Online Markov Decision Processes with Max-Min Fairness

Wang Chi Cheung [1]

Abstract

Consider an online Markov decision process with max-min fairness (MDP-MMF), which involves maximizing $K \geq 2$ types of rewards in a latent MDP. The agent maximizes the minimum total reward among the K total rewards. We design a learning algorithm that converges to the optimum as the number of time steps grows, despite the model uncertainty. Our algorithm involves a novel thresholding criterion that controls the number of switches among stationary policies, while promoting max-min fairness.

1. Model, and a Brief Literature Review

An instance of MDP-MMF is specified by the tuple $(\mathcal{S}, s_1, \mathcal{A}, T, \mathcal{O})$. The set $\mathcal{S}$ is a finite state space, and $s_1 \in \mathcal{S}$ is the starting state. In the collection $\mathcal{A} = \{\mathcal{A}_s\}_{s\in\mathcal{S}}$, each set $\mathcal{A}_s$ is finite, and contains the actions associated with state s. The quantity T is the number of time steps.

When the agent takes action $a \in \mathcal{A}_s$ at state s, he receives the array of stochastic outcomes $(s', U(s,a))$, governed by the outcome distribution $\mathcal{O}(s,a)$. The quantity $s' \in \mathcal{S}$ is the subsequent state he transits to. The quantity $U = (U_k(s,a))_{k=1}^K$ is a random vector lying in $[0,1]^K$, where $K \geq 2$. For each $1 \leq k \leq K$, the random variable $U_k(s,a) \in [0,1]$ is the amount of type-k reward he receives. We allow the random variables $s', U_1(s,a), \ldots U_K(s,a)$ to be arbitrarily correlated. For subsequent discussion, we denote $p(s'|s,a)$ as the probability of transiting to s' from s, a, and denote $v_k(s,a) = \mathbf{E}[U_k(s,a)]$.

While the agent knows[1] $\mathcal{S}, s_1, \mathcal{A}, T$, he does not know the outcome distribution $\mathcal{O}$. In addition, he does not know p, v. The dynamics of MDP-MMF is formalized as follows. At time $t \in \{1, \ldots T\}$, the agent observes his current state s_t. Then, he selects an action $a_t \in \mathcal{A}_{s_t}$ using a non-anticipatory policy. The choice of a_t only depends on s_t and the observations during time 1 to $t-1$. After that, he receives the stochastic feedback $(s_{t+1}, V_t(s_t,a_t))$ (distributed as $\mathcal{O}(s_t,a_t)$), where we denote $V_t(s_t,a_t) = (V_{t,k}(s_t,a_t))_{k=1}^K$ with $V_{t,k}(s_t,a_t)$ being the type-k reward received at time t. The agent's objective is to maximize

$$\min_k \bar{V}_{1:T,k}, \text{ where } \bar{V}_{1:T,k} = \frac{1}{T}\sum_{t=1}^{T} V_{t,k}(s_t,a_t),$$

where $\min_{k\in\{1,\ldots,K\}}$ is written as $\min_k$. The agent maximizes the minimum average reward among the K types of rewards.

We impose the following Assumption 1 to ensure learnability. For any $s, s' \in \mathcal{S}$ and any stationary policy π, the travel time from s to s' under π is equal to the random variable $\Lambda(s'|\pi, s) := \min\{t : s_{t+1} = s', s_1 = s, s_{\tau+1} \sim p(\cdot|s_\tau, \pi(s_\tau))\ \forall \tau\}$.

Assumption 1. *The MDP-MMF instance is communicating, that is, the quantity* $D := \max_{s,s'\in\mathcal{S}} \min_{\text{stationary } \pi} \mathbb{E}[\Lambda(s'|\pi,s)]$ *is finite. We call* D *the diameter of* $\mathcal{M}$.

The same reachability assumption is made in (Jaksch et al., 2010). Since p is not known, the underlying diameter D is also not known. To measure the effectiveness of a policy, we rephrase the objective as the minimization of regret: $\text{Reg} := \text{opt}(\mathsf{P}) - \min_k \bar{V}_{1:T,k}$. The offline benchmark $\text{opt}(\mathsf{P})$ is the optimum of the convex optimization problem (P) (Puterman, 1994; Altman, 1999):

$$\max_x \min_{k\in\{1,\ldots,K\}} \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} v_k(s,a)x(s,a)$$

$$\text{s.t.} \sum_{a\in\mathcal{A}_s} x(s,a) = \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}_{s'}} p(s|s',a')x(s',a') \quad \forall s \in \mathcal{S}$$

$$\sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} x(s,a) = 1$$

$$x(s,a) \geq 0 \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}_s \qquad .$$

The variables $\{x(s,a)\}_{s,a}$ form a probability distribution over the state-action pairs. The first set of constraints requires the rates of transiting into and out of each state s to be equal. The benchmark $\text{opt}(\mathsf{P})$ is justified by the following:

Theorem 1 ((Cheung, 2019b)). *For any non-anticipatory policy, it holds that* $\mathbb{E}[\min_k\{\bar{V}_{1:T,k}\}] \leq \text{opt}(\mathsf{P}) + 2D/T$.

[1]National University of Singapore. Correspondence to: Wang Chi Cheung <isecwc@nus.edu.sg>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

[1]Unknown T can be resolved by the doubling trick, see the full version of this work (Cheung, 2019b).

Literature Review. This work takes a different perspective from (Cheung, 2019a), who studied multi-objective optimization on online MDPs via target set objectives. MDP-MMF is an example of online learning problems with vectorial rewards and global aggregate objectives. They are studied in bandit settings (Agrawal & Devanur, 2014; Busa-Fekete et al., 2017; Berthet & Perchet, 2017), constrained bandit settings (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014). Online MDPs with scalar rewards are studied in (Jaksch et al., 2010; Fruit et al., 2018). A recent stream of works studied MDPs with vectorial rewards and global aggregate objectives (Achiam et al., 2017; Hazan et al., 2018; Tessler et al., 2019; Miryoosefi et al., 2019; Tarbouriech & Lazaric, 2019; Rosenberg & Mansour, 2019; Efroni et al., 2020; Brantley et al., 2020). While these works consider maximizing $g(\mathbb{E}[\bar{V}_{1:T}])$ for some concave g, we consider maximizing $g(\bar{V}_{1:T})$ for the concave function $g(w) = \min_k w_k$. Please see (Cheung, 2019b) for a full review.

2. Algorithm TMWU-UCRL2

We propose Algorithm TMWU-UCRL2, displayed in Algorithm 1. Let's start with a high level view. The algorithm runs in episodes. An episode $m \in \{1, 2, \dots\}$ starts at time $\tau(m)$ and ends at time $\tau(m+1) - 1$, and $\tau(1), \tau(2), \dots$ are determined adaptively.

At the start of episode m, the agent first compute confidence regions H_m^v, H_m^p for the latent v, p, using historical observation. Then, he scalarizes the vectorial rewards using the Multiplicative Weight Update (MWU, surveyed in (Arora et al., 2012)), which promotes max-min fairness. After that, he solves a latent MDP problem with scalarized rewards by the Extended Value Iteration (EVI) algorithm (Jaksch et al., 2010), which conducts optimistic exploration on the latent model, and outputs a stationary policy $\tilde{\pi}_m$.

The policy $\tilde{\pi}_m$ is run throughout the episode m, which ends when one of two stopping criteria is reached. These criteria are based on the change in the scalarization across time, and the frequency of visiting each state action pair.

Confidence regions $H_m^v = \{H_m^v(s,a)\}_{s,a}$, $H_m^p = \{H_m^p(s,a)\}_{s,a}$. We first define the count statistic $N_m^+(s,a) = \max\{1, N_m(s,a)\}$, where

$$N_m(s,a) = \sum_{t=1}^{\tau(m)-1} 1(s_t = s, a_t = a), \tag{2}$$

for each s, a. Define the estimates

$$\hat{v}_m(s,a) = \frac{\sum_{t=1}^{\tau(m)-1} V_t(s_t,a_t) 1(s_t = s, a_t = a)}{N_m^+(s,a)},$$

$$\hat{p}_m(s'|s,a) = \frac{\sum_{t=1}^{\tau(m)-1} 1(s_t = s, a_t = a, s_{t+1} = s')}{N_m^+(s,a)}.$$

