Policy Optimization Achieves Data-Dependent Regret Bounds in MDPs with Unknown Transitions

Mingyi Li
Department of Mathematical Informatics, The University of Tokyo
mingyi-mike@g.ecc.u-tokyo.ac.jp

Taira Tsuchiya
Department of Mathematical Informatics, The University of Tokyo
Center for Advanced Intelligence Project, RIKEN
tsuchiya@mist.i.u-tokyo.ac.jp

Kenji Yamanishi
Department of Mathematical Informatics, The University of Tokyo
yamanishi@g.ecc.u-tokyo.ac.jp

Abstract

We study policy optimization for online episodic tabular Markov decision processes with unknown transition kernels, aiming for best-of-both-worlds guarantees together with data-dependent regret bounds. Recent work (Dann et al., 2023; Li et al., 2026) has shown that policy optimization can adapt to both adversarial and stochastic losses with first-order, second-order, and path-length bounds, but only under known transitions, leaving open whether such data-dependent guarantees are achievable by policy optimization when the transition kernel is unknown. We resolve this by developing a new algorithm based on optimistic follow-the-regularized-leader that attains these guarantees under unknown transitions. The key ingredient is a new design of optimistic Q-function estimators together with a data-dependent transition bonus that controls estimator bias through the loss-prediction error. Our analysis further identifies an unavoidable transition-dependent complexity term that captures the intrinsic cost of estimating the transition kernel. As a result, we obtain first-order, second-order, and path-length bounds with the transition-dependent complexity term while simultaneously achieving gap-dependent $\mathrm{polylog}(T)$ regret in the stochastic regime.

1 Introduction

We study policy optimization for online finite-horizon episodic tabular Markov decision processes (MDPs) with unknown transitions. In this setting, the learner interacts with an MDP over T episodes. In each episode, the environment selects a loss function while the learner selects a policy. The learner then executes it and receives feedback on the loss function. The goal is to minimize the regret, the gap between the cumulative loss of the learner and that of the best fixed policy in hindsight. Policy optimization directly updates the policy and is a central paradigm in reinforcement learning (e.g., Schulman et al. 2015, 2017). For tabular MDPs, this approach updates a separate action distribution for each state (Shani et al., 2020; Luo et al., 2021), which is often more computationally efficient and practical than global optimization over the occupancy-measure space.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of regret upper bounds based on policy optimization. Here, $\Lambda(\pi) = \min\{L(\pi), HT - L(\pi), Q_\infty, V_1\}$ is the loss-dependent complexity, $Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)$ is the transition-dependent complexity satisfying $\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)] \le H(L(\pi) + \mathrm{Reg}_T(\pi))$ for any policy π, and FO indicates that only first-order adaptivity is achieved. For the stochastic regime with adversarial corruption, we report only the leading term U in bounds of the form $\mathcal{O}(U + \sqrt{UC})$.

Reference	Adversarial regime	Stochastic regime with adversarial corruption	Unknown transition	Data-dependent	Bandit feedback
Luo et al. (2021)	$\widetilde{\mathcal{O}}(\sqrt{H^4S^2AT})$	–	✓		✓
Dann et al. (2023, Theorem 4.2)	$\widetilde{\mathcal{O}}(\sqrt{H^4S^2AT})$	$\frac{H^4S^2A\log^2(T)}{\Delta_{\min}}$	✓		✓
Dann et al. (2023, Theorem 4.3)	$\widetilde{\mathcal{O}}(\sqrt{H^2SAL(\pi)})$	$\sum_s \sum_{a\neq\pi^\star(s)} \frac{H^2\log^2(T)}{\Delta(s,a)}$		FO	✓
Li et al. (2026)	$\widetilde{\mathcal{O}}(\sqrt{H^2SA\Lambda(\pi)})$	$\sum_s \sum_{a\neq\pi^\star(s)} \frac{H^2\log^2(T)}{\Delta(s,a)}$		✓	✓
This work (Theorem 4.1)	$\widetilde{\mathcal{O}}(\sqrt{H^2SA\Lambda(\pi)} + \sqrt{S^2A\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]})$	$\frac{H^2S^2A\log^2(T)}{\Delta_{\min}}$	✓	✓	
This work (Theorem 4.2)	$\widetilde{\mathcal{O}}(\sqrt{H^3S^2A\Lambda(\pi)} + \sqrt{S^2A\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]})$	$\frac{H^4S^2A\log^2(T)}{\Delta_{\min}}$	✓	✓	✓

The difficulty of tabular MDPs depends on how the loss sequence is generated. In the adversarial regime, the worst-case regret scales as $\widetilde{\mathcal{O}}(\sqrt{T})$ (Jin et al., 2020; Luo et al., 2021). In contrast, in the stochastic regime, one can often achieve much faster gap-dependent $\mathrm{polylog}(T)$ regret (Simchowitz and Jamieson, 2019; Chen et al., 2025). This contrast has motivated best-of-both-worlds algorithms, which aim to achieve near-optimal guarantees in both the adversarial and stochastic regimes with a single algorithm (Jin and Luo, 2020; Jin et al., 2021; Dann et al., 2023).

Beyond adapting to the underlying regime, another natural goal is to obtain regret bounds that scale with the difficulty of the realized loss sequence in the adversarial regime. Algorithms only with worst-case regret guarantees ignore the structure of losses and are often overly pessimistic in practical environments. Motivated by this, data-dependent bounds, such as first-order, second-order, and path-length bounds, have been actively studied in online learning, including prediction with expert advice and multi-armed bandits (Cesa-Bianchi et al., 1996; Allenberg et al., 2006; Neu, 2015; Wei and Luo, 2018; Bubeck et al., 2019). These depend, respectively, on the comparator's cumulative loss, the magnitude of the loss fluctuations, and the temporal variation of the losses.

Such data-dependent bounds have also been studied in episodic MDPs, but existing data-dependent guarantees in episodic MDPs are limited either in the algorithmic approach or in the transition model. Under unknown transitions, Lee et al. (2020) obtained a first-order bound, but only through occupancy-measure-based global optimization rather than policy optimization. For policy optimization, such data-dependent bounds have so far been obtained only under *known transitions*: Dann et al. (2023) established best-of-both-worlds guarantees with a first-order bound, and Li et al. (2026) provided algorithms achieving first-order, second-order, and path-length bounds simultaneously. Consequently, whether policy optimization can attain any data-dependent bound under unknown transitions, let alone combine first-order, second-order, and path-length adaptivity with best-of-both-worlds guarantees, was left open by Dann et al. (2023); Li et al. (2026). This raises the following question.

Can policy optimization achieve data-dependent bounds such as first-order, second-order, and path-length bounds in episodic MDPs with unknown transitions?

Contributions of this paper. We answer this question affirmatively by developing a policy optimization algorithm for episodic tabular MDPs with unknown transitions under *bandit feedback*, the harder setting in which the learner must estimate both the losses and the transition kernel from the realized trajectories. In the adversarial regime, its regret adapts to $\min\{L(\pi), HT - L(\pi), Q_\infty, V_1\}$, covering first-order, second-order, and path-length complexities, together with a transition-dependent complexity term $Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)$ in (2) that captures the intrinsic cost of estimating the transition kernel and can be connected to the first-order complexity $L(\pi)$ (see Appendix B.2 for details). The same algorithm also enjoys gap-dependent $\mathrm{polylog}(T)$ regret in the stochastic regime, achieving the best-of-both-worlds guarantee (see Theorem 4.2). To our knowledge, this is the first data-dependent guarantee for policy optimization under unknown transitions. It is worth noting that this is also the first algorithm for tabular MDPs achieving best-of-both-worlds and data-dependent bounds simultaneously under unknown transitions. A detailed comparison with prior policy-optimization results is given in Table 1.

As a by-product, our algorithm carries over to the easier full-information setting, where it retains these data-dependent and best-of-both-worlds guarantees and, in the worst case, recovers the $\widetilde{\mathcal{O}}(\sqrt{H^2S^2AT})$ rate of occupancy-measure-based methods (Rosenberg and Mansour, 2019; Jin et al., 2021) (see Theorem 4.1). This implies that the extra H factor often attributed to Q-function estimation is not inherent for policy optimization in the full-information setting under unknown transitions.

Technically, our algorithms build on the optimistic follow-the-regularized-leader framework, with new prediction terms and optimistic Q-function estimators designed for unknown transitions tailored for each feedback setting (see Section 3). For the full-information setting, the Q-function estimator uses the most optimistic transition kernel in the confidence set. For the bandit setting, we incorporate a new data-dependent transition bonus defined in (12) and (13) into the Q-function estimator. This bonus keeps the Q-function estimator optimistic while reducing the estimator bias to the loss-prediction error, with the remaining worst-case terms contributing only lower-order terms.

Our analysis identifies the new transition-dependent complexity term $Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)$, which captures the cost of estimating the unknown transition kernel. We show that this term is unavoidable under unknown transitions: even for a time-invariant loss sequence with $Q_\infty = V_1 = 0$, any algorithm still incurs regret from estimating the unknown transition kernel (see Proposition 4.3). This explains why data-dependent regret bounds under unknown transitions cannot be characterized solely by the loss-complexity measures used in the known-transition setting of Li et al. (2026). Due to space constraints, we defer a more detailed discussion of related work to Appendix B.1.

2 Preliminaries

Episodic tabular MDPs. We study an episodic tabular Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, H, s_0)$, where $\mathcal{S}$ is a finite state space with $S = |\mathcal{S}|$, $\mathcal{A}$ is a finite action space with $A = |\mathcal{A}|$, H is the horizon length, and s_0 is the initial state. The transition kernel P is fixed but unknown to the learner. For each state-action pair (s, a), $P(s' \mid s, a)$ denotes the probability of moving to state s' after taking action a in state s. We adopt the standard layered MDP assumption (Neu et al., 2010; Jin et al., 2020; Luo et al., 2021), where the state space is partitioned into $H + 1$ disjoint layers $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_H$. Here, $\mathcal{S}_0 = \{s_0\}$ is the initial layer and $\mathcal{S}_H = \{s_H\}$ is a terminal absorbing layer. For simplicity, we exclude s_H from $\mathcal{S}$, so that $H \le S$. Transitions occur only between consecutive layers. Specifically, for any $(s, a) \in \mathcal{S}_h \times \mathcal{A}$ with $h \in \{0, \dots, H-1\}$, the distribution $P(\cdot \mid s, a)$ is supported on $\mathcal{S}_{h+1}$. We write $h(s)$ for the layer index of state s.

The learner interacts with the MDP for T episodes, indexed by $t = 1, \dots, T$. At the beginning of episode t, the environment selects a loss function $\ell_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$. The learner then selects a policy π_t based on past observations, executes it from s_0 under P, and incurs the losses along the realized trajectory. Here, a stochastic policy maps each state s to a distribution $\pi(\cdot \mid s)$ over actions, and for a deterministic policy π, we write $\pi(s) \in \mathcal{A}$ for its selected action at state s. At the end of the episode, the learner receives feedback on ℓ_t. In the *full-information* setting, the feedback is the entire loss function ℓ_t, whereas in the *bandit* setting, it is limited to the realized losses $\{(s_{t,h}, a_{t,h}, \ell_t(s_{t,h}, a_{t,h}))\}_{h=0}^{H-1}$. We assume $T \ge \max\{2, S, A\}$ for notational convenience.[1]

For a policy π and a loss function ℓ, the value functions are defined recursively, with terminal condition $V^\pi(s_H; \ell) = 0$. Specifically, the value function $V^\pi(s; \ell)$ and the Q-function $Q^\pi(s, a; \ell)$ satisfy $V^\pi(s; \ell) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s, a; \ell)]$ and $Q^\pi(s, a; \ell) = \ell(s, a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s'; \ell)]$. We may overload notation by allowing any function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to play the role of the loss function, in which case $V^\pi(s; g)$ and $Q^\pi(s, a; g)$ are defined accordingly. The goal is to minimize the regret $\mathrm{Reg}_T(\pi) := \mathbb{E}\big[\sum_{t=1}^T V^{\pi_t}(s_0; \ell_t) - \sum_{t=1}^T V^\pi(s_0; \ell_t)\big]$ against any fixed comparator policy π, and we denote by $\mathring{\pi} \in \arg\min_\pi \mathbb{E}\big[\sum_{t=1}^T V^\pi(s_0; \ell_t)\big]$ an optimal policy in hindsight.

Additional notation. For $N \in \mathbb{N}$, let $[N] := \{1, \dots, N\}$, given a vector x, let $\|x\|_p$ be its ℓ_p-norm for $p \in [1, \infty]$, and for a scalar z, let $[z]_+ := \max\{z, 0\}$. For a set $\mathcal{K}$, let $\Delta(\mathcal{K})$ denote the set of probability distributions over $\mathcal{K}$. If f and g are functions with $g(x) > 0$, we write $f \lesssim g$ or $f = \mathcal{O}(g)$ when $f(x) \le c\,g(x)$ holds on the relevant domain for some universal constant $c > 0$, and use $\widetilde{\mathcal{O}}(\cdot)$

[1] The assumptions $T \ge S$ and $T \ge A$ are not essential. If they fail, the analysis remains valid after replacing $\log(T)$ with $\log(SAT)$ or $\log(AT)$.

to hide logarithmic factors. Let $\{\mathcal{F}_t\}_{t\geq 0}$ be the natural filtration generated by all observations up to the end of episode $t-1$, and define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t, \ell_t]$. We use $\mathbb{1}[\cdot]$ for indicators, which equal 1 on the event and 0 otherwise. Let $\mathbb{I}_t(s,a) = \mathbb{1}[(s_{t,h}, a_{t,h}) = (s,a), \exists h \in \{0,\dots,H-1\}]$ denote whether a state-action pair (s,a) is visited in episode t under policy π_t and transition kernel P, and let $\mathbb{I}_t(s) = \sum_a \mathbb{I}_t(s,a)$. For each layer h, write $\ell_t(h) \in [0,1]^{\mathcal{S}_h \times \mathcal{A}}$ for the restriction of ℓ_t to $\mathcal{S}_h \times \mathcal{A}$, and use the same convention for other functions on $\mathcal{S} \times \mathcal{A}$.

2.1 Confidence sets of the transition

Following Jin et al. (2020), we construct confidence sets for the unknown transition kernel. Let $n_t(s,a,s')$ be the number of observed transitions from (s,a) to s' up to episode $t-1$, and let $n_t(s,a)$ be the corresponding visitation count. Whenever $n_t(s,a) = 0$, we set it to 1 in any denominator. The empirical transition estimate is then defined by $\widehat{P}_t(s' \mid s,a) = \frac{n_t(s,a,s')}{n_t(s,a)}$ for all h and $(s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$. The confidence set of the true transition at episode t is then defined as

$$\mathcal{P}_t = \Bigg\{ \widetilde{P} : \Big|\widetilde{P}(s' \mid s,a) - \widehat{P}_t(s' \mid s,a)\Big| \leq \min\Bigg\{ 2\sqrt{\frac{\widehat{P}_t(s' \mid s,a)\iota}{n_t(s,a)}} + \frac{14\iota}{3n_t(s,a)}, 1 \Bigg\}, \\ \forall (s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1},\ \forall h = 0,\dots,H-1,\ \widetilde{P}(\cdot \mid s,a) \in \Delta(\mathcal{S}_{h+1}) \Bigg\}, \tag{1}$$

where $\iota = \log(SAT/\delta)$. Throughout the paper, we set $\delta = 1/T^2$.

By Jin et al. (2020, Lemma 2), the true transition kernel P belongs to $\mathcal{P}_t$ for all episodes $t = 1,\dots,T$ with probability at least $1 - 4\delta$. We note that the set $\mathcal{P}_t$ is rectangular, since its constraints are imposed independently for each state-action pair (s,a).

For a transition kernel $\widetilde{P}$ and a policy π, let $q^{\widetilde{P},\pi}(s,a)$ be the probability of visiting (s,a) within an episode. We also use $q^{\widetilde{P},\pi}(s' \mid s,a)$ and $q^{\widetilde{P},\pi}(s',a' \mid s,a)$ for the corresponding conditional occupancy measures given that (s,a) has been visited and set them to zero whenever $h(s') < h(s)$. For each state s, define $q^{\widetilde{P},\pi}(s) := \sum_a q^{\widetilde{P},\pi}(s,a)$, so that $q^{\widetilde{P},\pi}(s,a) = q^{\widetilde{P},\pi}(s)\pi(a \mid s)$. When $\widetilde{P} = P$, we simply write q^π. We also define the upper and lower occupancies $\overline{q}_t^\pi(s)$ and $\underline{q}_t^\pi(s)$ with respect to the confidence set $\mathcal{P}_t$, and for the learner's policy π_t, the smoothed occupancy $q_t(s)$ and relative occupancy width $\rho_t(s)$ using an exploration rate $\gamma_t > 0$:

$$\overline{q}_t^\pi(s) := \max_{\widetilde{P} \in \mathcal{P}_t} q^{\widetilde{P},\pi}(s), \quad \underline{q}_t^\pi(s) := \min_{\widetilde{P} \in \mathcal{P}_t} q^{\widetilde{P},\pi}(s), \quad q_t(s) := \overline{q}_t^{\pi_t}(s) + \gamma_t, \quad \rho_t(s) := \frac{q_t(s) - \underline{q}_t^{\pi_t}(s)}{q_t(s)}.$$

These quantities can be computed efficiently by the dynamic-programming procedure of Jin et al. (2020, Algorithm 3). Similarly, for any transition kernel $\widetilde{P}$, policy π, and function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we write $V^{\widetilde{P},\pi}(s;g)$ and $Q^{\widetilde{P},\pi}(s,a;g)$ for the value and Q-functions computed under $\widetilde{P}$ and π. When $\widetilde{P} = P$, we again omit the transition kernel and write $V^\pi(s;g)$ and $Q^\pi(s,a;g)$.

2.2 Regime of environments

We consider three regimes for the loss sequence $\ell_1,\dots,\ell_T$. In the adversarial regime, we impose no assumption on the losses. At the beginning of each episode t, the environment selects an arbitrary loss function $\ell_t \in [0,1]^{S \times A}$, which may depend on the learner's algorithm, as well as on past trajectories and observed losses, but not on the learner's internal randomness.

In the stochastic regime, the loss functions $\ell_1,\dots,\ell_T$ are sampled i.i.d. from a fixed but unknown distribution $\mathcal{D}$. The stochastic regime with adversarial corruption interpolates between these two settings. In this regime, there is an underlying (non-corrupted) loss sequence $\ell'_1,\dots,\ell'_T$ sampled i.i.d. from an unknown distribution $\mathcal{D}$, while the learner observes a possibly corrupted loss sequence $\ell_1,\dots,\ell_T$. The corruptions may depend arbitrarily on the past trajectories and observed losses, and their total magnitude is measured by $\mathcal{C} := \mathbb{E}\big[\sum_{t=1}^T \sum_{h=0}^{H-1} \|\ell'_t(h) - \ell_t(h)\|_\infty\big] \in [0, HT]$. When $\mathcal{C} = 0$, this regime reduces to the stochastic regime; when $\mathcal{C}$ is arbitrary, it coincides with the adversarial regime. For the stochastic and corrupted stochastic regimes, let $\mu(s,a) := \mathbb{E}_{\ell' \sim \mathcal{D}}[\ell'(s,a)]$ be the

mean uncorrupted loss function. Assume that the optimal policy for μ is unique, and let it be denoted by $\pi^\star$. We define the suboptimality gap of a state-action pair (s,a) by $\Delta\colon \mathcal{S}\times\mathcal{A}\to[0,H]$ as $\Delta(s,a) \coloneqq Q^{\pi^\star}(s,a;\mu) - \min_{a'\in\mathcal{A}} Q^{\pi^\star}(s,a';\mu)$ and let $\Delta_{\min} \coloneqq \min_{s,a\neq\pi^\star(s)} \Delta(s,a) > 0$.

2.3 Complexity measures in online MDPs

We use the following data-dependent complexity measures to state refined regret bounds. The first-order quantity introduced by Lee et al. (2020) is $L(\pi) \coloneqq \mathbb{E}\big[\sum_{t=1}^T V^\pi(s_0;\ell_t)\big]$, which is the cumulative loss of a fixed comparator policy π. The second-order quantity $Q_\infty \coloneqq \min_{\ell^\star\in[0,1]^{S\times A}} \mathbb{E}\big[\sum_{t=1}^T \sum_{h=0}^{H-1} \|\ell_t(h) - \ell^\star(h)\|_\infty^2\big]$, which is small when the losses stay close to a fixed baseline loss, and the path-length quantity $V_1 \coloneqq \mathbb{E}\big[\sum_{t=1}^{T-1} \|\ell_{t+1} - \ell_t\|_1\big]$, which is small when the losses vary slowly over time, are defined as in Li et al. (2026).

Under unknown transitions, our regret bounds contain a transition-dependent complexity term that reflects the difficulty of estimating the transition kernel. For a loss sequence $\ell = \{\ell_t\}_{t=1}^T$ and a policy sequence $\pi_{1:T} = \{\pi_t\}_{t=1}^T$, define

$$Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) \coloneqq \sum_{t=1}^T \sum_{s,a} q^{\pi_t}(s,a)\mathsf{Var}_{s'\sim P(\cdot|s,a)}[V^{\pi_t}(s';\ell_t)] \in [0,H^2T]. \tag{2}$$

This quantity measures the variance of future cumulative losses across next states. Thus, transition uncertainty affects regret only when possible next states have different values $V^{\pi_t}(s';\ell_t)$. This mirrors transition-variance terms in stochastic MDPs, such as $\mathsf{Var}_{s'\sim P(\cdot|s,a)}\big[V^{\pi^\star}(s';\mu)\big]$ (Zanette and Brunskill, 2019; Simchowitz and Jamieson, 2019). Our definition replaces the optimal stochastic value $V^{\pi^\star}(\,\cdot\,;\mu)$ with learner-induced value functions so that the same idea applies to arbitrary loss sequences. As discussed later, this transition-dependent term is unavoidable under unknown transitions, since learning the transition kernel can itself cause regret even when $Q_\infty = V_1 = 0$. Furthermore, this term recovers a first-order regret bound through a self-bounding argument. In particular, the relation $\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)] \le H(L(\pi)+\mathrm{Reg}_T(\pi))$, together with a regret bound of the form $\mathrm{Reg}_T(\pi) \lesssim \sqrt{\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} + J$, implies $\mathrm{Reg}_T(\pi) \lesssim \sqrt{HL(\pi)} + J$. Thus, the transition-dependent complexity is at least as adaptive as the first-order complexity, while it can be sharper because it only counts transition uncertainty that affects future cumulative losses. We defer a detailed explanation and a simple illustrative example to Appendix B.2.

3 Algorithm and technical components

We are now ready to present our policy optimization algorithm for both the full-information and bandit settings. The common template is given in Algorithm 1, with the feedback-dependent definitions specified in Figure 1. Both algorithms follow the same policy optimization template and use new Q-function estimators and prediction terms to handle unknown transitions and adversarial losses. We describe the key components below.

3.1 Common policy optimization template

Optimistic follow-the-regularized-leader. Our algorithms rely on the optimistic follow-the-regularized-leader (OFTRL) framework (Chiang et al., 2012; Rakhlin and Sridharan, 2013; Steinhardt and Liang, 2014), which incorporates a prediction of the next loss vector into the standard FTRL. Our policy update is based on the OFTRL policy-optimization scheme of Li et al. (2026), which was developed for known transitions, and we modify it to handle unknown transitions.

Policy optimization reduces the MDP problem to a separate multi-armed bandit problem at each state. By the performance-difference lemma (Kakade and Langford, 2002), the regret can be written as $\mathrm{Reg}_T(\pi) = \mathbb{E}\big[\sum_s q^\pi(s)\sum_t \langle \pi_t(\cdot\mid s) - \pi(\cdot\mid s), Q^{\pi_t}(s,\,\cdot\,;\ell_t)\rangle\big]$. Thus, the Q-function plays the role of the loss vector for updating the action distribution at each state. Under unknown transitions, we first form the loss prediction $m_t \in [0,1]^{S\times A}$ and then compute its Q-function using an optimistic transition kernel from the confidence set. The OFTRL update with a convex regularizer ψ_t is defined,

for each state $s \in \mathcal{S}$, by

$$\pi_t(\cdot \mid s) = \operatorname*{arg\,min}_{\pi(\cdot|s)\in\Delta(\mathcal{A})} \left\{ \left\langle \pi(\cdot \mid s), \sum_{\tau=1}^{t-1} \left(\widehat{Q}_\tau(s,\cdot) - B_\tau(s,\cdot) \right) + Q^{\underline{P}_t^m,\pi_t}(s,\cdot;m_t) \right\rangle + \psi_t(\pi(\cdot \mid s)) \right\}. \tag{3}$$

Here, $\widehat{Q}_\tau$ is a Q-function estimator, B_τ is a dilated exploration bonus, and $Q^{\underline{P}_t^m,\pi_t}(s,\cdot;m_t)$ is the optimistic prediction of $Q^{\pi_t}(s,\cdot;\ell_t)$. The transition kernel $\underline{P}_t^m \in \mathcal{P}_t$ is chosen to minimize $Q^{\widetilde{P},\pi_t}(s,a;m_t)$ over $\widetilde{P} \in \mathcal{P}_t$ for each state-action pair (s,a). Both $\underline{P}_t^m$ and $Q^{\underline{P}_t^m,\pi_t}(s,\cdot;m_t)$ are computable by backward dynamic programming over the layers, since for $s \in \mathcal{S}_h$, computing this Q-function only requires later-layer values and policies. In our algorithms, we use the log-barrier regularizer $\psi_t(\pi(\cdot \mid s)) = \sum_{a\in\mathcal{A}} \frac{1}{\eta_t(s,a)} \log\left(\frac{1}{\pi(a|s)}\right)$, where $\eta_t(s,a) > 0$ are state-action-wise time-varying, data-dependent learning rates. In both the full-information and bandit settings, the algorithm uses the same update rule (3), and differs only in the construction of $\widehat{Q}_t$, B_t, and $\eta_t(s,a)$.

The optimistic prediction $Q^{\underline{P}_t^m,\pi_t}(s,\cdot;m_t)$ is obtained from a loss prediction sequence m_t. We construct m_t from the realized trajectory, starting with $m_1(s,a) = 1/2$ for all (s,a) and setting

$$m_{t+1}(s,a) = \begin{cases} (1-\xi)m_t(s,a) + \xi\ell_t(s,a), & \text{if } \mathbb{I}_t(s,a) = 1, \\ m_t(s,a), & \text{if } \mathbb{I}_t(s,a) = 0, \end{cases} \tag{4}$$

where $\xi \in (0, 1/2)$ is a step size. Since the transition kernel is unknown, we use this predictor even in the full-information setting, where it is useful for deriving path-length bounds.

Dilated exploration bonus. In policy optimization, a naive state-wise update of the action distributions can lead to insufficient exploration. To address this, Luo et al. (2021) introduced a dilated exploration bonus $B_t(s,a)$, which is constructed in the same form as a Q-function,

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

Since B_t is subtracted from the Q-function estimator in (3), a state-action pair (s,a) with a larger dilated exploration bonus is evaluated more optimistically and is therefore more likely to be visited. This leads to the following lemma.

Lemma 3.1 (Luo et al. 2021, Lemma B.1)**.** *Suppose that $P \in \mathcal{P}_t$ for all $t \in [T]$. Suppose that $b_t(s)$ is a nonnegative function. Suppose also that, for a comparator policy π, there exists $J^\pi \in \mathbb{R}$ such that*

$$\begin{aligned} &\mathbb{E}\left[\sum_s q^\pi(s) \sum_{t=1}^T \sum_a (\pi_t(a \mid s) - \pi(a \mid s))(Q^{\pi_t}(s,a;\ell_t) - B_t(s,a))\right] \\ &\le J^\pi + \mathbb{E}\left[\sum_{t=1}^T \sum_s q^\pi(s) b_t(s)\right] + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^T \sum_s \sum_a q^\pi(s)\pi_t(a \mid s) B_t(s,a)\right]. \end{aligned} \tag{6}$$

Then, it holds that $\mathrm{Reg}_T(\pi) \le J^\pi + 3\,\mathbb{E}\left[\sum_{t=1}^T V^{\overline{P}_t^B,\pi_t}(s_0;b_t)\right]$*, where $\overline{P}_t^B \in \mathcal{P}_t$ simultaneously attains the maxima in (5) for all (s,a).*

From this lemma, once the condition (6) is verified, it remains only to evaluate the value function $V^{\overline{P}_t^B,\pi_t}(s_0;b_t)$ of the bonus under the learner's own policies π_t, rather than a comparator policy π. For further intuition on the dilated bonus, we refer the reader to Luo et al. (2021); Dann et al. (2023).

3.2 Feedback-dependent Q-function estimator

The main new component is the Q-function estimator. The estimator differs between the full-information and bandit settings, and this choice also affects the learning-rate update and the local exploration bonus b_t. We collect the definitions in Figure 1 and explain the idea below. Note that the quantities $\underline{P}_t^\ell$, $\underline{P}_t^m$, B_t, and C_t in Figure 1 are efficiently computable by backward dynamic programming with a state-action-wise greedy procedure over the confidence set, as in Jin et al. (2020); Luo et al. (2021).

Algorithm 1 Data-Dependent Policy Optimization with Unknown Transitions

Input: $\gamma_t = \frac{1}{t}$, $\eta_1 = \frac{1}{12H^3}$, $m_1 = \frac{1}{2}$, $\xi = \frac{1}{4}$, $\psi_t(\pi(\cdot \mid s)) = \sum_{a \in \mathcal{A}} \frac{1}{\eta_t(s,a)} \log\left(\frac{1}{\pi(a|s)}\right)$

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

- Update $\mathcal{P}_t$ by (1).
- For each $s \in \mathcal{S}$, compute π_t by the OFTRL update (3).
- Set $Y_t \leftarrow 1$ in the full-information setting, and $Y_t \leftarrow \mathbb{1}\left[\max_{s,a} \frac{\eta_t(s,a)}{q_t(s)} \leq \frac{1}{50\sqrt{H^3 S^3 A}}\right]$ in the bandit setting.
- If $Y_t = 0$, insert a virtual episode and shift the indices of real episodes.
- If $Y_t = 1$, execute π_t, observe $\{(s_{t,h}, a_{t,h})\}_{h=0}^{H-1}$, and receive the corresponding loss feedback.
- Compute $\widehat{Q}_t$, η_{t+1}, and b_t as in Figure 1, and compute B_t by (5).
- Compute loss prediction $m_{t+1}(s,a)$ by (4).

Full-information feedback. In the full-information setting, the learner observes the entire loss function ℓ_t after each episode, so the Q-function estimator can be formed directly from the entire loss ℓ_t. We choose $\underline{P}_t^{\ell} \in \mathcal{P}_t$ so that it minimizes $Q^{\widetilde{P},\pi_t}(s,a;\ell_t)$ over $\widetilde{P} \in \mathcal{P}_t$ for all state-action pairs (s,a), and design the new optimistic Q-function estimator by (7). On the event that $P \in \mathcal{P}_t$ for all t, this estimator is optimistic, and the resulting bias is one-sided. This is the only additional bias caused by unknown transitions, and it is controlled by the transition-dependent complexity $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$ in the regret analysis.

