Online Resource Allocation in Episodic Markov Decision Processes

Duksang Lee [1] Dabeen Lee [1 *]

May 18, 2023

Abstract

This paper studies a long-term resource allocation problem over multiple periods where each period requires a multi-stage decision-making process. We formulate the problem as an online resource allocation problem in an episodic finite-horizon Markov decision process with unknown non-stationary transitions and stochastic non-stationary reward and resource consumption functions for each episode. We provide an equivalent online linear programming reformulation based on occupancy measures, for which we develop an online mirror descent algorithm. Our online dual mirror descent algorithm for resource allocation deals with uncertainties and errors in estimating the true feasible set, which is of independent interest. We prove that under stochastic reward and resource consumption functions, the expected regret of the online mirror descent algorithm is bounded by $O(\rho^{-1}H^{3/2}S\sqrt{AT})$ where $\rho \in (0,1)$ is the budget parameter, H is the length of the horizon, S and A are the numbers of states and actions, and T is the number of episodes.

1 Introduction

We consider a long-term online resource allocation problem where requests for service arrive sequentially over episodes and the decision-maker chooses an action that generates a reward and consumes a certain amount of resources for each request. Such resource allocation problems arise in revenue management and online advertising. Hotels and airlines receive requests for a room or a flight, and they decide how to process the requests in real-time based on their availability of remaining rooms and flight seats [48]. For search engines, when a user arrives with a keyword, they collect bids from relevant advertisers and decide which ad to show to the user [38]. The decision-maker is informed of the reward function and the resource consumption function of each arriving request with respect to the associated action, while the decision-maker makes actions in an online fashion without the knowledge of future requests.

The online resource allocation problem has been studied in the context of or under the name of the AdWords

[1]Department of Industrial and Systems Engineering, KAIST, Daejeon 34141, Republic of Korea

[*]Correspondence to <dabeenl@kaist.ac.kr>

problem [10, 19], repeated auctions with budgets [6], online stochastic matching [23, 33], online linear programming [3, 24, 27], assortment optimization with limited inventories [25], online convex programming [2], online binary programming [35], online stochastic optimization [30], online resource allocation with concave rewards [5] and nonlinear rewards [7]. For a more comprehensive literature review on online resource allocation, we refer to [7] and references therein.

Although the above literature assumes that the decision-maker makes a *single* action for a request, many of the modern service systems allow *multi-stage* decision-making processes and interactions with customers based on user feedback. For example, medical processes [47] involve sequential decision-making while the prices and costs of medical operations are often predetermined. Multi-stage second-price repeated auctions [26] consider a bidder who is willing to participate in auctions multiple times until winning an item. For these applications, actions taken in multiple stages are not necessarily independent, and therefore, it is natural to group a multi-stage decision-making process as an episode. Then this brings about online resource allocation problems where an episode itself involves a sequential decision-making process. Therefore, to model these scenarios, we need a framework to capture multi-stage actions for requests and interactions between service systems and customers.

Motivated by this, we extend the existing framework to consider online resource allocation over an episodic Markov decision process (MDP), generalizing a single action for an episode to multi-stage actions. More precisely, at the beginning of each episode, the reward and the resource consumption of any action at a state throughout the episode are revealed to the decision-maker. Transitions between states are governed by an unknown transition kernel. After an episode, the decision-maker observes the cumulative reward and resource consumption accrued over the episode. There is a budget for the total resource consumption for the entire process, so the decision-maker can keep track of the remaining budget but cannot observe the reward and resource consumption functions of future episodes. Therefore, the problem is to prepare a policy based on the given reward and resource consumption functions of the past and current episodes, the remaining resource budget, and the estimation of the unknown transition kernel. The main challenge here is to deal with uncertainty in not only the reward and resource consumption functions of future episodes but also the unknown transition function of the underlying MDP.

Our Contributions This paper initiates the study of online resource allocation problems where each episode itself involves a multi-stage decision-making process. As a first step, we consider the formulation whose episodes are given by a finite-horizon Markov decision process.

- We show an online linear programming reformulation of the online resource allocation problem in an episodic finite-horizon Markov decision process. Based on this reduction, we develop an online dual mirror descent algorithm. Unlike the existing online resource allocation frameworks, we have to deal with uncertainties and errors in estimating the true feasible set.
- We prove that if the reward and resource consumption functions are i.i.d. over episodes, then the online dual mirror descent algorithm guarantees that the expected regret is bounded above by $O\left(\rho^{-1}H^{3/2}S\sqrt{AT}\,(\ln HSAT)^2\right)$ where $\rho \in (0,1)$ is the budget parameter, H is the length of the horizon for each episode, S is the number of states, A is the number of actions, and T is the number of episodes. The resource consumption constraint is

satisfied without any violation.

Our work is closely related to episodic MDPs with adversarial rewards in that we use an occupancy-measure-based formulation and develop a gradient-based policy optimization method. Three main directions are the loop-free setting [17, 29, 31, 32, 39, 40, 44, 45, 56], the stochastic shortest path problem [8, 12, 13, 46, 49], and the ergodic infinite-horizon case [20, 22, 41, 43, 53] (see also [1, 18, 34, 52, 54]). Another closely related setting is finite-horizon constrained MDPs [11, 14, 21, 36, 42, 50, 51].

Our framework extends the online dual mirror descent algorithm of [7] to the finite-horizon episodic MDP setting. We adapt and apply analytical tools developed for episodic MDPs with adversarial rewards under unknown transitions due to [12, 13, 31].

2 Problem Setting

Finite-Horizon Episodic MDP We model the online resource allocation problem with a *finite-horizon episodic MDP*. A finite-horizon MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^{H-1}, p)$ where $\mathcal{S}$ is the finite state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the finite action space with $|\mathcal{A}| = A$, H is the finite horizon, $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel at step $h \in [H]$, and p is the initial distribution of the states. Here, $P_h(s' \mid s, a)$ is the probability of transitioning to state s' from state s when the chosen action is a at step $h \in [H-1]$. Equivalently, we may define a single *non-stationary* transition kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0, 1]$ with $P(s' \mid s, a, h) = P_h(s' \mid s, a)$ and $P(s' \mid s, a, H) = p(s')$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H-1]$. We assume that the transition kernels $\{P_h\}_{h=1}^{H-1}$ and thus P are *unknown*. Each episode starts at a state in $\mathcal{S}$ with a stationary distribution $p : \mathcal{S} \to [0, 1]$ and terminates after making H actions.

Before an episode begins, the decision-maker prepares a *stochastic policy* $\pi : \mathcal{S} \times [H] \times \mathcal{A} \to [0, 1]$ where $\pi(a \mid s, h)$ is the probability of selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at step h. Here, π can be viewed as a *non-stationary policy* as it may change over the horizon, and this is due to the non-stationarity of the transition kernels over steps $h \in [H]$.

The reward and resource consumption functions of episode $t \in [T]$ is given by $f_t, g_t : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$, i.e., choosing action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ at step h generates a reward $f_t(s, a, h)$ and consumes resources of amount $g_t(s, a, h)$. Here, functions f_t and g_t are non-stationary over $h \in [H]$. Given a policy π_t for episode $t \in [T]$, the MDP proceeds with trajectory $\{s_h^{P,\pi_t}, a_h^{P,\pi_t}\}_{h \in [H]}$ generated by the stationary distribution p over the states in $\mathcal{S}$ and the transition kernels $\{P_h\}_{h \in [H]}$.

Comparisons to Episodic Loop-Free MDPs and Stochastic Shortest Paths The finite-horizon setting we consider allows non-stationary transitions and non-stationary reward and resource consumption functions in an episode. Hence, the setting sits between the *loop-free* MDP setting and the *stochastic shortest path* problem.

The finite-horizon MDP naturally translates to a loop-free MDP with $O(H)$ layers and $O(SH)$ states. That said, an alternate approach is to reduce the finite-horizon MDP to a loop-free MDP and adapt the framework of [44]. However, this would lead to a suboptimal dependence on the horizon parameter H, as the number of states in the reduced loop-free MDP is $O(SH)$.

Another related setting is the episodic stochastic shortest path problem. A successful technique due to [12, 13] is to consider a reduction of the problem to a finite-horizon episodic MDP. However, the finite-horizon reduction has a stationary transition kernel, i.e., P_h there is invariant over $h \in [H]$.

Hence, the main challenge for the finite-horizon MDP setting is to tighten the dependency on parameters S and H in the performance bound given the existence of non-stationary environments.

Online Resource Allocation in an Episodic Finite-Horizon MDP The budget for the total resource consumption over the T episodes is given by $TH\rho$ for some $\rho \in (0,1)$. Hence, the problem is to find policies $\pi_1, \ldots, \pi_T$ for the T episodes to maximize the total cumulative reward

$$\text{Reward}\left(\vec{\gamma}, \vec{\pi}\right) := \sum_{t=1}^{T}\sum_{h=1}^{H} f_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right)$$

while satisfying the resource consumption budget constraint

$$\text{Resource}\left(\vec{\gamma}, \vec{\pi}\right) := \sum_{t=1}^{T}\sum_{h=1}^{H} g_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right) \leq TH\rho.$$

Here, we use short-hand notations $\vec{\gamma} = (\gamma_1, \ldots, \gamma_T)$ where $\gamma_t = (f_t, g_t)$ and $\vec{\pi} = (\pi_1, \ldots, \pi_T)$.

The decision-maker selects policies $\pi_1, \ldots, \pi_T$ in an *online* fashion because the decision-maker is oblivious to the reward and resource consumption functions of *future* episodes as well as the true transition kernel P. The performance of the decision-maker can be compared to the best possible performance achievable when $\vec{\gamma}$ and P are all available in advance, which is given by

$$\begin{aligned} \text{OPT}(\vec{\gamma}) := \max_{\pi_1,\ldots,\pi_T} \quad & \mathbb{E}\left[\text{Reward}\left(\vec{\gamma}, \vec{\pi}\right) \mid \vec{\gamma}, \vec{\pi}, P\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\text{Resource}\left(\vec{\gamma}, \vec{\pi}\right) \mid \vec{\gamma}, \vec{\pi}, P\right] \leq TH\rho. \end{aligned} \tag{1}$$

Here, the expectations are taken with respect to the randomness of the trajectories of episodes. Hence, the goal is to design an algorithm to learn the optimal resource allocation strategy to maximize the total cumulative reward. Upon observing f_t and g_t before episode t starts, the decision-maker prepares a policy π_t based on the history $(\gamma_1, \ldots, \gamma_t)$, the remaining budget on resource consumption, and the estimated transition kernels. To measure the performance of a learning algorithm that produces policies $\pi_1, \ldots, \pi_T$, we consider

$$\text{Regret}\left(\vec{\gamma}, \vec{\pi}\right) := \text{OPT}(\vec{\gamma}) - \text{Reward}\left(\vec{\gamma}, \vec{\pi}\right).$$

We do not allow violating the resource consumption constraint. Basically, an algorithm needs to stop if the remaining budget is not enough for taking an action that generates a reward. To model this, we assume the following.

Assumption 1. *There exists an action $a^\star \in \mathcal{A}$ such that $f_t(s, a^*, h) = g_t(s, a^*, h) = 0$ for any $(s, h, t) \in \mathcal{S} \times [H] \times [T]$.*

Hence, an algorithm would take action a^* for the remaining steps if the budget for resource consumption is tight.

3 Reformulation

Our framework for the long-term online resource allocation problem over an episodic finite-horizon MDP is based on reducing the policy optimization problem given in (1) to *online linear programming*, and more generally, to the *online resource allocation* problem. First, we adapt the idea of *occupancy measures* [4, 16, 44, 55]. Given a policy π and a transition kernel P, let $\bar{q}^{P,\pi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ be defined as

$$\bar{q}^{P,\pi}(s,a,s',h) = \mathbb{P}\left[s_h^{P,\pi} = s,\ a_h^{P,\pi} = a,\ s_{h+1}^{P,\pi} = s' \mid \pi, P\right] \tag{2}$$

for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Note that any $\bar{q}$ defined as in (2) has the following properties.

$$\sum_{(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}} \bar{q}(s,a,s',h) = 1, \quad h \in [H] \tag{C1}$$

$$\sum_{(s',a) \in \mathcal{S}\times\mathcal{A}} \bar{q}(s,a,s',h) = \sum_{(s',a) \in \mathcal{S}\times\mathcal{A}} \bar{q}(s',a,s,h-1), \quad s \in \mathcal{S},\ h = 2,\ldots,H. \tag{C2}$$

The *occupancy measure* $q^{P,\pi} : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ associated with policy π and transition kernel P is defined as

$$q^{P,\pi}(s,a,h) = \sum_{s' \in \mathcal{S}} \bar{q}^{P,\pi}(s,a,s',h). \tag{C3}$$

Then it follows that

$$q^{P,\pi}(s,a,h) = \mathbb{P}\left[s_h^{P,\pi} = s,\ a_h^{P,\pi} = a \mid \pi, P\right].$$

Hence, if a policy π is chosen, then the occupancy measure for a loop-free MDP with transition kernel P is determined. Conversely, any $q \in \mathcal{S}\times\mathcal{A}\times[H] \to [0,1]$ with $\bar{q} : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H] \to [0,1]$ satisfying (C1), (C2), (C3) induces a transition kernel P^q and a policy π^q given as follows:

$$P^q(s' \mid s,a,h) = \frac{\bar{q}(s,a,s',h)}{\sum_{s'' \in \mathcal{S}} \bar{q}(s,a,s'',h)}, \quad \pi^q(a \mid s,h) = \frac{q(s,a,h)}{\sum_{b \in \mathcal{A}} q(s,b,h)}. \tag{3}$$

Lemma 3.1. *Let $q : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$. Then q is a valid occupancy measure that induces transition kernel P if and only if there exists $\bar{q} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ that satisfies (C1), (C2), (C3), and $P^q = P$.*

Therefore, there is a one-to-one correspondence between the set of policies and the set of occupancy measures that give rise to transition kernel P. Moreover, the cumulative reward for episode t under reward function f_t, policy π_t, and transition kernel P can be written in terms of occupancy measure q^{P,π_t} associated with π_t and P.

$$\mathbb{E}\left[\sum_{h=1}^{H} f_t\left(s_h^{\pi_t}(t), a_h^{\pi_t}(t), h\right) \mid \vec{\gamma}, \vec{\pi}, P\right] = \sum_{(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]} q^{P,\pi_t}(s,a,h)\, f_t(s,a,h).$$

We may express occupancy measure q^{P,π_t} as an $(S \times A \times H)$-dimensional vector $\boldsymbol{q^{P,\pi_t}}$ whose entries are given by $q^{P,\pi_t}(s,a,h)$ for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Similarly, we define $\boldsymbol{\bar{q}^{P,\pi_t}}$ as the vector whose entries are $\bar{q}^{P,\pi_t}(s,a,s',h)$ for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Moreover, we define vector $\boldsymbol{f_t}$ whose entry corresponding to $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ is given by $f_t(s,a,h)$. Then the right-hand side of the above equation is equal to $\langle \boldsymbol{f_t}, \boldsymbol{q^{P,\pi_t}} \rangle$, the inner product of $\boldsymbol{f_t}$ and $\boldsymbol{q^{P,\pi_t}}$. Likewise, we define vector $\boldsymbol{g_t}$ to represent the resource consumption function g_t. Consequently,

$$\mathbb{E}\left[\text{Reward}(\vec{\gamma}, \vec{\pi}) \mid \vec{\gamma}, \vec{\pi}, P\right] = \sum_{t=1}^{T} \langle \boldsymbol{f_t}, \boldsymbol{q^{P,\pi_t}} \rangle, \quad \mathbb{E}\left[\text{Resource}(\vec{\gamma}, \vec{\pi}) \mid \vec{\gamma}, \vec{\pi}, P\right] = \sum_{t=1}^{T} \langle \boldsymbol{g_t}, \boldsymbol{q^{P,\pi_t}} \rangle.$$

Then the policy optimization problem (1) can be reformulated as

$$\mathrm{OPT}(\vec{\gamma}) = \max_{\boldsymbol{q_1},\ldots,\boldsymbol{q_T}\in\Delta(P)} \quad \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \boldsymbol{q_t}\rangle \quad \text{s.t.} \quad \sum_{t=1}^{T}\langle \boldsymbol{g_t}, \boldsymbol{q_t}\rangle \leq TH\rho \tag{4}$$

where $\Delta(P)$ is the set of all valid occupancy measures inducing transition kernel P. More precisely, $\Delta(P)$ is defined as

$$\Delta(P) = \left\{\boldsymbol{q}\in[0,1]^{S\times A\times H} : \exists\bar{\boldsymbol{q}}\in[0,1]^{S\times A\times S\times H} \text{ satisfying (C1), (C2), (C3), } P^{q}=P\right\}$$

where $P^{\bar{q}}$ is defined as in (3). Hence, $\Delta(P)$ is a polytope, and therefore, (4) corresponds to an online linear programming instance.