Algorithm 1 TMWU-UCRL2

Input: Parameters $\delta \in (0,1)$, initial state s_1, horizon T, MWU rate $\eta = 1/T^{2/3}$.
Set $t = 1$.
for episode $m = 1, 2, \dots$ **do**
  Set $\tau(m) = t$
  Initialize $N_m(s,a)$ with Eq. (2) for all s, a.
  Define conf regions H_m^v, H_m^p with Eq. (3, 4).
  Compute the scalarized optimistic reward $\{\tilde{r}_m(s,a)\}_{s,a}$ with Eq. (6).
  Compute stationary policy $\tilde{\pi}_m$ with EVI.
  Set $\nu_m(s,a) = 0$ for all s, a.
  Set $\theta^{\mathsf{ref}} = \theta_{\tau(m)}$, and $\Psi = 0$.
  while $\Psi \le 1$ and $\nu_m(s_t, \tilde{\pi}_m(s_t)) < N_m^+(s_t, \tilde{\pi}_m(s_t))$ **do**
    Choose action $a_t = \tilde{\pi}_m(s_t)$.
    Observe s_{t+1} and $V_t(s_t, a_t)$.
    Compute weight vector θ_{t+1} based on (5).
    Update $\Psi \leftarrow \Psi + \|\theta_{t+1} - \theta^{\mathsf{ref}}\|_1$.
    Update $\nu_m(s_t, a_t) \leftarrow \nu_m(s_t, a_t) + 1$.
    Update $t \leftarrow t + 1$.
    if $t > T$ **then**
      Break.
    end if
  end while
end for

The confidence regions $H_m^v(s,a), H_m^p(s,a)$ are

$$H_m^v(s,a) = \big\{\bar{v} \in [0,1]^K : |\bar{v}_k - \hat{v}_{m,k}(s,a)| \le \mathsf{rad}_{m,k}^v(s,a)\ \forall 1 \le k \le K\big\}, \tag{3}$$

$$H_m^p(s,a) = \big\{\bar{p} \in \Delta^{\mathcal{S}} : |\bar{p}(s') - \hat{p}_m(s'|s,a)| \le \mathsf{rad}_m^p(s'|s,a)\ \forall s' \in \mathcal{S}\big\} \tag{4}$$

respectively, where confidence radii $\mathsf{rad}_{m,k}^v(s,a), \mathsf{rad}_m^p(s'|s,a)$ (Fruit et al., 2018) are defined in Appendix A.

Scalarization by MWU. For explanation sake, let's suppose that the mean reward vectors $\{v(s,a)\}_{s,a}$ are known. The scalarized reward for $v(s,a)$ is equal to $\theta_{\tau(m)}^\top v(s,a)$, where for time t we define

$$\theta_{t,k} = \frac{\exp\left[-\eta \sum_{q=1}^{t-1} V_{q,k}(s_q,a_q)\right]}{\sum_{\kappa=1}^K \exp\left[-\eta \sum_{q=1}^{t-1} V_{q,k}(s_q,a_q)\right]}, \tag{5}$$

and $\eta = 1/T^{2/3}$. The learning rate η is different from the conventional choice of $\Theta(1/T^{1/2})$ (Arora et al., 2012), due to our handling of the episode changes for the MDP. The choice of η is motivated in our subsequent analysis.

The scalarization promotes max-min fairness as follows. For two reward types k, k' with $\sum_{q=1}^{t-1} V_{q,k}(s_q,a_q) >$

$\sum_{q=1}^{t-1} V_{q,k'}(s_q, a_q)$, i.e. the amount of type k reward is larger than that of type k', we have $\theta_{t,k'} > \theta_{t,k}$, meaning that a higher weight is assigned to reward type k' than type k, hence maximizing the minimum.

Going back to the actual case of unknown $\{v(s,a)\}_{s,a}$, the agent conducts optimistic exploration by the *optimistic scalarized reward* $\tilde{r}_m(s,a)$ for each s, a:

$$\tilde{r}_m(s,a) = \max_{\bar{v}(s,a) \in H_m^v(s,a)} \theta_{\tau(m)}^\top \bar{v}(s,a). \tag{6}$$

EVI. First, we define $\mathsf{ave}(r,p)$ as the optimum average reward in an MDP with scalar rewards $r = \{r(s,a)\}_{s,a}$ and[2] transition probability $p = \{p(\cdot|s,a)\}_{s,a}$. By folklore, the optimum can be achieved by a stationary policy.

At the start of episode m, the EVI (Jaksch et al., 2010), which is fully displayed in Appendix B, inputs $\tilde{r}_m$, H_m^p. The EVI outputs a stationary policy $\tilde{\pi}_m$ that satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{q=1}^{T} \tilde{r}_m(s_q, \tilde{\pi}_m(s_q)) \geq \max_{\bar{p}} \{\mathsf{ave}(\tilde{r}_m, \bar{p})\} - \frac{1}{\sqrt{\tau(m)}},$$

where the maximum is over $\{\bar{p} : \bar{p}(s,a) \in H_m^p(s,a) \forall\, s,a\}$. The choice of $\tilde{\pi}_m$ leads to an optimistic exploration of the latent p in the confidence region H_m^p.

Stopping Criteria. The agent follows $\tilde{\pi}_m$ until the end of episode m, which is triggered by two criteria. One is when a state-action pair is visited sufficiently often, in the sense that $\nu_m(s_t, \tilde{\pi}_m(s_t)) \geq N_m^+(s_t, \tilde{\pi}_m(s_t))$. This criterion is due to (Jaksch et al., 2010).

Our novel contribution is another criterion, which we call the *thresholding criterion.* It is triggered when $\Psi \geq 1$, where Ψ measures the cumulative change on the scalarizations $\theta_{\tau(m)}, \theta_{\tau(m)+1}, \ldots$ during episode m. For a time index t during episode m, note that the scalarization is θ_t, while the agent is employing policy $\tilde{\pi}_m$ that is based on the scalarization $\theta_{\tau(m)}$. Essentially, the criterion is triggered when the cumulative discrepancy between θ_t , $\theta_{\tau(m)}$ (measured in the ℓ_1 norm) exceeds the threshold 1. The threshold 1 can be replaced by any absolute positive constant.

The thresholding criterion incorporates the balancing effect by the scalarization $\theta_{\tau(m)}$, while controls the frequency of switching among stationary policies. To illustrate the importance of the thresholding criterion, consider the instance in Fig 1. The figure depicts an instance with $K = 2$, and each arc represents a deterministic transition. For example, the arc from s^0 to s^1 represents the action a^{01} under which $p(s^1|s^0, a^{01}) = 1$. Likewise, the loop at s^1 represents the action a^{11} under which $p(s^1|s^1, a^{11}) = 1$. Let's first run TMWU-UCRL2 with the thresholding criterion silenced by replacing $\Psi \leq 1$ with $\Psi \leq 0$.

[2] $r(s,a)$ is a real number, whereas $v(s,a)$ which is a vector.

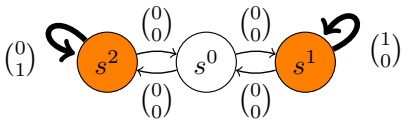

Figure 1. 3-state example

Assume v, p are known. Without loss of generality, suppose $s_1 = s^0$, and the agent always loops at s^2 when the scalarized rewards for the two loops are equal. When the thresholding criterion is silenced ($\Psi \leq 0$), the resulting trajectory is the recurrence of:

$$s^0 \xrightarrow{\binom{0}{0}} s^2 \xrightarrow{\binom{0}{1}} s^2 \xrightarrow{\binom{0}{0}} s^0 \xrightarrow{\binom{0}{0}} s^1 \xrightarrow{\binom{1}{0}} s^1 \xrightarrow{\binom{0}{0}} s^0 \ldots,$$

resulting in a $1/6 - O(1/T)$ max-min reward since $\bar{V}_{1:T} = \binom{1/6-O(1/T)}{1/6-O(1/T)}$. However, the optimal max-min reward is in fact $1/2 - 2/T$, simply by looping at s^2 for $T/2 - 1$ times then looping at s^1 for $T/2 - 1$ times . Altogether, silencing the thresholding criterion leads to $\mathrm{Reg} = \Omega(1)$.

By contrast, with the thresholding criterion, the agent loops at s^1 for $\Theta(t^{1/3})$ times when he moves from s^0 to s^1 at time t, and likewise for s^2. The resulting max-min reward is $1/2 - O(1/T^{1/3})$, which tends to the optimum as T grows.

3. Main Results and Analysis

Define $S = |\mathcal{S}|$, $A = \frac{1}{S}\sum_{s\in\mathcal{S}} |\mathcal{A}_s|$, and $\Gamma = \max_{s,a} \sum_{s'\in\mathcal{S}} \mathbf{1}(p(s'|s,a) > 0)$, which is at most S.

Theorem 2. *Algorihtm* TMWU-UCRL2 *satisfies*

$$\mathit{Reg} = \tilde{O}\left(\frac{D}{T^{1/3}} + \frac{D\sqrt{\Gamma S A}}{\sqrt{T}}\right)$$

with probability $1 - O(\delta)$*, where* $\tilde{O}(\cdot)$ *notation hides multiplicative factors logarithmic in* $S, A, K, T, 1/\delta$.

