A. Omitted Algorithm for Player 2 in Section 3

Algorithm 4 Optimistic Policy Optimization for Player 2 with Factored Independent Transition

Initialize: For all $h \in [H]$, $(s^1, s^2, a, b) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{A} \times \mathcal{B}$: $\mu_h^0(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$, $\widehat{\mathcal{P}}_h^{1,0}(\cdot|s^1, a) = \mathbf{1}/|\mathcal{S}_1|$, $\widehat{\mathcal{P}}_h^{2,0}(\cdot|s^2, b) = \mathbf{1}/|\mathcal{S}_2|$, $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^0(\cdot,\cdot,\cdot) = \mathbf{0}$.
for episode $k = 1, \ldots, K$ **do**
  Observe Player 1's policy $\{\mu_h^{k-1}\}_{h=1}^H$.
  Start from state $s_1 = (s_1^1, s_1^2)$, set $\overline{V}_{H+1}^{k-1}(\cdot) = \mathbf{0}$.
  for step $h = H, H-1, \ldots, 1$ **do**
    Estimate the transition and reward function by $\widehat{\mathcal{P}}_h^{k-1}(\cdot|\cdot,\cdot)$ and $\widehat{r}_h^{k-1}(\cdot,\cdot,\cdot)$ as (11).
    Update Q-function $\forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$:

$$\underline{Q}_h^{k-1}(s, a, b) = \min\{(\widehat{r}_h^{k-1} + \widehat{\mathcal{P}}_h^{k-1}\underline{V}_{h+1}^{k-1} - \beta_h^{k-1})(s, a, b), H - h + 1\}^+.$$

    Update value-function $\forall s \in \mathcal{S}$:

$$\underline{V}_h^{k-1}(s) = \left[\mu_h^{k-1}(\cdot|s)\right]^\top \underline{Q}_h^{k-1}(s, \cdot, \cdot)\nu_h^{k-1}(\cdot|s).$$

  end for
  Compute the empirical state reaching probability $d_h^{\mu^k, \widehat{\mathcal{P}}^{1,k}}(s^2)$ of Player 1 under $\mu^k, \widehat{\mathcal{P}}^{1,k}$, $\forall h \in [H]$.
  Update policy $\nu_h^k(b|s^2)$ by solving (15), $\forall (s^2, b, h)$.
  Take actions following $b_h^k \sim \nu_h^k(\cdot|s_h^{2,k}),\ \forall h \in [H]$.
  Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
end for

Based on the empirical state reaching probability, the policy improvement step is associated with solving the following optimization problem

$$\max_{\mu} \sum_{h=1}^{H} [\underline{G}_h^{k-1}(\nu_h) + \gamma^{-1} D_{\mathrm{KL}}(\nu_h(\cdot|s^2), \nu_h^k(\cdot|s^2))], \tag{15}$$

where we define the linear function as $\overline{G}_h^{k-1}(\mu_h) := \langle \nu_h(\cdot|s^2) - \nu_h^k(\cdot|s^2), \sum_{s^1 \in \mathcal{S}_1} F_h^{2,k}(s, \cdot) d_h^{\mu^k, \widehat{\mathcal{P}}^{1,k}}(s^1)\rangle_{\mathcal{B}}$ with $F_h^{2,k}(s, b) = \langle \underline{Q}_h^k(s, \cdot, b), \mu_h^k(\cdot|s^1)\rangle_{\mathcal{A}}$. Here (15) is a standard mirror descent step and admits a closed-form solution as $\nu_h^k(b|s^2) = (\widetilde{Y}_h^{k-1})^{-1} \nu_h^{k-1}(b \,|\, s^2) \cdot \exp\{-\gamma \sum_{s^1 \in \mathcal{S}_1} F_h^{2,k}(s, b) d_h^{\mu^k, \widehat{\mathcal{P}}^{1,k}}(s^1)\rangle_{\mathcal{A}}\}$, where $\widetilde{Y}_h^{k-1}$ is a probability normalization term.

B. Proofs for Section 3

Lemma B.1. *At the k-th episode, the difference between value functions $V_1^{\mu^*, \nu^k}(s_1)$ and $V_1^{\mu^k, \nu^k}(s_1)$ is*

$$\begin{aligned}
&V_1^{\mu^*, \nu^k}(s_1) - V_1^{\mu^k, \nu^k}(s_1) \\
&\quad = \overline{V}_1^k(s_1) - V_1^{\mu^k, \nu^k}(s_1) + \sum_{h=1}^{H} \mathbb{E}_{\mu^*, \mathcal{P}, \nu^k}\left\{[\mu_h^*(\cdot|s_h)]^\top \bar{\iota}_h^k(s_h, \cdot, \cdot)\nu_h^k(\cdot|s_h) \,\middle|\, s_1\right\} \\
&\qquad + \sum_{h=1}^{H} \mathbb{E}_{\mu^*, \mathcal{P}^1}\left\{\left\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), \sum_{s_h^2 \in \mathcal{S}_2} F_h^k(s_h^1, s_h^2, \cdot) d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s_h^2)\right\rangle_{\mathcal{A}} \,\middle|\, s_1^1, s_1^2\right\} \\
&\qquad + 2H \sum_{h=1}^{H} \sum_{s_h^2 \in \mathcal{S}_2} \left| q_h^{\nu^k, \mathcal{P}^2}(s_h^2) - d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s_h^2)\right|,
\end{aligned} \tag{16}$$