The corresponding learning-rate update and local exploration bonus are given in (8) and (9). The update of the learning rate is designed so that the penalty term $(1/\eta_{t+1}(s,a) - 1/\eta_t(s,a))\log(T)$ matches the stability term $\eta_t(s,a)\zeta_t(s,a)$, where ζ_t is defined in (10). Thus, both terms are controlled through the same data-dependent quantity ζ_t. The local exploration bonus b_t is directly induced by this learning-rate update, with transition-estimation error included through ζ_t, so no separate transition bonus is needed in the full-information setting. Since the estimator uses the observed loss directly, the stability condition typically needed when analyzing OFTRL is easy to satisfy without introducing virtual episodes. Accordingly, Algorithm 1 sets $Y_t = 1$ throughout the full-information setting.

Bandit feedback. In the bandit setting, the learner observes losses only along the realized trajectory. To deal with this, our new Q-function estimator (11) therefore combines three ingredients: the optimistic prediction $Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)$, estimator of the prediction-error $Q^{\pi_t}(s,a;\ell_t - m_t)$, and the transition bonus $C_t(s,a)$. The main challenge is designing the transition bonus. It has to account for both loss-estimation error from bandit feedback and transition-estimation error from the unknown transition kernel. More precisely, the local transition bonus $c_t(s,a)$ must upper bound the estimator bias, while $c_t(s,a)$ itself remains controllable by data-dependent quantities. In particular, it should satisfy a relation of the form $\mathbb{E}_t[c_t(s,a)] \geq \rho_t(s)|Q^{\pi_t}(s,a;\ell_t - m_t)|$ for all t,s,a. The simpler local transition bonus $c_t(s) = \rho_t(s)H$ used in Dann et al. (2023) controls the estimator bias only through a worst-case $\widetilde{\mathcal{O}}(\sqrt{T})$ bound and thus does not yield the desired data-dependent bound.

To tackle this challenge, we instead define the data-dependent local transition bonus $c_t(s,a)$ in (12) and its recursive version $C_t(s,a)$ in (13), both constructed from the bandit feedback. The key point in our estimator is that, in the local transition bonus $c_t(s,a)$, the part that does not depend on the prediction error scales as $\rho_t(s)^2$ rather than $\rho_t(s)$. This makes the worst-case contribution of the estimator bias only $\mathrm{polylog}(T)$, and the other terms can be bounded by data-dependent quantities. In addition, we use the prediction term $Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)$, which is not automatically a lower estimate of $Q^{\pi_t}(s,a;\ell_t)$, as also noted by Li et al. (2026). We therefore define C_t with an extra margin, so that the estimator remains optimistic without losing data-dependent control.

The learning-rate update and local exploration bonus are given in (14) and (15). As in the full-information setting, the penalty term of regret $(1/\eta_{t+1}(s,a) - 1/\eta_t(s,a))\log(T)$ is chosen to match the log-barrier stability term $\eta_t(s,a)\zeta_t(s,a)/q_t(s)^2$, where ζ_t is defined in (16). The local exploration bonus b_t contains the term induced by this learning-rate update, together with an additional term involving $C_t(s,a)^2$, which accounts for transition-estimation error.

Full-information.
Q-function estimator.

$$\widehat{Q}_t(s,a) = Q^{\underline{P}_t^\ell,\pi_t}(s,a;\ell_t) \qquad \text{with} \quad \underline{P}_t^\ell \in \operatorname*{arg\,min}_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;\ell_t) \text{ for all } (s,a). \tag{7}$$

Learning rate and exploration bonus.

$$\frac{1}{\eta_{t+1}(s,a)} = \frac{1}{\eta_t(s,a)} + \frac{\eta_t(s,a)\zeta_t(s,a)}{\log(T)}, \tag{8}$$

$$b_t(s) = 7\sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T), \tag{9}$$

$$\text{where} \quad \zeta_t(s,a) = \pi_t(a\mid s)(1-\pi_t(a\mid s))\sum_b \pi_t(b\mid s)\Big(Q^{\underline{P}_t^\ell,\pi_t}(s,b;\ell_t) - Q^{\underline{P}_t^m,\pi_t}(s,b;m_t)\Big)^2. \tag{10}$$

Bandit feedback.
Q-function estimator. Let $L_{t,h} = \sum_{h'=h}^{H-1}\ell_t(s_{t,h'},a_{t,h'})$, $M_{t,h} = \sum_{h'=h}^{H-1} m_t(s_{t,h'},a_{t,h'})$, and $D_{t,h} = \sum_{h'=h}^{H-1}|\ell_t(s_{t,h'},a_{t,h'}) - m_t(s_{t,h'},a_{t,h'})|$, and define

$$\widehat{Q}_t(s,a) = Q^{\underline{P}_t^m,\pi_t}(s,a;m_t) + \frac{\mathbb{I}_t(s,a)\big(L_{t,h(s)} - M_{t,h(s)}\big)}{q_t(s)\pi_t(a\mid s)}Y_t - C_t(s,a)Y_t, \tag{11}$$

$$\text{where} \quad c_t(s,a) = \rho_t(s)\frac{\mathbb{I}_t(s,a)D_{t,h(s)}}{q_t(s)\pi_t(a\mid s)} + \rho_t(s)^2 H, \tag{12}$$

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

Learning rate and exploration bonus. Let $(s_t^\dagger, a_t^\dagger)$ be any maximizer of $\eta_t(s,a)/q_t(s)$, and define

$$\frac{1}{\eta_{t+1}(s,a)} = \begin{cases} \frac{1}{\eta_t(s,a)} + \frac{\eta_t(s,a)\zeta_t(s,a)}{q_t(s)^2\log(T)}, & \text{if } t \text{ is a real episode,} \\ \frac{1}{\eta_t(s,a)}\left(1 + \frac{\mathbb{1}\{(s_t^\dagger,a_t^\dagger)=(s,a)\}}{324H\log(T)}\right), & \text{if } t \text{ is a virtual episode,} \end{cases} \tag{14}$$

$$b_t(s) = 6\sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) + 3\sum_a \eta_t(s,a)\pi_t(a\mid s)^2 C_t(s,a)^2 Y_t, \tag{15}$$

$$\text{where} \quad \zeta_t(s,a) = (\mathbb{I}_t(s,a) - \pi_t(a\mid s)\mathbb{I}_t(s))^2\big(L_{t,h(s)} - M_{t,h(s)}\big)^2. \tag{16}$$

Figure 1: Definitions of the Q-function estimator $\widehat{Q}_t$, the learning-rate update η_{t+1}, and the local exploration bonus b_t used in Algorithm 1.

Virtual episodes. In the bandit setting, $B_t(s,a)$ can scale as $1/q_t(s)^2$ through (15), so the resulting bonus-induced stability term cannot be controlled by the adaptive learning-rate update alone when $q_t(s)$ is small. Therefore, following Dann et al. (2023), we insert virtual episodes to control the ratio $\eta_t(s,a)/q_t(s)$. If $\max_{s,a}\eta_t(s,a)/q_t(s) > 1/(50\sqrt{H^3S^3A})$ at the beginning of episode t, we set $Y_t = 0$, make episode t virtual, and shift the indices of all subsequent real episodes. The algorithm then uses the zero loss function and only updates the pair $(s_t^\dagger, a_t^\dagger) \in \arg\max_{s,a}\eta_t(s,a)/q_t(s)$, multiplying $1/\eta_t(s_t^\dagger,a_t^\dagger)$ by $1+\frac{1}{324H\log(T)}$. It still computes $\widehat{Q}_t$, b_t, and B_t, but keeps the prediction m_t and the confidence set $\mathcal{P}_t$ unchanged. Since the number of virtual episodes is $\mathcal{O}(HSA\log^2(T))$, their effect is lower-order, and we still use T to denote the total number of episodes.[2]

4 Main results

Regret upper bounds. We now state the guarantees of Algorithm 1. The first theorem is for the full-information setting, and the second is for the bandit setting. In both cases, the algorithm achieves first-order, second-order, and path-length bounds in the adversarial regime, and at the same time achieves a gap-dependent $\mathrm{polylog}(T)$ bound in the stochastic regime. The transition-dependent

[2]The data-dependent complexity measures in the main statements are defined only over real episodes.

complexity $Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)$ captures transition uncertainty that affects future cumulative losses and can imply first-order dependence on $L(\pi)$ through self-bounding.

Theorem 4.1. *In the full-information setting, for any comparator policy π, Algorithm 1 guarantees*

$$\mathrm{Reg}_T(\pi) \lesssim \sqrt{H^2SA\log(T)\min\{L(\pi), HT - L(\pi), Q_\infty, V_1\}} + \sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]} + \left(H^4 + HS^2\right)A\log^2(T).$$

Under the stochastic regime with adversarial corruption, it simultaneously ensures $\mathrm{Reg}_T(\pi) \lesssim U + \sqrt{UC} + \left(H^4 + HS^2\right)A\log^2(T)$, *where* $U = \frac{H^2S^2A\log^2(T)}{\Delta_{\min}}$.

The above bound recovers the $\widetilde{\mathcal{O}}(\sqrt{H^2S^2AT})$ worst-case rate of occupancy-measure-based algorithms (Rosenberg and Mansour, 2019; Jin et al., 2021). To our knowledge, this is the first policy-optimization result to remove the extra H factor (Luo et al., 2021; Dann et al., 2023). In the stochastic regime, the leading gap-dependent term also has sharper dependence on H, S, and A than that of Jin et al. (2021), up to logarithmic factors.

Theorem 4.2. *In the bandit setting, for any comparator policy π, Algorithm 1 guarantees*

$$\mathrm{Reg}_T(\pi) \lesssim \sqrt{H^3S^2A\log^2(T)\min\{L(\pi), HT - L(\pi), Q_\infty, V_1\}} + \sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]} + \left(H^3S^2A + H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T).$$

Under the stochastic regime with adversarial corruption, it simultaneously ensures $\mathrm{Reg}_T(\pi) \lesssim U + \sqrt{UC} + \left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T)$, *where* $U = \frac{H^4S^2A\log^2(T)}{\Delta_{\min}}$.

In the worst case, the bound for the adversarial regime recovers the worst-case rate of Dann et al. (2023). For the first-order bound, the $Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)$ term is absorbed by the first-order term through the self-bounding argument. Thus, the final bound does not contain an explicit transition-dependent term. In the stochastic regime, the leading gap-dependent term matches the $H^4S^2A/\Delta_{\min}$ dependence of Dann et al. (2023). However, the main contribution of Theorems 4.1 and 4.2 is their data-dependent regret bounds under unknown transitions. Together, these theorems give the first policy-optimization guarantees under unknown transitions with first-order, second-order, and path-length adaptivity.

Lower bounds. We show that the data-dependent terms in our adversarial guarantees cannot be avoided in general. First, by running a standard hard instance for adversarial MDPs with unknown transitions (Jin et al., 2018; Domingues et al., 2021) for the first $T' \le T$ episodes and setting all later losses to zero, the complexities satisfy $L(\pi) = \mathcal{O}(HT')$, $Q_\infty = \mathcal{O}(HT')$, $V_1 = \mathcal{O}(SAT')$, and $\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)] = \mathcal{O}(H^2T')$, while $\max_\pi \mathrm{Reg}_T(\pi) \ge \Omega(\sqrt{H^2SAT'})$. Thus, one can obtain $\Omega(\sqrt{HSAL(\pi)})$, $\Omega(\sqrt{HSAQ_\infty})$, $\Omega(\sqrt{H^2V_1})$ and $\Omega(\sqrt{SA\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]})$ for the adversarial regime. Closing the remaining gaps is not specific to our data-dependent setting, and remains open even for worst-case policy optimization (Luo et al., 2021).

The following proposition further shows that a transition-dependent term is necessary under unknown transitions. The lower-bound constructions of Jin et al. (2018, Theorem 4) and Domingues et al. (2021, Theorem 9) use time-invariant loss sequences, so the regret arises solely from transition uncertainty.

Proposition 4.3. *For any algorithm, there exists an episodic MDP with unknown transitions and a time-invariant loss sequence such that $Q_\infty = V_1 = 0$, and the algorithm suffers* $\max_\pi \mathrm{Reg}_T(\pi) = \Omega(\sqrt{H^2SAT})$. *This lower bound holds even under full-information feedback.*

5 Proof overview

We sketch the proof for the bandit setting, which contains the main ideas also used in the full-information setting. The complete proofs are deferred to Appendix F for the full-information setting and Appendix G for the bandit setting. We use Lemma 3.1 as the main reduction for policy optimization. Recall that once the inequality in (6) is verified, the lemma reduces the regret

analysis to bounding $\sum_{t=1}^{T} V^{\overline{P}_t^B,\pi_t}(s_0;b_t)$. We prove (6) by decomposing its left-hand side into $\mathbb{E}\left[\sum_s q^\pi(s)(\mathbf{reg}^\pi(s)+\mathbf{bias}^\pi(s)+\mathbf{error}^\pi(s))\right]$, where

$$\mathbf{reg}^\pi(s)=\sum_{t,a}(\pi_t(a\mid s)-\pi(a\mid s))\left(\widehat{Q}_t(s,a)-B_t(s,a)\right),$$
$$\mathbf{bias}^\pi(s)=\sum_{t,a}(\pi_t(a\mid s)-\pi(a\mid s))\left(Q^{\pi_t}(s,a;\ell_t)-\widehat{Q}_t(s,a)+Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)-Q^{\pi_t}(s,a;m_t)\right),$$
$$\mathbf{error}^\pi(s)=\sum_{t,a}(\pi_t(a\mid s)-\pi(a\mid s))\left(Q^{\pi_t}(s,a;m_t)-Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)\right).$$

Here, $\mathbf{bias}^\pi(s)$ is the estimator bias, while $\mathbf{error}^\pi(s)$ comes from evaluating the prediction using $\underline{P}_t^m$ instead of the true transition kernel. The standard analysis in Lemma G.7 bounds $\mathbf{reg}^\pi(s)$ as

$$\mathbb{E}[\mathbf{reg}^\pi(s)]\le\mathcal{O}\left(HS^2A\log(T)\right)+\mathbb{E}\left[\sum_{t=1}^{T}b_t(s)\right]+\mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_a\pi_t(a\mid s)B_t(s,a)\right].$$

The remaining terms are controlled in Lemmas G.8 and G.9 by

$$\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{bias}^\pi(s)\right]\lesssim\widetilde{\mathcal{O}}\left(\sqrt{H^2S^2A\iota^2\mathbb{E}\left[\sum_{t:Y_t=1}\sum_{s,a}q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)\right]}\right),$$
$$\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{error}^\pi(s)\right]\lesssim\widetilde{\mathcal{O}}\left(\sqrt{S^2A\iota^2\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(m)]}\right).$$

The key point is that $\mathbf{bias}^\pi(s)$ is reduced to the prediction error ℓ_t-m_t. This follows from the design of $c_t(s,a)$, where the part independent of the prediction error scales with $\rho_t(s)^2$ rather than $\rho_t(s)$, and $\sum_{t:Y_t=1}\sum_s q^{\pi_t}(s)\rho_t(s)^2$ is only polylogarithmic in T (see Lemma G.5). The error term is captured by $Q_{\mathrm{trans}}^{\pi_{1:T}}(m)$, which can be reduced to $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$ through $Q_{\mathrm{trans}}^{\pi_{1:T}}(m)\le 2Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)+2H\sum_{t:Y_t=1}V^{\pi_t}(s_0;(\ell_t-m_t)^2)$ in Lemma D.12. Thus, all remaining terms reduce to the prediction error ℓ_t-m_t and $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$, yielding the desired first-order, second-order, and path-length bounds.

For the stochastic regime with adversarial corruption, we use the same decomposition as above. We show in Lemmas G.8 and G.9 that the individual terms can be bounded by factors of the form $\sum_s\sum_{a\neq\pi^\star(s)}\sqrt{\mathbb{E}\left[\sum_{t:Y_t=1}q^{\pi_t}(s,a)\right]}$ or $\sqrt{\mathbb{E}\left[\sum_{t:Y_t=1}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\mathring{\pi}}(s,a)]_+\right]}$. These quantities are then controlled by a self-bounding argument following Dann et al. (2023), yielding the desired gap-dependent bound under adversarial corruption.

6 Conclusion

In this work, we resolved the open problem raised by Dann et al. (2023) on data-dependent regret bounds for policy optimization in MDPs with unknown transitions. Our algorithms achieve refined best-of-both-worlds guarantees, adapting to $L(\pi)$, Q_∞, and V_1 in the adversarial regime while attaining a gap-dependent $\mathrm{polylog}(T)$ bound in the stochastic regime with adversarial corruption. A natural future direction is to obtain variance-aware gap-dependent bounds as in Chen et al. (2025), and to investigate whether the refined data-dependent guarantees can be achieved under aggregate feedback.

Acknowledgements

TT is supported by JSPS KAKENHI Grant Number JP26K21297 and KY is partially supported by JSPS KAKENHI Grant Number JP24H00703.

References

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 229–243. Springer, 2006.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165. Springer Berlin Heidelberg, 2007.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory*, volume 49, pages 116–120. PMLR, 2016.

Peter Auer, Nicoló Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 263–272. PMLR, 2017.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 508–528. PMLR, 2019.

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

Nicolo Cesa-Bianchi, Philip M Long, and Manfred K Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996.

Shulun Chen, Runlong Zhou, Zihan Zhang, Maryam Fazel, and Simon Shaolei Du. Sharp gap-dependent variance-aware regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 6.1–6.20. PMLR, 2012.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

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

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132, pages 578–598. PMLR, 2021.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(35):1287–1311, 2011.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 2552–2583. PMLR, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Conference on Learning Theory*, pages 1421–1422. PMLR, 2022.

Shinji Ito, Kevin Jamieson, Haipeng Luo, Arnab Maiti, and Taira Tsuchiya. Adapting to stochastic and adversarial losses in episodic MDPs with aggregate bandit feedback. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, pages 4863–4873. Curran Associates, Inc., 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 4860–4869. PMLR, 2020.

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

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

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

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

Tal Lancewicki and Yishay Mansour. Near-optimal regret using policy optimization in online MDPs with aggregate bandit feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 32467–32491. PMLR, 2025.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. In *Advances in Neural Information Processing Systems*, volume 33, pages 15522–15533. Curran Associates, Inc., 2020.

Mingyi Li, Taira Tsuchiya, and Kenji Yamanishi. Data- and variance-dependent regret bounds for online tabular MDPs. In *International Conference on Machine Learning*, 2026.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

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

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the Tsallis-INF algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 3330–3350. PMLR, 2021.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of the 28th Conference on Learning Theory*, volume 40, pages 1360–1375. PMLR, 2015.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *Proceedings of the 23rd Conference on Learning Theory*, pages 231–243, 2010.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 993–1019. PMLR, 2013.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5478–5486. PMLR, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1287–1295. PMLR, 2014.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8604–8613. PMLR, 2020.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, volume 32, pages 1153–1162. Curran Associates, Inc., 2019.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1593–1601. PMLR, 2014.

Daniil Tiapkin, Evgenii Chzhen, and Gilles Stoltz. Narrowing the gap between adversarial and stochastic MDPs via policy optimization. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1263–1291. PMLR, 2018.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular MDPs via adaptive multi-step bootstrap. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 4438–4472. PMLR, 2021.

Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. In *Recent Advances in Reinforcement Learning*, pages 268–281. Springer Berlin Heidelberg, 2008.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26, pages 1583–1591. Curran Associates, Inc., 2013.

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

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *the Twentieth International Conference on Machine Learning*, pages 928–935, 2003.

Contents

1 Introduction **1**

2 Preliminaries **3**

2.1 Confidence sets of the transition 4

2.2 Regime of environments 4

2.3 Complexity measures in online MDPs 5

3 Algorithm and technical components **5**

3.1 Common policy optimization template 5

3.2 Feedback-dependent Q-function estimator 6

4 Main results **8**

5 Proof overview **9**

6 Conclusion **10**

A Summary of notation **15**

B Additional related work and discussion **16**

B.1 Additional related work 16

B.2 Discussion of the transition-dependent complexity 17

C High probability event **18**

D Auxiliary lemmas **19**

D.1 Technical lemmas 19

D.2 Data-dependent and self-bounding lemmas 28

E General lemma of optimistic follow-the-regularized-leader **33**

F Regret analysis for the full-information setting (Theorem 4.1) **34**

F.1 Auxiliary lemmas 34

F.2 Common regret analysis 36

F.3 Proof for the adversarial regime 43

F.4 Proof for the stochastic regime with adversarial corruption 45

G Regret analysis for the bandit setting (Theorem 4.2) **45**

G.1 Auxiliary lemmas 46

G.2 Common regret analysis 52

G.3 Proof for the adversarial regime 65

G.4 Proof for the stochastic regime with adversarial corruption 67

A Summary of notation

For the reader's convenience, Table 2 collects the main notation used throughout the paper.

Table 2: Summary of notation.

Symbol	Meaning
Tabular MDPs	
$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, H, s_0)$	Episodic tabular MDP
$\mathcal{S}, S$	State space and its size $S = \lvert\mathcal{S}\rvert$
$\mathcal{A}, A$	Action space and its size $A = \lvert\mathcal{A}\rvert$
P	Transition kernel
H	Horizon length
T	Number of episodes
$h(s)$	Layer index of state s
$s_{t,h},\ a_{t,h}$	State / action at step h in episode t
$\ell_t(s,a)$	Loss assigned to (s,a) in episode t
π_t	Policy in episode t
$\mathrm{Reg}_T(\pi)$	Regret over T episodes with comparator π
$\mathbb{I}_t(s,a)$	Visitation indicator of (s,a) in episode t
$n_t(s,a)$	Number of visits to (s,a) up to episode $t-1$
$V^\pi(s;\ell)$	Value function under policy π from state s with loss ℓ
$Q^\pi(s,a;\ell)$	Q-function under policy π from (s,a) with loss ℓ
$q^\pi(s),\ q^\pi(s,a)$	Occupancy measure under policy π
$q^\pi(s',a' \mid s,a)$	Conditional occupancy from (s,a) to (s',a') under π
$\hat{\pi}$	Optimal policy for the regret
$\ell'_t(s,a)$	Uncorrupted i.i.d. loss
$\mathcal{C}$	Corruption budget
$\mu(s,a)$	Mean of $\ell'_t(s,a)$
$\pi^\star$	Optimal deterministic policy under μ
$\Delta(s,a), \Delta_{\min}$	Suboptimality gap at (s,a), and minimum positive gap
Data-dependent complexity measures	
$L(\pi)$	First-order complexity
Q_∞	Second-order complexity
V_1	Path-length complexity
$Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)$	Transition-dependent complexity in (2)
Notation for Algorithm 1	
ψ_t	Regularizer in episode t
$\eta_t(s,a)$	Learning rate for (s,a) in episode t
$m_t(s,a)$	Loss prediction for (s,a)
$\zeta_t(s,a)$	Data-dependent term for updating $\eta_t(s,a)$
$\widehat{Q}_t(s,a)$	Q-function estimator
$b_t(s), B_t(s,a)$	Exploaration bonus and its recursive Q-function version
$c_t(s,a), C_t(s,a)$	Transition bonus and its recursive Q-function version
$Y_t \in \{0,1\}$	Episode indicator ($Y_t = 1$ real, $Y_t = 0$ virtual)
$\mathcal{T}_r,\ \mathcal{T}_v$	Sets of real and virtual episodes
$\overline{q}^\pi_t(s), \underline{q}^\pi_t(s)$	Upper and lower occupancy measures
$\gamma_t, q_t(s) \coloneqq \overline{q}^{\pi_t}_t(s) + \gamma_t$	Exploration rate and smoothed state occupancy
$\rho_t(s) \coloneqq (q_t(s) - \underline{q}^{\pi_t}_t(s))/q_t(s)$	relative occupancy width
$L_{t,h(s)}, M_{t,h(s)}$	Realized / predicted suffix loss from layer $h(s)$
$D_{t,h(s)}$	Suffix prediction error from layer $h(s)$
$\mathcal{P}_t$	Confidence set at episode t
$\underline{P}^\ell_t$	Transition in $\mathcal{P}_t$ minimizing $Q(s,a;\ell_t)$ for all (s,a)
$\underline{P}^m_t$	Transition in $\mathcal{P}_t$ minimizing $Q(s,a;m_t)$ for all (s,a)
$\overline{P}^B_t$	Transition in $\mathcal{P}_t$ maximizing $B_t(s,a)$ for all (s,a)
$\overline{P}^C_t$	Transition in $\mathcal{P}_t$ maximizing $C_t(s,a)$ for all (s,a)

We formalize the conditional occupancy measure $q^{\pi_t}(s', a' \mid s, a)$ as follows:

$$\begin{cases} 0 & \text{if } h(s') < h(s), \\ 0 & \text{if } h(s') = h(s) \text{ and } (s', a') \neq (s, a), \\ 1 & \text{if } (s', a') = (s, a), \\ \Pr\{(s_{t,h(s')}, a_{t,h(s')}) = (s', a') \mid (s_{t,h(s)}, a_{t,h(s)}) = (s, a)\}, & \text{if } h(s') > h(s). \end{cases}$$

For $\underline{P}_t^{\ell}$, $\underline{P}_t^{m}$, $\overline{P}_t^{B}$, and $\overline{P}_t^{C}$, we choose kernels in $\mathcal{P}_t$ that attain the required state-action-wise minimum or maximum for all (s, a) simultaneously. This is possible since $\mathcal{P}_t$ is rectangular, and these kernels are efficiently computable by backward dynamic programming with a state-action-wise greedy procedure over the confidence set, as in Jin et al. (2020, Algorithm 3) and Luo et al. (2021, Algorithms 4–6).

B Additional related work and discussion

B.1 Additional related work

Tabular MDPs. Adversarial MDPs have been studied since the works of Even-Dar et al. (2009); Yu et al. (2008), and the episodic formulation was later introduced by Zimin and Neu (2013). Algorithms for online tabular MDPs can be broadly divided into two approaches: occupancy-measure based *global optimization* and *policy optimization*. Global optimization solves an optimization problem over the set of occupancy measures, and has led to strong regret guarantees in several settings (Zimin and Neu, 2013; Jin et al., 2020). However, global optimization can be computationally demanding because it optimizes over the occupancy measure space of the MDP. Policy optimization takes a more local approach by optimizing an action distribution at each state, which is often practical and computationally efficient (Shani et al., 2020; Luo et al., 2021).

Under known transitions and bandit feedback, global optimization achieves the minimax-optimal regret rate $\widetilde{\mathcal{O}}(\sqrt{HSAT})$ (Zimin and Neu, 2013). In contrast, the best known bound for policy optimization is $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$, achieved via dilated exploration (Luo et al., 2021). Compared with the global optimization rate, this bound incurs an extra factor of H, often attributed to Q-function estimation.

Unknown transitions introduce an additional difficulty, since the learner must also estimate the transition kernel. Among global optimization, the best known regret bounds were obtained by Rosenberg and Mansour (2019) for full-information feedback and Jin et al. (2020) for bandit feedback, both of order $\widetilde{\mathcal{O}}(\sqrt{H^2S^2AT})$. For policy optimization under bandit feedback, the best known guarantee was established by Luo et al. (2021), with regret $\widetilde{\mathcal{O}}(\sqrt{H^4S^2AT})$. On the other hand, minimax lower bounds of order $\Omega(\sqrt{H^2SAT})$ were established by Jin et al. (2018); Domingues et al. (2021), leaving gaps in the dependence on the horizon and the number of states. A line of work has sought to close these gaps, and Tiapkin et al. (2025) proposed a policy-optimization algorithm under full-information feedback with regret $\widetilde{\mathcal{O}}(\sqrt{H^6SAT})$. More recent work has studied policy optimization under more challenging feedback models, including aggregate feedback (Lancewicki and Mansour, 2025; Ito et al., 2025).

In the stochastic setting, near-optimal regret guarantees have been established by both model-based and value-based methods (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019). Beyond such gap-independent guarantees, a line of work has studied gap-dependent $\mathrm{polylog}(T)$ regret bounds for episodic tabular MDPs (Simchowitz and Jamieson, 2019; Dann et al., 2021; Xu et al., 2021).

Best-of-both-worlds algorithms. The best-of-both-worlds guarantee aims to obtain near-optimal regret in both adversarial and stochastic regimes using a single algorithm. This idea was first studied for multi-armed bandits by Bubeck and Slivkins (2012), and was later developed in several directions (Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017). A common approach is to use follow-the-regularized-leader with carefully chosen regularizers, which allows the algorithm to remain robust in the adversarial regime while improving automatically in the stochastic regime (Wei and Luo, 2018; Zimmert and Seldin, 2021; Ito, 2021). The stochastic-regime analysis is

often based on a self-bounding technique (Zimmert and Seldin, 2021; Masoudian and Seldin, 2021), and the same idea is useful for handling stochastic losses with adversarial corruption.

Best-of-both-worlds guarantees have also been studied in tabular MDPs. For global optimization, Jin and Luo (2020); Jin et al. (2021) first developed best-of-both-worlds algorithms, with subsequent extensions to adversarial transitions and aggregate feedback (Jin et al., 2023; Ito et al., 2025). In policy optimization, Dann et al. (2023) established best-of-both-worlds guarantees under bandit feedback for several regularizers, including Tsallis entropy, Shannon entropy, and log-barrier regularization, while Li et al. (2026) later obtained refined data-dependent guarantees using the OFTRL framework under known transitions.

Data-dependent bounds. Data-dependent regret bounds adapt to the realized loss sequence instead of depending only on worst-case parameters. They have been widely studied in online learning, including learning with expert advice (Littlestone and Warmuth, 1994), multi-armed bandits (Auer et al., 2002), and online convex optimization (Zinkevich, 2003). Typical examples in adversarial regimes include first-order bounds based on the comparator's cumulative loss (Cesa-Bianchi et al., 1996; Allenberg et al., 2006; Neu, 2015), second-order bounds based on loss fluctuations (Hazan and Kale, 2011; Wei and Luo, 2018; Ito, 2021), and path-length bounds based on temporal variation (Wei and Luo, 2018; Bubeck et al., 2019; Ito, 2021).

In tabular MDPs, it is also important to obtain regret bounds that adapt to data-dependent complexity measures. The first such result was obtained by Lee et al. (2020), who derived a first-order bound for MDPs with unknown transitions via global optimization. For policy optimization, Dann et al. (2023) established best-of-both-worlds guarantees under unknown transitions and first-order data-dependent guarantees under known transitions, leaving open whether data-dependent guarantees are possible under unknown transitions. More recently, Li et al. (2026) showed that, under known transitions, OFTRL-based policy optimization can adapt to multiple complexity measures, including first-order, second-order, and path-length measures.