In the regret bound, the first term accounts for the error due to the thresholding criterion, while the second term accounts for the error incurred in learning the latent model. The second term matches the regret bound for online MDPs with scalarized rewards in (Jaksch et al., 2010). To prove Theorem 2, denote an optimal solution of (P) as x^*, and the corresponding vector of average rewards is $v^* = (v_k^*)_{k=1}^K = \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} v(s,a)x^*(s,a)$. Clearly, $\mathrm{opt}(\mathsf{P}) = \min_k v_k^*$.

$$\begin{aligned}
\min_k \bar{V}_{1:T,k} &\geq \frac{1}{T}\sum_{t=1}^{T} \theta_t^\top V_t(s_t,a_t) - \left(\eta + \frac{\log K}{\eta T}\right) \\
&\overset{(\dagger)}{=} \frac{1}{T}\sum_{t=1}^{T} \theta_t^\top v^* - \frac{1}{T}\sum_{t=1}^{T} \theta_t^\top [v^* - V_t(s_t,a_t)] - \left(\eta + \frac{\log K}{\eta T}\right) \\
&\overset{(\ddagger)}{\geq} \frac{1}{T}\sum_{t=1}^{T} \min_k\{v_k^*\} - \frac{1}{T}\sum_{t=1}^{T} \theta_t^\top [v^* - V_t(s_t,a_t)] - \left(\eta + \frac{\log K}{\eta T}\right)
\end{aligned}$$

$$=\mathsf{opt}(\mathsf{P}) - \frac{1}{T}\sum_{t=1}^{T}\theta_t^\top[v^* - V_t(s_t,a_t)] - \left(\eta + \frac{\log K}{\eta T}\right). \tag{7}$$

Equality (†) is by the classical result on MWU, extracted from Theorem 2.3 in the survey (Arora et al., 2012).

Theorem 3 ((Arora et al., 2012)). *Let $W_1,\ldots,W_T$ be an arbitrary sequence of vectors, where $W_t = (W_{t,k})_{k=1}^K \in [0,1]^K$ for each t, and let $\eta \in (0,1]$. Consider the sequence of vectors $\vartheta_1,\ldots,\vartheta_T$, where $\vartheta_t = (\vartheta_{t,k})_{k=1}^K \in [0,1]^K$ is defined as $\vartheta_{t,k} = \frac{\exp\left[-\eta\sum_{q=1}^{t-1} W_{q,k}\right]}{\sum_{\kappa=1}^K \exp\left[-\eta\sum_{q=1}^{t-1} W_{q,k}\right]}$ for each t,k (we define $\sum_{q=1}^0 W_{q,k} = 0$). Then*

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_{t,k}W_{t,k} \le \min_k\left\{\frac{1}{T}\sum_{t=1}^{T}W_{t,k}\right\} + \eta + \frac{\log K}{\eta T}.$$

We apply the Theorem with $W_t = V_t(s_t,a_t)$, and $\vartheta_t = \theta_t$ in the algorithm. Inequality (‡) is by the fact that θ_t forms a probability distribution on $\{1,\ldots,K\}$ for each t. Altogether, we have the regret bound

$$\mathrm{Reg} \le \frac{1}{T}\sum_{t=1}^{T}\theta_t^\top[v^* - V_t(s_t,a_t)] + \eta + \frac{\log K}{\eta T}.$$

Theorem 2 is proved by combining (7) with Lemma 1 and Proposition 1 provided in the following. For time t, denote $m(t)$ as the index of the episode containing t, and thus $m(T)$ is the number of episodes during the horizon. Lemma 1 provides a deterministic upper bound $M(T)$ on $m(T)$:

Lemma 1. *With certainty, it holds that*

$$m(T) \le M(T) = \sqrt{2\eta}T + SA(1+\log_2 T).$$

Proposition 1 bounds the average $\frac{1}{T}\sum_{t=1}^{T}\theta_t^\top[v^* - V_t(s_t,a_t)]$ in terms of $M(T)$.

Proposition 1. *With probability at least $1-O(\delta)$ we have*

$$\sum_{t=1}^{T}\theta_t^\top[v^* - V_t(s_t,a_t)] = \tilde{O}\left(D\cdot M(T) + D\sqrt{\Gamma SAT}\right).$$

The Proposition is proved in Appendix C. The first term accounts for the lag between $\theta_{\tau(m)}$ and θ_t for each t in each episode m, while the second term accounts for the learning error on v,p. Altogether, the regret bound is

$$\mathrm{Reg} = \tilde{O}\left(D\sqrt{\eta} + \frac{D\sqrt{\Gamma SA}}{\sqrt{T}} + \eta + \frac{1}{\eta T}\right),$$

which explains the choice of η and proves the Theorem.

Proof of Lemma 1. We first partition the index set of episode indexes $\{1,\ldots,M(T)\}$ into two sets:

$$\mathcal{M}_\Psi = \{m\text{: episode } m+1 \text{ started by } \Psi \ge 1\},$$
$$\mathcal{M}_\nu = \{m\text{: episode } m+1 \text{ started by } \nu_m(s_t,\tilde{\pi}_m(s_t)) \ge N_m^+(s_t,\tilde{\pi}_m(s_t)) \text{ for some } t \ge \tau(m)\}.$$

Define $n_\Psi := |\mathcal{M}_\Psi|, n_\nu := |\mathcal{M}_\nu|$, so that $M(T) = n_\Psi + n_\nu$. Following (Jaksch et al., 2010), we can show that $n_\nu \le SA(1+\log_2 T)$ with certainty. We next assert

$$n_\Psi \le \sqrt{2\eta}T.$$

Let's express $\mathcal{M}_\Psi = \{m_1, m_2, \ldots, m_{n_\Psi}\}$, where $m_1 < m_2 < \ldots < m_{n_\Psi}$. Now, observe that $\tau(m_{j+1}) > \tau(m_j) + 1/\sqrt{2\eta}$. Indeed,

$$\sum_{q=\tau(m_j)}^{\tau(m_j)+\lfloor\frac{1}{\sqrt{2\eta}}\rfloor}\left\|\theta_q - \theta_{\tau(m_j)}\right\|_1 \le 2\eta\sum_{q=\tau(m_j)}^{\tau(m_j)+\lfloor\frac{1}{\sqrt{2\eta}}\rfloor} q - \tau(m_j)$$
$$\le 2\eta\frac{(1/\sqrt{2\eta})(1/\sqrt{2\eta}+1)}{2} \le 1. \tag{8}$$

The first inequality is by the claim that, for time indexes $t' < t$ such that $\eta(t-t') \le 1$, we have

$$\|\theta_{t'} - \theta_t\|_1 = \sum_{k=1}^{K}|\theta_{t',k} - \theta_{t,k}| \le 2\eta(t-t'). \tag{9}$$

We postpone the proof of (9) to the end. Now, eq. (8) implies that $\tau(m_{j+1}) > \tau(m_j) + 1/\sqrt{2\eta}$ for each j. Therefore, by the pigeonhole's principle, the count n_Ψ is at most $T/(1/\sqrt{2\eta}) = \sqrt{2\eta}T$. Finally, we return to the proof of (9). For each k, consider two cases. If $\theta_{t',k} < \theta_{t,k}$, then

$$|\theta_{t',k} - \theta_{t,k}| = \theta_{t,k} - \theta_{t',k}$$
$$= \frac{\theta_{t',k}\exp[-\eta\sum_{q=t'}^{t-1}V_{q,k}(s_q,a_q)]}{\sum_{\kappa=1}^K\theta_{t',\kappa}\exp[-\eta\sum_{q=t'}^{t-1}V_{q,\kappa}(s_q,a_q)]} - \theta_{t',k}$$
$$\le \frac{\theta_{t',k}}{\left(\sum_{\kappa=1}^K\theta_{t',\kappa}\right)\exp[-\eta(t-t')]} - \theta_{t',k} \tag{10}$$
$$= \theta_{t',k}\left(\exp[\eta(t-t')] - 1\right) \le 2\theta_{t',k}\eta(t-t'). \tag{11}$$

Step (10) is by $\exp[-\eta(t-t')] \le \exp[-\eta\sum_{q=t'}^{t-1}V_{q,k}(s_q,a_q)] \le 1$ for each k. Step (11) is because $\exp(a) \le 1+2a$ for $a \in [0,1]$. Otherwise, if $\theta_{t',k} \ge \theta_{t,k}$, then

$$|\theta_{t',k} - \theta_{t,k}| = \theta_{t',k} - \theta_{t,k}$$
$$= \theta_{t',k} - \frac{\theta_{t',k}\exp[-\eta\sum_{q=t'}^{t-1}V_{q,k}(s_q,a_q)]}{\sum_{\kappa=1}^K\theta_{t',\kappa}\exp[-\eta\sum_{q=t'}^{t-1}V_{q,\kappa}(s_q,a_q)]}$$
$$\le \theta_{t',k} - \frac{\theta_{t',k}\exp[-\eta(t-t')]}{\sum_{\kappa=1}^K\theta_{t',\kappa}}$$
$$= \theta_{t',k}\left(1 - \exp[-\eta(t-t')] - 1\right) \le \theta_{t',k}\eta(t-t').$$

Thus, the equation (9) is shown since $\sum_k \theta_{t',k} = 1$. □

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 22–31. JMLR.org, 2017.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *ACM Conference on Economics and Computation*, 2014.