In contrast to the standard online resource allocation problem (see [5, 7]), the feasible set $\Delta(P)$ is unknown as we do not have access to the true transition kernel P. To remedy this issue, we obtain *relaxations* $\Delta(P,t)$ for $t\in[T]$ over time by building *confidence sets* for the true transition kernel P. We cannot apply the existing methods for online resource allocation directly because it is difficult to make the relaxations $\Delta(P,t)$ i.i.d. due to the underlying dependency of the estimation process. Our framework is to modify and extend the framework of [7] to deal with the uncertainty in $\Delta(P)$.

4 Our Algorithm

In this section, we present Algorithm 1 for online resource allocation in episodic finite-horizon MDPs. As explained in Section 3, the online resource allocation problem can be reformulated as an instance of online linear programming where each decision is encoded by an occupancy measure that corresponds to a policy for an episode. Then we adapt the *online dual mirror descent* algorithm by [7] that was originally developed for nonlinear reward and resource consumption functions. However, as mentioned earlier, the issue with directly applying the online dual mirror descent algorithm to the formulation (4) is that the feasible set $\Delta(P)$ is not given to us because the true transition kernel P is unknown. To remedy this issue, we obtain empirical transition kernels $\bar{P}_1,\ldots,\bar{P}_T$ to estimate the true transition kernel P, based on which we construct relaxations $\Delta(P,1),\ldots,\Delta(P,T)$ of the feasible set $\Delta(P)$. This gives rise to a relaxation of (4), to which we may apply the online dual mirror descent algorithm. However, the relaxations $\Delta(P,1),\ldots,\Delta(P,T)$ are not i.i.d., which is assumed for the analysis given by [7]. Instead, we will show and use the property that $\Delta(P,1),\ldots,\Delta(P,T)$ contain $\Delta(P)$ with high probability, which turns out to be sufficient to provide the desired performance guarantee on Algorithm 1. We explain in greater detail the part of estimating the true transition kernel in Section 4.1 and the part of applying the online dual mirror descent algorithm in Section 4.2.

4.1 Confidence Sets

To construct confidence sets for estimating the non-stationary transition kernel P, we extend the framework of [31] developed for the loop-free setting to our finite-horizon setting. Following [12], we update the confidence set for each episode $t\in[T]$, in contrast to [31] where the confidence set is updated over *epochs* and an epoch may consist of multiple episodes.

Algorithm 1 Online dual mirror descent for resource allocation in episodic finite-horizon MDPs

Initialize: dual variable λ_1, initial budget $B = TH\rho$, episode counter $t = 1$, counters $N(s, a, h) = 0$ and $M(s, a, s', h) = 0$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, and step size $\eta > 0$.

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

(Confidence set construction)

Based on the counters $N_t \leftarrow N$ and $M_t \leftarrow M$, compute the empirical transition kernel $\bar{P}_t$, the confidence radius ϵ_t, and the confidence set $\mathcal{P}_t$ defined as in (5), (6), and (7), respectively.

(Policy update)

Observe reward function f_t and resource consumption function g_t.

Deduce policy $\pi_t = \pi^{\widehat{q}_t}$ defined as in (3) where $\widehat{\boldsymbol{q}}_{\boldsymbol{t}} \in \operatorname{argmax}_{\boldsymbol{q} \in \Delta(P,t)} \{\langle \boldsymbol{f_t}, \boldsymbol{q}\rangle - \lambda_t \langle \boldsymbol{g_t}, \boldsymbol{q}\rangle\}$ and $\Delta(P, t)$ is defined as in (8).

(Policy execution)

Sample state s_1 from distribution $p(\cdot)$.

for $h = 1, \ldots, H$ **do**

Sample action a_h from policy $\pi_t(\cdot \mid s_h, h)$ and accrue reward $f_t(s_h, a_h, h)$.

Update the remaining budget $B \leftarrow B - g_t(s_h, a_h, h)$.

if $B < 1$ **then**

Return

end if

Observe the next state s_{h+1} determined by distribution $P(\cdot \mid s_h, a_h, h)$.

Update counters $N(s, a, h) \leftarrow N(s, a, h) + 1$ and $M(s, a, s', h) \leftarrow M(s, a, s', h) + 1$.

end for

(Dual update)

Update $\lambda_{t+1} = \operatorname{argmin}_{\lambda \in \mathbb{R}_+} \left\{ \eta \left(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} \rangle \right)^\top \lambda + D(\lambda, \lambda_t) \right\}$ where $D(\cdot, \cdot)$ is the Bregman divergence associated with a reference function.

end for

To estimate the transition kernel, we maintain counters to keep track of the number of visits to each tuple (s, a, h) and tuple (s, a, s', h). For each $t \in [T]$, we define $N_t(s, a, h)$ and $M_t(s, a, s', h)$ as the number of visits to tuple (s, a, h) and the number of visits to tuple (s, a, s', h) up to the first $t - 1$ episodes, respectively, for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Given $N_t(s, a, h)$ and $M_t(s, a, s', h)$, we define the empirical transition kernel $\bar{P}_t$ for episode t as

$$\bar{P}_t(s' \mid s, a, h) = \frac{M_t(s, a, s', h)}{\max\{1, N_t(s, a, h)\}}. \tag{5}$$

Next, for some confidence parameter $\delta \in (0, 1)$, we define the confidence radius $\epsilon_t(s' \mid s, a, h)$ for $(s, a, s', h) \in$

$\mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ and $t \in [T]$ as

$$\epsilon_t(s' \mid s,a,h) = 2\sqrt{\frac{\bar{P}_t(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)-1\}}} + \frac{14\ln(HSAT/\delta)}{3\max\{1, N_t(s,a,h)-1\}}. \tag{6}$$

Based on the empirical transition kernel and the radius, we define the confidence set for episode t as

$$\mathcal{P}_t = \left\{\widehat{P} : \left|\widehat{P}(s' \mid s,a,h) - \bar{P}_t(s' \mid s,a,h)\right| \leq \epsilon_t(s' \mid s,a,h) \ \ \forall (s,a,s',h)\right\}. \tag{7}$$

Then, by the empirical Bernstein inequality due to [37], we show the following.

Lemma 4.1. *With probability at least $1-4\delta$, the true transition kernel P is contained in the confidence set $\mathcal{P}_t$ for every episode $t \in [T]$.*

For episode $t \in [T]$, we define $\Delta(P,t)$ as

$$\Delta(P,t) = \left\{\boldsymbol{q} \in [0,1]^{S\times A\times H} : \ \exists \bar{\boldsymbol{q}} \in [0,1]^{S\times A\times S\times H} \text{ satisfying (C1), (C2), (C3), } P^q \in \mathcal{P}_t\right\}. \tag{8}$$

As a direct consequence of Lemma 4.1, we deduce the following result.

Lemma 4.2. *With probability at least $1-4\delta$, $\Delta(P) \subseteq \Delta(P,t)$ for every episode $t \in [T]$.*

We remark that our procedure of taking the empirical estimation of the true transition kernel is different from that of [12] because the transition kernel P in our setting is allowed to be non-stationary over steps $h \in [H]$.

4.2 Online Dual Mirror Descent

Based on Lemma 4.2, we may consider a relaxation of (4) where $\Delta(P)$ is replaced by $\Delta(P,t)$ for $t \in [T]$ as follows.

$$\max_{\boldsymbol{q_1}\Delta(P,1),\ldots,\boldsymbol{q_T}\in\Delta(P,T)} \quad \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \boldsymbol{q_t}\rangle \quad \text{s.t.} \quad \sum_{t=1}^{T}\langle \boldsymbol{g_t}, \boldsymbol{q_t}\rangle \leq TH\rho. \tag{9}$$

Lemma 4.2 implies that Equation (9) is a relaxation of Equation (4) with probability at least $1-4\delta$. However, we cannot directly apply the analysis of the online dual descent algorithm by [7] because the empirical distributions $\bar{P}_1, \ldots, \bar{P}_T$, and thus $\Delta(P,1), \ldots, \Delta(P,T)$, are dependent and not identically distributed. Nevertheless, we will show that the online dual mirror descent algorithm equipped with the relaxations $\Delta(P,1), \ldots, \Delta(P,T)$, given by Algorithm 1, still guarantees the desired regret bound.

Algorithm 1 proceeds with four parts in each episode. At the beginning of each episode $t \in [T]$, it first obtains the feasible set $\Delta(P,t)$ by constructing the confidence set $\mathcal{P}_t$. Second, the algorithm prepares a policy π_t based on the current dual solution λ_t, reward function f_t, resource consumption function g_t, and the set $\Delta(P,t)$. Third, the algorithm runs the episode with policy π_t. Lastly, the algorithm prepares the dual solution λ_{t+1} for the next episode based on the outcomes of episode t.

To be more specific, the policy update part works as follows. Given the dual solution λ_t prepared before episode t starts, we take

$$\widehat{\boldsymbol{q}}_{\boldsymbol{t}} \in \operatorname*{argmax}_{\boldsymbol{q}\in\Delta(P,t)} \left\{\langle \boldsymbol{f_t}, \boldsymbol{q}\rangle - \lambda_t \langle \boldsymbol{g_t}, \boldsymbol{q}\rangle\right\}.$$

Here $f_t - \lambda_t g_t$ is the reward function f_t penalized by the resource consumption function g_t. Then based on (3), we deduce policy $\pi^{\widehat{q}_t}$ associated with the occupancy measure $\widehat{q}_t$ whose vector representation is $\widehat{\boldsymbol{q}}_{\boldsymbol{t}}$ as in (3). For ease of notation, we denote $\pi_t = \pi^{\widehat{q}_t}$.

Next, the algorithm executes policy π_t for episode t. The algorithm stops if the remaining amount of resources becomes less than 1. Remember that $g_t(s, a, h) \in [0, 1]$ for any $(s, a, h, t) \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$. Hence, we would not violate the resource consumption constraint if we run the process only when the remaining resource budget is greater than or equal to 1.

At the end of each episode, the algorithm updates the dual variable for the resource consumption constraint. The dual update rule

$$\lambda_{t+1} = \underset{\lambda \in \mathbb{R}_+}{\operatorname{argmin}} \left\{ \eta \left(H\rho - \langle \boldsymbol{g}_{\boldsymbol{t}}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} \rangle \right)^\top \lambda + D(\lambda, \lambda_t) \right\}$$

follows the online dual mirror descent algorithm of [7]. Here, $D(\lambda, \lambda_t)$ is given by

$$D(\lambda, \lambda_t) = \psi(\lambda) - \psi(\lambda_t) - \nabla\psi(\lambda)^\top (\lambda - \lambda_t)$$

where ψ is a reference function.

Assumption 2. *For some fixed constant C, $D(\lambda, \lambda') \leq C(\lambda - \lambda')^2$ for any $\lambda, \lambda' \in \mathbb{R}_+$.*

The assumption is satisfied for $\|\lambda\|_2^2/2$ and the negative entropy function. Note that we have a single resource consumption constraint, in which case we have a single dual variable λ and $H\rho - \langle \boldsymbol{g}_{\boldsymbol{t}}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} \rangle, \nabla\psi(\lambda)$ are scalars. In fact, our framework easily extends to multiple resource constraints, for which we use a vector of dual variables $\boldsymbol{\lambda} \in \mathbb{R}_+^m$ where m is the number of resource constraints.

5 Regret Analysis

Let $T^\star$ be the episode in or after which Algorithm 1 terminates. For $t > T^\star$, we set $\pi_t(a^\star \mid s, h) = 1$ for any $(s, h) \in \mathcal{S} \times [H]$ where action $a^\star$ is given in Assumption 1. Moreover, if Algorithm 1 terminates after step $h^\star \in [H]$ in episode T^*, then we take action $a_h = a^*$ for step $h > h^\star$. For $t \in [T^\star]$, let π_t be the policy deduced by occupancy measure $\widehat{q}_t$, and let q_t be the occupancy measure associated with policy π_t and the true transitional kernel P, i.e., $q_t = q^{P,\pi_t}$. For $t > T^\star$, let q_t and $\widehat{q}_t$ correspond to the policy $\pi(a^\star \mid s, h) = 1$ for any $(s, h) \in \mathcal{S} \times [H]$. Then it follows that

$$\begin{aligned}
&\text{Regret}\,(\vec{\gamma}, \vec{\pi}) = \text{OPT}(\vec{\gamma}) - \text{Reward}\,(\vec{\gamma}, \vec{\pi}) \\
&= \underbrace{\text{OPT}(\vec{\gamma}) - \sum_{t=1}^{T} \langle \boldsymbol{f}_{\boldsymbol{t}}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} \rangle}_{\text{(I)}} + \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{f}_{\boldsymbol{t}}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} - \boldsymbol{q}_{\boldsymbol{t}} \rangle}_{\text{(II)}} + \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{f}_{\boldsymbol{t}}, \boldsymbol{q}_{\boldsymbol{t}} \rangle - \sum_{t=1}^{T} \sum_{h=1}^{H} f_t \left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h \right)}_{\text{(III)}}
\end{aligned}$$

Here, we control the regret term (I) by the online dual mirror descent algorithm. One thing to consider is that $\Delta(P)$ and $\Delta(P, t)$ for $t \in [T]$ are different, but we use Lemma 4.2 that $\Delta(P) \subseteq \Delta(P, t)$ for $t \in [T]$ with probability at least $1 - 4\delta$. The regret term (II) is due to the error in estimating the true transition kernel and constructing the confidence sets. The regret term (III) is the sum of a martingale difference sequence where $\langle \boldsymbol{f_t}, \boldsymbol{q_t} \rangle$ is the expected reward accrued in episode t while $\sum_{h=1}^{H} f_t(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h)$ is the realized reward in episode t.

Theorem 1. *Under stochastic reward and resource consumption functions, Algorithm 1 with step size $\eta = 1/(\rho H\sqrt{T})$ guarantees that*

$$\mathbb{E}\left[\text{Regret}\left(\vec{\gamma}, \vec{\pi}\right) \mid P\right] = O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right) (\ln HSAT)^2\right)$$

where the expectation is taken with respect to the randomness of the reward and resource consumption functions and the randomness in the trajectories of episodes.

Here, Theorem 1 provides a guarantee on the expected regret of Algorithm 1. In fact, the regret term (III) is zero in expectation, as it gives rise to a martingale difference sequence. Nevertheless, we show bounds on the terms (II) and (III) that hold with high probability, which is of independent interest.

