangle &= \langle\ell^*,q-q^*\rangle-\left(\delta+\frac{c}{2}\right)(q(\tilde{s},\tilde{a})-q^*(\tilde{s},\tilde{a}))\\
&= \langle\Delta,q\rangle-\left(\delta+\frac{c}{2}\right)q(\tilde{s},\tilde{a})\ge c\max\{0,q(\tilde{s})-\bar{q}\}-\frac{c}{2}q(\tilde{s},\tilde{a})\\
&\ge c\max\{0,q(\tilde{s},\tilde{a})-\bar{q}\}-\frac{c}{2}q(\tilde{s},\tilde{a})=\frac{c}{2}\left(|q(\tilde{s},\tilde{a})-\bar{q}|-\bar{q}\right)
\end{aligned}\tag{53}$$

Further, as there exists $q \in \mathcal{Q}$ such that $\left\langle \tilde{\ell}, q - q^* \right\rangle = -\frac{c\bar{q}}{2}$, we have

$$\min_{q \in \mathcal{Q}} \left\langle \tilde{\ell}, q \right\rangle = \left\langle \tilde{\ell}, q^* \right\rangle - \frac{c\bar{q}}{2}. \tag{54}$$

In fact, such an occupancy measure q can be constructed from a corresponding policy $\tilde{\pi} : S \to A$, by modifying a policy $\pi^* : S \to A$ corresponding to $q^* \in \mathcal{Q}^*$ such that $q^*(s) = \bar{q}$. We set $\tilde{\pi}(s) = \pi^*(s)$ for all $s \neq \tilde{s}$ and set $\tilde{\pi}(\tilde{s}) = \tilde{a}$. An occupancy measure q corresponding to $\tilde{\pi}$ satisfies $q(\tilde{s}, \tilde{a}) = \bar{q}$ and $\langle \ell^*, q \rangle = \bar{q}\delta$, which implies that $\left\langle \tilde{\ell}, q \right\rangle = \frac{c\bar{q}}{2}$. Hence, the regret for the environment given by $\tilde{\ell}$ satisfies

$$\begin{aligned}
\tilde{\mathrm{Reg}}_T &= \max_{q \in \mathcal{Q}} \tilde{\mathbf{E}} \left[\sum_{t=1}^{T} \left\langle \tilde{\ell}, q_t - q \right\rangle \right] \\
&= \tilde{\mathbf{E}} \left[\sum_{t=1}^{T} \left(\left\langle \tilde{\ell}, q_t - q^* \right\rangle + \frac{c\bar{q}}{2} \right) \right] && \text{(from (54))} \\
&\geq \frac{c}{2} \tilde{\mathbf{E}} \left[\sum_{t=1}^{T} |q_t(\tilde{s}, \tilde{a}) - \bar{q}| \right] && \text{(from (53))} \\
&\geq \frac{c}{2} \tilde{\mathbf{E}} \left[\sum_{t=1}^{T} \max\{0, \bar{q} - q_t(\tilde{s}, \tilde{a})\} \right], && (55)
\end{aligned}$$

where $\tilde{\mathbf{E}}[\cdot]$ represents the expected value when feedback is generated by an environment associated with $\tilde{\ell}$. On the other hand, the regret for the environment given by ℓ^* satisfies

$$\mathrm{Reg}_T \geq \delta \cdot \mathbf{E} \left[\sum_{t=1}^{T} q_t(\tilde{s}, \tilde{a}) \right] \geq \delta \cdot \mathbf{E} \left[\sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \right]. \tag{56}$$

Let TV denote the total variation distance between trajectories $((q_t, p_t, c_t))_{t=1}^T$ for environments with ℓ^* and $\tilde{\ell}$. Then, as we have $\frac{1}{\bar{q}T} \sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \in [0, 1]$, we have

$$\left| \tilde{\mathbf{E}} \left[\frac{1}{\bar{q}T} \sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \right] - \mathbf{E} \left[\frac{1}{\bar{q}T} \sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \right] \right| \leq \mathrm{TV}\,.$$

We hence have

$$\begin{aligned}
1 - \mathrm{TV} &\leq 1 - \tilde{\mathbf{E}} \left[\frac{1}{\bar{q}T} \sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \right] + \mathbf{E} \left[\frac{1}{\bar{q}T} \sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \right] \\
&= \tilde{\mathbf{E}} \left[\frac{1}{\bar{q}T} \sum_{t=1}^{T} \max\{0, \bar{q} - q_t(\tilde{s}, \tilde{a})\} \right] + \mathbf{E} \left[\frac{1}{\bar{q}T} \sum_{t=1}^{T} \min\{\bar{q}, q_t(\tilde{s}, \tilde{a})\} \right] \\
&\leq \frac{1}{\bar{q}T} \left(\frac{2}{c} \tilde{\mathrm{Reg}}_T + \frac{1}{\delta} \mathrm{Reg}_T \right),
\end{aligned}$$

where the last inequality follows from (55) and (56). Here, from the Bretagnolle-Huber inequality (e.g., [Canonne [2022], Corollary 4]) and the chain rule of the KL divergence, we have

$$\begin{aligned}
1 - \mathrm{TV} &\geq \frac{1}{2} \exp\left(\sum_{t=1}^{T} \mathbf{E} \left[D_{\mathrm{KL}} \left(\langle \ell^*, p_t \rangle \,||\, \langle \tilde{\ell}, p_t \rangle \right) \right] \right) \\
&= \frac{1}{2} \exp\left(\sum_{t=1}^{T} \mathbf{E} \left[q_t(\tilde{s}, \tilde{a})\, D_{\mathrm{KL}} \left(\langle \ell^*, p_t \rangle \,||\, \langle \ell^*, p_t \rangle - \delta - \frac{c}{2} \right) \right] \right) \\
&\geq \frac{1}{2} \exp\left(-5 \left(\delta + \frac{c}{2} \right)^2 \cdot \mathbf{E} \left[\sum_{t=1}^{T} q_t(\tilde{s}, \tilde{a}) \right] \right).
\end{aligned}$$

Combining the above inequalities and $c \in (0, \delta]$, we obtain

$$\mathbf{E}\left[\sum_{t=1}^{T} q_t(\tilde{s}, \tilde{a})\right] \gtrsim \frac{1}{\delta^2} \log \frac{1}{2(1 - \mathrm{TV})} \geq \frac{1}{\delta^2} \log \frac{\bar{q}cT}{4(\mathrm{Reg}_T + \widetilde{\mathrm{Reg}}_T)}$$
$$\geq \frac{1}{\delta^2} \log \frac{\bar{q}cT}{8MT^{1-\varepsilon}} = \frac{1}{\delta^2} \left(\varepsilon \log T + \log\left(\frac{\bar{q}c}{8M}\right)\right)$$

Consequently, we have

$$\liminf_{T \to \infty} \frac{1}{\log T} \mathbf{E}\left[\sum_{t=1}^{T} q_t(\tilde{s}, \tilde{a})\right] \gtrsim \frac{\varepsilon}{\delta^2} = \frac{\varepsilon}{\Delta(\tilde{s}, \tilde{a})^2}$$

for any $\tilde{s} \in S^*$ and $\tilde{a} \in A$ such that $\Delta(\tilde{s}, \tilde{a}) > 0$. By combining this with (52), we obtain (51). □

Lemma 16. *Suppose that $s \in S^*$ and denote $\bar{q} = \max_{q \in \mathcal{Q}^*} \{q(s)\}$. Then, there exists $c > 0$ such that the following holds for all $q \in \mathcal{Q}$:*

$$\langle \Delta, q \rangle \geq c \max\{0, q(s) - \bar{q}\} + \sum_{a \in A} \Delta(s, a) q(s, a).$$

Proof. Suppose that $s \in S^* \cap S_k$. Decompose Δ as $\Delta = \Delta_{<k} + \Delta_{\geq k}$, where

$$(\Delta_{<k}(s, a), \Delta_{\geq k}(s, a)) = \begin{cases} (\Delta(s, a), 0) & \text{if } s \in \bigcup_{k' < k} S_{k'} \\ (0, \Delta(s, a)) & \text{if } s \in \bigcup_{k' \geq k} S_{k'} \end{cases}.$$

We define $q_{<k}$ and $q_{\geq k}$ in the same way. Define $f(x) = \inf_{q \in \mathcal{Q}: q(s) = x} \langle \Delta, q \rangle$. We then have

$$f(x) = \inf_{q \in \mathcal{Q}: q(s) = x} \langle \Delta, q \rangle = \inf_{q \in \mathcal{Q}: q(s) = x} \langle \Delta_{<k} + \Delta_{\geq k}, q \rangle = \inf_{q \in \mathcal{Q}: q(s) = x} \langle \Delta_{<k}, q \rangle .$$

The last equality follows from the fact that, for any $q \in \mathcal{Q}$ (corresponding to $\pi \in \Pi$) such that $q(s) = x$, there exists $q' \in \mathcal{Q}$ such that $q(s) = x$, $\langle \Delta_{<k}, q' \rangle = \langle \Delta_{<k}, q \rangle$, and $\langle \Delta_{\geq k}, q' \rangle = 0$. Such an occupancy measure q' can be constructed by a policy $\pi' \in \Pi$ given as $\pi'(s, a) = \pi(s, a)$ for $s \in \bigcup_{k' < k} S_{k'}$ and $\pi'(s, a) = \pi^*(s, a)$ for $s \in \bigcup_{k' \geq k} S_{k'}$.