where s_h, a_h, b_h are random variables for state and actions, $U_h^k(s, a) := \langle \overline{Q}_h^k(s, a, \cdot), \nu_h^k(\cdot \,|\, s)\rangle_{\mathcal{B}}$, and we define the model prediction error of Q-function as

$$\bar{\iota}_h^k(s, a, b) = r_h(s, a, b) + \mathcal{P}_h \overline{V}_{h+1}^k(s, a, b) - \overline{Q}_h^k(s, a, b). \tag{17}$$

Proof. The proof of this lemma starts with decomposing the value function difference as

$$V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1) = V_1^{\mu^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) + \overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1). \tag{18}$$

Here the term $\overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1)$ is the bias between the estimated value function $\overline{V}_1^k(s_1)$ generated by Algorithm 1 and the value function $V_1^{\mu^k,\nu^k}(s_1)$ under the true transition model $\mathcal{P}$ at the k-th episode.

We first analyze the term $V_1^{\mu^*,\nu^k}(s_1) - \overline{V}_1^k(s_1)$. For any h and s, we consider to decompose the term $V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s)$, which gives

$$\begin{aligned}
&V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s)\\
&\quad = [\mu_h^*(\cdot|s)]^\top Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\\
&\quad = [\mu_h^*(\cdot|s)]^\top Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - [\mu_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\\
&\qquad + [\mu_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\\
&\quad = [\mu_h^*(\cdot|s)]^\top \big[Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s)\\
&\qquad + \big[\mu_h^*(\cdot|s) - \mu_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s),
\end{aligned} \tag{19}$$

where the first inequality is by the definition of $V_h^{\mu^*,\nu^k}$ in (1) and the definition of $\overline{V}_h^k$ in Line 1 of Algorithm 1. In addition, by the definition of $Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot)$ in (2) and the definition of the model prediction error $\overline{\iota}_h^k$ for Player one in (36), we have

$$\begin{aligned}
&[\mu_h^*(\cdot|s)]^\top \big[Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s)\\
&\quad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\Bigg[\sum_{s'\in\mathcal{S}} \mathcal{P}_h(s'|s,a,b)\big[V_{h+1}^{\mu^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] + \overline{\iota}_h^k(s,a,b)\Bigg]\nu_h^k(b|s)\\
&\quad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\Bigg[\sum_{s'\in\mathcal{S}} \mathcal{P}_h(s'|s,a,b)\big[V_{h+1}^{\mu^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big]\Bigg]\nu_h^k(b|s) + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\overline{\iota}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned}$$

Combining this equality with (19) gives

$$\begin{aligned}
V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s) &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\Bigg[\sum_{s'\in\mathcal{S}} \mathcal{P}_h(s'|s,a,b)\big[V_{h+1}^{\mu^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big]\Bigg]\nu_h^k(b|s)\\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\overline{\iota}_h^k(s,a,b)\nu_h^k(b|s)\\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \big[\mu_h^*(a|s) - \mu_h^k(a|s)\big]\overline{Q}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned} \tag{20}$$

The inequality (20) indicates a recursion of the value function difference $V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s)$. As we have defined $V_{H+1}^{\mu^*,\nu^k}(s) = 0$ and $\overline{V}_{H+1}^k(s) = 0$, by recursively applying (20) from $h = 1$ to H, we obtain

$$\begin{aligned}
&V_1^{\mu^*,\nu^k}(s_1) - \overline{V}_1^k(s_1)\\
&\quad = \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big\{[\mu_h^*(\cdot|s_h)]^\top \overline{\iota}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\}\\
&\qquad + \underbrace{\sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big\{\big[\mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\}}_{\text{Term(I)}},
\end{aligned} \tag{21}$$

where s_h are a random variables denoting the state at the h-th step following a distribution determined jointly by $\mu^*, \mathcal{P}, \nu^k$. Note that we have the factored independent transition model structure $\mathcal{P}_h(s'|s,a,b) = \mathcal{P}_h^1(s^{1\prime}|s^1,a)\mathcal{P}_h^2(s^{2\prime}|s^2,b)$ with

$s = (s^1, s^2)$ and $s' = (s^{1\prime}, s^{2\prime})$, and $\mu_h(a|s) = \mu_h(a|s^1)$ as well as $\nu_h(b|s) = \nu_h(b|s^2)$. Here we also have the state reaching probability $q^{\nu^k,\mathcal{P}^2}(s^2) = \big\{q_h^{\mu^k,\mathcal{P}^2}(s^2)\big\}_{h=1}^H$ under μ^k and true transition $\mathcal{P}^2$ for Player 2, and define the empirical reaching probability $d^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2) = \{d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\}_{h=1}^H$ under the empirical transition model $\widehat{\mathcal{P}}^{2,k}$ for Player 2, where we let $\widehat{\mathcal{P}}_h^k(s'|s,a,b) = \widehat{\mathcal{P}}_h^{1,k}(s^{1\prime}|s^1,a)\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime}|s^2,b)$. Then, for Term(I), we have

$$\begin{aligned}\text{Term(I)} &= \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big\{\big[\mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\}\\ &= \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1,\mathcal{P}^2,\nu^k}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)\,\big|\, s_1^1,s_1^2\big\}\\ &= \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2}\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)q_h^{\nu^k,\mathcal{P}^2}(s_h^2)\,\big|\, s_1^1,s_1^2\big\}.\end{aligned} \tag{22}$$