Altman, E. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Arora, S., Hazan, E., and Kale, S. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

Audibert, J., Munos, R., and Szepesvári, C. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902, 2009.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, pp. 207–216. IEEE Computer Society, 2013. ISBN 978-0-7695-5135-7.

Berthet, Q. and Perchet, V. Fast rates for bandit optimization with upper-confidence frank-wolfe. In *Advances in Neural Information Processing Systems 30*, pp. 2225–2234. 2017.

Brantley, K., Dudík, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., and Sun, W. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *CoRR*, abs/2006.05051, 2020. URL https://arxiv.org/abs/2006.05051.

Busa-Fekete, R., Szörényi, B., Weng, P., and Mannor, S. Multi-objective bandits: Optimizing the generalized Gini index. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 625–634. PMLR, 2017.

Cheung, W. C. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pp. 724–734, 2019a.

Cheung, W. C. Exploration-exploitation trade-off in reinforcement learning on online markov decision processes with global concave rewards. *CoRR*, abs/1905.06466, 2019b. URL http://arxiv.org/abs/1905.06466.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained mdps. *CoRR*, abs/2003.02189, 2020. URL https://arxiv.org/abs/2003.02189.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018.

Hazan, E., Kakade, S. M., Singh, K., and Soest, A. V. Provably efficient maximum entropy exploration. *CoRR*, 2018. URL http://arxiv.org/abs/1812.02690.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010. ISSN 1532-4435.

Miryoosefi, S., Brantley, K., Daume III, H., Dudik, M., and Schapire, R. E. Reinforcement learning with convex constraints. In *Advances in Neural Information Processing Systems 32*, pp. 14093–14102. Curran Associates, Inc., 2019.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019.

Tarbouriech, J. and Lazaric, A. Active exploration in markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 974–982. PMLR, 16–18 Apr 2019.

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

A. Definitions of $\mathsf{rad}^v_{m,k}(s,a)$, $\mathsf{rad}^p_m(s'|s,a)$

We define $(\log\text{-}v)_m := \log(12KSA\tau^2(m)/\delta)$, $(\log\text{-}p)_m := \log(12S^2A\tau^2(m)/\delta)$. The radii are defined as

$$\mathsf{rad}^v_{m,k}(s,a) := \sqrt{\frac{2\hat{v}_{m,k}(s,a)\cdot(\log\text{-}v)_m}{N^+_m(s,a)}} + \frac{3\cdot(\log\text{-}v)_m}{N^+_m(s,a)},$$

$$\mathsf{rad}^p_m(s'|s,a) := \sqrt{\frac{2\hat{p}_m(s'|s,a)\cdot(\log\text{-}p)_m}{N^+_m(s,a)}} + \frac{3\cdot(\log\text{-}p)_m}{N^+_m(s,a)}.$$

B. EVI (Jaksch et al., 2010)

The Algorithm is displayed in Algorithm 2. At the start of each episode m, the EVI inputs rewards $\tilde{r}_m$, conf region H^p_m, tolerance parameter $1/\sqrt{\tau(m)}$, and outputs the stationary policy $\tilde{\pi}_m$, and the dual variables $(\tilde{\phi}_m, \tilde{\gamma}_m)$.

Algorithm 2 EVI, extracted from (Jaksch et al., 2010)

Input: rewards $\tilde{r}$, conf region H^p, tolerance parameter ϵ.
Initialize VI record $u_0 \in \mathbb{R}^{\mathcal{S}}$ as $u_0(s) = 0$ for all $s \in \mathcal{S}$.
for $i = 0, 1, \ldots$ **do**
  For each $s \in \mathcal{S}$, compute VI record

$$u_{i+1}(s) = \max_{a\in\mathcal{A}_s} \tilde{\Upsilon}_i(s,a), \text{ where } \tilde{\Upsilon}_i(s,a) = \tilde{r}(s,a) + \max_{\bar{p}\in H^p(s,a)} \left\{ \sum_{s'\in\mathcal{S}} u_i(s')\bar{p}(s') \right\}.$$

  if $\max_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\} - \min_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\} \le \epsilon$ **then**
    Break the **for** loop.
  end if
end for
Define stationary policy $\tilde{\pi} : \mathcal{S} \to \mathcal{A}_s$ as $\tilde{\pi}(s) = \text{argmax}_{a\in\mathcal{A}_s} \tilde{\Upsilon}_i(s,a)$.
Define an optimistic dual solution $\tilde{\phi} = \max_{s\in\mathcal{S}}\{u_{i+1}(s) - u_i(s)\}$, $\tilde{\gamma} = u_i$.
Return policy $\tilde{\pi}$ and dual variables $(\tilde{\phi}, \tilde{\gamma})$.

C. Proof of Proposition 1

The Proposition can be proved by proving a series of Lemmas, Lemmas 2 – 7. Lemma 2 justifies the confidence regions H^v_m, H^p_m.

Lemma 2. *Consider events $\mathcal{E}^v, \mathcal{E}^p$, which quantify the accuracy in estimating v, p:*

$$\mathcal{E}^v := \{v(s,a) \in H^v_m(s,a) \text{ for all } m \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A}_s\}, \tag{12}$$

$$\mathcal{E}^p := \{p(\cdot|s,a) \in H^p_m(s,a) \text{ for all } m \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A}_s\}. \tag{13}$$

It holds that $\mathbb{P}[\mathcal{E}^v] \ge 1 - \delta/2, \mathbb{P}[\mathcal{E}^p] \ge 1 - \delta/2$.

Lemma 3 decomposes the average $\frac{1}{T}\sum_{t=1}^T \theta_t^\top [v^* - V_t(s_t,a_t)]$ into 5 parts ($\clubsuit, \diamondsuit, \heartsuit, \spadesuit, \P$):

Lemma 3. *Let t be a time index, and let m be its episode index: $\tau(m) \le t < \tau(m+1)$. Conditional on events $\mathcal{E}^v, \mathcal{E}^p$, the following inequality holds:*

$$\theta_t^\top [v^* - V_t(s_t,a_t)] \le (\clubsuit_t) + (\diamondsuit_t) + (\heartsuit_t) + (\spadesuit_t) + (\P_t),$$

where

$$(\clubsuit_t) := \left[\theta_{\tau(m)} - \theta_t\right]^\top V_t(s_t,a_t)$$

$$(\diamondsuit_t) := \tilde{r}_m(s_t, a_t) - \theta_{\tau(m)}^\top V_t(s_t, a_t),$$
$$(\heartsuit_t) := \left[\theta_t - \theta_{\tau(m)}\right]^\top v^*$$
$$(\spadesuit_t) := \max_{\bar{p} \in H_m^p(s_t, a_t)} \left\{ \sum_{s' \in \mathcal{S}} \tilde{\gamma}_m(s') \bar{p}(s') \right\} - \tilde{\gamma}_m(s_t),$$
$$(\P_t) := 1/\sqrt{\tau(m)}.$$

Finally, Lemmas 4 – 7 bound the 5 parts. The proofs of Lemmas 5 – 7 largely follow (Jaksch et al., 2010).

Lemma 4. *With certainty,* $\sum_{t=1}^T (\clubsuit_t) \le M(T), \qquad \sum_{t=1}^T (\heartsuit_t) \le M(T).$

Lemma 5. *With certainty,* $\sum_{t=1}^T (\P_t) \le \left(\sqrt{2}+1\right)\sqrt{T}.$

Lemma 6. *Conditional on $\mathcal{E}^v$, with probability at least $1-\delta$ we have* $\sum_{t=1}^T (\diamondsuit_t) = \tilde{O}\left(\sqrt{SAT} + SA\right).$

Lemma 7. *Conditional on $\mathcal{E}^p$, with probability at least $1-\delta$ we have*

$$\sum_{t=1}^T (\spadesuit_t) = \tilde{O}\left(D \cdot M(T)\right) + \tilde{O}\left(D\sqrt{\Gamma SAT} + DS^2A\right).$$

To prepare for the proofs of these Lemmas, we first provide the auxiliary results needed in Section C.1. Then, we prove Lemma 2 in Section C.2. After that, we prove Lemma 3 in Section C.3. Finally, we Lemmas 4, 5, 6, 7 in Sections C.4, C.5, C.6, C.7 respectively.