5.1 Upper bounds on the regret terms (II) and (III)

Let $n_t(s, a, h)$ be defined as

$$n_t(s,a,h) = \begin{cases} 1, & \text{if the state-action pair } (s,a) \text{ is visited at step } h \text{ of episode } t, \\ 0, & \text{otherwise} \end{cases}$$

for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. By definition,

$$\mathbb{E}\left[n_t(s,a,h) \mid \pi_t, P\right] = q_t(s,a,h). \tag{10}$$

Moreover, we have

$$\sum_{h=1}^{H} f_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right) = \sum_{h=1}^{H} n_t(s,a,h) f_t(s,a,h) = \langle \boldsymbol{n_t}, \boldsymbol{f_t} \rangle$$

where $\boldsymbol{n_t}$ is the vector representation of $n_t : \mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}$. Given two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{S\times A\times H}$, let $\boldsymbol{u} \odot \boldsymbol{v}$ be defined as the vector obtained from coordinate-wise products of $\boldsymbol{u}$ and $\boldsymbol{v}$, i.e. $(\boldsymbol{u} \odot \boldsymbol{v})_i = u_i \odot v_i$ for $i \in [SAH]$. Moreover, we define $\vec{h}$ as an SAH-dimensional vector whose coordinate for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ is h. The following lemma is from [12], and it is useful to bound the variance of $\langle \boldsymbol{n_t}, \boldsymbol{f_t} \rangle$.

Lemma 5.1. [12, Lemma 2] *Let π_t be the policy for episode t, and let q_t denote the occupancy measure q^{P,π_t}. Let $f : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$ be an arbitrary reward function. Then*

$$\mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f} \rangle^2 \mid f, \pi_t, P\right] \leq 2\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f} \rangle$$

where $\boldsymbol{q_t}, \boldsymbol{n_t}, \boldsymbol{f}$ are the vector representations of $q_t, n_t, f : \mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}$.

Next, we provide Lemma 5.2, which is a modification of [12, Lemma 9] to our finite-horizon MDP setting.

Lemma 5.2. *Let π_t be the policy for episode t, and let P_t be any transition kernel from $\mathcal{P}_t$. Let $q_t, \widehat{q}_t$ denote the occupancy measures $q^{P,\pi_t}, q^{P_t,\pi_t}$, respectively. Let $f_t : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$ be an arbitrary reward function for* *episode $t \in [T]$. Then with probability at least $1-4\delta$, we have*

$$\sum_{t=1}^{T} |\langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t} \rangle| = O\left(\left(\sqrt{HS^2A\left(\sum_{t=1}^{T}\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f}\rangle + H^3\sqrt{T}\right)} + H^2S^2A \right) \left(\ln \frac{HSAT}{\delta}\right)^2 \right).$$

Based on Lemmas 5.1 and 5.2, we can bound the regret terms (II) and (III) as follows.

Lemma 5.3. *With probability at least $1-4\delta$, the regret term (II) is bounded by*

$$\sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} - \boldsymbol{q_t}\rangle = O\left(\left(H^{3/2}S\sqrt{AT} + H^{5/2}S^2A \right) \left(\ln \frac{HSAT}{\delta}\right)^2 \right).$$

Lemma 5.4. *With probability at least $1-5\delta$, the regret term (III) is bounded by*

$$\sum_{t=1}^{T}\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle - \sum_{t=1}^{T}\sum_{h=1}^{H} f_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right) = O\left(\left(H\sqrt{T} + H^2S\sqrt{A} \right) \left(\ln \frac{HSAT}{\delta}\right)^2 \right).$$

5.2 Bounding the regret term (I)

We consider

$$\widehat{L}_t(\lambda) = \max_{\boldsymbol{q}\in\Delta(P,t)} \left\{\langle \boldsymbol{f_t}, \boldsymbol{q}\rangle + \lambda(H\rho - \langle \boldsymbol{g_t}, \boldsymbol{q}\rangle)\right\}, \quad L_t(\lambda) = \max_{\boldsymbol{q}\in\Delta(P)} \left\{\langle \boldsymbol{f_t}, \boldsymbol{q}\rangle + \lambda(H\rho - \langle \boldsymbol{g_t}, \boldsymbol{q}\rangle)\right\}.$$

Lemma 5.5. [7, Proposition 1] *For any $\lambda \in \mathbb{R}_+$, we have* $\mathrm{OPT}(\vec{\gamma}) \leq \sum_{t=1}^{T} L_t(\lambda)$.

Moreover, it follows from Lemma 4.2 that with probability at least $1-4\delta$,

$$\widehat{L}_t(\lambda_t) \geq L_t(\lambda_t), \quad \forall t \in [T],$$

which implies that

$$\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \geq L_t(\lambda_t) - \lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle), \quad \forall t \in [T].$$

Based on this, we show the following upper bound on the regret term (III).

Lemma 5.6. *The following holds for the regret term (III).*

$$\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] = O\left(\left(\frac{H^{3/2}}{\rho}S\sqrt{AT} + \frac{H^{5/2}}{\rho}S^2A \right) (\ln HSAT)^2 \right)$$

where the expectation is taken with respect to the randomness of the reward and resource consumption functions and the randomness in the trajectories of episodes.

The proof of Lemma 5.6 follows that of [7, Theorem 1]. The main challenge in our setting is that $\mathrm{OPT}(\vec{\gamma})$ is defined with the feasible set $\Delta(P)$ while $\widehat{\boldsymbol{q}}_{\boldsymbol{t}}$ is chosen from $\Delta(P,t)$ for $t \in [T]$. Due to the error in estimating $\bar{P}_t$ and constructing $\mathcal{P}_t$ for $t \in [T]$, the regret terms (II) and (III) also appear when providing an upper bound on the regret term (I). Then we apply the bounds given by Lemma 5.3 and Lemma 5.4.

Finally, combining Lemmas 5.6, 5.3, and 5.4 providing bounds on the regret terms (I), (II), and (III), respectively, we prove Theorem 1.

Acknowledgements This research is supported, in part, by the KAIST Starting Fund (KAIST-G04220016), the FOUR Brain Korea 21 Program (NRF-5199990113928), the National Research Foundation of Korea (NRF-2022M3J6A1063021).

References

[1] Yasin Abbasi Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvari. Online learning in markov decision processes with adversarially chosen transition probability distributions. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/4f284803bd0966cc24fa8683a34afc6e-Paper.pdf.

[2] Shipra Agrawal and Nikhil R. Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the 2015 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1405–1424, 2014. doi: 10.1137/1.9781611973730.93. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973730.93.

[3] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014. doi: 10.1287/opre.2014.1289. URL https://doi.org/10.1287/opre.2014.1289.

[4] Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

[5] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Dual mirror descent for online allocation problems. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 613–628. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/balseiro20a.html.

[6] Santiago R. Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019. doi: 10.1287/mnsc.2018.3174. URL https://doi.org/10.1287/mnsc.2018.3174.

[7] Santiago R. Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119, 2023. doi: 10.1287/opre.2021.2242. URL https://doi.org/10.1287/opre.2021.2242.

[8] Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991. doi: 10.1287/moor.16.3.580. URL https://doi.org/10.1287/moor.16.3.580.

[9] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15

of *Proceedings of Machine Learning Research*, pages 19–26, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

[10] Niv Buchbinder, Kamal Jain, and Joseph Seffi Naor. Online primal-dual algorithms for maximizing ad-auctions revenue. In *Proceedings of the 15th Annual European Conference on Algorithms*, ESA'07, page 253–264, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540755195.

[11] Archana Bura, Aria Hasanzadezonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=U4BUMoVTrB2.

[12] Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: the adversarial cost and unknown transition case. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1651–1660. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/chen21l.html.

[13] Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1180–1215. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/chen21e.html.

[14] Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward Markov decision process with constraints. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3246–3270. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/chen22i.html.

[15] Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8210–8219. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rosenberg20a.html.

[16] Alon Cohen, Haim Kaplan, Tomer Koren, and Yishay Mansour. Online markov decision processes with aggregate bandit feedback. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1301–1329. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/cohen21a.html.

[17] Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial markov decision processes with bandit feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11437–11449. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4a5c76c63f83ea45fb136d62db6c7104-Paper-Conference.pdf.

[18] Ofer Dekel and Elad Hazan. Better rates for any adversarial deterministic mdp. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 675–683, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/dekel13.html.

[19] Nikhil R. Devanur and Thomas P. Hayes. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, page 71–78, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584584. doi: 10.1145/1566374.1566384. URL https://doi.org/10.1145/1566374.1566384.

[20] Travis Dick, András György, and Csaba Szepesvári. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–512–I–520. JMLR.org, 2014.

[21] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps, 2020.

[22] Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/40538442.

[23] Jon Feldman, Aranyak Mehta, Vahab Mirrokni, and S. Muthukrishnan. Online stochastic matching: Beating 1-1/e. In *Proceedings of the 2009 50th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '09, page 117–126, USA, 2009. IEEE Computer Society. ISBN 9780769538501. doi: 10.1109/FOCS.2009.72. URL https://doi.org/10.1109/FOCS.2009.72.

[24] Fred Glover, Randy Glover, Joe Lorenzo, and Claude McMillan. The passenger-mix problem in the scheduled airlines. *Interfaces*, 12(3):73–80, 1982. doi: 10.1287/inte.12.3.73. URL https://doi.org/10.1287/inte.12.3.73.

[25] Negin Golrezaei, Hamid Nazerzadeh, and Paat Rusmevichientong. Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551, 2014. doi: 10.1287/mnsc.2014.1939. URL https://doi.org/10.1287/mnsc.2014.1939.

[26] Ramakrishna Gummadi, Peter Key, and Alexandre Proutiere. A repeated auctions under budget constraints. In *Proceedings of the Eighth Ad Auction Workshop*, volume 4. Citeseer, 2012.

[27] Anupam Gupta and Marco Molinaro. How the experts algorithm can help solve lps online. *Mathematics of Operations Research*, 41(4):1404–1431, 2016. doi: 10.1287/moor.2016.0782. URL https://doi.org/10.1287/moor.2016.0782.

[28] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL http://dx.doi.org/10.1561/2400000013.

[29] Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4259–4280. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/he22a.html.

[30] Jiashuo Jiang, Xiaocheng Li, and Jiawei Zhang. Online stochastic optimization with wasserstein based non-stationarity, 2022.

[31] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20c.html.

[32] Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16557–16566. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c0f971d8cd24364f2029fcb9ac7b71f5-Paper.pdf.

[33] R. M. Karp, U. V. Vazirani, and V. V. Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*, STOC '90, page 352–358, New York, NY, USA, 1990. Association for Computing Machinery. ISBN 0897913612. doi: 10.1145/100216.100262. URL https://doi.org/10.1145/100216.100262.

[34] Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7281–7289, Jun. 2022. doi: 10.1609/aaai.v36i7.20690. URL https://ojs.aaai.org/index.php/AAAI/article/view/20690.

[35] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. Simple and fast algorithm for binary integer and online linear programming. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9412–9421. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6abba5d8ab1f4f32243e174beb754661-Paper.pdf.

[36] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Nl7VO_Y7K4Q.

[37] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[38] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5):22–es, oct 2007. ISSN 0004-5411. doi: 10.1145/1284320.1284321. URL https://doi.org/10.1145/1284320.1284321.

[39] Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10407–10417. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5631e6ee59a4175cd06c305840562ff3-Paper.pdf.

[40] Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. pages 231–243, 01 2010.

[41] Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59(3):676–691, 2014. doi: 10.1109/TAC.2013.2292137.

[42] Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15277–15287. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ae95296e27d7f695f891cd26b4f37078-Paper.pdf.

[43] Adrian Rivera Cardoso, He Wang, and Huan Xu. Large scale markov decision processes with changing rewards. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[44] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rosenberg19a.html.

[45] Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a0872cc5b5ca4cc25076f3d868e1bdf8-Paper.pdf.

[46] Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2936–2942. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/404. URL https://doi.org/10.24963/ijcai.2021/404. Main Track.

[47] Frank A. Sonnenberg and J. Robert Beck. Markov models in medical decision making: A practical guide. *Medical Decision Making*, 13(4):322–338, 1993. doi: 10.1177/0272989X9301300409. URL https://doi.org/10.1177/0272989X9301300409. PMID: 8246705.

[48] Kalyan T. Talluri and Garrett J. van Ryzin. *The Theory and Practice of Revenue Management*, volume 68 of *International Series in Operations Research & Management Science*. Springer, 2004.

[49] Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9428–9437. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/tarbouriech20a.html.

[50] Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3274–3307. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/wei22a.html.

[51] Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and Xingyu Zhou. Provably efficient model-free algorithms for non-stationary cmdps. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6527–6570. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wei23b.html.

[52] Jia Yuan Yu and Shie Mannor. Online learning in markov decision processes with arbitrarily changing rewards and transitions. In *2009 International Conference on Game Theory for Networks*, pages 314–322, 2009. doi: 10.1109/GAMENETS.2009.5137416.

[53] Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/40538443.

[54] Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online Markov decision processes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26865–26894. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhao22c.html.

[55] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

[56] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

A Valid Occupancy Measures

In this section, we prove Lemma 3.1 that characterizes valid occupancy measures for a finite-horizon MDP. The proof is based on the reduction to the loop-free MDP setting.

Proof of Lemma 3.1. Given the finite-horizon MDP associated with transition kernel P, we may define a loop-free MDP as follows. We define its state space as $\mathcal{S}' := \mathcal{S} \times [H+1]$, which can be viewed as $H+1$ layers $\mathcal{S} \times \{h\}$ for $h \in [H+1]$. Its transition kernel P' is given by $P'((s', h+1) \mid (s,h), a) = P(s' \mid s, a, h)$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Next, given $\bar{q}$, we may define an occupancy measure q' for the loop-free MDP as $q'((s,h), a, (s', h+1)) = \bar{q}(s, a, s', h)$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Then it follows from [44, Lemma 3.1] that q' is a valid occupancy measure for the loop-free MDP with transition kernel P' if and only if q' satisfies

(C1') $\sum_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} q'((s,h), a, (s', h+1)) = 1$ for $h = 1, \ldots, H$,

(C2') $\sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} q'((s,h), a, (s', h+1)) = \sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} q'((s', h-1), a, (s,h))$ for any $s \in \mathcal{S}$ and $h = 2, \ldots, H$,

and $P^{q'} = P'$ where $P^{q'}$ is given by

$$P^{q'}((s', h+1) \mid (s,h), a) = \frac{q'((s,h), a, (s', h+1))}{\sum_{s'' \in \mathcal{S}} q'((s,h), a, (s'', h+1))} = \frac{\bar{q}(s, a, s', h)}{\sum_{s'' \in \mathcal{S}} \bar{q}(s, a, s'', h)}.$$

Here, the conditions are equivalent to (C1), (C2), and $P^{\bar{q}} = P$. Moreover, q' is a valid occupancy measure with P' if and only if q is a valid occupancy measure with P, as required. $\square$

B Auxiliary Measures and Notations

In this section, we define some auxiliary measures and functions that are useful for the analysis of Algorithm 1.

Given a policy π, we define the *reward-to-go function* for a state $s \in \mathcal{S}$ at step h with reward function $f : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ and transition kernel P as follows.

$$J^{P,\pi,f}(s,h) = \mathbb{E}\left[\sum_{\ell=h}^{H} f\left(s_\ell^{P,\pi}, a_\ell^{P,\pi}, \ell\right) \mid f, \pi, P, s_h^{P,\pi} = s\right]. \tag{11}$$

Similarly, we define the *state-action value function* for $(s,a) \in \mathcal{S} \times \mathcal{A}$ at step h with reward function $f : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ and transition kernel P as follows.

$$Q^{P,\pi,f}(s,a,h) = \mathbb{E}\left[\sum_{\ell=h}^{H} f\left(s_\ell^{P,\pi}, a_\ell^{P,\pi}, \ell\right) \mid f, \pi, P, s_h^{P,\pi} = s, a_h^{P,\pi} = a\right]. \tag{12}$$

Furthermore, given a policy π and a transition kernel P, we define $q^{P,\pi}(s, a, h \mid s', m)$ as

$$q^{P,\pi}(s, a, h \mid s', m) = \mathbb{P}\left[s_h^{P,\pi} = s,\ a_h^{P,\pi} = a \mid \pi, P, s_m^{P,\pi} = s'\right] \tag{13}$$

for $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $1 \leq m \leq h \leq H$.