The last term of the above inequality (22) can be further bounded as

$$\begin{aligned}&\sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2}\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)q_h^{\nu^k,\mathcal{P}^2}(s_h^2)\,\big|\, s_1^1,s_1^2\big\}\\ &\quad= \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2}\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\,\big|\, s_1^1,s_1^2\big\}\\ &\qquad+ \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2}\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)[q_h^{\nu^k,\mathcal{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)]\,\big|\, s_1^1,s_1^2\big\}\\ &\quad\le \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2}\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\,\big|\, s_1^1,s_1^2\big\}\\ &\qquad+ 2H\sum_{h=1}^H\sum_{s_h^2\in\mathcal{S}_2}\Big|q_h^{\nu^k,\mathcal{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big|,\end{aligned}$$

where the factor H in the last term is due to $|\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)| \le H$. Combining the above inequality with (22), we have

$$\begin{aligned}\text{Term(I)} &\le \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\big\{\big[\mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2}\overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\,\big|\, s_1^1,s_1^2\big\}\\ &\quad+ 2H\sum_{h=1}^H\sum_{s_h^2\in\mathcal{S}_2}\Big|q_h^{\nu^k,\mathcal{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big|.\end{aligned} \tag{23}$$

Further combining (23) with (18), we eventually have

$$\begin{aligned}&V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1)\\ &\quad\le \overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1) + \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big\{[\mu_h^*(\cdot|s_h)]^\top \bar{\iota}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\}\\ &\qquad+ \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}^1}\Big\{\Big\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s_h^1,s_h^2,\cdot)d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}}\,\Big|\, s_1^1,s_1^2\Big\}\\ &\qquad+ 2H\sum_{h=1}^H\sum_{s_h^2\in\mathcal{S}_2}\Big|q_h^{\nu^k,\mathcal{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big|,\end{aligned}$$

where we denote $F_h^{1,k}(s_h^1, s_h^2, a) := \langle \overline{Q}_h^k(s_h^1, s_h^2, a, \cdot), \nu_h^k(\cdot|s_h^2)\rangle_{\mathcal{B}}$ for any $a \in \mathcal{A}$. This completes our proof. □

Lemma B.2. *With setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, the mirror ascent steps of Algorithm 1 lead to*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P}^1}\left\{\left\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s_h^1, s_h^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\right\rangle_{\mathcal{A}} \,\middle|\, s_1^1, s_1^2\right\} \leq \mathcal{O}\left(\sqrt{H^4 K\log|\mathcal{A}|}\right).$$

Proof. As shown in (10), the mirror ascent step at the k-th episode is to solve the following maximization problem

$$\underset{\mu\in\Delta(\mathcal{A}\,|\,\mathcal{S}_1,H)}{\text{maximize}} \sum_{h=1}^{H}\left\langle \mu_h(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right\rangle_{\mathcal{A}} - \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\mu_h(\cdot|s^1), \mu_h^k(\cdot|s^1)\big),$$

with $F_h^{1,k}(s^1, s^2, a) := \langle \overline{Q}_h^k(s^1, s^2, a, \cdot), \nu_h^k(\cdot|s^2)\rangle_{\mathcal{B}}$. We equivalently rewrite this maximization problem to a minimization problem as

$$\underset{\mu\in\Delta(\mathcal{A}\,|\,\mathcal{S}_1,H)}{\text{minimize}} -\sum_{h=1}^{H}\left\langle \mu_h(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right\rangle_{\mathcal{A}} + \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\mu_h(\cdot|s^1), \mu_h^k(\cdot|s^1)\big).$$

Note that the closed-form solution μ_h^{k+1} to this minimization problem is guaranteed to stay in the relative interior of its feasible set $\Delta(\mathcal{A}\,|\,\mathcal{S}_1, H)$ if initializing $\mu_h^0(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$. Thus, we apply Lemma C.12 and obtain that for any $\mu = \{\mu_h\}_{h=1}^H$, the following inequality holds

$$\begin{aligned}&-\eta\left\langle \mu_h^{k+1}(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right\rangle_{\mathcal{A}} + \eta\left\langle \mu_h(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right\rangle_{\mathcal{A}}\\ &\quad\leq D_{\mathrm{KL}}\big(\mu_h(\cdot|s^1), \mu_h^k(\cdot|s^1)\big) - D_{\mathrm{KL}}\big(\mu_h(\cdot|s^1), \mu_h^{k+1}(\cdot|s^1)\big) - D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s^1), \mu_h^k(\cdot|s^1)\big).\end{aligned}$$