C.1. Auxiliary Results

Theorem 4 ((Audibert et al., 2009))**.** *Let random variables $Y_1, \ldots, Y_N \in [0,1]$ be independently and identically distributed. Consider their sample mean $\hat{Y}_N$ and their sample variance $\hat{\sigma}^2_{Y,N}$:*

$$\hat{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i, \quad \hat{\sigma}^2_{Y,N} = \frac{1}{N}\sum_{i=1}^N (Y_i - \hat{Y})^2.$$

For any $\delta \in (0,1)$, the following inequality holds:

$$\Pr\left(\left|\hat{Y}_N - \mathbb{E}[Y_1]\right| \le \sqrt{\frac{2\hat{\sigma}^2_{Y,N}\log(1/\delta)}{N}} + \frac{3\log(1/\delta)}{N}\right) \ge 1 - 3\delta.$$

Theorem 5 ((Hoeffding, 1963))**.** *Let random variables $X_1, \ldots, X_T$ constitute a martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_t\}_{t=1}^T$, that is, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$ for all $1 \le t \le T$. Also, suppose that $|X_t| \le B$ almost surely for all t. Then the following inequality holds for any $0 < \delta < 1$:*

$$\Pr\left[\frac{1}{T}\sum_{t=1}^T X_t \le B\sqrt{\frac{2\log(1/\delta)}{T}}\right] \ge 1 - \delta.$$

Next, we present auxiliary results, mostly from (Jaksch et al., 2010). Theorem 6 is useful for analyzing the EVI oracle EVI. Lemmas 8, 9 and Lemma 10 are useful for proving the convergence of TMWU-UCRL2.

Theorem 6 ((Jaksch et al., 2010))**.** *Consider applying EVI (Algorithm 2) with input $(\tilde{r}, H^p; \epsilon)$, where the underlying transition kernel p of lies in H^p, and the underlying instance is communicating with diameter D. Then (i) EVI$(\tilde{r}, H^p; \epsilon)$ terminates in finite time, (ii) the output dual variables $\tilde{\gamma}$ satisfies $\max_{s\in\mathcal{S}} \tilde{\gamma}_s - \min_{s\in\mathcal{S}} \tilde{\gamma}_s \le D \cdot \max_{s,\in\mathcal{S}, a\in\mathcal{A}} |\tilde{r}(s,a)|$.*

Lemma 8 (Lemma 19 in (Jaksch et al., 2010))**.** *For any sequence of numbers $z_1, \ldots, z_n$ with $0 \le z_m \le Z_{m-1} := \max\{1, \sum_{i=1}^{m-1} z_i\}$, we have*

$$\sum_{m=1}^n \frac{z_m}{\sqrt{Z_{m-1}}} \le \left(\sqrt{2}+1\right)\sqrt{Z_n}.$$

Lemma 9 ((Jaksch et al., 2010)). *The following inequality holds with certainty:*

$$\sum_{t=1}^{T} \frac{1}{\sqrt{N^{+}_{m(t)}(s_t, a_t)}} \leq \left(\sqrt{2}+1\right)\sqrt{SAT}.$$

Lemma 10. *The following inequality holds with certainty:*

$$\sum_{t=1}^{T} \frac{1}{N^{+}_{m(t)}(s_t, a_t)} \leq SA\left(1+2\log T\right).$$

Proof of Lemma 10. To start the proof, first denote $\nu'_{m(T)}(s,a) = \sum_{t=\tau(m(T))}^{T} \mathbf{1}((s_t,a_t)=(s,a))$. Essentially $\nu'_{m(T)}(s,a)$ is $\nu_{m(T)}(s,a)$ capped at the end of time step T. In addition, denote $N^{+'}_{m(T)+1}(s,a) = \sum_{t=1}^{T} \mathbf{1}((s_t,a_t)=(s,a))$. Similar to $\nu'_{m(T)}(s,a)$, $N^{+'}_{m(T)+1}(s,a)$ denotes the version of $N^{+}_{m(T)+1}(s,a)$ capped at the end of time step T. Now,

$$\sum_{t=1}^{T} \frac{1}{N^{+}_{m(t)}(s_t, a_t)} = \sum_{m=1}^{m(T)-1} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \frac{\nu_m(s,a)}{N^{+}_{m}(s,a)} + \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \frac{\nu'_{m(T)}(s,a)}{N^{+}_{m(T)}(s,a)}.$$

Now, for every state-action pair s, a, we assert that

$$\sum_{m=1}^{m(T)-1} \frac{\nu_m(s,a)}{N^{+}_{m}(s,a)} + \frac{\nu'_{m(T)}(s,a)}{N^{+}_{m(T)}(s,a)} \leq 1 + 2\log\left(N^{+'}_{m(T)+1}(s,a)\right). \tag{14}$$

Indeed, the asserted inequality can be proved by drawing the following general fact: For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_m \leq Z_{m-1} := \max\{1, \sum_{i=1}^{m-1} z_i\}$, we have

$$\sum_{m=1}^{n} \frac{z_m}{Z_{m-1}} \leq 1 + 2\log Z_n. \tag{15}$$

We prove the inequality (15) by induction on n. The case for $n = 1$ is clearly true. Now, suppose the inequality is true for n. Then it is also true for $n+1$, since

$$\sum_{m=1}^{n+1} \frac{z_m}{Z_{m-1}} \leq 1 + 2\log Z_n + \frac{z_{n+1}}{Z_n} \leq 1 + 2\log Z_n + 2\log\left(1+\frac{z_{n+1}}{Z_n}\right) = 1 + 2\log Z_{n+1},$$

where we use the fact that $x \leq 2\log(1+x)$ for $x \in [0,1]$. Hence, the induction is established and the (15) is proved for general n.

Given (15) for general n, we can readily establish (14) by applying (15) with $n = m(T)$, $z_m = \nu_m(s,a)$ for $1 \leq m \leq n-1$ and $z_n = \nu'_n(s,a)$. Altogether, noting that $N^{+'}_{m(T)+1}(s,a) \leq T$, we achieve the required inequality. □

C.2. Proof of Lemma 2

We first analyze event $\mathcal{E}^v$. Consider a fixed objective index k, a fixed state s and a fixed action a. We assert that

$$\mathbb{P}\left[|\hat{v}_{m,k}(s,a) - v_k(s,a)| \leq \mathsf{rad}^{v}_{m,k}(s,a) \text{ for all } m\right] \geq 1 - \frac{\delta}{2KSA}. \tag{16}$$

Assuming inequality (16), the bound $\Pr[\mathcal{E}^v] \geq 1 - \delta/2$ is established by taking a union bound over $s \in \mathcal{S}$, $a \in \mathcal{A}_s$ and $k \in [K]$.

We establish inequality (16) by applying Theorem 4 and the union bound. First, note that $\hat{v}_{m,k}(s,a)$ is the sample mean of $N_m(s,a)$ i.i.d. random variables, which are distributed as $V_k(s,a)$. Let $\hat{\sigma}^2_{v,m,k}$ be the sample variance of these $N_m(s,a)$ i.i.d random variables. To apply the union bounds, we also consider $\Upsilon^V_1, \ldots, \Upsilon^V_T$, which are T i.i.d samples with the

same distribution as $V_k(s,a)$. Denote $\hat{\Upsilon}^V_t$, $\hat{\sigma}^2_{\Upsilon^V,t}$ respectively as the sample mean and variance of $\Upsilon^V_1,\ldots,\Upsilon^V_t$. Let $\delta^v(t)=\delta/(12KSAt^2)$. Now,

$$\begin{aligned}
&\mathbb{P}\left[|\hat{v}_{m,k}(s,a)-v_k(s,a)|\le\sqrt{\frac{2\hat{\sigma}^2_{v,m,k}\log(1/\delta^v(N^+_m(s,a)))}{N^+_m(s,a)}}+\frac{3\log(1/\delta^v(N^+_m(s,a)))}{N^+_m(s,a)}\forall\, m\right]\\
\ge&\mathbb{P}\left[\left|\hat{\Upsilon}^V_t-v_k(s,a)\right|\le\sqrt{\frac{2\hat{\sigma}^2_{\Upsilon^V,t}\log(1/\delta^v(t))}{t}}+\frac{3\log(1/\delta^v(t))}{t}\text{ for all } t\in[T]\right] \qquad (17)\\
\ge&1-3\sum_{t=1}^{T}\delta^v(t)=1-\frac{\delta}{4KSA}\sum_{t=1}^{T}\frac{1}{t^2}\ge 1-\frac{\delta}{2KSA}. \qquad (18)
\end{aligned}$$

Step (17) is by applying a union bound over all possible values of $N^+_m(s,a)$s. Step (18) is by applying Theorem 4. Finally, note that $\hat{\sigma}^2_{v,m,k}\le\hat{v}_{m,k}(s,a)$, since $V(s_t,a_t)\in[0,1]$. Putting in the definition of $\delta^v(t)$ yields

$$\mathsf{rad}^v_{m,k}(s,a)\ge\sqrt{\frac{2\hat{\sigma}^2_{v,m,k}\log(1/\delta^v(N^+_m(s,a)))}{N^+_m(s,a)}}+\frac{3\log(1/\delta^v(N^+_m(s,a)))}{N^+_m(s,a)}.$$

Altogether, the required inequality for $\mathcal{E}^v$ is shown.