Define $\underline{x} = \min_{q \in \mathcal{Q}} \{q(s)\}$ and $\bar{x} = \max_{q \in \mathcal{Q}} \{q(s)\}$. We note that $\underline{x} \leq \bar{q} \leq \bar{x}$ and $f(x) < +\infty$ if and only if $\underline{x} \leq x \leq \bar{x}$. As $\mathcal{Q}$ is a polytope, $f(x)$ is a piecewise linear function in x, i.e., there exists a finite sequence of real numbers $x_0 = \underline{x} < x_1 < x_2 < \cdots < x_n = \bar{x} \in \mathbb{R}$ such that $f(x)$ is an affine function in each interval $[x_i, x_{i+1}]$. From the definition of $\mathcal{Q}^*$ and $\bar{q}$, we have $f(x) > 0$ for any $x > \bar{q}$. Indeed, if $q(s) > \bar{q}$ then $q \notin \mathcal{Q}^*$, which means that $\langle \Delta, q \rangle > 0$. Hence, $c \in \mathbb{R}$ defined as

$$c = \inf \left\{ \frac{f(x_i)}{x_i - \bar{q}} \mid i \in [n], x_i > \bar{q} \right\}$$

is positive. (When $\bar{q} = \bar{x}$, i.e., when there is no $x_i > \bar{q}$, we let c be an arbitrary positive number.) Further, as $f(x) \geq 0$ for all x and $f(x)$ is affine in each interval $[x_i, x_{i+1}]$, we have $f(x) \geq c \max\{0, x - \bar{q}\}$ for all $x \leq \bar{x}$. From this, we have

$$\langle \Delta, q \rangle = \langle \Delta_{<k}, q \rangle + \langle \Delta_{\geq k}, q \rangle \geq f(q(s)) + \langle \Delta_{\geq k}, q \rangle \geq c \max\{0, q(s) - \bar{q}\} + \sum_{a \in A} \Delta(s, a) q(s, a).$$

□

E Algorithm for MDPs with unknown transition

In this section, we present the details for our best-of-both-worlds algorithm for MDPs with unknown transitions. Similar to Jin et al. [2021], our algorithm proceeds in epochs. In each epoch, we execute FTRL using our novel loss estimator and the current empirical estimates of the transitions. At the end of the epoch, we update these empirical estimates. We refer the reader to Algorithm 1 for full details. In this section, we use $\mathbf{E}_t[\cdot]$ to denote the conditional expectation $\mathbf{E}[\cdot | \mathcal{F}_{t-1}]$, where $\mathcal{F}_{t-1}$ is the past filtration.

E.1 Confidence set of the true transition

In this section, we use the same confidence sets used by prior works Jin et al. [2020, 2021].

For each epoch i, we define the empirical transition $\bar{P}_i$ as:

$$\bar{P}_i(s'|s,a) = \frac{m_i(s,a,s')}{m_i(s,a)},\ \forall (s,a,s') \in S_k \times A \times S_{k+1}, k = 0,\dots,L-1. \tag{57}$$

For each epoch i, we define the confidence width B_i as follows:

$$B_i(s,a,s') = \min\left\{2\sqrt{\frac{\bar{P}_i(s'|s,a)\ln\left(\frac{T|S||A|}{\delta}\right)}{m_i(s,a)} + \frac{14\ln\left(\frac{T|S||A|}{\delta}\right)}{m_i(s,a)}}, 1\right\}, \tag{58}$$

where δ is some confidence parameter.

For each epoch i, the confidence set $\mathcal{P}_i$ of the true transition is defined as follows:

$$\mathcal{P}_i = \left\{\widehat{P} : \left|\widehat{P}(s'|s,a) - \widehat{P}_i(s'|s,a)\right| \le B_i(s,a,s'), \forall (s,a,s') \in S_k \times A \times S_{k+1}, k < L\right\}. \tag{59}$$

As shown in Lemma 2 of Jin et al. [2020], the true transition P lies in the confidence $\mathcal{P}_i$ for all epoch i with probability at least $1 - 4\delta$.

E.2 Loss estimator and Regularizer

We begin by presenting our novel loss estimator:

$$\ell_t^u(s,a) = \frac{c_t \cdot \mathbb{I}_t(s,a) + (1 - \pi_t(a \mid s) - c_t) \cdot \mathbb{I}_t(s)\pi_t(a \mid s)}{u_t(s,a)} - (1 - \pi_t(a \mid s)), \tag{60}$$

where $u_t(s,a)$ denotes the upper occupancy measure of (s,a) under policy π_t, and is defined as

$$u_t(s,a) = \max_{\widehat{P} \in \mathcal{P}_{i(t)}} q^{\widehat{P},\pi_t}(s,a), \tag{61}$$

where $i(t)$ denotes the epoch to which round t belongs. Note that $u_t(s,a)$ can be efficiently computed using the COMP-UOB procedure proposed in [Jin et al., 2020].

To build intuition for why this estimator enables best-of-both-worlds guarantees, we now consider the corresponding loss estimator in the setting with known transitions.

$$\ell_t^q(s,a) = \frac{c_t \cdot \mathbb{I}_t(s,a) + (1 - \pi_t(a|s) - c_t) \cdot \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)} - (1 - \pi_t(a|s)) \tag{62}$$

In the unknown transition case under semi-bandit feedback, Jin et al. [2021] considered the loss estimator $\frac{\ell_t(s,a)\cdot\mathbb{I}_t(s,a)}{u_t(s,a)}$, which ensures that $\ell_t(s,a) - \mathbf{E}_t\left[\frac{\ell_t(s,a)\cdot\mathbb{I}_t(s,a)}{u_t(s,a)}\right] \ge 0$ whenever $u_t(s,a) \ge q_t(s,a)$. This inequality plays a key role in establishing their best-of-both-worlds result.

In our setting, the role of $\ell_t(s,a)$ is played by the pseudo-loss $\bar{\ell}_t(s,a) := \mathbf{E}_t[\ell_t^q(s,a)]$. When $u_t(s,a) \ge q_t(s,a)$, we show an analogous inequality: $\bar{\ell}_t(s,a) - \mathbf{E}_t[\ell_t^u(s,a)] \ge 0$, which is similarly crucial for our analysis.

We begin by analyzing the pseudo-loss $\bar{\ell}_t(s,a)$ as follows:

$$\begin{aligned}
\bar{\ell}_t(s,a) &:= \mathbf{E}_t\left[\ell_t^q(s,a)\right] \\
&= \mathbf{E}_t\left[\frac{c_t \cdot \mathbb{I}_t(s,a) - c_t \cdot \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\right] + \mathbf{E}_t\left[\frac{(1 - \pi_t(a|s))\mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\right] - (1 - \pi_t(a|s)) \\
&= Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s) - (1 - \pi_t(a|s)) \\
&= Q^{\pi_t}(s,a) - V^{\pi_t}(s).
\end{aligned}$$

By Lemma 1, we have $\langle \bar{\ell}_t, q_t - q^* \rangle = \langle \ell_t, q_t - q^* \rangle$, implying that the pseudo-loss $\bar{\ell}_t$ can indeed play the role of ℓ_t in our setting.

Next, we compute the expectation of $\ell_t^u(s,a)$ as follows:

$$\mathbf{E}_t\left[\ell_t^u(s,a)\right] = \frac{q_t(s)}{u_t(s)}\left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)).$$

To analyze this expression, observe that:

$$\begin{aligned} Q^{\pi_t}(s,a) - V^{\pi_t}(s) &= Q^{\pi_t}(s,a) - \sum_{a' \in A} \pi_t(a' \mid s) Q^{\pi_t}(s,a') \\ &= (1-\pi_t(a|s)) Q^{\pi_t}(s,a) - \sum_{a' \neq a} \pi_t(a' \mid s) Q^{\pi_t}(s,a') \\ &\geq -(1-\pi_t(a|s)), && \text{(as } Q^{\pi_t}(s,a) \in [0,1]) \end{aligned}$$

and thus:

$$Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s) \geq 0.$$

Now, under the condition that $u_t(s) \geq q_t(s)$, we have $\frac{q_t(s)}{u_t(s)} \leq 1$, and therefore:

$$\begin{aligned} \mathbf{E}_t\left[\ell_t^u(s,a)\right] &= \frac{q_t(s)}{u_t(s)}\left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) \\ &\leq \left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) \\ &= Q^{\pi_t}(s,a) - V^{\pi_t}(s) \\ &= \bar{\ell}_t(s,a). \end{aligned}$$

That is, $\ell_t^u(s,a)$ is an optimistic estimator of the pseudo-loss $\bar{\ell}_t(s,a)$, which plays a crucial role in our regret analysis.