A related line of work has studied gap-dependent bounds in stochastic regimes. For multi-armed bandits, Audibert et al. (2007) introduced variance-aware gap-dependent bounds, and related ideas have been incorporated into best-of-both-worlds analyses (*e.g.,* Ito et al. 2022). For tabular MDPs, recent works have also studied variance-aware gap-dependent $\mathrm{polylog}(T)$ regret bounds (Simchowitz and Jamieson, 2019; Dann et al., 2021; Chen et al., 2025; Li et al., 2026).

These studies left open the case of data-dependent policy optimization under unknown transitions, especially when combined with best-of-both-worlds guarantees. Our work addresses these gaps by showing that policy optimization can achieve refined data-dependent and best-of-both-worlds guarantees under unknown transitions, up to an additional transition-dependent complexity term.

B.2 Discussion of the transition-dependent complexity

In this part, we discuss the role of the transition-dependent complexity $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$, which captures the difficulty caused by unknown transitions. By Lemma D.6, we have

$$Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) \leq H \sum_{t=1}^{T} \sum_{s,a} q^{\pi_t}(s,a) \ell_t(s,a)^2. \tag{17}$$

Since $\ell_t(s,a) \in [0,1]$, taking expectations gives

$$\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)] \leq H \mathbb{E}\left[\sum_{t=1}^{T} \sum_{s,a} q^{\pi_t}(s,a) \ell_t(s,a)\right] = H(L(\pi) + \mathrm{Reg}_T(\pi)) \tag{18}$$

for each policy π. From this inequality, we can show that the transition-dependent bound recovers a regret bound depending on the first-order complexity. Suppose that for the any comparator policy π, there exists an absolute constant $c > 0$ such that

$$\mathrm{Reg}_T(\pi) \leq c\sqrt{S^2 A \log^2(T) \mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} + J$$

for some $J > 0$, as the regret bounds in Theorems 4.1 and 4.2. Then, using the inequality (18), the regret is upper bounded as

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\le c\sqrt{HS^2A\log^2(T)(L(\pi)+\mathrm{Reg}_T(\pi))} + J \\ &\le c\sqrt{HS^2A\log^2(T)L(\pi)} + c\sqrt{HS^2A\log^2(T)\mathrm{Reg}_T(\pi)} + J \\ &\le c\sqrt{HS^2A\log^2(T)L(\pi)} + \frac{1}{2}\mathrm{Reg}_T(\pi) + \frac{1}{2}c^2HS^2A\log^2(T) + J\end{aligned}$$

where the last inequality follows from the AM–GM inequality. Thus, we obtain

$$\mathrm{Reg}_T(\pi) \lesssim \sqrt{HS^2A\log^2(T)L(\pi)} + HS^2A\log^2(T) + J.$$

This implies that $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$ is at least as adaptive as the first-order complexity. However, the comparison in (17) can be loose. Transition uncertainty contributes to $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$ only when it creates variation in the future cumulative loss.

The following example shows that $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$ can be small even when the cumulative loss is large. Consider an $H = 3$ layered MDP with layers $\mathcal{S}_0 = \{s_0\}$, $\mathcal{S}_1 = \{x_1, x_2\}$, and $\mathcal{S}_2 = \{y\}$, followed by the terminal state. For each action a, the transition $P(\cdot \mid s_0, a)$ may be stochastic and unknown over $\{x_1, x_2\}$. For every action, the transitions from x_1 and x_2 go deterministically to y, and the transition from y goes deterministically to the terminal state.

Let the losses at s_0, x_1 and x_2 be zero, and let the loss at y be one:

$$\ell_t(s_0, a) = 0, \qquad \ell_t(x_i, a) = 0 \quad (i \in \{1, 2\}), \qquad \ell_t(y, a) = 1, \qquad \forall a \in \mathcal{A},\ \forall t \in [T].$$

Since every trajectory passes through y and the loss at y is one, the cumulative loss is linear in T, and in particular

$$\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s, a)\ell_t(s, a)^2 = T.$$

However, the actual losses satisfy

$$V^{\pi_t}(x_1; \ell_t) = V^{\pi_t}(x_2; \ell_t).$$

Hence, for every action a,

$$\mathsf{Var}_{s' \sim P(\cdot|s_0,a)}\left[V^{\pi_t}(s'; \ell_t)\right] = 0.$$

The other transitions are deterministic, so no transition contributes to $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$, and thus $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) = 0$.

In this example, the right-hand side of (17) is of order HT, whereas $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) = 0$. By adding deterministic layers after y, the same construction gives an $\mathcal{O}(H^2T)$ term on the right-hand side while keeping $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) = 0$. This shows that $Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)$ can be much smaller than the upper bound in (17) and (18).

C High probability event

We define the following good event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, which holds with high probability as shown in Definition C.4.

$$\begin{aligned}\mathcal{E}_1 &= \{P \in \mathcal{P}_t,\ \forall t = 1, \dots, T\}, \\ \mathcal{E}_2 &= \left\{\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} \frac{q^{\pi_t}(s, a)}{n_t(s, a)} \lesssim |\mathcal{S}_h|A\log T + \log(H/\delta),\ \forall h = 0, \dots, H-1\right\}.\end{aligned}$$

Lemma C.1 (Event $\mathcal{E}_1$, Jin et al. 2020, Lemma 2)**.** *With probability at least* $1-4\delta$*, we have* $P \in \mathcal{P}_t$ *for all* $t = 1, \dots, T$.

Lemma C.2 (Jin et al. 2021, Corollary D.3.4)**.** *If* $P \in \mathcal{P}_t$*, then for all* $\widetilde{P} \in \mathcal{P}_t$

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

Lemma C.3 (Event $\mathcal{E}_2$, Jin et al. 2020, Lemma 10)**.** *We have with probability at least* $1-\delta$

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

for all layers h.

Definition C.4. Define $\mathcal{E}$ to be the event that $P \in \mathcal{P}_t$ for all t and the bound in Lemma C.3 holds. In this case, $\Pr(\mathcal{E}) \ge 1 - 5\delta$.

D Auxiliary lemmas

D.1 Technical lemmas

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

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

Lemma D.2 (Dann et al. 2023, Lemma C.2)**.** *For any policies* π_1 *and* π_2*, and any function* $L : S \times A \to \mathbb{R}$*, define*

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

Then, for all (s,a)*,* $Q^{\pi_1}(s,a;\ell) = L(s,a)$*. Consequently,*

$$\sum_s q^{\pi_2}(s) \sum_a (\pi_1(a \mid s) - \pi_2(a \mid s)) L(s,a) = V^{\pi_1}(s_0;\ell) - V^{\pi_2}(s_0;\ell).$$

Lemma D.3 (Occupancy measure difference, Jin et al., 2021, Lemma D.3.1)**.** *For any transition functions* P_1, P_2 *and any policy* π*,*

$$\begin{aligned} q^{P_1,\pi}(s) - q^{P_2,\pi}(s) &= \sum_{h=0}^{h(s)-1} \sum_{u\in\mathcal{S}_h} \sum_{v\in\mathcal{A}} \sum_{w\in\mathcal{S}_{h+1}} q^{P_1,\pi}(u,v)(P_1(w \mid u,v) - P_2(w \mid u,v)) q^{P_2,\pi}(s \mid w) \\ &= \sum_{h=0}^{h(s)-1} \sum_{u\in\mathcal{S}_h} \sum_{v\in\mathcal{A}} \sum_{w\in\mathcal{S}_{h+1}} q^{P_2,\pi}(u,v)(P_1(w \mid u,v) - P_2(w \mid u,v)) q^{P_1,\pi}(s \mid w). \end{aligned}$$

Lemma D.4. *Suppose that the event* $\mathcal{E}$ *holds. Let* $\widetilde{P}_t \in \mathcal{P}_t$ *and* $F : \mathcal{S}_{h+1} \to [-\overline{F}, \overline{F}]$*. Then,*

$$\left|\sum_{s'\in S_{h+1}} \left(\widetilde{P}_t(s' \mid s,a) - P(s' \mid s,a)\right) F(s')\right| \lesssim \sqrt{\frac{|\mathcal{S}_{h+1}| \mathsf{Var}_{s'\sim P(\cdot|s,a)}[F(s')]\iota}{n_t(s,a)}} + \frac{|\mathcal{S}_{h+1}|\overline{F}\iota}{n_t(s,a)}$$

for all state-action pairs $(s,a) \in \mathcal{S}_h \times \mathcal{A}$.

Proof. Since $\sum_{s'\in S_{h+1}} \left(\widetilde{P}_t(s' \mid s,a) - P(s' \mid s,a)\right) = 0$, we may subtract any constant from F. Let

$$c = \mathbb{E}_{s'\sim P(\cdot|s,a)}[F(s')].$$

Then,

$$\sum_{s'\in S_{h+1}}\Big(\widetilde{P}_t(s'\mid s,a)-P(s'\mid s,a)\Big)F(s')=\sum_{s'\in S_{h+1}}\Big(\widetilde{P}_t(s'\mid s,a)-P(s'\mid s,a)\Big)(F(s')-c)$$

Therefore,

$$\begin{aligned}
&\left|\sum_{s'\in S_{h+1}}\Big(\widetilde{P}_t(s'\mid s,a)-P(s'\mid s,a)\Big)F(s')\right| \\
&\le \sum_{s'\in S_{h+1}}\left|\widetilde{P}_t(s'\mid s,a)-P(s'\mid s,a)\right||F(s')-c| \\
&\lesssim \sum_{s'\in S_{h+1}}\sqrt{\frac{P(s'\mid s,a)\iota}{n_t(s,a)}}|F(s')-c|+\sum_{s'\in S_{h+1}}\frac{\iota}{n_t(s,a)}|F(s')-c| \\
&\qquad\qquad\text{(by Lemma C.2 and the assumption that } \mathcal{E} \text{ holds)} \\
&\le \sqrt{\frac{|\mathcal{S}_{h+1}|\sum_{s'\in S_{h+1}}P(s'\mid s,a)(F(s')-c)^2\iota}{n_t(s,a)}}+\frac{2|\mathcal{S}_{h+1}|\overline{F}\iota}{n_t(s,a)} \\
&\qquad\qquad\text{(by the Cauchy–Schwarz inequality and } |F(s')-c|\le 2\overline{F}) \\
&\le \sqrt{\frac{|\mathcal{S}_{h+1}|\mathsf{Var}_{s'\sim P(\cdot|s,a)}[F(s')]\iota}{n_t(s,a)}}+\frac{2|\mathcal{S}_{h+1}|\overline{F}\iota}{n_t(s,a)}.
\end{aligned}$$

□

Lemma D.5. *Suppose that the event $\mathcal{E}$ holds. Let $\widetilde{P}_t$ be a transition kernel in $\mathcal{P}_t$, and let $g_t(s,a)\in[-G,G]$. Then, it holds that*

$$\left|\sum_t\big(V^{\pi_t}(s_0;g_t)-V^{\widetilde{P}_t,\pi_t}(s_0;g_t)\big)\right|\lesssim\sqrt{S^2A\iota^2Q^{\pi_{1:T}}_{\mathrm{trans}}(g)}+HS^2AG\iota^2.$$

Proof. For each layer $h\in\{0,\dots,H-1\}$, let $W_h:=\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}$. Throughout the proof, whenever summation indices are omitted, expressions such as $\sum_{u,v,w}$ or $\sum_{x,y,z}$ stand for $\sum_{h=0}^{H-1}\sum_{(s,a,s')\in W_h}$.

According to Lemma D.3, we have

$$\begin{aligned}
&V^{\pi_t}(s_0;g_t)-V^{\widetilde{P}_t,\pi_t}(s_0;g_t) \\
&=\sum_{s,a}\Big(q^{\pi_t}(s)-q^{\widetilde{P}_t,\pi_t}(s)\Big)\pi_t(a\mid s)g_t(s,a) \\
&=\sum_{s,a}\sum_{u,v,w}q^{\pi_t}(u,v)\Big(P(w\mid u,v)-\widetilde{P}_t(w\mid u,v)\Big)q^{\widetilde{P}_t,\pi_t}(s\mid w)\pi_t(a\mid s)g_t(s,a) \\
&=\sum_{s,a}\sum_{u,v,w}q^{\pi_t}(u,v)\Big(P(w\mid u,v)-\widetilde{P}_t(w\mid u,v)\Big)q^{\pi_t}(s\mid w)\pi_t(a\mid s)g_t(s,a) \\
&\quad+\sum_{s}\sum_{a}\sum_{u,v,w}q^{\pi_t}(u,v)\Big(P(w\mid u,v)-\widetilde{P}_t(w\mid u,v)\Big)\Big(q^{\widetilde{P}_t,\pi_t}(s\mid w)-q^{\pi_t}(s\mid w)\Big)\pi_t(a\mid s)g_t(s,a) \\
&=\underbrace{\sum_{s,a}\sum_{u,v,w}q^{\pi_t}(u,v)\Big(P(w\mid u,v)-\widetilde{P}_t(w\mid u,v)\Big)q^{\pi_t}(s\mid w)\pi_t(a\mid s)g_t(s,a)}_{\textbf{term}_1} \\
&\quad+\underbrace{\sum_{u,v,w}q^{\pi_t}(u,v)\Big(P(w\mid u,v)-\widetilde{P}_t(w\mid u,v)\Big)\Big(V^{\widetilde{P}_t,\pi_t}(w;g_t)-V^{\pi_t}(w;g_t)\Big)}_{\textbf{term}_2}.
\end{aligned}$$

We first bound $\mathbf{term}_1$.

$$\begin{aligned}
\mathbf{term}_1 &\le \sum_{u,v,w} q^{\pi_t}(u,v)\Big(P(w \mid u,v) - \widetilde{P}_t(w \mid u,v)\Big) \sum_{s,a} q^{\pi_t}(s,a \mid w) g_t(s,a) \\
&= \sum_{u,v,w} q^{\pi_t}(u,v)\Big(P(w \mid u,v) - \widetilde{P}_t(w \mid u,v)\Big) V^{\pi_t}(w; g_t) \\
&\le \sum_{u,v} q^{\pi_t}(u,v) \left| \sum_{w} \Big(P(w \mid u,v) - \widetilde{P}_t(w \mid u,v)\Big) V^{\pi_t}(w; g_t) \right| \\
&\lesssim \sum_{h=0}^{H-1} \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} q^{\pi_t}(u,v) \left(\sqrt{\frac{|\mathcal{S}_{h+1}| \mathsf{Var}_{w\sim P(\cdot|u,v)}[V^{\pi_t}(w; g_t)]\iota}{n_t(u,v)}} + \frac{|\mathcal{S}_{h+1}| H G \iota}{n_t(u,v)} \right) \\
&\qquad \text{(by Lemma D.4)} \\
&= \underbrace{\sum_{h=0}^{H-1} \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} q^{\pi_t}(u,v) \left(\sqrt{\frac{|\mathcal{S}_{h+1}| \mathsf{Var}_{w\sim P(\cdot|u,v)}[V^{\pi_t}(w; g_t)]\iota}{n_t(u,v)}} \right)}_{\mathbf{term}_{1a}} \\
&\quad + \underbrace{\sum_{h=0}^{H-1} |\mathcal{S}_{h+1}| H G \iota \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} \frac{q^{\pi_t}(u,v)}{n_t(u,v)}}_{\mathbf{term}_{1b}}
\end{aligned}$$

For $\mathbf{term}_{1a}$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbf{term}_{1a} &\le \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} q^{\pi_t}(u,v) \left(\sqrt{\frac{|\mathcal{S}_{h+1}| \mathsf{Var}_{w\sim P(\cdot|u,v)}[V^{\pi_t}(w; g_t)]\iota}{n_t(u,v)}} \right) \\
&\le \sqrt{\sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} |\mathcal{S}_{h+1}| \frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)}} \\
&\qquad \times \sqrt{\sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} q^{\pi_t}(u,v) \mathsf{Var}_{w\sim P(\cdot|u,v)}[V^{\pi_t}(w; g_t)]} \\
&\qquad \text{(by the Cauchy–Schwarz inequality)} \\
&\lesssim \sqrt{\left(\sum_{h=0}^{H-1} |\mathcal{S}_{h+1}||\mathcal{S}_h| A \iota^2 \right) \left(\sum_{t=1}^{T} \sum_{u,v} q^{\pi_t}(u,v) \mathsf{Var}_{w\sim P(\cdot|u,v)}[V^{\pi_t}(w; g_t)] \right)} \\
&\qquad \text{(by Lemma C.3 and the assumption that } \mathcal{E} \text{ holds)} \\
&\le \sqrt{S^2 A \iota^2 \left(\sum_{t=1}^{T} \sum_{u,v} q^{\pi_t}(u,v) \mathsf{Var}_{w\sim P(\cdot|u,v)}[V^{\pi_t}(w; g_t)] \right)}, \qquad (19)
\end{aligned}$$

where the last inequality uses $\sum_{h=0}^{H-1} |\mathcal{S}_{h+1}||\mathcal{S}_h| \le S^2$.

For $\mathbf{term}_{1b}$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbf{term}_{1b} &\le \sum_{t=1}^{T} \sum_{h=0}^{H-1} |\mathcal{S}_{h+1}| H G \iota \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} \frac{q^{\pi_t}(u,v)}{n_t(u,v)} \\
&\lesssim \sum_{h=0}^{H-1} |\mathcal{S}_h||\mathcal{S}_{h+1}| H A G \iota^2 \qquad \text{(by Lemma C.3 and the assumption that } \mathcal{E} \text{ holds)} \\
&\le H S^2 A G \iota^2. \qquad (20)
\end{aligned}$$

Thus, combining (19) and (20), we obtain

$$\sum_{t=1}^{T} \mathbf{term}_1 \lesssim \sqrt{S^2 A \iota^2 \left(\sum_{t=1}^{T} \sum_{u,v} q^{\pi_t}(u,v) \mathsf{Var}_{w \sim P(\cdot \mid u,v)}[V^{\pi_t}(w; g_t)] \right)} + HS^2 AG\iota^2.$$

We next bound $\mathbf{term}_2$. First, for $(x,y) \in \mathcal{S}_j \times \mathcal{A}$, we have

$$\begin{aligned}
\sum_z \left| P(z \mid x,y) - \widetilde{P}_t(z \mid x,y) \right| &\le 4\sqrt{\frac{\iota}{n_t(x,y)}} \sum_z \sqrt{P(z \mid x,y)} + \frac{40|\mathcal{S}_{j+1}|\iota}{3n_t(x,y)} \quad \text{(by Lemma C.2)} \\
&\le 4\sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}} + \frac{40|\mathcal{S}_{j+1}|\iota}{3n_t(x,y)}. \\
&\qquad \text{(by the Cauchy–Schwarz inequality)}
\end{aligned}$$

Combining this with $\sum_z \left| P(z \mid x,y) - \widetilde{P}_t(z \mid x,y) \right| \le \sum_z \left(P(z \mid x,y) + \widetilde{P}_t(z \mid x,y) \right) = 2$ gives

$$\begin{aligned}
\sum_z \left| P(z \mid x,y) - \widetilde{P}_t(z \mid x,y) \right| &\le \min\left\{ 2,\ 4\sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}} + \frac{40|\mathcal{S}_{j+1}|\iota}{3n_t(x,y)} \right\} \\
&\le 4\sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}} + \sqrt{\frac{80|\mathcal{S}_{j+1}|\iota}{3n_t(x,y)}} \lesssim \sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}}, \qquad (21)
\end{aligned}$$

where we used $\min\{2, a+b\} \le a + \min\{2,b\}$ and $\min\{2,b\} \le \sqrt{2b}$ for $a, b \ge 0$. Next, for $w \in \mathcal{S}_{h+1}$, Lemma D.3 gives

$$\begin{aligned}
&\left| V^{\widetilde{P}_t, \pi_t}(w; g_t) - V^{\pi_t}(w; g_t) \right| \\
&\le \sum_{j=h+1}^{H-1} \sum_{(x,y) \in \mathcal{S}_j \times \mathcal{A}} q^{\pi_t}(x,y \mid w) \sum_z \left| P(z \mid x,y) - \widetilde{P}_t(z \mid x,y) \right| \left| V^{\widetilde{P}_t, \pi_t}(z; g_t) \right| \\
&\le HG \sum_{j=h+1}^{H-1} \sum_{(x,y) \in \mathcal{S}_j \times \mathcal{A}} q^{\pi_t}(x,y \mid w) \sum_z \left| P(z \mid x,y) - \widetilde{P}_t(z \mid x,y) \right| \\
&\lesssim HG \sum_{j=h+1}^{H-1} \sum_{(x,y) \in \mathcal{S}_j \times \mathcal{A}} q^{\pi_t}(x,y \mid w) \sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}}.
\end{aligned}$$

Using this bound, we obtain

$$\begin{aligned}
|\mathbf{term}_2| &\le \sum_{u,v,w} q^{\pi_t}(u,v) \left| P(w \mid u,v) - \widetilde{P}_t(w \mid u,v) \right| \left| V^{\widetilde{P}_t, \pi_t}(w; g_t) - V^{\pi_t}(w; g_t) \right| \\
&\lesssim \sum_{u,v,w} q^{\pi_t}(u,v) \left(\sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} + \frac{\iota}{n_t(u,v)} \right) \left| V^{\widetilde{P}_t, \pi_t}(w; g_t) - V^{\pi_t}(w; g_t) \right| \\
&\qquad \text{(by Lemma C.2)} \\
&\lesssim HG \underbrace{\sum_{0 \le h < j \le H-1} \sum_{(u,v,w) \in W_h} \sum_{(x,y) \in \mathcal{S}_j \times \mathcal{A}} q^{\pi_t}(u,v) \sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} q^{\pi_t}(x,y \mid w) \sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}}}_{\mathbf{term}_{2a}} \\
&\quad + HG \underbrace{\sum_{h=0}^{H-1} |\mathcal{S}_{h+1}| \sum_{(u,v) \in \mathcal{S}_h \times \mathcal{A}} \frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)}}_{\mathbf{term}_{2b}}.
\end{aligned}$$

For $\mathbf{term}_{2a}$, applying the Cauchy–Schwarz inequality for each pair $h < j$ before summing over the layers gives

$$\begin{aligned}
\sum_{t=1}^{T} \mathbf{term}_{2a} &\le \sum_{0\le h<j\le H-1} \sqrt{\sum_{t=1}^{T} \sum_{\substack{(u,v,w)\in W_h \\ (x,y)\in \mathcal{S}_j\times\mathcal{A}}} \frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)} q^{\pi_t}(x,y \mid w)} \\
&\quad \times \sqrt{\sum_{t=1}^{T} \sum_{\substack{(u,v,w)\in W_h \\ (x,y)\in \mathcal{S}_j\times\mathcal{A}}} q^{\pi_t}(u,v) P(w \mid u,v) q^{\pi_t}(x,y \mid w) \frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}} \\
&\le \sum_{0\le h<j\le H-1} \sqrt{|\mathcal{S}_{h+1}| \sum_{t=1}^{T} \sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}} \frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)}} \sqrt{|\mathcal{S}_{j+1}| \sum_{t=1}^{T} \sum_{(x,y)\in\mathcal{S}_j\times\mathcal{A}} \frac{q^{\pi_t}(x,y)\iota}{n_t(x,y)}} \\
&\lesssim A\iota^2 \sum_{0\le h<j\le H-1} \sqrt{|\mathcal{S}_h||\mathcal{S}_{h+1}||\mathcal{S}_j||\mathcal{S}_{j+1}|} \\
&\le A\iota^2 \left(\sum_{h=0}^{H-1} \sqrt{|\mathcal{S}_h||\mathcal{S}_{h+1}|} \right)^2 \\
&\le A\iota^2 \left(\frac{1}{2} \sum_{h=0}^{H-1} (|\mathcal{S}_h| + |\mathcal{S}_{h+1}|) \right)^2 \qquad \text{(by the AM–GM inequality)} \\
&= S^2 A\iota^2, \qquad (22)
\end{aligned}$$

where the second inequality follows from $\sum_{(x,y)\in\mathcal{S}_j\times\mathcal{A}} q^{\pi_t}(x,y \mid w) \le 1$ and $\sum_{(u,v,w)\in W_h} q^{\pi_t}(u,v) P(w \mid u,v) q^{\pi_t}(x,y \mid w) = q^{\pi_t}(x,y)$ for $j > h$, and the third inequality follows from Lemma C.3. For $\mathbf{term}_{2b}$, Lemma C.3 yields

$$\sum_{t=1}^{T} \mathbf{term}_{2b} \lesssim A\iota^2 \sum_{h=0}^{H-1} |\mathcal{S}_h||\mathcal{S}_{h+1}| \le S^2 A\iota^2. \qquad (23)$$

Combining (22) and (23) , we obtain

$$\sum_{t=1}^{T} |\mathbf{term}_2| \lesssim HS^2 AG\iota^2.$$

Combining the bounds for $\mathbf{term}_1$ and $\mathbf{term}_2$ and using the definition of $Q^{\pi_{1:T}}_{\mathrm{trans}}(g)$, we obtain

$$\sum_{t=1}^{T} \big(V^{\pi_t}(s_0; g_t) - V^{\widetilde{P}_t, \pi_t}(s_0; g_t) \big) \lesssim \sqrt{S^2 A\iota^2 Q^{\pi_{1:T}}_{\mathrm{trans}}(g)} + HS^2 AG\iota^2.$$

Applying the same bound to $-g_t$ completes the proof. □

Lemma D.6. *For arbitrary signed functions $g_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,*

$$Q^{\pi_{1:T}}_{\mathrm{trans}}(g) \le H \sum_{t,s,a} q^{\pi_t}(s,a) g_t(s,a)^2.$$

Proof. It suffices to prove the following inequality for each t:

$$\sum_{s,a} q^{\pi_t}(s,a) \mathsf{Var}_{s'\sim P(\cdot|s,a)}[V^{\pi_t}(s'; g_t)] \le H \sum_{s,a} q^{\pi_t}(s,a) g_t(s,a)^2.$$

Let $(X_{t,0}, A_{t,0}, X_{t,1}, A_{t,1}, \ldots, X_{t,H})$ be the random trajectory generated by (P, π_t). For $h = 0, 1, \ldots, H$ and $x \in \mathcal{S}_h$, define

$$\mathcal{G}_{t,h} \coloneqq \sum_{h'=h}^{H-1} g_t(X_{t,h'}, A_{t,h'}), \quad \mathcal{V}_{t,h}(x) \coloneqq \mathbb{E}_{\pi_t}[\mathcal{G}_{t,h} \mid X_{t,h} = x], \quad \mathcal{M}_{t,h}(x) \coloneqq \mathbb{E}_{\pi_t}[\mathcal{G}^2_{t,h} \mid X_{t,h} = x].$$

In particular, $\mathcal{V}_{t,H} = \mathcal{M}_{t,H} = 0$.

Since $\mathcal{G}_{t,h} = g_t(X_{t,h}, A_{t,h}) + \mathcal{G}_{t,h+1}$, we have for any state x

$$\mathcal{V}_{t,h}(x) = \sum_a \pi_t(a \mid x)\left(g_t(x,a) + \sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right),$$

and

$$\begin{aligned}\mathcal{M}_{t,h}(x) = \sum_a \pi_t(a \mid x)\left(g_t(x,a)^2 + 2g_t(x,a)\sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right) \\ + \sum_a \pi_t(a \mid x)\sum_{x'} P(x' \mid x,a)\mathcal{M}_{t,h+1}(x').\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mathcal{M}_{t,h}(x) - \mathcal{V}_{t,h}(x)^2 \\
&= \sum_a \pi_t(a \mid x)\left(g_t(x,a)^2 + 2g_t(x,a)\sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x') + \sum_{x'} P(x' \mid x,a)\mathcal{M}_{t,h+1}(x')\right) \\
&\quad - \left(\sum_a \pi_t(a \mid x)\left(g_t(x,a) + \sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right)\right)^2 \\
&= \sum_a \pi_t(a \mid x)\sum_{x'} P(x' \mid x,a)\big(\mathcal{M}_{t,h+1}(x') - \mathcal{V}_{t,h+1}(x')^2\big) \\
&\quad + \sum_a \pi_t(a \mid x)\left(\sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')^2 - \left(\sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right)^2\right) \\
&\quad + \sum_a \pi_t(a \mid x)\left(g_t(x,a) + \sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right)^2 \\
&\quad - \left(\sum_a \pi_t(a \mid x)\left(g_t(x,a) + \sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right)\right)^2 \\
&\geq \sum_a \pi_t(a \mid x)\sum_{x'} P(x' \mid x,a)\big(\mathcal{M}_{t,h+1}(x') - \mathcal{V}_{t,h+1}(x')^2\big) \\
&\quad + \sum_a \pi_t(a \mid x)\mathsf{Var}_{x' \sim P(\cdot \mid x,a)}[\mathcal{V}_{t,h+1}(x')],
\end{aligned}$$

where the last inequality follows from Jensen's inequality

$$\begin{aligned}\left(\sum_a \pi_t(a \mid x)\left(g_t(x,a) + \sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right)\right)^2 \\ \leq \sum_a \pi_t(a \mid x)\left(g_t(x,a) + \sum_{x'} P(x' \mid x,a)\mathcal{V}_{t,h+1}(x')\right)^2.\end{aligned}$$

Thus, we have

$$\begin{aligned}\sum_a \pi_t(a \mid x)\mathsf{Var}_{x' \sim P(\cdot \mid x,a)}[\mathcal{V}_{t,h+1}(x')] \leq \big(\mathcal{M}_{t,h}(x) - \mathcal{V}_{t,h}(x)^2\big) \\ - \sum_a \pi_t(a \mid x)\sum_{x'} P(x' \mid x,a)\big(\mathcal{M}_{t,h+1}(x') - \mathcal{V}_{t,h+1}(x')^2\big).\end{aligned}$$

Multiplying by $q^{\pi_t}(x)$ and summing over $x \in \mathcal{S}_h$, we obtain

$$\sum_{(x,a)\in\mathcal{S}_h\times\mathcal{A}} q^{\pi_t}(x,a)\mathsf{Var}_{x'\sim P(\cdot|x,a)}[\mathcal{V}_{t,h+1}(x')] \le \sum_{x\in\mathcal{S}_h} q^{\pi_t}(x)\big(\mathcal{M}_{t,h}(x) - \mathcal{V}_{t,h}(x)^2\big) - \sum_{x'\in\mathcal{S}_{h+1}} q^{\pi_t}(x')\big(\mathcal{M}_{t,h+1}(x') - \mathcal{V}_{t,h+1}(x')^2\big).$$

Summing over $h = 0, 1, \dots, H-1$ and using $\mathcal{V}_{t,h+1}(x') = V^{\pi_t}(x'; g_t)$, we obtain

$$\sum_{s,a} q^{\pi_t}(s,a)\mathsf{Var}_{s'\sim P(\cdot|s,a)}[V^{\pi_t}(s'; g_t)] \le \mathcal{M}_{t,0}(s_0) - \mathcal{V}_{t,0}(s_0)^2 \le \mathcal{M}_{t,0}(s_0). \tag{24}$$

Finally,

$$\begin{aligned}\mathcal{M}_{t,0}(s_0) &= \mathbb{E}_{\pi_t}\left[\left(\sum_{h=0}^{H-1} g_t(X_{t,h}, A_{t,h})\right)^2\right] \\ &\le H\mathbb{E}_{\pi_t}\left[\sum_{h=0}^{H-1} g_t(X_{t,h}, A_{t,h})^2\right] = H\sum_{s,a} q^{\pi_t}(s,a) g_t(s,a)^2, \end{aligned} \tag{25}$$

where the inequality follows from the Cauchy–Schwarz inequality. Combining (24) and (25) completes the proof. □

From the proof of Lemma D.8, we obtain the following bound, whose lower-order term improves that of Dann et al. (2023, Lemma C.4).