Next, we analyze the event $\mathcal{E}^p$ by in a similar way. Consider fixed states $s',s\in\mathcal{S}$ and a fixed action $a\in\mathcal{A}_s$. We assert that

$$\mathbb{P}\left[|\hat{p}_m(s'|s,a)-p(s'|s,a)|\le\mathsf{rad}^p_m(s'|s,a)\text{ for all } m\right]\ge 1-\frac{\delta}{2S^2A}. \qquad (19)$$

Assuming inequality (19), the bound $\Pr[\mathcal{E}^p]\ge 1-\delta/2$ is established by taking a union bound over $s',s\in\mathcal{S}$ and $a\in\mathcal{A}_s$. Let $\Upsilon^p_1,\ldots,\Upsilon^p_T$ be T i.i.d. Bernoulli random variables with the common mean $p(s'|s,a)$. For each $t\in[T]$, denote $\hat{\Upsilon}^p_t,\hat{\sigma}^2_{\Upsilon^p,t}$ respectively as the sample mean and sample variance of $\Upsilon^p_1,\ldots\Upsilon^p_t$. In addition, let $\delta^p(t)=\delta/(12S^2At^2)$. We have

$$\begin{aligned}
&\mathbb{P}\left[|\hat{p}_m(s'|s,a)-p(s'|s,a)|\le\mathsf{rad}^p_m(s'|s,a)\text{ for all } m\right]\\
\ge&\mathbb{P}\left[\left|\hat{\Upsilon}^p_t-p(s'|s,a)\right|\le\sqrt{\frac{2\hat{\sigma}^2_{\Upsilon^p,t}\log(1/\delta^p(t))}{t}}+\frac{3\log(1/\delta^p(t))}{t}\text{ for all } t\in[T]\right]\\
\ge&1-3\sum_{t=1}^{T}\delta^p(t)=1-\frac{\delta}{4S^2A}\sum_{t=1}^{T}\frac{1}{t^2}\ge 1-\frac{\delta}{2S^2A}.
\end{aligned}$$

Hence, the Lemma is proved.

C.3. Proof of Lemma 3

First, by the definitions of $(\clubsuit_t),(\diamondsuit_t)$, it is clear that

$$\theta_t^\top V_t(s_t,a_t)=\tilde{r}_m(s_t,a_t)-\left[(\clubsuit_t)+(\diamondsuit_t)\right].$$

Thus, it suffices to show that

$$\tilde{r}_m(s_t,a_t)\ge\theta_t^\top v^*-\left[(\heartsuit_t)+(\spadesuit_t)+(\P_t)\right]. \qquad (20)$$

To prove (20), we first focus on the application of the EVI oracle for episode m. By Assumption 1 and by assuming the event $\mathcal{E}^p$, we know that the oracle terminates in finite time, by virtue of item (i) in Theorem 6. Thus, the output policy $\tilde{\pi}_m$ and the output dual variables $(\tilde{\phi}_m,\tilde{\gamma}_m)$ are well-defined. Now, we assert that

$$\tilde{r}_m(s_t,a_t)\ge\tilde{\phi}_m-\left[(\spadesuit_t)+(\P_t)\right]. \qquad (21)$$

To show (21), we let $\tilde{u}_{\iota+1}, \tilde{u}_{\iota} \in \mathbb{R}^{\mathcal{S}}$ respectively be the terminating and the penultimate VI records, when EVI$(\tilde{r}_m, H_p^m, 1/\sqrt{\tau(m)})$ is applied. Now, we have

$$\begin{aligned}
\tilde{\phi}_m - (\P_t) &= \max_{s\in\mathcal{S}} \{\tilde{u}_{\iota+1}(s) - \tilde{u}_{\iota}(s)\} - \frac{1}{\sqrt{\tau(m)}} && (22)\\
&\le \min_{s\in\mathcal{S}} \{\tilde{u}_{\iota+1}(s) - \tilde{u}_{\iota}(s)\} && (23)\\
&\le \tilde{u}_{\iota+1}(s_t) - \tilde{u}_{\iota}(s_t)\\
&= \max_{a\in\mathcal{A}_{s_t}} \left\{ \tilde{r}_m(s_t,a) + \max_{\bar{p}\in H_m^p(s_t,a)} \left\{ \sum_{s'\in\mathcal{S}} \tilde{u}_{\iota}(s')\bar{p}(s') \right\} \right\} - \tilde{u}_{\iota}(s_t)\\
&= \tilde{r}_m(s_t,a_t) + \max_{\bar{p}\in H_m^p(s_t,a_t)} \left\{ \sum_{s'\in\mathcal{S}} \tilde{u}_{\iota}(s')\bar{p}(s') \right\} - \tilde{u}_{\iota}(s_t) && (24)\\
&= \tilde{r}_m(s_t,a_t) + (\spadesuit_t), && (25)
\end{aligned}$$

where step (22) is by the definition of $\tilde{\phi}_m$, step (23) is by the terminating condition of EVI, and step (24) is by the definition of $\tilde{\pi}_m$, and step (25) is by the definition of $\tilde{\gamma}_m$.

In order to prove the inequality (20) and complete the proof of the Lemma, it suffices to show

$$\tilde{\phi}_m \ge \theta_t^\top v^* - (\heartsuit_t) = \theta_t^\top \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} v(s,a)x^*(s,a) - (\heartsuit_t). \tag{26}$$

To this end, we first claim that the output dual variables $(\tilde{\phi}_m, \tilde{\gamma}_m)$ are feasible to the following linear program (lin-D$_m$):

$$\begin{aligned}
(\text{lin-D}_m)\text{: } \min\ & \phi\\
\text{s.t. } & \phi + \gamma(s) \ge \tilde{r}_m(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)\gamma(s') \quad \forall s\in\mathcal{S}, a\in\mathcal{A}_s\\
& \phi, \gamma(s) \text{ free} \qquad\qquad \forall s\in\mathcal{S}.
\end{aligned}$$

Indeed, for any $s\in\mathcal{S}$, $a\in\mathcal{A}_s$, we have

$$\begin{aligned}
\tilde{\phi}_m + \tilde{\gamma}_m(s) &\ge \tilde{u}_{\iota+1}(s) - \tilde{u}_{\iota}(s) + \tilde{u}_{\iota}(s) = \tilde{u}_{\iota+1}(s)\\
&\ge \tilde{r}_m(s,a) + \max_{\bar{p}\in H_m^p(s,a)} \left\{ \sum_{s'\in\mathcal{S}} \tilde{u}_{\iota}(s')\bar{p}(s') \right\} && (28)\\
&\ge \tilde{r}_m(s,a) + \sum_{s'\in\mathcal{S}} \tilde{u}_{\iota}(s')p(s'|s,a) = \tilde{r}_m(s,a) + \sum_{s'\in\mathcal{S}} \tilde{\gamma}_m(s')p(s'|s,a),
\end{aligned}$$

where step (28) is by the assumption that $p\in H_m^p$, since we condition on the event $\mathcal{E}^p$. Therefore, we have $\tilde{\phi}_m \ge$ opt(lin-D$_m$) = opt(lin-P$_m$), where the linear program

$$\begin{aligned}
(\text{lin-P}_m)\text{: } \max\ & \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} \tilde{r}_m(s,a)x(s,a)\\
\text{s.t. } & \sum_{a\in\mathcal{A}_s} x(s,a) = \sum_{s'\in\mathcal{S},a'\in\mathcal{A}_{s'}} P(s|s',a')x(s',a') \quad \forall s\in\mathcal{S}\\
& \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} x(s,a) = 1\\
& x(s,a) \ge 0 \qquad\qquad \forall s\in\mathcal{S}, a\in\mathcal{A}_s
\end{aligned}$$

is a dual formulation of (lin-D$_m$). The optimal solution x^* of the offline benchmark problem (P) is feasible to the problem (lin-P$_m$), since both (P), (lin-P$_m$) have the same feasible region.