Then, by rearranging the terms and letting $\mu_h = \mu_h^*$, we have

$$\begin{aligned}&\eta\left\langle \mu_h^*(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\right\rangle_{\mathcal{A}}\\ &\quad\leq D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s^1), \mu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s), \mu_h^k(\cdot|s)\big)\\ &\qquad + \eta\left\langle \mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)\right\rangle_{\mathcal{A}}.\end{aligned} \tag{24}$$

Due to Pinsker's inequality, we have

$$-D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s^1), \mu_h^k(\cdot|s^1)\big) \leq -\frac{1}{2}\big\|\mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1)\big\|_1^2.$$

Further by Cauchy-Schwarz inequality, we have

$$\eta\left\langle \mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right\rangle_{\mathcal{A}} \leq \eta H\big\|\mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1)\big\|_1.$$

since we have

$$\begin{aligned}&\left\|\sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right\|_\infty\\ &\quad= \max_{a\in\mathcal{A}}\sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, a) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\\ &\quad= \max_{a\in\mathcal{A}}\sum_{s^2\in\mathcal{S}_2} \langle \overline{Q}_h^k(s^1, s^2, a, \cdot), \nu_h^k(\cdot|s^2)\rangle_{\mathcal{B}} \cdot d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\\ &\quad\leq \sum_{s^2\in\mathcal{S}_2} H\cdot d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2) = H.\end{aligned}$$

Thus, we further obtain

$$
\begin{aligned}
&- D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s^1), \mu_h^k(\cdot|s^1)\big) + \eta\big\langle \mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s^2)\big\rangle_{\mathcal{A}} \\
&\quad \le -\frac{1}{2}\big\|\mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1)\big\|_1^2 + \eta H \big\|\mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1)\big\|_1 \\
&\quad \le \frac{1}{2}\eta^2 H^2,
\end{aligned}
\tag{25}
$$

where the last inequality is by viewing $\big\|\mu_h^{k+1}(\cdot|s^1) - \mu_h^k(\cdot|s^1)\big\|_1$ as a variable x and finding the maximal value of $-1/2 \cdot x^2 + \eta H x$ to obtain the upper bound $1/2 \cdot \eta^2 H^2$.

Thus, combing (25) with (24), the policy improvement step in Algorithm 1 implies

$$
\begin{aligned}
&\eta\Big\langle \mu_h^*(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^k(s^1, s^2, \cdot) d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s^2)\Big\rangle_{\mathcal{A}} \\
&\quad \le D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s^1), \mu_h^k(\cdot|s^1)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s^1), \mu_h^{k+1}(\cdot|s^1)\big) + \frac{1}{2}\eta^2 H^2,
\end{aligned}
$$

which further leads to

$$
\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\Big\{\Big\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s_h^1, s_h^2, \cdot) d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \,\Big|\, s_1^1, s_1^2\Big\} \\
&\quad \le \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s_h^1), \mu_h^k(\cdot|s_h^1)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s_h^1), \mu_h^{k+1}(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta H^3.
\end{aligned}
$$

Taking summation from $k = 1$ to K of both sides, we obtain

$$
\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\Big\{\Big\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s_h^1, s_h^2, \cdot) d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \,\Big|\, s_1^1, s_1^2\Big\} \\
&\quad \le \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s_h^1), \mu_h^1(\cdot|s_h^1)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s_h^1), \mu_h^{K+1}(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta K H^3 \\
&\quad \le \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s_h^1), \mu_h^1(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta K H^3,
\end{aligned}
$$

where the last inequality is by non-negativity of KL divergence. With the initialization in Algorithm 1, it is guaranteed that $\mu_h^1(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$, which thus leads to $D_{\mathrm{KL}}\left(\mu_h^*(\cdot|s^1), \mu_h^1(\cdot|s^1)\right) \le \log|\mathcal{A}|$ for any s^1. Then, with setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, we bound the last term as

$$
\frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s_h^1), \mu_h^1(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta K H^3 \le \mathcal{O}\left(\sqrt{H^4 K \log|\mathcal{A}|}\right),
$$

which gives

$$
\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\mu^*, \mathcal{P}^1}\Big\{\Big\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s_h^1, s_h^2, \cdot) d_h^{\nu^k, \widehat{\mathcal{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \,\Big|\, s_1^1, s_1^2\Big\} \le \mathcal{O}\left(\sqrt{H^4 K \log|\mathcal{A}|}\right).
$$

This completes the proof. □

Lemma B.3. *For any $k \in [K]$, $h \in [H]$ and all $(s, a, b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, with probability at least $1-\delta$, we have*

$$
\big|\widehat{r}_h^k(s, a, b) - r_h(s, a, b)\big| \le \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s, a, b), 1\}}}.
$$