Lemma D.7. *Suppose that the event $\mathcal{E}$ 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*

$$\sum_{t=1}^{T}\sum_{s}\left|q^{\pi_t}(s) - q^{\widetilde{P}_t^s,\pi_t}(s)\right| g_t(s) \lesssim \sqrt{HS^2A\iota^2\sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s) g_t(s)^2} + S^3AG\iota^2.$$

Lemma D.8. *Suppose that the event $\mathcal{E}$ holds. Let $\widetilde{P}_t^{s,\widetilde{s}}$ be a transition kernel in $\mathcal{P}_t$ which may depend on $s, \widetilde{s}$, and let $g_t(s) \in [0, G]$. Then*

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{\widetilde{s}} q^{\pi_t}(\widetilde{s})\sum_{s}\left|q^{\pi_t}(s \mid \widetilde{s}) - q^{\widetilde{P}_t^{s,\widetilde{s}},\pi_t}(s \mid \widetilde{s})\right| g_t(s) \\ &\lesssim \sqrt{H^3S^2A\iota^2\sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s) g_t(s)^2} + HS^3AG\iota^2.\end{aligned}$$

Proof. For each layer $h \in \{0, \dots, H-1\}$, let $W_h := \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$. Throughout the proof, whenever summation indices are omitted, expressions such as $\sum_{u,v,w}$ or $\sum_{x,y,z}$ stand for $\sum_{h=0}^{H-1}\sum_{(s,a,s')\in W_h}$.

By Lemma D.3, for any $s, \widetilde{s}$ and policy π,

$$\begin{aligned}&q^{\pi}(s \mid \widetilde{s}) - q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid \widetilde{s}) \\ &= \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s})\Big(P(w \mid u,v) - \widetilde{P}_t^{s,\widetilde{s}}(w \mid u,v)\Big) q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid w) \\ &= \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s})\Big(P(w \mid u,v) - \widetilde{P}_t^{s,\widetilde{s}}(w \mid u,v)\Big) q^{\pi}(s \mid w) \\ &\quad + \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s})\Big(P(w \mid u,v) - \widetilde{P}_t^{s,\widetilde{s}}(w \mid u,v)\Big)\Big(q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid w) - q^{\pi}(s \mid w)\Big).\end{aligned}$$

Taking absolute values, multiplying by $g_t(s)$, and summing over s, we obtain

$$\begin{aligned}
&\sum_s \left| q^{\pi}(s \mid \widetilde{s}) - q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid \widetilde{s}) \right| g_t(s) \\
&\lesssim \sum_s \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s}) \left(\sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} + \frac{\iota}{n_t(u,v)} \right) q^{\pi}(s \mid w) g_t(s) \\
&\quad + \sum_s \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s}) \left(\sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} + \frac{\iota}{n_t(u,v)} \right) \left| q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid w) - q^{\pi}(s \mid w) \right| g_t(s) \\
&\qquad\qquad \text{(by Lemma C.2)} \\
&\lesssim \sum_s \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s}) \sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} q^{\pi}(s \mid w) g_t(s) \\
&\quad + SG \sum_{h=0}^{H-1} |\mathcal{S}_{h+1}| \sum_{(u,v) \in \mathcal{S}_h \times \mathcal{A}} \frac{q^{\pi}(u,v \mid \widetilde{s})\iota}{n_t(u,v)} \\
&\quad + \sum_s \sum_{u,v,w} q^{\pi}(u,v \mid \widetilde{s}) \sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} \left| q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid w) - q^{\pi}(s \mid w) \right| g_t(s), \qquad (26)
\end{aligned}$$

where we used $q^{\pi}(s \mid w) + \left| q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid w) - q^{\pi}(s \mid w) \right| \le 2$ and $g_t(s) \le G$. For $w \in \mathcal{S}_{h+1}$, we have

$$\begin{aligned}
&\left| q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid w) - q^{\pi}(s \mid w) \right| \\
&\le \sum_{j=h+1}^{H-1} \sum_{(x,y) \in \mathcal{S}_j \times \mathcal{A}} q^{\pi}(x,y \mid w) \sum_z \left| P(z \mid x,y) - \widetilde{P}_t^{s,\widetilde{s}}(z \mid x,y) \right| q^{\widetilde{P}_t^{s,\widetilde{s}},\pi}(s \mid z) \qquad \text{(by Lemma D.3)} \\
&\lesssim \sum_{j=h+1}^{H-1} \sum_{(x,y) \in \mathcal{S}_j \times \mathcal{A}} q^{\pi}(x,y \mid w) \sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}}. \qquad \text{(by (21))}
\end{aligned}$$

Combining these bounds gives

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{\widetilde{s}} q^{\pi_t}(\widetilde{s}) \sum_s \left| q^{\pi_t}(s \mid \widetilde{s}) - q^{\widetilde{P}_t^{s,\widetilde{s}},\pi_t}(s \mid \widetilde{s}) \right| g_t(s) \\
&\lesssim \sum_{t=1}^{T} \sum_s \sum_{u,v,w} \left(\sum_{\widetilde{s}} q^{\pi_t}(\widetilde{s}) q^{\pi_t}(u,v \mid \widetilde{s}) \right) \sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} q^{\pi_t}(s \mid w) g_t(s) \\
&\quad + SG \sum_{t=1}^{T} \sum_{h=0}^{H-1} |\mathcal{S}_{h+1}| \sum_{(u,v) \in \mathcal{S}_h \times \mathcal{A}} \left(\sum_{\widetilde{s}} q^{\pi_t}(\widetilde{s}) q^{\pi_t}(u,v \mid \widetilde{s}) \right) \frac{\iota}{n_t(u,v)} \\
&\quad + SG \sum_{t=1}^{T} \sum_{0 \le h < j \le H-1} \sum_{\substack{(u,v,w) \in W_h \\ (x,y) \in \mathcal{S}_j \times \mathcal{A}}} \left(\sum_{\widetilde{s}} q^{\pi_t}(\widetilde{s}) q^{\pi_t}(u,v \mid \widetilde{s}) \right) \sqrt{\frac{P(w \mid u,v)\iota}{n_t(u,v)}} q^{\pi_t}(x,y \mid w) \sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}} \\
&\qquad\qquad \text{(by (26) and } \textstyle\sum_s g_t(s) \le SG\text{)}
\end{aligned}$$

$$\lesssim \underbrace{H\sum_{t=1}^{T}\sum_{s}\sum_{u,v,w} q^{\pi_t}(u,v)\sqrt{\frac{P(w\mid u,v)\iota}{n_t(u,v)}}q^{\pi_t}(s\mid w)g_t(s)}_{\textbf{term}_1}$$
$$+ \underbrace{HSG\sum_{t=1}^{T}\sum_{h=0}^{H-1}|\mathcal{S}_{h+1}|\sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}}\frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)}}_{\textbf{term}_2}$$
$$+ \underbrace{HSG\sum_{t=1}^{T}\sum_{0\le h<j\le H-1}\sum_{\substack{(u,v,w)\in W_h\\(x,y)\in\mathcal{S}_j\times\mathcal{A}}} q^{\pi_t}(u,v)\sqrt{\frac{P(w\mid u,v)\iota}{n_t(u,v)}}q^{\pi_t}(x,y\mid w)\sqrt{\frac{|\mathcal{S}_{j+1}|\iota}{n_t(x,y)}}}_{\textbf{term}_3}.$$

(by $\sum_{\widetilde{s}} q^{\pi_t}(\widetilde{s})q^{\pi_t}(u,v\mid\widetilde{s}) \le Hq^{\pi_t}(u,v)$)

We first bound $\textbf{term}_1$. Fix h. Then

$$H\sum_{t=1}^{T}\sum_{s}\left(\sum_{(u,v,w)\in W_h} q^{\pi_t}(u,v)\sqrt{\frac{P(w\mid u,v)\iota}{n_t(u,v)}}q^{\pi_t}(s\mid w)\right)g_t(s)$$
$$\le H\sum_{t=1}^{T}\sum_{s}\left(\sum_{(u,v,w)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}} q^{\pi_t}(u,v)\left(\alpha P(w\mid u,v)g_t(s)^2+\frac{\iota}{\alpha n_t(u,v)}\right)q^{\pi_t}(s\mid w)\right)$$

(holds for any $\alpha>0$ by the AM-GM inequality)

$$= \alpha H\sum_{t=1}^{T}\sum_{s}\sum_{(u,v,w)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}} q^{\pi_t}(u,v)P(w\mid u,v)q^{\pi_t}(s\mid w)g_t(s)^2$$
$$+\frac{H}{\alpha}\sum_{t=1}^{T}\sum_{s}\sum_{(u,v,w)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}}\frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)}q^{\pi_t}(s\mid w)$$
$$\le \alpha H\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)g_t(s)^2+\frac{H^2|\mathcal{S}_{h+1}|}{\alpha}\sum_{t=1}^{T}\sum_{(u,v)\in\mathcal{S}_h\times\mathcal{A}}\frac{q^{\pi_t}(u,v)\iota}{n_t(u,v)}$$
$$\lesssim \alpha H\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)g_t(s)^2+\frac{H^2|\mathcal{S}_{h+1}||\mathcal{S}_h|A\iota^2}{\alpha}$$

(by Lemma C.3 and the assumption that $\mathcal{E}$ holds)

$$\lesssim \sqrt{H^3|\mathcal{S}_h||\mathcal{S}_{h+1}|A\iota^2\sum_{t=1}^{T}\sum_{s}q^{P,\pi_t}(s)g_t(s)^2} \qquad \text{(by picking the optimal } \alpha\text{)}$$
$$\le (|\mathcal{S}_h|+|\mathcal{S}_{h+1}|)\sqrt{H^3A\iota^2\sum_{t=1}^{T}\sum_{s}q^{P,\pi_t}(s)g_t(s)^2}. \qquad \text{(by the AM–GM inequality)}$$

Summing over h gives

$$\sum_{h=0}^{H-1}(|\mathcal{S}_h|+|\mathcal{S}_{h+1}|)\sqrt{H^3A\iota^2\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)g_t(s)^2}\lesssim\sqrt{H^3S^2A\iota^2\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)g_t(s)^2}. \quad (27)$$

For $\textbf{term}_2$ and $\textbf{term}_3$, applying (23) and (22), respectively, gives

$$\textbf{term}_2+\textbf{term}_3\lesssim HSG\cdot S^2A\iota^2=HS^3AG\iota^2. \quad (28)$$

Combining (27) and (28) proves Lemma D.8. For Lemma D.7, set $\widetilde{s}=s_0$ in (26) and sum over t. The same estimates, with the prefactor H in each of $\textbf{term}_1$, $\textbf{term}_2$, and $\textbf{term}_3$ omitted, prove Lemma D.7. □

D.2 Data-dependent and self-bounding lemmas

Lemma D.9. *It holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)\right] \lesssim H\mathbb{E}\left[\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2)\right] + HS^3A\iota^2.$$

Proof. On the event $\bar{\mathcal{E}}$, $\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2) = \mathcal{O}(H^2T)$ gives a contribution $\mathcal{O}(H^2T\delta)$, which is absorbed by taking $\delta = 1/T^2$.

Hence, it suffices to prove the bound on $\mathcal{E}$. For each t, let $\overline{P}_t^Q \in \mathcal{P}_t$ be such that

$$Q^{\overline{P}_t^Q,\pi_t}(s,a;(\ell_t-m_t)^2) = \max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)$$

for all (s,a). Then,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)Q^{\overline{P}_t^Q,\pi_t}(s,a;(\ell_t-m_t)^2)\\
&\le \sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)Q^{\pi_t}(s,a;(\ell_t-m_t)^2)\\
&\quad + \sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\Big(Q^{\overline{P}_t^Q,\pi_t}(s,a;(\ell_t-m_t)^2) - Q^{\pi_t}(s,a;(\ell_t-m_t)^2)\Big). \qquad (29)
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)Q^{\pi_t}(s,a;(\ell_t-m_t)^2)\\
&= \sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\sum_{s',a'} q^{\pi_t}(s',a'\mid s,a)(\ell_t(s',a')-m_t(s',a'))^2\\
&\le H\sum_{t=1}^{T}\sum_{s',a'} q^{\pi_t}(s',a')(\ell_t(s',a')-m_t(s',a'))^2\\
&= H\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2) \qquad (30)
\end{aligned}$$

For the second term, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\Big(Q^{\overline{P}_t^Q,\pi_t}(s,a;(\ell_t-m_t)^2) - Q^{\pi_t}(s,a;(\ell_t-m_t)^2)\Big)\\
&= \sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\sum_{s'}\Big(q^{\overline{P}_t^Q,\pi_t}(s'\mid s,a) - q^{\pi_t}(s'\mid s,a)\Big)\sum_{a'}\pi_t(a'\mid s')(\ell_t(s',a')-m_t(s',a'))^2\\
&= \sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s)\sum_{s'}\Big(q^{\overline{P}_t^Q,\pi_t}(s'\mid s) - q^{\pi_t}(s'\mid s)\Big)\sum_{a'}\pi_t(a'\mid s')(\ell_t(s',a')-m_t(s',a'))^2\\
&\lesssim \sqrt{H^3S^2A\iota^2\sum_{t=1}^{T}\sum_{s',a'} q^{\pi_t}(s',a')(\ell_t(s',a')-m_t(s',a'))^2} + HS^3A\iota^2 \qquad \text{(by Lemma D.8)}\\
&= \sqrt{H^3S^2A\iota^2\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2)} + HS^3A\iota^2. \qquad (31)
\end{aligned}$$

Combining (29)–(31), we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a) Q^{\overline{P}_t^Q,\pi_t}(s,a;(\ell_t-m_t)^2) \\
&\lesssim H\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2) + \sqrt{H^3S^2A\iota^2\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2) + HS^3A\iota^2} \\
&\lesssim H\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2) + HS^3A\iota^2,
\end{aligned}$$

where the last line follows from the AM–GM inequality since $ax+\sqrt{bx} \le ax + ax + \frac{b}{2a}$ for $a>0$ and $b, x \ge 0$. Taking expectations completes the proof. □

Lemma D.10 (Li et al. 2026, Lemma E.9)**.** *It holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\mathbb{I}_t(s,a)(L_{t,h(s)}-M_{t,h(s)})^2\right] \le H^2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\mathbb{I}_t(s,a)(\ell_t(s,a)-m_t(s,a))^2\right].$$

Lemma D.11 (Li et al. 2026, Lemma F.12)**.** *Suppose m_t is defined in (4) with some constant $\xi \in (0, \frac{1}{2})$. Then, for any comparator policy π, it holds that*

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\mathbb{I}_t(s,a)(\ell_t(s,a)-m_t(s,a))^2\right] \\
&\lesssim \min\{L(\pi)+\mathrm{Reg}_T(\pi), HT-L(\pi)-\mathrm{Reg}_T(\pi), Q_\infty, V_1\} + SA.
\end{aligned}$$

Lemma D.12. *For any sequence of policies $\{\pi_t\}_{t=1}^T$ and any prediction sequence $\{m_t\}_{t=1}^T$, it holds that*

$$Q_{\mathrm{trans}}^{\pi_{1:T}}(m) \le 2Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) + 2H\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2).$$

Proof. For any state s,

$$V^{\pi_t}(s;m_t) = V^{\pi_t}(s;\ell_t) + V^{\pi_t}(s;m_t-\ell_t).$$

By using $\mathsf{Var}(X+Y) \le 2\mathsf{Var}(X) + 2\mathsf{Var}(Y)$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\mathsf{Var}_{s'\sim P(\cdot|s,a)}[V^{\pi_t}(s';m_t)] \\
&\le 2\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\mathsf{Var}_{s'\sim P(\cdot|s,a)}[V^{\pi_t}(s';\ell_t)] + 2\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\mathsf{Var}_{s'\sim P(\cdot|s,a)}[V^{\pi_t}(s';m_t-\ell_t)] \\
&\le 2Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) + 2Q_{\mathrm{trans}}^{\pi_{1:T}}(m-\ell).
\end{aligned}$$

Thus, we have

$$Q_{\mathrm{trans}}^{\pi_{1:T}}(m) \le 2Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) + 2Q_{\mathrm{trans}}^{\pi_{1:T}}(m-\ell).$$

Finally, Lemma D.6 with $g_t = m_t - \ell_t$ gives

$$Q_{\mathrm{trans}}^{\pi_{1:T}}(m-\ell) \le H\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2).$$

Therefore,

$$Q_{\mathrm{trans}}^{\pi_{1:T}}(m) \le 2Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell) + 2H\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t-m_t)^2).$$

□

Lemma D.13 (Dann et al. (2023, Lemma C.5)). *For any two policies π_1 and π_2, it holds that*

$$\sum_{s,a}|q^{\pi_1}(s,a)-q^{\pi_2}(s,a)| \le H\sum_{s,a} q^{\pi_1}(s)|\pi_1(a\mid s)-\pi_2(a\mid s)|.$$

Lemma D.14. *For any policy π and layer h, it holds that*

$$\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left|q^{\pi}(s,a)-q^{\pi^\star}(s,a)\right| \le 2\sum_{h'=0}^{h}\sum_{s\in\mathcal{S}_{h'}}\sum_{a\neq\pi^\star(s)} q^{\pi}(s,a).$$

Proof. Since $\pi^\star$ is deterministic, we have

$$\begin{aligned}
&\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left|q^{\pi}(s,a)-q^{\pi^\star}(s,a)\right| \\
&= \sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left|q^{\pi}(s)\pi(a\mid s)-q^{\pi^\star}(s)\mathbb{1}[a=\pi^\star(s)]\right| \\
&\le \sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}|q^{\pi}(s)\pi(a\mid s)-q^{\pi}(s)\mathbb{1}[a=\pi^\star(s)]| + \sum_{s\in\mathcal{S}_h}\left|q^{\pi}(s)-q^{\pi^\star}(s)\right| \\
&\qquad\qquad\qquad

Combining (32) and (34) yields

$$\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left|q^{\pi}(s,a)-q^{\pi^\star}(s,a)\right| \le 2\sum_{h'=0}^{h}\sum_{s\in\mathcal{S}_{h'}}\sum_{a\neq\pi^\star(s)} q^{\pi}(s,a).$$

This completes the proof. □

Lemma D.15 (Jin et al. 2020, Section 2.1)**.** *Under the stochastic regime with adversarial corruption, for any sequence of policies $\{\pi_t\}_{t=1}^T$, the regret satisfies the following $(\Delta, 2\mathcal{C}, T)$ self-bounding constraint:*

$$\mathrm{Reg}_T(\mathring{\pi}) \ge \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} q^{\pi_t}(s,a)\Delta(s,a)\right] - 2\mathcal{C}.$$

Lemma D.16. *Let $G : S\times A\to\mathbb{R}_{\ge 0}$ be any nonnegative function. Under the stochastic regime with adversarial corruption, for any $\alpha>0$, it holds that*

$$\sum_{s}\sum_{a\neq\pi^\star(s)} G(s,a)\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} q^{\pi_t}(s,a)\right]} \le \alpha(\mathrm{Reg}_T(\mathring{\pi})+2\mathcal{C}) + \sum_{s}\sum_{a\neq\pi^\star(s)}\frac{G(s,a)^2}{4\alpha\Delta(s,a)}.$$

Proof. For each (s,a) with $a\neq\pi^\star(s)$, the AM–GM inequality and Lemma D.15 imply

$$\begin{aligned}
&\sum_{s}\sum_{a\neq\pi^\star(s)} G(s,a)\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} q^{\pi_t}(s,a)\right]} \\
&\le \alpha\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} q^{\pi_t}(s,a)\Delta(s,a)\right] + \sum_{s}\sum_{a\neq\pi^\star(s)}\frac{G(s,a)^2}{4\alpha\Delta(s,a)} \\
&\qquad\text{(by the AM–GM inequality with } \alpha>0) \\
&\le \alpha(\mathrm{Reg}_T(\mathring{\pi})+2\mathcal{C}) + \sum_{s}\sum_{a\neq\pi^\star(s)}\frac{G(s,a)^2}{4\alpha\Delta(s,a)}. \qquad\text{(by Lemma D.15)}
\end{aligned}$$

□

Lemma D.17. *Let $G(h)$ be any nonnegative function defined on layer h. Under the stochastic regime with adversarial corruption, for any $\alpha>0$, it holds that*

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sqrt{G(h)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left[q^{\pi_t}(s,a)-q^{\mathring{\pi}}(s,a)\right]_+\right]} \\
&\qquad\le \alpha(\mathrm{Reg}_T(\mathring{\pi})+4\mathcal{C}) + \frac{\left(\sum_{h=0}^{H-1}\sqrt{G(h)}\right)^2}{4\alpha\Delta_{\min}}.
\end{aligned}$$

Proof. Applying Lemma D.14, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left[q^{\pi_t}(s,a)-q^{\mathring{\pi}}(s,a)\right]_+\right] \\
&= \mathbb{E}\left[\frac{1}{2}\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left|q^{\pi_t}(s,a)-q^{\mathring{\pi}}(s,a)\right|\right] \\
&\le \mathbb{E}\left[\frac{1}{2}\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\left(\left|q^{\pi_t}(s,a)-q^{\pi^\star}(s,a)\right|+\left|q^{\mathring{\pi}}(s,a)-q^{\pi^\star}(s,a)\right|\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h'=0}^{h}\sum_{s\in\mathcal{S}_{h'}}\sum_{a\neq\pi^\star(s)}\left(q^{\pi_t}(s,a)+q^{\mathring{\pi}}(s,a)\right)\right].
\end{aligned}$$

We next bound the comparator contribution. Using the uncorrupted mean loss function μ, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} q^{\mathring{\pi}}(s,a)\Delta(s,a)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\Big(V^{\mathring{\pi}}(s_0;\mu) - V^{\pi^\star}(s_0;\mu)\Big)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\big(V^{\mathring{\pi}}(s_0;\mu) - V^{\mathring{\pi}}(s_0;\ell_t)\big)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\Big(V^{\mathring{\pi}}(s_0;\ell_t) - V^{\pi^\star}(s_0;\ell_t)\Big)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\Big(V^{\pi^\star}(s_0;\ell_t) - V^{\pi^\star}(s_0;\mu)\Big)\right].
\end{aligned}$$

The middle term is nonpositive by the definition of $\mathring{\pi}$. Moreover, since $\mathbb{E}[\ell'_t] = \mu$, for any fixed policy π,

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{t=1}^{T}\big(V^{\pi}(s_0;\ell_t) - V^{\pi}(s_0;\mu)\big)\right]\right| &= \left|\mathbb{E}\left[\sum_{t=1}^{T}\big(V^{\pi}(s_0;\ell_t) - V^{\pi}(s_0;\ell'_t)\big)\right]\right| \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\|\ell_t(h) - \ell'_t(h)\|_\infty\right] = \mathcal{C}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} q^{\mathring{\pi}}(s,a)\Delta(s,a)\right] \leq 2\mathcal{C}. \tag{35}$$

Combining (35) with Lemma D.15, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} \big(q^{\pi_t}(s,a) + q^{\mathring{\pi}}(s,a)\big)\Delta(s,a)\right] \leq \mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}.$$

Then, for any $\alpha > 0$, we have

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sqrt{G(h)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)]_+\right]} \\
&\leq \sum_{h=0}^{H-1}\sqrt{G(h)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)}(q^{\pi_t}(s,a) + q^{\mathring{\pi}}(s,a))\right]} \\
&\leq \sum_{h=0}^{H-1}\sqrt{G(h)}\sqrt{\frac{1}{\Delta_{\min}}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)}(q^{\pi_t}(s,a) + q^{\mathring{\pi}}(s,a))\Delta(s,a)\right]} \\
&\leq \sum_{h=0}^{H-1}\sqrt{G(h)}\sqrt{\frac{\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}}{\Delta_{\min}}} \\
&\leq \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}) + \frac{\left(\sum_{h=0}^{H-1}\sqrt{G(h)}\right)^2}{4\alpha\Delta_{\min}},
\end{aligned}$$

where we use the AM–GM inequality. □

Lemma D.18. *Let $G > 0$. Under the stochastic regime with adversarial corruption, for any $\alpha > 0$, it holds that*

$$\sqrt{G\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}[q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)]_+\right]} \le \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}) + \frac{HG}{4\alpha\Delta_{\min}}.$$

Proof. Applying Lemma D.13 to compare any policy π with $\pi^\star$, we get

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{s,a}\left|q^{\pi}(s,a) - q^{\pi^\star}(s,a)\right| &\le H\sum_{t=1}^{T}\sum_{s,a} q^{\pi}(s)|\pi(a \mid s) - \pi^\star(a \mid s)| \\
&\le H\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} q^{\pi}(s)\pi(a \mid s) \\
&\quad + H\sum_{t=1}^{T}\sum_{s} q^{\pi}(s)(1 - \pi(\pi^\star(s) \mid s)) \\
&\le 2H\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)} q^{\pi}(s)\pi(a \mid s).
\end{aligned}$$

Using this bound with $\pi = \pi_t$ and $\pi = \mathring{\pi}$, for any $\alpha > 0$ we obtain

$$\begin{aligned}
&\sqrt{G\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}[q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)]_+\right]} \\
&= \sqrt{\frac{1}{2}G\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}|q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)|\right]} \\
&\le \sqrt{\frac{1}{2}G\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}(|q^{\pi_t}(s,a) - q^{\pi^\star}(s,a)| + |q^{\mathring{\pi}}(s,a) - q^{\pi^\star}(s,a)|)\right]} \\
&\le \sqrt{HG\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)}(q^{\pi_t}(s,a) + q^{\mathring{\pi}}(s,a))\right]} \\
&\le \sqrt{HG\frac{1}{\Delta_{\min}}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}\sum_{a\neq\pi^\star(s)}(q^{\pi_t}(s,a) + q^{\mathring{\pi}}(s,a))\Delta(s,a)\right]} \\
&\le \sqrt{HG\frac{\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}}{\Delta_{\min}}} \qquad \text{(by Lemma D.15 and (35))} \\
&\le \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}) + \frac{HG}{4\alpha\Delta_{\min}},
\end{aligned}$$

where the last inequality follows from the AM–GM inequality. □

E General lemma of optimistic follow-the-regularized-leader

In this section, we provide a regret analysis of optimistic follow-the-regularized-leader (OFTRL) for the MDP setting.

Lemma E.1 (OFTRL with policy optimization, Li et al. 2026, Lemma C.4)**.** *Suppose that a sequence of probability vectors $p_1, \dots, p_T \in \triangle(\mathcal{A})$ is given by OFTRL in*

$$p_t = \operatorname*{arg\,min}_{p\in\Delta(\mathcal{A})}\left\{\left\langle p, \sum_{\tau=1}^{t-1}\ell_\tau + m_t\right\rangle + \psi_t(p)\right\}, \qquad \psi_t(p) = \sum_a \frac{1}{\eta_t(a)}\log\left(\frac{1}{p(a)}\right)$$

with $\eta_1(a) = \eta_1$ for all a, and positive nonincreasing learning rates, and let losses $\{\ell_t\}_{t=1}^T$, loss predictions $\{m_t\}_{t=1}^{T+1}$ and $\{x_t\}_{t=1}^T$ be such that

$$\eta_t(a)p_t(a)(\ell_t(a) - m_t(a) + x_t) \geq -\frac{1}{2}$$

for all t, a. Then, for any $u \in \triangle(\mathcal{A})$, the OFTRL algorithm achieves

$$\begin{aligned}\sum_{t=1}^T \langle p_t - u, \ell_t\rangle \leq \frac{A\log(AT^2)}{\eta_1} &+ \sum_{t=1}^T \sum_a \left(\frac{1}{\eta_{t+1}(a)} - \frac{1}{\eta_t(a)}\right)\log(AT^2) \\ &+ \sum_{t=1}^T \sum_a \eta_t(a)p_t(a)^2(\ell_t(a) - m_t(a) + x_t)^2 \\ &+ \frac{1}{T^2}\sum_{t=1}^T \left\langle -u + \frac{1}{A}\mathbf{1}, \ell_t\right\rangle + 2\|m_{T+1}\|_\infty.\end{aligned}$$

F Regret analysis for the full-information setting (Theorem 4.1)

In this section, we prove the best-of-both-worlds results for the full-information setting. We present the bound for the adversarial regime in Theorem F.10 and the bound for the stochastic regime with adversarial corruption in Theorem F.11. Together, these results establish Theorem 4.1.

F.1 Auxiliary lemmas

Building on the policy optimization framework of Luo et al. (2021); Dann et al. (2023), we use the following key lemma to derive our regret bounds.

Lemma F.1 (Restatement of Lemma 3.1)**.** *Suppose that $P \in \mathcal{P}_t$ for all $t \in [T]$. Suppose that $b_t(s)$ is a nonnegative loss function. Suppose also that, for a comparator policy π, there exists $J^\pi \in \mathbb{R}$ such that*

$$\begin{aligned}&\mathbb{E}\left[\sum_s q^\pi(s)\sum_{t=1}^T\sum_a(\pi_t(a \mid s) - \pi(a \mid s))(Q^{\pi_t}(s,a;\ell_t) - B_t(s,a))\right] \\ &\leq J^\pi + \mathbb{E}\left[\sum_{t=1}^T\sum_s q^\pi(s)b_t(s)\right] + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^T\sum_s\sum_a q^\pi(s)\pi_t(a \mid s)B_t(s,a)\right].\end{aligned} \tag{36}$$

Then, it holds that

$$\mathrm{Reg}_T(\pi) \leq J^\pi + 3\,\mathbb{E}\left[\sum_{t=1}^T V^{\overline{P}_t^B,\pi_t}(s_0;b_t)\right],$$

where $\overline{P}_t^B \in \mathcal{P}_t$ simultaneously attains the maxima in (5) for all (s,a).

To show (36), we choose b_t in (9) and decompose the LHS of (36) as

$$\begin{aligned}&\sum_s q^\pi(s)\sum_{t,a}(\pi_t(a \mid s) - \pi(a \mid s))(Q^{\pi_t}(s,a;\ell_t) - B_t(s,a)) \\ &= \sum_s q^\pi(s)\underbrace{\sum_{t,a}(\pi_t(a \mid s) - \pi(a \mid s))\left(\widehat{Q}_t(s,a) - B_t(s,a)\right)}_{\mathbf{reg}^\pi(s)} \\ &\quad + \sum_s q^\pi(s)\underbrace{\sum_{t,a}(\pi_t(a \mid s) - \pi(a \mid s))\left(Q^{\pi_t}(s,a;\ell_t) - \widehat{Q}_t(s,a)\right)}_{\mathbf{bias}^\pi(s)}.\end{aligned} \tag{37}$$

Lemma F.2 (full-information setting)**.** *The variables $b_t(s)$ in (9) and $B_t(s,a)$ in (5) satisfy*

$$B_t(s,a) \leq 2$$

for all episodes t and state-action pairs (s,a).