Finally, we prove the inequality (26), and hence completing the proof of the Lemma. In the following derivation, we denote $\tilde{v}_m(s,a)$ as an optimal solution to the optimization problem (6) for computing the optimistic reward $\tilde{r}_m(s,a)$:

$$\begin{aligned}
\tilde{\phi}_m &\geq \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} \tilde{r}_m(s,a)x^*(s,a) \\
&= \theta_{\tau(m)}^\top \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} \tilde{v}_m(s,a)x^*(s,a) \\
&= \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} x^*(s,a)\left[\theta_{\tau(m)}^\top \tilde{v}_m(s,a) - \theta_{\tau(m)}^\top v(s,a)\right] \\
&\quad + \left[-\theta_t + \theta_{\tau(m)}\right]^\top \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} v(s,a)x^*(s,a) + \theta_t^\top \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} v(s,a)x^*(s,a) \qquad (30) \\
&\geq -(\heartsuit_t) + \theta_t^\top v^*,
\end{aligned}$$

where step (30) holds, since we condition on the event $\mathcal{E}^v$, which ensures that $\theta_{\tau(m)}^\top \tilde{v}_m(s,a) - \theta_{\tau(m)}^\top v(s,a) \geq 0$ for each $s\in\mathcal{S}, a\in\mathcal{A}_s$. Therefore, the first sum in (30) is non-negative, hence the step is justified. Altogether, inequality (26) is shown, and the Lemma is proved.

C.4. Proof of Lemma 4, which bounds ($\clubsuit$, $\heartsuit$)

Now,

$$\begin{aligned}
\sum_{t=1}^{T}(\clubsuit_t) &= \sum_{m=1}^{m(T)-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left[\theta_{\tau(m)} - \theta_t\right]^\top V_t(s_t,a_t) + \sum_{t=\tau(m(T))}^{T}\left[\theta_{\tau(m)} - \theta_t\right]^\top V_t(s_t,a_t) \\
&\leq \sum_{m=1}^{m(T)-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left\|\theta_{\tau(m)} - \theta_t\right\|_1 \left\|V_t(s_t,a_t)\right\|_\infty + \sum_{t=\tau(m(T))}^{T}\left\|\theta_{\tau(m)} - \theta_t\right\|_1 \left\|V_t(s_t,a_t)\right\|_\infty \qquad (31) \\
&\leq \sum_{m=1}^{m(T)-1} 1\cdot 1 + Q\max_{t\in\{1,\ldots,T\}} 1\cdot 1 = M(T). \qquad (32)
\end{aligned}$$

Step (31) is by the triangle inqeuality and the Cauchy-Scwhartz inequality. Step (32) is by our terminating criteria, which require $\Psi \leq 1$ for each episode. Similar to the above, we also have:

$$\begin{aligned}
&\sum_{t=1}^{T}(\heartsuit_t) = \left\{\sum_{m=1}^{m(T)-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left[\theta_t - \theta_{\tau(m)}\right] + \sum_{t=\tau(m(T))}^{T}\left[\theta_t - \theta_{\tau(m)}\right]\right\}^\top \sum_{s\in\mathcal{S},a\in\mathcal{A}_s} v(s,a)x^*(s,a) \\
&\leq \left\{\sum_{m=1}^{m(T)-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\left\|\theta_t - \theta_{\tau(m)}\right\|_1 + \sum_{t=\tau(m(T))}^{T}\left\|\theta_t - \theta_{\tau(m)}\right\|_1\right\}\left\|\sum_{s\in\mathcal{S},a\in\mathcal{A}_s} v(s,a)x^*(s,a)\right\|_\infty \\
&\leq \left[\sum_{m=1}^{m(T)-1} 1\right] + 1 = M(T).
\end{aligned}$$

Altogether, the Lemma is proved.

C.5. Proof of Lemma 5, which bounds (¶)

The proof uses Lemma 8. Let's apply $n = m(T)$, as well as

$$z_m = \begin{cases} \tau(m+1) - \tau(m) & \text{if } 1 \leq m < m(T) \\ T - \tau(m) & \text{if } m = m(T) \end{cases},$$

where we set $\tau(0) = 0$. Now, $Z_0 = 1$, $Z_m = \tau(m)$ for $1 \le m < m(T)$, and $Z_{m(T)} = T$. Therefore,

$$\sum_{t=1}^{T}(\P_t) = \sum_{m=1}^{m(T)-1}\sum_{t=\tau(m)}^{\tau(m+1)-1}\frac{1}{\sqrt{\tau(m)}} + \sum_{t=\tau(m(T))}^{T}\frac{1}{\sqrt{\tau(m(T))}}$$
$$= \sum_{m=1}^{n}\frac{z_m}{\sqrt{Z_{m-1}}} \le \left(\sqrt{2}+1\right)\sqrt{Z_{m(T)}} = \left(\sqrt{2}+1\right)\sqrt{T}.$$

Hence the claim is proved.

C.6. Proof of Lemma 6, which bounds $(\diamondsuit)$

The proof of the Lemma uses the Azuma-Hoeffding inequality in Theorem 5, as well as Lemma 9 by (Jaksch et al., 2010) and Lemma 10.

To start the proof, we define $\tilde{v}_m(s,a)$ and $m(t)$. We express $\tilde{r}_m(s,a) = \theta_{\tau(m)}^{\top}\tilde{v}_m(s,a)$, where $\tilde{v}_m(s,a)$ is an optimal solution to the optimization problem (6). For each t, we define $m(t)$ to be the episode index such that $\tau(m(t)) \le t < \tau(m(t)+1) - 1$. We first decompose $\sum_{t=1}^{T}(\diamondsuit_t)$ as follows:

$$\sum_{t=1}^{T}(\diamondsuit_t) \le \sum_{t=1}^{T}\tilde{r}_{m(t)}(s_t,a_t) - \theta_{\tau(m(t))}^{\top}V_t(s_t,a_t)$$
$$= \underbrace{\sum_{t=1}^{T}\theta_{\tau(m(t))}^{\top}\left[\tilde{v}_{m(t)}(s_t,a_t) - v(s_t,a_t)\right]}_{(\dagger_v)} + \underbrace{\sum_{t=1}^{T}\theta_{\tau(m(t))}^{\top}\left[v(s_t,a_t) - V_t(s_t,a_t)\right]}_{(\ddagger_v)}.$$

We bound the sums $(\dagger_v, \ddagger_v)$ as follows:

Bounding $(\dagger_v)$. We bound this term by invoking the confidence bounds asserted by the event $\mathcal{E}^v$. Define the notation $(\log\text{-}v) := \log(12KSAT^2/\delta)$. We have

$$(\dagger_v) = \sum_{t=1}^{T}\theta_{\tau(m(t))}^{\top}\left[\tilde{v}_{m(t)}(s_t,a_t) - \hat{v}_{m(t)}(s_t,a_t) + \hat{v}_{m(t)}(s_t,a_t) - v(s_t,a_t)\right]$$
$$\le \sum_{t=1}^{T}\left\|\theta_{\tau(m(t))}\right\|_1\left[\left\|\tilde{v}_{m(t)}(s_t,a_t) - \hat{v}_{m(t)}(s_t,a_t)\right\|_1 + \left\|\hat{v}_{m(t)}(s_t,a_t) - v(s_t,a_t)\right\|_1\right] \tag{33}$$
$$\le 2\sum_{t=1}^{T}\left\|(\mathsf{rad}^v_{m(t),k}(s_t,a_t))_{k=1}^{K}\right\|_\infty \tag{34}$$
$$\le 4\left[\sqrt{(\log\text{-}v)}\cdot\sum_{t=1}^{T}\frac{1}{\sqrt{N^+_{m(t)}(s_t,a_t)}} + 3\cdot(\log\text{-}v)\cdot\sum_{t=1}^{T}\frac{1}{N^+_{m(t)}(s_t,a_t)}\right] \tag{35}$$
$$\le 4\left[\left(\sqrt{2}+1\right)\sqrt{SAT\cdot(\log\text{-}v)} + 3\cdot(\log\text{-}v)\cdot SA\left(1+2\log T\right)\right].$$

Step (33) is by the Cauchy-Schwartz inequality, step (34) is by the assumption that the event $\mathcal{E}^v$ holds. Step (35) is by Lemma 9 and Lemma 10, as well as $(\log\text{-}v) \ge (\log\text{-}v)_m$ for all m.

Bounding $(\ddagger_v)$. Consider random variable $X_t = \theta_{\tau(m(t))}^{\top}\left[v(s_t,a_t) - V_t(s_t,a_t)\right]$ and filtration $\mathcal{F}_t = \sigma(\{s_t,a_t,V_t(s_t,a_t),\theta_{t+1}\}_{\tau=1}^{t})$. Now, $|X_t| \le 1$, X_t is $\mathcal{F}_t$-measurable with $\mathbf{E}[X_t|\mathcal{F}_{t-1}] = 0$. Thus, we apply Theorem 5 to conclude that, with probability $\ge 1-\delta$,

$$(\ddagger_v) \le \sqrt{2T\log(1/\delta)}.$$

Altogether, we have, with probability at least $1-\delta$,

$$\sum_{t=1}^{T}(\diamondsuit_t) \le \left[\left(5\sqrt{2}+4\right)\sqrt{\sum_{s\in\mathcal{S}}|\mathcal{A}_s|T\cdot(\text{log-}v)}+12\cdot(\text{log-}v)\cdot\sum_{s\in\mathcal{S}}|\mathcal{A}_s|\log T\right] \tag{36}$$
$$= O\left(\sqrt{SAT\log\frac{KSAT}{\delta}}+SA\log^2\frac{KSAT}{\delta}\right).$$

Hence, the Lemma is proved.