In addition to bounding the expectation, we also analyze the second moment of the loss estimator. This analysis is facilitated by the careful introduction of the term $(1-\pi_t(a \mid s))$ into our loss estimator, which yields the following bound:

$$\begin{aligned} \mathbf{E}_t\left[\ell_t^u(s,a)^2\right] &\lesssim \mathbf{E}_t\left[\frac{1}{u_t(s,a)^2}\left(c_t^2 \cdot (\mathbb{I}_t(s,a) - \mathbb{I}_t(s)\pi_t(a \mid s))^2 + (1-\pi_t(a \mid s))^2 \cdot \mathbb{I}_t(s)\pi_t(a \mid s)^2\right)\right] \\ &\quad + (1-\pi_t(a \mid s))^2 \\ &\lesssim (1-\pi_t(a \mid s))\left(\frac{q_t(s)}{u_t(s)} \cdot \frac{1}{u_t(s,a)} + 1 - \pi_t(a \mid s)\right). \end{aligned}$$

Using this bound, we obtain a control on the stability term in the regret analysis:

$$\mathbf{E}_t\left[\widehat{q}_t(s,a)^{3/2} \cdot \ell_t^u(s,a)^2\right] \lesssim \widehat{q}_t(s,a)^{1/2}(1-\pi_t(a \mid s)).$$

This upper bound plays a crucial role in enabling a regret analysis based on self-bounding terms. In particular, it allows us to bypass the loss-shifting technique employed in Jin et al. [2021], while still controlling the stability term.

To leverage this upper bound, we use the following regularizer in epoch i:

$$\phi_t(q) = -\frac{1}{\eta_t}\sum_{s \neq s_L}\sum_{a \in A}\sqrt{q(s,a)} - \beta \sum_{s \neq s_L}\sum_{a \in A} \ln q(s,a), \tag{63}$$

where $\eta_t = \frac{1}{\sqrt{t-t_i+1}}$ and $\beta = 1024L$. The log-barrier term is included to stabilize the updates, following the approach of Jin and Luo [2020].

E.3 Main Result

Theorem 10. *In the bandit feedback setting, Algorithm 1 with $\delta = \frac{1}{T^3}$ and $\iota = \frac{|S||A|T}{\delta}$ guarantees $\mathrm{Reg}_T(\pi^\star) = \tilde{\mathcal{O}}\left(L|S|\sqrt{|A|T} + |S||A|\sqrt{LT} + L^2|S|^3|A|^2\right)$ and simultaneously $\mathrm{Reg}_T(\pi^\star) = \mathcal{O}\left(U + \sqrt{UC} + V\right)$ under Condition (4), where $V = L^2|S|^3|A|^2\ln^2\iota$ and U is defined as*

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

We defer the proof of the above theorem to Appendix F.

F Analysis of BOBW with unknown transitions

For any time-step t, let $i(t)$ denote the epoch the time-step t is part of. Let $\mathbf{E}_t[\cdot] := \mathbf{E}[\cdot|\mathcal{F}_{t-1}]$ be the conditional expectation, where $\mathcal{F}_{t-1}$ is the past filtration. Recall the definition of ℓ_t^u and ℓ_t^q from Eq. (60) and Eq. (62) respectively. Also recall that $\widehat{\ell}_t = \ell_t^u - B_{i(t)}$. Let $\bar{\ell}_t$ be a pseudo-loss such that $\bar{\ell}_t(s,a) := \mathbf{E}_t[\ell_t^q(s,a)] = Q^{\pi_t}(s,a) - V^{\pi_t}(s)$.

Now we define the conditional expectation of $\widehat{\ell}_t$ as follows:

$$\tilde{\ell}_t(s,a) := \mathbf{E}_t[\widehat{\ell}_t(s,a)] = \frac{q_t(s)}{u_t(s)}\left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) - B_{i(t)}(s,a). \tag{64}$$

Definition 4. For any policy π, the estimated state-action and state value functions associated with $\bar{P}_{i(t)}$ and loss function $\tilde{\ell}_t$ are defined as:

$$\begin{aligned}
\widetilde{Q}_t^\pi(s,a) &= \tilde{\ell}_t(s,a) + \sum_{s' \in \mathcal{S}_{K(s)+1}} \bar{P}_{i(t)}(s'|s,a)\widetilde{V}_t^\pi(s'), \quad \forall (s,a) \in (S \setminus \{s_L\}) \times A, \\
\widetilde{V}_t^\pi(s) &= \sum_{a \in A} \pi(a|s)\widetilde{Q}_t^\pi(s,a), \quad \forall s \in S, \\
\widetilde{Q}_t^\pi(s_L,a) &= 0, \quad \forall a \in A.
\end{aligned}$$

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

$$\begin{aligned}
Q_t^\pi(s,a) &= \ell_t(s,a) + \sum_{s' \in \mathcal{S}_{K(s)+1}} P(s'|s,a)V_t^\pi(s'), \quad \forall (s,a) \in (S \setminus \{s_L\}) \times A, \\
V_t^\pi(s) &= \sum_{a \in A} \pi(a|s)Q_t^\pi(s,a), \quad \forall s \in S, \\
Q_t^\pi(s_L,a) &= 0, \quad \forall a \in A.
\end{aligned}$$

where P denotes the true transition function.

Moreover, we define pseudo state-action and value functions as follows:

$$\begin{aligned}
\bar{Q}_t^\pi(s,a) &= \bar{\ell}_t(s,a) + \sum_{s' \in \mathcal{S}_{K(s)+1}} P(s'|s,a)\bar{V}_t^\pi(s'), \quad \forall (s,a) \in (S \setminus \{s_L\}) \times A, \\
\bar{V}_t^\pi(s) &= \sum_{a \in A} \pi(a|s)\bar{Q}_t^\pi(s,a), \quad \forall s \in S, \\
\bar{Q}_t^\pi(s_L,a) &= 0, \quad \forall a \in A.
\end{aligned}$$

Let $\mathcal{A}$ be the event that $P \in \mathcal{P}_i$ for all epochs $i \geq 1$. Moreover, we also define $\mathcal{A}_i$ to be the event $P \in \mathcal{P}_i$.Note that the value of $\mathbb{1}\{\mathcal{A}_i\}$ gets determined based on observations prior to epoch i only. Let $\iota = \frac{T|\mathcal{S}||\mathcal{A}|}{\delta}$ and let $\delta = \frac{1}{T^3} \in (0,1)$.

We decompose the regret against policy π as follows:

$$\begin{aligned}
\mathrm{Reg}(\pi) &= \mathbf{E}\left[\sum_{t=1}^T V_t^{\pi_t}(s_0) - V_t^\pi(s_0)\right] \\
&= \mathbf{E}\left[\sum_{t=1}^T \bar{V}_t^{\pi_t}(s_0) - \bar{V}_t^\pi(s_0)\right] && \text{(due to Lemma 1)} \\
&= \mathbf{E}\Bigg[\underbrace{\sum_{t=1}^T \bar{V}_t^{\pi_t}(s_0) - \widetilde{V}_t^{\pi_t}(s_0)}_{\textsc{Err}_1}\Bigg] + \mathbf{E}\Bigg[\underbrace{\sum_{t=1}^T \widetilde{V}_t^{\pi_t}(s_0) - \widetilde{V}_t^{\pi}(s_0)}_{\textsc{EstReg}}\Bigg] + \mathbf{E}\Bigg[\underbrace{\sum_{t=1}^T \widetilde{V}_t^{\pi}(s_0) - \bar{V}_t^{\pi}(s_0)}_{\textsc{Err}_2}\Bigg].
\end{aligned} \tag{65}$$

Note that, the second term (restated below) is controlled by the FTRL process.

$$\mathbf{E}[\textsc{EstReg}] = \mathbf{E}\left[\sum_{t=1}^{T}\left\langle q^{\bar{P}_{i(t)},\pi_t} - q^{\bar{P}_{i(t)},\pi}, \tilde{\ell}_t\right\rangle\right] = \mathbf{E}\left[\sum_{t=1}^{T}\left\langle q^{\bar{P}_{i(t)},\pi_t} - q^{\bar{P}_{i(t)},\pi}, \widehat{\ell}_t\right\rangle\right]. \tag{66}$$

F.1 Auxiliary lemmas

We often use the following lemma to handle the small-probability event $\mathcal{A}^c$ while taking the expectation.

Lemma 17 (Jin et al. [2021])**.** *Suppose that a random variable X satisfies the following conditions:*

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

Then, we have

$$\mathbf{E}\left[X\right] \leq D \cdot \Pr\left[\mathcal{E}^c\right] + \mathbf{E}\left[Y\right].$$

We next restate the performance difference lemma.

Lemma 18 (Performance difference lemma)**.** *Suppose $\bar{\ell}$ is defined by $\bar{\ell}(s,a) = Q^{\pi}(s,a;\ell) - V^{\pi}(s;\ell)$ for some $\pi \in \Pi$ and for all $s \in S$ and $a \in A$. We then have*

$$V^{\pi'}(s;\bar{\ell}) = V^{\pi'}(s;\ell) - V^{\pi}(s;\ell), \quad Q^{\pi'}(s,a;\bar{\ell}) = Q^{\pi'}(s,a;\ell) - V^{\pi}(s;\ell), \tag{67}$$

for any $\pi' \in \Pi$, $s \in S$ and $a \in A$.

We immediately get following corollary.

Corollary 5. *$\forall (s,a) \in S \times A$, we have $-1 \leq \bar{V}_t^{\pi}(s) \leq 1$ and $-1 \leq \bar{Q}_t^{\pi}(s,a) \leq 1$.*

We next state the following lemma.

