Best of Both Worlds Policy Optimization

Christoph Dann [1] **Chen-Yu Wei** [2] **Julian Zimmert** [1]

Abstract

Policy optimization methods are popular reinforcement learning algorithms in practice. Recent works have built theoretical foundation for them by proving $\sqrt{T}$ regret bounds even when the losses are adversarial. Such bounds are tight in the worst case but often overly pessimistic. In this work, we show that in tabular Markov decision processes (MDPs), by properly designing the regularizer, the exploration bonus and the learning rates, one can achieve a more favorable $\text{polylog}(T)$ regret when the losses are stochastic, without sacrificing the worst-case guarantee in the adversarial regime. To our knowledge, this is also the first time a gap-dependent $\text{polylog}(T)$ regret bound is shown for policy optimization. Specifically, we achieve this by leveraging a Tsallis entropy or a Shannon entropy regularizer in the policy update. Then we show that under known transitions, we can further obtain a first-order regret bound in the adversarial regime by leveraging the log barrier regularizer.

1. Introduction

Policy optimization methods have seen great empirical success in various domains (Schulman et al., 2017; Levine & Koltun, 2013). An appealing property of policy optimization methods is the local-search nature, which lends itself to an efficient implementation as a search over the whole MDP is avoided. However, this property also makes it difficult to obtain global optimality guarantees for these algorithms and a large portion of the literature postulates strong and often unrealistic assumptions to ensure global exploration (see e.g., Abbasi-Yadkori et al., 2019; Agarwal et al., 2020b; Neu & Olkhovskaya, 2021; Wei et al., 2021). Recently, the need for extra assumptions has been overcome by adding exploration bonuses to the update (Cai et al., 2020; Shani et al., 2020; Agarwal et al., 2020a; Zanette et al., 2020; Luo et al., 2021). These works demonstrate an additional robustness property of policy optimization, which is able to handle adversarial losses or some level of corruption. Luo et al. (2021) and Chen et al. (2022) even managed to obtain the optimal $\sqrt{T}$ rate.

However, when the losses are in fact stochastic, the $\sqrt{T}$ minimax regret is often overly pessimistic and $\log(T)$ with problem-dependent factors is the optimal rate (Lai et al., 1985). Recently, Jin et al. (2021) obtained a best-of-both-worlds algorithm that automatically adapts to the nature of the environment, a method which relies on FTRL with a global regularizer over the occupancy measure.

In this work, we show that by properly assigning the bonus and tuning the learning rates, policy optimization can also achieve the best of both worlds, which gives a more computationally favorable solution than Jin et al. (2021) for the same setting. Specifically, we show that policy optimization with Tsallis entropy or Shannon entropy regularizer achieves $\sqrt{T}$ regret in the adversarial regime and $\text{polylog}(T)$ regret in the stochastic regime. The $\sqrt{T}$ can further be improved to $\sqrt{L}$ if the transition is known and if a log-barrier regularizer is used, where L is the cumulative loss of the best policy. Though corresponding results in multi-armed bandits have been well-studied, new challenges arise in the MDP setting which require non-trivial design for the exploration bonus and the learning rate scheduling. The techniques we develop to address these issues constitute the main contribution of this work.

2. Related Work

For multi-armed bandits, the question whether there is a single algorithm achieving near-optimal regret bounds in both the adversarial and the stochastic regimes was first asked by Bubeck & Slivkins (2012). A series of followup works refined the bounds through different techniques (Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019; Ito, 2021). One of the most successful approaches is developed by Wei & Luo (2018); Zimmert & Seldin (2019); Ito (2021), who demonstrated that a simple Online Mirror Descent (OMD) or Follow the Regularized Leader (FTRL) algorithm, which was originally designed only for the adversarial case, is able

[1]Google Research [2]MIT Institute for Data, Systems, and Society. Correspondence to: Chen-Yu Wei <chenyuw@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

to achieve the best of both worlds. This approach has been adopted to a wide range of problems including semi-bandits (Zimmert et al., 2019), graph bandits (Erez & Koren, 2021; Ito et al., 2022), partial monitoring (Tsuchiya et al., 2022), multi-armed bandits with switching costs (Rouyer et al., 2021; Amir et al., 2022), tabular MDPs (Jin & Luo, 2020; Jin et al., 2021), and others. Though under a similar framework, each of them addresses new challenges that arises in their specific setting.

Previous works that achieve the best of both worlds in tabular MDPs (Jin & Luo, 2020; Jin et al., 2021) are based on FTRL over the *occupancy measure* space. This approach has several shortcomings, making it less favorable in practice. First, the feasible set of occupancy measure depends on the transition kernel, so the extension to a model-free version is difficult. Second, since the occupancy measure space is a general convex set that may change over time as the learner gains more knowledge about transitions, it requires solving a different convex programming in each round. In contrast, policy optimization is easier to extend to settings where transitions are hard to learn, and it is computationally simple — in tabular MDPs, it is equivalent to running an individual multi-armed bandit algorithm on each state.

Due to its local search nature, exploration under policy optimization is non-trivial, especially when coupled with bandit feedback and adversarial losses. In a simpler setting where the loss feedback has full information, He et al. (2022); Cai et al. (2020) showed $\sqrt{T}$ regret for linear mixture MDPs using policy optimization. In another simpler setting where the loss is stochastic, Agarwal et al. (2020a); Zanette et al. (2021) showed poly$(1/\epsilon)$ sample complexity for linear MDPs. The work by Shani et al. (2020) first studied policy optimization with bandit feedback and adversarial losses, and obtained a $T^{2/3}$ regret for tabular MDPs. Luo et al. (2021) improved it to the optimal $\sqrt{T}$, and provided extensions to linear-Q and linear MDPs. In this work, we demonstrate another power of policy optimization by showing a best-of-both-world regret bound in tabular MDPs. To our knowledge, this is also the first time a gap-dependent polylog(T) regret bound is shown for policy optimization.

We also note that a first-order bound has been shown for adversarial MDPs by Lee et al. (2020). Their algorithm is based on regularization on the occupancy measure, and does not rely on knowledge of the transition kernel. On the other hand, our first-order bound currently relies on the learner knowing the transitions. Whether it can be achieved under unknown transitions is an open question.

3. Notation and Setting

Notation For $f \in \mathbb{R}$ and $g \in \mathbb{R}_+$, we use $f \lesssim g$ or $f \leq O(g)$ to mean that $f \leq c \cdot g$ for some absolute constant $c > 0$. $[x]_+ \triangleq \max\{x, 0\}$. $\triangle(\mathcal{X})$ denotes the probability simplex over the set $\mathcal{X}$.

3.1. MDP setting

We consider episodic fixed-horizon MDPs. Let T be the total number of episodes. The MDP is described by a tuple $(\mathcal{S}, \mathcal{A}, H, P, \{\ell_t\}_{t=1}^T)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, H is the horizon length, $P : \mathcal{S} \times \mathcal{A} \to \triangle(\mathcal{S})$ is the transition kernel so that $P(s'|s,a)$ is the probability of moving to state s' after taking action a on state s, and $\ell_t : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the loss function in episode t. We define $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, which are both assumed to be finite. Without loss of generality, we assume $A \leq T$. A policy $\pi : \mathcal{S} \to \triangle(\mathcal{A})$ describes how the player interacts with the MDP, with $\pi(\cdot|s) \in \triangle(\mathcal{A})$ being the action distribution the player uses to select actions in state s. If for all s, $\pi(\cdot|s)$ is only supported on one action, we call π a deterministic policy, and we abuse the notation $\pi(s) \in \mathcal{A}$ to denote the action π chooses on state s.

Without loss of generality, we assume that the state space can be partitioned into $H+1$ disjoint layers $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$, and the transition is only possible from one layer to the next (i.e., $P(\cdot|s,a)$ is only supported on $\mathcal{S}_{h+1}$ if $s \in \mathcal{S}_h$)[1]. Without loss of generality, we assume that $\mathcal{S}_0 = \{s_0\}$ (initial state) and $\mathcal{S}_H = \{s_H\}$ (terminal state). Also, since there is at least one state on each layer, it holds that $H \leq S$. Let $h(s)$ denotes the layer where state s lies.

The environment decides P and $\{\ell_t\}_{t=1}^T$ ahead of time. In episode t, the learner decides on a policy π_t. Starting from the initial state $s_{t,0} = s_0$, the learner repeatedly draws action $a_{t,h}$ from $\pi_t(\cdot|s_{t,h})$ and transitions to the next state $s_{t,h+1} \in \mathcal{S}_{h+1}$ following $s_{t,h+1} \sim P(\cdot|s_{t,h}, a_{t,h})$, until it reaches the terminal state $s_{t,H} = s_H$. The learner receives $\{\ell_t(s_{t,h}, a_{t,h})\}_{h=0}^{H-1}$ at the end of episode t.

For a policy π and a loss function ℓ, we define $V^\pi(s_H; \ell) = 0$ and recursively define

$$Q^\pi(s,a;\ell) = \ell(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V^\pi(s';\ell)\right],$$
$$V^\pi(s;\ell) = \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s,a;\ell), \tag{1}$$

which are the standard *state-action value function* and *state value function* under policy π and loss function ℓ.

The learner's *regret* with respect to a policy π is defined as

$$\mathrm{Reg}(\pi) = \mathbb{E}\left[\sum_{t=1}^T (V^{\pi_t}(s_0;\ell_t) - V^\pi(s_0;\ell_t))\right].$$

[1] This setting follows previous work on adversarial MDPs (Jin et al., 2020; Luo et al., 2021).

3.2. Known and unknown transition

Following Jin & Luo (2020); Jin et al. (2021), we consider both scenarios where the learner knows the transition kernel P and where he does not know it.

The empirical transition is defined by the following:

$$\widehat{P}_t(s'|s,a) = \frac{n_t(s,a,s')}{n_t(s,a)}$$

where $n_t(s,a)$ is the number of visits to (s,a) prior to episode t, and $n_t(s,a,s')$ is the number of visits to s' after visiting (s,a), prior to episode t. If $n_t(s,a)=0$, we define $\widehat{P}_t(\cdot|s,a)$ to be uniform over the states on layer $h(s)+1$.

In the unknown transition case, we define the confidence set of the transition:

$$\mathcal{P}_t = \Bigg\{ \widetilde{P} : \forall h, \forall (s,a)\in\mathcal{S}_h\times\mathcal{A},\ \ \widetilde{P}(\cdot|s,a)\in\triangle(\mathcal{S}_{h+1}),$$
$$\left|\widetilde{P}(s'|s,a)-\widehat{P}_t(s'|s,a)\right| \le 2\sqrt{\frac{\widehat{P}_t(s'|s,a)]\iota}{n_t(s,a)}} + \frac{14\iota}{3n_t(s,a)} \Bigg\} \tag{2}$$

where $\iota = \ln(SAT/\delta)$. As shown in (Jin & Luo, 2020), $P \in \bigcap_{t=1}^{T}\mathcal{P}_t$ with probability at least $1-4\delta$. Through out the paper, we use $\delta = \frac{1}{T^3}$.

For an arbitrary transition kernel $\widetilde{P}$, define

$$\mu^{\widetilde{P},\pi}(s,a) = \sum_{h=0}^{H} \Pr(s_h=s, a_h=a \mid \pi, \widetilde{P}),$$

where $\Pr(\cdot|\pi,\widetilde{P})$ denotes the probability measure induced by policy π and transition kernel $\widetilde{P}$. Furthermore, define $\mu^{\widetilde{P},\pi}(s) = \sum_a \mu^{\widetilde{P},\pi}(s,a)$. We write $\mu^{\pi}(s) = \mu^{P,\pi}(s)$ and $\mu^{\pi}(s,a) = \mu^{P,\pi}(s,a)$ where P is the true transition. Define the upper and lower confidence measure as

$$\overline{\mu}_t^{\pi}(s) = \max_{\widetilde{P}\in\mathcal{P}_t} \mu^{\widetilde{P},\pi}(s), \qquad \underline{\mu}_t^{\pi}(s) = \min_{\widetilde{P}\in\mathcal{P}_t} \mu^{\widetilde{P},\pi}(s).$$

Finally, define $V^{\widetilde{P},\pi}(s;\ell)$ and $Q^{\widetilde{P},\pi}(s,a;\ell)$ to be similar to (1), with the transition kernel replaced by $\widetilde{P}$.

3.3. Adversarial versus stochastic regimes

We analyze our algorithm in two regimes: the *adversarial* regime and the *stochastic* regime. In both regimes, the transition P is fixed throughout all episodes. In the *adversarial* regime, the loss functions $\{\ell_t\}_{t=1}^T$ are determined arbitrarily ahead of time. In the *stochastic* regime, ℓ_t are generated randomly, and there exists a deterministic policy $\pi^\star$, a gap function $\Delta : \mathcal{S}\times\mathcal{A}\to\mathbb{R}_{\ge 0}$, and $\{\lambda_t(\pi)\}_{t,\pi}\subset\mathbb{R}$ such that for any policy π and any t,

$$\mathbb{E}\left[V^{\pi}(s_0;\ell_t) - V^{\pi^\star}(s_0;\ell_t)\right] = \sum_s \sum_{a\neq\pi^\star(s)} \mu^{\pi}(s,a)\Delta(s,a) - \lambda_t(\pi).$$

If $\lambda_t(\pi)\le 0$ for all π, the condition above certifies that $\pi^\star$ is the optimal policy in episode t, and every time π visits state s and chooses an action $a\neq\pi^\star(s)$, the incurred regret against $\pi^\star$ is at least $\Delta(s,a)$. The amount $[\lambda_t(\pi)]_+$ thus quantifies how much the condition above is violated. The stochastic regime captures the standard RL setting (i.e., $\{\ell_t\}$ are i.i.d.) with $\lambda_t(\pi)=0$ and $\Delta(s,a) = \mathbb{E}\left[Q^{\pi^\star}(s,a;\ell_t) - V^{\pi^\star}(s;\ell_t)\right]$. Define $\Delta_{\min} = \min_s \min_{a\neq\pi^\star(s)}\Delta(s,a)$. Also, define $\mathcal{C} = \left(\mathbb{E}\left[\sum_{t=1}^T \lambda_t(\pi_t)\right]\right)_+$ and $\mathcal{C}(\pi) = \left(\sum_{t=1}^T \lambda_t(\pi)\right)_+$.

4. Main Results and Techniques Overview

Our main results with Tsallis entropy and log barrier regularizers are the following (see Section 6 and Appendix H for results with Shannon entropy):

Theorem 4.1. *Under known transitions, Algorithm 1 with Tsallis entropy regularizer ensures for any π*

$$\mathrm{Reg}(\pi) \lesssim \sqrt{H^3SAT\ln(T)} + \mathit{poly}(H,S,A)\ln(T)$$

in the adversarial regime, and

$$\mathrm{Reg}(\pi) \lesssim U + \sqrt{U\mathcal{C}} + \mathit{poly}(H,S,A)\ln(T) \tag{3}$$

in the stochastic regime, where $U = \sum_s\sum_{a\neq\pi^\star(s)} \frac{H^2\ln(T)}{\Delta(s,a)}$.[2]

Our bounds in both regimes are similar to those of Jin et al. (2021) up to the definition of U under their parameter $\gamma = \frac{1}{H}$ (tuning γ trades their bounds between the two regimes; see their Appendix A.3). Compared with our definition of U, theirs involves an additional additive term $\frac{\mathrm{poly}(H)S\ln(T)}{\Delta_{\min}}$ even under the assumption that the optimal action is unique on all states.

Theorem 4.2. *Under unknown transitions, Algorithm 1 with Tsallis entropy regularizer ensures for any π*

$$\mathrm{Reg}(\pi) \lesssim \sqrt{H^4S^2AT\ln(T)\iota} + \mathit{poly}(H,S,A)\ln(T)\iota$$

in the adversarial regime, and

$$\mathrm{Reg}(\pi) \lesssim U + \sqrt{U(\mathcal{C}+\mathcal{C}(\pi))} + \mathit{poly}(H,S,A)\ln(T)\iota \tag{4}$$

[2] A lower bound in (Xu et al., 2021) shows that an $S/\Delta_{\min}$ dependence is inevitable even when the transition is known. However, this lower bound only holds when there exist multiple optimal actions on $\Omega(S)$ of the states, while our gap bound is finite only when the optimal action is unique on all states. Therefore, our upper bound does not violate their lower bound.

in the stochastic regime, where $U = \frac{H^4 S^2 A \ln(T)\iota}{\Delta_{\min}}$ *and* $\iota = \ln(SAT)$.

In Jin et al. (2021), for the stochastic case under unknown transition, a similar guarantee as (4) is proven only for $\pi = \pi^\star$, with the case for general π left open. We generalize their result by resolving some technical difficulties in their analysis. Overall, our bound in the stochastic regime improves that of Jin et al. (2021), and the bound in the adversarial regime matches that of Luo et al. (2021). Notice that comparing (4) with (3), the bound under unknown transition involves an additional term $\mathcal{C}(\pi)$. It remains open whether it can be removed.

Finally, we provide a first-order best-of-both-world result under known transition.

Theorem 4.3. *Under known transitions, Algorithm 1 with log barrier regularizer ensures for any* π,

$$\mathrm{Reg}(\pi) \lesssim \sqrt{H^2 SA \sum_{t=1}^{T} V^{\pi}(s_0;\ell_t)\ln^2(T)} + \mathit{poly}(H,S,A)\ln^2(T)$$

in the adversarial regime, and

$$\mathrm{Reg}(\pi) \lesssim U + \sqrt{U\mathcal{C}} + \mathit{poly}(H,S,A)\ln^2(T)$$

in the stochastic regime, where $U = \sum_s \sum_{a\neq\pi^\star(s)} \frac{H^2\ln^2(T)}{\Delta(s,a)}$.

In the next two subsections, we overview the techniques we used and challenges we faced in obtaining our results.

4.1. Exploration bonus for policy optimization

In the tabular case, a policy optimization algorithm can be viewed as running an individual bandit algorithm on each state. Our algorithm is built upon the policy optimization framework developed by Luo et al. (2021), who achieve near-optimal worst-case regret in adversarial MDPs. Their key idea is summarized in the next lemma.

Lemma 4.4 (Lemma B.1 of Luo et al. (2021))**.** *Suppose that for some* $\{b_t\}_{t=1}^T$ *and* $\{\mathcal{P}_t\}_{t=1}^T$, *where each* $b_t : \mathcal{S} \to \mathbb{R}_{\geq 0}$ *is a non-negative bonus function and each* $\mathcal{P}_t$ *is a set of transitions, it holds that*

$$B_t(s,a) = b_t(s) + \left(1+\frac{1}{H}\right)\max_{\widetilde{P}\in\mathcal{P}_t} \mathbb{E}_{s'\sim\widetilde{P}(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[B_t(s',a')\right]. \quad (5)$$

Also, suppose that the following holds for a policy π *and a function* $X^\pi : \mathcal{S}\to\mathbb{R}$:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_a \left(\pi_t(a|s)-\pi(a|s)\right)\left(Q^{\pi_t}(s,a;\ell_t)-B_t(s,a)\right)\right] \leq X^\pi(s) + \mathbb{E}\left[\sum_{t=1}^{T} b_t(s) + \frac{1}{H}\sum_{t=1}^{T}\sum_a \pi_t(a|s)B_t(s,a)\right] \quad (6)$$

Then $\mathrm{Reg}(\pi)$ *is upper bounded by*

$$\sum_s \mu^\pi(s)X^\pi(s) + 3\mathbb{E}\left[\sum_{t=1}^{T} V^{\widetilde{P}_t,\pi_t}(s_0;b_t)\right] + \mathbb{E}[\mathcal{F}] \quad (7)$$

where $\widetilde{P}_t$ *is the* $\widetilde{P}$ *that attains the maximum in* (5)*, and* $\mathcal{F} = HT\mathbb{I}\{\exists t\in[T], P\notin\mathcal{P}_t\}$.

The intuition about Lemma 4.4 is explained below (the reader may also refer to Section 3 of Luo et al. (2021) for a more complete explanation.

We start by analyzing a vanilla policy optimization algorithm without adding bonus (i.e., feeding $\widehat{Q}_t(s,a)$, an estimator of $Q^{\pi_t}(s,a;\ell_t)$, to the bandit algorithm on state s). By the value difference lemma (Kakade & Langford, 2002) and standard analysis for the mirror descent algorithm, we run into the following form of regret:

$$\begin{aligned}\mathrm{Reg}(\pi) &= \mathbb{E}\left[\sum_t \left(V^{\pi_t}(s_0;\ell_t) - V^\pi(s_0;\ell_t)\right)\right] \\ &= \mathbb{E}\left[\sum_s \mu^\pi(s)\sum_{t,a}\left(\pi_t(a|s)-\pi(a|s)\right)Q^{\pi_t}(s,a;\ell_t)\right] \\ &\leq \mathbb{E}\left[\sum_s \mu^\pi(s)\left(X^\pi(s)+\sum_t b_t(s)\right)\right] \quad (8)\\ &= \mathbb{E}\left[\sum_s \mu^\pi(s)X^\pi(s) + \sum_t V^\pi(s_0;b_t)\right]\end{aligned}$$

where $X^\pi(s)+\sum_t b_t(s)$ is the regret bound of the bandit algorithm on state s, with $X^\pi(s)$ related to the regularization penalty, and $b_t(s)$ related to the stability of the algorithm. Specifically, in the known transition case, the standard choice is to use $\widehat{Q}_t(s,a) = \frac{L_t(s,a)\mathbb{I}\{(s,a) \text{ is visited in episode } t\}}{\mu^{\pi_t}(s,a)}$ as an unbiased loss estimator for $Q^{\pi_t}(s,a;\ell_t)$, where $L_t(s,a)$ is the cumulative loss starting from state-action (s,a) in episode t. Using exponential weights with learning rate η on every state, one can derive a regret bound of (8) with $X^\pi(s) = \frac{\ln A}{\eta}$ and $b_t(s) = O\left(\frac{\eta H^2}{\mu^{\pi_t}(s)}\right)$. This makes the quantity $V^\pi(s_0;b_t)$ involve the *distribution mismatch coefficient* $\sum_s \frac{\mu^\pi(s)}{\mu^{\pi_t}(s)}$ (Agarwal et al., 2020b; Wei et al., 2020) that can be prohibitively large.