Proof. The proof for this theorem is a direct application of Hoeffding's inequality. For $k \geq 1$, the definition of $\widehat{r}_h^k$ in (11) indicates that $\widehat{r}_h^k(s,a,b)$ is the average of $N_h^k(s,a,b)$ samples of the observed rewards at (s,a,b) if $N_h^k(s,a,b) > 0$. Then, for fixed $k \in [K], h \in [H]$ and state-action tuple $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, when $N_h^k(s,a,b) > 0$, according to Hoeffding's inequality, with probability at least $1-\delta'$ where $\delta' \in (0,1]$, we have

$$\big|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)\big| \leq \sqrt{\frac{\log(2/\delta')}{2N_h^k(s,a,b)}},$$

where we also use the facts that the observed rewards $r_h^k \in [0,1]$ for all k and h, and $\mathbb{E}\big[\widehat{r}_h^k\big] = r_h$ for all k and h. For the case where $N_h^k(s,a,b) = 0$, by (11), we know $\widehat{r}_h^k(s,a,b) = 0$ such that $|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)| = |r_h(s,a,b)| \leq 1$. On the other hand, we have $\sqrt{2\log(2/\delta')} \geq 1 > |\widehat{r}_h^k(s,a,b) - r_h(s,a,b)|$. Thus, combining the above results, with probability at least $1-\delta'$, for fixed $k \in [K], h \in [H]$ and state-action tuple $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, we have

$$\big|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)\big| \leq \sqrt{\frac{2\log(2/\delta')}{\max\{N_h^k(s,a,b),1\}}}.$$

Moreover, by the union bound, letting $\delta = |\mathcal{S}||\mathcal{A}||\mathcal{B}|HK\delta'/2$, assuming $K > 1$, with probability at least $1-\delta$, for any $k \in [K], h \in [H]$ and any state-action tuple $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, we have

$$\big|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)\big| \leq \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}.$$

This completes the proof. ∎

In (9), we actually factor the state as $s = (s^1, s^2)$ such that we have $|\mathcal{S}| = |\mathcal{S}_1||\mathcal{S}_2|$. Thus, we set $\beta_h^{r,k}(s,a,b) = \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}} = \sqrt{\frac{4\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^1,s^2,a,b),1\}}}$, which equals the bound in Lemma B.3. The counter $N_h^k(s,a,b)$ is equivalent to $N_h^k(s^1,s^2,a,b)$.

Lemma B.4. *For any $k \in [K]$, $h \in [H]$ and all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$, we have*

$$\Big\|\widehat{\mathcal{P}}_h^k(\cdot \,|\, s,a,b) - \mathcal{P}_h(\cdot \,|\, s,a,b)\Big\|_1 \leq \sqrt{\frac{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}},$$

where we have a factored state space $s = (s^1, s^2)$, $s' = (s^{1\prime}, s^{2\prime})$, and an independent state transition $\mathcal{P}_h(s' \,|\, s,a,b) = \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\mathcal{P}_h^1(s^{2\prime} \,|\, s^2,b)$ and $\widehat{\mathcal{P}}_h^k(\cdot \,|\, s,a,b) = \widehat{\mathcal{P}}_h^{1,k}(s^{1\prime} \,|\, s^1,a)\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b)$.

Proof. Since the state space and the transition model are factored, we need to decompose the term as follows

$$\begin{aligned}
&\Big\|\widehat{\mathcal{P}}_h^k(\cdot \,|\, s,a,b) - \mathcal{P}_h(\cdot \,|\, s,a,b)\Big\|_1 \\
&\quad = \sum_{s^{1\prime},s^{2\prime}} \Big|\widehat{\mathcal{P}}_h^{1,k}(s^{1\prime} \,|\, s^1,a)\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) - \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\mathcal{P}_h^2(s^{2\prime} \,|\, s^2,b)\Big| \\
&\quad = \sum_{s^{1\prime},s^{2\prime}} \Big|\Big[\widehat{\mathcal{P}}_h^{1,k}(s^{1\prime} \,|\, s^1,a) - \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\Big]\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) + \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\Big[\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) - \mathcal{P}_h^2(s^{2\prime} \,|\, s^2,b)\Big]\Big| \\
&\quad \leq \sum_{s^{1\prime},s^{2\prime}} \Big\{\Big|\widehat{\mathcal{P}}_h^{1,k}(s^{1\prime} \,|\, s^1,a) - \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\Big|\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) + \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\Big|\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) - \mathcal{P}_h^2(s^{2\prime} \,|\, s^2,b)\Big|\Big\} \\
&\quad \leq \sum_{s^{1\prime}} \Big|\widehat{\mathcal{P}}_h^{1,k}(s^{1\prime} \,|\, s^1,a) - \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\Big| + \sum_{s^{2\prime}} \Big|\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) - \mathcal{P}_h^2(s^{2\prime} \,|\, s^2,b)\Big| \\
&\quad = \Big\|\widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1,a) - \mathcal{P}_h^1(\cdot \,|\, s^1,a)\Big\|_1 + \Big\|\widehat{\mathcal{P}}_h^{2,k}(\cdot \,|\, s^2,b) - \mathcal{P}_h^2(\cdot \,|\, s^2,b)\Big\|_1
\end{aligned}$$

where the last inequality is due to $\sum_{s^{2\prime}} \widehat{\mathcal{P}}_h^{2,k}(s^{2\prime} \,|\, s^2,b) = 1$ and $\sum_{s^{1\prime}} \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a) = 1$. Thus, we need to bound the two terms $\|\widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1,a) - \mathcal{P}_h^1(s^{1\prime} \,|\, s^1,a)\|_1$ and $\|\widehat{\mathcal{P}}_h^{2,k}(\cdot \,|\, s^2,b) - \mathcal{P}_h^2(\cdot \,|\, s^2,b)\|_1$ separately.