Lemma 19. *If event $\mathcal{A}$ holds, then $\sum_{s' \in S_{k(s)+1}} \left(\bar{P}_{i(t)}(s'|s,a) - P(s'|s,a)\right) \bar{V}_t^{\pi}(s') - B_{i(t)}(s,a) \leq 0$.*

Proof. When $B_{i(t)}(s,a) = 2$, we have

$$\begin{aligned}
&\sum_{s' \in S_{k(s)+1}} \left(\bar{P}_{i(t)}(s'|s,a) - P(s'|s,a)\right) \bar{V}_t^{\pi}(s') - B_{i(t)}(s,a) \\
&\leq \sum_{s' \in S_{k(s)+1}} \bar{P}_{i(t)}(s'|s,a) \cdot 1 + \sum_{s' \in S_{k(s)+1}} P_{i(t)}(s'|s,a) \cdot 1 - 2 = 0,
\end{aligned}$$

where the inequality follows from the fact $-1 \leq \bar{V}_t^{\pi}(s') \leq 1$.

On the other hand, when $\sum_{s' \in S_{k(s)+1}} B_{i(t)}(s,a,s') = B_{i(t)}(s,a)$, we have

$$\begin{aligned}
&\sum_{s' \in S_{k(s)+1}} \left(\bar{P}_{i(t)}(s'|s,a) - P(s'|s,a)\right) \bar{V}_t^{\pi}(s') - B_{i(t)}(s,a) \\
&\leq \sum_{s' \in S_{k(s)+1}} B_{i(t)}(s,a,s') \cdot 1 - B_{i(t)}(s,a) = 0,
\end{aligned}$$

where the second line follows from the definition of event $\mathcal{A}$.

□

We next state the following proposition

Proposition 1. *For all $(s,a) \in S \times A$, we have $0 \leq Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s) \leq 2$.*

Proof. As we have

$$Q^{\pi_t}(s,a) - V^{\pi_t}(s) = Q^{\pi_t}(s,a) - \sum_{a'} \pi_t(a'|s) Q^{\pi_t}(s,a') \tag{68}$$

$$= (1-\pi_t(a|s)) Q^{\pi_t}(s,a) - \sum_{a' \neq a} \pi_t(a'|s) Q^{\pi_t}(s,a') \geq -(1-\pi_t(a|s)), \tag{69}$$

we get $Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s) \geq 0$.
As $0 \leq Q^{\pi_t}(s,a) \leq 1$, we also have $Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s) \leq Q^{\pi_t}(s,a) + 1 \leq 2$. □

We next state the following proposition.

Proposition 2. *If event $\mathcal{A}$ holds, $\tilde{\ell}_t(s,a) \leq \bar{\ell}_t(s,a) - B_{i(t)}(s,a)$ for all $(s,a) \in S \times A$.*

Proof. The following holds by Proposition 1, given that the event $\mathcal{A}$ occurs:

$$\begin{aligned}
\tilde{\ell}_t(s,a) &= \frac{q_t(s)}{u_t(s)} \left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) - B_{i(t)}(s,a) \\
&\leq \left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) - B_{i(t)}(s,a) \\
&= \bar{\ell}_t(s,a) - B_{i(t)}(s,a).
\end{aligned}$$

□

We next state the following proposition.

Proposition 3. *If event $\mathcal{A}$ holds, $-3 \leq \tilde{\ell}_t(s,a) \leq 1$ for all $(s,a) \in S \times A$.*

Proof. Due to Proposition 2 and the total loss of any trajectory is between 0 and 1, we have $\tilde{\ell}_t(s,a) \leq \bar{\ell}_t(s,a) - B_{i(t)}(s,a) \leq Q^{\pi_t}(s,a) \leq 1$. On the otherhand, due to Proposition 1:

$$\begin{aligned}
\tilde{\ell}_t(s,a) &= \frac{q_t(s)}{u_t(s)} \left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) - B_{i(t)}(s,a) \\
&\geq -(1-\pi_t(a|s)) - B_{i(t)}(s,a) \\
&\geq -3.
\end{aligned}$$

□

We next state the following lemma.

Lemma 20. *If event $\mathcal{A}$ holds, the following holds:*

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

Specifically, we have:

$$\left\langle q^{\bar{P}_{i(t),\pi}}, \tilde{\ell}_t \right\rangle = \tilde{V}_t^{\pi}(s_0) \leq \bar{V}_t^{\pi}(s_0) = \left\langle q^{P,\pi}, \bar{\ell}_t \right\rangle.$$

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

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

Induction step: Assume that $\widetilde{Q}_t^\pi(s,a) \le Q_t^\pi(s,a)$ holds for all states s with $k(s) > h$. Then, for any state s with $k(s) = h$, we have

$$\begin{aligned}
\widetilde{Q}_t^\pi(s,a) &= \bar{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} \bar{P}_{i(t)}(s'|s,a)\widetilde{V}_t^\pi(s') - B_{i(t)}(s,a) && \text{(due to Proposition 2)} \\
&\le \bar{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} \bar{P}_{i(t)}(s'|s,a)\bar{V}_t^\pi(s') - B_{i(t)}(s,a) && \text{(induction hypothesis)} \\
&\le \bar{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} P(s'|s,a)\bar{V}_t^\pi(s') \\
&\quad + \sum_{s' \in S_{k(s)+1}} \left(\bar{P}_{i(t)}(s'|s,a) - P(s'|s,a)\right)\bar{V}_t^\pi(s') - B_{i(t)}(s,a) \\
&\le \bar{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} P(s'|s,a)\bar{V}_t^\pi(s') && \text{(due to Lemma 19)} \\
&\le \bar{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} P(s'|s,a)\bar{V}^\pi(s') = \bar{Q}_t^\pi(s,a).
\end{aligned}$$

This concludes the induction. □

The following lemma follows directly from Lemma C.1.2 in Jin et al. [2021].

Lemma 21 (Jin and Luo [2020])**.** *Algorithm 1 ensures $u_t(s) \ge \frac{1}{|S|t}$ for all t and s.*

Lemma 22. *Algorithm 1 ensures the following:*

$$\left|\widehat{\ell}_t(s,a)\right| \le 3 + \frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)} \cdot |S|t.$$

We also have:

$$\mathbf{E}\left[\frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\middle|\mathcal{A}_{i(t)}\right] = \mathbf{E}\left[\frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\middle|\mathcal{A}_{i(t)}^c\right] = 2.$$

Proof. Due to Lemma 21, we have the following:

$$\begin{aligned}
\left|\widehat{\ell}_t(s,a)\right| &\le 3 + \frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{u_t(s)\cdot\pi_t(a|s)} \\
&\le 3 + \frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s)\cdot\pi_t(a|s)} \cdot |S|t \\
&= 3 + \frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)} \cdot |S|t.
\end{aligned}$$

Next, we have:

$$\mathbf{E}\left[\frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\middle|\mathcal{A}_{i(t)}\right] = \mathbf{E}\left[\mathbf{E}_t\left[\frac{\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\right]\middle|\mathcal{A}_{i(t)}\right] = \mathbf{E}\left[2\middle|\mathcal{A}_{i(t)}\right] = 2$$

Similarly, we can show that $\mathbf{E}\left[\frac{\mathbb{I}_t(s,a)+\mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\middle|\mathcal{A}_{i(t)}^c\right] = 2$. □

Lemma 23. *Algorithm 1 ensures the following:*

$$\left|\tilde{\ell}_t(s,a)\right| \le 6|S|t,\ \forall (s,a) \in S \times A, t \in [T].$$

Proof. Due to Eq. (64), we have the following:

$$\begin{aligned}
\left|\tilde{\ell}_t(s,a)\right| &= \left|\frac{q_t(s)}{u_t(s)}\left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right) - (1-\pi_t(a|s)) - B_{i(t)}(s,a)\right| \\
&\le \frac{2q_t(s)}{u_t(s)} + 3 && \text{(due to Proposition 1)} \\
&\le 2|S|t + 3 && \text{(due to Lemma 21)} \\
&\le 6|S|t.
\end{aligned}$$

□

We immediately get the following corollary.

Corollary 6. *Algorithm 1 ensures the following:*

$$\left|\tilde{Q}_t^{\pi}(s,a)\right| \leq 6L|S|t,\ \forall (s,a) \in S \times A, t \in [T].$$

Let $\phi_H(q) = -\sum_{s \neq s_L} \sum_{a \in A} \sqrt{q(s,a)}$ and $\phi_L(q) = -\beta \sum_{s \neq s_L} \sum_{a \in A} \ln q(s,a)$. Recall that $\phi_t(q) = \phi_H(q) + \phi_L(q)$ and $\beta = 1024L$. Now we prove the following proposition.