On the other hand, an observation is that the problematic quantity $V^{\pi}(s_0; b_t)$ is nicely bounded if π is π_t. This motivates (Luo et al., 2021) to use $\ell_t(s,a) - b_t(s)$ as the loss, where $b_t(s)$ can be viewed as a *bonus* term that encourages the learner to visit states that have been seldom visited before. To see why this works, assume for a moment that the regret bound still roughly holds when we replace the loss ℓ_t by $\ell_t - b_t$. Then similar to (8), we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(V^{\pi_t}(s_0; \ell_t - b_t) - V^{\pi}(s_0; \ell_t - b_t)\right)\right] \lesssim \mathbb{E}\left[\sum_s \mu^{\pi}(s) X^{\pi}(s) + \sum_{t=1}^{T} V^{\pi}(s_0; b_t)\right] \tag{9}$$

which implies

$$\mathrm{Reg}(\pi) = \mathbb{E}\left[\sum_{t=1}^{T}\left(V^{\pi_t}(s_0; \ell_t) - V^{\pi}(s_0; \ell_t)\right)\right] \lesssim \mathbb{E}\left[\sum_{s,a} \mu^{\pi}(s) X^{\pi}(s) + \sum_{t=1}^{T} V^{\pi_t}(s_0; b_t)\right] \tag{10}$$

by rearranging and the linearity of the value function $V^{\pi}(s_0; \ell_t - b_t) = V^{\pi}(s_0; \ell_t) - V^{\pi}(s_0; b_t)$ for any π. Now since the regret only involves $V^{\pi_t}(s_0; b_t)$, there will be no distribution mismatch coefficient in the regret bound.

The caveat of the discussion above is the assumption of (9). After adding the bonus b_t, the original regret bound can be affected, that is, the b_t on the right-hand side of (9) can be something larger, breaking the desired cancellation effect to achieve (10). To resolve this issue, Luo et al. (2021) proposed to use the *dilated bonus* defined in (5) in the policy optimization update. In (5), the bonus-to-go function is not constructed through a standard Bellman equation, but through a dilated version that includes an additional $1 + \frac{1}{H}$ factor for future steps. The additional amount of bonus-to-go can be used to cancel the additional regret due to the inclusion of b_t.

Lemma 4.4 gives a general recipe to design the exploration bonus for policy optimization algorithms. Roughly speaking, the bonus function $b_t(s)$ is chosen to be the instantaneous regret of the bandit algorithm on state s, which scales inversely with the probability of visiting state s (i.e., $\frac{1}{\mu^{\pi_t}(s)}$). Lemma 4.4 suggests that the bandit algorithm on state s should update itself using $Q^{\pi_t}(s,a;\ell_t) - B_t(s,a)$ as the loss, where $B_t(s,a)$ is the dilated bonus-to-go.

The bonus function $b_t(s)$ we use is slightly different from that in Luo et al. (2021) though. We notice that the $b_t(s)$ defined in Luo et al. (2021) has two parts: the first part is *FTRL regret overhead*, which comes from the regret bound of the FTRL algorithm under the given loss estimator, and the second part comes from the *estimation error* in estimating the transition kernel. In order to apply the self-bounding technique to obtain the best-of-both-worlds result, the second term in (7) can only involve the the first part (FTRL regret overhead) but not the second part (estimation error). Therefore, we split their bonus into two: our $b_t(s)$ only includes the first part, and $c_t(s)$ only includes the second part. This allows us to use self-bounding on the second term in (7). Our $c_t(s)$ goes to the first term in (7) instead and is handled differently from Luo et al. (2021). More details are given in Section 5 and Section 6.

4.2. Adaptive learning rate tuning and bonus design

Our algorithm heavily relies on carefully tuning the learning rates and assigning a proper amount of bonus. These two tasks are intertwined with each other and introduce new challenges that are not seen in the global regularization approach (Jin et al., 2021) or policy optimization approach that only aims at a worst-case bound (Luo et al., 2021). Below we give a high-level overview for the challenges.

In the FTRL analysis, a major challenge is to handle losses that are overly negative[3]. Typically, if the learning rate is η and the negative loss of action a has a magnitude of R, we need $\eta p(a)^{\beta} R \le 1$ in order to keep the algorithm stable, where $p(a)$ is the probability of choosing action a, and $\beta \in [0,1]$ is a parameter related to the choice of the regularizer ($\frac{1}{2}$ for Tsallis entropy, 0 for Shannon entropy, and 1 for log barrier). In our case, there are two places we potentially encounter overly negative losses. One is when applying the standard *loss-shifting* technique for best-of-both-world bounds (see Jin et al. (2021)). The loss-shifting effectively creates a negative loss in the analysis. The other overly negative loss is the bonus we use to obtain the first-order bound.

For the first case, we develop a simple trick that only performs loss-shifting when the introduced negative loss is not too large, and further show that the extra penalty due to "not performing loss-shifting" is well-controlled. This is explained in Section 5.1. For the second case, we develop an even more general technique (which can also cover the first case). This technique can be succinctly described as "inserting virtual episodes" when $\eta p(a)^{\beta} R$ is potentially too large. In virtual episodes, the losses are assumed to be all-zero (because the learner actually does not interact with the environment in these episodes) and the algorithm only updates over some bonus term. The goal of the virtual episodes is solely to tune down the learning rate η and prevent $\eta p(a)^{\beta} R$ from being to large in real episodes. Similarly, we are able to show that the extra penalty due to virtual episodes is well-controlled. This is explained in Section 5.3.

[3] Losses here refer not only to the loss from the environment, but also loss estimators or bonuses constructed by the algorithm.

Algorithm 1 Policy Optimization

Define: $\psi_t(\pi; s)$, $b_t(s)$ are defined according to Figure 1, $\gamma_t \triangleq \min\left\{\frac{10^6 H^4 A^2}{t}, 1\right\}$.

for $t = 1, 2, \ldots$ **do**

$$\pi_t(\cdot|s) = \operatorname*{argmin}_{\pi \in \triangle(\mathcal{A})} \left\{ \sum_{\tau=1}^{t-1} \sum_a \pi(a) \left(\widehat{Q}_\tau(s,a) - B_\tau(s,a) - C_\tau(s,a) \right) + \psi_t(\pi; s) \right\} \tag{11}$$

Add a virtual round if needed (only when aiming to get a first-order bound with log barrier — see Section 5.3 (29)).
Execute π_t in episode t, and receive $\{\ell_t(s_{t,h}, a_{t,h})\}_{h=0}^{H-1}$.
Define $\mathcal{P}_t$: Under known transition, define $\mathcal{P}_t = \{P\}$. Under unknown transition, define $\mathcal{P}_t$ by (2).
Define $\widehat{Q}_t$: For $s \in \mathcal{S}_h$, let $\mathbb{I}_t(s,a) = \mathbb{I}\{(s_{t,h}, a_{t,h}) = (s,a)\}$, $L_{t,h} = \sum_{h'=h}^{H-1} \ell_t(s_{t,h'}, a_{t,h'})$, and

$$\widehat{Q}_t(s,a) = \frac{\mathbb{I}_t(s,a) L_{t,h}}{\mu_t(s)\pi_t(a|s)}, \qquad \text{where } \mu_t(s) = \overline{\mu}_t^{\pi_t}(s) + \gamma_t. \tag{12}$$

Define C_t: Let $c_t(s) = \frac{\mu_t(s) - \underline{\mu}_t^{\pi_t}(s)}{\mu_t(s)} H$, and compute $C_t(s,a)$ by

$$C_t(s,a) = \max_{\widetilde{P} \in \mathcal{P}_t} \mathbb{E}_{s' \sim \widetilde{P}(\cdot|s,a), a' \sim \pi_t(\cdot|s')} \left[c_t(s') + C_t(s', a') \right], \tag{13}$$

Define B_t: Compute $B_t(s,a)$ by (5) using the $b_t(s)$ defined in Figure 1.

5. Algorithm

The template of our algorithm is Algorithm 1, in which we can plug different regularizers. The template applies to both known transition and unknown transition cases — the only difference is in the definition of the confidence set $\mathcal{P}_t$.

The policy update (11) is equivalent to running individual FTRL on each state with an adaptive learning rate. The loss estimator $\widehat{Q}_t(s,a)$ defined in (12) is similar to that in Luo et al. (2021): if (s,a) is visited, it is the cumulative loss starting from (s,a) divided by the *upper occupancy measure* (Jin et al., 2020) of (s,a); otherwise it is zero. One difference is that the "implicit exploration" factor γ_t added to the denominator is of order $\frac{1}{t}$ in our case, while it is of order $\frac{1}{\sqrt{t}}$ in Luo et al. (2021). This smaller γ_t allows us to achieve logarithmic regret in the stochastic regime.

There are two bonus functions $c_t(s)$ and $b_t(s)$ defined in (13) and Figure 1, respectively. As discussed in Section 4.1, the bonus functions are defined to be the instantaneous regret of the bandit algorithm on state s. The first bonus function $c_t(s)$ comes from the bias of the loss estimator. Our choice of $c_t(s)$ is such that $\forall a, Q^{\pi_t}(s,a;\ell_t) - \mathbb{E}[\widehat{Q}_t(s,a)] \le c_t(s)$. The second bonus function $b_t(s)$ is related to the regret of the FTRL algorithm under the given loss estimator, which is regularizer dependent. We will elaborate how to choose $b_t(s)$ for different regularizers later in this section.

Finally, dynamic programming are used to obtain $C_t(s,a)$ and $B_t(s,a)$, which are trajectory sums of $c_t(s)$ and $b_t(s)$, with an $(1 + \frac{1}{H})$ dilation on $B_t(s,a)$. They are then used in the policy update (11). In the following subsections, we discuss how we choose $b_t(s)$ and tune the learning rate for each regularizer.

5.1. Tsallis entropy

$b_t(s)$ corresponds to the instantaneous regret of the bandit algorithm on state s under the given loss estimator. To obtain its form, we first analyze the regret assuming $B_t(s,a)$ is not included, i.e., only update on $\widehat{Q}_t(s,a) - C_t(s,a)$ ($B_t(s,a)$ will be added back for analysis after the form of $b_t(s)$ is decided). Inspired by Zimmert & Seldin (2019) for multi-armed bandits, our target is to show that the instantaneous regret (see Appendix D for details) on state s is upper bounded by

$$\underbrace{\left(\frac{1}{\eta_t(s)} - \frac{1}{\eta_{t-1}(s)}\right)\xi_t(s)}_{\text{penalty term}} + \underbrace{\frac{H^2 \eta_t(s) \xi_t(s)}{\mu_t(s)}}_{\text{stability term}} + \nu_t(s) \tag{26}$$

where $\xi_t(s) = \sum_a \sqrt{\pi_t(a|s)}(1 - \pi_t(a|s)) \le \sqrt{A}$, and $\nu_t(s)$ is some overhead due to the inclusion of $-C_t(s,a)$. The factor $\xi_t(s)$ allows us to use the self-bounding technique that leads to best-of-both-worlds bounds, which cannot be relaxed to $\sqrt{A}$ in general. Compared to the bound for multi-armed bandits in (Zimmert & Seldin, 2019), the extra $\frac{1}{\mu_t(s)}$ scaling in the stability term comes from importance weighting because state s is visited with probability roughly $\mu_t(s)$.

Figure 1. Definitions of $\psi_t(\pi; s)$ and $b_t(s)$ for different regularizers (to be used in Algorithm 1).

Tsallis entropy:

$$\psi_t(\pi; s) = -\frac{2}{\eta_t(s)} \sum_a \sqrt{\pi(a)}, \tag{14}$$

$$b_t(s) = 4\left(\frac{1}{\eta_t(s)} - \frac{1}{\eta_{t-1}(s)}\right)\left(\xi_t(s) + \sqrt{A} \cdot \mathbb{I}\left[\frac{\eta_t(s)}{\mu_t(s)} > \frac{1}{8H}\right]\right) + \nu_t(s), \tag{15}$$

where

$$\eta_t(s) = \frac{1}{1600H^4\sqrt{A} + 4H\sqrt{\sum_{\tau=1}^{t} \frac{1}{\mu_\tau(s)}}}, \quad \xi_t(s) = \sum_a \sqrt{\pi_t(a|s)}(1 - \pi_t(a|s)), \quad \nu_t(s) = 8\eta_t(s) \sum_a \pi_t(a|s) C_t(s,a)^2. \tag{16}$$

Shannon entropy:

$$\psi_t(\pi; s) = \sum_a \frac{1}{\eta_t(s,a)} \pi(a) \ln \pi(a), \tag{17}$$

$$b_t(s) = 8 \sum_a \left(\frac{1}{\eta_t(s,a)} - \frac{1}{\eta_{t-1}(s,a)}\right)\left(\xi_t(s,a) + 1 - \frac{\min_{\tau\in[t]} \mu_\tau(s)}{\min_{\tau\in[t-1]} \mu_\tau(s)}\right) + \nu_t(s), \tag{18}$$

where

$$\frac{1}{\eta_t(s,a)} = \frac{1}{\eta_{t-1}(s,a)} + 4\left(\frac{H}{\mu_t(s)\sqrt{\sum_{\tau=1}^{t-1} \frac{\xi_\tau(s,a)}{\mu_\tau(s)} + \frac{1}{\mu_t(s)}}} + \frac{H}{\sqrt{t}}\right)\sqrt{\ln T}, \quad \text{with } \frac{1}{\eta_0(s,a)} = 1600H^4 A\sqrt{\ln T}, \tag{19}$$

$$\xi_t(s,a) = \min\{\pi_t(a|s)\ln(T), 1 - \pi_t(a|s)\}, \qquad \nu_t(s) = 8\sum_a \eta_t(s,a)\pi_t(a|s)C_t(s,a)^2. \tag{20}$$

Log barrier (for first-order bound under known transition):

$$\psi_t(\pi; s) = \sum_a \frac{1}{\eta_t(s,a)} \ln \frac{1}{\pi(a)}, \tag{21}$$

$$b_t(s) = 8 \sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) + \nu_t(s), \tag{22}$$

where

$$(s_t^\dagger, a_t^\dagger) = \operatorname*{argmax}_{s,a} \frac{\eta_t(s,a)}{\mu_t(s)} \qquad \text{(break tie arbitrarily)}$$

$$\frac{1}{\eta_{t+1}(s,a)} = \begin{cases} \frac{1}{\eta_t(s,a)} + \frac{4\eta_t(s,a)\zeta_t(s,a)}{\mu_t(s)^2 \log(T)} & \text{if } t \text{ is a real episode} \\ \frac{1}{\eta_t(s,a)}\left(1 + \frac{1}{24H\log T}\right) & \text{if } t \text{ is a virtual episode and } (s_t^\dagger, a_t^\dagger) = (s,a) \\ \frac{1}{\eta_t(s,a)} & \text{if } t \text{ is a virtual episode and } (s_t^\dagger, a_t^\dagger) \neq (s,a) \end{cases} \tag{23}$$

$$\frac{1}{\eta_1(s,a)} = 4H^4, \tag{24}$$

$$\zeta_t(s,a) = \left(\mathbb{I}_t(s,a) - \pi_t(a|s)\mathbb{I}_t(s)\right)^2 L_{t,h}^2 \qquad \text{where } \mathbb{I}_t(s) = \sum_a \mathbb{I}_t(s,a), \qquad \text{(suppose that } s \in \mathcal{S}_h\text{)}$$

$$\nu_t(s) = 8\sum_a \eta_t(s,a)\pi_t(a|s)C_t(s,a)^2. \tag{25}$$

This desired bound suggests a learning rate scheduling of

$$\eta_t(s) \approx \frac{1}{H\sqrt{\sum_{\tau=1}^{t} \frac{1}{\mu_\tau(s)}}} \tag{27}$$

to balance the penalty and the stability terms. This is exactly how we tune $\eta_t(s)$ in (16). However, to obtain the $\xi_t(s)$ factor in the stability term in (26), we need to perform "loss-shifting" in the analysis, which necessitates the condition $\frac{\eta_t(s)H}{\mu_t(s)} \lesssim 1$ as discussed in Section 4.2. From the choice of $\eta_t(s)$ in (27), this condition may not always hold, but every time it is violated, $\eta_t(s)$ is decreased by a relatively large factor in the next episode.

Our strategy is that whenever the condition $\frac{H\eta_t(s)}{\mu_t(s)} \lesssim 1$ is violated, we do not perform loss-shifting. This still allows us to prove a stability term of $\frac{H^2\eta_t(s)\sqrt{A}}{\mu_t(s)}$ for that episode. The key in the analysis is to show that the extra cost due to "not performing loss-shifting" is only logarithmic in T (see the proof of Lemma 6.3). Combining this idea with the instantaneous regret bound in (26) and the choice of $\eta_t(s)$ in (27), we are able to derive the form of $b_t(s)$ in (15). After figuring out the form of $b_t(s)$ assuming $B_t(s,a)$ is not incorporated in the updates, we incorporate it back and re-analyze the stability term. The extra stability term due to $b_t(s)$ leads to a separate quantity $\frac{1}{H}\pi_t(a|s)B_t(s,a)$, which is an overhead allowed by (6).

5.2. Shannon entropy

The design of $b_t(s)$ under Shannon entropy follows similar procedures as in Section 5.1, except that the tuning of the learning rate is inspired by Ito et al. (2022). One improvement over theirs is that we adopt coordinate-dependent learning rates that can give us a refined gap-dependent bound in the stochastic regime (in multi-armed bandits, this improves their $\max_a \frac{A}{\Delta(a)}$ dependence to $\sum_a \frac{1}{\Delta(a)}$). With Shannon entropy, there is less learning rate tuning issue because its optimal learning rate decreases faster than other regularizers, and there is no need to perform loss-shifting (Ito et al., 2022). The regret bound under Shannon entropy is overall worse than that of Tsallis entropy by a $\ln^2(T)$ factor.

5.3. Log barrier

As shown by Wei & Luo (2018); Ito (2021), FTRL with a log barrier regularizer is also able to achieve the best of both worlds, with the additional benefit of having *data-dependent* bounds. In this subsection, we demonstrate the possibility of this by showing that under *known* transition, Algorithm 1 is able to achieve a first-order bound in the adversarial regime, while achieving $\mathrm{polylog}(T)$ regret in the stochastic regime.

To get a first-order best-of-both-world bound with log barrier, inspired by Ito (2021), we need to prove the following instantaneous regret for the bandit algorithm on s:

$$\underbrace{\sum_a \left(\frac{1}{\eta_t(s,a)} - \frac{1}{\eta_{t-1}(s,a)}\right)\ln T}_{\text{penalty term}} + \underbrace{\sum_a \frac{\eta_t(s,a)\zeta_t(s,a)}{\mu_t(s)^2}}_{\text{stability term}} + \nu_t(s) \tag{28}$$

where $\zeta_t(s,a) = \left(\mathbb{I}_t(s,a) - \pi(a|s)\mathbb{I}_t(s)\right)^2 L_{t,h}^2$ for $s \in \mathcal{S}_h$. This suggests a learning rate scheduling of $1/\eta_{t+1}(s,a) = 1/\eta_t(s,a) + \eta_t(s,a)\zeta_t(s,a)/\mu_t(s)^2$. Similar to the Tsallis entropy case, obtaining the desired stability term in (28) requires loss-shifting, so we encounter the same issue as before and can resolve it in the same way. With this choice of $\eta_t(s,a)$, we can derive the desired form of $b_t(s)$ from (28). However, the magnitude of this bonus is larger than in the Tsallis entropy case because of the $\frac{1}{\mu_t(s)^2}$ scaling here. Therefore, an additional problem arises: the $B_t(s,a)$ derived from this $b_t(s)$ can be large that makes $\eta_t(s,a)\pi_t(a|s)B_t(s,a) > \frac{1}{H}$ happen, which violates the condition under which we can bound the extra stability term (due to the inclusion of $b_t(s)$) by $\frac{1}{H}\pi_t(a|s)B_t(s,a)$. Notice that this was not an issue under Tsallis entropy.

To resolve this, we note that $\eta_t(s,a)\pi_t(a|s)B_t(s,a)$ can be as large as $\mathrm{poly}(H,S)\max_{s',a'}\frac{\eta_t(s',a')^2}{\mu_t(s')^2}$ (Lemma E.3), so all we need is to make $\frac{\eta_t(s,a)}{\mu_t(s)} \le \frac{1}{\mathrm{poly}(H,S)}$ for all s,a.