For $k \geq 1$, we have $\|\widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a)\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(s^{1\prime} \,|\, s^1, a), \mathbf{z} \rangle_{\mathcal{S}_1}$ by the duality. We construct an ε-covering net for the set $\{\mathbf{z} \in \mathbb{R}^{|\mathcal{S}_1|} : \|\mathbf{z}\|_\infty \leq 1\}$ with the distance induced by $\|\cdot\|_\infty$, denoted as $\mathcal{N}_\varepsilon$, such that for any $\mathbf{z} \in \mathbb{R}^{|\mathcal{S}_1|}$, there always exists $\mathbf{z}' \in \mathcal{N}_\varepsilon$ satisfying $\|\mathbf{z} - \mathbf{z}'\|_\infty \leq \varepsilon$. The covering number is $|\mathcal{N}_\varepsilon| = 1/\varepsilon^{|\mathcal{S}_1|}$. Thus, we know that for any $(s^1, a) \in \mathcal{S}_1 \times \mathcal{A}$ and any $\mathbf{z}$ with $\|\mathbf{z}\|_\infty \leq 1$, there exists $\mathbf{z}' \in \mathcal{N}_\varepsilon$ such that $\|\mathbf{z}' - \mathbf{z}\|_\infty \leq \varepsilon$ and

$$
\begin{aligned}
&\big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z} \big\rangle_{\mathcal{S}_1} \\
&\quad = \big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \big\rangle_{\mathcal{S}_1} + \big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z} - \mathbf{z}' \big\rangle_{\mathcal{S}_1} \\
&\quad \leq \big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \big\rangle_{\mathcal{S}_1} + \varepsilon \Big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \Big\|_1,
\end{aligned}
$$

such that we further have

$$
\begin{aligned}
&\Big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \Big\|_1 \\
&\quad = \max_{\|\mathbf{z}\|_\infty \leq 1} \big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a)), \mathbf{z} \big\rangle_{\mathcal{S}_1} \\
&\quad \leq \max_{\mathbf{z}' \in \mathcal{N}_\varepsilon} \big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \big\rangle_{\mathcal{S}_1} + \varepsilon \Big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \Big\|_1.
\end{aligned} \tag{26}
$$

By Hoeffding's inequality and union bound over all $\mathbf{z}' \in \mathcal{N}_\varepsilon$, when $N_h^k(s^1, a) > 0$, with probability at least $1 - \delta'$ where $\delta' \in (0, 1]$,

$$
\max_{\mathbf{z}' \in \mathcal{N}_\varepsilon} \big\langle \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \big\rangle_{\mathcal{S}_1} \leq \sqrt{\frac{|\mathcal{S}_1| \log(1/\varepsilon) + \log(1/\delta')}{2N_h^k(s^1, a)}}. \tag{27}
$$

Letting $\varepsilon = 1/2$, by (26) and (27), with probability at least $1 - \delta'$, we have

$$
\Big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \Big\|_1 \leq 1\sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2N_h^k(s^1, a)}}.
$$

When $N_h^k(s^1, a) = 0$, we have $\big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \big\|_1 = \|\mathcal{P}_h^1(\cdot \,|\, s^1, a)\|_1 = 1$ such that $2\sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2}} > 1 = \big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \big\|_1$ always holds. Thus, with probability at least $1 - \delta'$,

$$
\Big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \Big\|_1 \leq 2\sqrt{\frac{|\mathcal{S}_1| \log 2 + \log(1/\delta')}{2 \max\{N_h^k(s^1, a), 1\}}} \leq \sqrt{\frac{2|\mathcal{S}_1| \log(2/\delta')}{\max\{N_h^k(s^1, a), 1\}}}.
$$

Then, by union bound, assuming $K > 1$, letting $\delta'' = |\mathcal{S}_1||\mathcal{A}|HK\delta'/2$, with probability at least $1 - \delta''$, for any $(s^1, a) \in \mathcal{S}_1 \times \mathcal{A}$ and any $h \in [H]$ and $k \in [K]$, we have

$$
\Big\| \widehat{\mathcal{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathcal{P}_h^1(\cdot \,|\, s^1, a) \Big\|_1 \leq \sqrt{\frac{2|\mathcal{S}_1| \log(|\mathcal{S}_1||\mathcal{A}|HK/\delta'')}{\max\{N_h^k(s^1, a), 1\}}}.
$$

Similarly, we can also obtain that with probability at least $1 - \delta''$, for any $(s^2, a) \in \mathcal{S}_2 \times \mathcal{B}$ and any $h \in [H]$ and $k \in [K]$, we have

$$
\Big\| \widehat{\mathcal{P}}_h^{2,k}(\cdot \,|\, s^2, b) - \mathcal{P}_h^2(\cdot \,|\, s^2, b) \Big\|_1 \leq \sqrt{\frac{2|\mathcal{S}_2| \log(|\mathcal{S}_2||\mathcal{B}|HK/\delta'')}{\max\{N_h^k(s^2, b), 1\}}}.
$$