Proposition 4. *If event $\mathcal{A}$ holds, then $||\widehat{\ell}_t||_{(\nabla^2 \phi_t(\widehat{q}_t))^{-1}} \leq \frac{1}{8}$.*

Proof. We have the following:

$$\begin{aligned}
||\widehat{\ell}_t||^2_{(\nabla^2 \phi_t(\widehat{q}_t))^{-1}} &\leq ||\widehat{\ell}_t||^2_{(\nabla^2 \phi_L(\widehat{q}_t))^{-1}} && (\text{as } \nabla^2 \phi_L(\widehat{q}_t) \preceq \nabla^2 \phi_t(\widehat{q}_t)) \\
&\leq \frac{1}{\beta} \cdot \sum_{s \neq s_L} \sum_{a \in A} \left(\frac{2(\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi(s,a))^2}{u_t(s,a)^2} + 2(-(1-\pi_t(a|s)) + B_i(s,a))^2 \right) \cdot \widehat{q}_t(s,a)^2 \\
& && (\text{as } (a+b)^2 \leq 2a^2 + 2b^2) \\
&\leq \frac{1}{\beta} \cdot \sum_{s \neq s_L} \sum_{a \in A} \left(\frac{4(\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi(s,a))}{u_t(s,a)^2} + 8 \right) \cdot \widehat{q}_t(s,a)^2 \\
& \quad (\text{as } \mathbb{I}_t(s), \mathbb{I}_t(s,a), \pi(s,a) \in [0,1] \text{ and } -(1-\pi_t(a|s)) + B_i(s,a) \in [-1,2]) \\
&\leq \frac{1}{\beta} \cdot \sum_{s \neq s_L} \sum_{a \in A} (4(\mathbb{I}_t(s,a) + \mathbb{I}_t(s)\pi(s,a)) + 8\widehat{q}_t(s,a)) \quad (\text{as } u_t(s,a) \leq \widehat{q}_t(s,a)) \\
&= \frac{1}{\beta} \cdot \sum_{s \neq s_L} (8\mathbb{I}_t(s) + 8\widehat{q}_t(s)) \\
&= \frac{1}{\beta} \cdot (16L) \\
&= \frac{1}{64}.
\end{aligned}$$

□

We now state the following lemma, which follows from arguments identical to those in Lemma C.1.8 of Jin et al. [2021], and provides an upper bound for $\mathbf{E}\left[\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \widehat{q}_t(s,a) \cdot B_{i(t)}(s,a)^2\right]$.

Lemma 24. *Algorithm 1 ensures the following:*

$$\mathbf{E}\left[\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \widehat{q}_t(s,a) \cdot B_{i(t)}(s,a)^2\right] = \mathcal{O}\left(L^2|S|^3|A|^2 \ln^2 \iota + |S||A|T \cdot \delta\right). \tag{70}$$

Finally, we state the following lemma on the learning rates and the number of epochs.

Lemma 25 (Jin et al. [2021])**.** *According to the design of the learning rate $\eta_t = \frac{1}{\sqrt{t - t_{i(t)} + 1}}$, the following inequalities hold:*

$$\sum_{t=1}^{T} \eta_t^2 \leq \mathcal{O}\left(|S||A| \log^2 T\right), \tag{71}$$

$$\sum_{t=1}^{T} \eta_t \leq \mathcal{O}\left(\sqrt{|S||A|T \log T}\right). \tag{72}$$

$$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \leq \eta_t \quad \forall t \geq 2 \tag{73}$$

Moreover, Algorithm 1 ensures that $N \leq 4|S||A|(\log T + 1)$ where N is the number of epochs.

F.2 Technical lemmas to analyze EstReg

We defined the estimated regret in each epoch i as follows:

$$\textsc{EstReg}_i(\pi) = \mathbf{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\left\langle q^{\bar{P}_i,\pi_t} - q^{\bar{P}_i,\pi}, \widehat{\ell}_t\right\rangle\right] = \mathbf{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\left\langle \widehat{q}_t - q^{\bar{P}_i,\pi}, \widehat{\ell}_t\right\rangle\right].$$

Lemma 26. *With $\beta = 1024L$, for any epoch i, Algorithm 1 ensures*

$$\textsc{EstReg}_i(\pi) \le \mathcal{O}\left(\mathbf{E}\left[\sqrt{L|S||A|}\cdot\sum_{t=t_i}^{t_{i+1}-1}\eta_t\right] + L|S||A|\log T + \delta\cdot\mathbf{E}\left[L|S|T\left(t_{i+1}-t_i\right)\right]\right) \tag{74}$$

for any policy π, and simultaneously

$$\begin{aligned}\mathbf{E}\left[\textsc{EstReg}_i(\pi)\right] \le{}& \mathcal{O}\left(\mathbf{E}\left[\sqrt{L|S|}\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sqrt{\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\widehat{q}_t(s,a)}\right]\right)\\ &+\mathcal{O}\left(\mathbf{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)}\right]\right)\\ &+\mathcal{O}\left(\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\cdot B_{i(t)}(s,a)^2\right)\\ &+\mathcal{O}\left(L|S||A|\log T + \delta\cdot\mathbf{E}\left[L|S|T\left(t_{i+1}-t_i\right)\right]\right)\end{aligned} \tag{75}$$

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

Proof. If the event $\mathcal{A}_i$ does not hold, we have the following:

$$\begin{aligned}\sum_{t=t_i}^{t_{i+1}-1}\left\langle \widehat{q}_t - q, \widehat{\ell}_t\right\rangle &\le \sum_{t=t_i}^{t_{i+1}-1}\sum_{s\neq s_L}\sum_{a\in A}\left(\widehat{q}_t(s,a)+q(s,a)\right)\cdot\left|\widehat{\ell}_t(s,a)\right|\\ &\le \sum_{t=t_i}^{t_{i+1}-1}\sum_{s\neq s_L}\sum_{a\in A}\left(\widehat{q}_t(s,a)+q(s,a)\right)\cdot\left(3+\frac{\mathbb{I}_t(s,a)+\mathbb{I}_t(s)\pi_t(a|s)}{u_t(s,a)}\cdot|S|t\right)\\ &\qquad\text{(due to Lemma 22)}\\ &\le |S|T\cdot\sum_{t=t_i}^{t_{i+1}-1}\sum_{s\neq s_L}\sum_{a\in A}\left(\widehat{q}_t(s,a)+q(s,a)\right)\cdot\frac{\mathbb{I}_t(s,a)+\mathbb{I}_t(s)\pi_t(a|s)}{q_t(s,a)}\\ &\quad+6L\cdot\left(t_{i+1}-t_i\right).\end{aligned}$$

Due to the second part of Lemma 22, we have the following:

$$\begin{aligned}&\mathbf{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\left\langle \widehat{q}_t - q, \widehat{\ell}_t\right\rangle\middle|\mathcal{A}_i^c\right]\\ &\le \mathbf{E}\left[(6L+4L|S|T)\cdot\left(t_{i+1}-t_i\right)|\mathcal{A}_i^c\right]\\ &\le \mathcal{O}\Big(\mathbf{E}\left[L|S|T\cdot\left(t_{i+1}-t_i\right)|\mathcal{A}_i^c\right]\Big).\end{aligned}$$

Recall that $\phi_H(q) = -\sum_{s\neq s_L}\sum_{a\in A}\sqrt{q(s,a)}$. Also recall that $\widehat{Q}_t(s,a) = \widehat{\ell}_t(s,a) + \mathbf{E}_{s'\sim\bar{P}(\cdot|s,a)}[\widehat{V}_t(s')]$ and $\widehat{V}_t(s) = \mathbf{E}_{a\sim\pi_t(\cdot|s)}[\widehat{Q}_t(s,a)]$ (with $\widehat{V}_L(s_L) = 0$). Due to Proposition 4,

we get the following by using the same argument as [Jin and Luo, 2020, Lemma 5]:

$$\begin{aligned}
&\sum_{t=t_i}^{t_{i+1}-1} \left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle \\
&= \mathcal{O}\left(L|S||A|\log T\right) + \sum_{t=t_i+1}^{t_{i+1}-1} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\left(\phi_H(q) - \phi_H(\widehat{q}_t)\right) \\
&\quad + 8\sum_{t=t_i}^{t_{i+1}-1} \eta_t \sum_{s\neq s_L}\sum_{a\in A} \widehat{q}_t(s,a)^{3/2}\widehat{\ell}_t(s,a)^2. \qquad (76)
\end{aligned}$$

Now, we condition on the event $\mathcal{A}_i$. Recall the definition of ℓ_t^u from Eq. (60). Now we have the following:

$$\begin{aligned}
\widehat{q}_t(s,a)^{3/2}\ell_t^u(s,a)^2 &\leq \frac{4\widehat{q}_t(s,a)^{3/2}}{u_t(s,a)^2}\left(c_t^2\cdot\left(\mathbb{I}_t(s,a) - \mathbb{I}_t(s)\pi_t(a|s)\right)^2 + (1-\pi_t(a|s))^2\mathbb{I}_t(s)\pi_t(a|s)^2\right) \\
&\quad + 2\widehat{q}_t(s,a)^{3/2}(1-\pi_t(a|s))^2 \qquad \text{(as } (a+b)^2 \leq 2a^2+2b^2\text{)} \\
&\leq \frac{4\widehat{q}_t(s,a)^{1/2}}{q_t(s,a)}\left(\left(\mathbb{I}_t(s,a) - \mathbb{I}_t(s)\pi_t(a|s)\right)^2 + (1-\pi_t(a|s))\mathbb{I}_t(s)\pi_t(a|s)\right) \\
&\quad + 2\widehat{q}_t(s,a)^{3/2}(1-\pi_t(a|s))^2, \qquad (77)
\end{aligned}$$

where we get the last inequality as as $u_t(s) \leq q_t(s)$ and $u_t(s,a) \leq \widehat{q}_t(s,a)$.