Our solution is to insert *virtual episodes* when $\frac{\eta_t(s,a)}{\mu_t(s)}$ is too large on some (s,a). In virtual episodes, the learner does not actually interact with the environment; instead, the goal is purely to tune down $\eta_t(s,a)$. To decide whether to insert a virtual episode, in episode t, after the learner computes $\pi_t(\cdot|s)$ on all states, he checks if

$$\max_{s,a} \frac{\eta_t(s,a)}{\mu_t(s)} > \frac{1}{60\sqrt{H^3S}}. \tag{29}$$

If so, then episode t is made a virtual episode in which the losses are assumed to be zero everywhere.[4] In a virtual episode, let $(s_t^\dagger, a_t^\dagger) = \mathrm{argmax}_{s,a}\frac{\eta_t(s,a)}{\mu_t(s)}$, and we tune down $\eta_t(s_t^\dagger, a_t^\dagger)$ by a factor of $(1+\frac{1}{24H\log T})$. Also, a bonus $b_t(s)$ is assigned to $s_t^\dagger$ to reflect the increased penalty term on state $s_t^\dagger$ due to the decrease in learning rate (by (28)). Combinig all the above, we get the bonus and learning rate specified in (22) and (23). Again, since every time a virtual episode happens, there exists some $\eta_t(s,a)$ decreased by a significant factor, it cannot happen too many times.

[4] Inserting a virtual episode shifts the index of future real episodes. Since there are only $O(HSA\log^2 T)$ virtual episodes, we still use T to denote the total number of episodes.

6. Sketch of Regret Analysis

Our goal is to show (6) and bound the right-hand side of (7) (for all regularizers and known/unknown transitions). To show (6), for a fixed π, we do the following decomposition:

$$\sum_{t,a}(\pi_t(a|s)-\pi(a|s))\left(Q^{\pi_t}(s,a;\ell_t)-B_t(s,a)\right) \tag{30}$$
$$=\underbrace{\sum_{t,a}(\pi_t(a|s)-\pi(a|s))\left(\widehat{Q}_t(s,a)-B_t(s,a)-C_t(s,a)\right)}_{\textbf{ftrl-reg}^{\pi}(s)}$$
$$+\underbrace{\sum_{t,a}(\pi_t(a|s)-\pi(a|s))\left(Q^{\pi_t}(s,a;\ell_t)-\widehat{Q}_t(s,a)+C_t(s,a)\right)}_{\textbf{bias}^{\pi}(s)}.$$

The next lemma bounds the expectation of $\textbf{ftrl-reg}^{\pi}(s)$.

Lemma 6.1. *$\mathbb{E}\left[\textbf{ftrl-reg}^{\pi}(s)\right]$ is upper bounded by*

$$O(H^4SA\ln(T))+\mathbb{E}\left[\sum_{t=1}^{T}b_t(s)+\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a|s)B_t(s,a)\right].$$

The proof of Lemma 6.1 is in Appendix E. Notice that depending on the regularizers and whether the transition is known/unknown, the definitions of $b_t(s)$ are different, so we prove it individually for each case.

Combining (30) with Lemma 6.1, we see that the condition in Lemma 4.4 is satisfied with $X^{\pi}(s)=O(H^4SA\ln(T))+\mathbb{E}[\textbf{bias}^{\pi}(s)]$. By Lemma 4.4, we can upper bound $\mathbb{E}[\mathrm{Reg}(\pi)]$ by the order of

$$H^5SA\ln(T)+\mathbb{E}\left[\sum_{s}\mu^{\pi}(s)\textbf{bias}^{\pi}(s)+\sum_{t=1}^{T}V^{\widetilde{P}_t,\pi_t}(s_0;b_t)\right]. \tag{31}$$

The next lemma bounds the bias part in (31). See Appendix F for the proof.

Lemma 6.2. *With known transitions, $\mathbb{E}\left[\sum_s\mu^{\pi}(s)\textbf{bias}^{\pi}(s)\right]\lesssim H^5SA^2\ln(T)$, and with unknown transitions,*

$$\mathbb{E}\left[\sum_{s}\mu^{\pi}(s)\textbf{bias}^{\pi}(s)\right]\lesssim H^2S^4A^2\ln(T)\iota+$$
$$\sqrt{H^3S^2A\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+\right]\ln(T)\iota}.$$

Next, we bound the bonus part in (31) for all regularizers we consider. The proofs are in Appendix G.

Lemma 6.3. *Using Tsallis entropy as the regularizer, with known transitions,*

$$\mathbb{E}\left[\sum_{t=1}^{T}V^{\widetilde{P}_t,\pi_t}(s_0;b_t)\right]\lesssim H^4SA^2\ln(T)+$$
$$H\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\mu_t(s)\pi_t(a|s)(1-\pi_t(a|s))\right]\ln(T)}.$$

With unknown transitions, the right-hand side above further has an additional term $O(HS^4A^2\ln(T)\iota)$.

Lemma 6.4. *Using Shannon entropy as the regularizer, With known transitions,*

$$\mathbb{E}\left[\sum_{t=1}^{T}V^{\widetilde{P}_t,\pi_t}(s_0;b_t)\right]\lesssim H^4SA^2\sqrt{\ln^3(T)}+$$
$$H\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\mu_t(s)\pi_t(a|s)(1-\pi_t(a|s))\right]\ln^3(T)}.$$

With unknown transitions, the right-hand side above further has an additional term $O(HS^4A^2\ln(T)\iota)$.

Lemma 6.5. *Using log barrier as the regularizer, with known transitions,*

$$\mathbb{E}\left[\sum_{t=1}^{T}V^{\pi_t}(s_0;b_t)\right]\lesssim H^3S^2A^2\ln(T)\ln(SAT)+$$
$$\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}(\mathbb{I}_t(s,a)-\pi_t(a|s)\mathbb{I}_t(s))^2L_{t,h(s)}^2\right]\ln^2(T)}.$$

Final regret bounds To obtain the final regret bounds, we combine Lemma 6.2 with each of Lemma 6.3, Lemma 6.4, and Lemma 6.5 based on (31). Then we use the standard self-bounding technique to derive the bounds for each regime. The details are provided in Appendix H.

7. Conclusion

In this work, we develop policy optimization algorithms for tabular MDPs that achieves *the best of both worlds*. Compared to previous solutions with a similar guarantee (Jin & Luo, 2020; Jin et al., 2021), our algorithm is computationally much simpler; compared to most existing RL algorithms, our algorithm is more robust (handling adversarial losses) and more adaptive (achieving fast rate in stochastic environments) simultaneously. Built upon the flexible policy optimization framework, our work paves a way towards developing more robust and adaptive algorithms for more general settings. Future directions include obtaining data-dependent bounds under unknown transitions, and incorporating function approximation.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702. PMLR, 2019.

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020a.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020b.

Amir, I., Azov, G., Koren, T., and Livni, R. Better best of both worlds bounds for bandits with switching costs. *arXiv preprint arXiv:2206.03098*, 2022.

Auer, P. and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 116–120. PMLR, 2016.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chen, L., Luo, H., and Wei, C.-Y. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pp. 1216–1259. PMLR, 2021.

Chen, L., Luo, H., and Rosenberg, A. Policy optimization for stochastic shortest path. *arXiv preprint arXiv:2202.03334*, 2022.

Erez, L. and Koren, T. Best-of-all-worlds bounds for online learning with feedback graphs. *arXiv preprint arXiv:2107.09572*, 2021.

He, J., Zhou, D., and Gu, Q. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 4259–4280. PMLR, 2022.

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

Ito, S., Tsuchiya, T., and Honda, J. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. *arXiv preprint arXiv:2206.00873*, 2022.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33: 16557–16566, 2020.

Jin, T., Huang, L., and Luo, H. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press (preprint), 2018.

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 2020.

Levine, S. and Koltun, V. Guided policy search. In *International conference on machine learning*, pp. 1–9. PMLR, 2013.

Luo, H. Homework 3 solution, introduction to online optimization/learning. http://haipeng-luo.net/courses/CSCI659/2022_fall/homework/HW3_solutions.pdf, November 2022.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612v2*, 2021.

Rouyer, C., Seldin, Y., and Cesa-Bianchi, N. An algorithm for stochastic and adversarial bandits with switching costs. In *International Conference on Machine Learning*, pp. 9127–9135. PMLR, 2021.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 1743–1759. PMLR, 2017.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pp. 1287–1295. PMLR, 2014.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Tsuchiya, T., Ito, S., and Honda, J. Best-of-both-worlds algorithms for partial monitoring. *arXiv preprint arXiv:2207.14550*, 2022.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, 2018.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International Conference on Machine Learning*, pp. 10170–10180. PMLR, 2020.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Xu, H., Ma, T., and Du, S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 467–475. PMLR, 2019.

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

A. Additional Definitions

Define $\mu^{\widetilde{P},\pi}(s'|s,a)$ as the probability of visiting s' conditioned on that (s,a) is already visited, under transition kernel $\widetilde{P}$ and policy π. In other words, $\mu^{\widetilde{P},\pi}(s'|s,a)$ is defined as

$$\begin{cases} 0 & \text{if } h(s') < h(s), \\ 0 & \text{if } h(s) = h(s'),\ s \neq s', \\ 1 & \text{if } s' = s, \\ \Pr\{s_{h(s')} = s' \mid (s_h, a_h) = (s,a)\} & \text{if } h(s') > h(s). \end{cases}$$

Further define $\mu^{\widetilde{P},\pi}(s'|s) = \sum_a \mu^{\widetilde{P},\pi}(s'|s,a)\pi(a|s)$. We write $\mu^{\pi}(s'|s,a) = \mu^{P,\pi}(s'|s,a)$ and $\mu^{\pi}(s'|s) = \mu^{P,\pi}(s'|s)$ where P is the true transition.

B. Concentration Bounds

Lemma B.1. *If $P \in \mathcal{P}_t$, then for all $\widetilde{P} \in \mathcal{P}_t$,*

$$\left|\widetilde{P}(s'|s,a) - P(s'|s,a)\right| \leq \min\left\{4\sqrt{\frac{P(s'|s,a)\iota}{n_t(s,a)}} + \frac{40\iota}{3n_t(s,a)},\ 1\right\}.$$

Lemma B.2 (Lemma D.3.7 of Jin et al. (2021))**.** *With probability at least $1-\delta$, for any h,*

$$\sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} \frac{\mu^{\pi_t}(s,a)}{n_t(s,a)} \lesssim |\mathcal{S}_h| A \ln T + \ln(1/\delta)$$

Definition B.3. Define $\mathcal{E}$ to be the event that $P \in \mathcal{P}_t$ for all t and the bound in Lemma B.2 holds. By (2) and Lemma B.2, $\Pr\{\mathcal{E}\} \geq 1 - 5H\delta$.

C. Difference Lemmas

Lemma C.1 (Performance difference)**.** *For any policies π_1 and π_2, and any loss function $\ell : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$,*

$$V^{\pi_1}(s_0;\ell) - V^{\pi_2}(s_0;\ell) = \sum_s \mu^{\pi_2}(s)(\pi_1(a|s) - \pi_2(a|s))Q^{\pi_1}(s,a;\ell).$$

Lemma C.2. *For any policies π_1 and π_2 and any function $L : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$,*

$$\sum_s \mu^{\pi_2}(s)(\pi_1(a|s) - \pi_2(a|s))L(s,a) = V^{\pi_1}(s_0;\ell) - V^{\pi_2}(s_0;\ell)$$

where

$$\ell(s,a) \triangleq L(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi_1(\cdot|s')}[L(s',a')].$$

Proof. This is simply a different way to write the performance difference lemma (Lemma C.1). One only needs to verify that $Q^{\pi_1}(s,a;\ell) = L(s,a)$. This can be shown straightforwardly by backward induction from $s \in \mathcal{S}_H$ to $s \in \mathcal{S}_0$ and using the definition of $\ell(s,a)$. □

Lemma C.3 (Occupancy measure difference, Lemma D.3.1 of (Jin et al., 2021))**.**

$$\begin{aligned} \mu^{P_1,\pi}(s) - \mu^{P_2,\pi}(s) &= \sum_{(u,v,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \mu^{P_1,\pi}(u,v)\left[P_1(w|u,v) - P_2(w|u,v)\right]\mu^{P_2,\pi}(s|w) \\ &= \sum_{(u,v,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \mu^{P_2,\pi}(u,v)\left[P_1(w|u,v) - P_2(w|u,v)\right]\mu^{P_1,\pi}(s|w) \end{aligned}$$

Lemma C.4 (Generalized version of Lemma 4 in (Jin et al., 2020)). *Suppose the high probability event $\mathcal{E}$ defined in Definition B.3 holds. Let $\widetilde{P}_t^s$ be a transition kernel in $\mathcal{P}_t$ which may depend on s, and let $g_t(s) \in [0, G]$. Then*

$$\sum_{t=1}^{T}\sum_{s}\left|\mu^{\pi_t}(s) - \mu^{\widetilde{P}_t^s,\pi_t}(s)\right| g_t(s) \lesssim \sqrt{HS^2A\ln(T)\iota\sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s)g_t(s)^2} + HS^4AG\ln(T)\iota.$$

Proof. We first show that for any t, s,

$$\left|\mu^{\pi}(s) - \mu^{\widetilde{P}_t^s,\pi}(s)\right| \lesssim \sum_{(u,v,w)\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \mu^{\pi}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi}(s|w) + HS^2\sum_{(u,v)\times\mathcal{S}\times\mathcal{A}}\frac{\mu^{\pi}(u,v)\iota}{n_t(u,v)}. \tag{32}$$

Below, the summation range of (u, w, v) and (x, y, z) are both $\bigcup_{h=0}^{H-1}(\mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1})$ if without specifying.

$$\begin{aligned}
&\left|\mu^{\pi}(s) - \mu^{\widetilde{P}_t^s,\pi}(s)\right| \\
&\le \sum_{u,v,w}\mu^{\pi}(u,v)\left|P(w|u,v) - \widetilde{P}_t^s(w|u,v)\right|\mu^{\widetilde{P}_t^s,\pi}(s|w) && \text{(by Lemma C.3)} \\
&= \sum_{u,v,w}\mu^{\pi}(u,v)\left|P(w|u,v) - \widetilde{P}_t^s(w|u,v)\right|\mu^{\pi}(s|w) \\
&\quad + \sum_{u,v,w}\mu^{\pi}(u,v)\left|P(w|u,v) - \widetilde{P}_t^s(w|u,v)\right|\left(\mu^{\widetilde{P}_t^s,\pi}(s|w) - \mu^{\pi}(s|w)\right) \\
&\le \sum_{u,v,w}\mu^{\pi}(u,v)\left|P(w|u,v) - \widetilde{P}_t^s(w|u,v)\right|\mu^{\pi}(s|w) \\
&\quad + \sum_{u,v,w}\mu^{\pi}(u,v)\left|P(w|u,v) - \widetilde{P}_t^s(w|u,v)\right|\sum_{x,y,z}\mu^{\pi}(x,y|w)\left|\widetilde{P}_t^s(z|x,y) - P(z|x,y)\right|\mu^{\widetilde{P}_t^s,\pi}(s|z) && \text{(by Lemma C.3)} \\
&\lesssim \sum_{u,v,w}\mu^{\pi}(u,v)\left(\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}} + \frac{\iota}{n_t(u,v)}\right)\mu^{\pi}(s|w) \\
&\quad + \sum_{u,v,w}\sum_{x,y,z}\mu^{\pi}(u,v)\left(\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}} + \frac{\iota}{n_t(u,v)}\right)\mu^{\pi}(x,y|w)\min\left\{\sqrt{\frac{P(z|x,y)\iota}{n_t(x,y)}} + \frac{\iota}{n_t(x,y)},\ 1\right\} \\
&\qquad\qquad \text{(by Lemma B.1 and the assumption that } \mathcal{E} \text{ holds)} \\
&\le \sum_{u,v,w}\mu^{\pi}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi}(s|w) \\
&\quad + \sum_{u,v,w}\mu^{\pi}(u,v)\frac{\iota}{n_t(u,v)}\mu^{\pi}(s|w) && (=: \mathbf{term}_1) \\
&\quad + \sum_{u,v,w}\sum_{x,y,z}\mu^{\pi}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi}(x,y|w)\sqrt{\frac{P(z|x,y)\iota}{n_t(x,y)}} && (=: \mathbf{term}_2) \\
&\quad + \sum_{u,v,w}\sum_{x,y,z}\mu^{\pi}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi}(x,y|w)\min\left\{\frac{\iota}{n_t(x,y)},\ 1\right\} && (=: \mathbf{term}_3) \\
&\quad + \sum_{u,v,w}\sum_{x,y,z}\mu^{\pi}(u,v)\frac{\iota}{n_t(u,v)}\mu^{\pi}(x,y|w) && (=: \mathbf{term}_4)
\end{aligned}$$

We bound $\mathbf{term}_1$ to $\mathbf{term}_4$ separately as below:

$$\mathbf{term}_1 \le \sum_{u,v,w}\frac{\mu^{\pi}(u,v)\iota}{n_t(u,v)} \le S\sum_{u,v}\frac{\mu^{\pi}(u,v)\iota}{n_t(u,v)}.$$

$$
\begin{aligned}
\textbf{term}_2 &= \sum_{u,v,w}\sum_{x,y,z}\sqrt{\frac{\mu^\pi(u,v)P(z|x,y)\mu^\pi(x,y|w)\iota}{n_t(u,v)}}\sqrt{\frac{\mu^\pi(u,v)P(w|u,v)\mu^\pi(x,y|w)\iota}{n_t(x,y)}} \\
&\le \sqrt{\sum_{u,v,w}\sum_{x,y,z}\frac{\mu^\pi(u,v)P(z|x,y)\mu^\pi(x,y|w)\iota}{n_t(u,v)}}\sqrt{\sum_{u,v,w}\sum_{x,y,z}\frac{\mu^\pi(u,v)P(w|u,v)\mu^\pi(x,y|w)\iota}{n_t(x,y)}} \qquad \text{(AM-GM)} \\
&\le \sqrt{H\sum_{u,v,w}\frac{\mu^\pi(u,v)\iota}{n_t(u,v)}}\sqrt{H\sum_{x,y,z}\frac{\mu^\pi(x,y)\iota}{n_t(x,y)}} \\
&\le HS\sum_{u,v}\frac{\mu^\pi(u,v)\iota}{n_t(u,v)}.
\end{aligned}
$$

$$
\begin{aligned}
\textbf{term}_3 &\le \sum_{u,v,w}\sum_{x,y,z}\mu^\pi(u,v)\left(P(w|u,v)+\frac{\iota}{n_t(u,v)}\right)\mu^\pi(x,y|w)\min\left\{\frac{\iota}{n_t(x,y)},\,1\right\} \\
&\le \sum_{u,v,w}\sum_{x,y,z}\mu^\pi(u,v)P(w|u,v)\mu^\pi(x,y|w)\frac{\iota}{n_t(x,y)}+\sum_{u,v,w}\sum_{x,y,z}\mu^\pi(u,v)\frac{\iota}{n_t(u,v)}\mu^\pi(x,y|w) \\
&\le H\sum_{x,y,z}\mu^\pi(x,y)\frac{\iota}{n_t(x,y)}+HS\sum_{u,v,w}\mu^\pi(u,v)\frac{\iota}{n_t(u,v)} \\
&\le HS\sum_{x,y}\frac{\mu^\pi(x,y)\iota}{n_t(x,y)}+HS^2\sum_{u,v}\frac{\mu^\pi(u,v)\iota}{n_t(u,v)}.
\end{aligned}
$$

Similarly,

$$
\textbf{term}_4 \le HS\sum_{u,v,w}\mu^\pi(u,v)\frac{\iota}{n_t(u,v)} \le HS^2\sum_{u,v}\frac{\mu^\pi(u,v)\iota}{n_t(u,v)}.
$$

Collecting all terms we obtain (32). Thus,

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{s}\left|\mu^{\pi_t}(s)-\mu^{\widetilde{P}_t^s,\pi_t}(s)\right|g_t(s) \\
&\le \sum_{t=1}^{T}\sum_{s}\left[\sum_{u,v,w}\mu^{\pi_t}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi_t}(s|w)+HS^2\sum_{u,v}\frac{\mu^{\pi_t}(u,v)\iota}{n_t(u,v)}\right]g_t(s) \\
&\le \underbrace{\sum_{t=1}^{T}\sum_{s}\left[\sum_{u,v,w}\mu^{\pi_t}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi_t}(s|w)\right]g_t(s)}_{(\star)}+HS^3G\sum_{t=1}^{T}\sum_{u,v}\frac{\mu^{\pi_t}(u,v)\iota}{n_t(u,v)} \qquad (33)
\end{aligned}
$$

Fix an h, we consider the summation $(\star)$ restricted to $(u,v,w)\in\mathcal{T}_h\triangleq\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}$. That is,

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{s}\left[\sum_{(u,v,w)\in\mathcal{T}_h}\mu^{\pi_t}(u,v)\sqrt{\frac{P(w|u,v)\iota}{n_t(u,v)}}\mu^{\pi_t}(s|w)\right]g_t(s) \\
&\le \sum_{t=1}^{T}\sum_{s}\left[\sum_{(u,v,w)\in\mathcal{T}_h}\mu^{\pi_t}(u,v)\left(\alpha P(w|u,v)g_t(s)^2+\frac{\iota}{\alpha n_t(u,v)}\right)\mu^{\pi_t}(s|w)\right] \qquad \text{(holds for any } \alpha>0 \text{ by AM-GM)} \\
&\le \alpha\sum_{t=1}^{T}\sum_{s}\sum_{(u,v,w)\in\mathcal{T}_h}\mu^{\pi_t}(u,v)P(w|u,v)\mu^{\pi_t}(s|w)g_t(s)^2+\frac{1}{\alpha}\sum_{t=1}^{T}\sum_{s}\sum_{(u,v,w)\in\mathcal{T}_h}\frac{\mu^{\pi_t}(u,v)\iota}{n_t(u,v)}\mu^{\pi_t}(s|w)
\end{aligned}
$$

$$\leq \alpha \sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) g_t(s)^2 + \frac{H|\mathcal{S}_{h+1}|}{\alpha} \sum_{t=1}^{T} \sum_{u,v} \frac{\mu^{\pi_t}(u,v)\iota}{n_t(u,v)}$$

$$\lesssim \alpha \sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) g_t(s)^2 + \frac{H|\mathcal{S}_{h+1}||\mathcal{S}_h|A\ln(T)\iota}{\alpha} + \frac{H|\mathcal{S}_{h+1}|\ln(1/\delta)\iota}{\alpha}$$