C.7. Proof of Lemma 7, which bounds ($\spadesuit$)

First, recall that $\tilde{phi}_m, \tilde{\gamma}_m$ are the dual variables output by the EVI at the start of episode m. Observe that

$$\max_{s\in\mathcal{S}}\{\tilde{\gamma}_m(s)\}-\min_{s\in\mathcal{S}}\{\tilde{\gamma}_m(s)\}\le D$$

for each m. Indeed, conditioned on $\mathcal{E}^p$, we have $p\in H_m^p$ for all m. In addition, $\max_{s,a}\tilde{r}_m(s,a)\le 1$. The desired inequality follows from item (ii) in Theorem 6. For each episode m and state s, consider replacing $\tilde{\gamma}_m(s)$ by $\gamma_m(s):=\tilde{\gamma}_m(s)-\min_{s'\in\mathcal{S}}\{\tilde{\gamma}_m(s')\}$. Now, $0\le\max_{m,s}\{\gamma_m(s)\}\le D$, and the value of each ($\spadesuit_t$) is preserved:

$$(\spadesuit_t)=\max_{\bar{p}\in H^p_{m(t)}(s_t,a_t)}\left\{\sum_{s'\in\mathcal{S}}\tilde{\gamma}_m(s')\bar{p}(s')\right\}-\tilde{\gamma}_m(s_t)$$
$$=\max_{\bar{p}\in H^p_{m(t)}(s_t,a_t)}\left\{\sum_{s'\in\mathcal{S}}\gamma_{m(t)}(s')\bar{p}(s')\right\}-\gamma_{m(t)}(s_t),$$

where $m(t)$ is the episode index such that $\tau(m(t))\le t<\tau(m(t)+1)$. Consider the following decomposition:

$$\sum_{t=1}^{T}(\spadesuit_t)=\underbrace{\sum_{t=1}^{T}\left[\max_{\bar{p}\in H^p_m(s_t,a_t)}\left\{\sum_{s'\in\mathcal{S}}\gamma_{m(t)}(s')\bar{p}(s')\right\}-\sum_{s\in\mathcal{S}}\gamma_{m(t)}(s)p(s|s_t,a_t)\right]}_{(\dagger_p)}$$
$$+\underbrace{\sum_{t=1}^{T}\left[\sum_{s\in\mathcal{S}}\gamma_{m(t)}(s)p(s|s_t,a_t)-\gamma_{m(t)}(s_t)\right]}_{(\ddagger_p)}.$$

Bounding $(\dagger_p)$**.** We proceed by unraveling H_m^p. Now, denote $(\text{log-}p):=\log(12S^2AT^2/\delta)$.

$$(\ddagger_p)\le\sum_{t=1}^{T}\left[\max_{\bar{p}\in H^p_{m(t)}(s_t,a_t)}\left\{\sum_{s\in\mathcal{S}}\gamma_{m(t)}(s)\bar{p}(s)\right\}-\min_{\bar{p}\in H^p_{m(t)}(s_t,a_t)}\left\{\sum_{s\in\mathcal{S}}\gamma_{m(t)}(s)\bar{p}(s)\right\}\right]$$
$$\le 2\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\gamma_{m(t)}(s)\mathsf{rad}^p_{m(t)}(s|s_t,a_t)$$
$$\le 2D\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\left[\sqrt{\frac{2\hat{p}_{m(t)}(s'|s,a)\cdot(\text{log-}p)}{N^+_{m(t)}(s,a)}}+\frac{3(\text{log-}p)}{N^+_{m(t)}(s,a)}\right]$$
$$\le 2D\sum_{t=1}^{T}\left[\sqrt{\frac{2\Gamma\cdot(\text{log-}p)}{N^+_{m(t)}(s,a)}}+\frac{3\cdot S(\text{log-}p)}{N^+_{m(t)}(s,a)}\right] \tag{37}$$
$$\le 2(\sqrt{2}+1)D\sqrt{2\Gamma SAT\cdot(\text{log-}p)}+6DS^2A(1+2\log T)(\text{log-}p). \tag{38}$$

We justify step (37) as follows. Now, recall $\Gamma = \max_{s\in\mathcal{S},a\in\mathcal{A}_s} \sum_{s'\in\mathcal{S}} \mathbf{1}(p(s'|s,a) > 0) = \|p(\cdot|s,a)\|_0$. With certainty, we have $\|\hat{p}_m(\cdot|s,a)\|_0 \le \|p(\cdot|s,a)\|_0 \le \Gamma$. Indeed, for each $s' \in \mathcal{S}$, $p(s'|s,a) = 0$ implies that $\hat{p}_m(s'|s,a) = 0$ with certainty. By the Cauchy-Schwartz inequality,

$$\begin{aligned}\sum_{s'\in\mathcal{S}} \sqrt{\hat{p}_m(s'|s,a)} &= \sum_{s'\in\mathcal{S}} \sqrt{\hat{p}_m(s'|s,a)\cdot \mathbf{1}(p(s'|s,a)>0)} \\ &\le \sqrt{\left[\sum_{s'\in\mathcal{S}} \hat{p}_m(s'|s,a)\right]\left[\sum_{s'\in\mathcal{S}} \mathbf{1}(p(s'|s,a)>0)\right]} = \sqrt{\|p(\cdot|s,a)\|_0} = \sqrt{\Gamma}.\end{aligned}$$

Step (38) is by Proposition 9 and Lemma 10.

Bounding $(\ddagger_p)$**.** We analyze the term by accounting for the number of episodes:

$$\begin{aligned}(\ddagger_p) &= \left[\gamma_{m(T+1)}(s_{T+1}) - \gamma_{m(1)}(s_1)\right] + \sum_{t=1}^{T}\left[\sum_{s\in\mathcal{S}} \gamma_{m(t)}(s)p(s|s_t,a_t) - \gamma_{m(t+1)}(s_{t+1})\right] \\ &= \left[\gamma_{m(T+1)}(s_{T+1}) - \gamma_{m(1)}(s_1)\right] + \sum_{t=1}^{T}\left[\gamma_{m(t)}(s_{t+1}) - \gamma_{m(t+1)}(s_{t+1})\right] \\ &\quad + \sum_{t=1}^{T}\left[\sum_{s\in\mathcal{S}} \gamma_{m(t)}(s)p(s|s_t,a_t) - \gamma_{m(t)}(s_{t+1})\right] \quad \text{w.p. } 1 \qquad (39) \\ &\le \max_{t,s}\{\gamma_{m(t)}(s)\}(M(T)+1) + \sum_{t=1}^{T}\left[\sum_{s\in\mathcal{S}} \gamma_{m(t)}(s)p(s|s_t,a_t) - \gamma_{m(t)}(s_{t+1})\right] \qquad (40) \\ &\le \max_{t,s}\{\gamma_{m(t)}(s)\}(M(T)+1) + \max_{t,s}\{\gamma_{m(t)}(s)\}\sqrt{2T\log(1/\delta)} \quad \text{w.p. } 1-\delta \\ &\le D(M(T)+1) + D\sqrt{2T\log(1/\delta)}.\end{aligned}$$

Step (40) is shown by analyzing the second summation in (39), which is $\sum_{t=1}^{T} \gamma_{m(t)}(s_{t+1}) - \gamma_{m(t+1)}(s_{t+1})$. In the summation, at most $m(T) \le M(T)$ summands are non-zero, and each non-zero summand is less than or equal to $\max_{t,s}\{\gamma_{m(t)}(s)\} \le D$.

Combining the bounds for $(\dagger_p, \ddagger_p)$, with probability at least $1-\delta$ we have

$$\begin{aligned}\sum_{t=1}^{T}(\spadesuit_t) &\le (2\sqrt{2}+3)D\sqrt{2\Gamma SAT\cdot(\text{log-}p)} + D(M(T)+1) \qquad (41) \\ &\quad + 6DS^2A(1+2\log T)\cdot(\text{log-}p) \\ &= O\left(D\cdot M(T)\right) + O\left(D\sqrt{\Gamma SAT\log\frac{SAT}{\delta}} + DS^2A\log^2\frac{SAT}{\delta}\right).\end{aligned}$$

Altogether, the Lemma is proved.