As $\mathbf{E}_t[(\mathbb{I}_t(s,a) - \mathbb{I}_t(s)\pi_t(a|s))^2] = q(s)(1-\pi_t(a|s))\pi_t^2(a|s) + q(s)\pi_t(a|s)(1-\pi_t(a|s))^2$ and $\mathbf{E}_t[\mathbb{I}_t(s)] = q_t(s)$, we have the following:

$$\begin{aligned}
&\mathbf{E}_t\left[\frac{4\widehat{q}_t(s,a)^{1/2}}{q_t(s,a)}\left(\left(\mathbb{I}_t(s,a) - \mathbb{I}_t(s)\pi_t(a|s)\right)^2 + (1-\pi_t(a|s))\mathbb{I}_t(s)\pi_t(a|s)\right) + 2\widehat{q}_t(s,a)^{3/2}(1-\pi_t(a|s))^2\right] \\
&\leq \frac{12\widehat{q}_t(s,a)^{2/2}\cdot q_t(s)\cdot(1-\pi_t(a|s))\cdot\pi_t(a|s)}{q_t(s,a)} + 2\widehat{q}_t(s,a)^{3/2}(1-\pi_t(a|s))^2 \\
&\leq 14\widehat{q}_t(s,a)^{1/2}\cdot(1-\pi_t(a|s)). \qquad (78)
\end{aligned}$$

We now proceed to prove Eq. (74) and Eq. (75).

Proving Eq. (74) In this case, we consider the second term inside the minimum in Eq. (76), and derive a straightforward upper bound to $\phi_H(q) - \phi_H(\widehat{q}_t)$ using the Cauchy-Schwarz inequality, yielding $\phi_H(q) - \phi_H(\widehat{q}_t) \leq \sum_{s\neq s_L}\sum_{a\in A}\sqrt{\widehat{q}_t(s,a)} \leq \sqrt{L|S||A|}$. This gives

$$\begin{aligned}
&\sum_{t=t_i}^{t_{i+1}-1} \left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle \\
&\leq \mathcal{O}\left(L|S||A|\log T\right) + \sqrt{L|S||A|}\cdot\sum_{t=t_i}^{t_{i+1}-1}\eta_t + 8\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\widehat{\ell}_t(s,a)^2 \\
&\qquad\qquad \left(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} \leq \eta_t \text{ since } \tfrac{1}{\eta_t} = \sqrt{t-t_i+1}\right) \\
&\leq \mathcal{O}\left(L|S||A|\log T\right) + \sqrt{L|S||A|}\cdot\sum_{t=t_i}^{t_{i+1}-1}\eta_t + 16\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\ell_t^u(s,a)^2 \\
&\quad + 16\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\cdot B_{i(t)}(s,a)^2 \\
&\leq \mathcal{O}\left(L|S||A|\log T + \sqrt{L|S||A|}\cdot\sum_{t=t_i}^{t_{i+1}-1}\eta_t + \sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\ell_t^u(s,a)^2\right). \\
&\qquad\qquad \text{(as } B_{i(t)}(s,a)^2 \leq 4\text{)}
\end{aligned}$$

Therefore, by Lemma 17, Eq. (77), Eq. (78), and tower rule, we have for any policy π that,

$$\begin{aligned}\mathbf{E}\left[\textsc{EstReg}_i(\pi)\right] &\leq \mathcal{O}\left(\mathbf{E}\left[\sqrt{L|S||A|}\cdot\sum_{t=t_i}^{t_{i+1}-1}\eta_t + \sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{1/2}\cdot(1-\pi_t(a|s))\right]\right)\\ &\quad+\mathcal{O}\left(L|S||A|\log T+\delta\cdot\mathbf{E}\left[L|S|T\left(t_{i+1}-t_i\right)\right]\right)\\ &\leq \mathcal{O}\left(\mathbf{E}\left[\sqrt{L|S||A|}\cdot\sum_{t=t_i}^{t_{i+1}-1}\eta_t\right]+L|S||A|\log T+\delta\cdot\mathbf{E}\left[L|S|T\left(t_{i+1}-t_i\right)\right]\right),\end{aligned}$$

where the second step follows from $\sum_{s\neq s_L}\sum_{a\in A}\sqrt{\widehat{q}_t(s,a)}\leq\sqrt{L|S||A|}$. This completes the proof of Eq. (74).

Proving Eq.(75) In this case, note that since π is a deterministic policy, it follows that

$$\phi_H(q)-\phi_H(\widehat{q}_t)=\sum_{s\neq s_L}\sqrt{\widehat{q}_t(s)}\left(\sum_{a\in A}\sqrt{\pi_t(a|s)}-1\right)+\sum_{s\neq s_L}\left(\sqrt{\widehat{q}_t(s)}-\sqrt{q(s)}\right).$$

By applying [Jin and Luo, 2020, Lemma 16] with $\alpha=0$ to upper bound the first term, and using [Jin and Luo, 2020, Lemma 19] to upper bound the second term, we arrive at

$$\phi_H(q)-\phi_H(\widehat{q}_t)=\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)}+\sqrt{L|S|\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\widehat{q}_t(s,a)}.$$

Therefore, considering the Eq. (76) and using the inequalities $\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\leq\eta_t$ from Lemma 25 and $(a+b)^2\leq 2a^2+2b^2$, we have

$$\begin{aligned}\sum_{t=t_i}^{t_{i+1}-1}\left\langle\widehat{q}_t-q,\widehat{\ell}_t\right\rangle &\leq \sqrt{L|S|}\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sqrt{\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\widehat{q}_t(s,a)}\\ &\quad+\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)}\\ &\quad+16\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\ell_t^u(s,a)^2\\ &\quad+16\sum_{t=t_i}^{t_{i+1}-1}\eta_t\cdot\sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\cdot B_{i(t)}(s,a)^2\\ &\quad+\mathcal{O}\left(L|S||A|\log T\right).\end{aligned}\tag{79}$$

Next, observe that for a deterministic policy π, we have:

$$\begin{aligned}\sum_{s\neq s_L}\sum_{a\in A}\sqrt{\widehat{q}_t(s,a)}\left(1-\pi_t(a|s)\right) &\leq \sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)}+\sum_{s\neq s_L}\sqrt{\widehat{q}_t(s)}\left(1-\pi_t(\pi(s)|s)\right)\\ &=\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)}+\sum_{s\neq s_L}\sqrt{\widehat{q}_t(s)}\left(\sum_{a\neq\pi(s)}\pi_t(a|s)\right)\\ &\leq 2\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)},\end{aligned}\tag{80}$$

Therefore, by Eq. (79), Lemma 17, Eq. (77), Eq. (78), Eq. (80) and tower rule, we have for any deterministic policy $\pi : S \to A$ that,

$$\begin{aligned}
\mathbf{E}\left[\textsc{EstReg}_i(\pi)\right] \leq \mathcal{O}\left(\mathbf{E}\left[\sqrt{L|S|}\sum_{t=t_i}^{t_{i+1}-1}\eta_t \cdot \sqrt{\sum_{s\neq s_L}\sum_{a\neq\pi(s)}\widehat{q}_t(s,a)}\right]\right) \\
+ \mathcal{O}\left(\mathbf{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\eta_t \cdot \sum_{s\neq s_L}\sum_{a\neq\pi(s)}\sqrt{\widehat{q}_t(s,a)}\right]\right) \\
+ \mathcal{O}\left(\sum_{t=t_i}^{t_{i+1}-1}\eta_t \cdot \sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2}\cdot B_{i(t)}(s,a)^2\right) \\
+ \mathcal{O}\left(L|S||A|\log T + \delta\cdot\mathbf{E}\left[L|S|T\left(t_{i+1}-t_i\right)\right]\right)
\end{aligned}$$

□

F.3 Adversarial regret guarantee

Recall from Eq. (65) that the regret decomposes as:

$$\mathrm{Reg}(\pi) = \mathbf{E}\Bigg[\underbrace{\sum_{t=1}^{T}\bar{V}_t^{\pi_t}(s_0) - \widetilde{V}_t^{\pi_t}(s_0)}_{\textsc{Err}_1}\Bigg] + \mathbf{E}\Bigg[\underbrace{\sum_{t=1}^{T}\widetilde{V}_t^{\pi_t}(s_0) - \widetilde{V}_t^{\pi}(s_0)}_{\textsc{EstReg}}\Bigg] + \mathbf{E}\Bigg[\underbrace{\sum_{t=1}^{T}\widetilde{V}_t^{\pi}(s_0) - \bar{V}_t^{\pi}(s_0)}_{\textsc{Err}_2}\Bigg].$$

We now analyse each term separately.

First, we have the following:

$$\begin{aligned}
\textsc{Err}_1 &= \sum_{t=1}^{T}\left\langle q_t, \bar{\ell}_t\right\rangle - \left\langle \widehat{q}_t, \tilde{\ell}_t\right\rangle \\
&= \sum_{t=1}^{T}\sum_{s\neq s_L}\sum_{a\in A}\frac{(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s))\widehat{q}_t(s,a)}{u_t(s,a)}\cdot\left(u_t(s,a) - q_t(s,a)\right) \\
&\quad + \sum_{t=1}^{T}\left\langle q_t - \widehat{q}_t, \bar{\ell}_t\right\rangle + \sum_{t=1}^{T}\left\langle \widehat{q}_t, B_{i(t)}\right\rangle,
\end{aligned}$$

where the second equality follows from Eq. (64) and $\bar{\ell}_t(s,a) = Q^{\pi_t}(s,a) - V^{\pi_t}(s)$.