(by Lemma B.2 and the assumption that $\mathcal{E}$ holds)

$$= \sqrt{H|\mathcal{S}_h||\mathcal{S}_{h+1}|A\ln(T)\iota \sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) g_t(s)^2} \qquad \text{(picking the optimal } \alpha \text{ and using our choice of } \delta = \tfrac{1}{T^3}\text{)}$$

$$\leq (|\mathcal{S}_h| + |\mathcal{S}_{h+1}|) \sqrt{HA\ln(T)\iota \sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) g_t(s)^2}.$$

Continue from (33):

$$\sum_{t=1}^{T} \sum_{s} \left| \mu^{\pi_t}(s) - \mu^{\widetilde{P}_t^s, \pi_t}(s) \right| g_t(s)$$

$$\lesssim \sum_{h} (|\mathcal{S}_h| + |\mathcal{S}_{h+1}|) \sqrt{HA\ln(T)\iota \sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) g_t(s)^2} + HS^4AG\ln(T)\iota$$

(by Lemma B.2 and the assumption that $\mathcal{E}$ holds)

$$\lesssim S \sqrt{HA\ln(T)\iota \sum_{t=1}^{T} \sum_{s} \mu^{\pi_t}(s) g_t(s)^2} + HS^4AG\ln(T)\iota.$$

□

Lemma C.5. *For any* π_1, π_2,

$$\sum_{s,a} |\mu^{\pi_1}(s,a) - \mu^{\pi_2}(s,a)| \leq H \sum_{s,a} \mu^{\pi_1}(s) \left| \pi_1(a|s) - \pi_2(a|s) \right|$$

Proof. For any s, a, we can view $\mu^{\pi}(s,a)$ as $V^{\pi}(s_0; \mathbf{1}_{s,a})$ where $\mathbf{1}_{s,a}$ is the loss function that takes the value of 1 on (s,a) and 0 on other state-actions. By the performance difference lemma (Lemma C.1),

$$|\mu^{\pi_1}(s,a) - \mu^{\pi_2}(s,a)| \leq \sum_{s',a'} \mu^{\pi_1}(s') \left| \pi_1(a'|s') - \pi_2(a'|s') \right| Q^{\pi_2}(s',a'; \mathbf{1}_{s,a}).$$

Therefore,

$$\sum_{s,a} |\mu^{\pi_1}(s,a) - \mu^{\pi_2}(s,a)| \leq \sum_{s',a'} \mu^{\pi_1}(s') \left| \pi_1(a'|s') - \pi_2(a'|s') \right| \sum_{s,a} Q^{\pi_2}(s',a'; \mathbf{1}_{s,a})$$

$$= \sum_{s',a'} \mu^{\pi_1}(s') \left| \pi_1(a'|s') - \pi_2(a'|s') \right| Q^{\pi_2}(s',a'; \mathbf{1})$$

($\mathbf{1}$ is the loss function that takes a constant value 1)

$$\leq H \sum_{s',a'} \mu^{\pi_1}(s') \left| \pi_1(a'|s') - \pi_2(a'|s') \right|.$$

□

D. FTRL Regret Bounds

The lemmas in this section are standard results for FTRL, which can be found in e.g. Lattimore & Szepesvári (2018); Zimmert & Seldin (2019); Ito (2021); Luo (2022). We list the results here for completeness.

Lemma D.1. *The FTRL algorithm:*

$$p_t = \operatorname*{argmin}_{p\in\Omega} \left\{ \left\langle p, \sum_{\tau=1}^{t-1} \ell_\tau \right\rangle + \psi_t(p) \right\}$$

guarantees the following:

$$\sum_{t=1}^{T} \langle p_t - u, \ell_t \rangle \le \underbrace{\psi_0(u) - \min_{p\in\Omega} \psi_0(p) + \sum_{t=1}^{T} \left(\psi_t(u) - \psi_t(p_t) - \psi_{t-1}(u) + \psi_{t-1}(p_t) \right)}_{\textit{penalty term}}$$
$$+ \underbrace{\sum_{t=1}^{T} \max_{p\in\Omega} \left(\langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t) \right)}_{\textit{stability term}}.$$

Proof. Let $L_t \triangleq \sum_{\tau=1}^{t} \ell_\tau$. Define $F_t(p) = \langle p, L_{t-1} \rangle + \psi_t(p)$ and $G_t(p) = \langle p, L_t \rangle + \psi_t(p)$. Therefore, p_t is the minimizer of F_t. Let p'_{t+1} be minimizer of G_t. Then by the first-order optimality condition, we have

$$F_t(p_t) - G_t(p'_{t+1}) \le F_t(p'_{t+1}) - G_t(p'_{t+1}) - D_{\psi_t}(p'_{t+1}, p_t) = -\langle p'_{t+1}, \ell_t \rangle - D_{\psi_t}(p'_{t+1}, p_t). \tag{34}$$

By definition, we also have

$$G_t(p'_{t+1}) - F_{t+1}(p_{t+1}) \le G_t(p_{t+1}) - F_{t+1}(p_{t+1}) = \psi_t(p_{t+1}) - \psi_{t+1}(p_{t+1}). \tag{35}$$

Thus,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle p_t, \ell_t \rangle \\
&\le \sum_{t=1}^{T} \left(\langle p_t - p'_{t+1}, \ell_t \rangle - D_{\psi_t}(p'_{t+1}, p_t) + G_t(p'_{t+1}) - F_t(p_t) \right) && \text{(by (34))} \\
&= \sum_{t=1}^{T} \left(\langle p_t - p'_{t+1}, \ell_t \rangle - D_{\psi_t}(p'_{t+1}, p_t) + G_{t-1}(p'_t) - F_t(p_t) \right) + G_T(p'_{T+1}) - G_0(p'_1) \\
&\le \sum_{t=1}^{T} \left(\max_p \left\{ \langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t) \right\} - \psi_t(p_t) + \psi_{t-1}(p_t) \right) + G_T(u) - \min_p \psi_0(p) \\
&&& \text{(by (35), using that } p'_{T+1} \text{ is the minimizer of } G_T\text{)} \\
&= \sum_{t=1}^{T} \left(\max_p \left\{ \langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t) \right\} - \psi_t(p_t) + \psi_{t-1}(p_t) \right) + \sum_{t=1}^{T} \langle u, \ell_t \rangle + \psi_T(u) - \min_p \psi_0(p) \\
&= \sum_{t=1}^{T} \left(\max_p \left\{ \langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t) \right\} + \psi_t(u) - \psi_t(p_t) - \psi_{t-1}(u) + \psi_{t-1}(p_t) \right) + \sum_{t=1}^{T} \langle u, \ell_t \rangle + \psi_0(u) - \min_p \psi_0(p).
\end{aligned}$$

Re-arranging finishes the proof. □

Lemma D.2 (Stability under Tsallis entropy)**.** *Let* $\psi_t(p) = -\frac{2}{\eta_t} \sum_a \sqrt{p(a)}$*, and let* $\ell_t \in \mathbb{R}^A$ *be such that* $\eta_t \sqrt{p(a)} \ell_t(a) \ge -\frac{1}{2}$*. Then*

$$\max_{p\in\triangle(\mathcal{A})} \{ \langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t) \} \le 2\eta_t \sum_a p_t(a)^{\frac{3}{2}} \ell_t(a)^2.$$

Proof. The proof can be found in the Problem 1 of Luo (2022). □

Lemma D.3 (Stability under Shannon entropy). *Let $\psi_t(p) = \sum_a \frac{1}{\eta_t(a)} p(a) \ln p(a)$, and let $\ell_t \in \mathbb{R}^A$ be such that $\eta(a)\ell_t(a) \geq -1$. Then*

$$\max_{p \in \triangle(\mathcal{A})} \{\langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t)\} \leq \sum_a \eta_t(a) p_t(a) \ell_t(a)^2.$$

Proof. The proof can be found in the Proof of Lemma 1 in Chen et al. (2021). □

Lemma D.4 (Stability under log barrier). *Let $\psi_t(p) = \sum_a \frac{1}{\eta_t(a)} \ln \frac{1}{p(a)}$, and let $\ell_t \in \mathbb{R}^A$ be such that $\eta_t(a)p(a)\ell_t(a) \geq -\frac{1}{2}$. Then*

$$\max_{p \in \triangle(\mathcal{A})} \{\langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t)\} \leq \sum_a \eta_t(a) p_t(a)^2 \ell_t(a)^2.$$

Proof.

$$\max_{p \in \triangle(\mathcal{A})} \{\langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t)\} \leq \max_{q \in \mathbb{R}^A_+} \{\langle p_t - q, \ell_t \rangle - D_{\psi_t}(q, p_t)\}$$

Define $f(q) = \langle p_t - q, \ell_t \rangle - D_{\psi_t}(q, p_t)$. Let $q^\star$ be the solution in the last expression. Next, we verify that under the specified conditions, we have $\nabla f(q^\star) = 0$. It suffices to show that there exists $q \in \mathbb{R}^A_+$ such that $\nabla f(q) = 0$ since if such q exists, then it must the maximizer of f and thus $q^\star = q$.

$$[\nabla f(q)]_a = -\ell_t(a) - [\nabla \psi_t(q)]_a + [\nabla \psi_t(p_t)]_a = -\ell_t(a) + \frac{1}{\eta_t(a) q(a)} - \frac{1}{\eta_t(a) p_t(a)}$$

By the condition, we have $-\frac{1}{\eta_t(a)p_t(a)} - \ell_t(a) < 0$ for all a. and so $\nabla f(q) = \mathbf{0}$ has solution in $\mathbb{R}_+$, which is $q(a) = \left(\frac{1}{p_t(a)} + \eta_t(a)\ell_t(a)\right)^{-1}$.

Therefore, $\nabla f(q^\star) = -\ell_t - \nabla \psi_t(q^\star) + \nabla \psi_t(p_t) = 0$, and we have

$$\max_{q \in \mathbb{R}^A_+} \{\langle p_t - q, \ell_t \rangle - D_{\psi_t}(q, p_t)\} = \langle p_t - q^\star, \nabla \psi_t(p_t) - \nabla \psi_t(q^\star) \rangle - D_{\psi_t}(q^\star, p_t) = D_{\psi_t}(p_t, q^\star).$$

It remains to bound $D_{\psi_t}(p_t, q^\star)$, which by definition can be written as

$$D_{\psi_t}(p_t, q^\star) = \sum_a \frac{1}{\eta_t(a)} h\left(\frac{p_t(a)}{q^\star(a)}\right)$$

where $h(x) = x - 1 - \ln(x)$. By the relation between $q^\star(a)$ and $p_t(a)$ we just derived, it holds that $\frac{p_t(a)}{q^\star(a)} = 1 + \eta_t(a)p_t(a)\ell_t(a)$. By the fact that $\ln(1+x) \geq x - x^2$ for all $x \geq -\frac{1}{2}$, we have

$$h\left(\frac{p_t(a)}{q^\star(a)}\right) = \eta_t(a)p_t(a)\ell_t(a) - \ln(1 + \eta_t(a)p_t(a)\ell_t(a)) \leq \eta_t(a)^2 p_t(a)^2 \ell_t(a)^2$$

which gives the desired bound.

□

Lemma D.5 (FTRL with Tsallis entropy). *Let $\psi_t(p) = -\frac{2}{\eta_t} \sum_a \sqrt{p(a)}$ for non-increasing η_t, and let x_t be such that $\eta_t \sqrt{p_t(a)}(\ell_t(a) + x_t) \geq -\frac{1}{2}$ for all t, a. Then the FTRL algorithm in Lemma D.1 ensures for any $u \in \triangle(\mathcal{A})$,*

$$\sum_{t=1}^T \langle p_t - u, \ell_t \rangle \leq \frac{2\sqrt{A}}{\eta_0} + 2\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \xi_t + 2\sum_{t=1}^T \eta_t \sum_a p_t(a)^{\frac{3}{2}} \left(\ell_t(a) + x_t\right)^2,$$

where $\xi_t = \sum_a \sqrt{p_t(a)}(1 - p_t(a))$.

Proof. We use Lemma D.1, and bound the penalty term and stability individually.

$$\begin{aligned}\text{penalty term} &= \frac{2}{\eta_0}\max_{p\in\triangle(\mathcal{A})}\sum_a\left(\sqrt{p(a)}-\sqrt{u(a)}\right)+2\sum_{t=1}^T\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\sum_a\left(\sqrt{p_t(a)}-\sqrt{u(a)}\right)\\ &\le \frac{2\sqrt{A}}{\eta_0}+2\sum_{t=1}^T\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\left(\sum_a\sqrt{p_t(a)}-1\right)\\ &\le \frac{2\sqrt{A}}{\eta_0}+2\sum_{t=1}^T\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\xi_t.\end{aligned}$$

Bounding the stability term:

$$\text{stability term} = \sum_{t=1}^T\max_{p\in\triangle(\mathcal{A})}\left\{\langle p_t-p,\ell_t+x_t\mathbf{1}\rangle-D_{\psi_t}(p,p_t)\right\}\le 2\sum_{t=1}^T\eta_t\sum_a p_t(a)^{\frac{3}{2}}\left(\ell_t(a)+x_t\right)^2$$

where the first equality is because $\langle p_t-p,\mathbf{1}\rangle=0$ for $p_t,p\in\triangle(\mathcal{A})$, and the last inequality is by Lemma D.2. □

Lemma D.6 (FTRL with Shannon entropy)**.** *Let $\psi_t(p)=\sum_a\frac{1}{\eta_t(a)}p(a)\ln p(a)$, for non-increasing $\eta_t(a)$ such that $\eta_0(a)=\eta_0$ for all a. Assume that $\eta_t(a)\ell_t(a)\ge -1$ for all t,a, and assume $A\le T$. Then for any $u\in\triangle(\mathcal{A})$,*

$$\sum_{t=1}^T\langle p_t-u,\ell_t\rangle\le\frac{\ln A}{\eta_0}+6\sum_{t=1}^T\sum_a\left(\frac{1}{\eta_t(a)}-\frac{1}{\eta_{t-1}(a)}\right)\xi_t(a)+\sum_{t=1}^T\sum_a\eta_t(a)p_t(a)\ell_t(a)^2+\frac{1}{T^2}\sum_{t=1}^T\left\langle -u+\frac{1}{A}\mathbf{1},\ell_t\right\rangle.$$

where $\xi_t(a)=\min\{p_t(a)\ln(T),1-p_t(a)\}$.

Proof. Let $u'=\left(1-\frac{1}{T^2}\right)u+\frac{1}{AT^2}\mathbf{1}$. We use Lemma D.1, and bound the penalty term and stability individually (with respect to u').

penalty term

$$\begin{aligned}&=\frac{1}{\eta_0}\max_p\sum_a\left(p(a)\ln\frac{1}{p(a)}-u'(a)\ln\frac{1}{u'(a)}\right)+\sum_{t=1}^T\sum_a\left(\frac{1}{\eta_t(a)}-\frac{1}{\eta_{t-1}(a)}\right)\left(p_t(a)\ln\frac{1}{p_t(a)}-u'(a)\ln\frac{1}{u'(a)}\right)\\ &\le\frac{\ln A}{\eta_0}+\sum_{t=1}^T\sum_a\left(\frac{1}{\eta_t(a)}-\frac{1}{\eta_{t-1}(a)}\right)\left(p_t(a)\ln\frac{1}{p_t(a)}-u'(a)\ln\frac{1}{u'(a)}\right).\end{aligned}$$

To bound $p_t(a)\ln\frac{1}{p_t(a)}-u'(a)\ln\frac{1}{u'(a)}$, first observe that $p_t(a)\ln\frac{1}{p_t(a)}=p_t(a)\ln\left(1+\frac{1-p_t(a)}{p_t(a)}\right)\le p_t(a)\cdot\frac{1-p_t(a)}{p_t(a)}\le 1-p_t(a)$ because $\ln(1+x)\le x$. By the definition of u', we have

$$u'(a)\ln\frac{1}{u'(a)}\ge\min\left\{\frac{1}{AT^2}\ln(AT^2),\left(1-\frac{1}{T^2}\right)\ln\frac{1}{1-\frac{1}{T^2}}\right\}\ge\min\left\{\frac{1}{AT^2},\left(1-\frac{1}{T^2}\right)\frac{1}{T^2}\right\}=\frac{1}{AT^2}.$$

If $p_t(a)\le\frac{1}{A^2T^4}$, then

$$p_t(a)\ln\frac{1}{p_t(a)}-u'(a)\ln\frac{1}{u'(a)}\le\frac{1}{A^2T^4}\ln(A^2T^4)-\frac{1}{AT^2}=\frac{2\ln(AT^2)-AT^2}{A^2T^4}\le 0$$

where the first inequality is because $x\ln(x)$ is increasing for $x\le e^{-1}$, and last inequality is because $2\ln(x)-x<0$ for all $x\in\mathbb{R}$. If $p_t(a)>\frac{1}{A^2T^4}$, then $p_t(a)\ln\frac{1}{p_t(a)}\le p_t(a)\ln(A^2T^4)\le 6p_t(a)\ln(T)$ by the assumption $A\le T$. Combining all arguments above, we get

$$\text{penalty term}\le\frac{\ln A}{\eta_0}+6\sum_{t=1}^T\sum_a\left(\frac{1}{\eta_t(a)}-\frac{1}{\eta_t(a)}\right)\min\{1-p_t(a),p_t(a)\ln(T)\}.$$

Bounding the stability term:

$$\text{stability term} = \sum_{t=1}^{T} \max_{p \in \triangle(\mathcal{A})} \left\{ \langle p_t - p, \ell_t \rangle - D_{\psi_t}(p, p_t) \right\} \leq \sum_{t=1}^{T} \sum_{a} \eta_t(a) p_t(a) \ell_t(a)^2$$

where the last inequality is by Lemma D.3.

Therefore,

$$\sum_{t=1}^{T} \langle p_t - u', \ell_t \rangle \leq \frac{\ln A}{\eta_0} + 6 \sum_{t=1}^{T} \sum_{a} \left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)} \right) \xi_t(a) + \sum_{t=1}^{T} \sum_{a} \eta_t(a) p_t(a) \ell_t(a)^2$$

Then noticing that

$$\begin{aligned} \sum_{t=1}^{T} \langle p_t - u, \ell_t \rangle &= \sum_{t=1}^{T} \langle p_t - u', \ell_t \rangle + \sum_{t=1}^{T} \langle u' - u, \ell_t \rangle \\ &= \sum_{t=1}^{T} \langle p_t - u', \ell_t \rangle + \frac{1}{T^2} \sum_{t=1}^{T} \left\langle -u + \frac{1}{A} \mathbf{1}, \ell_t \right\rangle \end{aligned}$$

finishes the proof. □

Lemma D.7 (FTRL with log barrier)**.** *Let $\psi_t(p) = \sum_a \frac{1}{\eta_t(a)} \ln \frac{1}{p(a)}$ for non-increasing $\eta_t(a)$ with $\eta_0(a) = \eta_0$ for all a, and let x_t be such that $\eta_t(a) p_t(a) (\ell_t(a) + x_t) \geq -\frac{1}{2}$ for all t, a. Then for any $u \in \triangle(\mathcal{A})$,*

$$\sum_{t=1}^{T} \langle p_t - u, \ell_t \rangle \leq \frac{3A \ln T}{\eta_0} + 4 \sum_{a} \left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)} \right) \ln(T) + \sum_{t=1}^{T} \sum_{a} \eta_t(a) p_t(a) \ell_t(a)^2 + \frac{1}{T^3} \sum_{t=1}^{T} \left\langle -u + \frac{1}{A} \mathbf{1}, \ell_t \right\rangle.$$

Proof. Let $u' = \left(1 - \frac{1}{T^3}\right) u + \frac{1}{AT^3} \mathbf{1}$. We use Lemma D.1, and bound the penalty term and stability individually (with respect to u').

$$\begin{aligned} \text{penalty term} &\leq \frac{A \ln(T^3)}{\eta_0} + \sum_{t=1}^{T} \sum_{a} \left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)} \right) \left(\ln \frac{1}{u'(a)} - \ln \frac{1}{p_t(a)} \right) \\ &\leq \frac{3A \ln T}{\eta_0} + \sum_{t=1}^{T} \sum_{a} \left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)} \right) \ln(AT^3) \\ &\leq \frac{3A \ln T}{\eta_0} + 4 \sum_{t=1}^{T} \sum_{a} \left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)} \right) \ln(T) \qquad \text{(because } A \leq T\text{)} \end{aligned}$$

Bounding the stability term:

$$\text{stability term} = \sum_{t=1}^{T} \max_{p \in \triangle(\mathcal{A})} \left\{ \langle p_t - p, \ell_t + x_t \mathbf{1} \rangle - D_{\psi_t}(p, p_t) \right\} \leq \sum_{t=1}^{T} \sum_{a} \eta_t(a) p_t(a)^2 \left(\ell_t(a) + x_t \right)^2$$

where the first equality is because $\langle p_t - p, \mathbf{1} \rangle = 0$, and the last inequality is by Lemma D.4. Then noticing that

$$\begin{aligned} \sum_{t=1}^{T} \langle p_t - u, \ell_t \rangle &= \sum_{t=1}^{T} \langle p_t - u', \ell_t \rangle + \sum_{t=1}^{T} \langle u' - u, \ell_t \rangle \\ &= \sum_{t=1}^{T} \langle p_t - u', \ell_t \rangle + \frac{1}{T^3} \sum_{t=1}^{T} \left\langle -u + \frac{1}{A} \mathbf{1}, \ell_t \right\rangle \end{aligned}$$

finishes the proof. □

E. Analysis for FTRL Regret Bound (Lemma 6.1)

E.1. Tsallis entropy

Proof of Lemma 6.1 (Tsallis entropy). We focus on a particular s, and use $\pi_t(a)$, $\widehat{Q}_t(a)$, $B_t(a), C_t(a)$, η_t, μ_t, ξ_t, b_t to denote $\pi_t(a|s)$, $\widehat{Q}_t(s,a)$, $B_t(s,a)$, $C_t(s,a)$, $\eta_t(s)$, $\mu_t(s)$, $\xi_t(s)$, $b_t(s)$, respectively.