Given two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{S\times A\times H}$, let $\boldsymbol{u} \odot \boldsymbol{v}$ be defined as the vector obtained from coordinate-wise products of $\boldsymbol{u}$ and $\boldsymbol{v}$, i.e. $(\boldsymbol{u} \odot \boldsymbol{v})_i = u_i \odot v_i$ for $i \in [SAH]$. Let $\vec{h}$ be an $(S \times A \times H)$-dimensional vector all of whose coordinates are h.

For $t \in [T]$, let $\mathcal{F}_t$ be the σ-algebra of events up to the beginning of episode t. More precisely, we define ξ_1 as $\xi_1 = \{f_1, g_1\}$ and for $t \geq 2$, we define ξ_t as

$$\left\{\text{trajectory } \left(s_1^{P,\pi_{t-1}}, a_1^{P,\pi_{t-1}}, \ldots, s_h^{P,\pi_{t-1}}, a_h^{P,\pi_{t-1}}\right), f_t, g_t\right\}$$

where π_{t-1} denotes the policy for episode $t-1$. Then $\mathcal{F}_t$ is defined as the σ-algebra generated by $\{\xi_1, \ldots, \xi_t\}$.

C Confidence Sets for the True Transition Kernel

Lemma 4.1 is a modification of [31, Lemma 2] to our finite-horizon MDP setting. We prove Lemma 4.1 using the empirical Bernstein inequality provided in Theorem 2.

Proof of Lemma 4.1. We will show that with probability at least $1 - 4\delta$,

$$\left|P(s' \mid s,a,h) - \bar{P}_t(s' \mid s,a,h)\right| \leq \epsilon_t(s' \mid s,a,h) \tag{14}$$

where

$$\epsilon_t(s' \mid s,a,h) = 2\sqrt{\frac{\bar{P}_t(s' \mid s,a,h)\ln\left(HSAT/\delta\right)}{\max\{1, N_t(s,a,h)-1\}}} + \frac{14\ln\left(HSAT/\delta\right)}{3\max\{1, N_t(s,a,h)-1\}}$$

holds for every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ and every episode $t \in [T]$.

Let us first consider the case $N_t(s,a,h) \leq 1$. As we may assume that $HSAT \geq 2$, it follows that

$$\epsilon_t(s' \mid s,a,h) = \frac{14\ln\left(HSAT/\delta\right)}{3\max\{1, N_t(s,a,h)-1\}} \geq \frac{14}{3}\ln 2 > 1.$$

Then (14) holds because $0 \leq P(s' \mid s,a,h), \bar{P}_t(s' \mid s,a,h) \leq 1$.

Assume that $n = N_t(s,a,h) \geq 2$. Then we define $Z_1, \ldots, Z_n$ as follows.

$$Z_j = \begin{cases} 1, & \text{if the transition after the } j\text{th visit to } (s,a,h) \text{ is } s', \\ 0, & \text{otherwise.} \end{cases}$$

Then $Z_1, \ldots, Z_n$ are i.i.d. with mean $P(s' \mid s,a,h)$, and we have

$$\sum_{j=1}^{n} Z_j = M_t(s,a,s',h).$$

Moreover, the sample variance V_n of $Z_1, \ldots, Z_n$ is given by

$$\begin{aligned} V_n &= \frac{1}{N_t(s,a,h)(N_t(s,a,h)-1)} M_t(s,a,s',h)\left(N_t(s,a,h) - M_t(s,a,s',h)\right) \\ &= \frac{N_t(s,a,h)}{(N_t(s,a,h)-1)}\bar{P}_t(s' \mid s,a,h)\left(1 - \bar{P}_t(s' \mid s,a,h)\right). \end{aligned} \tag{15}$$

Then it follows from Theorem 2 that with probability at least $1 - 2\delta/(HS^2AT)$,

$$\begin{aligned} &P(s' \mid s,a,h) - \bar{P}_t(s' \mid s,a,h) \\ &\leq \sqrt{\frac{2\bar{P}_t(s' \mid s,a,h)\left(1-\bar{P}_t(s' \mid s,a,h)\right)\ln\left(HS^2AT/\delta\right)}{N_t(s,a,h)-1}} + \frac{7\ln\left(HS^2AT/\delta\right)}{3(N_t(s,a,h)-1)}. \end{aligned} \tag{16}$$

Here, as we assumed that $N_t(s,a,h) \geq 2$, we have $N_t(s,a,h) - 1 = \max\{1, N_t(s,a,h)-1\}$. In addition, we know that $1 - \bar{P}_t(s' \mid s,a,h) \leq 1$ and that $\ln\left(HS^2AT/\delta\right) \leq 2\ln\left(HSAT/\delta\right)$. Then (16) implies that with probability at least $1 - 2\delta/(HS^2AT)$,

$$P(s' \mid s,a,h) - \bar{P}_t(s' \mid s,a,h) \leq \epsilon_t(s' \mid s,a,h) \tag{17}$$

Next, we apply Theorem 2 to variables $1 - Z_1, \ldots, 1 - Z_n$ that are i.i.d. and have mean $1 - \bar{P}_t(s' \mid s,a,h)$. Moreover, the sample variance of $1 - Z_1, \ldots, 1 - Z_n$ is also equal to V_n defined as in (15). Therefore, based on the same argument, we deduce that with probability at least $1 - 2\delta/(HS^2AT)$,

$$-P(s' \mid s,a,h) + \bar{P}_t(s' \mid s,a,h) \leq \epsilon_t(s' \mid s,a,h). \tag{18}$$

By applying union bound to (17) and (18), with probability at least $1 - 4\delta/(HS^2AT)$, (14) holds for (s,a,s',h). Furthermore, by applying union bound over all $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, it follows that with probability at least $1 - 4\delta$, (14) holds for every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, as required. $\square$

Lemma 4.1 bounds the difference between the true transition kernel P and the empirical transition kernels $\bar{P}_t$. Based on Lemma 4.1, the next lemma bounds the difference between the true transition kernel P and any $\widehat{P}$ contained in the confidence sets $\mathcal{P}_t$. Lemma C.1 is a modification of [31, Lemma 8] to our finite-horizon MDP setting.

Lemma C.1. *Let $t \in [T]$. Assume that the true transition kernel satisfies $P \in \mathcal{P}_t$. Then we have*

$$\left|\widehat{P}(s' \mid s,a,h) - P(s' \mid s,a,h)\right| \leq \epsilon_t^*(s' \mid s,a,h) \tag{19}$$

where

$$\epsilon_t^\star(s' \mid s,a,h) = 6\sqrt{\frac{P(s' \mid s,a,h)\ln\left(HSAT/\delta\right)}{\max\{1, N_t(s,a,h)\}}} + 94\frac{\ln\left(HSAT/\delta\right)}{\max\{1, N_t(s,a,h)\}}$$

for every $\widehat{P} \in \mathcal{P}_t$ and every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.

Proof. We follow the proof of [15, Lemma B.13]. Note that

$$\max\{1, N_t(s,a,h)-1\} \geq \frac{1}{2} \cdot \max\{1, N_t(s,a,h)\}$$

holds for any value of $N_t(s,a,h)$. As we assumed that $P \in \mathcal{P}_t$, we have that

$$\bar{P}_t(s' \mid s,a,h) \leq P(s' \mid s,a,h) + \sqrt{\frac{8\bar{P}_t(s' \mid s,a,h)\ln\left(HSAT/\delta\right)}{\max\{1, N_t(s,a,h)\}}} + \frac{28\ln\left(HSAT/\delta\right)}{3\max\{1, N_t(s,a,h)\}}.$$

We may view this as a quadratic inequality in terms of $x = \sqrt{\bar{P}_t(s' \mid s,a,h)}$. Note that $x^2 \leq ax + b + c$ for any

$a, b, c \geq 0$ implies that $x \leq a + \sqrt{b} + \sqrt{c}$. Therefore, we deduce that

$$\sqrt{\bar{P}_t(s' \mid s,a,h)} \leq \sqrt{P(s' \mid s,a,h)} + \left(2\sqrt{2} + \sqrt{\frac{28}{3}}\right) \sqrt{\frac{\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}}$$
$$\leq \sqrt{P(s' \mid s,a,h)} + 13\sqrt{\frac{\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}}.$$

Using this bound on $\sqrt{\bar{P}_t(s' \mid s,a,h)}$, we obtain the following.

$$\begin{aligned}
\epsilon_t(s' \mid s,a,h) &\leq \sqrt{\frac{8\bar{P}_t(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}} + \frac{28\ln(HSAT/\delta)}{3\max\{1, N_t(s,a,h)\}} \\
&\leq \sqrt{\frac{8P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}} + \left(13\sqrt{8} + \frac{28}{3}\right)\frac{\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}} \\
&\leq 3\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}} + 47\frac{\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}} \\
&= \frac{1}{2} \cdot \epsilon_t^\star(s' \mid s,a,h)
\end{aligned} \tag{20}$$

Since we assumed that $P \in \mathcal{P}_t$,

$$\left|P(s' \mid s,a,h) - \bar{P}_t(s' \mid s,a,h)\right| \leq \frac{1}{2} \cdot \epsilon_t^\star(s' \mid s,a,h).$$

Moreover, for any $\widehat{P} \in \mathcal{P}_t$, we have

$$\left|\widehat{P}(s' \mid s,a,h) - \bar{P}_t(s' \mid s,a,h)\right| \leq \epsilon_t(s' \mid s,a,h) \leq \frac{1}{2} \cdot \epsilon_t^\star(s' \mid s,a,h).$$

By the triangle inequality, it follows that

$$\left|\widehat{P}(s' \mid s,a,h) - P(s' \mid s,a,h)\right| \leq \epsilon_t^\star(s' \mid s,a,h),$$

as required. □

D Two Technical Lemmas

In this section, we prove two technical lemmas, Lemma 5.1 and Lemma 5.2, that are crucial in proving the desired upper bound on the regret.

D.1 Proof of Lemma 5.1

In this section, we provide the proof of Lemma 5.1.

Proof of Lemma 5.1. For ease of notation, let $\mathbb{E}_t[\cdot]$ denotes $\mathbb{E}[\cdot \mid f, \pi_t, P]$, and let s_h and a_h denote s_h^{P,π_t} and

a_h^{P,π_t}, respectively for $h \in [H]$. Note that

$$\begin{aligned}
\mathbb{E}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f}\rangle^2\right] &= \mathbb{E}_t\left[\left(\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_t(s,a,h) f(s,a,h)\right)^2\right] \\
&= \mathbb{E}_t\left[\left(\sum_{h=1}^{H} f(s_h,a_h,h)\right)^2\right] \\
&\le 2\mathbb{E}_t\left[\sum_{h=1}^{H} f(s_h,a_h,h)\left(\sum_{m=h}^{H} f(s_m,a_m,m)\right)\right] \\
&= 2\mathbb{E}_t\left[\sum_{h=1}^{H} \mathbb{E}_t\left[f(s_h,a_h,h)\left(\sum_{m=h}^{H} f(s_m,a_m,m)\right) \mid s_h, a_h\right]\right] \\
&= 2\mathbb{E}_t\left[\sum_{h=1}^{H} f(s_h,a_h,h)\mathbb{E}_t\left[\sum_{m=h}^{H} f(s_m,a_m,m) \mid s_h, a_h\right]\right] \\
&= 2\mathbb{E}_t\left[\sum_{h=1}^{H} f(s_h,a_h,h) Q^{P,\pi_t,f}(s_h,a_h,h)\right] \\
&= 2\mathbb{E}_t\left[\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_t(s,a,h) f(s,a,h) Q^{P,\pi_t,f}(s,a,h)\right]
\end{aligned}$$

where the first inequality holds because $(\sum_{h=1}^{H} x_h)^2 \le 2\sum_{h=1}^{H} x_h(\sum_{m=h}^{H} x_h)$. Moreover,

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_t(s,a,h) f(s,a,h) Q^{P,\pi_t,f}(s,a,h)\right] \\
&= \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(s,a,h) Q^{P,\pi_t,f}(s,a,h)\mathbb{E}_t\left[n_t(s,a,h)\right] \\
&= \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(s,a,h) Q^{P,\pi_t,f}(s,a,h) q_t(s,a,h) \\
&= \langle \boldsymbol{q_t}, \boldsymbol{f} \odot \boldsymbol{Q^{P,\pi_t,f}}\rangle.
\end{aligned}$$

Therefore, it follows that

$$\mathbb{E}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f}\rangle^2\right] \le \langle \boldsymbol{q_t}, \boldsymbol{f} \odot \boldsymbol{Q^{P,\pi_t,f}}\rangle.$$

Next, observe that

$$\begin{aligned}
\langle \boldsymbol{q_t}, \boldsymbol{f} \odot \boldsymbol{Q^{P,\pi_t,f}} \rangle &\leq \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} Q^{P,\pi_t,f}(s,a,h) q_t(s,a,h) \\
&= \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi(a \mid s,h) Q^{P,\pi_t,f}(s,a,h) \left(\sum_{a'\in\mathcal{A}} q_t(s,a',h) \right) \\
&= \sum_{h=1}^{H} \sum_{s\in\mathcal{S}} J^{P,\pi_t,f}(s,h) \left(\sum_{a'\in\mathcal{A}} q_t(s,a',h) \right) \\
&= \sum_{h=1}^{H} \sum_{s\in\mathcal{S}} \left(\sum_{m=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} q_t(s',a',m \mid s,h) f(s',a',m) \right) \left(\sum_{a'\in\mathcal{A}} q_t(s,a',h) \right) \\
&= \sum_{h=1}^{H} \sum_{m=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \sum_{s\in\mathcal{S}} q_t(s',a',m \mid s,h) \left(\sum_{a'\in\mathcal{A}} q_t(s,a',h) \right) f(s',a',m) \\
&= \sum_{h=1}^{H} \sum_{m=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} q_t(s',a',m) f(s',a',m) \\
&= \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} h \cdot q_t(s,a,h) f(s,a,h) \\
&= \langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f} \rangle
\end{aligned}$$

where the first inequality holds because $f(s,a,h) \leq 1$ for any (s,a,h), the first equality holds because

$$q_t(s,a,h) = \pi(a \mid s,h) \sum_{a'\in\mathcal{A}} q_t(s,a',h),$$

the fifth equality follows from

$$\sum_{s\in\mathcal{S}} q_t(s',a',m \mid s,h) \left(\sum_{a'\in\mathcal{A}} q_t(s,a',h) \right) = q_t(s',a',m).$$

Therefore, we get that $\langle \boldsymbol{q_t}, \boldsymbol{f} \odot \boldsymbol{Q^{P,\pi_t,f}} \rangle \leq \langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f} \rangle$, as required. □

D.2 Proof of Lemma 5.2

In this section, we provide the proof of Lemma 5.2.

The following lemma is from the first statement of Lemma 7 in [12] with a few modifications to adapt the proof to our setting.