Proof. We first upper bound the dilated bonus-to-go $B_t(s,a)$ by unrolling the dilated recursion. Since $(1+1/H)^H \le 3$, we obtain

$$B_t(s,a) \le 3 \sum_{s'} q^{\overline{P}_t^B, \pi_t}(s' \mid s,a)\, b_t(s'). \tag{38}$$

By the definition of b_t in (9) and the learning-rate update (8),

$$\begin{aligned} b_t(s) &= 7 \sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)} \right) \log(T) \\ &\le 7 \sum_a \eta_t(s,a) \zeta_t(s,a) \\ &\le 7 \eta_1 H^2, \end{aligned}$$

where we used $\sum_a \zeta_t(s,a) \le H^2$ and $\eta_t(s,a) \le \eta_1$.

Using $\eta_1 = \frac{1}{12H^3}$ and (38), we get

$$B_t(s,a) \le 3 \sum_{s'} q^{\overline{P}_t^B, \pi_t}(s' \mid s,a) b_t(s') \le 21 \eta_1 H^3 \le 2.$$

□

Lemma F.3 (Dann et al. 2023, 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,

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

Lemma F.4 (full-information setting)**.** *Suppose that the learning rates are updated according to (8). Then, it holds*

$$\eta_t(s,a) \le \frac{2\sqrt{\log(T)}}{\sqrt{\sum_{\tau=1}^t \zeta_\tau(s,a)}}$$

for all episodes t and state-action pairs (s,a).

Proof. Fix (s,a) and define $\phi_t(s,a) = \frac{\zeta_t(s,a)}{\log(T)}$. Then the update rule of learning rates can be written as

$$\frac{1}{\eta_{t+1}(s,a)} = \frac{1}{\eta_t(s,a)} + \eta_t(s,a) \phi_t(s,a).$$

To apply Lemma F.3, we verify that $\phi_t(s,a) \le \frac{1}{\eta_t(s,a)^2}$. Indeed,

$$\phi_t(s,a) \eta_t(s,a)^2 = \frac{\eta_t(s,a)^2 \zeta_t(s,a)}{\log(T)} \le \frac{\eta_1^2 H^2}{\log(T)} \le \frac{H^2}{\log(T)} \cdot \frac{1}{144 H^6} \le 1,$$

which follows from $\eta_1 = \frac{1}{12H^3}$.

Then, by Lemma F.3, we have

$$\eta_t(s,a) \le \frac{2}{\sqrt{\sum_{\tau=1}^t \phi_\tau}} \le \frac{2\sqrt{\log(T)}}{\sqrt{\sum_{\tau=1}^t \zeta_\tau(s,a)}}.$$

□

F.2 Common regret analysis

In this part, we upper bound the right-hand side of (36).

Lemma F.5 (full-information setting)**.** *For each state $s \in \mathcal{S}$ and any comparator policy π, it holds that*

$$\mathbf{reg}^{\pi}(s) \le O\big(H^3 A \log(T)\big) + \sum_{t=1}^{T} b_t(s) + \frac{1}{H}\sum_{t=1}^{T}\sum_{a} \pi_t(a \mid s) B_t(s,a).$$

Proof. We will apply Lemma E.1 with $p_t = \pi_t(\cdot \mid s)$ and $\ell_t = \widehat{Q}_t(s,a) - B_t(s,a)$ for each $s \in \mathcal{S}$. To do so, in what follows, we will check the conditions of Lemma E.1. Let

$$\begin{aligned} x_t &= \Big\langle -\pi_t(\cdot \mid s), \widehat{Q}_t(s,\cdot) - Q^{\underline{P}_t^m, \pi_t}(s,\cdot; m_t) \Big\rangle \\ &= -\sum_{a} \pi_t(a \mid s)\Big(Q^{\underline{P}_t^{\ell}, \pi_t}(s,a;\ell_t) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t)\Big) \end{aligned}$$

and verify that for all (s,a), $\eta_t(s,a)\pi_t(a \mid s)\Big(\widehat{Q}_t(s,a) - B_t(s,a) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) + x_t\Big) \ge -1/2$. Indeed,

$$\begin{aligned} &\eta_t(s,a)\pi_t(a \mid s)\Big(\widehat{Q}_t(s,a) - B_t(s,a) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) + x_t\Big) \\ &\ge \eta_t(s,a)\pi_t(a \mid s)(-H - B_t(s,a) - H) \\ &\ge -4\eta_1 H \ge -\frac{1}{2}, \end{aligned}$$

where the bounds in the second lines use $-H \le \widehat{Q}_t(s,a) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) \le H$, $B_t(s,a) \le 2 \le 2H$ from Lemma F.2.

Thus, Lemma E.1 implies that, for any comparator policy π,

$$\begin{aligned} &\mathbf{reg}^{\pi}(s) \\ &\le \frac{A\log(AT^2)}{\eta_1} + \sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(AT^2) \\ &\quad + \sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2\Big(Q^{\underline{P}_t^{\ell}, \pi_t}(s,a;\ell_t) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) - B_t(s,a) + x_t\Big)^2 \\ &\quad + \frac{1}{T^2}\sum_{t=1}^{T}\Big\langle -\pi(\cdot \mid s) + \frac{1}{A}\mathbf{1}, \widehat{Q}_t(s,\cdot) - B_t(s,\cdot)\Big\rangle + 2\Big\|Q^{\underline{P}_{T+1}^m, \pi_{T+1}}(s,\cdot; m_{T+1})\Big\|_{\infty} \\ &\le \frac{3A\log(T)}{\eta_1} + \frac{6H}{T} + 2H + 3\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) \\ &\qquad\qquad (\Big\|Q^{\underline{P}_{T+1}^m, \pi_{T+1}}(s,\cdot; m_{T+1})\Big\|_{\infty} \le H) \\ &\quad + 2\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2\Big(Q^{\underline{P}_t^{\ell}, \pi_t}(s,a;\ell_t) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) + x_t\Big)^2 \\ &\quad + 2\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2 B_t(s,a)^2 \\ &\le \mathcal{O}(H^3 A\log(T)) + 3\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) \\ &\quad + \underbrace{2\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2\Big(Q^{\underline{P}_t^{\ell}, \pi_t}(s,a;\ell_t) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) + x_t\Big)^2}_{\text{stability-term}} \end{aligned}$$

$$+\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a\mid s)B_t(s,a). \tag{39}$$

Here, the second inequality follows from

$$\begin{aligned}
&\frac{1}{T^2}\sum_{t=1}^{T}\left\langle -\pi(\cdot\mid s)+\frac{1}{A}\mathbf{1},\widehat{Q}_t(s,\cdot)-B_t(s,\cdot)\right\rangle\\
&\le \frac{1}{T^2}\left\|-\pi(\cdot\mid s)+\frac{1}{A}\mathbf{1}\right\|_1\left\|\sum_{t=1}^{T}\Big(\widehat{Q}_t(s,\cdot)-B_t(s,\cdot)\Big)\right\|_\infty\\
&\le \frac{2T\cdot 3H}{T^2}=\frac{6H}{T},
\end{aligned}$$

where we used $\left\|-\pi(\cdot\mid s)+\frac{1}{A}\mathbf{1}\right\|_1\le 2$, $\left|\widehat{Q}_t(s,a)\right|\le H$ and $B_t(s,a)\le 2$ from Lemma F.2.

It remains to bound the stability-term. Define

$$\mathcal{Q}(s,a)\coloneqq Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t).$$

Since

$$x_t=-\sum_b \pi_t(b\mid s)\mathcal{Q}(s,b),$$

we have

$$\begin{aligned}
\left(\mathcal{Q}(s,a)-\sum_b\pi_t(b\mid s)\mathcal{Q}(s,b)\right)^2 &= \left((1-\pi_t(a\mid s))\mathcal{Q}(s,a)-\sum_{b\neq a}\pi_t(b\mid s)\mathcal{Q}(s,b)\right)^2\\
&\le 2(1-\pi_t(a\mid s))^2\mathcal{Q}(s,a)^2+2\left(\sum_{b\neq a}\pi_t(b\mid s)\mathcal{Q}(s,b)\right)^2\\
&\le 2(1-\pi_t(a\mid s))^2\mathcal{Q}(s,a)^2\\
&\quad+2(1-\pi_t(a\mid s))\sum_{b\neq a}\pi_t(b\mid s)\mathcal{Q}(s,b)^2,
\end{aligned}$$

where the last inequality uses the Cauchy–Schwarz inequality. Thus, we obtain

$$\pi_t(a\mid s)^2\left(\mathcal{Q}(s,a)-\sum_b\pi_t(b\mid s)\mathcal{Q}(s,b)\right)^2\le 2\pi_t(a\mid s)(1-\pi_t(a\mid s))\sum_b\pi_t(b\mid s)\mathcal{Q}(s,b)^2$$

Therefore,

$$\begin{aligned}
&\eta_t(s,a)\pi_t(a\mid s)^2\Big(Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t)+x_t\Big)^2\\
&\le 2\eta_t(s,a)\pi_t(a\mid s)(1-\pi_t(a\mid s))\sum_b\pi_t(b\mid s)\Big(Q^{\underline{P}_t^{\ell},\pi_t}(s,b;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,b;m_t)\Big)^2\\
&\le 2\eta_t(s,a)\zeta_t(s,a),
\end{aligned}$$

where the last equality follows from the definition of $\zeta_t(s,a)=\pi_t(a\mid s)(1-\pi_t(a\mid s))\sum_b\pi_t(b\mid s)\Big(Q^{\underline{P}_t^{\ell},\pi_t}(s,b;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,b;m_t)\Big)^2$.

Then, stability-term is evaluated as

$$\begin{aligned}\text{stability-term} &= \sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a\mid s)^2\Big(Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t)+x_t\Big)^2\\ &\le 2\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\zeta_t(s,a)\\ &\le 2\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)}-\frac{1}{\eta_t(s,a)}\right)\log(T),\end{aligned}$$

where the last inequality follows from (8).

Combining the above bound on the stability-term with (39), we obtain, for any policy π,

$$\begin{aligned}\mathbf{reg}^{\pi}(s) &\le \mathcal{O}\big(H^3A\log(T)\big)+7\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)}-\frac{1}{\eta_t(s,a)}\right)\log(T)\\ &\quad+\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a\mid s)B_t(s,a)\\ &= \mathcal{O}\big(H^3A\log(T)\big)+\sum_{t=1}^{T}b_t(s)+\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a\mid s)B_t(s,a).\end{aligned}$$

□

Lemma F.6 (full-information setting)**.** *Suppose that the event $\mathcal{E}$ holds. Then, for any comparator policy π, it holds that*

$$\sum_{s}q^{\pi}(s)\mathbf{bias}^{\pi}(s)\lesssim\sqrt{S^2A\iota^2Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)}+HS^2A\iota^2. \tag{40}$$

Also,

$$\begin{aligned}\sum_{s}q^{\pi}(s)\mathbf{bias}^{\pi}(s)\lesssim\sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\iota^2\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}\\ +HS^2A\iota^2.\end{aligned} \tag{41}$$

Proof.

$$\begin{aligned}\sum_{s}q^{\pi}(s)\mathbf{bias}^{\pi}(s) &= \sum_{s}q^{\pi}(s)\sum_{t,a}(\pi_t(a\mid s)-\pi(a\mid s))\Big(Q^{\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)\Big)\\ &= \sum_{t=1}^{T}\sum_{s,a}(q^{\pi_t}(s,a)-q^{\pi}(s,a))z_t(s,a),\end{aligned}$$

where, for each t, Lemma D.2 is applied with

$$L(s,a)=Q^{\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t).$$

Thus,

$$\begin{aligned}z_t(s,a) &= Q^{\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)\\ &\quad-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\Big[Q^{\pi_t}(s',a';\ell_t)-Q^{\underline{P}_t^{\ell},\pi_t}(s',a';\ell_t)\Big].\end{aligned}$$

Expanding the definition of $z_t(s,a)$, we obtain

$$\begin{aligned}z_t(s,a) &= \mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[Q^{\pi_t}(s',a';\ell_t)]-\mathbb{E}_{s'\sim\underline{P}_t^{\ell}(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\Big[Q^{\underline{P}_t^{\ell},\pi_t}(s',a';\ell_t)\Big]\\ &\quad-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\Big[Q^{\pi_t}(s',a';\ell_t)-Q^{\underline{P}_t^{\ell},\pi_t}(s',a';\ell_t)\Big]\\ &= \mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\Big[Q^{\underline{P}_t^{\ell},\pi_t}(s',a';\ell_t)\Big]-\mathbb{E}_{s'\sim\underline{P}_t^{\ell}(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\Big[Q^{\underline{P}_t^{\ell},\pi_t}(s',a';\ell_t)\Big]\\ &= \mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[V^{\underline{P}_t^{\ell},\pi_t}(s';\ell_t)\Big]-\mathbb{E}_{s'\sim\underline{P}_t^{\ell}(\cdot|s,a)}\Big[V^{\underline{P}_t^{\ell},\pi_t}(s';\ell_t)\Big].\end{aligned}$$

By the definition of $\underline{P}_t^\ell$, which minimizes $Q(s,a;\ell_t)$ for all (s,a), we have $z_t(s,a) \geq 0$.

First, we show (40). By the fact that $z_t(s,a) \geq 0$,

$$\begin{aligned}
\sum_{t=1}^T \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a)) z_t(s,a) &\leq \sum_{t=1}^T V^{\pi_t}(s_0; z_t) \\
&\leq \sum_{t=1}^T \Big(V^{\pi_t}(s_0;\ell_t) - V^{\underline{P}_t^\ell, \pi_t}(s_0;\ell_t) \Big) \qquad \text{(by Lemma D.2)} \\
&= \sum_{t=1}^T \sum_s \Big(q^{\pi_t}(s) - q^{\underline{P}_t^\ell, \pi_t}(s) \Big) \sum_a \pi_t(a \mid s) \ell_t(s,a). \qquad (42)
\end{aligned}$$

Applying Lemma D.5 to (42), we obtain

$$\sum_{t=1}^T \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a)) z_t(s,a) \lesssim \sqrt{S^2 A \iota^2 Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)} + HS^2 A \iota^2.$$

This proves (40).

Next, we show (41). By Lemma D.4,

$$\begin{aligned}
z_t(s,a) &= \sum_{s'} \Big(P(s' \mid s,a) - \underline{P}_t^\ell(s' \mid s,a) \Big) V^{\underline{P}_t^\ell, \pi_t}(s';\ell_t) \\
&\lesssim \sqrt{\frac{|S_{h(s)+1}| \mathsf{Var}_{s' \sim P(\cdot|s,a)} \big[V^{\underline{P}_t^\ell, \pi_t}(s';\ell_t) \big] \iota}{n_t(s,a)}} + \frac{|S_{h(s)+1}| H \iota}{n_t(s,a)}.
\end{aligned}$$

Using $z_t(s,a) \geq 0$, we further have

$$\begin{aligned}
&\sum_{t=1}^T \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a)) z_t(s,a) \\
&\leq \sum_{t=1}^T \sum_{s,a} [q^{\pi_t}(s,a) - q^\pi(s,a)]_+ z_t(s,a) \\
&\lesssim \sum_{t=1}^T \sum_{s,a} [q^{\pi_t}(s,a) - q^\pi(s,a)]_+ \left(\sqrt{\frac{|S_{h(s)+1}| \mathsf{Var}_{s' \sim P(\cdot|s,a)} \big[V^{\underline{P}_t^\ell, \pi_t}(s';\ell_t) \big] \iota}{n_t(s,a)}} + \frac{|S_{h(s)+1}| H \iota}{n_t(s,a)} \right) \\
&\leq \sum_{h=0}^{H-1} \sum_{t=1}^T \sum_{(s,a) \in \mathcal{S}_h \times \mathcal{A}} [q^{\pi_t}(s,a) - q^\pi(s,a)]_+ \sqrt{\frac{|S_{h+1}| \mathsf{Var}_{s' \sim P(\cdot|s,a)} \big[V^{\underline{P}_t^\ell, \pi_t}(s';\ell_t) \big] \iota}{n_t(s,a)}} \\
&\quad + \sum_{t=1}^T \sum_{s,a} q^{\pi_t}(s,a) \frac{|S_{h(s)+1}| H \iota}{n_t(s,a)}. \qquad (43)
\end{aligned}$$

Here, by the Cauchy–Schwarz inequality, the first term in (43) is bounded as

$$\begin{aligned}
&\sum_{t=1}^T \sum_{(s,a) \in \mathcal{S}_h \times \mathcal{A}} [q^{\pi_t}(s,a) - q^\pi(s,a)]_+ \sqrt{\frac{|S_{h+1}| \mathsf{Var}_{s' \sim P(\cdot|s,a)} \big[V^{\underline{P}_t^\ell, \pi_t}(s';\ell_t) \big] \iota}{n_t(s,a)}} \\
&\leq \sqrt{\sum_{t=1}^T \sum_{(s,a) \in \mathcal{S}_h \times \mathcal{A}} [q^{\pi_t}(s,a) - q^\pi(s,a)]_+} \\
&\quad \times \sqrt{\sum_{t=1}^T \sum_{(s,a) \in \mathcal{S}_h \times \mathcal{A}} [q^{\pi_t}(s,a) - q^\pi(s,a)]_+ \frac{|S_{h+1}| \mathsf{Var}_{s' \sim P(\cdot|s,a)} \big[V^{\underline{P}_t^\ell, \pi_t}(s';\ell_t) \big] \iota}{n_t(s,a)}}.
\end{aligned}$$

The term inside the second square root is further bounded as

$$\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\frac{|S_{h+1}|\mathsf{Var}_{s'\sim P(\cdot|s,a)}\left[V^{\underline{P}_t^{\ell},\pi_t}(s';\ell_t)\right]\iota}{n_t(s,a)}$$
$$\le H^2|S_{h+1}|\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\frac{q^{\pi_t}(s,a)\iota}{n_t(s,a)}$$
$$\lesssim H^2|S_h||S_{h+1}|A\iota^2. \qquad \text{(by Lemma C.3)}$$

The second term in (43) is bounded as

$$\sum_{t=1}^{T}\sum_{s,a}q^{\pi_t}(s,a)\frac{|S_{h(s)+1}|H\iota}{n_t(s,a)} \le \sum_{h=0}^{H-1}|S_{h+1}|\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}q^{\pi_t}(s,a)\frac{H\iota}{n_t(s,a)}$$
$$\lesssim H\sum_{h=0}^{H-1}|S_{h+1}||S_h|A\iota^2$$
$$\le HS^2A\iota^2.$$

Therefore,

$$\sum_{t=1}^{T}\sum_{s,a}(q^{\pi_t}(s,a)-q^{\pi}(s,a))z_t(s,a)$$
$$\lesssim \sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\iota^2\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}+HS^2A\iota^2$$

This proves (41), and hence completes the proof. □

Lemma F.7 (full-information setting)**.** *Suppose that the event $\mathcal{E}$ holds. Then, it holds that*

$$\sum_{t=1}^{T}V^{\overline{P}_t^{B},\pi_t}(s_0;b_t)\lesssim\sqrt{\log(T)}\sum_{s,a}\sqrt{\sum_{t=1}^{T}q^{\pi_t}(s)\zeta_t(s,a)}+S^2A\iota^2.$$

Proof. First, we have

$$\sum_{t=1}^{T}V^{\overline{P}_t^{B},\pi_t}(s_0;b_t)=\sum_{t=1}^{T}\sum_{s}q^{\overline{P}_t^{B},\pi_t}(s)b_t(s)$$
$$=\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)b_t(s)+\sum_{t=1}^{T}\sum_{s}\left(q^{\overline{P}_t^{B},\pi_t}(s)-q^{\pi_t}(s)\right)b_t(s). \qquad (44)$$

To bound the second term, define

$$g_t(s,a):=b_t(s)\le 7\eta_1H^2=\frac{7}{12H}:=G.$$

Applying Lemmas D.5 and D.6 with this choice of g_t yields

$$\sum_{t=1}^{T}\sum_{s}\left(q^{\overline{P}_t^{B},\pi_t}(s)-q^{\pi_t}(s)\right)b_t(s)\lesssim\sqrt{HS^2A\iota^2\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)b_t(s)^2}+HS^2AG\iota^2$$
$$\le\sqrt{HS^2AG\iota^2\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)b_t(s)}+HS^2AG\iota^2$$
$$\lesssim\sqrt{S^2A\iota^2\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)b_t(s)}+S^2A\iota^2,$$

where we used $b_t(s)^2 \le G b_t(s)$ and $HG = 7/12$. Combining this with (44), we obtain

$$\sum_{t=1}^{T}\sum_{s} q^{\overline{P}_t^B,\pi_t}(s)b_t(s) \lesssim \sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s)b_t(s) + \sqrt{S^2A\iota^2\sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s)b_t(s)} + S^2A\iota^2$$
$$\lesssim \sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s)b_t(s) + S^2A\iota^2, \tag{45}$$

where the last inequality follows from the AM–GM inequality .

We next bound $\sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s)b_t(s)$.

By Lemma F.4, we have

$$\sum_{t=1}^{T}\sum_{s} q^{\pi_t}(s)b_t(s) = 7\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s)\eta_t(s,a)\zeta_t(s,a)$$
$$\lesssim \sqrt{\log(T)}\sum_{t=1}^{T}\sum_{s,a}\frac{q^{\pi_t}(s)\sqrt{\zeta_t(s,a)}\times\sqrt{\zeta_t(s,a)}}{\sqrt{\sum_{\tau=1}^{t}\zeta_\tau(s,a)}}$$
$$\le \sqrt{\log(T)}\sum_{s,a}\sum_{t=1}^{T}\frac{q^{\pi_t}(s)\zeta_t(s,a)}{\sqrt{\sum_{\tau=1}^{t}q^{\pi_\tau}(s)\zeta_\tau(s,a)}}$$
$$\lesssim \sqrt{\log(T)}\sum_{s,a}\sqrt{\sum_{t=1}^{T}q^{\pi_t}(s)\zeta_t(s,a)}. \tag{46}$$

Here we used $q^{\pi_t}(s) \le 1$ and, for $x_t \ge 0$,

$$\sum_{t=1}^{T}\frac{x_t}{\sqrt{\sum_{\tau=1}^{t}x_\tau}} \le 2\sum_{t=1}^{T}\left(\sqrt{\sum_{\tau=1}^{t}x_\tau} - \sqrt{\sum_{\tau=1}^{t-1}x_\tau}\right) = 2\sqrt{\sum_{t=1}^{T}x_t}.$$

By (45) and (46), we obtain

$$\sum_{t=1}^{T}V^{\overline{P}_t^B,\pi_t}(s_0;b_t) \lesssim \sqrt{\log(T)}\sum_{s,a}\sqrt{\sum_{t=1}^{T}q^{\pi_t}(s)\zeta_t(s,a)} + S^2A\iota^2.$$

This completes the proof. □

Lemma F.8 (full-information setting)**.** *For any comparator policy π, Algorithm 1 guarantees*

$$\mathrm{Reg}_T(\pi) \lesssim A(H^4+HS^2)\iota^2 + \sum_{s,a}\sqrt{\log(T)\mathbb{E}\left[\sum_{t=1}^{T}q^{\pi_t}(s)\zeta_t(s,a)\right]}$$
$$+ \min\Bigg\{\sqrt{S^2A\iota^2\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]},$$
$$\sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\iota^2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\right]}\Bigg\}.$$

Proof. By the definition of the regret decomposition in (37),

$$\begin{aligned}
&\mathbb{E}\left[\sum_s q^\pi(s)\sum_{t,a}(\pi_t(a\mid s)-\pi(a\mid s))(Q^{\pi_t}(s,a;\ell_t)-B_t(s,a))\right]\\
&=\mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{reg}^\pi(s)\right]+\mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{bias}^\pi(s)\right]\\
&\le \mathcal{O}\big(H^4A\log(T)\big)+\mathbb{E}\left[\sum_{t=1}^T\sum_s q^\pi(s)b_t(s)\right]+\mathbb{E}\left[\frac{1}{H}\sum_{t=1}^T\sum_s q^\pi(s)\sum_a \pi_t(a\mid s)B_t(s,a)\right]\\
&\quad+\mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{bias}^\pi(s)\right],
\end{aligned}$$

where we used Lemma F.5. Absorbing the $\mathcal{O}(H^2T\delta)$ contribution of $\bar{\mathcal{E}}_1$ into $\mathcal{O}(H^4A\log(T))$, we obtain

$$\mathrm{Reg}_T(\pi)\le \mathcal{O}\big(H^4A\log(T)\big)+3\mathbb{E}\left[\sum_{t=1}^T V^{\overline{P}_t^B,\pi_t}(s_0;b_t)\right]+\mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{bias}^\pi(s)\right].$$

Here, by Lemma F.7 and Jensen's inequality, with $\delta=1/T^2$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^T V^{\overline{P}_t^B,\pi_t}(s_0;b_t)\right]\lesssim \sqrt{\log(T)}\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^T q^{\pi_t}(s)\zeta_t(s,a)\right]}+S^2A\iota^2+\mathcal{O}(\delta)\times\mathcal{O}(HT),$$

where we used the trivial bound $\sum_{t=1}^T V^{\overline{P}_t^B,\pi_t}(s_0;b_t)\le\mathcal{O}(HT)$ on $\bar{\mathcal{E}}$ from Lemma F.2.

Then, using $\delta=\frac{1}{T^2}$, we have

$$\begin{aligned}
\mathrm{Reg}_T(\pi)&\lesssim H^4A\log(T)+S^2A\iota^2+\sqrt{\log(T)}\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^T q^{\pi_t}(s)\zeta_t(s,a)\right]}\\
&\quad+\mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{bias}^\pi(s)\right],
\end{aligned}\tag{47}$$

By (47) and Lemma F.6 and Jensen's inequality, together with the trivial bound $\mathcal{O}(H^2T)$ on $\bar{\mathcal{E}}$ and the choice $\delta=\mathcal{O}(1/T^2)$, we obtain, for any policy π,

$$\begin{aligned}
\mathrm{Reg}_T(\pi)&\lesssim H^4A\log(T)+S^2A\iota^2\\
&\quad+\sqrt{\log(T)}\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t=1}^T q^{\pi_t}(s)\zeta_t(s,a)\right]}\\
&\quad+\min\Bigg\{\sqrt{S^2A\iota^2\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]},\\
&\qquad\sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\iota^2\mathbb{E}\left[\sum_{t=1}^T\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^\pi(s,a)]_+\right]}\Bigg\}\\
&\quad+HS^2A\iota^2+\mathcal{O}(\delta)\times\mathcal{O}(H^2T)\\
&\lesssim A(H^4+HS^2)\iota^2+\sum_{s,a}\sqrt{\log(T)\mathbb{E}\left[\sum_{t=1}^T q^{\pi_t}(s)\zeta_t(s,a)\right]}\\
&\quad+\min\Bigg\{\sqrt{S^2A\iota^2\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]},
\end{aligned}$$

$$\sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\iota^2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\right]}\Bigg\}.$$

This completes the proof. □

F.3 Proof for the adversarial regime

Lemma F.9 (full-information setting)**.** *It holds that*

$$\begin{aligned}&\sum_{s,a}\sqrt{\log(T)\mathbb{E}\left[\sum_{t=1}^{T}q^{\pi_t}(s)\zeta_t(s,a)\right]}\\&\lesssim\sqrt{HSA\log(T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)\right]}.\end{aligned}$$

Proof. By the definition of $\zeta_t(s,a)$, we have

$$\begin{aligned}&\sum_{s,a}\sqrt{\log(T)\mathbb{E}\left[\sum_{t=1}^{T}q^{\pi_t}(s)\zeta_t(s,a)\right]}\\&=\sum_{s,a}\sqrt{\log(T)\mathbb{E}\left[\sum_{t=1}^{T}q^{\pi_t}(s,a)(1-\pi_t(a\mid s))\sum_{b}\pi_t(b\mid s)\big(Q^{\underline{P}_t^{\ell},\pi_t}(s,b;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,b;m_t)\big)^2\right]}\\&\le\sqrt{SA\log(T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s}q^{\pi_t}(s)\sum_{b}\pi_t(b\mid s)\big(Q^{\underline{P}_t^{\ell},\pi_t}(s,b;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,b;m_t)\big)^2\right]},\end{aligned}\tag{48}$$

where the inequality follows from the Cauchy–Schwarz inequality.

We next bound $Q^{\underline{P}_t^{\ell},\pi_t}(s,b;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,b;m_t)$.

$$\begin{aligned}&Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t)\\&=\Big(Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;\ell_t)\Big)+\Big(Q^{\underline{P}_t^{m},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t)\Big)\\&\le 0+Q^{\underline{P}_t^{m},\pi_t}(s,a;\ell_t-m_t)\\&\le\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|).\end{aligned}$$

Similarly,

$$Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t)\ge-\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)$$

Here, for any transition kernel $\widetilde{P}$, we have

$$\begin{aligned}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)&=\sum_{s',a'}q^{\widetilde{P},\pi_t}(s',a'\mid s,a)|\ell_t(s',a')-m_t(s',a')|\\&\le\sqrt{\sum_{s',a'}q^{\widetilde{P},\pi_t}(s',a'\mid s,a)}\sqrt{\sum_{s',a'}q^{\widetilde{P},\pi_t}(s',a'\mid s,a)(\ell_t(s',a')-m_t(s',a'))^2}\\&\le\sqrt{HQ^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Big(Q^{\underline{P}_t^{\ell},\pi_t}(s,a;\ell_t)-Q^{\underline{P}_t^{m},\pi_t}(s,a;m_t)\Big)^2&\le\left(\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)\right)^2\\&\le H\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)\end{aligned}\tag{49}$$

Combining this with (48) completes the proof. □

Theorem F.10 (full-information setting)**.** *For any comparator policy π, Algorithm 1 guarantees*

$$\begin{aligned}\mathrm{Reg}_T(\pi) \lesssim \sqrt{H^2SA\log(T)\min\{L(\pi), HT - L(\pi), Q_\infty, V_1\}} \\ + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} + A(H^4 + HS^2)\log^2(T).\end{aligned}$$

Proof. By Lemmas D.9, F.8 and F.9 and the choice $\delta = 1/T^2$, which implies $\iota \lesssim \log(T)$, we have, for any policy π,

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\lesssim A(H^4 + HS^2)\log^2(T) \\ &\quad + \sqrt{HSA\log(T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a} q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t - m_t)^2)\right]} \\ &\quad + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} \\ &\lesssim A(H^4 + HS^2)\log^2(T) + \sqrt{H^2SA\log(T)\mathbb{E}\left[\sum_{t=1}^{T} V^{\pi_t}(s_0;(\ell_t - m_t)^2)\right]} \\ &\quad + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]}.\end{aligned}$$

Combining this with Lemma D.11, we obtain

$$\begin{aligned}\mathrm{Reg}_T(\pi) \lesssim \sqrt{H^2SA\log(T)\min\{L(\pi) + \mathrm{Reg}_T(\pi), HT - L(\pi) - \mathrm{Reg}_T(\pi), Q_\infty, V_1\}} \\ + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} + A(H^4 + HS^2)\log^2(T). \qquad (50)\end{aligned}$$

Now, we derive the first-order bound. Assume $\mathrm{Reg}_T(\pi) \geq 0$. By (50), there exists an absolute constant $c > 0$ such that

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\leq cA(H^4 + HS^2)\log^2(T) + c\sqrt{H^2SA\log(T)L(\pi)} \\ &\quad + c\sqrt{H^2SA\log(T)\mathrm{Reg}_T(\pi)} \\ &\quad + c\sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} \\ &\leq cA(H^4 + HS^2)\log^2(T) + c\sqrt{H^2SA\log(T)L(\pi)} \\ &\quad + \frac{c^2}{2}H^2SA\log(T) + \frac{1}{2}\mathrm{Reg}_T(\pi) \\ &\quad + c\sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]},\end{aligned}$$

where the last inequality follows from the AM–GM inequality. Thus, we obtain

$$\begin{aligned}\mathrm{Reg}_T(\pi) \lesssim A(H^4 + HS^2)\log^2(T) + \sqrt{H^2SA\log(T)L(\pi)} \\ + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]}. \qquad (51)\end{aligned}$$

Furthermore, we derive the bound in terms of $HT - L(\pi)$. If $\mathrm{Reg}_T(\pi) < 0$, the desired upper bound is trivial. Otherwise, $\mathrm{Reg}_T(\pi) \geq 0$ and (50) gives

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\lesssim A(H^4 + HS^2)\log^2(T) + \sqrt{H^2SA\log(T)(HT - L(\pi) - \mathrm{Reg}_T(\pi))} \\ &\quad + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} \\ &\leq A(H^4 + HS^2)\log^2(T) + \sqrt{H^2SA\log(T)(HT - L(\pi))} \\ &\quad + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]}. \qquad (52)\end{aligned}$$

Combining (50)–(52) completes the proof. □

F.4 Proof for the stochastic regime with adversarial corruption

Theorem F.11 (full-information setting)**.** *Under the stochastic regime with adversarial corruption, Algorithm 1 guarantees that, for any policy π,*

$$\mathrm{Reg}_T(\pi) \lesssim U + \sqrt{U\mathcal{C}} + A(H^4 + HS^2)\log^2(T),$$

where $U = \frac{H^2S^2A\log^2(T)}{\Delta_{\min}}$.