By Lemma D.5, we have for any π

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \pi_t - \pi, \widehat{Q}_t - B_t - C_t\rangle\right] \tag{36}$$

$$\leq \frac{2\sqrt{A}}{\eta_0} + \mathbb{E}\left[2\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\xi_t + 2\sum_{t=1}^{T}\sum_{a}\eta_t\pi_t(a)^{\frac{3}{2}}\left(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t\right)^2\right] \tag{37}$$

for arbitrary $x_t \in \mathbb{R}$ such that $\eta_t\sqrt{\pi_t(a|s)}(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t) \geq -\frac{1}{2}$ for all t, a. Our choice of x_t is the following:

$$x_t = -\left\langle \pi_t, \widehat{Q}_t\right\rangle Y_t. \tag{38}$$

with $Y_t \triangleq \mathbb{I}\left[\frac{\eta_t}{\mu_t} \leq \frac{1}{8H}\right]$. Below, we verify that $\eta_t\sqrt{\pi_t(a)}\left(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t\right) \geq -\frac{1}{2}$:

$$\begin{aligned}
&\eta_t\sqrt{\pi_t(a)}\left(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t\right) \\
&\geq \eta_t\sqrt{\pi_t(a)}\left(-B_t(a) - C_t(a) - \left\langle \pi_t, \widehat{Q}_t\right\rangle Y_t\right) && \text{(using (38) and } \widehat{Q}_t(a) \geq 0) \\
&\geq -\eta_t B_t(a) - \eta_t C_t(a) - \eta_t\sum_{a'}\pi_t(a')\frac{H\mathbb{I}_t(s,a')}{\mu_t\pi_t(a')}Y_t && \text{(by the definition of } \widehat{Q}_t(a)) \\
&\geq -\frac{1}{8H} - \frac{1}{4H^2} - \frac{H\eta_t}{\mu_t}Y_t && \text{(using Lemma E.1, } C_t(a) \leq H^2 \text{ and } \eta_t \leq \tfrac{1}{4H^4}) \\
&\geq -\frac{1}{2}. && \text{(by the definition of } Y_t = \mathbb{I}\left[\tfrac{\eta_t}{\mu_t} \leq \tfrac{1}{8H}\right])
\end{aligned}$$

Continued from (37) with the choice of x_t:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle \pi_t - \pi, \widehat{Q}_t - B_t - C_t\rangle\right] \\
&\leq \frac{2\sqrt{A}}{\eta_0} + \mathbb{E}\left[2\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\xi_t + 2\sum_{t=1}^{T}\sum_{a}\eta_t\pi_t(a)^{\frac{3}{2}}\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle Y_t - B_t(a) - C_t(a)\right)^2\right] \\
&\leq O(H^4A) + \mathbb{E}\left[2\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\xi_t + 8\sum_{t=1}^{T}\sum_{a}\eta_t\pi_t(a)^{\frac{3}{2}}\left(\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle\right)^2 + \widehat{Q}_t(a)^2Y_t' + B_t(a)^2 + C_t(a)^2\right)\right] \\
&\qquad\qquad \text{(define } Y_t' = 1 - Y_t) \\
&\leq O(H^4A) + \mathbb{E}\left[2\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\xi_t + \underbrace{8\sum_{t=1}^{T}\sum_{a}\eta_t\pi_t(a)^{\frac{3}{2}}\left(\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle\right)^2 + \widehat{Q}_t(a)^2Y_t'\right)}_{\textbf{term}_1}\right] \\
&\quad + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a)B_t(a)\right] + \mathbb{E}\left[8\sum_{t=1}^{T}\sum_{a}\eta_t\pi_t(a)C_t(a)^2\right]. && \text{(using Lemma E.1)}
\end{aligned} \tag{39}$$

To bound $\textbf{term}_1$, notice that

$$\mathbb{E}_t\left[\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle\right)^2\right] = \mathbb{E}_t\left[\left(\frac{\mathbb{I}_t(s,a)L_{t,h}}{\mu_t\pi_t(a)} - \frac{\mathbb{I}_t(s)L_{t,h}}{\mu_t}\right)^2\right] \qquad \text{(assume } s \in \mathcal{S}_h)$$

$$\leq \mu_t \pi_t(a) \left(\frac{H}{\mu_t \pi_t(a)} - \frac{H}{\mu_t} \right)^2 + \mu_t (1 - \pi_t(a)) \left(\frac{H}{\mu_t} \right)^2$$
$$= \frac{1}{\mu_t \pi_t(a)} (1 - \pi_t(a))^2 H^2 + \frac{1}{\mu_t} (1 - \pi_t(a)) H^2$$
$$= \frac{1 - \pi_t(a)}{\mu_t \pi_t(a)} H^2$$

and that

$$\mathbb{E}_t \left[\widehat{Q}_t(a)^2 Y_t' \right] = \mathbb{E}_t \left[\left(\frac{\mathbb{I}_t(s,a) L_{t,h}}{\mu_t \pi_t(a)} \right)^2 \right] Y_t' \leq \frac{H^2}{\mu_t \pi_t(a)} Y_t'.$$

Therefore,

$$\mathbb{E}[\mathbf{term}_1] \leq \mathbb{E} \left[8H^2 \sum_{t=1}^{T} \sum_a \eta_t \pi_t(a)^{\frac{3}{2}} \left(\frac{1 - \pi_t(a)}{\mu_t \pi_t(a)} + \frac{1}{\mu_t \pi_t(a)} Y_t' \right) \right]$$
$$\leq \mathbb{E} \left[8H^2 \sum_{t=1}^{T} \frac{\eta_t}{\mu_t} \sum_a \left(\sqrt{\pi_t(a)} (1 - \pi_t(a)) + \sqrt{\pi_t(a)} Y_t' \right) \right]$$
$$\leq \mathbb{E} \left[8H^2 \sum_{t=1}^{T} \frac{\eta_t}{\mu_t} \left(\xi_t + \sqrt{A} Y_t' \right) \right].$$

Notice that

$$\frac{8H^2 \eta_t}{\mu_t} \leq 2H \frac{\frac{1}{\mu_t}}{\sqrt{\sum_{\tau=1}^{t} \frac{1}{\mu_\tau}}} \leq 4H \left(\sqrt{\sum_{\tau=1}^{t} \frac{1}{\mu_\tau}} - \sqrt{\sum_{\tau=1}^{t-1} \frac{1}{\mu_\tau}} \right) \leq \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}$$

Thus

$$\mathbb{E}[\mathbf{term}_1] \leq \mathbb{E} \left[\sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \left(\xi_t + \sqrt{A} Y_t' \right) \right],$$

and continuing from (39) we have

$$\mathbb{E} \left[\sum_{t=1}^{T} \langle \pi_t - \pi, \widehat{Q}_t - B_t - C_t \rangle \right]$$
$$\leq O(H^4 A) + 3\mathbb{E} \left[\sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \left(\xi_t + \sqrt{A} Y_t' \right) \right] + \mathbb{E} \left[\frac{1}{H} \sum_{t=1}^{T} \sum_a \pi_t(a) B_t(a) \right] + \mathbb{E} \left[8 \sum_{t=1}^{T} \sum_a \eta_t \pi_t(a) C_t(a)^2 \right]$$
$$\leq O(H^4 A) + \mathbb{E} \left[\sum_{t=1}^{T} b_t \right] + \mathbb{E} \left[\frac{1}{H} \sum_{t=1}^{T} \sum_a \pi_t(a) B_t(a) \right]$$

with b_t defined in (15). This finishes the proof.

□

Lemma E.1 (Tsallis entropy)**.** $\eta_t(s) B_t(s,a) \leq \frac{1}{8H}$.

Proof. By the definition of $b_t(s)$ in (15), we have

$$b_t(s) \leq 8\sqrt{A} \left(\frac{1}{\eta_t(s)} - \frac{1}{\eta_{t-1}(s)} \right) + 8\eta_t(s) H^4 \qquad (C_t(s,a) \leq H^2)$$

$$
\begin{aligned}
&= 32H\sqrt{A}\left(\sqrt{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}} - \sqrt{\sum_{\tau=1}^{t-1}\frac{1}{\mu_\tau(s)}}\right) + 8\eta_t(s)H^4 \\
&\le 32H\sqrt{A} \times \frac{\frac{1}{\mu_t(s)}}{\sqrt{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}} + 8 \times \frac{1}{4H^4} \times H^4 \\
&\le 32H\sqrt{\frac{A}{\mu_t(s)}} + 2 \le 34H\sqrt{\frac{A}{\gamma_t}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\eta_t(s)B_t(s,a) &\le \eta_t(s)\left(1+\frac{1}{H}\right)^H H \max_{s'} b_t(s') \\
&\le \min\left\{\frac{1}{1600H^4\sqrt{A}}, \frac{1}{4H\sqrt{t}}\right\} \times 34eH^2\sqrt{\frac{A}{\gamma_t}} \\
&\le 100H^2 \min\left\{\frac{1}{1600H^4\sqrt{A}}, \frac{1}{4H\sqrt{t}}\right\} \times \max\left\{\sqrt{\frac{At}{10^6H^4A^2}}, \sqrt{A}\right\} && \text{(by the definition of } \gamma_t\text{)} \\
&\le \frac{1}{8H}.
\end{aligned}
$$

□

E.2. Shannon entropy

Proof of Lemma 6.1 (Shannon entropy). We focus on a particular s, and use $\pi_t(a)$, $\widehat{Q}_t(a)$, $B_t(a)$, $\eta_t(a)$, μ_t, b_t to denote $\pi_t(a|s)$, $\widehat{Q}_t(s,a)$, $B_t(s,a)$, $\eta_t(s,a)$, $\mu_t(s)$, $b_t(s)$, respectively.

Notice that for any t, a, since $\widehat{Q}_t(a) \ge 0$, $\eta_t(a)B_t(a) \le \frac{1}{4H}$ (by Lemma E.2), and $\eta_t(a)C_t(a) \le \frac{1}{4H^4} \times H^2 = \frac{1}{4H^2}$ (because $\eta_t(a) \le \eta_0(a) = \frac{1}{4H^4}$ and $C_t(a) \le H^2$), we have

$$
\eta_t(a)(\widehat{Q}_t(a) - B_t(a) - C_t(a)) \ge -\frac{1}{4H} - \frac{1}{4H^2} \ge -1.
$$

Besides, for any a,

$$
\begin{aligned}
\left|\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Q}_t(a)\right]\right| &\le \mathbb{E}\left[\sum_{t=1}^{T}\frac{H}{\mu_t}\right] \le \sum_{t=1}^{T}\frac{H}{\gamma_t} \le HT^2 && \text{(by the definition of } \gamma_t\text{)} \\
\mathbb{E}\left[\sum_{t=1}^{T}B_t(a)\right] &\le 400T^2\sqrt{\log T} && \text{(by Lemma E.2)} \\
\mathbb{E}\left[\sum_{t=1}^{T}C_t(a)\right] &\le H^2T && (C_t(a) \le H^2)
\end{aligned}
$$

With these inequalities, by Lemma D.6, the following holds for any π:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle \pi_t - \pi, \widehat{Q}_t - B_t - C_t\rangle\right] \\
&\le \sum_a \frac{\ln A}{\eta_0(a)} + \mathbb{E}\left[6\sum_{t=1}^{T}\sum_a\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\xi_t(a) + \sum_{t=1}^{T}\sum_a \eta_t(a)\pi_t(a)\left(\widehat{Q}_t(a) - B_t(a) - C_t(a)\right)^2\right] \qquad (40) \\
&\quad + \frac{2}{T^2}\max_a\left|\mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{Q}_t(a) - B_t(a) - C_t(a)\right)\right]\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq O(H^4 A \ln(T)) + \mathbb{E}\left[6\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\xi_t(a) + 3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\pi_t(a)\left(\widehat{Q}_t(a)^2 + B_t(a)^2 + C_t(a)^2\right)\right] \\
&\leq O(H^4 A \ln(T)) + \mathbb{E}\left[6\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\xi_t(a) + 3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\left(\frac{H^2}{\mu_t} + \pi_t(a)B_t(a)^2 + \pi_t(a)C_t(a)^2\right)\right] \\
&\leq O(H^4 A \ln(T)) + \mathbb{E}\left[6\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\xi_t(a) + 3\sum_{t=1}^{T}\sum_{a}\frac{H^2\eta_t(a)}{\mu_t}\right] \\
&\quad + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a)B_t(a) + 3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\pi_t(a)C_t(a)^2\right] \qquad \text{(by Lemma E.2)}
\end{aligned} \tag{41}
$$

By the update $\eta_t(a)$,

$$
\frac{1}{\eta_t(a)} \geq 4H\sqrt{\log T}\sum_{\tau=1}^{t}\frac{1}{\mu_\tau} \times \frac{1}{\sqrt{\sum_{\tau=1}^{T}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[T]}\frac{1}{\mu_\tau}}}.
$$

Therefore,

$$
\begin{aligned}
3H^2\sum_{t=1}^{T}\sum_{a}\frac{\eta_t(a)}{\mu_t} &\leq \frac{H}{\sqrt{\log T}}\sum_{a}\sqrt{\sum_{\tau=1}^{T}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[T]}\frac{1}{\mu_\tau}} \times \sum_{t=1}^{T}\frac{\frac{1}{\mu_t}}{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau}} \\
&\leq 2H\sqrt{\log T}\sum_{a}\sqrt{\sum_{\tau=1}^{T}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[T]}\frac{1}{\mu_\tau}} \\
&= 2H\sqrt{\log T}\sum_{a}\sum_{t=1}^{T}\left(\sqrt{\sum_{\tau=1}^{t}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[t]}\frac{1}{\mu_\tau}} - \sqrt{\sum_{\tau=1}^{t-1}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[t-1]}\frac{1}{\mu_\tau}}\right) \\
&\leq 2H\sqrt{\log T}\sum_{a}\sum_{t=1}^{T}\frac{\frac{\xi_t(a)}{\mu_t} + \max_{\tau\in[t]}\frac{1}{\mu_\tau} - \max_{\tau\in[t-1]}\frac{1}{\mu_\tau}}{\sqrt{\sum_{\tau=1}^{t}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[t]}\frac{1}{\mu_\tau}}} \\
&= 2H\sqrt{\log T}\sum_{a}\sum_{t=1}^{T}\frac{\frac{\xi_t(a)}{\mu_t} + \frac{1}{\mu_t}\left(1 - \frac{\min_{\tau\in[t]}\mu_\tau}{\min_{\tau\in[t-1]}\mu_\tau}\right)}{\sqrt{\sum_{\tau=1}^{t}\frac{\xi_\tau(a)}{\mu_\tau} + \max_{\tau\in[t]}\frac{1}{\mu_\tau}}} \\
&\leq \sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\left(\xi_t(a) + 1 - \frac{\min_{\tau\in[t]}\mu_\tau}{\min_{\tau\in[t-1]}\mu_\tau}\right)
\end{aligned}
$$

where we use (19) in the last inequality. Using this in (41), we get

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle\pi_t - \pi, \widehat{Q}_t - B_t - C_t\rangle\right] &\leq O(H^4 A \ln(T)) + \mathbb{E}\left[7\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\left(\xi_t(a) + 1 - \frac{\min_{\tau\in[t]}\mu_\tau}{\min_{\tau\in[t-1]}\mu_\tau}\right)\right] \\
&\quad + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a)B_t(a) + 3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\pi_t(a)C_t(a)^2\right] \\
&\leq O(H^4 A \ln(T)) + \mathbb{E}\left[\sum_{t=1}^{T}b_t + \frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a)B_t(a)\right],
\end{aligned}
$$

where we use the definition of b_t in (18). This finishes the proof.

□

Lemma E.2 (Shannon entropy). *$\eta_t(s,a)B_t(s,a) \le \frac{1}{4H}$ and $B_t(s,a) \le 400\sqrt{T\log T}$.*

Proof. By the definition of $b_t(s)$ in (18), we have

$$\begin{aligned}
b_t(s) &\le 16\sum_a \left(\frac{1}{\eta_t(s,a)} - \frac{1}{\eta_{t-1}(s,a)}\right) + 8\sum_a \eta_t(s,a)\pi_t(a|s)H^4 && (C_t(s,a) \le H^2)\\
&\le 64\sum_a \left(\frac{H}{\mu_t(s)\sqrt{\sum_{\tau=1}^{t-1}\frac{\xi_\tau(s,a)}{\mu_\tau(s)} + \frac{1}{\mu_t(s)}}} + \frac{H}{\sqrt{t}}\right)\sqrt{\log T} + 2 && (\text{using (19) and } \eta_t(s,a) \le \tfrac{1}{4H^4})\\
&\le 64\left(\frac{HA}{\sqrt{\mu_t(s)}} + \frac{HA}{\sqrt{t}}\right)\sqrt{\log T} + 2 \le \frac{132HA\sqrt{\log T}}{\sqrt{\gamma_t}}.
\end{aligned}$$

Further notice that

$$\frac{1}{\eta_t(s,a)} \ge 4\sum_{\tau=1}^{t}\frac{H\sqrt{\log T}}{\sqrt{\tau}} \ge 4H\sqrt{t\log T}.$$

Therefore,

$$\begin{aligned}
B_t(s,a) &\le H\left(1+\frac{1}{H}\right)^H \max_s b_t(s) \le \frac{396H^2A\sqrt{\log T}}{\sqrt{\gamma_t}} \le 400H^2A\sqrt{T\log T}\\
\eta_t(s,a)B_t(s,a) &\le \min\left\{\frac{1}{1600H^4A\sqrt{\log T}}, \frac{1}{4H\sqrt{t\log T}}\right\} \times \frac{396H^2A\sqrt{\log T}}{\sqrt{\gamma_t}}\\
&\le \min\left\{\frac{1}{1600H^4A\sqrt{\log T}}, \frac{1}{4H\sqrt{t\log T}}\right\} \times \max\left\{\frac{396H^2A\sqrt{t\log T}}{\sqrt{10^6H^4A^2}}, 396H^2A\sqrt{\log T}\right\}\\
&\qquad\qquad (\text{by the definition of } \gamma_t)\\
&\le \frac{1}{4H}
\end{aligned}$$

by the definition of γ_t. □

E.3. Log barrier

Proof of Lemma 6.1 (log barrier). We focus on a particular s, and use $\pi_t(a)$, $\widehat{Q}_t(a)$, $B_t(a), C_t(a)$, η_t, μ_t, $\zeta_t(a)$, to denote $\pi_t(a|s)$, $\widehat{Q}_t(s,a)$, $B_t(s,a)$, $C_t(s,a)$, $\eta_t(s)$, $\mu_t(s)$, $\zeta_t(s,a)$, respectively.

By Lemma D.7,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle \pi_t - \pi, \widehat{Q}_t - B_t - C_t\rangle\right]\\
&\le O(H^4A\ln(T)) + \mathbb{E}\left[4\sum_{t=1}^{T}\sum_a\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\log(T) + \sum_{t=1}^{T}\sum_a \eta_t(a)\pi_t(a)^2\left(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t\right)^2\right]\\
&\quad + \frac{2}{T^3}\max_a\left|\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Q}_t(a) - B_t(a) - C_t(a)\right]\right| \qquad (42)
\end{aligned}$$

for arbitrary $x_t \in \mathbb{R}$ such that $\eta_t(a)\pi_t(a)(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t) \ge -1$. Recall that with log barrier, there are real episodes and virtual episodes in which $\ell_t(s,a) = 0$ for all (s,a). Let $Y_t = 0$ if t is a virtual episode, and $Y_t = 1$ otherwise.