Lemma D.1. [12, Lemma 7] *Let π be a policy, and let $\widetilde{P}, \widehat{P}$ be two different transition kernels. We denote by $\widetilde{q}$ the occupancy measure $q^{\widetilde{P},\pi}$ associated with $\widetilde{P}$ and π, and we denote by $\widehat{q}$ the occupancy measure $q^{\widehat{P},\pi}$ associated with $\widehat{P}$ and π. Then*

$$\begin{aligned}
&\widehat{q}(s,a,h) - \widetilde{q}(s,a,h) \\
&= \sum_{(s',a',s'')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \sum_{m=1}^{h-1} \widetilde{q}(s',a',m) \left(\widehat{P}(s'' \mid s',a',m) - \widetilde{P}(s'' \mid s',a',m) \right) \widehat{q}(s,a,h \mid s'',m+1).
\end{aligned}$$

Proof. We prove the first statement by induction on h. When $h = 1$, note that

$$\widehat{q}(s,a,h) = \widetilde{q}(s,a,h) = \pi(a \mid s, 1) \cdot p(s).$$

Hence, both the left-hand side and right-hand side are equal to 0. Next assume that the equality holds with $h-1 \geq 1$. Then we consider h. By the definition of occupancy measures,

$$\begin{aligned}
&\widehat{q}(s,a,h) - \widetilde{q}(s,a,h) \\
&= \pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} (\widehat{P}(s \mid s',a',h-1)\widehat{q}(s',a',h-1) - \widetilde{P}(s \mid s',a',h-1)\widetilde{q}(s',a',h-1)) \\
&= \underbrace{\pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \widehat{P}(s \mid s',a',h-1)(\widehat{q}(s',a',h-1) - \widetilde{q}(s',a',h-1))}_{\text{Term 1}} \\
&\quad \underbrace{+\, \pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \widetilde{q}(s',a',h-1)(\widehat{P}(s \mid s',a',h-1) - \widetilde{P}(s \mid s',a',h-1))}_{\text{Term 2}}.
\end{aligned}$$

To provide an upper bound on Term 1, we use the induction hypothesis for $h-1$:

$$\begin{aligned}
&\widehat{q}(s',a',h-1) - \widetilde{q}(s',a',h-1) \\
&= \sum_{(s'',a'',s''') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-2} \widetilde{q}(s'',a'',m) \left((\widehat{P} - \widetilde{P})(s''' \mid s'',a'',m)\right) \widehat{q}(s',a',h-1 \mid s''',m+1)
\end{aligned}$$

where

$$(\widehat{P} - \widetilde{P})(s''' \mid s'',a'',m) = \widehat{P}(s''' \mid s'',a'',m) - \widetilde{P}(s''' \mid s'',a'',m).$$

In addition, observe that

$$\pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \widehat{P}(s \mid s',a',h-1)\widehat{q}(s',a',h-1 \mid s''',m+1) = \widehat{q}(s,a,h \mid s''',m+1).$$

Therefore, it follows that Term 1 is equal to

$$\begin{aligned}
&\sum_{(s'',a'',s''') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-2} \widetilde{q}(s'',a'',m) \left((\widehat{P} - \widetilde{P})(s''' \mid s'',a'',m)\right) \widehat{q}(s,a,h \mid s''',m+1) \\
&= \sum_{(s',a',s'') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-2} \widetilde{q}(s',a',m) \left(\widehat{P}(s'' \mid s',a',m) - \widetilde{P}(s'' \mid s',a',m)\right) \widehat{q}(s,a,h \mid s'',m+1).
\end{aligned}$$

Next, we upper bound Term 2. Note that

$$\widehat{q}(s,a,h \mid s'',h) = \pi(a \mid s'',h) \cdot \mathbf{1}\left[s'' = s\right].$$

Then it follows that

$$\begin{aligned}
&\pi(a \mid s,h)(\widehat{P}(s \mid s',a',h-1) - \widetilde{P}(s \mid s',a',h-1)) \\
&= \sum_{s'' \in \mathcal{S}} \mathbf{1}\left[s'' = s\right] \cdot \pi(a \mid s'',h)(\widehat{P}(s'' \mid s',a',h-1) - \widetilde{P}(s'' \mid s',a',h-1)) \\
&= \sum_{s'' \in \mathcal{S}} \widehat{q}(s,a,h \mid s'',h)(\widehat{P}(s'' \mid s',a',h-1) - \widetilde{P}(s'' \mid s',a',h-1)),
\end{aligned}$$

implying in turn that Term 2 equals

$$\sum_{(s',a',s'')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \widetilde{q}(s',a',h-1)(\widehat{P}(s''\mid s',a',h-1)-\widetilde{P}(s''\mid s',a',h-1))\widehat{q}(s,a,h\mid s'',h).$$

Adding the equivalent expression of Term 1 and that of Term 2 that we have obtained, we get the right-hand side of the statement. $\square$

Based on Lemma C.1 and Lemma D.1, we show the following lemma, which is a modification of [12, Lemma 7, the second statement].

Lemma D.2. *Let π be a policy, and let $\widetilde{P}, \widehat{P}$ be two different transition kernels. We denote by $\widetilde{q}$ the occupancy measure $q^{\widetilde{P},\pi}$ associated with $\widetilde{P}$ and π, and we denote by $\widehat{q}$ the occupancy measure $q^{\widehat{P},\pi}$ associated with $\widehat{P}$ and π. If $\widehat{P}, \widetilde{P} \in \mathcal{P}_t$, then we have*

$$\begin{aligned}
&|\langle \widehat{\boldsymbol{q}}-\widetilde{\boldsymbol{q}}, \boldsymbol{f}\rangle| \\
&= \left|\sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widetilde{q}(s,a,h)\left(\widehat{P}(s'\mid s,a,h)-\widetilde{P}(s'\mid s,a,h)\right)J^{\widehat{P},\pi,f}(s',h+1)\right| \\
&\leq H \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widetilde{q}(s,a,h)\epsilon_t^\star(s'\mid s,a,h)
\end{aligned}$$

where $\widehat{\boldsymbol{q}}, \widetilde{\boldsymbol{q}}, \boldsymbol{f}$ are the vector representations of $\widehat{q}, \widetilde{q}, f : \mathcal{S}\times\mathcal{A}\times[H]\to\mathbb{R}$.

Proof. First, observe that

$$\langle \widehat{\boldsymbol{q}}-\widetilde{\boldsymbol{q}}, \boldsymbol{f}\rangle = \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \left(\widehat{q}(s,a,h)-\widetilde{q}(s,a,h)\right)f(s,a,h).$$

By Lemma D.1, the right-hand side can be rewritten so that we obtain the following.

$$\begin{aligned}
&\langle \widehat{\boldsymbol{q}}-\widetilde{\boldsymbol{q}}, \boldsymbol{f}\rangle \\
&= \sum_{(s,a,h)}\sum_{(s',a',s'')}\sum_{m=1}^{h-1} \widetilde{q}(s',a',m)\left((\widehat{P}-\widetilde{P})(s''\mid s',a',m)\right)\widehat{q}(s,a,h\mid s'',m+1)f(s,a,h) \\
&= \sum_{m=1}^{H}\sum_{(s',a',s'')} \widetilde{q}(s',a',m)\left((\widehat{P}-\widetilde{P})(s''\mid s',a',m)\right)\sum_{(s,a,h):h>m}\widehat{q}(s,a,h\mid s'',m+1)f(s,a,h) \\
&= \sum_{m=1}^{H}\sum_{(s',a',s'')} \widetilde{q}(s',a',m)\left((\widehat{P}-\widetilde{P})(s''\mid s',a',m)\right)J^{\widehat{P},\pi,f}(s'',m+1) \\
&= \sum_{h=1}^{H}\sum_{(s',a',s'')} \widetilde{q}(s',a',h)\left(\widehat{P}(s''\mid s',a',h)-\widetilde{P}(s''\mid s',a',h)\right)J^{\widehat{P},\pi,f}(s'',h+1).
\end{aligned}$$

Since $\widehat{P}, P \in \mathcal{P}_t$, Lemma C.1 implies that

$$
\begin{aligned}
|\langle \widehat{\boldsymbol{q}} - \widetilde{\boldsymbol{q}}, \boldsymbol{f} \rangle| &\leq \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \left| \widehat{P}(s'' \mid s',a',h) - \widetilde{P}(s'' \mid s',a',h) \right| J^{\widehat{P},\pi,f}(s'',h+1) \\
&\leq \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \epsilon_t^{\star}(s'' \mid s',a',h) J^{\widehat{P},\pi,f}(s'',h+1) \\
&\leq H \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \epsilon_t^{\star}(s'' \mid s',a',h) \\
&= H \sum_{(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]} \widetilde{q}(s,a,h) \epsilon_t^{\star}(s' \mid s,a,h)
\end{aligned}
$$

where the third inequality holds because $J^{\widehat{P},\pi,f}(s'',h+1) \leq H$, as required. □

Lemma D.3. *Let π be a policy, and let $\widetilde{P}, \widehat{P}$ be two different transition kernels. We denote by $\widetilde{q}$ the occupancy measure $q^{\widetilde{P},\pi}$ associated with $\widetilde{P}$ and π, and we denote by $\widehat{q}$ the occupancy measure $q^{\widehat{P},\pi}$ associated with $\widehat{P}$ and π. Let $(s,h) \in \mathcal{S} \times [H]$, and consider $\widetilde{q}(\cdot \mid s,h), \widehat{q}(\cdot \mid s,h) : \mathcal{S} \times \mathcal{A} \times \{h, \ldots, H\}$. If $\widehat{P}, \widetilde{P} \in \mathcal{P}_t$, then we have*

$$
\left| \langle \widehat{\boldsymbol{q}}_{\boldsymbol{(s,h)}} - \widetilde{\boldsymbol{q}}_{\boldsymbol{(s,h)}}, \boldsymbol{f}_{\boldsymbol{(h)}} \rangle \right| \leq H \sum_{(s',a',s'',m) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \{h,\ldots,H\}} \widetilde{q}(s',a',m \mid s,h) \epsilon_t^{\star}(s'' \mid s',a',m)
$$

where $\widehat{\boldsymbol{q}}_{\boldsymbol{(s,h)}}, \widetilde{\boldsymbol{q}}_{\boldsymbol{(s,h)}}, \boldsymbol{f}_{\boldsymbol{(h)}}$ are the vector representations of $\widehat{q}(\cdot \mid s,h), \widetilde{q}(\cdot \mid s,h), f : \mathcal{S} \times \mathcal{A} \times \{h, \ldots, H\} \to \mathbb{R}$.

Proof. The proof follows the same argument of Lemma D.1 and Lemma D.2. □

The following lemma is from [12] after some changes to adapt to our setting.

Lemma D.4. [12, Lemma 4] *Let π_t be the policy for episode t, and let q_t denote the occupancy measure q^{P,π_t}. Let $f : \mathcal{S} \times \mathcal{A} \times [H] \to [0,\infty)$ be an arbitrary reward function, and define $\mathbb{V}_t(s,a,h) = \mathrm{Var}_{s' \sim P(\cdot|s,a,h)} \left[J^{P,\pi_t,f}(s',h+1) \right]$. Then*

$$
\langle \boldsymbol{q_t}, \mathbb{V}_{\boldsymbol{t}} \rangle \leq \mathrm{Var}\left[\langle \boldsymbol{n_t}, \boldsymbol{f} \rangle \mid f, \pi_t, P \right]
$$

where $\boldsymbol{q_t}, \mathbb{V}_{\boldsymbol{t}}, \boldsymbol{n_t}, \boldsymbol{f}$ are the vector representations of $q_t, \mathbb{V}_t, n_t, f : \mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}$.

Proof. For ease of notation, let s_h and a_h denote s_h^{P,π_t} and a_h^{P,π_t}, respectively for $h \in [H]$. Moreover, let $J(s,h)$ denote $J^{P,\pi_t,f}(s,h)$ for $(s,h) \in \mathcal{S} \times [H]$. Note that

$$
\langle \boldsymbol{n_t}, \boldsymbol{f} \rangle = \sum_{(s,a,h) \mathcal{S} \times \mathcal{A} \times [H]} f(s,a,h) n_t(s,a,h) = \sum_{h=1}^{H} f(s_h, a_h, h) .
$$

For ease of notation, let $\mathbb{E}_t[\cdot]$ and $\mathrm{Var}_t[\cdot]$ denote $\mathbb{E}[\cdot \mid f, \pi_t, P]$ and $\mathrm{Var}[\cdot \mid f, \pi_t, P]$, respectively. Then

$$
\begin{aligned}
\mathbb{E}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f} \rangle \right] &= \mathbb{E}_t\left[\sum_{h=1}^{H} f(s_h, a_h, h) \right] \\
&= \mathbb{E}_t\left[\mathbb{E}\left[\sum_{h=1}^{H} f(s_h, a_h, h) \mid f, \pi_t, P, s_1 \right] \right] \\
&= \mathbb{E}_t\left[J(s_1, 1) \right]
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\mathrm{Var}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f}\rangle\right] \\
&= \mathbb{E}_t\left[\left(\sum_{h=1}^{H} f\left(s_h, a_h, h\right) - \mathbb{E}_t\left[J(s_1,1)\right]\right)^2\right] \\
&= \mathbb{E}_t\left[\left(\sum_{h=1}^{H} f\left(s_h, a_h, h\right) - J(s_1,1) + J(s_1,1) - \mathbb{E}_t\left[J(s_1,1)\right]\right)^2\right] \\
&= \mathbb{E}_t\left[\left(\sum_{h=1}^{H} f\left(s_h, a_h, h\right) - J(s_1,1)\right)^2\right] + \mathbb{E}_t\left[\left(J(s_1,1) - \mathbb{E}_t\left[J(s_1,1)\right]\right)^2\right] \\
&\quad + 2\mathbb{E}_t\left[\left(\sum_{h=1}^{H} f\left(s_h, a_h, h\right) - J(s_1,1)\right)\left(J(s_1,1) - \mathbb{E}_t\left[J(s_1,1)\right]\right)\right] \\
&\geq \mathbb{E}_t\left[\left(\sum_{h=1}^{H} f\left(s_h, a_h, h\right) - J(s_1,1)\right)^2\right]
\end{aligned}
$$

where the inequality is by $\mathbb{E}_t\left[J(s_1,1) - \mathbb{E}_t\left[J(s_1,1)\right] \mid s_1\right] = 0$ and $\left(J(s_1,1) - \mathbb{E}_t\left[J(s_1,1)\right]\right)^2 \geq 0$. Therefore,

$$
\begin{aligned}
&\mathrm{Var}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f}\rangle\right] \\
&\geq \mathbb{E}_t\left[\left(\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2) + f\left(s_1, a_1, 1\right) + J(s_2,2) - J(s_1,1)\right)^2\right].
\end{aligned}
$$

Note that

$$
\mathbb{E}_t\left[\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2) \mid s_1, a_1, s_2\right] = \mathbb{E}_t\left[\sum_{h=2}^{H} f\left(s_h, a_h, h\right) \mid s_2\right] - J(s_2,2) = 0. \tag{21}
$$

Then

$$
\begin{aligned}
&\mathrm{Var}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f}\rangle\right] \\
&\geq \mathbb{E}_t\left[\left(\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2)\right)^2\right] + \mathbb{E}_t\left[\left(f\left(s_1, a_1, 1\right) + J(s_2,2) - J(s_1,1)\right)^2\right] \\
&\quad + 2\mathbb{E}_t\left[\mathbb{E}_t\left[\left(\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2)\right)\left(f\left(s_1, a_1, 1\right) + J(s_2,2) - J(s_1,1)\right) \mid s_1, a_1, s_2\right]\right] \\
&= \mathbb{E}_t\left[\left(\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2)\right)^2\right] + \mathbb{E}_t\left[\left(f\left(s_1, a_1, 1\right) + J(s_2,2) - J(s_1,1)\right)^2\right] \\
&\quad + 2\mathbb{E}_t\left[\left(f\left(s_1, a_1, 1\right) + J(s_2,2) - J(s_1,1)\right)\mathbb{E}_t\left[\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2) \mid s_1, a_1, s_2\right]\right] \\
&= \mathbb{E}_t\left[\left(\sum_{h=2}^{H} f\left(s_h, a_h, h\right) - J(s_2,2)\right)^2\right] + \mathbb{E}_t\left[\left(f\left(s_1, a_1, 1\right) + J(s_2,2) - J(s_1,1)\right)^2\right]
\end{aligned}
$$

where the last equality follows from (21). Here, the second term from the right-most side can be bounded from below as follows.