Proof. Since $\mathrm{Reg}_T(\pi) \le \mathrm{Reg}_T(\mathring{\pi})$ for all π, it suffices to bound $\mathrm{Reg}_T(\mathring{\pi})$. From Lemma F.8 and the choice $\delta = 1/T^2$, which implies $\iota \lesssim \log(T)$, we have

$$\mathrm{Reg}_T(\mathring{\pi}) \lesssim A(H^4 + HS^2)\log^2(T) + 2H\log(T)\sum_s \sum_{a\neq\pi^\star(s)} \sqrt{\mathbb{E}\left[\sum_{t=1}^T q^{\pi_t}(s,a)\right]} + \sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\log^2(T)\mathbb{E}\left[\sum_{t=1}^T \sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} [q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)]_+\right]}.$$

Applying Lemmas D.16 and D.17, for any $\alpha > 0$, we get

$$\begin{aligned}
&\mathrm{Reg}_T(\mathring{\pi}) \\
&\le \alpha(2\mathrm{Reg}_T(\mathring{\pi}) + 6\mathcal{C}) + \mathcal{O}\left(\sum_s \sum_{a\neq\pi^\star(s)} \frac{H^2\log^2(T)}{\alpha\Delta(s,a)} + A(H^4+HS^2)\log^2(T)\right) \\
&\quad + \mathcal{O}\left(\frac{\left(\sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\log^2(T)}\right)^2}{4\alpha\Delta_{\min}}\right).
\end{aligned}$$

Here,

$$\begin{aligned}
\left(\sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\log^2(T)}\right)^2 &\le \left(\sum_{h=0}^{H-1}\frac{1}{2}(|S_h|+|S_{h+1}|)\sqrt{H^2A\log^2(T)}\right)^2 \\
&\qquad \text{(by the AM–GM inequality)} \\
&= \left(S\sqrt{H^2A\log^2(T)}\right)^2 = H^2S^2A\log^2(T). \qquad (53)
\end{aligned}$$

Then, using (53) and choosing $\alpha = \min\left\{\frac{1}{4}, \sqrt{\frac{U}{\mathcal{C}}}\right\}$ with $U = \frac{H^2S^2A\log^2(T)}{\Delta_{\min}}$, we get

$$\mathrm{Reg}_T(\mathring{\pi}) \lesssim U + \sqrt{U\mathcal{C}} + A(H^4+HS^2)\log^2(T).$$

Therefore, since $\mathrm{Reg}_T(\pi) \le \mathrm{Reg}_T(\mathring{\pi})$ for all π by the definition of $\mathring{\pi}$, the desired bound follows.

□

G Regret analysis for the bandit setting (Theorem 4.2)

In this section, we prove the best-of-both-worlds results for the bandit setting. We present the bound for the adversarial regime in Theorem G.12 and the bound for the stochastic regime with adversarial corruption in Theorem G.13. Together, these results establish Theorem 4.2.

Throughout this section, virtual episodes are inserted into the episode sequence, shifting the indices of subsequent real episodes.

We denote by $\mathcal{T}_r = \{t \in [T] : Y_t = 1\}$ and $\mathcal{T}_v = \{t \in [T] : Y_t = 0\}$ the sets of real and virtual episodes, with sizes $|\mathcal{T}_r|$ and $|\mathcal{T}_v|$, respectively.

G.1 Auxiliary lemmas

Following the full-information setting, Lemma F.1 reduces the regret analysis to proving (36). With the choice of b_t in (15), we decompose the left-hand side of (36) as

$$\begin{aligned}
&\sum_s q^\pi(s) \sum_{t,a} (\pi_t(a \mid s) - \pi(a \mid s))(Q^{\pi_t}(s,a;\ell_t) - B_t(s,a)) \\
&= \underbrace{\sum_s q^\pi(s) \sum_{t,a} (\pi_t(a \mid s) - \pi(a \mid s)) \Big(\widehat{Q}_t(s,a) - B_t(s,a) \Big)}_{\mathbf{reg}^\pi(s)} \\
&\quad + \underbrace{\sum_s q^\pi(s) \sum_{t,a} (\pi_t(a \mid s) - \pi(a \mid s)) \Big(Q^{\pi_t}(s,a;\ell_t) - \widehat{Q}_t(s,a) + Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) - Q^{\pi_t}(s,a;m_t) \Big)}_{\mathbf{bias}^\pi(s)} \\
&\quad + \underbrace{\sum_s q^\pi(s) \sum_{t,a} (\pi_t(a \mid s) - \pi(a \mid s)) \Big(Q^{\pi_t}(s,a;m_t) - Q^{\underline{P}_t^m, \pi_t}(s,a;m_t) \Big)}_{\mathbf{error}^\pi(s)}. \qquad (54)
\end{aligned}$$

Lemma G.1 (bandit setting)**.** *It holds that*

$$q_t(s)\pi_t(a \mid s) C_t(s,a) \le 6HS, \qquad \mathbb{E}_t[C_t(s,a)] \le 2HS(t+1). \qquad (55)$$

for any episode t and state-action pair (s,a) on every history.

Proof. We first prove a conditional occupancy property. For any $\widetilde{P} \in \mathcal{P}_t$, any states s, s' with $h(s') > h(s)$, and any actions a, a', we have

$$q_t(s)\pi_t(a \mid s) q^{\widetilde{P}, \pi_t}(s', a' \mid s, a) \le q_t(s')\pi_t(a' \mid s'). \qquad (56)$$

Hence,

$$\overline{q}_t^{\pi_t}(s') \ge \overline{q}_t^{\pi_t}(s)\pi_t(a \mid s) q^{\widetilde{P}, \pi_t}(s' \mid s, a).$$

Using $q_t(x) = \overline{q}_t^{\pi_t}(x) + \gamma_t$ and $\pi_t(a \mid s) q^{\widetilde{P}, \pi_t}(s' \mid s, a) \le 1$, we obtain

$$\begin{aligned}
q_t(s')\pi_t(a' \mid s') &= (\overline{q}_t^{\pi_t}(s') + \gamma_t)\pi_t(a' \mid s') \\
&\ge (\overline{q}_t^{\pi_t}(s) + \gamma_t)\pi_t(a \mid s) q^{\widetilde{P}, \pi_t}(s' \mid s, a)\pi_t(a' \mid s') \\
&= q_t(s)\pi_t(a \mid s) q^{\widetilde{P}, \pi_t}(s', a' \mid s, a),
\end{aligned}$$

which proves (56).

Now, by the definition of C_t and (56) with $\widetilde{P} = \overline{P}_t^C$,

$$\begin{aligned}
&q_t(s)\pi_t(a \mid s) C_t(s,a) \\
&\le 2q_t(s)\pi_t(a \mid s) \sum_{s',a'} q^{\overline{P}_t^C, \pi_t}(s', a' \mid s, a) \rho_t(s') \frac{\mathbb{I}_t(s',a') D_{t,h(s')}}{q_t(s')\pi_t(a' \mid s')} \\
&\quad + 2H q_t(s)\pi_t(a \mid s) \sum_{s',a'} q^{\overline{P}_t^C, \pi_t}(s', a' \mid s, a) \rho_t(s')^2 \\
&\le 2 \sum_{s',a'} \rho_t(s') \mathbb{I}_t(s',a') D_{t,h(s')} + 2H \sum_{s',a'} q_t(s')\pi_t(a' \mid s') \rho_t(s')^2. \qquad (57)
\end{aligned}$$

Since $\rho_t(s') \le 1$, $D_{t,h(s')} \le H$,

$$2 \sum_{s',a'} \rho_t(s') \mathbb{I}_t(s',a') D_{t,h(s')} \le 2H^2.$$

Moreover, using $\rho_t(s') \le 1$ and $\sum_{s',a'} q_t(s')\pi_t(a' \mid s') \le S + S\gamma_t \le 2S$,

$$2H \sum_{s',a'} q_t(s')\pi_t(a' \mid s')\rho_t(s')^2 \le 4HS.$$

Combining the last three displays gives

$$q_t(s)\pi_t(a \mid s)C_t(s,a) \le 6HS.$$

This proves the first inequality in (55).

We next prove the second inequality. Let $\overline{P}_t^C$ be a measurable transition kernel that attains the maximum in (13). Then, it holds that

$$C_t(s,a) \le 2 \sum_{s',a'} q^{\overline{P}_t^C,\pi_t}(s',a' \mid s,a) c_t(s',a').$$

Since $q_t(s) \ge 1/t$, $\rho_t(s) \le 1$, and $D_{t,h(s)} \le H$,

$$\mathbb{E}_t[c_t(s,a)] \le \rho_t(s)\frac{Hq^{\pi_t}(s)}{q_t(s)} + H\rho_t(s)^2 \le H(t+1).$$

By using $q^{\overline{P}_t^C,\pi_t}(s',a' \mid s,a) \le \pi_t(a' \mid s')$ for $h(s') > h(s)$, we get

$$\mathbb{E}_t[C_t(s,a)] \le 2HS(t+1),$$

which proves the second inequality. □

Lemma G.2 (bandit setting)**.** *The variables $b_t(s)$ in (15) and $B_t(s,a)$ in (5) satisfy*

$$\eta_t(s,a)\pi_t(a \mid s)B_t(s,a) \le \frac{1}{6H}, \quad B_t(s,a) \le \frac{8\sqrt{HS^3A}}{\gamma_t}$$

for any episode t and state-action pair (s,a).

Proof. Let $R_t = \max_{s,a} \frac{\eta_t(s,a)}{q_t(s)}$.

We first consider the case when t is a real episode. In real episodes, by the definition of b_t,

$$b_t(s) = 6\sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) + 3\sum_a \eta_t(s,a)\pi_t(a \mid s)^2 C_t(s,a)^2$$

For the first term, the learning-rate update gives

$$\begin{aligned}
&6\sum_a \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T) \\
&= \frac{6}{q_t(s)^2}\sum_a \eta_t(s,a)\zeta_t(s,a) \\
&= \frac{6}{q_t(s)^2}\sum_a \eta_t(s,a)(\mathbb{I}_t(s,a) - \pi_t(a \mid s)\mathbb{I}_t(s))^2\big(L_{t,h(s)} - M_{t,h(s)}\big)^2 \\
&\le \frac{6H^2}{q_t(s)^2}\max_a \eta_t(s,a)\sum_a (\mathbb{I}_t(s,a) - \pi_t(a \mid s)\mathbb{I}_t(s))^2 \\
&\le \frac{12H^2}{q_t(s)^2}\max_a \eta_t(s,a) \\
&\le \frac{12H^2}{q_t(s)}\max_a \frac{\eta_t(s,a)}{q_t(s)}.
\end{aligned} \tag{58}$$

For the second term, by Lemma G.1,

$$\begin{aligned}
3\sum_a \eta_t(s,a)\pi_t(a\mid s)^2 C_t(s,a)^2 &= 3\sum_a \frac{\eta_t(s,a)q_t(s)^2\pi_t(a\mid s)^2 C_t(s,a)^2}{q_t(s)^2} \\
&\le \frac{108H^2S^2\sum_a \eta_t(s,a)}{q_t(s)^2} \qquad \text{(by Lemma G.1)} \\
&\le \frac{108H^2S^2A}{q_t(s)}\max_a \frac{\eta_t(s,a)}{q_t(s)}. \qquad (59)
\end{aligned}$$

Combining (58) and (59), we obtain

$$b_t(s) \le \frac{12H^2+108H^2S^2A}{q_t(s)}\max_a \frac{\eta_t(s,a)}{q_t(s)} \le \frac{120H^2S^2A}{q_t(s)}\max_a \frac{\eta_t(s,a)}{q_t(s)} \qquad (60)$$

Using (38) and (60), we have

$$\begin{aligned}
B_t(s,a) &\le 3\sum_{h=h(s)}^{H-1}\sum_{s'\in\mathcal{S}_h} q^{\overline{P}_t^B,\pi_t}(s'\mid s,a)b_t(s') \\
&\le 360H^2S^2AR_t\left(\sum_{h=h(s)}^{H-1}\sum_{s'\in\mathcal{S}_h} q^{\overline{P}_t^B,\pi_t}(s'\mid s,a)\frac{1}{q_t(s')}\right) \\
&\le 360H^2S^2AR_t\left(\sum_{h=h(s)}^{H-1}\sum_{s'\in\mathcal{S}_h} q^{\overline{P}_t^B,\pi_t}(s'\mid s,a)\frac{1}{\overline{q}_t^{\pi_t}(s)\pi_t(a\mid s)q^{\overline{P}_t^B,\pi_t}(s'\mid s,a)+\gamma_t}\right) \\
&\le 360H^2S^2AR_t\left(\sum_{h=h(s)}^{H-1}\sum_{s'\in\mathcal{S}_h} \frac{1}{\overline{q}_t^{\pi_t}(s)\pi_t(a\mid s)+\gamma_t}\right) \\
&\le 360H^2S^3AR_t\cdot\frac{1}{\overline{q}_t^{\pi_t}(s)\pi_t(a\mid s)+\gamma_t} \qquad (61) \\
&\le \frac{8\sqrt{HS^3A}}{\gamma_t},
\end{aligned}$$

where in the last line we used $R_t \le \frac{1}{50\sqrt{H^3S^3A}}$ that holds in real episodes.

By using (61), we also have

$$\begin{aligned}
\eta_t(s,a)\pi_t(a\mid s)B_t(s,a) &\le 360H^2S^3AR_t\cdot\frac{\eta_t(s,a)\pi_t(a\mid s)}{\overline{q}_t^{\pi_t}(s)\pi_t(a\mid s)+\gamma_t} \\
&\le 360H^2S^3AR_t\cdot\frac{\eta_t(s,a)}{q_t(s)} \\
&\le 360H^2S^3AR_t^2 \\
&\le \frac{1}{6H},
\end{aligned}$$

where we used $\overline{q}_t^{\pi_t}(s)\pi_t(a\mid s)+\gamma_t \ge q_t(s)\pi_t(a\mid s)$, $\frac{\eta_t(s,a)}{q_t(s)} \le R_t$.

We next consider the case when t is a virtual episode. In a virtual episode, only the single pair $(s_t^\dagger, a_t^\dagger)$ is updated, and thus

$$\begin{aligned}
b_t(s) &= 6\sum_a\left(\frac{1}{\eta_{t+1}(s,a)}-\frac{1}{\eta_t(s,a)}\right)\log(T) \\
&\le \sum_a \frac{\mathbb{1}\{(s_t^\dagger,a_t^\dagger)=(s,a)\}}{54\eta_t(s,a)H} \\
&= \sum_a \frac{\mathbb{1}\{(s_t^\dagger,a_t^\dagger)=(s,a)\}}{54Hq_t(s)}\cdot\frac{1}{\max_{s',a'}\frac{\eta_t(s',a')}{q_t(s')}} \qquad \text{(since } (s_t^\dagger,a_t^\dagger)\in\arg\max_{s,a}\tfrac{\eta_t(s,a)}{q_t(s)}\text{)}
\end{aligned}$$

$$= \frac{\mathbb{1}\{s_t^\dagger = s\}}{q_t(s)} \cdot \frac{1}{54H \max_{s',a'} \frac{\eta_t(s',a')}{q_t(s')}}. \tag{62}$$

Using (38) and (62), we have

$$\begin{aligned}
B_t(s,a) &\le 3 \sum_{h=h(s)}^{H-1} \sum_{s' \in \mathcal{S}_h} q^{\overline{P}_t^B, \pi_t}(s' \mid s,a) b_t(s') \\
&\le \frac{1}{18HR_t} \sum_{h=h(s)}^{H-1} \sum_{s' \in \mathcal{S}_h} q^{\overline{P}_t^B, \pi_t}(s' \mid s,a) \frac{\mathbb{1}\{s_t^\dagger = s'\}}{q_t(s')} \\
&\le \frac{1}{18HR_t} \sum_{h=h(s)}^{H-1} \sum_{s' \in \mathcal{S}_h} q^{\overline{P}_t^B, \pi_t}(s' \mid s,a) \frac{\mathbb{1}\{s_t^\dagger = s'\}}{\overline{q}_t^{\pi_t}(s)\pi_t(a \mid s) q^{\overline{P}_t^B, \pi_t}(s' \mid s,a) + \gamma_t} \\
&\le \frac{1}{18HR_t} \sum_{h=h(s)}^{H-1} \sum_{s' \in \mathcal{S}_h} \frac{\mathbb{1}\{s_t^\dagger = s'\}}{\overline{q}_t^{\pi_t}(s)\pi_t(a \mid s) + \gamma_t} \\
&\le \frac{1}{18HR_t} \frac{1}{\overline{q}_t^{\pi_t}(s)\pi_t(a \mid s) + \gamma_t} \\
&\le \frac{3\sqrt{HS^3A}}{\gamma_t},
\end{aligned} \tag{63}$$

where in the last inequality we used $R_t > \frac{1}{50\sqrt{H^3S^3A}}$ in a virtual episode. By using (63), we also have

$$\begin{aligned}
\eta_t(s,a)\pi_t(a \mid s) B_t(s,a) &\le \frac{1}{18HR_t} \frac{\eta_t(s,a)\pi_t(a \mid s)}{\overline{q}_t^{\pi_t}(s)\pi_t(a \mid s) + \gamma_t} \\
&\le \frac{1}{18HR_t} \frac{\eta_t(s,a)}{q_t(s)} \\
&\le \frac{1}{18H} \le \frac{1}{6H},
\end{aligned}$$

where we used $\overline{q}_t^{\pi_t}(s)\pi_t(a \mid s) + \gamma_t \ge q_t(s)\pi_t(a \mid s)$, $\frac{\eta_t(s,a)}{q_t(s)} \le R_t$. This completes the proof. □

The following two lemmas are based on the Li et al. (2026, Lemmas E.5 and E.6). We adapt the proof to our choice of γ_t and our virtual-episode rule, and include the details for completeness.

Lemma G.3 (bandit setting)**.** *Suppose that the learning rates are updated according to* (14) *. Then, it holds*

$$\eta_t(s,a) \le \frac{2\sqrt{\log(T)}}{\sqrt{\sum_{\tau \le t : \tau \in \mathcal{T}_r} \frac{\zeta_\tau(s,a)}{q_\tau(s)^2}}}$$

for any episode t and state-action pair (s,a).

Proof. Let $\phi_t(s,a) = \frac{\zeta_t(s,a)}{q_t(s)^2 \log(T)}$ in real episodes and $\phi_t(s,a) = \frac{\mathbb{I}\{(s_t^\dagger, a_t^\dagger) = (s,a)\}}{324\eta_t(s,a)^2 H \log(T)}$ in virtual episodes. Then the update rule of learning rates can be written as

$$\frac{1}{\eta_{t+1}(s,a)} = \frac{1}{\eta_t(s,a)} + \eta_t(s,a)\phi_t(s,a).$$

To apply Lemma F.3, it suffices to show that $\phi_t(s,a) \le \frac{1}{\eta_t(s,a)^2}$. This is clear for virtual episodes. For real episodes,

$$\phi_t(s,a)\eta_t(s,a)^2 = \frac{\eta_t(s,a)^2 \zeta_t(s,a)}{q_t(s)^2 \log(T)} \le \frac{H^2}{\log(T)} \left(\frac{\eta_t(s,a)}{q_t(s)} \right)^2 \le \frac{H^2}{\log(T)} \cdot \frac{1}{50^2 H^3 S^3 A} \le 1,$$

which follows from $\zeta_t(s,a) \le H^2$ and $\frac{\eta_t(s,a)}{q_t(s)} \le \frac{1}{50\sqrt{H^3S^3A}}$ in real episodes.

Then, by Lemma F.3, we have

$$\eta_t(s,a) \le \frac{2}{\sqrt{\sum_{\tau \le t} \phi_\tau}} \le \frac{2\sqrt{\log(T)}}{\sqrt{\sum_{\tau \le t : \tau \in \mathcal{T}_r} \frac{\zeta_\tau(s,a)}{q_\tau(s)^2}}}.$$

□

In the following lemma only, we use T to denote the number of real episodes rather than the total number of episodes.

Lemma G.4 (bandit setting)**.** *When the number of real episodes is T, the number of virtual episodes $|\mathcal{T}_v|$ is upper bounded by*

$$|\mathcal{T}_v| \lesssim HSA \log^2(T).$$

Proof. By the definition of virtual episodes, whenever $t \in \mathcal{T}_v$, there exists a pair (s,a) such that $\frac{\eta_t(s,a)}{q_t(s)} > \frac{1}{50\sqrt{H^3S^3A}}$. Moreover, in virtual episodes, the corresponding learning rate will shrink by a factor of $\left(1 + \frac{1}{324H\log(|\mathcal{T}_r|+|\mathcal{T}_v|)}\right)$ for a state-action pair $(s_t^\dagger, a_t^\dagger)$.

Therefore, for each fixed (s,a), a virtual update on this pair can occur only while

$$\eta_t(s,a) > \frac{1}{50\sqrt{H^3S^3A}(|\mathcal{T}_r| + |\mathcal{T}_v|)}$$

Thus, the number of virtual updates on this pair is at most the number of multiplicative shrink steps needed to reduce $\eta_t(s,a)$ from η_1 to $1/(50\sqrt{H^3S^3A}(|\mathcal{T}_r| + |\mathcal{T}_v|))$. Hence,

$$\begin{aligned}
|\mathcal{T}_v| &\lesssim SA \frac{\log\left(\eta_1\sqrt{H^3S^3A}(|\mathcal{T}_r| + |\mathcal{T}_v|)\right)}{\log\left(1 + \frac{1}{324H\log(|\mathcal{T}_r|+|\mathcal{T}_v|)}\right)} \\
&\lesssim HSA \log^2(|\mathcal{T}_r| + |\mathcal{T}_v|),
\end{aligned}$$

where we used $\log(1+x) \ge x/2$ for $x \in (0,1]$, $\eta_1 = 1/(12H^3)$, and $H, S, A \le |\mathcal{T}_r|$. This implies

$$|\mathcal{T}_v| \lesssim HSA \log^2(|\mathcal{T}_r|).$$

Since $|\mathcal{T}_r| = T$, the desired bound follows. □

Lemma G.5 (bandit setting)**.** *Suppose that the event $\mathcal{E}$ holds. Then, it holds that*

$$\sum_{t \in \mathcal{T}_r} \sum_s q^{\pi_t}(s) \rho_t(s)^2 \lesssim S^3 A \iota^2.$$

Proof. Recall that $\rho_t(s) = \dfrac{q_t(s) - \underline{q}_t^{\pi_t}(s)}{q_t(s)}$ and $q_t(s) = \overline{q}_t^{\pi_t}(s) + \gamma_t$. Then, we have

$$\begin{aligned}
\sum_{t \in \mathcal{T}_r} \sum_s q^{\pi_t}(s) \rho_t(s)^2 &\le \sum_{t \in \mathcal{T}_r} \sum_s q^{\pi_t}(s) \left(\frac{\overline{q}_t^{\pi_t}(s) - \underline{q}_t^{\pi_t}(s) + \gamma_t}{q_t(s)} \right)^2 \\
&\le \sum_{t \in \mathcal{T}_r} \sum_s q^{\pi_t}(s) \left(2\left(\frac{\overline{q}_t^{\pi_t}(s) - \underline{q}_t^{\pi_t}(s)}{q_t(s)} \right)^2 + 2\left(\frac{\gamma_t}{q_t(s)} \right)^2 \right) \\
&\le 2 \sum_{t \in \mathcal{T}_r} \sum_s q^{\pi_t}(s) \left(\frac{\overline{q}_t^{\pi_t}(s) - \underline{q}_t^{\pi_t}(s)}{q_t(s)} \right)^2 + 2 \sum_{t \in \mathcal{T}_r} \sum_s \gamma_t
\end{aligned}$$

We bound the first term. By Lemma D.7, we obtain

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_r}\sum_s q^{\pi_t}(s)\left(\frac{\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)}{q_t(s)}\right)^2 \\
&\le \sum_{t\in\mathcal{T}_r}\sum_s \left(\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)\right)\frac{\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)}{q_t(s)} \\
&\le c\sqrt{HS^2A\iota^2\sum_{t\in\mathcal{T}_r}\sum_s q^{\pi_t}(s)\left(\frac{\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)}{q_t(s)}\right)^2} + cS^3A\iota^2. \quad \text{(for some absolute constant } c\text{)}
\end{aligned}$$

Now use the fact that if $x \le a\sqrt{x}+b$ with $x,a,b\ge 0$, then $x\le 2a^2+2b$, we obtain

$$\sum_{t\in\mathcal{T}_r}\sum_s q^{\pi_t}(s)\left(\frac{\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)}{q_t(s)}\right)^2 \lesssim HS^2A\iota^2 + S^3A\iota^2 \lesssim S^3A\iota^2.$$

We also have

$$2\sum_{t\in\mathcal{T}_r}\sum_s \gamma_t \le 2S(1+\log(T))$$

Combining the above bounds gives

$$\sum_{t\in\mathcal{T}_r}\sum_s q^{\pi_t}(s)\rho_t(s)^2 \lesssim S^3A\iota^2.$$

□

Lemma G.6 (bandit setting)**.** *Suppose that the event $\mathcal{E}$ holds. Then, it holds that*

$$\sum_{t\in\mathcal{T}_r}\sum_s q_t(s)\rho_t(s)^2 \lesssim S^3A\iota^2.$$

Proof. We decompose

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_r}\sum_{s'} q_t(s')\rho_t(s')^2 \\
&= \underbrace{\sum_{t\in\mathcal{T}_r}\sum_{s'} q^{\pi_t}(s')\rho_t(s')^2}_{\textbf{term}_a} + \underbrace{\sum_{t\in\mathcal{T}_r}\sum_{s'}(\overline{q}_t^{\pi_t}(s')-q^{\pi_t}(s'))\rho_t(s')^2}_{\textbf{term}_b} + \underbrace{\sum_{t\in\mathcal{T}_r}\sum_{s'}\gamma_t\rho_t(s')^2}_{\textbf{term}_c}.
\end{aligned}$$

For $\textbf{term}_a$, by Lemma G.5,

$$\textbf{term}_a \lesssim S^3A\iota^2.$$

For $\textbf{term}_b$, applying Lemma D.7 with $g_t(s)=\rho_t(s)^2$ and Lemma G.5, we obtain

$$\begin{aligned}
\textbf{term}_b &\lesssim \sqrt{HS^2A\iota^2\sum_{t\in\mathcal{T}_r}\sum_s q^{\pi_t}(s)\rho_t(s)^4} + S^3A\iota^2 \\
&\lesssim \sqrt{HS^2A\iota^2\cdot S^3A\iota^2} + S^3A\iota^2 \\
&\lesssim S^3A\iota^2,
\end{aligned}$$

For $\textbf{term}_c$, since $\rho_t(s')\le 1$,

$$\textbf{term}_c \le \sum_{t\in\mathcal{T}_r}\sum_{s'}\gamma_t \lesssim S\log(T).$$

Thus, we have

$$\sum_{t\in\mathcal{T}_r}\sum_s q_t(s)\rho_t(s)^2 \lesssim S^3A\iota^2.$$

□

G.2 Common regret analysis

Now we upper bound the right-hand side of (54).

Lemma G.7 (bandit setting)**.** *For each state $s \in \mathcal{S}$ and any comparator π, it holds that*

$$\mathbb{E}[\mathbf{reg}^{\pi}(s)] \le \mathcal{O}\big(HS^2A\log(T)\big) + \mathbb{E}\left[\sum_{t=1}^{T} b_t(s)\right] + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a \mid s)B_t(s,a)\right].$$

Proof. We will apply Lemma E.1 with $p_t = \pi_t(\cdot \mid s)$ and $\ell_t = \widehat{Q}_t(s,a) - B_t(s,a)$ for each $s \in \mathcal{S}$. To do so, in what follows, we will check the conditions of Lemma E.1. Let

$$x_t = \left\langle -\pi_t(\cdot \mid s), \widehat{Q}_t(s,\cdot) - Q^{P_t^m,\pi_t}(s,\cdot;m_t) + C_t(s,\cdot)Y_t \right\rangle = -\frac{\mathbb{I}_t(s)(L_{t,h(s)} - M_{t,h(s)})}{q_t(s)}Y_t$$

and verify that for all (s,a), $\eta_t(s,a)\pi_t(a \mid s)\Big(\widehat{Q}_t(s,a) - B_t(s,a) - Q^{P_t^m,\pi_t}(s,a;m_t) + x_t\Big) \ge -1/2$. Recall that in a virtual episode we have $Y_t = 0$ and $\ell_t(s,a) = 0$ for all state-action pairs (s,a).