We define

$$x_t = -\left\langle \pi_t, \widehat{Q}_t\right\rangle. \qquad (43)$$

Below, we verify that $\eta_t(a)\pi_t(a)\left(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t\right) \geq -\frac{1}{2}$:

$$\begin{aligned}
&\eta_t(a)\pi_t(a)\left(\widehat{Q}_t(a) - B_t(a) - C_t(a) + x_t\right) \\
&\geq \eta_t(a)\pi_t(a)\left(-B_t(a) - C_t(a) - \left\langle \pi_t, \widehat{Q}_t \right\rangle\right) && \text{(using (43) and } \widehat{Q}_t(a) \geq 0) \\
&\geq -\eta_t(a)\pi_t(a|s)B_t(a) - \eta_t(a)C_t(a) - \eta_t(a)\sum_{a'}\pi_t(a')\frac{H\mathbb{I}_t(s,a')}{\mu_t\pi_t(a')}Y_t && \text{(when } Y_t = 0,\ \widehat{Q}_t(a) = 0) \\
&\geq -\frac{1}{8H} - \frac{1}{4H^2} - \frac{H\eta_t}{\mu_t}Y_t && \text{(by Lemma E.3 and that } C_t(a) \leq H^2 \text{ and } \eta_t(a) \leq \tfrac{1}{4H^4}) \\
&\geq -\frac{1}{2}. && \text{(when } Y_t = 1 \text{ (real episode), } \tfrac{\eta_t(a)}{\mu_t} \leq \tfrac{1}{8H})
\end{aligned}$$

Besides, for any a,

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{t=1}^{T}\widehat{Q}_t(a)\right]\right| &\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{H}{\mu_t}\right] \leq \sum_{t=1}^{T}\frac{H}{\gamma_t} \leq HT^2 && \text{(by the definition of } \gamma_t) \\
\mathbb{E}\left[\sum_{t=1}^{T}B_t(a)\right] &\leq 15ST^2 && \text{(by Lemma E.3)} \\
\mathbb{E}\left[\sum_{t=1}^{T}C_t(a)\right] &\leq H^2T && (C_t(a) \leq H^2)
\end{aligned}$$

Below, we continue from (42) with our choice of x_t:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle \pi_t - \pi, \widehat{Q}_t - B_t - C_t\rangle\right] \\
&\leq O(H^4SA\ln(T)) + \mathbb{E}\left[4\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\log(T)\right. \\
&\quad \left. + 3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\pi_t(a)^2\left(\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle\right)^2 + B_t(a)^2 + C_t(a)^2\right)\right] \\
&\leq O(H^4SA\ln(T)) + \mathbb{E}\Bigg[\underbrace{4\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_t(a)} - \frac{1}{\eta_{t-1}(a)}\right)\log(T)}_{\textbf{term}_1} + \underbrace{3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\pi_t(a)^2\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle\right)^2}_{\textbf{term}_2}\Bigg] \\
&\quad + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a)B_t(a)\right] + \mathbb{E}\Bigg[\underbrace{3\sum_{t=1}^{T}\sum_{a}\eta_t(a)\pi_t(a)C_t(a)^2}_{\textbf{term}_3}\Bigg]. && \text{(by Lemma E.3)}
\end{aligned}$$

We further manipulate $\textbf{term}_2$ (suppose that $s \in \mathcal{S}_h$). In virtual episodes, $\textbf{term}_2 = 0$, and in real episodes,

$$\begin{aligned}
\eta_t(a)\pi_t(a)^2\left(\widehat{Q}_t(a) - \langle \pi_t, \widehat{Q}_t\rangle\right)^2 &= \eta_t(a)\pi_t(a|s)^2\left(\frac{\mathbb{I}_t(s,a)L_{t,h}}{\mu_t\pi_t(a)} - \frac{\mathbb{I}_t(s)L_{t,h}}{\mu_t}\right)^2 \\
&= \eta_t(a)\left(\frac{\mathbb{I}_t(s,a)L_{t,h}}{\mu_t} - \frac{\pi_t(a|s)\mathbb{I}_t(s)L_{t,h}}{\mu_t}\right)^2 \\
&= \frac{\eta_t(a)}{\mu_t^2}\left(\mathbb{I}_t(s,a) - \pi_t(a|s)\mathbb{I}_t(s)\right)L_{t,h}^2 \\
&= \frac{\eta_t(a)\zeta_t(a)}{\mu_t^2}
\end{aligned}$$

$$\leq \frac{\log T}{4}\left(\frac{1}{\eta_{t+1}(a)} - \frac{1}{\eta_t(a)}\right) \qquad \text{(by Eq. (23))}$$

By the definition of b_t in (22), we have $\mathbb{E}[\mathbf{term}_1 + \mathbf{term}_2 + \mathbf{term}_3] \leq \mathbb{E}\left[\sum_{t=1}^{T} b_t\right]$, which finishes the proof.

□

Lemma E.3 (log barrier)**.** *$\eta_t(s,a)\pi_t(a|s)B_t(s,a) \leq \frac{1}{8H}$ and $B_t(s,a) \leq 15ST$.*

Proof. If t is a real episode,

$$\begin{aligned} b_t(s) &= 8\sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right) \\ &= 32\sum_a \frac{\eta_t(s,a)\mathbb{I}_t(s,a)L_t(s,a)^2}{\mu_t(s)^2} \leq 32H^2 \times \max_a \frac{\eta_t(s,a)}{\mu_t(s)} \times \frac{1}{\mu_t(s)} \leq \frac{1}{\mu_t(s)} \times 32H^2 \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right). \end{aligned} \tag{44}$$

Therefore,

$$\begin{aligned} B_t(s,a) &\leq b_t(s) + 3\sum_{s':h(s')>h(s)} \mu^{\widetilde{P}_t,\pi_t}(s'|s,a)b_t(s') \\ &\leq \left(\frac{1}{\mu_t(s)} + 3\sum_{s':h(s')>h(s)} \mu^{\widetilde{P}_t,\pi_t}(s'|s,a)\frac{1}{\mu_t(s')}\right) \times 32H^2 \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right) \\ &\leq \left(\frac{1}{\mu_t(s)} + 3\sum_{s':h(s')>h(s)} \mu^{\widetilde{P}_t,\pi_t}(s'|s,a) \times \frac{1}{\overline{\mu}_t^{\pi_t}(s)\pi_t(a|s)\mu^{\widetilde{P}_t,\pi_t}(s'|s,a) + \gamma_t}\right) \times 32H^2 \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right) \\ &\leq 3\sum_{s'} \frac{1}{\mu^{\pi_t}(s)\pi_t(a|s) + \gamma_t} \times 32H^2 \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right) \\ &\leq \frac{S}{\mu_t(s)\pi_t(a|s)} \times 96H^2 \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right) \end{aligned} \tag{45}$$

and thus

$$\eta_t(s,a)\pi_t(a|s)B_t(s,a) \leq 96H^2 S \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right)^2 \leq \frac{1}{8H} \tag{46}$$

where the last inequality is because $\frac{\eta_t(s',a')}{\mu_t(s')} \leq \frac{1}{60\sqrt{H^3S}}$ in real episodes.

From the second-to-last step in (45), we also have

$$B_t(s,a) \leq \frac{3S}{\gamma_t} \times 32H^2 \max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right) \leq \frac{2\sqrt{HS}}{\gamma_t} \leq 2ST.$$

In virtual episodes,

$$\begin{aligned} b_t(s) &\leq \sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) \\ &\leq \sum_a \frac{\mathbb{I}\{(s_t^\dagger, a_t^\dagger) = (s,a)\}}{24\eta_t(s,a)H\log T} \times \log T \\ &= \sum_a \frac{\mathbb{I}\{(s_t^\dagger, a_t^\dagger) = (s,a)\}}{24\mu_t(s)H} \times \frac{1}{\max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right)} \qquad \text{(by the definition of } (s_t^\dagger, a_t^\dagger)) \end{aligned}$$

$$\leq \frac{\mathbb{I}\{s_t^\dagger = s\}}{24\mu_t(s)H} \times \frac{1}{\max_{s',a'}\left(\frac{\eta_t(s',a')}{\mu_t(s')}\right)}$$
$$\leq \frac{\mathbb{I}\{s_t^\dagger = s\}}{\mu_t(s)} \times \frac{1}{24HM_t}$$

where we define $M_t = \max_{s',a'} \frac{\eta_t(s',a')}{\mu_t(s')}$. Similar to (45):

$$\begin{aligned}
B_t(s,a) &\leq b_t(s) + 3 \sum_{s':h(s')>h(s)} \mu^{\widetilde{P}_t,\pi_t}(s'|s,a) b_t(s') \\
&\leq \left(\frac{\mathbb{I}\{s_t^\dagger = s\}}{\mu_t(s)} + 3\sum_{s'} \mu^{\widetilde{P}_t,\pi_t}(s'|s,a) \frac{\mathbb{I}\{s_t^\dagger = s'\}}{\mu_t(s')} \right) \times \frac{1}{24HM_t} \\
&\leq \left(\frac{\mathbb{I}\{s_t^\dagger = s\}}{\mu_t(s)} + 3 \sum_{s':h(s')>h(s)} \mu^{\widetilde{P}_t,\pi_t}(s'|s,a) \times \frac{\mathbb{I}\{s_t^\dagger = s'\}}{\overline{\mu}_t^{\pi_t}(s)\pi_t(a|s)\mu^{\widetilde{P}_t,\pi_t}(s'|s,a) + \gamma_t} \right) \times \frac{1}{24HM_t} \\
&\leq 3\sum_{s'} \frac{\mathbb{I}\{s_t^\dagger = s'\}}{\mu^{\pi_t}(s)\pi_t(a|s) + \gamma_t} \times \frac{1}{24HM_t} \\
&\leq \frac{1}{\mu_t(s)\pi_t(a|s)} \times \frac{1}{8HM_t} \qquad (47)
\end{aligned}$$

and thus

$$\eta_t(s,a)\pi_t(a|s)B_t(s,a) \leq \frac{\eta_t(s,a)}{\mu_t(s)} \times \frac{1}{8HM_t} \leq \frac{1}{8H}$$

where the last step uses the definition of M_t.

From the second-to-last step in (47) , we also have

$$B_t(s,a) \leq \frac{1}{8\gamma_t HM_t} \leq \frac{15\sqrt{HS}}{\gamma_t} \leq 15ST$$

where we use that $M_t \geq \frac{1}{60\sqrt{H^3S}}$ in vitrual episodes. □

F. Analysis for the Bias (Lemma 6.2)

Proof of Lemma 6.2.

$$\begin{aligned}
&\mathbb{E}\left[\sum_s \mu^\pi(s)\mathbf{bias}^\pi(s)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} \mu^\pi(s)\left(\pi_t(a|s) - \pi(a|s)\right)\left(Q^{\pi_t}(s,a;\ell_t) - \widehat{Q}_t(s,a) + C_t(s,a)\right)\right] \qquad (48) \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} \mu^\pi(s)\left(\pi_t(a|s) - \pi(a|s)\right)\left(Q^{\pi_t}(s,a;\ell_t) - \frac{\mu^{\pi_t}(s)}{\mu_t(s)}Q^{\pi_t}(s,a;\ell_t) + C_t(s,a)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} \mu^\pi(s)\left(\pi_t(a|s) - \pi(a|s)\right)\left(\frac{\mu_t(s) - \mu^{\pi_t}(s)}{\mu_t(s)}Q^{\pi_t}(s,a;\ell_t) + C_t(s,a)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} (\mu^{\pi_t}(s,a) - \mu^\pi(s,a)) z_t(s,a)\right] \qquad (49)
\end{aligned}$$

with $z_t(s,a)$ defined as the following based on Lemma C.2:

$$z_t(s,a) \triangleq \frac{\mu_t(s)-\mu^{\pi_t}(s)}{\mu_t(s)}Q^{\pi_t}(s,a;\ell_t)+C_t(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\frac{\mu_t(s')-\mu^{\pi_t}(s')}{\mu_t(s')}Q^{\pi_t}(s',a';\ell_t)+C_t(s',a')\right]$$

Recall the high probability event $\mathcal{E}$ defined in Definition B.3. Notice that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a))z_t(s,a)\right]\\
&=\Pr(\mathcal{E})\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a))z_t(s,a)\,\middle|\,\mathcal{E}\right]+\Pr(\overline{\mathcal{E}})\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a))z_t(s,a)\,\middle|\,\overline{\mathcal{E}}\right]\\
&\le\Pr(\mathcal{E})\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a))z_t(s,a)\,\middle|\,\mathcal{E}\right]+O(H\delta)\times O(TH\times TH^2)\\
&\qquad\qquad\text{(because } |z_t(s,a)|\le O(TH^2) \text{ almost surely)}\\
&\le\Pr(\mathcal{E})\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a))z_t(s,a)\,\middle|\,\mathcal{E}\right]+O\left(\frac{H^4}{T}\right). \qquad (\delta=\tfrac{1}{T^3})
\end{aligned} \tag{50}$$

From now on, it suffices to bound $\sum_{t=1}^{T}\sum_{s,a}(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a))z_t(s,a)$ assuming $\mathcal{E}$ holds (i.e., $P\in\mathcal{P}_t$ for all t).

By the definition of $C_t(s,a)$, we have

$$\begin{aligned}
z_t(s,a)&=\frac{\mu_t(s)-\mu^{\pi_t}(s)}{\mu_t(s)}Q^{\pi_t}(s,a;\ell_t)+\max_{\widetilde{P}\in\mathcal{P}_t}\mathbb{E}_{s'\sim\widetilde{P}(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\frac{\mu_t(s')-\underline{\mu}_t^{\pi_t}(s')}{\mu_t(s')}H+C_t(s',a')\right]\\
&\qquad-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\frac{\mu_t(s')-\mu^{\pi_t}(s')}{\mu_t(s')}Q^{\pi_t}(s',a';\ell_t)+C_t(s',a')\right]\\
&\ge\frac{\mu_t(s)-\mu^{\pi_t}(s)}{\mu_t(s)}Q^{\pi_t}(s,a;\ell_t)\ge 0.
\end{aligned} \tag{51}$$

On the other hand,

$$\begin{aligned}
z_t(s,a)&\le\frac{\mu_t(s)-\mu^{\pi_t}(s)}{\mu_t(s)}H+C_t(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[C_t(s',a')\right]\\
&\le c_t(s)+\mathbb{E}_{s'\sim\overline{P}_t(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[c_t(s')+C_t(s',a')\right]-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[C_t(s',a')\right]\\
&\qquad\qquad\text{(let } \overline{P}_t \text{ be the transition that attains the maximum in (13))}\\
&\le c_t(s)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[c_t(s')]+\sum_{s',a'}\left|\overline{P}_t(s'|s,a)-P(s'|s,a)\right|\pi_t(a'|s')\left(c_t(s')+C_t(s',a')\right)\\
&\le c_t(s)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[c_t(s')]+\sum_{s',a'}e_t(s'|s,a)\pi_t(a'|s')(c_t(s')+C_t(s',a'))
\end{aligned} \tag{52}$$

where we define $e_t(s'|s,a)=\left|\overline{P}_t(s'|s,a)-P(s'|s,a)\right|$.

Observe that by the definition of $C_t(s,a)$, it holds that

$$C_t(s,a)=\sum_{s':h(s')>h(s)}\mu^{\overline{P}_t,\pi_t}(s'|s,a)c_t(s'),$$

and therefore,

$$c_t(s)+C_t(s,a)=\sum_{s'}\mu^{\overline{P}_t,\pi_t}(s'|s,a)c_t(s')$$

and

$$\sum_a \pi_t(a|s)\left(c_t(s)+C_t(s,a)\right)=\sum_{s'}\mu^{\overline{P}_t,\pi_t}(s'|s)c_t(s').$$

Thus we can thus rewrite (52) as

$$z_t(s,a)\le c_t(s)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[c_t(s')]+\sum_{s'}e_t(s'|s,a)\sum_{s''}\mu^{\overline{P}_t,\pi_t}(s''|s')c_t(s''). \tag{53}$$

Continue from the previous calculation in (50):

$$\begin{aligned}
&\sum_{t=1}^T\sum_{s,a}\left(\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right)z_t(s,a)\\
&\le\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+ z_t(s,a) && \text{(by (51))}\\
&\le\underbrace{\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+ c_t(s)}_{\textbf{term}_1}+\underbrace{\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+ \mathbb{E}_{s'\sim P(\cdot|s,a)}[c_t(s')]}_{\textbf{term}_2}\\
&\quad+\underbrace{\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+\sum_{s'}e_t(s'|s,a)\sum_{s''}\mu^{\overline{P}_t,\pi_t}(s''|s')c_t(s'')}_{\textbf{term}_3}. && \text{(by (53))}
\end{aligned}$$

Known transition case

For the known transition case, we have

$$c_t(s)\le\frac{\mu^{\pi_t}(s)+\gamma_t-\mu^{\pi_t}(s)}{\mu_t(s)}H=\frac{\gamma_t}{\mu_t(s)}H$$

and $e_t(s'|s,a)=0$. Thus,

$$\mathbb{E}\left[\sum_s\mu^{\pi}(s)\textbf{bias}^{\pi}(s)\right]\lesssim\sum_{t=1}^T\sum_s\mu^{\pi_t}(s)\times\frac{\gamma_t}{\mu_t(s)}H\le HS\sum_{t=1}^T\gamma_t=O\left(H^5SA^2\ln(T)\right).$$

Unknown transition case

Upper bounding term$_1$. By the definition of $c_t(s)$,

$$\begin{aligned}
\textbf{term}_1&\le H\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+\left(\frac{\overline{\mu}_t^{\pi_t}(s)-\underline{\mu}_t^{\pi_t}(s)+\gamma_t}{\mu_t(s)}\right)\\
&\le\underbrace{H\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+\left(\frac{\overline{\mu}_t^{\pi_t}(s)-\mu^{\pi_t}(s)}{\mu_t(s)}\right)}_{\textbf{term}_{1a}}\\
&\quad+\underbrace{H\sum_{t=1}^T\sum_{s,a}\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+\left(\frac{\mu^{\pi_t}(s)-\underline{\mu}_t^{\pi_t}(s)}{\mu_t(s)}\right)}_{\textbf{term}_{1b}}+\underbrace{\sum_{t=1}^T\sum_{s,a}\mu^{\pi_t}(s,a)\left(\frac{H\gamma_t}{\mu_t(s)}\right)}_{\textbf{term}_{1c}}.
\end{aligned}$$

To bound $\textbf{term}_{1a}$, we apply Lemma C.4 with

$$g_t(s)=\sum_a\frac{\left[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)\right]_+}{\mu_t(s)}\le 1,$$

which gives

$$\begin{aligned}\textbf{term}_{1a} &\le \sqrt{H^3S^2A\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\frac{[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+}{\mu_t(s)}\ln(T)\iota} + H^2S^4A\ln(T)\iota \\ &\le \sqrt{H^3S^2A\sum_{t=1}^{T}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\ln(T)\iota} + H^2S^4A\ln(T)\iota.\end{aligned}$$

$\textbf{term}_{1b}$ can be bound in the same way and admits the same upper bound. $\textbf{term}_{1c} \le HS\sum_{t=1}^{T}\gamma_t = O\left(H^5SA^2\ln(T)\right)$. Combining $\textbf{term}_1, \textbf{term}_2, \textbf{term}_3$, we get

$$\textbf{term}_1 \lesssim \sqrt{H^3S^2A\sum_{t=1}^{T}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\ln(T)\iota} + H^2S^4A^2\ln(T)\iota.$$

Upper bounding $\textbf{term}_2$. This is very similar to the procedure of bounding $\textbf{term}_1$. We perform a similar decomposition:

$$\begin{aligned}\textbf{term}_2 \le\ & \underbrace{H\sum_{t=1}^{T}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\frac{\overline{\mu}_t^{\pi_t}(s')-\mu^{\pi_t}(s')}{\mu_t(s')}\right]}_{\textbf{term}_{2a}} \\ &+ \underbrace{H\sum_{t=1}^{T}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\frac{\mu^{\pi_t}(s')-\underline{\mu}_t^{\pi_t}(s')}{\mu_t(s')}\right]}_{\textbf{term}_{2b}} + \underbrace{\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\frac{H\gamma_t}{\mu_t(s')}\right]}_{\textbf{term}_{2c}}.\end{aligned}$$

To bound $\textbf{term}_{2a}$, we apply Lemma C.4 with

$$g_t(s') = \sum_{s,a}\frac{[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+}{\mu_t(s')}P(s'|s,a) \le \frac{\sum_{s,a}\mu^{\pi_t}(s,a)P(s'|s,a)}{\mu_t(s')} \le \frac{\mu^{\pi_t}(s')}{\mu_t(s')} \le 1,$$

which gives

$$\begin{aligned}\textbf{term}_{2a} &\le \sqrt{H^3S^2A\sum_{t=1}^{T}\sum_{s'}\mu^{\pi_t}(s')\sum_{s,a}\frac{[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+}{\mu_t(s')}P(s'|s,a)\ln(T)\iota} + H^2S^4A\ln(T)\iota \\ &\le \sqrt{H^3S^2A\sum_{t=1}^{T}\sum_{s'}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+P(s'|s,a)\ln(T)\iota} + H^2S^4A\ln(T)\iota \\ &= \sqrt{H^3S^2A\sum_{t=1}^{T}\sum_{s,a}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\ln(T)\iota} + H^2S^4A\ln(T)\iota.\end{aligned}$$

which is same as the bound for $\textbf{term}_{1a}$. Also, $\textbf{term}_{2b}$ can be handled in the same way as $\textbf{term}_{2a}$, and $\textbf{term}_{2c} \le \sum_{t=1}^{T}\sum_{s'}\mu^{\pi_t}(s')\times\frac{H\gamma_t}{\mu_t(s')} \le HS\sum_{t=1}^{T}\gamma_t$. Overall, $\textbf{term}_2$ can be bounded by the same order as $\textbf{term}_1$.