Further by union bound, we have with probability at least $1 - \delta$ where $\delta = 2\delta''$,

$$
\Big\| \widehat{\mathcal{P}}_h^{k}(\cdot \,|\, s, a, b) - \mathcal{P}_h(\cdot \,|\, s, a, b) \Big\|_1 \leq \sqrt{\frac{2|\mathcal{S}_1| \log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s^1, a), 1\}}} + \sqrt{\frac{2|\mathcal{S}_2| \log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^2, b), 1\}}}.
$$

This completes the proof. $\square$

In (9), we set $\beta_h^{\mathcal{P},k}(s,a,b) = \sqrt{\frac{2H^2|\mathcal{S}_1|\log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s^1,a),1\}}} + \sqrt{\frac{2H^2|\mathcal{S}_2|\log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^2,b),1\}}}$, which equals the product of the upper bound in Lemma B.4 and the factor H.

Lemma B.5. *With probability at least $1-2\delta$, Algorithm 1 ensures that*

$$\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\left[\bar{\iota}_h^k(s_h,a_h,b_h)\,\middle|\, s_1\right] \le 0.$$

Proof. We prove the upper bound of the model prediction error term. We can write the instantaneous prediction error at the h-step of the k-th episode as

$$\bar{\iota}_h^k(s,a,b) = r_h(s,a,b) + \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \overline{Q}_h^k(s,a,b), \tag{28}$$

where the equality is by the definition of the prediction error in (17). By plugging in the definition of $\overline{Q}_h^k$ in Line (1) of Algorithm 1, for any (s,a,b), we bound the following term as

$$\begin{aligned} & r_h(s,a,b) + \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \overline{Q}_h^k(s,a,b) \\ &\quad \le r_h(s,a,b) + \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \min\left\{\widehat{r}_h^k(s,a,b) + \left\langle \widehat{\mathcal{P}}_h^k(\cdot|s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \beta_h^k, H-h+1\right\} \\ &\quad \le \max\left\{r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathcal{P}}_h^k(\cdot|s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \beta_h^k, 0\right\}, \end{aligned} \tag{29}$$

where the inequality holds because

$$\begin{aligned} & r_h(s,a,b) + \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} \\ &\quad \le r_h(s,a,b) + \left\|\mathcal{P}_h(\cdot\,|\,s,a,b)\right\|_1 \|\overline{V}_{h+1}^k(\cdot)\|_\infty \le 1 + \max_{s'\in\mathcal{S}}\left|\overline{V}_{h+1}^k(s')\right| \le 1+H-h, \end{aligned}$$

since $\left\|\mathcal{P}_h(\cdot\,|\,s,a,b)\right\|_1 = 1$ and also the truncation step as shown in Line 1 of Algorithm 1 for $\overline{Q}_{h+1}^k$ such that for any $s'\in\mathcal{S}$

$$\begin{aligned} \left|\overline{V}_{h+1}^k(s')\right| &= \left|\left[\mu_{h+1}^k(\cdot|s')\right]^\top \overline{Q}_{h+1}^k(s',\cdot,\cdot)\nu_{h+1}^k(\cdot|s')\right| \\ &\le \left\|\mu_{h+1}^k(\cdot|s')\right\|_1 \left\|\overline{Q}_{h+1}^k(s',\cdot,\cdot)\nu_{h+1}^k(\cdot|s')\right\|_\infty \\ &\le \max_{a,b}\left|\overline{Q}_{h+1}^k(s',a,b)\right| \le H. \end{aligned} \tag{30}$$

Combining (28) and (29) gives

$$\bar{\iota}_h^k(s,a,b) \le \max\left\{r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathcal{P}}_h^k(\cdot|s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \beta_h^k, 0\right\}. \tag{31}$$

Note that as shown in (9), we have

$$\beta_h^k(s,a,b) = \beta_h^{r,k}(s,a,b) + \beta_h^{\mathcal{P},k}(s,a,b).$$

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

$$\begin{aligned} & r_h(s,a,b) - \widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b) \\ &\quad \le \left|r_h(s,a,b) - \widehat{r}_h^k(s,a,b)\right| - \beta_h^{r,k}(s,a,b) \\ &\quad \le \beta_h^{r,k}(s,a,b) - \beta_h^{r,k}(s,a,b) = 0, \end{aligned}$$

where the last inequality is by Lemma B.3 and the setting of the bonus for the reward. Moreover, with probability at least $1-\delta$, we have

$$\begin{aligned} & \left\langle \mathcal{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathcal{P}}_h^k(\cdot|s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \beta_h^{\mathcal{P},k}(s,a,b) \\ &\quad \le \left\|\mathcal{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathcal{P}}_h^k(\cdot|s,a,b)\right\|_1 \left\|\overline{V}_{h+1}^k(\cdot)\right\|_\infty - \beta_h^{\mathcal{P},k}(s,a,b) \\ &\quad \le H\left\|\mathcal{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathcal{P}}_h^k(\cdot|s,a)\right\|_1 - \beta_h^{\mathcal{P},k}(s,a,b) \\ &\quad \le \beta_h^{\mathcal{P},k}(s,a,b) - \beta_h^{\mathcal{P},k}(s,a,b) = 0, \end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality is due to $\max_{s'\in\mathcal{S}} \big\|\overline{V}^k_{h+1}(s')\big\|_\infty \le H$ as shown in (30), and the last inequality is by the setting of $\beta^{\mathcal{P},k}_h$ in (9) and also Lemma B.4. Thus, with probability at least $1-2\delta$, the following inequality holds

$$r_h(s,a,b) - \widehat{r}^k_h(s,a,b) + \big\langle \mathcal{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathcal{P}}^k_h(\cdot|s,a,b), \overline{V}^k_{h+1}(\cdot)\big\rangle_{\mathcal{S}} - \beta^k_h(s,a,b) \le 0.$$