Then, we have

$$\begin{aligned}
&\eta_t(s,a)\pi_t(a \mid s)\Big(\widehat{Q}_t(s,a) - B_t(s,a) - Q^{P_t^m,\pi_t}(s,a;m_t) + x_t\Big) \\
&= \eta_t(s,a)\pi_t(a \mid s)\left(\frac{\mathbb{I}_t(s,a)(L_{t,h(s)} - M_{t,h(s)})}{q_t(s)\pi_t(a \mid s)}Y_t - \frac{\mathbb{I}_t(s)(L_{t,h(s)} - M_{t,h(s)})}{q_t(s)}Y_t \right. \\
&\qquad\qquad \left. - B_t(s,a) - C_t(s,a)Y_t\right) \\
&\ge -\frac{2\eta_t(s,a)H}{q_t(s)}Y_t - \eta_t(s,a)\pi_t(a \mid s)B_t(s,a) - \eta_t(s,a)\pi_t(a \mid s)C_t(s,a)Y_t \\
&\ge -\frac{1}{3} - \frac{1}{6H} \ge -\frac{1}{2}.
\end{aligned}$$

where we used $0 \le L_{t,h(s)}, M_{t,h(s)} \le H$ and $C_t \ge 0$. The last inequality uses $\eta_t(s,a)/q_t(s) \le 1/(50\sqrt{H^3S^3A})$ in real episodes, Lemmas G.1 and G.2. Indeed,

$$\begin{aligned}
\frac{2\eta_t(s,a)H}{q_t(s)} + \eta_t(s,a)\pi_t(a \mid s)C_t(s,a) &= \frac{\eta_t(s,a)}{q_t(s)}(2H + q_t(s)\pi_t(a \mid s)C_t(s,a)) \\
&\le \frac{2H + 6HS}{50\sqrt{H^3S^3A}} \le \frac{1}{3},
\end{aligned}$$

and

$$\eta_t(s,a)\pi_t(a \mid s)B_t(s,a) \le \frac{1}{6H}.$$

Hence, by Lemma E.1, we obtain

$$\begin{aligned}
&\mathbb{E}[\mathbf{reg}^{\pi}(s)] \\
&\le \frac{A\log(AT^2)}{\eta_1} + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(AT^2)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2\Big(\Big(\widehat{Q}_t(s,a) - Q^{P_t^m,\pi_t}(s,a;m_t)\Big) - B_t(s,a) + x_t\Big)^2\right] \\
&\quad + \mathbb{E}\left[\frac{1}{T^2}\sum_{t=1}^{T}\left\langle -\pi(\cdot \mid s) + \frac{1}{A}\mathbf{1}, \widehat{Q}_t(s,\cdot) - B_t(s,\cdot)\right\rangle\right] \\
&\quad + 2\mathbb{E}\Big[\Big\|Q^{P_{T+1}^m,\pi_{T+1}}(s,\cdot;m_{T+1})\Big\|_{\infty}\Big]
\end{aligned}$$

$$\leq \mathcal{O}\big(HS^2A\log(T)\big) + 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T)\right]$$

(by $\left\|Q^{P^m_{T+1},\pi_{T+1}}(s,\cdot;m_{T+1})\right\|_\infty \leq H$)

$$+ 3\mathbb{E}\underbrace{\left[\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a\mid s)^2\left(\frac{\mathbb{I}_t(s,a)(L_{t,h(s)}-M_{t,h(s)})}{q_t(s)\pi_t(a\mid s)}Y_t - \frac{\mathbb{I}_t(s)(L_{t,h(s)}-M_{t,h(s)})}{q_t(s)}Y_t\right)^2\right]}_{\text{stability-term}}$$

$$+ 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a\mid s)^2\big(C_t(s,a)Y_t\big)^2\right] + 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a\mid s)^2 B_t(s,a)^2\right]. \tag{64}$$

Here, the second inequality follows from

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T^2}\sum_{t=1}^{T}\left\langle -\pi(\cdot\mid s) + \frac{1}{A}\mathbf{1}, \widehat{Q}_t(s,\cdot) - B_t(s,\cdot)\right\rangle\right] \\
&= \frac{1}{T^2}\left\langle -\pi(\cdot\mid s) + \frac{1}{A}\mathbf{1}, \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t\left[\widehat{Q}_t(s,\cdot) - B_t(s,\cdot)\right]\right]\right\rangle \\
&\leq \frac{1}{T^2}\left\|-\pi(\cdot\mid s) + \frac{1}{A}\mathbf{1}\right\|_1 \left\|\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t\left[\widehat{Q}_t(s,\cdot) - B_t(s,\cdot)\right]\right]\right\|_\infty \\
&\leq \frac{2T(6HST + 8\sqrt{HS^3AT})}{T^2} = 12HS + 16\sqrt{HS^3A} \leq \mathcal{O}\big(HS^2A\log(T)\big),
\end{aligned}$$

where we used $\left\|-\pi(\cdot\mid s) + \frac{1}{A}\mathbf{1}\right\|_1 \leq 2$, $\left|\mathbb{E}_t[\widehat{Q}_t(s,a)]\right| \leq 6HST$ and $B_t(s,a) \leq \frac{8\sqrt{HS^3A}}{\gamma_t}$ from Lemmas G.1 and G.2, which together imply $\left\|\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t\left[\widehat{Q}_t(s,\cdot) - B_t(s,\cdot)\right]\right]\right\|_\infty \leq T(6HST + 8\sqrt{HS^3AT})$. The initial regularization term is also absorbed into $\mathcal{O}(HS^2A\log T)$ since $A/\eta_1 = 12H^3A \leq 12HS^2A$.

We further evaluate the stability-term in the last inequality as

$$\begin{aligned}
&\eta_t(s,a)\pi_t(a\mid s)^2\left(\frac{\mathbb{I}_t(s,a)(L_{t,h(s)}-M_{t,h(s)})}{q_t(s)\pi_t(a\mid s)}Y_t - \frac{\mathbb{I}_t(s)(L_{t,h(s)}-M_{t,h(s)})}{q_t(s)}Y_t\right)^2 \\
&= \eta_t(s,a)\left(\frac{\mathbb{I}_t(s,a)(L_{t,h(s)}-M_{t,h(s)})}{q_t(s)} - \frac{\pi_t(a\mid s)\mathbb{I}_t(s)(L_{t,h(s)}-M_{t,h(s)})}{q_t(s)}\right)^2 Y_t \\
&= \frac{\eta_t(s,a)}{q_t(s)^2}(\mathbb{I}_t(s,a) - \pi_t(a\mid s)\mathbb{I}_t(s))^2(L_{t,h(s)}-M_{t,h(s)})^2 Y_t \\
&= \frac{\eta_t(s,a)\zeta_t(s,a)}{q_t(s)^2}Y_t,
\end{aligned}$$

where the last equality follows from the definition of $\zeta_t(s,a) = (\mathbb{I}_t(s,a) - \pi_t(a\mid s)\mathbb{I}_t(s))^2(L_{t,h(s)} - M_{t,h(s)})^2$. Thus, we have

$$\begin{aligned}
\text{stability-term} &= \sum_{t=1}^{T}\sum_{a}\frac{\eta_t(s,a)(\zeta_t(s,a))}{q_t(s)^2}Y_t \\
&\leq \sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T),
\end{aligned}$$

where the last inequality follows from (14).

Then, together with (64) and Lemma G.2, we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{reg}^{\pi}(s)] &\le \mathcal{O}\big(HS^2A\log(T)\big) + 6\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right)\log(T)\right] \\ &\quad + 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2 C_t(s,a)^2 Y_t\right] \\ &\quad + 3\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\eta_t(s,a)\pi_t(a \mid s)^2 B_t(s,a)^2\right] \\ &\le \mathcal{O}\big(HS^2A\log(T)\big) + \mathbb{E}\left[\sum_{t=1}^{T} b_t(s)\right] + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_{a}\pi_t(a \mid s)B_t(s,a)\right].\end{aligned}$$

□

Lemma G.8 (bandit setting)**.** *For any comparator policy π, the following two bounds hold:*

$$\begin{aligned}&\mathbb{E}\left[\sum_{s} q^{\pi}(s)\mathbf{bias}^{\pi}(s)\right] \\ &\lesssim \sqrt{H^2S^2A\iota^2\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a} q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t - m_t)^2)\right]} + H^{3/2}S^3A\iota^2\end{aligned}$$

and

$$\mathbb{E}\left[\sum_{s} q^{\pi}(s)\mathbf{bias}^{\pi}(s)\right] \lesssim \sqrt{H^3S^2A\iota^2\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a) - q^{\pi}(s,a)]_+\right]} + H^{3/2}S^3A\iota^2.$$

Proof. By the definition of $c_t(s,a)$ (12),

$$\mathbb{E}_t[c_t(s,a)] = \rho_t(s)\frac{q^{\pi_t}(s)}{q_t(s)}Q^{\pi_t}(s,a;|\ell_t - m_t|) + \rho_t(s)^2 H. \tag{65}$$

When $P \in \mathcal{P}_t$ holds , we obtain

$$\begin{aligned}\mathbb{E}_t[c_t(s,a)] &\ge \rho_t(s)\frac{q^{\pi_t}(s)}{q_t(s)}|Q^{\pi_t}(s,a;\ell_t - m_t)| + \rho_t(s)\left(1 - \frac{q^{\pi_t}(s)}{q_t(s)}\right)|Q^{\pi_t}(s,a;\ell_t - m_t)| \\ &= \rho_t(s)|Q^{\pi_t}(s,a;\ell_t - m_t)| \qquad (66) \\ &\ge \frac{q_t(s) - q^{\pi_t}(s)}{q_t(s)}|Q^{\pi_t}(s,a;\ell_t - m_t)|. \qquad (67)\end{aligned}$$

Thus, by the definition of $\widehat{Q}_t$, we have

$$\begin{aligned}&\mathbb{E}_t\Big[Q^{\pi_t}(s,a;\ell_t) - \widehat{Q}_t(s,a) + Q^{\underline{P}_t^m,\pi_t}(s,a;m_t) - Q^{\pi_t}(s,a;m_t)\Big] \\ &= Q^{\pi_t}(s,a;\ell_t) - \left(Q^{\underline{P}_t^m,\pi_t}(s,a;m_t) + \frac{q^{\pi_t}(s)}{q_t(s)}Q^{\pi_t}(s,a;\ell_t - m_t)Y_t - \mathbb{E}_t[C_t(s,a)]Y_t\right) \\ &\quad + Q^{\underline{P}_t^m,\pi_t}(s,a;m_t) - Q^{\pi_t}(s,a;m_t) \\ &= \frac{q_t(s) - q^{\pi_t}(s)Y_t}{q_t(s)}Q^{\pi_t}(s,a;\ell_t - m_t) + \mathbb{E}_t[C_t(s,a)]Y_t \\ &= \begin{cases} -Q^{\pi_t}(s,a;m_t) & \text{if } Y_t = 0, \\ \dfrac{q_t(s) - q^{\pi_t}(s)}{q_t(s)}Q^{\pi_t}(s,a;\ell_t - m_t) + \mathbb{E}_t[C_t(s,a)] & \text{if } Y_t = 1. \end{cases}\end{aligned}$$

When $Y_t = 1$ and $P \in \mathcal{P}_t$, by (67) and $C_t(s,a) \geq c_t(s,a)$ pathwise,

$$\begin{aligned}
&\frac{q_t(s) - q^{\pi_t}(s)}{q_t(s)} Q^{\pi_t}(s,a;\ell_t - m_t) + \mathbb{E}_t[C_t(s,a)] \\
&\geq -\frac{q_t(s) - q^{\pi_t}(s)}{q_t(s)} |Q^{\pi_t}(s,a;\ell_t - m_t)| + \mathbb{E}_t[c_t(s,a)] \\
&\geq 0.
\end{aligned}$$

When $Y_t = 0$, using $Q^{\pi_t}(s,a;m_t) \leq H$ and the fact that the virtual episode has zero loss, we obtain

$$\begin{aligned}
-H &\leq \mathbb{E}_t\Big[Q^{\pi_t}(s,a;\ell_t) - \widehat{Q}_t(s,a) + Q^{\underline{P}_t^m,\pi_t}(s,a;m_t) - Q^{\pi_t}(s,a;m_t)\Big] \\
&\leq 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{bias}^\pi(s)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{s,a} q^\pi(s)(\pi_t(a \mid s) - \pi(a \mid s))\mathbb{E}_t\Big[Q^{\pi_t}(s,a;\ell_t) - \widehat{Q}_t(s,a) + Q^{\underline{P}_t^m,\pi_t}(s,a;m_t) - Q^{\pi_t}(s,a;m_t)\Big]\right] \\
&\leq \mathbb{E}\left[\sum_{t\in\mathcal{T}_r} \sum_{s,a} q^\pi(s)(\pi_t(a \mid s) - \pi(a \mid s))\left(\frac{q_t(s) - q^{\pi_t}(s)}{q_t(s)} Q^{\pi_t}(s,a;\ell_t - m_t) + \mathbb{E}_t[C_t(s,a)]\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t\in\mathcal{T}_v} \sum_{s,a} q^\pi(s)\pi(a \mid s)H\right] \\
&\leq \mathbb{E}\left[\sum_{t\in\mathcal{T}_r} \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a))z_t(s,a)\right] + H^2|\mathcal{T}_v|.
\end{aligned}$$

The additive term $H^2|\mathcal{T}_v| \lesssim H^3SA\log^2 T$ by Lemma G.4 is lower order and will be absorbed into the final lower-order term. The last inequality follows by applying Lemma D.2, for each real episode $t \in \mathcal{T}_r$, with

$$L(s,a) = \frac{q_t(s) - q^{\pi_t}(s)}{q_t(s)} Q^{\pi_t}(s,a;\ell_t - m_t) + \mathbb{E}_t[C_t(s,a)].$$

Thus,

$$\begin{aligned}
z_t(s,a) = &\frac{q_t(s) - q^{\pi_t}(s)}{q_t(s)} Q^{\pi_t}(s,a;\ell_t - m_t) + \mathbb{E}_t[C_t(s,a)] \\
&- \mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\frac{q_t(s') - q^{\pi_t}(s')}{q_t(s')} Q^{\pi_t}(s',a';\ell_t - m_t) + \mathbb{E}_t[C_t(s',a')]\right].
\end{aligned} \tag{68}$$

We first reduce the bound to the high-probability event $\mathcal{E}$. By $\mathbb{E}_t[C_t(s,a)] \leq \mathcal{O}(HST)$ from Lemma G.1, $|z_t(s,a)| \leq \mathcal{O}(HST)$ holds. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{T}_r} \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a))z_t(s,a)\right] \\
&= \Pr(\mathcal{E})\mathbb{E}\left[\sum_{t\in\mathcal{T}_r} \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a))z_t(s,a) \;\middle|\; \mathcal{E}\right] \\
&\quad + \Pr(\bar{\mathcal{E}})\mathbb{E}\left[\sum_{t\in\mathcal{T}_r} \sum_{s,a} (q^{\pi_t}(s,a) - q^\pi(s,a))z_t(s,a) \;\middle|\; \bar{\mathcal{E}}\right]
\end{aligned}$$

$$\leq \Pr(\mathcal{E})\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}(q^{\pi_t}(s,a)-q^{\pi}(s,a))z_t(s,a)\;\middle|\;\mathcal{E}\right]+\mathcal{O}(\delta)\times\mathcal{O}(HT)\times\mathcal{O}(HST)$$

$$\leq \Pr(\mathcal{E})\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}(q^{\pi_t}(s,a)-q^{\pi}(s,a))z_t(s,a)\;\middle|\;\mathcal{E}\right]+\mathcal{O}\big(H^2S\big). \tag{69}$$

In the following, we work on the event $\mathcal{E}$ and bound the first term on the right-hand side of (69).

On $\mathcal{E}$, since $P\in\mathcal{P}_t$, we can upper bound $\mathbb{E}_t[c_t(s,a)]$ as

$$\begin{aligned}\mathbb{E}_t[c_t(s,a)] &\leq \rho_t(s)Q^{\pi_t}(s,a;|\ell_t-m_t|)+\rho_t(s)^2H\\ &\leq \rho_t(s)\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)+\rho_t(s)^2H\\ &\leq \frac{\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)}{q_t(s)}\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)+\frac{\gamma_tH}{q_t(s)}+\rho_t(s)^2H.\end{aligned} \tag{70}$$

We next derive lower and upper bounds on $z_t(s,a)$. Since $P\in\mathcal{P}_t$ on $\mathcal{E}$, the definition of $C_t(s,a)$ (13) implies

$$\mathbb{E}_t[C_t(s,a)]\geq\mathbb{E}_t[c_t(s,a)]+\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[\mathbb{E}_t[c_t(s',a')+C_t(s',a')]]. \tag{71}$$

Combining (67), (68) and (71), we have for any real episode $t\in\mathcal{T}_r$

$$\begin{aligned}z_t(s,a)&\geq\frac{q_t(s)-q^{\pi_t}(s)}{q_t(s)}Q^{\pi_t}(s,a;\ell_t-m_t)+\mathbb{E}_t[c_t(s,a)]\\ &\quad+\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\mathbb{E}_t[c_t(s',a')]-\frac{q_t(s')-q^{\pi_t}(s')}{q_t(s')}Q^{\pi_t}(s',a';\ell_t-m_t)\right]\\ &\geq 0.\end{aligned} \tag{72}$$

On the other hand, let $\overline{P}_t^C$ be the measurable transition kernel that attains the maximum in (13) for the realized table c_t. Then, for any real episode $t\in\mathcal{T}_r$

$$\begin{aligned}z_t(s,a)&=\frac{q_t(s)-q^{\pi_t}(s)}{q_t(s)}Q^{\pi_t}(s,a;\ell_t-m_t)+\mathbb{E}_t[c_t(s,a)]\\ &\quad+\mathbb{E}_t\left[\mathbb{E}_{s'\sim\overline{P}_t^C(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[c_t(s',a')+C_t(s',a')]\right]\\ &\quad-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\frac{q_t(s')-q^{\pi_t}(s')}{q_t(s')}Q^{\pi_t}(s',a';\ell_t-m_t)+\mathbb{E}_t[C_t(s',a')]\right]\\ &\leq 2\mathbb{E}_t[c_t(s,a)]+\mathbb{E}_t\left[\mathbb{E}_{s'\sim\overline{P}_t^C(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[c_t(s',a')+C_t(s',a')]\right]\\ &\quad-\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[-\mathbb{E}_t[c_t(s',a')]+\mathbb{E}_t[C_t(s',a')]] \qquad \text{(by (66))}\\ &\leq 2\mathbb{E}_t[c_t(s,a)]+2\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[\mathbb{E}_t[c_t(s',a')]]\\ &\quad+\mathbb{E}_t\left[\sum_{s',a'}\left|\overline{P}_t^C(s'\mid s,a)-P(s'\mid s,a)\right|\pi_t(a'\mid s')(c_t(s',a')+C_t(s',a'))\right].\end{aligned} \tag{73}$$

By the definition of $C_t(s,a)$ (13), we have

$$\sum_a\pi_t(a\mid s)(c_t(s,a)+C_t(s,a))=2\sum_{s',a'}q^{\overline{P}_t^C,\pi_t}(s',a'\mid s)c_t(s',a'). \tag{74}$$

Combining (73) and (74), we obtain for any real episode $t\in\mathcal{T}_r$

$$\begin{aligned}z_t(s,a)&\leq 2\mathbb{E}_t[c_t(s,a)]+2\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[\mathbb{E}_t[c_t(s',a')]]\\ &\quad+2\mathbb{E}_t\left[\sum_{s'}\left|\overline{P}_t^C(s'\mid s,a)-P(s'\mid s,a)\right|\sum_{s'',a''}q^{\overline{P}_t^C,\pi_t}(s'',a''\mid s')c_t(s'',a'')\right].\end{aligned} \tag{75}$$

Therefore, by (72) and (75),

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}_r}\sum_{s,a}(q^{\pi_t}(s,a)-q^{\pi}(s,a))z_t(s,a)\\
&\le \sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+ z_t(s,a) && \text{(by (72))}\\
&\le \underbrace{2\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_t[c_t(s,a)]}_{\textbf{term}_1}\\
&\quad \underbrace{+2\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[\mathbb{E}_t[c_t(s',a')]]}_{\textbf{term}_2}\\
&\quad \underbrace{+2\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_t\left[\sum_{s'}\left|\overline{P}_t^{C}(s'\mid s,a)-P(s'\mid s,a)\right|\sum_{s'',a''}q^{\overline{P}_t^{C},\pi_t}(s'',a''\mid s')c_t(s'',a'')\right]}_{\textbf{term}_3}.
\end{aligned}
$$

Let

$$
G^{\pi_t}\coloneqq \min\left\{\sum_{s,a}q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2),\ H\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\right\}. \tag{76}
$$

We first bound $\textbf{term}_1$. By (70),

$$
\begin{aligned}
\textbf{term}_1 &= \sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_t[c_t(s,a)]\\
&\le \underbrace{\sum_{t\in\mathcal{T}_r}\sum_{s,a}\left(\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)\right)\frac{[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}{q_t(s)}\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)}_{\textbf{term}_{1a}}\\
&\quad \underbrace{+H\sum_{t\in\mathcal{T}_r}\sum_{s,a}\frac{\gamma_t}{q_t(s)}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}_{\textbf{term}_{1b}}+\underbrace{H\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\rho_t(s)^2}_{\textbf{term}_{1c}}.
\end{aligned}
$$

We first control $\textbf{term}_{1a}$. Define

$$
\begin{aligned}
g_t^{(1)}(s)&\coloneqq \sum_a\frac{[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}{q_t(s)}\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)\\
&\le H.
\end{aligned}
$$

Thus, by Lemma D.7,

$$
\begin{aligned}
\textbf{term}_{1a}&=\sum_{t\in\mathcal{T}_r}\sum_{s}\left(\overline{q}_t^{\pi_t}(s)-\underline{q}_t^{\pi_t}(s)\right)g_t^{(1)}(s)\\
&\lesssim \sqrt{HS^2A\iota^2\sum_{t\in\mathcal{T}_r}\sum_{s}q^{\pi_t}(s)\left(g_t^{(1)}(s)\right)^2}+HS^3A\iota^2.
\end{aligned}
$$

Here, by the Cauchy–Schwarz inequality,

$$
\begin{aligned}
&\sum_s q^{\pi_t}(s)\left(g_t^{(1)}(s)\right)^2\\
&\le \sum_s q^{\pi_t}(s)\left(\sum_a\frac{[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}{q_t(s)}\right)\left(\sum_a\frac{[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}{q_t(s)}\left(\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)\right)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&\le \sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\left(\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;|\ell_t-m_t|)\right)^2\\
&\le H\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2) \qquad \text{(by (49))}\\
&\le HG^{\pi_t},
\end{aligned}
$$

Therefore,

$$
\mathbf{term}_{1a} \lesssim \sqrt{H^2S^2A\iota^2\sum_{t\in\mathcal{T}_r} G^{\pi_t}} + HS^3A\iota^2. \tag{77}
$$

Next, $\mathbf{term}_{1b}$ can be bounded by

$$
\begin{aligned}
\mathbf{term}_{1b} &\le H\sum_{t\in\mathcal{T}_r}\sum_{s}\gamma_t\frac{q^{\pi_t}(s)}{q_t(s)}\\
&\le HS\sum_{t\in\mathcal{T}_r}\gamma_t\\
&\lesssim HS\log(T).
\end{aligned} \tag{78}
$$

Finally, $\mathbf{term}_{1c}$ can be bounded by using Lemma G.5,

$$
\begin{aligned}
\mathbf{term}_{1c} &\le H\sum_{t\in\mathcal{T}_r}\sum_{s} q^{\pi_t}(s)\rho_t(s)^2\\
&\lesssim HS^3A\iota^2.
\end{aligned} \tag{79}
$$

Combining (77)–(79), we obtain

$$
\mathbf{term}_1 \lesssim \sqrt{H^2S^2A\iota^2\sum_{t\in\mathcal{T}_r} G^{\pi_t}} + HS^3A\iota^2. \tag{80}
$$

We next bound $\mathbf{term}_2$. By (70), we have

$$
\begin{aligned}
\mathbf{term}_2 &= \sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}[\mathbb{E}_t[c_t(s',a')]]\\
&\le \underbrace{\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_t(\cdot|s')}\left[\frac{\overline{q}_t^{\pi_t}(s')-\underline{q}_t^{\pi_t}(s')}{q_t(s')}\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s',a';|\ell_t-m_t|)\right]}_{\mathbf{term}_{2a}}\\
&\quad+\underbrace{H\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\frac{\gamma_t}{q_t(s')}\right]}_{\mathbf{term}_{2b}}\\
&\quad+\underbrace{H\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\rho_t(s')^2\right]}_{\mathbf{term}_{2c}}.
\end{aligned}
$$

We first bound $\mathbf{term}_{2a}$. Define

$$
\begin{aligned}
g_t^{(2)}(s') &\coloneqq \frac{\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+P(s'\mid s,a)}{q_t(s')}\sum_{a'}\pi_t(a'\mid s')\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s',a';|\ell_t-m_t|)\\
&\le \sum_{a'}\pi_t(a'\mid s')\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s',a';|\ell_t-m_t|)\\
&\le H.
\end{aligned}
$$

Therefore, by Lemma D.7,

$$\begin{aligned}\textbf{term}_{2a} &= \sum_{t\in\mathcal{T}_r}\sum_{s'}\Big(\overline{q}_t^{\pi_t}(s') - \underline{q}_t^{\pi_t}(s')\Big)g_t^{(2)}(s') \\ &\lesssim \sqrt{HS^2A\iota^2\sum_{t\in\mathcal{T}_r}\sum_{s'}q^{\pi_t}(s')\Big(g_t^{(2)}(s')\Big)^2} + HS^3A\iota^2.\end{aligned}$$

Moreover, by using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}&\sum_{s'}q^{\pi_t}(s')\Big(g_t^{(2)}(s')\Big)^2 \\ &\le \sum_{s'}q^{\pi_t}(s')\left(\sum_{a'}\pi_t(a'\mid s')\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s',a';|\ell_t-m_t|)\right)^2 \\ &\le \sum_{s',a'}q^{\pi_t}(s',a')\left(\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s',a';|\ell_t-m_t|)\right)^2 \\ &\le H\sum_{s',a'}q^{\pi_t}(s',a')\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s',a';(\ell_t-m_t)^2), \qquad (81)\end{aligned}$$

where the last inequality follows from (49).

Simultaneously, it holds that

$$\begin{aligned}&\sum_{s'}q^{\pi_t}(s')\Big(g_t^{(2)}(s')\Big)^2 \\ &\le H\sum_{s'}q^{\pi_t}(s')\sum_{s,a,a'}P(s'\mid s,a)\pi_t(a'\mid s')\frac{[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+}{q_t(s')}\max_{\widetilde{P}\in\mathcal{P}_t}Q^{\widetilde{P},\pi_t}(s',a';|\ell_t-m_t|) \\ &\le H^2\sum_{s,a,s',a'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+P(s'\mid s,a)\pi_t(a'\mid s') \\ &\le H^2\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+. \qquad (82)\end{aligned}$$

From (81) and (82), we have

$$\sum_{s'}q^{\pi_t}(s')\Big(g_t^{(2)}(s')\Big)^2 \le HG^{\pi_t}.$$

Hence,

$$\textbf{term}_{2a} \lesssim \sqrt{H^2S^2A\iota^2\sum_{t\in\mathcal{T}_r}G^{\pi_t}} + HS^3A\iota^2. \qquad (83)$$

Next, $\textbf{term}_{2b}$ can be bounded by

$$\begin{aligned}\textbf{term}_{2b} &\le H\sum_{t\in\mathcal{T}_r}\sum_{s'}\gamma_t\frac{\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+P(s'\mid s,a)}{q_t(s')} \\ &\le HS\sum_{t\in\mathcal{T}_r}\gamma_t \\ &\lesssim HS\log(T). \qquad (84)\end{aligned}$$

Finally, Lemma G.5 gives

$$\begin{aligned}\textbf{term}_{2c} &\le H\sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+P(s'\mid s,a)\rho_t(s')^2 \\ &\le H\sum_{t\in\mathcal{T}_r}\sum_{s'}q^{\pi_t}(s')\rho_t(s')^2 \\ &\lesssim HS^3A\iota^2. \qquad (85)\end{aligned}$$

Combining (83)–(85), we obtain

$$\mathbf{term}_2 \lesssim \sqrt{H^2S^2A\iota^2 \sum_{t\in\mathcal{T}_r} G^{\pi_t}} + HS^3A\iota^2. \tag{86}$$

We finally bound $\mathbf{term}_3$. By Lemma C.2 and (65), we have

$\mathbf{term}_3$

$$\begin{aligned}
&= \sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\mathbb{E}_t\left[\sum_{s'}\left|\overline{P}_t^C(s'\mid s,a)-P(s'\mid s,a)\right|\sum_{s'',a''} q^{\overline{P}_t^C,\pi_t}(s'',a''\mid s')c_t(s'',a'')\right]\\
&\lesssim \sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\left(\sqrt{\frac{P(s'\mid s,a)\iota}{n_t(s,a)}}+\frac{\iota}{n_t(s,a)}\right)\mathbb{E}_t\left[\sum_{s'',a''} q^{\overline{P}_t^C,\pi_t}(s'',a''\mid s')c_t(s'',a'')\right]\\
&\le \sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\left(\sqrt{\frac{P(s'\mid s,a)\iota}{n_t(s,a)}}+\frac{\iota}{n_t(s,a)}\right)\\
&\quad\times\sum_{s''}\max_{\widetilde{P}\in\mathcal{P}_t} q^{\widetilde{P},\pi_t}(s''\mid s')\sum_{a''}\pi_t(a''\mid s'')\mathbb{E}_t[c_t(s'',a'')]\\
&\lesssim H\sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\left(\sqrt{\frac{P(s'\mid s,a)\iota}{n_t(s,a)}}+\frac{\iota}{n_t(s,a)}\right)\sum_{s''}\max_{\widetilde{P}\in\mathcal{P}_t} q^{\widetilde{P},\pi_t}(s''\mid s')\rho_t(s'').
\end{aligned} \tag{87}$$

Applying the Cauchy–Schwarz inequality to (87) gives

$$\begin{aligned}
\mathbf{term}_3 \lesssim\ & H\sqrt{\sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+\frac{\iota}{n_t(s,a)}}\\
&\times\sqrt{\sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+P(s'\mid s,a)\left(\sum_{s''}\max_{\widetilde{P}\in\mathcal{P}_t} q^{\widetilde{P},\pi_t}(s''\mid s')\rho_t(s'')\right)^2}\\
&+HS\sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}\frac{q^{\pi_t}(s,a)\iota}{n_t(s,a)}.
\end{aligned}$$

To bound the second factor, for each layer h, we have

$$\begin{aligned}
&\sum_{(s,a,s')\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}} q^{\pi_t}(s,a)P(s'\mid s,a)\max_{\widetilde{P}\in\mathcal{P}_t} q^{\widetilde{P},\pi_t}(s''\mid s')\\
&=\max_{\widetilde{P}\in\mathcal{P}_t}\sum_{(s,a,s')\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}} q^{\pi_t}(s,a)P(s'\mid s,a)q^{\widetilde{P},\pi_t}(s''\mid s')\le \overline{q}_t^{\pi_t}(s'').
\end{aligned} \tag{88}$$

Thus, we have

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_r}\sum_{s,a,s'}[q^{\pi_t}(s,a)-q^{\pi}(s,a)]_+P(s'\mid s,a)\left(\sum_{s''}\max_{\widetilde{P}\in\mathcal{P}_t} q^{\widetilde{P},\pi_t}(s''\mid s')\rho_t(s'')\right)^2\\
&\le S\sum_{t\in\mathcal{T}_r}\sum_{h=0}^{H-1}\sum_{(s,a,s')\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}} q^{\pi_t}(s,a)P(s'\mid s,a)\sum_{s''}\rho_t(s'')^2\max_{\widetilde{P}\in\mathcal{P}_t} q^{\widetilde{P},\pi_t}(s''\mid s')\\
&\qquad\qquad\text{(by the Cauchy–Schwarz inequality)}\\
&\le SH\sum_{t\in\mathcal{T}_r}\sum_{s''}\overline{q}_t^{\pi_t}(s'')\rho_t(s'')^2 \qquad\text{(by (88))}\\
&\lesssim HS^4A\iota^2. \qquad\text{(by Lemma G.6)}
\end{aligned}$$

Substituting these bounds and applying Lemma C.3, we obtain

$$\begin{aligned}
\mathbf{term}_3 &\lesssim H\sqrt{S^2A\iota^2\cdot HS^4A\iota^2}+HS^3A\iota^2\\
&\lesssim H^{3/2}S^3A\iota^2.
\end{aligned} \tag{89}$$

Combining (69), (80), (86) and (89) with the virtual-episode contribution and applying Jensen's inequality gives

$$\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{bias}^\pi(s)\right] \lesssim \sqrt{H^2S^2A\iota^2\mathbb{E}\left[\sum_{t\in\mathcal{T}_r} G^{\pi_t}\right]} + H^{3/2}S^3A\iota^2.$$

Using either entry of the minimum in (76) proves the two asserted bounds.