Due to the above equality, proposition 1, Lemma 17, the fact that $\widehat{q}_t(s,a) \leq u_t(s,a)$ under event $\mathcal{A}$ and analysis of $\textsc{Err}_1$ from Appendix C.2 Jin et al. [2021], we get the following:

$$\mathbf{E}[\textsc{Err}_1] = \tilde{\mathcal{O}}\left(L|S|\sqrt{|A|T} + L^2|S|^3|A|^2\right)$$

Next, we have the following due to Lemma 20, Lemma 17, Corollary 5, and Corollary 6:

$$\mathbf{E}[\textsc{Err}_2] = \tilde{\mathcal{O}}\left(L|S|T^2\cdot\delta\right)$$

According to Eq. (74) of Lemma 26, we have

$$\begin{aligned}
\textsc{EstReg}(\pi) &= \mathbf{E}\left[\sum_{t=1}^{T}\left\langle \widehat{q}_t - q^{\bar{P}_{i(t)},\pi}, \widehat{\ell}_t\right\rangle\right] = \mathbf{E}\left[\sum_{i=1}^{N}\textsc{EstReg}_i(\pi)\right] \\
&\leq \mathcal{O}\left(\mathbf{E}\left[\sqrt{L|S||A|}\cdot\sum_{i=1}^{N}\sum_{t=t_i}^{t_{i+1}-1}\eta_t\right] + L|S|^2|A|^2\log^2 T + \delta\cdot L|S|T^2\right) \\
&\qquad\qquad \text{(due to Eq. (74) and Lemma 25)} \\
&\leq \tilde{\mathcal{O}}\left(|S||A|\sqrt{LT} + L|S|^2|A|^2\right). \qquad \text{(due to Eq. (72))}
\end{aligned}$$

Finally, by combining the bounds for $\textsc{Err}_1$, $\textsc{Err}_2$, and $\textsc{EstReg}$, we obtain:

$$\mathrm{Reg}(\pi) = \tilde{\mathcal{O}}\left(L|S|\sqrt{|A|T} + |S||A|\sqrt{LT} + L^2|S|^3|A|^2\right).$$

F.4 Stochastic regret guarantee

F.4.1 Self-bounding terms and related lemma

In this section, we adopt the definition of self-bounding terms and the related lemmas from Jin et al. [2021].

Definition 5 (Self-bounding Terms)**.** For some mapping $\pi^\star : S \to A$, define the following:

$$\mathbb{Q}_1(J) = \sum_{t=1}^{T}\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)} q_t(s,a)\sqrt{\frac{J}{\max\left\{m_{i(t)}(s,a)\right\}}},$$

$$\mathbb{Q}_2(J) = \sum_{t=1}^{T}\sum_{s\neq s_L}\sum_{a=\pi^\star(s)} (q_t(s,a) - q_t^\star(s,a))\sqrt{\frac{J}{\max\left\{m_{i(t)}(s,a),1\right\}}},$$

$$\mathbb{Q}_3(J) = \sum_{t=1}^{T}\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\sum_{k=0}^{k(s)-1}\sum_{(u,v,w)\in T_k} q_t(u,v)\sqrt{\frac{P(w|u,v)\cdot J}{\max\left\{m_{i(t)}(u,v),1\right\}}}q_t(s,a|w),$$

$$\mathbb{Q}_4(J) = \sqrt{J\cdot\sum_{t=1}^{T}\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)} q_t(s,a)},$$

$$\mathbb{Q}_5(J) = \sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\sqrt{J\sum_{t=1}^{T} q_t(s,a)},$$

$$\mathbb{Q}_6(J) = \sum_{t=1}^{T}\sum_{s\neq s_L}\sum_{a=\pi^\star(s)}\frac{q_t(s,a)-q_t^\star(s,a)}{q_t(s,a)}\left(\sum_{k=0}^{k(s)-1}\sum_{(u,v,w)\in T_k} q_t(u,v)\sqrt{\frac{P(w|u,v)\cdot J}{\max\left\{m_{i(t)}(u,v),1\right\}}}q_t(s,a|w)\right).$$

Lemma 27. *Suppose Condition* (4) *holds. Then we have for any* $\alpha \in \mathbb{R}_+$*,*

$$\mathbf{E}\left[\mathbb{Q}_1(J)\right] \le \alpha\cdot(\mathrm{Reg}_T(\pi^\star)+C) + \frac{1}{\alpha}\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\frac{8J}{\Delta(s,a)}.$$

Lemma 28. *Suppose Condition* (4) *holds. Then we have for any* $\beta \in \mathbb{R}_+$*,*

$$\mathbf{E}\left[\mathbb{Q}_2(J)\right] \le \beta\cdot(\mathrm{Reg}_T(\pi^\star)+C) + \frac{1}{\beta}\cdot\frac{8|S|LJ}{\Delta_{\textsc{min}}}.$$

Lemma 29. *Suppose Condition* (4) *holds. Then we have for any* $\alpha,\beta \in \mathbb{R}_+$*,*

$$\mathbf{E}\left[\mathbb{Q}_3(J)\right] \le (\alpha+\beta)\cdot(\mathrm{Reg}_T(\pi^\star)+C) + \frac{1}{\alpha}\cdot\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\frac{8L^2|S|J}{\Delta(s,a)} + \frac{1}{\beta}\cdot\frac{8L^2|S|^2J}{\Delta_{\textsc{min}}}.$$

Lemma 30. *Suppose Condition* (4) *holds. Then we have for any* $\beta \in \mathbb{R}_+$*,*

$$\mathbf{E}\left[\mathbb{Q}_4(J)\right] \le \beta\cdot(\mathrm{Reg}_T(\pi^\star)+C) + \frac{1}{\beta}\cdot\frac{J}{4\Delta_{\textsc{min}}}.$$

Lemma 31. *Suppose Condition* (4) *holds. Then we have for any* $\alpha \in \mathbb{R}_+$*,*

$$\mathbf{E}\left[\mathbb{Q}_5(J)\right] \le \alpha\cdot(\mathrm{Reg}_T(\pi^\star)+C) + \sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\frac{J}{4\alpha\Delta(s,a)}.$$

Lemma 32. *Suppose Condition* (4) *holds. Then we have for any* $\beta \in \mathbb{R}_+$*,*

$$\mathbf{E}\left[\mathbb{Q}_6(J)\right] \le \beta\cdot(\mathrm{Reg}_T(\pi^\star)+C) + \frac{1}{\beta}\cdot\frac{8L^3|S|^2|A|\cdot J}{\Delta_{\textsc{min}}}.$$

F.4.2 Proof for the stochastic world

Similarly to the proof in Appendix C.3 of Jin et al. [2021], we decompose the sum of $\textsc{Err}_1$ and $\textsc{Err}_2$ into four terms ErrSub, ErrOpt, OccDiff and Bias:

$$\begin{aligned}
\textsc{Err}_1 + \textsc{Err}_2 = &\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} q_t(s,a) \bar{E}_t^{\pi^\star}(s,a) && \text{(ErrSub)} \\
&+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a = \pi^\star(s)} \left(q_t(s,a) - q_t^\star(s,a)\right) \bar{E}_t^{\pi^\star}(s,a) && \text{(ErrOpt)} \\
&+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \left(q_t(s,a) - \widehat{q}_t(s,a)\right) \left(\widetilde{Q}_t^{\pi^\star}(s,a) - \widetilde{V}_t^{\pi^\star}(s)\right) && \text{(OccDiff)} \\
&+ \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} q_t^\star(s,a) \left(\widetilde{V}_t^{\pi^\star}(s) - \bar{V}_t^{\pi^\star}(s)\right) && \text{(Bias)}
\end{aligned}$$

where $\bar{E}_t^\pi$ is defined as

$$\bar{E}_t^\pi(s,a) = \bar{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} P(s'|s,a) \widetilde{V}_t^\pi(s') - \widetilde{Q}_t^\pi(s,a).$$

We now begin by upper bounding $\mathbf{E}\left[\textsc{OccDiff}\right]$. First observe that we have the following:

$$\begin{aligned}
\textsc{OccDiff} &= \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \in A} \left(q_t(s,a) - \widehat{q}_t(s,a)\right) \left(\tilde{Q}_t^{\pi^\star}(s,a) - \tilde{V}_t^{\pi^\star}(s)\right) \\
&= \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \left(q_t(s,a) - \widehat{q}_t(s,a)\right) \left(\tilde{Q}_t^{\pi^\star}(s,a) - \tilde{V}_t^{\pi^\star}(s)\right) \\
&\leq \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \left|q_t(s,a) - \widehat{q}_t(s,a)\right| \cdot \left|\tilde{Q}_t^{\pi^\star}(s,a) - \tilde{V}_t^{\pi^\star}(s)\right|,
\end{aligned}$$

Under event $\mathcal{A}$, we further have $\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \left|q_t(s,a) - \widehat{q}_t(s,a)\right| \cdot \left|\tilde{Q}_t^{\pi^\star}(s,a) - \tilde{V}_t^{\pi^\star}(s)\right| \leq 5L \sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \left|q_t(s,a) - \widehat{q}_t(s,a)\right|$ as $-3 \leq \tilde{\ell}_t(s,a) \leq 1$ under event $\mathcal{A}$. If event $\mathcal{A}$ doesn't hold, we have $\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \left|q_t(s,a) - \widehat{q}_t(s,a)\right| \cdot \left|\tilde{Q}_t^{\pi^\star}(s,a) - \tilde{V}_t^{\pi^\star}(s)\right| \leq 12L|S|^2|A|T^2$ due to Corollary 6. Hence due to Lemma 17, we have the following

$$\begin{aligned}
\mathbf{E}[\textsc{OccDiff}] &\leq \mathcal{O}\left(L \cdot \mathbf{E}[\sum_{t=1}^{T} \sum_{s \neq s_L} \sum_{a \neq \pi^\star(s)} \left|q_t(s,a) - \widehat{q}_t(s,a)\right|] + \delta \cdot L|S|^2|A|T^2\right) \\
&\leq \mathcal{O}\left(L^3|S|^3|A|^2 \ln^2 \iota + \delta \cdot L|S|^2|A|T^2 + \mathbb{Q}_3(L^2 \ln \iota)\right)
\end{aligned}$$

where the last line follows from the definition of $\mathbb{Q}_3$ and Lemma D.3.10 of Jin et al. [2021].