Upper bounding $\textbf{term}_3$.

$$\begin{aligned}\textbf{term}_3 &\le \sum_{t=1}^{T}\sum_{s,a,s'}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\left(\sqrt{\frac{\overline{P}_t(s'|s,a)\iota}{n_t(s,a)}}+\frac{\iota}{n_t(s,a)}\right)\sum_{s''}\mu^{\overline{P}_t,\pi_t}(s''|s')c_t(s'') \\ &\lesssim \sum_{t=1}^{T}\sum_{s,a,s'}[\mu^{\pi_t}(s,a)-\mu^{\pi}(s,a)]_+\left(\overline{P}_t(s'|s,a)\alpha+\frac{\iota}{n_t(s,a)\alpha}\right)\sum_{s''}\mu^{\overline{P}_t,\pi_t}(s''|s')c_t(s'') \qquad \text{(for any } \alpha\in(0,1]\text{)}\end{aligned}$$

$$
\begin{aligned}
&= \alpha \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \overline{P}_t(s'|s,a) \sum_{s''} \mu^{\overline{P}_t,\pi_t}(s''|s') c_t(s'') \\
&\quad + \frac{1}{\alpha} \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \frac{\iota}{n_t(s,a)} \sum_{s''} \mu^{\overline{P}_t,\pi_t}(s''|s') c_t(s'') \\
&\le \alpha \sum_{t=1}^{T} \sum_{s,a} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \sum_{s'} \mu^{\overline{P}_t,\pi_t}(s'|s,a) c_t(s') + \frac{H^2}{\alpha} \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a))]_+ \frac{\iota}{n_t(s,a)} \\
&= \alpha H \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \mu^{\overline{P}_t,\pi_t}(s'|s,a) \frac{\mu_t(s') - \underline{\mu}_t(s')}{\mu_t(s')} + \frac{H^2 S}{\alpha} \sum_{t=1}^{T} \sum_{s,a} \frac{\mu^{\pi_t}(s,a)\iota}{n_t(s,a)} \\
&\le \underbrace{\alpha H \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \mu^{\overline{P}_t,\pi_t}(s'|s,a) \frac{\overline{\mu}_t^{\pi_t}(s') - \mu^{\pi_t}(s')}{\mu_t(s')}}_{\textbf{term}_{3a}} \\
&\quad + \underbrace{\alpha H \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \mu^{\overline{P}_t,\pi_t}(s'|s,a) \frac{\mu^{\pi_t}(s') - \underline{\mu}_t^{\pi_t}(s')}{\mu_t(s')}}_{\textbf{term}_{3b}} \\
&\quad + \underbrace{\alpha H \sum_{t=1}^{T} \sum_{s,a,s'} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \mu^{\overline{P}_t,\pi_t}(s'|s,a) \frac{\gamma_t}{\mu_t(s')}}_{\textbf{term}_{3c}} + \frac{H^2 S^2 A \ln(T)\iota}{\alpha}
\end{aligned}
$$

(by Lemma B.2 and the assumption that $\mathcal{E}$ holds.)

For $\textbf{term}_{3a}$ we apply Lemma C.4 with

$$
g_t(s') = \frac{\sum_{s,a} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \mu^{\overline{P}_t,\pi_t}(s'|s,a)}{\mu_t(s')} \le \frac{\sum_{s,a} \mu^{\pi_t}(s,a) \mu^{\overline{P}_t,\pi_t}(s'|s,a)}{\mu_t(s')} \le H,
$$

and we get

$$
\begin{aligned}
\textbf{term}_{3a} &\le \alpha H \sqrt{H^2 S^2 A \ln(T)\iota \sum_{t=1}^{T} \sum_{s'} \mu^{\pi_t}(s') \frac{\sum_{s,a} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+ \mu^{\overline{P}_t,\pi_t}(s'|s,a)}{\mu_t(s')}} + \alpha H \cdot H^2 S^4 A \ln(T)\iota \\
&\le \alpha H \sqrt{H^3 S^2 A \ln(T)\iota \sum_{t=1}^{T} \sum_{s,a} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+} + \alpha H^3 S^4 A \ln(T)\iota
\end{aligned}
$$

The same bound applies to $\textbf{term}_{3b}$, too.

$$
\textbf{term}_{3c} \le \alpha H \sum_{t=1}^{T} \sum_{s,a,s'} \mu^{\pi_t}(s,a) \mu^{\overline{P}_t,\pi_t}(s'|s,a) \frac{\gamma_t}{\mu_t(s')} \le \alpha H^2 \sum_{s'} \gamma_t \lesssim \alpha H^6 S A^2.
$$

Picking $\alpha = \frac{1}{H}$, combining $\textbf{term}_{3a}$ and $\textbf{term}_{3b}$, and using $H \le S$, we get

$$
\textbf{term}_3 \le \sqrt{H^3 S^2 A \ln(T)\iota \sum_{t=1}^{T} \sum_{s,a} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)]_+} + H^2 S^4 A^2 \ln(T)\iota
$$

which is also of the same order as $\textbf{term}_1$.

Combining $\textbf{term}_1, \textbf{term}_2, \textbf{term}_3$, we get that if $\mathcal{E}$ holds, then

$$
\sum_{t=1}^{T} \sum_{s,a} (\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)) z_t(s,a) \lesssim \sqrt{H^3 S^2 A \sum_{t=1}^{T} \sum_{s,a} [\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a))]_+ \ln(T)\iota} + H^2 S^4 A^2 \ln(T)\iota
$$

Using this in (50) finishes the proof.

□

G. Bounding $\sum_s V^{\pi_t}(s_0; b_t)$ (Lemma 6.3, Lemma 6.4, Lemma 6.5)

We first show Lemma G.1 and Lemma G.2 which are common among different regularizers.

Lemma G.1.

$$\mathbb{E}\left[\sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu_t(s)\pi_t(a|s)(1-\pi_t(a|s))}\right]$$
$$\lesssim \mathbb{E}\left[\sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(s)(1-\pi_t(a|s))}\right] + \sqrt{H^4S^2A^3\ln(T)} + \mathbb{I}\{\textit{unknown transition}\}\sqrt{HS^5A^3\ln(T)\iota}.$$

Proof. Define $\phi(s,a) = \pi_t(a|s)(1-\pi_t(a|s))$.

$$\sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu_t(s)\phi(s,a)} \le \sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu^{\pi_t}(s)\phi(s,a)} + \underbrace{\sum_{s,a}\sqrt{\sum_{t=1}^{T}\gamma_t\phi_t(s,a)}}_{\textbf{term}_1} + \underbrace{\sum_{s,a}\sqrt{\sum_{t=1}^{T}|\overline{\mu}_t^{\pi_t}(s)-\mu^{\pi_t}(s)|\,\phi_t(s,a)}}_{\textbf{term}_2}$$

$$\textbf{term}_1 \le \sum_{s,a}\sqrt{\sum_{t=1}^{T}\gamma_t\phi_t(s,a)} \le \sqrt{SA\sum_{t=1}^{T}\gamma_t\sum_{s,a}\phi_t(s,a)} \le S\sqrt{A\sum_{t=1}^{T}\gamma_t} \le \sqrt{H^4S^2A^3\ln(T)}.$$

$\textbf{term}_2$ is zero in the known transition case, and in the unknown transition case, if $\mathcal{E}$ defined in Definition B.3 holds, then

$$\begin{aligned}
\textbf{term}_2 &\le \sum_{s,a}\left(\alpha\sum_{t=1}^{T}(\overline{\mu}_t^{\pi_t}(s)-\mu^{\pi_t}(s))\phi_t(s,a) + \frac{1}{\alpha}\right) && \text{(for any } \alpha>0\text{)}\\
&\le \alpha\left(\sqrt{HS^2A\ln(T)\iota\sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s)\left(\sum_{a}\phi_t(s,a)\right)^2} + HS^4A\ln(T)\iota\right) + \frac{SA}{\alpha}\\
&&\text{(by Lemma C.4 with } g_t(s)=\textstyle\sum_a\phi_t(s,a)\text{)}\\
&\le \alpha\left(\sqrt{HS^2A\ln(T)\iota\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\phi_t(s,a)} + HS^4A\ln(T)\iota\right) + \frac{SA}{\alpha}\\
&\lesssim \sqrt{\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\phi_t(s,a)} + \sqrt{HS^5A^3\ln(T)\iota} && \text{(choosing } \alpha=\tfrac{1}{\sqrt{HS^3A\ln(T)\iota}}\text{)}\\
&\le \sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu^{\pi_t}(s)\phi_t(s,a)} + \sqrt{HS^5A^3\ln(T)\iota}.
\end{aligned}$$

If $\mathcal{E}$ does not hold (which happens with probability $\le O(H/T^3)$), then $\textbf{term}_2 \le O(SA\sqrt{T})$. Overall,

$$\mathbb{E}[\textbf{term}_2] \lesssim \mathbb{E}\left[\sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu^{\pi_t}(s)\phi_t(s,a)}\right] + \sqrt{HS^5A^3\ln(T)\iota} + \frac{HSA}{T^{2.5}}.$$

Collecting terms and using $H \le S$ finishes the proof. □

Lemma G.2. *With known transition,*

$$\sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s)\nu_t(s)\lesssim H^4SA^2\ln(T).$$

For Tsallis entropy or Shannon entropy with unknown transition,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s)\right]\le HS^4A^2\ln(T)\iota.$$

Proof. With known transition, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s}\mu^{\pi_t}(s)\nu_t(s)\\
&\le\frac{1}{H^4}\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\pi_t(a|s)C_t(s,a)^2 && (\eta_t(s,a)\le\tfrac{1}{H^4}\text{ or }\eta_t(s)\le\tfrac{1}{H^4})\\
&\le\frac{1}{H^2}\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\pi_t(a|s)C_t(s,a) && (C_t(s,a)\le H^2)\\
&\le\frac{1}{H}\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\pi_t(a|s)\sum_{s'}\mu^{\pi_t}(s'|s,a)\frac{\mu_t(s')-\mu^{\pi_t}(s')}{\mu_t(s')} && (\text{by the definition of } C_t(s,a))\\
&\le\sum_{t=1}^{T}\sum_{s'}\mu^{\pi_t}(s')\times\frac{\mu_t(s')-\mu^{\pi_t}(s')}{\mu_t(s')}\\
&\le\sum_{t=1}^{T}S\gamma_t\\
&\lesssim H^4SA^2\ln(T).
\end{aligned}$$

With unknown transitions, notice that for Tsallis entropy we have $\eta_t(s)\le\min\left\{\frac{1}{H^4},\frac{1}{H\sqrt{t}}\right\}$ and for Shannon entropy we have $\eta_t(s,a)\le\min\left\{\frac{1}{H^4},\frac{1}{H\sqrt{t}}\right\}$. Therefore, in both cases, suppose that $\mathcal{E}$ holds,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s)\\
&\le\sum_{t=1}^{T}\min\left\{\frac{1}{H^4},\frac{1}{H\sqrt{t}}\right\}\sum_{s,a}\mu^{\widetilde{P}_t,\pi_t}(s)\pi_t(a|s)C_t(s,a)^2\\
&\le\sum_{t=1}^{T}\min\left\{\frac{1}{H^2},\frac{H}{\sqrt{t}}\right\}\sum_{s,a}\mu^{\widetilde{P}_t,\pi_t}(s)\pi_t(a|s)C_t(s,a)\\
&\le\sum_{t=1}^{T}\min\left\{\frac{1}{H},\frac{H^2}{\sqrt{t}}\right\}\sum_{s,a}\mu^{\widetilde{P}_t,\pi_t}(s)\pi_t(a|s)\sum_{s'}\mu^{\overline{P}_t,\pi_t}(s'|s,a)\frac{\mu_t(s')-\underline{\mu}_t^{\pi_t}(s')}{\mu_t(s')}\\
&\qquad\qquad(\text{let }\overline{P}_t\text{ be the }\widetilde{P}\text{ attaining maximum in (13)})\\
&\le\sum_{t=1}^{T}\min\left\{1,\frac{H^3}{\sqrt{t}}\right\}\sum_{s'}\overline{\mu}_t^{\pi_t}(s')\times\frac{\mu_t(s')-\underline{\mu}_t^{\pi_t}(s')}{\mu_t(s')}\\
&\le\sum_{t=1}^{T}\min\left\{1,\frac{H^3}{\sqrt{t}}\right\}\sum_{s'}\left(\overline{\mu}_t^{\pi_t}(s')-\underline{\mu}_t^{\pi_t}(s')\right)+\sum_{t=1}^{T}S\gamma_t
\end{aligned}$$

$$\leq \sum_{t=1}^{T} \min\left\{1, \frac{H^3}{\sqrt{t}}\right\} \sum_{s'} \left(\overline{\mu}_t^{\pi_t}(s') - \underline{\mu}_t^{\pi_t}(s')\right) + H^4 S A^2 \ln(T)$$

By Lemma C.4, the first part above can be upper bounded by

$$\sqrt{HS^2A\ln(T)\iota \sum_{t=1}^{T}\sum_{s} \mu^{\pi_t}(s) \times \min\left\{1, \frac{H^6}{t}\right\}} + HS^4A\ln(T)\iota$$
$$\lesssim \sqrt{H^8S^2A\iota}\ln(T) + HS^4A\ln(T)\iota \lesssim HS^4A\ln(T)\iota$$

where we use $H \leq S$.

Suppose that $\mathcal{E}$ does not hold (happens with probability $O(H/T^3)$), we still have

$$\sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) \leq \sum_{t=1}^{T} \min\left\{\frac{1}{H^4}, \frac{1}{H\sqrt{t}}\right\} \sum_{s,a} \mu^{\widetilde{P}_t,\pi_t}(s)\pi_t(a|s)C_t(s,a)^2$$
$$\leq O\left(T \times \frac{1}{H^4} \times H\left(H^2\right)^2\right) \leq O(HT)$$

because $|C_t(s,a)| \leq H^2$ with probability 1.

Combining all terms and taking expectation, we conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s)\right] \lesssim HS^4A^2\ln(T)\iota.$$

□

G.1. Tsallis entropy

Proof of Lemma 6.3.

$$\sum_{t=1}^{T} V^{\pi_t,\widetilde{P}_t}(s_0; b_t)$$
$$= \sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s) b_t(s)$$
$$\leq \sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s) \left[\nu_t(s) + \left(\frac{1}{\eta_t(s)} - \frac{1}{\eta_{t-1}(s)}\right)\left(\xi_t(s) + \sqrt{A}\cdot\mathbb{I}\left[\frac{\eta_t(s)}{\mu_t(s)} > \frac{1}{8H}\right]\right)\right] \quad \text{(by (15))}$$
$$\lesssim \sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) + H\sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s) \frac{\frac{1}{\mu_t(s)}}{\sqrt{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}}\left(\xi_t(s) + \sqrt{A}\cdot\frac{8H\eta_t(s)}{\mu_t(s)}\right) \quad \text{(by (16))}$$
$$\lesssim \underbrace{\sum_{t=1}^{T}\sum_{s} \mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s)}_{\textbf{term}_1} + \underbrace{H\sum_{t=1}^{T}\sum_{s} \frac{\xi_t(s)}{\sqrt{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}}}_{\textbf{term}_2} + \underbrace{H^2\sqrt{A}\sum_{t=1}^{T}\sum_{s} \frac{\frac{1}{\mu_t(s)}\cdot\eta_t(s)}{\sqrt{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}}}_{\textbf{term}_3}$$

Bounding term$_1$. **term$_1$** can be bounded using Lemma G.2, which gives

$$\mathbb{E}[\textbf{term}_1] \lesssim H^4SA^2\ln(T) + \mathbb{I}\{\text{unknown transition}\}HS^4A^2\ln(T)\iota.$$

Bounding $\mathbf{term}_2$.

$$\begin{aligned}
\mathbf{term}_2 &\le H\sum_{t=1}^{T}\sum_{s}\sqrt{\mu_t(s)}\xi_t(s)\sqrt{\frac{\frac{1}{\mu_t(s)}}{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}} \\
&\le H\sum_{t=1}^{T}\sum_{s,a}\sqrt{\mu_t(s)\pi_t(a|s)}(1-\pi_t(a|s))\sqrt{\frac{\frac{1}{\mu_t(s)}}{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}} \\
&\le H\sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu_t(s)\pi_t(s,a)(1-\pi_t(a|s))}\sqrt{\sum_{t=1}^{T}\frac{\frac{1}{\mu_t(s)}}{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}}} \\
&\lesssim H\sqrt{\ln T}\sum_{s,a}\sqrt{\sum_{t=1}^{T}\mu_t(s)\pi_t(a|s)(1-\pi_t(a|s))}.
\end{aligned}$$

By Lemma G.1, we can bound the last expression by

$$\mathbb{E}\left[H\sum_{s,a}\sqrt{\ln(T)\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(s,a)(1-\pi_t(s,a))}\right]+\sqrt{H^6S^2A^3}\ln(T)+\mathbb{I}\{\text{unknown transition}\}\sqrt{H^3S^5A^3}\ln(T)\iota.$$

Bounding $\mathbf{term}_3$. By (16),

$$\mathbf{term}_3 \le H\sqrt{A}\sum_{t=1}^{T}\sum_{s}\frac{\frac{1}{\mu_t(s)}}{\sum_{\tau=1}^{t}\frac{1}{\mu_\tau(s)}} \le HS\sqrt{A}\ln(T).$$

Combining $\mathbf{term}_1, \mathbf{term}_2, \mathbf{term}_3$ finishes the proof.

□

G.2. Shannon entropy

Proof of Lemma 6.4.

$$\begin{aligned}
&\sum_{t=1}^{T}V^{\widetilde{P}_t,\pi_t}(s_0;b_t) \\
&\lesssim H\sqrt{\ln T}\sum_{t,s,a}\mu^{\widetilde{P}_t,\pi_t}(s)\left(\frac{1}{\mu_t(s)\sqrt{\sum_{\tau=1}^{t-1}\frac{\xi_\tau(s,a)}{\mu_\tau(s)}+\frac{1}{\mu_t(s)}}}+\frac{1}{\sqrt{t}}\right)\left(\xi_t(s,a)+1-\frac{\min_{\tau\in[t]}\mu_\tau(s)}{\min_{\tau\in[t-1]}\mu_\tau(s)}\right) \\
&\quad+\sum_{t,s}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) \\
&\le\sum_{t,s,a}\frac{H\sqrt{\ln T}}{\sqrt{\sum_{\tau=1}^{t-1}\frac{\xi_\tau(s,a)}{\mu_\tau(s)}+\frac{1}{\mu_t(s)}}}\xi_t(s,a)+\sum_{t,s,a}\frac{H\sqrt{\ln T}}{\sqrt{\sum_{\tau=1}^{t-1}\frac{\xi_\tau(s,a)}{\mu_\tau(s)}+\frac{1}{\mu_t(s)}}}\left(1-\frac{\min_{\tau\in[t]}\mu_\tau(s)}{\min_{\tau\in[t-1]}\mu_\tau(s)}\right) \\
&\quad+\sum_{t,s,a}\frac{H\sqrt{\ln T}\mu_t(s)}{\sqrt{t}}\xi_t(s,a)+H\sqrt{\ln T}\sqrt{\sum_{t,s,a}\mu^{\widetilde{P}_t,\pi_t}(s)\frac{1}{t}}\sqrt{\sum_{t,s,a}\mu^{\widetilde{P}_t,\pi_t}(s)\left(1-\frac{\min_{\tau\in[t]}\mu_\tau(s)}{\min_{\tau\in[t-1]}\mu_\tau(s)}\right)^2} \\
&\quad+\sum_{t,s}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) \\
&\le H\sqrt{\ln T}\sum_{s,a}\sqrt{\sum_{t}\frac{\frac{\xi_t(s,a)}{\mu_t(s)}}{\sum_{\tau=1}^{t-1}\frac{\xi_\tau(s,a)}{\mu_\tau(s)}+\frac{1}{\mu_t(s)}}}\sqrt{\sum_{t}\mu_t(s)\xi_t(s,a)}+H\sqrt{\ln T}\sum_{t,s,a}\ln\left(\frac{\min_{\tau\in[t-1]}\mu_\tau(s)}{\min_{\tau\in[t]}\mu_\tau(s)}\right)
\end{aligned}$$

$$
\begin{aligned}
&+ H\sqrt{\ln T}\sum_{s,a}\sqrt{\sum_t \frac{\mu_t(s)\xi_t(s,a)}{t}}\sqrt{\sum_t \mu_t(s)\xi_t(s,a)} \\
&+ H\sqrt{\ln T}\sqrt{HA\ln(T)}\sqrt{A\sum_{t,s}\ln\left(\frac{\min_{\tau\in[t-1]}\mu_\tau(s)}{\min_{\tau\in[t]}\mu_\tau(s)}\right)} + \sum_{t,s}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) \\
&\lesssim H\sqrt{\ln T}\sum_{s,a}\sqrt{\ln(T)\sum_t \mu_t(s)\xi_t(s,a)} + \sum_{t,s}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) + H^2SA\ln^{\frac{3}{2}}(T) \\
&\lesssim H\sum_{s,a}\sqrt{\ln^3(T)\sum_{t=1}^T \mu_t(s)\pi_t(a|s)(1-\pi_t(a|s))} + \sum_{s,t}\mu^{\widetilde{P}_t,\pi_t}(s)\nu_t(s) + H^2SA\ln^{\frac{3}{2}}(T)
\end{aligned}
$$

By Lemma G.1 and Lemma G.2, the expectation of this can be upper bounded by

$$
\begin{aligned}
&\mathbb{E}\left[H\sum_{s,a}\sqrt{\ln^3(T)\sum_{t=1}^T \mu^{\pi_t}(s)\pi_t(a|s)(1-\pi_t(a|s))}\right] \\
&\quad + \sqrt{H^6S^2A^3\ln^3(T)} + \mathbb{I}\{\text{unknown transition}\}\sqrt{H^3S^5A^3\ln^3(T)\iota} \\
&\quad + H^4SA^2\ln^{\frac{3}{2}}(T) + \mathbb{I}\{\text{unknown transition}\}HS^4A^2\ln(T)\iota \\
&\lesssim \mathbb{E}\left[H\sum_{s,a}\sqrt{\ln^3(T)\sum_{t=1}^T \mu^{\pi_t}(s)\pi_t(a|s)(1-\pi_t(a|s))}\right] \\
&\quad + H^4SA^2\sqrt{\ln^3(T)} + \mathbb{I}\{\text{unknown transition}\}HS^4A^2\ln(T)\iota. \qquad \text{(using } H\le S \text{ and } \log(T)\lesssim\iota\text{)}
\end{aligned}
$$

□

G.3. Log barrier

Lemma G.3. *Let $\eta_1 > 0, \eta_2, \eta_3, \ldots$ be updated by*

$$\frac{1}{\eta_{t+1}} = \frac{1}{\eta_t} + \eta_t\phi_t \qquad \forall t \ge 1$$

with $0 \le \phi_t \le \eta_t^{-2}$. Then

$$\frac{1}{\eta_{t+1}} \ge \frac{1}{2}\sqrt{\sum_{\tau=1}^{t+1}\phi_\tau}.$$

Proof. By the update rule,

$$\frac{1}{\eta_{t+1}^2} - \frac{1}{\eta_t^2} = \left(\frac{1}{\eta_{t+1}} + \frac{1}{\eta_t}\right)\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) = \left(\frac{1}{\eta_{t+1}} + \frac{1}{\eta_t}\right)\eta_t\phi_t \ge \phi_t,$$

which implies

$$\frac{1}{\eta_{t+1}} \ge \sqrt{\frac{1}{\eta_1^2} + \sum_{\tau=1}^t \phi_\tau} \ge \sqrt{\sum_{\tau=1}^t \phi_\tau}.$$

By the condition on ϕ_t, we also have

$$\frac{1}{\eta_{t+1}} \ge \sqrt{\phi_{t+1}}.$$

Combining the two inequalities finishes the proof. □

Proof of Lemma 6.5. In this proof we only focus on the know transition case. We use $\mathcal{T}_r$ and $\mathcal{T}_v$ to denote the set of real and virtual episodes, respectively.