$$
\begin{aligned}
&\mathbb{E}_t\left[\left(f\left(s_1,a_1,1\right)+J(s_2,2)-J(s_1,1)\right)^2\right]\\
&=\mathbb{E}_t\left[\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)+J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\right)^2\right]\\
&=\mathbb{E}_t\left[\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)\right)^2\right]\\
&\quad+\mathbb{E}_t\left[\left(J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\right)^2\right]\\
&\quad+2\mathbb{E}_t\left[\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)\right)\left(J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\right)\right]\\
&=\mathbb{E}_t\left[\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)\right)^2\right]\\
&\quad+\mathbb{E}_t\left[\left(J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\right)^2\right]\\
&\geq\mathbb{E}_t\left[\mathbb{V}_t(s_1,a_1,1)\right]
\end{aligned}
$$

where third equality holds because

$$
\begin{aligned}
&\mathbb{E}_t\left[\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)\right)\left(J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\right)\mid s_1,a_1\right]\\
&=\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)\right)\mathbb{E}_t\left[J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\mid s_1,a_1\right]\\
&=\left(f\left(s_1,a_1,1\right)+\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)-J(s_1,1)\right)\times 0
\end{aligned}
$$

and the last inequality holds because

$$
\mathbb{E}_t\left[\left(J(s_2,2)-\sum_{s'\in\mathcal{S}}P(s'\mid s_1,a_1,1)J(s',2)\right)^2\right]=\mathbb{E}_t\left[\mathbb{V}_t(s_1,a_1,1)\right].
$$

Then it follows that

$$
\begin{aligned}
\mathrm{Var}_t\left[\langle\boldsymbol{n_t},\boldsymbol{f}\rangle\right]&\geq\mathbb{E}_t\left[\left(\sum_{h=1}^{H}f\left(s_h,a_h,h\right)-J(s_1,1)\right)^2\right]\\
&\geq\mathbb{E}_t\left[\left(\sum_{h=2}^{H}f\left(s_h,a_h,h\right)-J(s_2,2)\right)^2\right]+\mathbb{E}_t\left[\mathbb{V}_t(s_1,a_1,1)\right].
\end{aligned}
$$

Repeating the same argument, we deduce that

$$\mathrm{Var}_t\left[\langle \boldsymbol{n_t}, \boldsymbol{f}\rangle\right] \geq \sum_{h=1}^{H} \mathbb{E}_t\left[\mathbb{V}_t(s_h, a_h, h)\right] = \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} q_t(s_h, a_h, h)\mathbb{V}_t(s_h, a_h, h) = \langle \boldsymbol{q_t}, \mathbb{V}_{\boldsymbol{t}}\rangle,$$

as required. $\square$

Next, using Theorem 3 that states the Bernstein-type concentration inequality for a martingale difference sequence, we prove the following lemma that is useful for our analysis. Lemma D.5 is a modification of [31, Lemma 10] and [12, Lemma 8] to our finite-horizon MDP setting.

Lemma D.5. *With probability at least* $1-2\delta$, *we have*

$$\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} = O\left(SAH\ln T + H\ln(H/\delta)\right) \tag{22}$$

$$\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}} = O\left(H\sqrt{SAT} + SAH\ln T + H\ln(H/\delta)\right) \tag{23}$$

Proof. Note that

$$\sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} = \sum_{t=1}^{T}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} + \sum_{t=1}^{T} Y_t \tag{24}$$

where

$$Y_t = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{-n_t(s,a,h) + q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}}.$$

As (10) holds for every $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, we know that $Y_1,\ldots,Y_T$ is a martingale difference sequence. We know that $Y_t\leq 1$ for each $t\in[T]$. Let $\mathbb{E}_t[\cdot]$ denote $\mathbb{E}[\cdot \mid \mathcal{F}_t, \pi_t, P]$. Then we deduce

$$\begin{aligned}
\mathbb{E}_t\left[Y_t^2\right] &= \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_t\left[(n_t(s,a,h) - q_t(s,a,h))(n_t(s',a',h) - q_t(s',a',h))\right]}{\max\{1, N_t(s,a,h)\}\cdot\max\{1, N_t(s',a',h)\}}\\
&= \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_t\left[n_t(s,a,h)n_t(s',a',h) - q_t(s,a,h)q_t(s',a',h)\right]}{\max\{1, N_t(s,a,h)\}\cdot\max\{1, N_t(s',a',h)\}}\\
&\leq \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_t\left[n_t(s,a,h)n_t(s',a',h)\right]}{\max\{1, N_t(s,a,h)\}\cdot\max\{1, N_t(s',a',h)\}}\\
&\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_t\left[n_t(s,a,h)\right]}{\max\{1, N_t(s,a,h)\}}\\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}}
\end{aligned}$$

where the second equality holds because (10) implies that

$$\mathbb{E}_t\left[q_t(s,a,h)n_t(s',a',h)\right] = \mathbb{E}_t\left[q_t(s',a',h)n_t(s,a,h)\right] = q_t(s,a,h)q_t(s',a',h),$$

the second inequality holds because $n_t(s,a,h)n_t(s',a',h) = 0$ if $(s,a)\neq(s',a')$, and the last equality is from (10).

Then we may apply Theorem 3 with $\lambda = 1/2$, and we deduce that with probability at least $1 - \delta/H$,

$$\sum_{t=1}^{T} Y_t \leq \frac{1}{2} \sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} + 2\ln(H/\delta).$$

Plugging this inequality to (24), it follows that

$$\sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} = 2\sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} + 4\ln(H/\delta).$$

Here, the first term on the right-hand side can be bounded as follows. We have

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{n_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} \\
&= \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\max\{1, N_{t+1}(s,a,h)\}} + \sum_{t=1}^{T} \left(\frac{n_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} - \frac{n_t(s,a,h)}{\max\{1, N_{t+1}(s,a,h)\}} \right) \\
&\leq \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\max\{1, N_{t+1}(s,a,h)\}} + \sum_{t=1}^{T} \left(\frac{1}{\max\{1, N_t(s,a,h)\}} - \frac{1}{\max\{1, N_{t+1}(s,a,h)\}} \right) \\
&\leq \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\max\{1, N_{t+1}(s,a,h)\}} + 1 \\
&= O(\ln T).
\end{aligned}$$

where the first inequality is due to $n_t(s,a,h) \leq 1$ and the last inequality holds because

$$n_t(s,a,h) = N_{t+1}(s,a,h) - N_t(s,a,h) \quad \text{and} \quad N_T(s,a,h) + n_T(s,a,h) \leq T.$$

Therefore, it follows that

$$\sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} = O(SA\ln T).$$

As a result, for any fixed $h \in [H]$,

$$\sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} = O\left(SA\ln T + \ln\left(H/\delta\right)\right)$$

holds with probability at least $1 - \delta/H$. By union bound, (22) holds with probability at least $1 - \delta$.

Next, we will show that (23) holds.

$$\sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}} = \sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}} + \sum_{t=1}^{T} Z_t \tag{25}$$

where

$$Z_t = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{-n_t(s,a,h) + q_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}}.$$

As (10) holds for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we know that $Z_1, \ldots, Z_T$ is a martingale difference sequence. We know that $Z_t \leq 1$ for each $t \in [T]$. Then we deduce

$$
\begin{aligned}
\mathbb{E}_t\left[Z_t^2\right] &\leq \sum_{(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_t\left[n_t(s,a,h) n_t(s',a',h)\right]}{\sqrt{\max\{1, N_t(s,a,h)\}} \cdot \sqrt{\max\{1, N_t(s',a',h)\}}} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_t\left[n_t(s,a,h)\right]}{\max\{1, N_t(s,a,h)\}} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}}
\end{aligned}
$$

where the first inequality is derived by the same argument when bounding $\mathbb{E}_t[Y_t^2]$, the first equality holds because $n_t(s,a,h)n_t(s',a',h) = 0$ if $(s,a) \neq (s',a')$, and the last equality is from (10). Then we may apply Theorem 3 with $\lambda = 1$, and we deduce that with probability at least $1 - \delta/H$,

$$
\sum_{t=1}^{T} Z_t \leq \sum_{t=1}^{T} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} + \ln(H/\delta).
$$

Then with probability at least $1 - 2\delta$, (22) holds and

$$
\begin{aligned}
\sum_{h\in[H]} \sum_{t=1}^{T} Z_t &\leq \sum_{t=1}^{T} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_t(s,a,h)}{\max\{1, N_t(s,a,h)\}} + H\ln(H/\delta) \\
&= O(SAH \ln T + H \ln(H/\delta)).
\end{aligned} \tag{26}
$$

holds. Moreover, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}} \\
&= \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_{t+1}(s,a,h)\}}} + \sum_{t=1}^{T} \left(\frac{n_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}} - \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_{t+1}(s,a,h)\}}} \right) \\
&\leq \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_{t+1}(s,a,h)\}}} + \sum_{t=1}^{T} \left(\frac{1}{\sqrt{\max\{1, N_t(s,a,h)\}}} - \frac{1}{\sqrt{\max\{1, N_{t+1}(s,a,h)\}}} \right) \\
&\leq \sum_{t=1}^{T} \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_{t+1}(s,a,h)\}}} + 1 \\
&= O(\sqrt{N_{T+1}(s,a,h)} + 1).
\end{aligned}
$$

where the last equality holds because $n_t(s,a,h) = N_{t+1}(s,a,h) - N_t(s,a,h)$. Then

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{n_t(s,a,h)}{\sqrt{\max\{1, N_t(s,a,h)\}}} &= O\left(\sum_{\mathcal{S}\times\mathcal{A}\times[H]} \left(\sqrt{N_{T+1}(s,a,h)} + 1 \right) \right) \\
&= O\left(\sqrt{SAH \sum_{\mathcal{S}\times\mathcal{A}\times[H]} N_{T+1}(s,a,h)} + SAH \right) \\
&= O\left(H\sqrt{SAT} + SAH \right)
\end{aligned}
$$

where the second equality is due to the Cauchy-Schwarz inequality. Then it follows from (25) and (26) that (23) holds. $\square$

Lemma D.6. *Assume that $P \in \mathcal{P}_t$ for every episode $t \in [T]$. Then*

$$\sum_{t=1}^{T} \left| \sum_{(s,a,s',h)} q_t(s,a,h) \left((P - P_t)(s' \mid s,a,h) \right) \left(\left(J^{P_t,\pi_t,f} - J^{P,\pi_t,f} \right)(s',h+1) \right) \right|$$
$$= O\left(H^2 S^2 A \left(\ln(HSAT/\delta) \right)^2 \right)$$

for any $P_t \in \mathcal{P}_t$ where $(P - P_t)(s' \mid s,a,h) = P(s' \mid s,a,h) - P_t(s' \mid s,a,h)$ and $\left(J^{P_t,\pi_t,f} - J^{P,\pi_t,f} \right)(s',h+1) = J^{P_t,\pi_t,f}(s',h+1) - J^{P,\pi_t,f}(s',h+1)$.

Proof. Let $\boldsymbol{q}^{\boldsymbol{P_t},\boldsymbol{\pi_t}}_{\boldsymbol{(s',h+1)}}, \boldsymbol{q}^{\boldsymbol{P},\boldsymbol{\pi_t}}_{\boldsymbol{(s',h+1)}}, \boldsymbol{f}$ be the vector representations of $q^{P_t,\pi_t}(\cdot \mid s',h+1), q^{P,\pi_t}(\cdot \mid s',h+1), f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, respectively. Note that

$$\begin{aligned}
&\sum_{t=1}^{T} \left| \sum_{(s,a,s',h)} q_t(s,a,h) \left((P - P_t)(s' \mid s,a,h) \right) \left(\left(J^{P_t,\pi_t,f} - J^{P,\pi_t,f} \right)(s',h+1) \right) \right| \\
&\le \sum_{t=1}^{T} \sum_{(s,a,s',h)} q_t(s,a,h) \epsilon_t^\star(s' \mid s,a,h) \left| \left(J^{P_t,\pi_t,f} - J^{P,\pi_t,f} \right)(s',h+1) \right| \\
&= \sum_{t=1}^{T} \sum_{(s,a,s',h)} q_t(s,a,h) \epsilon_t^\star(s' \mid s,a,h) \left| \langle \boldsymbol{q}^{\boldsymbol{P_t},\boldsymbol{\pi_t}}_{\boldsymbol{(s',h+1)}} - \boldsymbol{q}^{\boldsymbol{P},\boldsymbol{\pi_t}}_{\boldsymbol{(s',h+1)}}, \boldsymbol{f} \rangle \right| \\
&\le H \sum_{t=1}^{T} \sum_{(s,a,s',h)} q_t(s,a,h) \epsilon_t^\star(s' \mid s,a,h) \sum_{(s'',a'',s'''),m \ge h} q_t(s'',a'',m \mid s',h+1) \epsilon_t^\star(s''' \mid s'',a'',m)
\end{aligned}$$

where the first inequality is from Lemma C.1, the first equality holds because $J^{P_t,\pi_t,f}(s',h+1) = \langle \boldsymbol{q}^{\boldsymbol{P_t},\boldsymbol{\pi_t}}_{\boldsymbol{(s',h+1)}}, \boldsymbol{f} \rangle$ and $J^{P,\pi_t,f}(s',h+1) = \langle \boldsymbol{q}^{\boldsymbol{P},\boldsymbol{\pi_t}}_{\boldsymbol{(s',h+1)}}, \boldsymbol{f} \rangle$, the second inequality is due to Lemma D.3. Then plugging in the definition of ϵ_t^*, it follows that

$$
\begin{aligned}
&\frac{1}{H\ln(HSAT/\delta)}\sum_{t=1}^{T}\left|\sum_{(s,a,s',h)} q_t(s,a,h)\left((P-P_t)(s'\mid s,a,h)\right)\left(\left(J^{P_t,\pi_t,f}-J^{P,\pi_t,f}\right)(s',h+1)\right)\right| \\
&= O\left(\sum_{\substack{t,(s,a,s',h)\\(s'',a'',s''')\\m\geq h+1}} q_t(s,a,h)\sqrt{\frac{P(s'\mid s,a,h)}{\max\{1,n_t(s,a,h)\}}}q_t(s'',a'',m\mid s',h+1)\sqrt{\frac{P(s'''\mid s'',a'',m)}{\max\{1,n_t(s'',a'',m)\}}}\right) \\
&= O\left(\sum_{\substack{t,(s,a,s',h)\\(s'',a'',s''')\\m\geq h+1}} \sqrt{\tfrac{q_t(s,a,h)P(s'''\mid s'',a'',m)q_t(s'',a'',m\mid s',h+1)}{\max\{1,n_t(s,a,h)\}}}\sqrt{\tfrac{q_t(s,a,h)P(s'\mid s,a,h)q_t(s'',a'',m\mid s',h+1)}{\max\{1,n_t(s'',a'',m)\}}}\right) \\
&= O\left(\sqrt{\sum_{\substack{t,(s,a,s',h)\\(s'',a'',s''')\\m\geq h+1}} \tfrac{q_t(s,a,h)P(s'''\mid s'',a'',m)q_t(s'',a'',m\mid s',h+1)}{\max\{1,n_t(s,a,h)\}}}\sqrt{\sum_{\substack{t,(s,a,s',h)\\(s'',a'',s''')\\m\geq h+1}} \tfrac{q_t(s,a,h)P(s'\mid s,a,h)q_t(s'',a'',m\mid s',h+1)}{\max\{1,n_t(s'',a'',m)\}}}\right) \\
&= O\left(\sqrt{S\sum_{t=1}^{T}\sum_{(s,a,h)}\frac{q_t(s,a,h)}{\max\{1,n_t(s,a,h)\}}}\sqrt{S\sum_{t=1}^{T}\sum_{(s'',a'',m)}\frac{q_t(s'',a'',m)}{\max\{1,n_t(s'',a'',m)\}}}\right) \\
&= O\left(S^2AH\ln T + SH\ln\left(H/\delta\right)\right)
\end{aligned}
$$

where the third equality follows from the Cauchy-Schwarz inequality and the last equality is due to Lemma D.5. Therefore, we deduce that

$$
\begin{aligned}
&\sum_{t=1}^{T}\left|\sum_{(s,a,s',h)} q_t(s,a,h)\left((P-P_t)(s'\mid s,a,h)\right)\left(\left(J^{P_t,\pi_t,f}-J^{P,\pi_t,f}\right)(s',h+1)\right)\right| \\
&= O\left(H^2S^2A\ln T\ln(HSAT/\delta) + SH^2\ln\left(H/\delta\right)\ln(HSAT/\delta)\right) \\
&= O\left(H^2S^2A\left(\ln(HSAT/\delta)\right)^2\right),
\end{aligned}
$$

as required. □

Proof of Lemma 5.2. Let us define

$$
\mu_t(s,a,h) = \mathbb{E}_{s'\sim P(\cdot\mid s,a,h)}\left[J^{P,\pi_t,f_t}(s',h+1)\right].
$$