□

Lemma G.9 (bandit setting)**.** *For any comparator policy π, it holds that*

$$\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{error}^\pi(s)\right] \lesssim \sqrt{S^2A\iota^2\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(m)]} + H^2S^2A\iota^2. \tag{90}$$

Also,

$$\begin{aligned}&\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{error}^\pi(s)\right]\\ &\lesssim \sum_{h=0}^{H-1}\sqrt{H^2|S_h||S_{h+1}|A\iota^2\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a)-q^\pi(s,a)]_+\right]} + H^2S^2A\iota^2.\end{aligned} \tag{91}$$

Proof. The proof is identical to that of Lemma F.6, with the loss sequence ℓ_t replaced by the prediction sequence m_t. The only additional point is that we first separate the contribution of virtual episodes.

$$\begin{aligned}&\sum_s q^\pi(s)\mathbf{error}^\pi(s)\\ &= \sum_s q^\pi(s)\sum_{t=1}^{T}\sum_a(\pi_t(a\mid s)-\pi(a\mid s))\Big(Q^{\pi_t}(s,a;m_t)-Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)\Big)\\ &\le \sum_s q^\pi(s)\sum_{t\in\mathcal{T}_r}\sum_a(\pi_t(a\mid s)-\pi(a\mid s))\Big(Q^{\pi_t}(s,a;m_t)-Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)\Big) + 2H^2|\mathcal{T}_v|\\ &\le \sum_s q^\pi(s)\sum_{t\in\mathcal{T}_r}\sum_a(\pi_t(a\mid s)-\pi(a\mid s))\Big(Q^{\pi_t}(s,a;m_t)-Q^{\underline{P}_t^m,\pi_t}(s,a;m_t)\Big)\\ &\quad + \mathcal{O}(H^3SA\log^2(T)),\end{aligned}$$

where we used Lemma G.4. The resulting virtual-episode contribution is absorbed into the lower-order term. It remains to bound the first term in expectation, which follows from Lemma F.6 applied to the real episodes.

Applying the first bound of Lemma F.6 with ℓ_t replaced by m_t gives, on $\mathcal{E}$,

$$\sum_s q^\pi(s)\mathbf{error}^\pi(s) \lesssim \sqrt{S^2A\iota^2Q_{\mathrm{trans}}^{\pi_{1:T}}(m)} + H^2S^2A\iota^2.$$

Taking expectations, using Jensen's inequality, and combining with the trivial bound $\mathcal{O}(H^2T)$ on $\bar{\mathcal{E}}$ with $\delta=\mathcal{O}(1/T^2)$, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{error}^\pi(s)\right] &\lesssim \sqrt{S^2A\iota^2\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(m)]} + H^2S^2A\iota^2 + \delta H^2T\\ &\lesssim \sqrt{S^2A\iota^2\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(m)]} + H^2S^2A\iota^2,\end{aligned}$$

which proves (90).

Similarly, applying the second bound of Lemma F.6 with ℓ_t replaced by m_t gives, on $\mathcal{E}$,

$$\sum_s q^\pi(s)\mathbf{error}^\pi(s) \lesssim \sum_{h=0}^{H-1} \sqrt{H^2|S_h||S_{h+1}|A\iota^2 \sum_{t\in\mathcal{T}_r}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} [q^{\pi_t}(s,a)-q^\pi(s,a)]_+} + H^2S^2A\iota^2.$$

Taking expectations, applying Jensen's inequality, and combining with the trivial bound $\mathcal{O}(H^2T)$ on $\bar{\mathcal{E}}$ with $\delta=\mathcal{O}(1/T^2)$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_s q^\pi(s)\mathbf{error}^\pi(s)\right] \\
&\lesssim \sum_{h=0}^{H-1} \sqrt{H^2|S_h||S_{h+1}|A\iota^2\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} [q^{\pi_t}(s,a)-q^\pi(s,a)]_+\right]} + H^2S^2A\iota^2 + \delta H^2T \\
&\lesssim \sum_{h=0}^{H-1} \sqrt{H^2|S_h||S_{h+1}|A\iota^2\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} [q^{\pi_t}(s,a)-q^\pi(s,a)]_+\right]} + H^2S^2A\iota^2.
\end{aligned}$$

This proves (91). □

Lemma G.10 (bandit setting)**.** *It holds that*

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a} q^{\overline{P}_t^B,\pi_t}(s)\eta_t(s,a)\pi_t(a\mid s)^2C_t(s,a)^2\right] \lesssim H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{1}{2}}\iota^2.$$

Proof. By the decomposition of $q_t(s)\pi_t(a\mid s)C_t(s,a)$ in (57), we have

$$\begin{aligned}
&q_t(s)\pi_t(a\mid s)C_t(s,a) \\
&\le 2\sum_{s',a'} \frac{q_t(s)\pi_t(a\mid s)q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a)}{q_t(s')\pi_t(a'\mid s')}\rho_t(s')\mathbb{I}_t(s',a')D_{t,h(s')} \\
&\quad + 2Hq_t(s)\pi_t(a\mid s)\sum_{s',a'} q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a)\rho_t(s')^2.
\end{aligned}$$

Thus, by the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{s,a} q_t(s)^2\pi_t(a\mid s)^2C_t(s,a)^2\right] \\
&\lesssim H^3\mathbb{E}_t\left[\sum_{s',a'}\rho_t(s')^2\mathbb{I}_t(s',a')\sum_{s,a}\left(\frac{q_t(s)\pi_t(a\mid s)q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a)}{q_t(s')\pi_t(a'\mid s')}\right)^2\right] \\
&\quad + H^3\mathbb{E}_t\left[\sum_{s',a'}\rho_t(s')^2\sum_{s,a} q_t(s)^2\pi_t(a\mid s)^2q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a)\right]. \qquad (92)
\end{aligned}$$

For the first term of (92), we have

$$\sum_{s,a}\left(\frac{q_t(s)\pi_t(a\mid s)q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a)}{q_t(s')\pi_t(a'\mid s')}\right)^2 \le \sum_s\left(\sum_a \frac{q_t(s)\pi_t(a\mid s)q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a)}{q_t(s')\pi_t(a'\mid s')}\right)^2 \le S,$$

where the same argument as for (56) gives

$$\begin{aligned}
&\sum_a q_t(s)\pi_t(a\mid s)q^{\overline{P}_t^C,\pi_t}(s',a'\mid s,a) \\
&= (\overline{q}_t^{\pi_t}(s)+\gamma_t)\sum_a \pi_t(a\mid s)q^{\overline{P}_t^C,\pi_t}(s'\mid s,a)\pi_t(a'\mid s') \\
&\le (\overline{q}_t^{\pi_t}(s')+\gamma_t)\pi_t(a'\mid s') = q_t(s')\pi_t(a'\mid s').
\end{aligned}$$

For the second term of (92), since $q_t(s)\pi_t(a \mid s) \le 2$, we have

$$\sum_{s,a} q_t(s)^2 \pi_t(a \mid s)^2 q^{\overline{P}_t^C, \pi_t}(s', a' \mid s, a) \le 2S q_t(s') \pi_t(a' \mid s').$$

Hence, using $\mathbb{E}_t[\mathbb{I}_t(s', a')] = q^{\pi_t}(s', a')$,

$$\mathbb{E}_t\left[\sum_{s,a} q_t(s)^2 \pi_t(a \mid s)^2 C_t(s,a)^2\right] \lesssim H^3 S \sum_{s'} (q^{\pi_t}(s') + q_t(s'))\rho_t(s')^2. \tag{93}$$

Combining Lemma G.6 and Lemma G.5 with the trivial bound $\sum_{s'} (q^{\pi_t}(s') + q_t(s'))\rho_t(s')^2 \le 3S$, and (93), we get

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \sum_{s,a} q_t(s)^2 \pi_t(a \mid s)^2 C_t(s,a)^2\right] \lesssim H^3 S\big(S^3 A\iota^2 + \delta ST\big) \lesssim H^3 S^4 A \iota^2. \tag{94}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \sum_{s,a} q^{\overline{P}_t^B, \pi_t}(s) \eta_t(s,a) \pi_t(a \mid s)^2 C_t(s,a)^2\right] \\
&= \mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \sum_{s,a} q^{\overline{P}_t^B, \pi_t}(s) \frac{\eta_t(s,a) q_t(s)^2 \pi_t(a \mid s)^2 C_t(s,a)^2}{q_t(s)^2}\right] \\
&\le \mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \sum_{s,a} q_t(s)^2 \pi_t(a \mid s)^2 C_t(s,a)^2 \max_{s,a} \frac{\eta_t(s,a)}{q_t(s)}\right] \\
&\le \mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \sum_{s,a} q_t(s)^2 \pi_t(a \mid s)^2 C_t(s,a)^2\right] \cdot \frac{1}{50\sqrt{H^3 S^3 A}} && \text{(by } t \text{ is a real episode)} \\
&\lesssim H^3 S^4 A \iota^2 \cdot \frac{1}{\sqrt{H^3 S^3 A}} && \text{(by (94))} \\
&= H^{\frac{3}{2}} S^{\frac{5}{2}} A^{\frac{1}{2}} \iota^2.
\end{aligned}$$

□

Lemma G.11 (bandit setting)**.** *It holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T} V^{\overline{P}_t^B, \pi_t}(s_0; b_t)\right] \lesssim \log(T) \sum_{s,a} \sqrt{\mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \zeta_t(s,a)\right]} + H^{\frac{3}{2}} S^{\frac{5}{2}} A^{\frac{3}{2}} \iota^2.$$

Proof. By Lemma G.3 and $\eta_t(s,a) \le \eta_1 \le \sqrt{\log(T)}$,

$$\eta_t(s,a) \lesssim \frac{\sqrt{\log(T)}}{\sqrt{1 + \sum_{\tau \le t : \tau \in \mathcal{T}_r} \zeta_\tau(s,a)/q_\tau(s)^2}}. \tag{95}$$

We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} V^{\overline{P}_t^B, \pi_t}(s_0; b_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{s} q^{\overline{P}_t^B, \pi_t}(s) b_t(s)\right] \\
&\lesssim \mathbb{E}\left[\sum_{t \in \mathcal{T}_r} \sum_{s,a} q^{\overline{P}_t^B, \pi_t}(s) \left(\frac{1}{\eta_{t+1}(s,a)} - \frac{1}{\eta_t(s,a)}\right) \log(T)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\sum_{t\in\mathcal{T}_v}\sum_{s,a} q^{\overline{P}_t^B,\pi_t}(s)\left(\frac{1}{\eta_{t+1}(s,a)}-\frac{1}{\eta_t(s,a)}\right)\log(T)\right] \\
&+ \mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a} q^{\overline{P}_t^B,\pi_t}(s)\eta_t(s,a)\pi_t(a\mid s)^2 C_t(s,a)^2\right] \\
\lesssim\ & \mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a} q^{\overline{P}_t^B,\pi_t}(s)\frac{\eta_t(s,a)\zeta_t(s,a)}{q_t(s)^2}\right] + \mathbb{E}\left[\sum_{t\in\mathcal{T}_v}\sum_{s,a} q^{\overline{P}_t^B,\pi_t}(s)\frac{\mathbb{1}\{(s_t^\dagger,a_t^\dagger)=(s,a)\}}{\eta_t(s,a)H}\right] \\
&+ H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{1}{2}}\iota^2 && \text{(by Lemma G.10)} \\
\le\ & \mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}\frac{\eta_t(s,a)\zeta_t(s,a)}{q_t(s)}\right] + \mathbb{E}\left[\sum_{t\in\mathcal{T}_v}\frac{q^{\overline{P}_t^B,\pi_t}(s_t^\dagger)}{\eta_t(s_t^\dagger,a_t^\dagger)H}\right] + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{1}{2}}\iota^2 \\
\lesssim\ & \mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}\frac{\eta_t(s,a)\zeta_t(s,a)}{q_t(s)}\right] + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\iota^2 \\
& && \text{(by the virtual-episode condition and Lemma G.4)} \\
\lesssim\ & \sqrt{\log(T)}\,\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}\frac{\frac{\sqrt{\zeta_t(s,a)}}{q_t(s)}\sqrt{\zeta_t(s,a)}}{\sqrt{1+\sum_{\tau\le t:\tau\in\mathcal{T}_r}\frac{\zeta_\tau(s,a)}{q_\tau(s)^2}}}\right] + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\iota^2 && \text{(by (95))} \\
\lesssim\ & \sqrt{\log(T)}\,\mathbb{E}\left[\sum_{s,a}\sqrt{\sum_{t\in\mathcal{T}_r}\frac{\zeta_t(s,a)/q_t(s)^2}{1+\sum_{\tau\le t:\tau\in\mathcal{T}_r}\zeta_\tau(s,a)/q_\tau(s)^2}}\sqrt{\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)}\right] + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\iota^2 \\
& && \text{(by the Cauchy–Schwarz inequality)} \\
\lesssim\ & \log(T)\,\mathbb{E}\left[\sum_{s,a}\sqrt{\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)}\right] + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\iota^2,
\end{aligned}
$$

where the last inequality follows from

$$
\sqrt{\sum_{t\in\mathcal{T}_r}\frac{\zeta_t(s,a)/q_t(s)^2}{1+\sum_{\tau\le t:\tau\in\mathcal{T}_r}\zeta_\tau(s,a)/q_\tau(s)^2}} \le \sqrt{\log\left(1+\sum_{t\in\mathcal{T}_r}\frac{\zeta_t(s,a)}{q_t(s)^2}\right)} \lesssim \sqrt{\log(T)},
$$

using $q_t(s)\ge 1/T$ and $\zeta_t(s,a)\le H^2$. Finally, the desired bound follows from Jensen's inequality. □

We now summarize the bounds in the bandit setting proved above and plug them into the regret decomposition. By the definition of the regret decomposition in (54),

$$
\begin{aligned}
&\mathbb{E}\left[\sum_s q^\pi(s)\sum_{t,a}(\pi_t(a\mid s)-\pi(a\mid s))(Q^{\pi_t}(s,a;\ell_t)-B_t(s,a))\right] \\
&= \mathbb{E}\left[\sum_s q^\pi(s)\cdot\mathbf{reg}^\pi(s)\right] + \mathbb{E}\left[\sum_s q^\pi(s)\cdot\mathbf{bias}^\pi(s)\right] + \mathbb{E}\left[\sum_s q^\pi(s)\cdot\mathbf{error}^\pi(s)\right] \\
&\le \mathcal{O}\big(H^2S^2A\log(T)\big) + \mathbb{E}\left[\sum_{t=1}^T\sum_s q^\pi(s)b_t(s)\right] + \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^T\sum_{s,a} q^\pi(s)\pi_t(a\mid s)B_t(s,a)\right] \\
&\quad + \mathbb{E}\left[\sum_s q^\pi(s)\cdot\mathbf{bias}^\pi(s)\right] + \mathbb{E}\left[\sum_s q^\pi(s)\cdot\mathbf{error}^\pi(s)\right]
\end{aligned}
$$

where we used Lemma G.7. Combining this with Lemma F.1 and $\delta = \frac{1}{T^2}$, for any comparator policy π, Algorithm 1 guarantees

$$\begin{aligned}
\mathrm{Reg}_T(\pi) &\le \mathcal{O}\big(H^2S^2A\log(T)\big) + 3\mathbb{E}\left[\sum_{t=1}^{T} V^{\overline{P}_t^B,\pi_t}(s_0;b_t)\right] \\
&\quad + \mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{bias}^\pi(s)\right] + \mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{error}^\pi(s)\right] \\
&\lesssim H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\iota^2 + \log(T)\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)\right]} \\
&\quad + \mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{bias}^\pi(s)\right] + \mathbb{E}\left[\sum_s q^\pi(s)\cdot \mathbf{error}^\pi(s)\right],
\end{aligned} \tag{96}$$

The contribution of $\bar{\mathcal{E}}_1$ is $\mathcal{O}(H^{3/2}S^{3/2}\sqrt{A})$ by Lemma G.2 and $\delta = T^{-2}$, and is absorbed into $\mathcal{O}(H^2S^2A\log(T))$. The last inequality follows from Lemma G.11.

G.3 Proof for the adversarial regime

Theorem G.12 (bandit setting)**.** *For any comparator policy π, Algorithm 1 guarantees*

$$\begin{aligned}
\mathrm{Reg}_T(\pi) &\lesssim \sqrt{H^3S^2A\log^2(T)\min\{L(\pi), HT-L(\pi), Q_\infty, V_1\}} \\
&\quad + \sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]} \\
&\quad + \big(H^3S^2A + H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\big)\log^2(T).
\end{aligned}$$

Proof. By (96) and Lemmas G.8 and G.9, for any comparator policy π, we have

$$\begin{aligned}
\mathrm{Reg}_T(\pi) &\lesssim \big(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\big)\iota^2 + \log(T)\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)\right]} \\
&\quad + \sqrt{H^2S^2A\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a} q^{\pi_t}(s,a)\max_{\widetilde{P}\in\mathcal{P}_t} Q^{\widetilde{P},\pi_t}(s,a;(\ell_t-m_t)^2)\right]} \\
&\quad + \sqrt{S^2A\iota^2\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(m)]}.
\end{aligned}$$

Here, since $\delta = \mathcal{O}(1/T^2)$, we have $\iota \lesssim \log(T)$. Hence,

$$\begin{aligned}
\mathrm{Reg}_T(\pi) &\lesssim \big(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\big)\log^2(T) + \underbrace{\log(T)\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)\right]}}_{\mathbf{term}_1} \\
&\quad + \underbrace{\sqrt{H^3S^2A\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r} V^{\pi_t}(s_0;(\ell_t-m_t)^2)\right]}}_{\mathbf{term}_2} \\
&\quad + \underbrace{\sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(m)]}}_{\mathbf{term}_3},
\end{aligned} \tag{97}$$

where the last inequality uses Lemma D.9 on $\mathcal{E}$ together with the trivial bound $\mathcal{O}(H^2T)$ on $\bar{\mathcal{E}}$, whose contribution is absorbed since $\delta = \mathcal{O}(1/T^2)$.

For $\mathbf{term}_1$, by the definition of $\zeta_t(s,a)$, we have

$$\begin{aligned}
\mathbf{term}_1 &= \sum_{s,a}\sqrt{\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}(\mathbb{I}_t(s,a)-\pi_t(a\mid s)\mathbb{I}_t(s))^2(L_{t,h(s)}-M_{t,h(s)})^2\right]} \\
&\le \sqrt{SA\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}(\mathbb{I}_t(s,a)-\pi_t(a\mid s)\mathbb{I}_t(s))^2(L_{t,h(s)}-M_{t,h(s)})^2\right]} \\
&\qquad\qquad\text{(by the Cauchy–Schwarz inequality)} \\
&\le \sqrt{SA\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}2\mathbb{I}_t(s,a)(L_{t,h(s)}-M_{t,h(s)})^2\right]} \\
&\lesssim \sqrt{H^2SA\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}\mathbb{I}_t(s,a)(\ell_t(s,a)-m_t(s,a))^2\right]}, \qquad (98)
\end{aligned}$$

where the third line uses $\sum_a(\mathbb{I}_t(s,a)-\pi_t(a\mid s)\mathbb{I}_t(s))^2\le 2\mathbb{I}_t(s)$ for each fixed state s, and the last inequality follows from Lemma D.10.

Applying Lemma D.11 to (98), we obtain, for any policy π,

$$\begin{aligned}
\mathbf{term}_1 \lesssim{}& \sqrt{H^2SA\log^2(T)\min\{L(\pi)+\mathrm{Reg}_T(\pi), HT-L(\pi)-\mathrm{Reg}_T(\pi), Q_\infty, V_1\}} \\
&+HSA\log(T). \qquad (99)
\end{aligned}$$

For $\mathbf{term}_2$, Lemma D.11 similarly yields, for any policy π,

$$\begin{aligned}
\mathbf{term}_2 \lesssim{}& \sqrt{H^3S^2A\log^2(T)\min\{L(\pi)+\mathrm{Reg}_T(\pi), HT-L(\pi)-\mathrm{Reg}_T(\pi), Q_\infty, V_1\}} \\
&+\sqrt{H^3S^3}A\log(T). \qquad (100)
\end{aligned}$$

For $\mathbf{term}_3$, Lemma D.12 applied to the real episodes yields

$$\mathbf{term}_3 \lesssim \sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]} + \sqrt{HS^2A\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}V^{\pi_t}(s_0;(\ell_t-m_t)^2)\right]}.$$

Applying Lemma D.11 to the second term, we obtain, for any policy π,

$$\begin{aligned}
\mathbf{term}_3 \lesssim{}& \sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]} \\
&+\sqrt{HS^2A\log^2(T)\min\{L(\pi)+\mathrm{Reg}_T(\pi), HT-L(\pi)-\mathrm{Reg}_T(\pi), Q_\infty, V_1\}} \\
&+\sqrt{HS^3}A\log(T). \qquad (101)
\end{aligned}$$

Combining (99)–(101) with (97), we obtain

$$\begin{aligned}
\mathrm{Reg}_T(\pi) \lesssim{}& \sqrt{H^3S^2A\log^2(T)\min\{L(\pi)+\mathrm{Reg}_T(\pi), HT-L(\pi)-\mathrm{Reg}_T(\pi), Q_\infty, V_1\}} \\
&+\sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]}+\left(H^{\frac{3}{2}}S^3A+H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T). \qquad (102)
\end{aligned}$$

Below, we derive the first-order bound. By (102), there exists an absolute constant $c>0$ such that

$$\begin{aligned}
\mathrm{Reg}_T(\pi) &\le c\sqrt{H^3S^2A\log^2(T)L(\pi)}+c\sqrt{H^3S^2A\log^2(T)\mathrm{Reg}_T(\pi)} \\
&\quad+c\sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]}+c\left(H^{\frac{3}{2}}S^3A+H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T) \\
&\le c\sqrt{H^3S^2A\log^2(T)L(\pi)}+\frac{c^2}{2}H^3S^2A\log^2(T)+\frac{1}{2}\mathrm{Reg}_T(\pi) \\
&\quad+c\sqrt{S^2A\log^2(T)\mathbb{E}[Q^{\pi_{1:T}}_{\mathrm{trans}}(\ell)]}+c\left(H^{\frac{3}{2}}S^3A+H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T),
\end{aligned}$$

where the last inequality follows from the AM–GM inequality. Thus, we obtain

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\lesssim \sqrt{H^3S^2A\log^2(T)L(\pi)} + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} \\ &\quad + \left(H^3S^2A + H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T). \end{aligned} \tag{103}$$

Furthermore, we derive the bound in terms of $HT - L(\pi)$. (102) gives

$$\begin{aligned}\mathrm{Reg}_T(\pi) &\lesssim \sqrt{H^3S^2A\log^2(T)(HT - L(\pi) - \mathrm{Reg}_T(\pi))} + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} \\ &\quad + \left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T) \\ &\le \sqrt{H^3S^2A\log^2(T)(HT - L(\pi))} + \sqrt{S^2A\log^2(T)\mathbb{E}[Q_{\mathrm{trans}}^{\pi_{1:T}}(\ell)]} \\ &\quad + \left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T). \end{aligned} \tag{104}$$

Combining (102)–(104) completes the proof. □

G.4 Proof for the stochastic regime with adversarial corruption

Theorem G.13 (bandit setting)**.** *Under the stochastic regime with adversarial corruption, Algorithm 1 with $\delta = \frac{1}{T^2}$ guarantees that, for any policy π,*

$$\mathrm{Reg}_T(\pi) \lesssim U + \sqrt{U\mathcal{C}} + \left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T),$$

where $U = \frac{H^4S^2A\log^2(T)}{\Delta_{\min}}$.

Proof. Since $\mathrm{Reg}_T(\pi) \le \mathrm{Reg}_T(\mathring{\pi})$ for all π, it suffices to bound $\mathrm{Reg}_T(\mathring{\pi})$. By (96) and Lemma G.8, (91) and $\delta = \mathcal{O}(1/T^2)$, we have $\iota \lesssim \log(T)$. Hence, there exists an absolute constant $c > 0$ such that

$$\begin{aligned}\mathrm{Reg}_T(\mathring{\pi}) \le c\left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T) &+ \underbrace{c\log(T)\sum_{s,a}\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\zeta_t(s,a)\right]}}_{\textbf{term}_1} \\ &+ \underbrace{c\sqrt{H^3S^2A\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{s,a}[q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)]_+\right]}}_{\textbf{term}_2} \\ &+ \underbrace{c\sum_{h=0}^{H-1}\sqrt{H^2|\mathcal{S}_h||\mathcal{S}_{h+1}|A\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}[q^{\pi_t}(s,a) - q^{\mathring{\pi}}(s,a)]_+\right]}}_{\textbf{term}_3}. \end{aligned} \tag{105}$$

For $\textbf{term}_1$, by the definition of $\zeta_t(s,a)$, we have for any $\alpha > 0$,

$$\begin{aligned}\textbf{term}_1 &= c\sum_{s,a}\sqrt{\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}(\mathbb{I}_t(s,a) - \pi_t(a\mid s)\mathbb{I}_t(s))^2(L_{t,h(s)} - M_{t,h(s)})^2\right]} \\ &\le cH\sum_{s,a}\sqrt{\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}q^{\pi_t}(s,a)(1 - \pi_t(a\mid s))\right]} \\ &\le 2cH\sum_{s}\sum_{a\neq\pi^\star(s)}\sqrt{\log^2(T)\mathbb{E}\left[\sum_{t\in\mathcal{T}_r}q^{\pi_t}(s,a)\right]} \\ &\le \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 2\mathcal{C}) + \sum_{s}\sum_{a\neq\pi^\star(s)}\frac{c^2H^2\log^2(T)}{\alpha\Delta(s,a)}, \end{aligned} \tag{106}$$

where the last line follows from Lemma D.16.

For $\mathbf{term}_2$, Lemma D.18 gives

$$\mathbf{term}_2 \le \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}) + \frac{c^2 H^4 S^2 A \log^2(T)}{4\alpha\Delta_{\min}}. \tag{107}$$

For $\mathbf{term}_3$, Lemma D.17 gives

$$\mathbf{term}_3 \le \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}) + \frac{c^2\left(\sum_{h=0}^{H-1}\sqrt{H^2|\mathcal{S}_h||\mathcal{S}_{h+1}|A\log^2(T)}\right)^2}{4\alpha\Delta_{\min}}.$$

Moreover, by the AM–GM inequality,

$$\begin{aligned}\left(\sum_{h=0}^{H-1}\sqrt{H^2|\mathcal{S}_h||\mathcal{S}_{h+1}|A\log^2(T)}\right)^2 &\le \left(\sum_{h=0}^{H-1}\frac{1}{2}(|\mathcal{S}_h|+|\mathcal{S}_{h+1}|)\sqrt{H^2A\log^2(T)}\right)^2 \\ &\le \left(\sqrt{H^2S^2A\log^2(T)}\right)^2 = H^2S^2A\log^2(T).\end{aligned}$$

Thus, we obtain

$$\mathbf{term}_3 \le \alpha(\mathrm{Reg}_T(\mathring{\pi}) + 4\mathcal{C}) + \frac{c^2 H^2 S^2 A \log^2(T)}{4\alpha\Delta_{\min}}. \tag{108}$$

Combining (106)–(108) with (105), we obtain

$$\mathrm{Reg}_T(\mathring{\pi}) \le \alpha(3\mathrm{Reg}_T(\mathring{\pi}) + 10\mathcal{C}) + \mathcal{O}\left(\frac{H^4S^2A\log^2(T)}{4\alpha\Delta_{\min}} + \left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T)\right).$$

Finally, choosing $\alpha = \min\left\{\frac{1}{6}, \sqrt{\frac{U}{\mathcal{C}}}\right\}$ with $U = \frac{H^4S^2A\log^2(T)}{\Delta_{\min}}$, we get

$$\mathrm{Reg}_T(\mathring{\pi}) \lesssim U + \sqrt{U\mathcal{C}} + \left(H^{\frac{3}{2}}S^3A + H^{\frac{3}{2}}S^{\frac{5}{2}}A^{\frac{3}{2}}\right)\log^2(T).$$

Therefore, since $\mathrm{Reg}_T(\pi) \le \mathrm{Reg}_T(\mathring{\pi})$ for all π by the definition of $\mathring{\pi}$, the desired bound follows. □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the gap in prior work addressed by this paper and summarize the main regret guarantees.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations and future directions are discussed in Sections 4 and 6, including lower-bound gaps, variance-aware bounds, and aggregate-feedback extensions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The assumptions and theoretical results are stated in Sections 2 and 4, with a proof sketch in Section 5. Complete proofs are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

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

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

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

Answer: [N/A]

Justification: This paper does not include experiments.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [N/A]

Justification: This paper does not include experiments.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [N/A]

Justification: This paper does not include experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

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

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The research is theoretical and uses no human subjects or sensitive data.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This is a theoretical paper and does not have direct societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: This paper does not release new research assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method and theoretical results do not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.