Next, due to Lemma 17, Lemma 20, and Corollary 6, we have $\mathbf{E}\left[\textsc{Bias}\right] \leq \delta \cdot \mathcal{O}(L|S|^2AT^2)$. The first two terms ErrSub and ErrOpt are bounded differently. First, under event $\mathcal{A}$, we have

$$\begin{aligned}
\bar{E}_t^{\pi^\star}(s,a) &= \bar{\ell}_t(s,a) - \tilde{\ell}_t(s,a) + \sum_{s' \in S_{k(s)+1}} \left(P(s'|s,a) - \bar{P}_{i(t)}(s'|s,a)\right) \widetilde{V}_t^{\pi^\star}(s') \\
&= \left(Q^{\pi_t}(s,a) - V^{\pi_t}(s) + 1 - \pi_t(a|s)\right)\left(1 - \frac{q_t(s,a)}{u_t(s,a)}\right) \\
&\quad + B_{i(t)}(s,a) + \sum_{s' \in S_{k(s)+1}} \left(P(s'|s,a) - \bar{P}_{i(t)}(s'|s,a)\right) \widetilde{V}_t^{\pi^\star}(s') \\
&\leq \frac{2(u_t(s,a) - q_t(s,a))}{q_t(s,a)} + 4L \cdot B_{i(t)}(s,a)
\end{aligned}$$

where the last line uses the definition of the event $\mathcal{A}$ along with the fact that $q_t(s,a) \leq u_t(s,a)$ and $-3 \leq \tilde{\ell}_t(s,a) \leq 1$ under this event. Next, observe that the range of $\bar{E}_t^{\pi}$ is $\mathcal{O}(L|S|t)$, as established by Proposition 1 and Corollary 6, which implies that the range of both ErrSub and ErrOpt is $\mathcal{O}(L^2|S|T^2)$. Thus, it suffices to add a term of order $\mathcal{O}(\delta \cdot L^2|S|T^2)$ to account for the event $\mathcal{A}^c$.

By the exact same analysis as in Appendix C.3 and Appendix B.2 of Jin et al. [2021] and the fact that $|\tilde{V}(s)| \leq \mathcal{O}(L)$ under event $\mathcal{A}$, we have the following:

$$\begin{aligned}
\mathbf{E}\left[\textsc{ErrSub}\right] &= \mathcal{O}\left(\mathbf{E}\left[\mathbb{Q}_3(\ln \iota) + \mathbb{Q}_1(L^2|S|\ln \iota)\right] + L^2|S|^3|A|^2 \ln^2 \iota\right) \\
\mathbf{E}\left[\textsc{ErrOpt}\right] &= \mathcal{O}\left(\mathbf{E}\left[\mathbb{Q}_6(\ln \iota) + \mathbb{Q}_2(L^2|S|\ln \iota)\right] + L^2|S|^3|A|^2 \ln^2 \iota\right).
\end{aligned}$$

We are now left with bounding EstReg using self-bounding terms.

Term EstReg Based on Eq. (75) from Lemma 26, summing over all epochs gives the following upper bound for $\mathbf{E}\left[\textsc{EstReg}\right]$:

$$\begin{aligned}
&\mathcal{O}\left(\mathbf{E}\left[\sqrt{|S|L}\sum_{i=1}^{N}\sum_{t=t_i}^{t_{i+1}-1}\eta_t \cdot \sqrt{\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\widehat{q}_t(s,a)}\right] + \mathbf{E}\left[\sum_{i=1}^{N}\sum_{t=t_i}^{t_{i+1}-1}\eta_t \cdot \sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\sqrt{\widehat{q}_t(s,a)}\right]\right) \\
&\quad + \mathcal{O}\left(\sum_{i=1}^{N}\sum_{t=t_i}^{t_{i+1}-1}\eta_t \cdot \sum_{s\neq s_L}\sum_{a\in A}\widehat{q}_t(s,a)^{3/2} \cdot B_{i(t)}(s,a)^2\right) + \mathcal{O}\left(L|S|^2|A|^2\log^2 T + \delta \cdot L|S|T^2\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(due to Lemma 25)} \\
&= \mathcal{O}\left(\mathbf{E}\left[\sqrt{|S|L}\sum_{t=1}^{T}\eta_t \cdot \sqrt{\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\widehat{q}_t(s,a)}\right]\right) + \mathcal{O}\left(\mathbf{E}\left[\sum_{t=1}^{T}\eta_t \cdot \sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\sqrt{\widehat{q}_t(s,a)}\right]\right) \\
&\quad + \mathcal{O}\left(L^2|S|^3|A|^2\ln^2\iota\right). \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(due to Lemma 24)}
\end{aligned}$$

By the exact same analysis as in Appendix C.3 of Jin et al. [2021], we bound the first term as follows:

$$\begin{aligned}
&\mathbf{E}\left[\sqrt{|S|L}\sum_{t=1}^{T}\eta_t \cdot \sqrt{\sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\widehat{q}_t(s,a)}\right] \\
&= \mathcal{O}\Big(\mathbf{E}\left[\mathbb{Q}_4(L|S|^2|A|\log^2 T) + \mathbb{Q}_3\left(\ln\iota\right)\right] + L^2|S|^3|A|^2\ln^2\iota\Big).
\end{aligned}$$

Again by using the exact same analysis as in Appendix C.3 of Jin et al. [2021], we bound the second term as follows:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=1}^{T}\eta_t \cdot \sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\sqrt{\widehat{q}_t(s,a)}\right] \\
&= \mathcal{O}\Big(\mathbf{E}\left[\mathbb{Q}_5(|S||A|\log^2 T) + \mathbb{Q}_3\left(\ln\iota\right)\right] + L^2|S|^3|A|^2\ln^2\iota\Big).
\end{aligned}$$

Thus, we obtain the final bound on $\mathbf{E}\left[\textsc{EstReg}\right]$:

$$\mathbf{E}\left[\textsc{EstReg}\right] = \mathcal{O}\Big(\mathbf{E}\left[\mathbb{Q}_4\left(L|S|^2|A|\log^2 T\right) + \mathbb{Q}_5\left(|S||A|\log^2 T\right) + \mathbb{Q}_3\left(\ln \iota\right)\right] + L^2|S|^3|A|^2\ln^2\iota\Big)$$

Recall that $\delta = 1/T^3$ and $\iota = \frac{|S||A|T}{\delta}$. Finally, by combining the bounds of each term, we finally have

$$\begin{aligned}
\mathrm{Reg}_T(\pi^\star) \le \mathcal{O}\Big(&\mathbf{E}\left[\mathbb{Q}_1\left(L^2|S|\ln\iota\right) + \mathbb{Q}_3\left(\ln\iota\right)\right] && \text{(from ErrSub)}\\
&+ \mathbf{E}\left[\mathbb{Q}_2\left(L^2|S|\ln\iota\right) + \mathbb{Q}_6\left(\ln\iota\right)\right] && \text{(from ErrOpt)}\\
&+ \mathbf{E}\left[\mathbb{Q}_3\left(L^2\ln\iota\right)\right] && \text{(from OccDiff)}\\
&+ \mathbf{E}\left[\mathbb{Q}_4\left(L|S|^2|A|\ln^2\iota\right) + \mathbb{Q}_5\left(|S||A|\ln^2\iota\right) + \mathbb{Q}_3\left(\ln\iota\right)\right] && \text{(from EstReg)}\\
&+ L^2|S|^3|A|^2\ln^2\iota\Big).
\end{aligned}$$

Using the self-bounding lemmas (27-32) and the exact same analysis as in Appendix C.3 of Jin et al. [2021], we get $\mathrm{Reg}_T(\pi^\star)$ is bounded by $\mathcal{O}\left(U + \sqrt{UC} + V\right)$ where $V = L^2|S|^3|A|^2\ln^2\iota$ and U is defined as

$$U = \sum_{s\neq s_L}\sum_{a\neq\pi^\star(s)}\left[\frac{L^4|S|\ln\iota + |S||A|\ln^2\iota}{\Delta(s,a)}\right] + \left[\frac{(L^4|S|^2 + L^3|S|^2|A|)\ln\iota + L|S|^2|A|\ln^2\iota}{\Delta_{\textsc{min}}}\right].$$

This completes the entire proof.