Let $\phi_t(s,a) = \frac{4\zeta_t(s,a)}{\mu_t(s)^2 \log(T)}$ in real episodes and $\phi_t(s,a) = \frac{\mathbb{I}\{(s_t^\dagger,a_t^\dagger)=(s,a)\}}{24\eta_t(s,a)^2 H \log T}$ in virtual episodes. We first show that $\phi_t(s,a) \le \frac{1}{\eta_t(s,a)^2}$, which allows us to apply Lemma G.3 because $\frac{1}{\eta_{t+1}(s,a)} = \frac{1}{\eta_t(s,a)} + \eta_t(s,a)\phi_t(s,a)$ by our update rule. This is clear for virtual episodes. For real episodes,

$$\phi_t(s,a)\eta_t(s,a)^2 = \frac{4\eta_t(s,a)^2\zeta_t(s,a)}{\mu_t(s)^2 \log T} \le \frac{H^2}{\log T} \times \frac{1}{H^3 S} \le 1$$

because $\frac{\eta_t(s,a)}{\mu_t(s)} \le \frac{1}{60\sqrt{H^3 S}}$ in real episodes.

$$\begin{aligned}
&\sum_{t=1}^T V^{\pi_t}(s_0; b_t) \\
&\lesssim \sum_{t\in\mathcal{T}_r}\sum_s \mu^{\pi_t}(s)\sum_a \left(\frac{\eta_t(s,a)\zeta_t(s,a)}{\mu_t(s)^2\log(T)}\log(T)\right) + \sum_{t\in\mathcal{T}_v}\mu^{\pi_t}(s_t^\dagger)\frac{1}{\eta_t(s_t^\dagger,a_t^\dagger)H\log T}\log T + \sum_{t=1}^T\sum_s \mu^{\pi_t}(s)\nu_t(s) \\
&\lesssim \sum_{t\in\mathcal{T}_r}\sum_{s,a}\eta_t(s,a)\frac{\zeta_t(s,a)}{\mu_t(s)} + \frac{\sqrt{H^3S}}{H}|\mathcal{T}_v| + H^4SA\ln(T) \\
&\qquad\qquad \text{(in virtual episodes, } \tfrac{\eta_t(s_t^\dagger,a_t^\dagger)}{\mu_t(s_t^\dagger)} \ge \tfrac{1}{\sqrt{H^3S}}\text{, and we use Lemma G.2 to bound the last term)} \\
&\le \sqrt{\log(T)}\sum_{t\in\mathcal{T}_r}\sum_{s,a}\frac{\frac{\zeta_t(s,a)}{\mu_t(s)}}{\sqrt{\sum_{\tau\le t:\tau\in\mathcal{T}_r}\frac{\zeta_\tau(s,a)}{\mu_\tau(s)^2}}} + \sqrt{HS}|\mathcal{T}_v| + +H^4SA^2\ln(T) \\
&\qquad\qquad \text{(by Lemma G.3 and the condition verified at the beginning of the proof)} \\
&\le \sqrt{\log T}\sum_{s,a}\sqrt{\sum_{t\in\mathcal{T}_r}\frac{\frac{\zeta_t(s,a)}{\mu_t(s)^2}}{\sum_{\tau\le t:\tau\in\mathcal{T}_r}\frac{\zeta_\tau(s,a)}{\mu_\tau(s)^2}}}\sqrt{\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)} + \sqrt{HS}|\mathcal{T}_v| + +H^4SA^2\ln(T) \\
&\le \log(T)\sum_{s,a}\sqrt{\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)} + \sqrt{HS}|\mathcal{T}_v| + H^4SA^2\ln(T).
\end{aligned}$$

Now we bound the number of virtual episodes. Notice that each time a virtual episode happens, there exist s,a such that $\frac{\eta_t(s,a)}{\mu_t(s)} \ge \frac{1}{60\sqrt{H^3S}}$, and $\eta_t(s,a)$ will shrink by a factor of $(1+\frac{1}{24H\log T})$ after the virtual episode. Since $\mu_t(s) \ge \gamma_t$, this event cannot happen if $\eta_t(s,a) \le \frac{\gamma_t}{60\sqrt{H^3S}}$. Thus, the number of virtual episodes is upper bounded by

$$|\mathcal{T}_v| \lesssim SA \times \frac{\log\frac{60\sqrt{H^3S}}{\gamma_t}}{\log\left(1+\frac{1}{24H\log T}\right)} \lesssim HSA\ln(T)\ln(SAT).$$

Applying this bound in the last expression and using $H \le S$ finishes the proof. □

H. Final Regret Bounds through Self-Bounding (Theorem 4.1, Theorem 4.2, Theorem 4.3)

Proof of Theorem 4.1. Let $\mathring{\pi} = \operatorname{argmax}_\pi \mathrm{Reg}(\pi)$. By (31), Lemma 6.2, and Lemma 6.3, under known transition and Tsallis entropy, we have

$$\mathrm{Reg}(\mathring{\pi}) \lesssim H\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^T \mu^{\pi_t}(s)\pi_t(a|s)(1-\pi_t(a|s))\right]\ln(T)} + H^5SA^2\ln(T)$$

For the adversarial regime, we bound the above by

$$H\sqrt{SA\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\mu^{\pi_t}(s)\pi_t(a|s))\right]\ln(T)} + H^5SA\ln(T) = \sqrt{H^3SAT} + H^5SA^2\ln(T).$$

For the stochastic regime, notice that $\mathrm{Reg}(\mathring{\pi}) \geq \mathrm{Reg}(\pi^\star) \geq \mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(a|s)\Delta(s,a)\right] - \mathcal{C}$, and we have

$$\begin{aligned}
\mathrm{Reg}(\mathring{\pi}) &\leq c_1 H\sum_s\sum_{a\neq\pi^\star(s)}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(a|s)\right]\ln(T)} + c_2H^5SA^2\ln(T) \qquad \text{(for some universal constants } c_1, c_2\text{)} \\
&\leq H\sum_s\sum_{a\neq\pi^\star(s)}\left(\frac{\alpha}{H}\mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(a|s)\Delta(s,a)\right] + \frac{c_1^2H\ln(T)}{\alpha\Delta(s,a)}\right) + c_2H^5SA^2\ln(T) \quad \text{(for arbitrary } \alpha>0\text{)} \\
&\leq \alpha\mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(a|s)\Delta(s,a)\right] + O\left(\sum_s\sum_{a\neq\pi^\star(s)}\frac{H^2\ln(T)}{\alpha\Delta(s,a)} + H^5SA^2\ln(T)\right) \\
&\leq \alpha(\mathrm{Reg}(\mathring{\pi}) + \mathcal{C}) + O\left(\sum_s\sum_{a\neq\pi^\star(s)}\frac{H^2\ln(T)}{\alpha\Delta(s,a)} + H^5SA^2\ln(T)\right)
\end{aligned}$$

Picking $\alpha = \min\left\{\frac{1}{2}, \mathcal{C}^{-\frac{1}{2}}\left(\frac{H^2\ln(T)}{\Delta(s,a)}\right)^{\frac{1}{2}}\right\}$ leads to the bound

$$\mathrm{Reg}(\mathring{\pi}) \lesssim U + \sqrt{U\mathcal{C}} + H^5SA^2\ln(T)$$

where $U = \sum_s\sum_{a\neq\pi^\star(s)}\frac{H^2\ln(T)}{\Delta(s,a)}$. Finally, using that $\mathrm{Reg}(\pi) \leq \mathrm{Reg}(\mathring{\pi})$ for all π finishes the proof. □

Proof of Theorem 4.2. By (31), Lemma 6.2, and Lemma 6.3, under unknown transition and Tsallis entropy, we have

$$\begin{aligned}
\mathrm{Reg}(\pi) \leq\ & \underbrace{c_1\sqrt{H^3S^2A\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\left[\mu^{\pi_t}(s,a) - \mu^{\pi}(s,a)\right]_+\right]\ln(T)\iota}}_{\textbf{term}_1} \\
&+ \underbrace{c_2H\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(a|s)(1-\pi_t(a|s))\right]\ln(T)\iota}}_{\textbf{term}_2} + c_3H^2S^4A^2\ln(T)\iota \\
&\qquad\qquad \text{(for universal constants } c_1, c_2, c_3\text{)}
\end{aligned}$$

In the adversarial regime, we can bound it by the order of

$$\sqrt{H^4S^2AT\ln(T)\iota} + H^2S^4A^2\ln(T)\iota$$

To get a bound in the stochastic regime, we first argue that it suffices to show the desired bound for all π that satisfies $\mathrm{Reg}(\pi) \geq \mathrm{Reg}(\pi^\star)$. This is because we can then bound $\mathrm{Reg}(\pi)$ for π such that $\mathrm{Reg}(\pi) < \mathrm{Reg}(\pi^\star)$ by

$$\mathrm{Reg}(\pi) < \mathrm{Reg}(\pi^\star) \lesssim U + \sqrt{U(\mathcal{C} + \mathcal{C}(\pi^\star))} + \mathit{poly}(H,S,A)\ln(T)\iota = U + \sqrt{U\mathcal{C}} + \mathit{poly}(H,S,A)\ln(T)\iota$$

because $\mathcal{C}(\pi^\star) = 0$ by definition.

Below we assume that $\mathrm{Reg}(\pi) \geq \mathrm{Reg}(\pi^\star)$. Note that by Lemma C.5, for any π,

$$\sum_{s,a}\left|\mu^{\pi}(s,a) - \mu^{\pi^\star}(s,a)\right| \leq H\sum_{s,a}\mu^{\pi}(s)\left|\pi(a|s) - \pi^\star(a|s)\right|$$

$$= H \sum_s \sum_{a \neq \pi^\star(s)} \mu^\pi(s)\pi(a|s) + H \sum_s \mu^\pi(s)(1 - \pi(\,\pi^\star(s)\,|s))$$
$$= 2H \sum_s \sum_{a \neq \pi^\star(s)} \mu^\pi(s)\pi(a|s).$$

Hence,

$$\begin{aligned}
\textbf{term}_1 &\leq c_1 \sqrt{H^3 S^2 A \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} |\mu^{\pi_t}(s,a) - \mu^\pi(s,a)|\right] \ln(T)\iota} \\
&\leq c_1 \sqrt{H^3 S^2 A \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} |\mu^{\pi_t}(s,a) - \mu^{\pi^\star}(s,a)|\right] \ln(T)\iota} + c_1 \sqrt{H^3 S^2 A \sum_{t=1}^T \sum_{s,a} |\mu^{\pi}(s,a) - \mu^{\pi^\star}(s,a)| \ln(T)\iota} \\
&\leq c_1 \sqrt{2H^4 S^2 A \mathbb{E}\left[\sum_{t=1}^T \sum_s \sum_{a \neq \pi^\star(s)} \mu^{\pi_t}(s,a)\right] \ln(T)\iota} + c_1 \sqrt{2H^4 S^2 A \sum_{t=1}^T \sum_s \sum_{a \neq \pi^\star(s)} \mu^{\pi}(s,a) \ln(T)\iota\iota} \\
&\leq \alpha \mathbb{E}\left[\sum_{t=1}^T \sum_s \sum_{a \neq \pi^\star(s)} \mu^{\pi_t}(s,a)\Delta_{\min}\right] + \alpha \sum_{t=1}^T \sum_s \sum_{a \neq \pi^\star(s)} \mu^\pi(s,a)\Delta_{\min} + O\left(\frac{H^4 S^2 A \ln(T)\iota}{\alpha \Delta_{\min}}\right) \\
&\qquad \text{(by AM-GM)} \\
&\leq \alpha(\text{Reg}(\pi^\star) + \mathcal{C}) + \alpha(\text{Reg}(\pi^\star) - \text{Reg}(\pi) + \mathcal{C}(\pi)) + O\left(\frac{H^4 S^2 A \ln(T)\iota}{\alpha \Delta_{\min}}\right) \qquad \text{(see explanation below)} \\
&\leq \alpha \text{Reg}(\pi) + \alpha(\mathcal{C} + \mathcal{C}(\pi)) + O\left(\frac{H^4 S^2 A \ln(T)\iota}{\alpha \Delta_{\min}}\right) \qquad \text{(by the assumption } \text{Reg}(\pi^\star) \leq \text{Reg}(\pi))
\end{aligned}$$

where in the second-to-last inequality we use the property:

$$\begin{aligned}
\text{Reg}(\pi^\star) - \text{Reg}(\pi) &= \mathbb{E}\left[\sum_{t=1}^T V^\pi(s_0; \ell_t) - V^{\pi^\star}(s_0; \ell_t)\right] \\
&= \sum_{t=1}^T \sum_s \sum_{a \neq \pi^\star(s)} \mu^\pi(s,a)\Delta(s,a) - \sum_{t=1}^T \lambda_t(\pi) \\
&\geq \sum_{t=1}^T \sum_s \sum_{a \neq \pi^\star(s)} \mu^\pi(s,a)\Delta_{\min} - \mathcal{C}(\pi)
\end{aligned}$$

For $\textbf{term}_2$, similar to before,

$$\begin{aligned}
\textbf{term}_2 &\leq c_2 H \sum_{s,a} \sqrt{\mathbb{E}\left[\sum_{t=1}^T \mu^{\pi_t}(s)\pi_t(a|s)(1 - \pi_t(a|s))\right] \ln(T)} \\
&\leq \alpha(\text{Reg}(\pi) + \mathcal{C}) + O\left(\sum_s \sum_{a \neq \pi^\star(s)} \frac{H^2 \ln(T)}{\alpha \Delta(s,a)} + H^5 S A^2 \ln(T)\right) \\
&\leq \alpha(\text{Reg}(\pi) + \mathcal{C}) + O\left(\sum_s \sum_{a \neq \pi^\star(s)} \frac{H^2 \ln(T)}{\alpha \Delta(s,a)} + H^5 S A^2 \ln(T)\right) \\
&\leq \alpha(\text{Reg}(\pi) + \mathcal{C}) + O\left(\frac{H^4 S^2 A \ln(T)}{\alpha \Delta_{\min}} + H^2 S^4 A^2 \ln(T)\right)
\end{aligned}$$

Combining $\textbf{term}_1$ and $\textbf{term}_2$, we get

$$\text{Reg}(\pi) \leq 2\alpha \text{Reg}(\pi) + 2\alpha(\mathcal{C} + \mathcal{C}(\pi)) + O\left(\frac{H^4 S^2 A \ln(T)\iota}{\alpha \Delta_{\min}} + H^2 S^4 A^2 \ln(T)\iota\right)$$

Picking $\alpha = \min\left\{\frac{1}{4}, (\mathcal{C}+\mathcal{C}(\pi))^{-\frac{1}{2}}\left(\frac{H^4S^2A\ln(T)\iota}{\Delta_{\min}}\right)^{\frac{1}{2}}\right\}$ leads to the desired bound. □

Proof of Theorem 4.3.

$$\mathrm{Reg}(\pi) \lesssim \sum_{s,a}\sqrt{\ln^2(T)\mathbb{E}\left[\sum_{t=1}^{T}(\mathbb{I}_t(s,a)-\pi_t(a|s)\mathbb{I}_t(s))^2L_{t,h(s)}^2\right]} + H^3S^2A^2\ln(T)\ln(SAT)$$

In the adversarial regime,

$$\begin{aligned}\mathrm{Reg}(\pi) &\le \sqrt{HSA\ln^2(T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(\mathbb{I}_t(s,a)-\pi_t(a|s)\mathbb{I}_t(s))^2L_{t,h(s)}\right]} + H^3S^2A^2\ln(T)\ln(SAT)\\ &\le \sqrt{HSA\ln^2(T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\mathbb{I}_t(s,a)L_{t,h(s)}\right]} + H^3S^2A^2\ln(T)\ln(SAT)\\ &\le \sqrt{H^2SA\ln^2(T)\mathbb{E}\left[\sum_{t=1}^{T}V^{\pi_t}(s_0;\ell_t)\right]} + H^3S^2A^2\ln(T)\ln(SAT)\end{aligned}$$

On the other hand, $\mathrm{Reg}(\pi) = \mathbb{E}\left[\sum_{t=1}^{T}V^{\pi_t}(s_0;\ell_t) - \sum_{t=1}^{T}V^{\pi}(s_0;\ell_t)\right]$. Solving the inequality, we get

$$\mathrm{Reg}(\pi) \lesssim \sqrt{H^2SA\ln^2(T)\sum_{t=1}^{T}V^{\pi}(s_0;\ell_t)} + H^3S^2A^2\ln(T)\ln(SAT).$$

In the stochastic regime,

$$\begin{aligned}\mathrm{Reg}(\pi) &\lesssim \sum_{s,a}\sqrt{\ln^2(T)\mathbb{E}\left[\sum_{t=1}^{T}(\mathbb{I}_t(s,a)-\pi_t(a|s)\mathbb{I}_t(s))^2L_{t,h(s)}^2\right]} + H^3S^2A^2\ln(T)\ln(SAT)\\ &\le \sum_{s,a}\sqrt{H^2\ln^2(T)\mathbb{E}\left[\sum_{t=1}^{T}\mu^{\pi_t}(s)\pi_t(a|s)(1-\pi_t(a|s))\right]} + H^3S^2A^2\ln(T)\ln(SAT),\end{aligned}$$

which is similar to the stochastic bound in Theorem 4.1. Following the same self-bounding analysis in the proof of Theorem 4.1 we can get the desired bound. □

To get regret bounds for the Shannon entropy version under known and unknown transitions, we use Lemma 6.2 and Lemma 6.4 and follow exactly the same procedure as in the proofs of Theorem 4.1 and Theorem 4.2. This leads to the following guarantees:

Theorem H.1. *Under known transitions, Algorithm 1 with Shannon entropy regularizer ensures for any π*

$$\mathrm{Reg}(\pi) \lesssim \sqrt{H^3SAT\ln^3(T)} + \mathit{poly}(H,S,A)\ln^2(T)$$

in the adversarial case, and

$$\mathrm{Reg}(\pi) \lesssim U + \sqrt{U\mathcal{C}} + \mathit{poly}(H,S,A)\ln^2(T)$$

in the stochastic case, where $U = \sum_s\sum_{a\neq\pi^\star(s)}\frac{H^2\ln^3(T)}{\Delta(s,a)}$.

Theorem H.2. *Under unknown transitions, Algorithm 1 with Shannon entropy regularizer ensures for any* π

$$\mathrm{Reg}(\pi) \lesssim \sqrt{H^4 S^2 A T \ln^2(T)\iota} + \mathit{poly}(H, S, A)\ln(T)\iota$$

in the adversarial case, and

$$\mathrm{Reg}(\pi) \lesssim U + \sqrt{U(\mathcal{C} + \mathcal{C}(\pi))} + \mathit{poly}(H, S, A)\ln(T)\iota$$

in the stochastic case, where $U = \frac{H^4 S^2 A \ln^2(T)\iota}{\Delta_{\min}}$.