Note that

$$
\begin{aligned}
&\sum_{t=1}^{T} |\langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t} \rangle| \\
&= \sum_{t=1}^{T} \left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_t(s,a,h) \left(P(s' \mid s,a,h) - P_t(s' \mid s,a,h)\right) J^{P_t,\pi_t,f_t}(s',h+1) \right| \\
&\le \sum_{t=1}^{T} \left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_t(s,a,h) \left(P(s' \mid s,a,h) - P_t(s' \mid s,a,h)\right) J^{P,\pi_t,f_t}(s',h+1) \right| \\
&\quad + O\left(H^2 S^2 A \left(\ln(HSAT/\delta)\right)^2\right)
\end{aligned}
$$

where the first equality is due to Lemma D.1 and the first inequality is by Lemma D.6. Moreover,

$$
\begin{aligned}
&\sum_{t=1}^{T} \left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_t(s,a,h) \left(P(s' \mid s,a,h) - P_t(s' \mid s,a,h)\right) J^{P,\pi_t,f_t}(s',h+1) \right| \\
&= \sum_{t=1}^{T} \left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_t(s,a,h) \left(\left(P - P_t\right)(s' \mid s,a,h)\right) \left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right) \right| \\
&\le \sum_{t=1}^{T} \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_t(s,a,h) \epsilon_t^\star(s' \mid s,a,h) \left| J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h) \right| \\
&\le O\left(\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} q_t(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}} \left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \right) \\
&\quad + O\left(HS \sum_{t=1}^{T} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_t(s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}} \right) \\
&\le O\left(\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} q_t(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}} \left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \right) \\
&\quad + O\left(H^2 S^2 A \left(\ln(HSAT/\delta)\right)^2\right)
\end{aligned}
$$

where $\left(P - P_t\right)(s' \mid s,a,h) = P(s' \mid s,a,h) - P_t(s' \mid s,a,h)$, the first equality holds because $\sum_{s'\in\mathcal{S}} \left(P - P_t\right)(s' \mid s,a,h) = 0$ and $\mu_t(s,a,h)$ is independent of s', the first inequality is due to Lemma C.1, and the second inequality is from Lemma C.1 and $\left| J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h) \right| \le H$. Note that

$$
q_t(s,a,h) = \mathbb{E}\left[n_t(s,a,h) \mid \mathcal{F}_t, \pi_t, P\right],
$$

which implies that

$$
\begin{aligned}
&\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} q_t(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}} \left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \\
&= \sum_{t=1}^{T} \mathbb{E}\left[X_t \mid \mathcal{F}_t, \pi_t, P\right]
\end{aligned}
$$

where

$$X_t = \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2}.$$

Here, we have

$$0 \le X_t \le O\left(HS \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} n_t(s,a,h)\sqrt{\ln(HSAT/\delta)}\right) = O(H^2 S\sqrt{\ln(HSAT/\delta)}).$$

Then it follows from Lemma F.2 that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[X_t \mid \mathcal{F}_t, \pi_t, P\right] \\
&\le 2\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \\
&\quad + O\left(H^2 S \left(\ln(HSAT/\delta)\right)^{3/2}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \\
&\le \sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_{t+1}(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \\
&\quad + H\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\left(\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}} - \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_{t+1}(s,a,h)\}}}\right) \\
&\le \sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_{t+1}(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \\
&\quad + H\sqrt{S}\sum_{t=1}^{T} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \left(\sqrt{\frac{\ln(HSAT/\delta)}{\max\{1, N_t(s,a,h)\}}} - \sqrt{\frac{\ln(HSAT/\delta)}{\max\{1, N_{t+1}(s,a,h)\}}}\right) \\
&\le \sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAT/\delta)}{\max\{1, N_{t+1}(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2} \\
&\quad + O\left(H^2 S^{3/2} A\sqrt{\ln(HSAT/\delta)}\right).
\end{aligned}$$

where the first inequality holds because $\left|J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right| \le H$, the second inequality holds because $n_t(s,a,h) \le 1$ and the Cauchy-Schwarz inequality implies that

$$\sum_{s'\in\mathcal{S}} \sqrt{P(s' \mid s,a,h)} \le \sqrt{S\sum_{s'\in\mathcal{S}} P(s' \mid s,a,h)} = \sqrt{S},$$

and the third inequality follows from

$$\sum_{t=1}^{T}\left(\sqrt{\frac{1}{\max\{1,N_t(s,a,h)\}}}-\sqrt{\frac{1}{\max\{1,N_{t+1}(s,a,h)\}}}\right)\leq\sqrt{\frac{1}{\max\{1,N_1(s,a,h)\}}}=1.$$

Next, the Cauchy-Schwarz inequality implies the following.

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\sqrt{\frac{P(s'\mid s,a,h)\ln(HSAT/\delta)}{\max\{1,N_{t+1}(s,a,h)\}}\left(J^{P,\pi_t,f_t}(s',h+1)-\mu_t(s,a,h)\right)^2}\\
&\leq\sqrt{\sum_{t=1}^{T}\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)P(s'\mid s,a,h)\left(J^{P,\pi_t,f_t}(s',h+1)-\mu_t(s,a,h)\right)^2}\\
&\quad\times\sqrt{\sum_{t=1}^{T}\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)\frac{\ln(HSAT/\delta)}{\max\{1,N_{t+1}(s,a,h)\}}}
\end{aligned}$$

Here, the second term can be bounded as follows.

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} n_t(s,a,h)\frac{\ln(HSAT/\delta)}{\max\{1,N_{t+1}(s,a,h)\}}\\
&=S\ln\left(\frac{HSAT}{\delta}\right)\sum_{t=1}^{T}\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}\frac{n_t(s,a,h)}{\max\{1,N_{t+1}(s,a,h)\}}\\
&=S\ln\left(\frac{HSAT}{\delta}\right)\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}\sum_{t=1}^{T}\frac{n_t(s,a,h)}{\max\{1,N_{t+1}(s,a,h)\}}\\
&=O\left(HS^2A\left(\ln\left(HSAT/\delta\right)\right)^2\right).
\end{aligned}$$

$(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, we define

$$\mathbb{V}_t(s,a,h)=\operatorname*{Var}_{s'\sim P(\cdot|s,a,h)}\left[J^{P,\pi_t,f_t}(s',h+1)\right].$$

Then

$$\begin{aligned}
\mathbb{V}_t(s,a,h)&=\mathbb{E}_{s'\sim P(\cdot|s,a,h)}\left[\left(J^{P,\pi_t,f_t}(s',h+1)-\mu_t(s,a,h)\right)^2\right]\\
&=\sum_{s'\in\mathcal{S}}P(s'\mid s,a,h)\left(J^{P,\pi_t,f_t}(s',h+1)-\mu_t(s,a,h)\right)^2
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_t(s,a,h)P(s' \mid s,a,h)\left(J^{P,\pi_t,f_t}(s',h+1) - \mu_t(s,a,h)\right)^2 \\
&= \sum_{t=1}^{T} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} n_t(s,a,h)\mathbb{V}_t(s,a,h) \\
&= \sum_{t=1}^{T} \langle \boldsymbol{q_t}, \mathbb{V}_{\boldsymbol{t}} \rangle + \sum_{t=1}^{T} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} (n_t(s,a,h) - q_t(s,a,h))\mathbb{V}_t(s,a,h) \\
&\le \sum_{t=1}^{T} \mathrm{Var}\left[\langle n_t, f_t\rangle \mid f_t, \pi_t, P\right] + O\left(H^3\sqrt{T\ln(1/\delta)}\right)
\end{aligned}$$

where $\mathbb{V}_{\boldsymbol{t}} \in \mathbb{R}^{SAH}$ is the vector representation of $\mathbb{V}_t$ and the inequality follows from Lemma D.4, $\mathbb{V}_t(s,a,h) \le H^2$,

$$\begin{aligned}
\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} (n_t(s,a,h) - q_t(s,a,h))\mathbb{V}_t(s,a,h) &\le \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} (n_t(s,a,h) + q_t(s,a,h))H^2 \\
&\le 2H^3,
\end{aligned}$$

and Lemma F.1. Therefore, we finally have proved that

$$\begin{aligned}
&\sum_{t=1}^{T} |\langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t}\rangle| \\
&= O\left(\sqrt{HS^2A\left(\ln\frac{HSAT}{\delta}\right)^2\left(\sum_{t=1}^{T}\mathrm{Var}\left[\langle n_t, f_t\rangle \mid f_t, \pi_t, P\right] + H^3\sqrt{T\ln\frac{1}{\delta}}\right)}\right) \\
&\quad + O\left(H^2S^2A\left(\ln\frac{HSAT}{\delta}\right)^2\right).
\end{aligned}$$

Moreover, we know from Lemma 5.1 that

$$\mathrm{Var}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid f_t, \pi_t, P\right] \le \mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid f_t, \pi_t, P\right] \le 2\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t}\rangle,$$

and therefore, it follows that

$$\begin{aligned}
&\sum_{t=1}^{T} |\langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t}\rangle| \\
&= O\left(\left(\sqrt{HS^2A\left(\sum_{t=1}^{T}\langle \boldsymbol{q_t}, \vec{h}\odot \boldsymbol{f_t}\rangle + H^3\right)} + H^2S^2A\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right),
\end{aligned}$$

as required. □

Based on Lemmas 5.1 and 5.2, we can prove Lemma 5.4 that bounds the difference between the expected reward and the realized reward.

E Proof of the Main Theorem

E.1 Bounds on the Regret Terms (II) and (III)

Based on Lemmas 5.1 and 5.2, we can prove Lemma 5.3 that bounds the regret due to the estimation error and Lemma 5.4 that bounds the difference between the expected reward and the realized reward. We define filtration $\{\mathcal{F}_t\}_{t=0}^T$ as follows. $\mathcal{F}_0 = \{0, \Omega\}$.

Proof of Lemma 5.4. We closely follow the proof of [12, Theorem 6]. Recall that π_t is the policy for episode t and q_t denotes the occupancy measure q^{P,π_t}. Then Lemma 5.1 implies that

$$\mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid f_t, \pi_t, P\right] \leq 2\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t}\rangle$$

where $\boldsymbol{q_t}, \boldsymbol{n_t}, \boldsymbol{f_t}$ are the vector representations of $q_t, n_t, f_t : \mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}$. Since π_t is $\mathcal{F}_t$-measurable, it follows that

$$\mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid \mathcal{F}_t, P\right] \leq 2\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t}\rangle.$$

Then it follows that

$$\sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid \mathcal{F}_t, P\right] \leq 2\sum_{t=1}^T \langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \vec{h} \odot \boldsymbol{f_t}\rangle + 2\sum_{t=1}^T \langle \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \vec{h} \odot \boldsymbol{f_t}\rangle.$$

Note that the first term on the right-hand side can be bounded as follows.

$$\sum_{t=1}^T \langle \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \vec{h} \odot \boldsymbol{f_t}\rangle = O(H^2 T).$$

To upper bound the first term, we consider

$$\sum_{t=1}^T \langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \vec{h} \odot \boldsymbol{f_t}\rangle \leq H \sum_{t=1}^T \langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t}\rangle.$$

Applying Lemma 5.2 with function f_t, we deduce that with probability at least $1 - 4\delta$,

$$\begin{aligned}
&\sum_{t=1}^T \langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t}\rangle \\
&= O\left(\left(\sqrt{HS^2A\left(\sum_{t=1}^T \langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t}\rangle + H^3\sqrt{T}\right)} + H^2S^2A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right) \\
&= O\left(\left(\sqrt{HS^2A\left(H^2T + H^3\sqrt{T}\right)} + H^2S^2A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right) \\
&= O\left(\left(\sqrt{H^4S^2AT} + H^2S^2A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right) \\
&= O\left(\left(HT + H^3S^2A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right)
\end{aligned}$$

where the second equality holds because $\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t}\rangle = O(H^2)$ and the fourth equality holds because $\sqrt{H^4S^2AT} \leq H\sqrt{T} + H^3S^2A$. Then it follows that

$$\sum_{t=1}^{T}\langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \vec{h} \odot \boldsymbol{f_t}\rangle = O\left((H^2T + H^4S^2A)\left(\ln\frac{HSAT}{\delta}\right)^2\right).$$

Therefore, we obtain

$$\sum_{t=1}^{T}\mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid \mathcal{F}_t, P\right] = O\left((H^2T + H^4S^2A)\left(\ln\frac{HSAT}{\delta}\right)^2\right).$$

Next, we apply Theorem 3 with λ is set to

$$\lambda = \frac{1}{\sqrt{H^2T + H^4S^2A}} \leq \frac{1}{H} \leq \frac{1}{\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle}.$$

Then we get that with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}\langle \boldsymbol{q_t} - \boldsymbol{n_t}, \boldsymbol{f_t}\rangle &\leq \lambda\sum_{t=1}^{T}\mathbb{E}\left[\langle \boldsymbol{q_t} - \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid \mathcal{F}_t, P\right] + \frac{1}{\lambda}\ln\frac{1}{\delta}\\
&\leq \lambda\sum_{t=1}^{T}\mathbb{E}\left[\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle^2 \mid \mathcal{F}_t, P\right] + \frac{1}{\lambda}\ln\frac{1}{\delta}\\
&= O\left(\left(H\sqrt{T} + H^2S\sqrt{A}\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right).
\end{aligned}$$

Since we have

$$\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle = \sum_{h=1}^{H} f_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right),$$

it follows that

$$\sum_{t=1}^{T}\langle \boldsymbol{n_t}, \boldsymbol{f_t}\rangle - \sum_{t=1}^{T}\sum_{h=1}^{H} f_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right) = O\left(\left(H\sqrt{T} + H^2S\sqrt{A}\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right),$$

as required. □

Proof of Lemma 5.3. We closely follow the proof of [12, Theorem 6]. Recall that π_t is the policy for episode t and q_t denotes the occupancy measure q^{P,π_t}. Let P_t be the transition kernel induced by $\widehat{\boldsymbol{q}}_{\boldsymbol{t}}$ defined as in (3).

Applying Lemma 5.2 with function f_t, we deduce that with probability at least $1-4\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}\langle \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}, \boldsymbol{f_t}\rangle\\
&= O\left(\left(\sqrt{HS^2A\left(\sum_{t=1}^{T}\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t}\rangle + H^3\sqrt{T}\right)} + H^2S^2A\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right)\\
&= O\left(\left(\sqrt{H^3S^2AT + H^4S^2A\sqrt{T}} + H^2S^2A\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right)\\
&= O\left(\left(\sqrt{H^3S^2AT + H^3S^2AT + H^5S^2A} + H^2S^2A\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right)\\
&= O\left(\left(H^{3/2}S\sqrt{AT} + H^{5/2}S\sqrt{A} + H^2S^2A\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right)
\end{aligned}$$

where the second equality holds because $\langle \boldsymbol{q_t}, \vec{h} \odot \boldsymbol{f_t} \rangle = O(H^2)$ and the third equality holds because $H^4 S^2 A \sqrt{T} = O(H^3 S^2 AT + H^5 S^2 A)$. ☐

E.2 Bounds on the Regret Term (I)

In this section, we provide a proof of Lemma 5.6]. We follow the analysis of the online dual mirror descent algorithm due to [7, Theorem 1]. In our analysis, we need Lemma 5.3 and Lemma 5.4 that provide bounds on the regret terms (II) and (III). This is because our dual mirror descent algorithm works over $\Delta(P,1), \ldots, \Delta(P,T)$ instead of the true feasible set $\Delta(P)$.

Proof of Lemma 5.6. For $t \in [T]$, let G_t denote the amount of resource consumed in episode t. We define the stopping time τ of Algorithm 1 as

$$\min \left\{ t : \sum_{s=1}^{t} G_s + H \geq TH\rho \right\}.$$

By definition, we have $TH\rho - \sum_{t=1}^{\tau-1} G_t > H$. Since $G_t \leq H$ for any $t \in [T]$, it follows that $TH\rho - \sum_{t=1}^{\tau} G_t > 0$, and therefore, Algorithm 1 does not terminate until the end of episode τ. Then we have

$$G_t = \langle \boldsymbol{n_t}, \boldsymbol{g_t} \rangle, \quad t \leq \tau.$$

By Lemma 4.2, with probability at least $1 - 4\delta$, we have

$$\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_t \rangle = \widehat{L}_t(\lambda_t) - \lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_t \rangle) \geq L_t(\lambda_t) - \lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_t \rangle), \quad \forall t \in [T] \tag{27}$$

where

$$\widehat{L}_t(\lambda) = \max_{\boldsymbol{q} \in \Delta(P,t)} \left\{ \langle \boldsymbol{f_t}, \boldsymbol{q} \rangle + \lambda(H\rho - \langle \boldsymbol{g_t}, \boldsymbol{q} \rangle) \right\},$$
$$L_t(\lambda) = \max_{\boldsymbol{q} \in \Delta(P)} \left\{ \langle \boldsymbol{f_t}, \boldsymbol{q} \rangle + \lambda(H\rho - \langle \boldsymbol{g_t}, \boldsymbol{q} \rangle) \right\}.$$

Suppose that the pair (f_t, g_t) of reward and resource consumption functions follows a distribution Γ. Then we define $\bar{L}(\lambda)$ as

$$\bar{L}(\lambda) = \mathbb{E}_{(f,g)\sim\Gamma} \left[\max_{\boldsymbol{q} \in \Delta(P)} \left\{ \langle \boldsymbol{f}, \boldsymbol{q} \rangle + \lambda(H\rho - \langle \boldsymbol{g}, \boldsymbol{q} \rangle) \right\} \right].$$

Since (f_t, g_t) for $t \in [T]$ are i.i.d. with distribution Γ, it follows that

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} L_t(\lambda) \right] = \bar{L}(\lambda) = \mathbb{E}_{(f,g)\sim\Gamma} \left[\max_{\boldsymbol{q} \in \Delta(P)} \left\{ \langle \boldsymbol{f}, \boldsymbol{q} \rangle - \lambda \langle \boldsymbol{g}, \boldsymbol{q} \rangle \right\} \right] + \lambda H\rho.$$

Let $\mathcal{H}_0 = \{\emptyset, \Omega\}$ and $\mathcal{H}_t$ be defined as the σ-algebra generated by $\{f_1, g_1, \ldots, f_t, g_t\}$. Consider

$$Z_t = \sum_{s=1}^{t} \lambda_s(H\rho - \langle \boldsymbol{g_s}, \widehat{\boldsymbol{q}}_s \rangle) - \sum_{s=1}^{t} \mathbb{E}\left[\lambda_s(H\rho - \langle \boldsymbol{g_s}, \widehat{\boldsymbol{q}}_s \rangle) \mid \mathcal{H}_{s-1} \right]$$

for $t \in [T]$. Then Z_t is $\mathcal{H}_t$-measurable and $\mathbb{E}[Z_{t+1} \mid \mathcal{H}_t] = Z_t$. Therefore, $Z_1, \ldots, Z_T$ is a martingale. Since the stopping time τ is with respect to $\{\mathcal{H}_t\}_{t\in[T]}$ and τ is bounded, the Optional Stopping Theorem implies that

$$\mathbb{E}\left[\sum_{t=1}^{\tau} \lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_t \rangle) \right] = \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{E}\left[\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_t \rangle) \mid \mathcal{H}_{t-1} \right] \right].$$

Likewise, we can argue by the Optional Stopping Theorem that

$$\mathbb{E}\left[\sum_{t=1}^{\tau}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{\tau}\mathbb{E}\left[\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid \mathcal{H}_{t-1}\right]\right].$$

Taking the conditional expectation with respect to $\mathcal{H}_{t-1}$ of both sides of (27), it follows that

$$\begin{aligned}
&\mathbb{E}\left[\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid \mathcal{H}_{t-1}\right]\\
&\geq \mathbb{E}\left[\max_{\boldsymbol{q}\in\Delta(P)}\left\{\langle \boldsymbol{f_t}, \boldsymbol{q}\rangle + \lambda_t(H\rho - \langle \boldsymbol{g_t}, \boldsymbol{q}\rangle)\right\} \mid \mathcal{H}_{t-1}\right] - \mathbb{E}\left[\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \mid \mathcal{H}_{t-1}\right]\\
&= \mathbb{E}\left[\mathbb{E}_{(f_t,g_t)\sim\Gamma}\left[\max_{\boldsymbol{q}\in\Delta(P)}\left\{\langle \boldsymbol{f_t}, \boldsymbol{q}\rangle + \lambda_t(H\rho - \langle \boldsymbol{g_t}, \boldsymbol{q}\rangle)\right\}\right] \mid \mathcal{H}_{t-1}\right] - \mathbb{E}\left[\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \mid \mathcal{H}_{t-1}\right]\\
&= \mathbb{E}\left[\bar{L}(\lambda_t) \mid \mathcal{H}_{t-1}\right] - \mathbb{E}\left[\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \mid \mathcal{H}_{t-1}\right]\\
&= \bar{L}(\lambda_t) - \mathbb{E}\left[\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \mid \mathcal{H}_{t-1}\right]
\end{aligned}$$

where the first equality is due to the tower rule and the last equality holds because λ_t is $\mathcal{H}_{t-1}$-measurable. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau}\bar{L}(\lambda_t)\right] - \mathbb{E}\left[\sum_{t=1}^{\tau}\mathbb{E}\left[\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \mid \mathcal{H}_{t-1}\right]\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{\tau}\bar{L}(\lambda_t)\right] - \mathbb{E}\left[\sum_{t=1}^{\tau}\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle)\right]
\end{aligned}$$

where the second equality comes from the Optional Stopping Theorem. Furthermore, note that $L_t(\lambda)$ is the maximum of linear functions in terms of λ, so $L_t(\lambda)$ is convex for any $t \in [T]$. Then $\bar{L}(\lambda)$ is also convex with respect to λ, and therefore,

$$\mathbb{E}\left[\sum_{t=1}^{\tau}\bar{L}(\lambda_t)\right] \geq \mathbb{E}\left[\tau\bar{L}\left(\frac{1}{\tau}\sum_{t=1}^{\tau}\lambda_t\right)\right].$$

This implies that

$$\mathbb{E}\left[\sum_{t=1}^{\tau}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle\right] \geq \mathbb{E}\left[\tau\bar{L}\left(\frac{1}{\tau}\sum_{t=1}^{\tau}\lambda_t\right)\right] - \mathbb{E}\left[\sum_{t=1}^{\tau}\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle)\right].$$

Next, consider the second term on the right-hand side of this inequality:

$$\sum_{t=1}^{\tau}\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle).$$

Let $w_t(\lambda)$ be defined as

$$w_t(\lambda) = \lambda(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle).$$

Then the dual update rule

$$\lambda_{t+1} = \operatorname*{argmin}_{\lambda\in\mathbb{R}_+}\left\{\eta\left(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle\right)^\top\lambda + D(\lambda,\lambda_t)\right\}$$

corresponds to the online mirror descent algorithm applied to the linear functions $w_t(\lambda)$ for $t \in [T]$. Since ψ is 1-strongly convex and $|H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle| \leq 2H$, the standard analysis of online mirror descent (see [28]) gives us that

$$\sum_{t=1}^{\tau}\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) - \sum_{t=1}^{\tau}\lambda(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \leq 2H^2\eta\tau + \frac{1}{\eta}D(\lambda,\lambda_1) \leq 2H^2\eta T + \frac{1}{\eta}D(\lambda,\lambda_1).$$

Next, note that for any $\lambda \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma})\right] &= \frac{T-\tau}{T}\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma})\right] + \frac{\tau}{T}\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma})\right] \\ &\leq (T-\tau)H + \frac{\tau}{T}\mathbb{E}\left[\sum_{t=1}^{T} L_t\left(\lambda\right)\right] \\ &= (T-\tau)H + \tau\bar{L}\left(\lambda\right)\end{aligned}$$

where the second inequality is implied by $\mathrm{OPT}(\vec{\gamma}) \leq TH$ and Lemma 5.5. In particular, we set $\lambda = \frac{1}{\tau}\sum_{t=1}^{\tau}\lambda_t$, and obtain

$$\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma})\right] \leq (T-\tau)H + \tau\bar{L}\left(\frac{1}{\tau}\sum_{t=1}^{\tau}\lambda_t\right).$$

Then it follows that

$$\begin{aligned}&\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] \\ &\leq \mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{\tau}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] \\ &\leq \mathbb{E}\left[(T-\tau)H + \sum_{t=1}^{\tau}\lambda_t(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \mid P\right] \\ &\leq \mathbb{E}\left[(T-\tau)H + \sum_{t=1}^{\tau}\lambda(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) + 2H^2\eta T + \frac{1}{\eta}D(\lambda, \lambda_1) \mid P\right]\end{aligned}$$

where the last inequality is from the online mirror descent analysis.

If $\tau = T$, then we set $\lambda = 0$, in which case

$$\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] \leq 2H^2\eta T + \frac{1}{\eta}D(0, \lambda_1).$$

If $\tau < T$, then we have

$$\sum_{t=1}^{\tau} G_t + H = \sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{n_t}\rangle + H \geq TH\rho.$$

In this case, we set $\lambda = 1/\rho$. Then

$$\begin{aligned}&\sum_{t=1}^{\tau}\lambda(H\rho - \langle \boldsymbol{g_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle) \\ &= \tau H - \frac{1}{\rho}\sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{n_t}\rangle + \frac{1}{\rho}\sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{n_t} - \boldsymbol{q_t}\rangle + \frac{1}{\rho}\sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle.\end{aligned}$$

By Lemma 5.3 and Lemma 5.4, with probability at least $1 - 13\delta$, we have

$$\begin{aligned}&\tau H - \frac{1}{\rho}\sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{n_t}\rangle + \frac{1}{\rho}\sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{n_t} - \boldsymbol{q_t}\rangle + \frac{1}{\rho}\sum_{t=1}^{\tau}\langle \boldsymbol{g_t}, \boldsymbol{q_t} - \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \\ &\leq \tau H - TH + \frac{H}{\rho} + O\left(\left(\frac{H^{3/2}}{\rho}S\sqrt{AT} + \frac{H^{5/2}}{\rho}S^2 A\right)\left(\ln\frac{HSAT}{\delta}\right)^2\right).\end{aligned}$$

In this case, we deduce that

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] \\
&\leq 2H^2\eta T + \frac{1}{\eta} D\left(\frac{1}{\rho}, \lambda_1\right) + O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right) \\
&\leq 2H^2\eta T + \frac{C}{\eta}\left(\frac{1}{\rho} - \lambda_1\right)^2 + O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right) \\
&\leq 2H^2\eta T + \frac{2C}{\eta\rho^2}\left(1 + \lambda_1\right)^2 + O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right)
\end{aligned}
$$

where the second inequality follows from (2) and the third inequality holds because $\rho < 1$. Setting

$$
\eta = \frac{1}{\rho H\sqrt{T}},
$$

we deduce that

$$
\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] = O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln \frac{HSAT}{\delta}\right)^2\right).
$$

Now we may set

$$
\delta = \frac{1}{13HT}.
$$

Note that with probability at most $1/HT$,

$$
\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] \leq \mathrm{OPT}(\vec{\gamma}) \leq HT.
$$

Moreover, with probability at least $1 - 1/HT$,

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] &= O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln \frac{H^2SAT^2}{13}\right)^2\right) \\
&= O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln HSAT\right)^2\right).
\end{aligned}
$$

Then it follows that

$$
\mathbb{E}\left[\mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle \mid P\right] = O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right)\left(\ln HSAT\right)^2\right),
$$

as required. □

E.3 Proof of Theorem 1

Now we are ready to prove Theorem 1. Recall that

$$
\begin{aligned}
&\mathrm{Regret}\left(\vec{\gamma}, \vec{\pi}\right) = \mathrm{OPT}(\vec{\gamma}) - \mathrm{Reward}\left(\vec{\gamma}, \vec{\pi}\right) \\
&= \mathrm{OPT}(\vec{\gamma}) - \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}}\rangle + \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \widehat{\boldsymbol{q}}_{\boldsymbol{t}} - \boldsymbol{q_t}\rangle + \sum_{t=1}^{T}\langle \boldsymbol{f_t}, \boldsymbol{q_t}\rangle - \sum_{t=1}^{T}\sum_{h=1}^{H} f_t\left(s_h^{P,\pi_t}, a_h^{P,\pi_t}, h\right).
\end{aligned}
$$

Then it follows from Lemmas 5.6, 5.3, and 5.4 that

$$\mathbb{E}\left[\text{Regret}\left(\vec{\gamma}, \vec{\pi}\right) \mid P\right] = O\left(\left(\frac{H^{3/2}}{\rho} S\sqrt{AT} + \frac{H^{5/2}}{\rho} S^2 A\right) \left(\ln HSAT\right)^2\right),$$

as required.

F Concentration Inequalities

Theorem 2. [37, Theorem 4] *Let $Z_1, \ldots, Z_n \in [0,1]$ be i.i.d. random variables with mean z, and let $\delta > 0$. Then with probability at least $1-\delta$,*

$$z - \frac{1}{n}\sum_{j=1}^{n} Z_j \leq \lambda\sqrt{\frac{2V_n \ln(2/\delta)}{n}} + \frac{7\ln(2/\delta)}{3(n-1)}$$

where V_n is the sample variance given by

$$V_n = \frac{1}{n(n-1)} \sum_{1 \leq j < k \leq n} (Z_j - Z_k)^2.$$

Next, we need the following Bernstein-type concentration inequality for martingales due to [9]. We take the version used in [31, Lemma 9].

Theorem 3. [9, Theorem 1] *Let $Y_1, \ldots, Y_T$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1, \ldots, \mathcal{F}_T$. Assume that $Y_t \leq R$ almost surely for all $t \in [T]$. Then for any $\delta \in (0,1)$ and $\lambda \in (0, 1/R]$, with probability at least $1-\delta$, we have*

$$\sum_{t=1}^{T} Y_t \leq \lambda \sum_{t=1}^{T} \mathbb{E}\left[Y_t^2 \mid \mathcal{F}_t\right] + \frac{\ln(1/\delta)}{\lambda}.$$

Lemma F.1 (Azuma's inequality)**.** *Let $Y_1, \ldots, Y_T$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1, \ldots, \mathcal{F}_T$. Assume that $|Y_t| \leq B$ for $t \in [T]$. Then with probability at least $1-\delta$, we have*

$$\left|\sum_{t=1}^{T} Y_t\right| \leq B\sqrt{2T\ln(2/\delta)}.$$

Next, we need the following concentration inequality due to [15].

Lemma F.2. [15, Lemma D.4] *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables adapted to the filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$. Suppose that $0 \leq X_n \leq B$ holds almost surely for all n. Then with probability at least $1-\delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\sum_{i=1}^{n} \mathbb{E}\left[X_i \mid \mathcal{F}_i\right] \leq 2\sum_{i=1}^{n} X_i + 4B\ln\left(2n/\delta\right).$$