Combining the above inequality with (58), we have that with probability at least $1-2\delta$, for any $h\in[H]$ and $k\in[K]$, the following inequality holds

$$\overline{\iota}^k_h(s,a,b) \le 0, \ \ \forall (s,a,b)\in \mathcal{S}\times\mathcal{A}\times\mathcal{B},$$

which leads to

$$\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big[\overline{\iota}^k_h(s_h,a_h,b_h)\,\big|\,s_1\big] \le 0.$$

This completes the proof. ☐

Lemma B.6. *With probability at least $1-\delta$, Algorithm 1 ensures that*

$$\sum_{k=1}^K \overline{V}^k_1(s_1) - \sum_{k=1}^K V^{\mu^k,\nu^k}_1(s_1) \le \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}_1|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}_2|^2|\mathcal{B}|H^4K} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H^2K}).$$

Proof. We assume that a trajectory $\{(s^k_h,a^k_h,b^k_h,s^k_{h+1})\}_{h=1}^H$ for all $k\in[K]$ is generated according to the policies μ^k, ν^k, and the true transition model $\mathcal{P}$. Thus, we expand the bias term at the h-th step of the k-th episode, which is

$$\begin{aligned}
&\overline{V}^k_h(s^k_h) - V^{\mu^k,\nu^k}_h(s^k_h)\\
&\quad= \big[\mu^k_h(\cdot|s^k_h)\big]^\top\big[\overline{Q}^k_h(s^k_h,\cdot,\cdot) - Q^{\mu^k,\nu^k}_h(s^k_h,\cdot,\cdot)\big]\nu^k_h(\cdot|s^k_h)\\
&\quad= \zeta^k_h + \overline{Q}^k_h(s^k_h,a^k_h,b^k_h) - Q^{\mu^k,\nu^k}_h(s^k_h,a^k_h,b^k_h)\\
&\quad= \zeta^k_h + \big\langle \mathcal{P}_h(\cdot\,|\,s^k_h,a^k_h,b^k_h), \overline{V}^k_{h+1}(\cdot) - V^{\mu^k,\nu^k}_{h+1}(\cdot)\big\rangle_{\mathcal{S}} - \overline{\iota}^k_h(s^k_h,a^k_h,b^k_h)\\
&\quad= \zeta^k_h + \xi^k_h + \overline{V}^k_{h+1}(s^k_{h+1}) - V^{\mu^k,\nu^k}_{h+1}(s^k_{h+1}) - \overline{\iota}^k_h(s^k_h,a^k_h,b^k_h),
\end{aligned}\tag{32}$$

where the first equality is by Line 2 of Algorithm 2 and (1), the third equality is by plugging in (2) and (36). Specifically, in the above equality, we introduce two martingale difference sequence, namely, $\{\zeta^k_h\}_{h\ge 0,k\ge 0}$ and $\{\xi^k_h\}_{h\ge0,k\ge0}$, which are defined as

$$\begin{aligned}
\zeta^k_h &:= \big[\mu^k_h(\cdot|s^k_h)\big]^\top\big[\overline{Q}^k_h(s^k_h,\cdot,\cdot) - Q^{\mu^k,\nu^k}_h(s^k_h,\cdot,\cdot)\big]\nu^k_h(\cdot|s^k_h) - \big[\overline{Q}^k_h(s^k_h,a^k_h,b^k_h) - Q^{\mu^k,\nu^k}_h(s^k_h,a^k_h,b^k_h)\big],\\
\xi^k_h &:= \big\langle \mathcal{P}_h(\cdot\,|\,s^k_h,a^k_h,b^k_h), \overline{V}^k_{h+1}(\cdot) - V^{\mu^k,\nu^k}_{h+1}(\cdot)\big\rangle_{\mathcal{S}} - \big[\overline{V}^k_{h+1}(s^k_{h+1}) - V^{\mu^k,\nu^k}_{h+1}(s^k_{h+1})\big],
\end{aligned}$$

such that

$$\mathbb{E}_{a^k_h\sim\mu^k_h(\cdot|s^k_h),b^k_h\sim\nu^k_h(\cdot|s^k_h)}\big[\zeta^k_h\,\big|\,\mathcal{F}^k_h\big]=0, \qquad \mathbb{E}_{s^k_{h+1}\sim\mathcal{P}_h(\cdot\,|\,s^k_h,a^k_h,b^k_h)}\big[\xi^k_h\,\big|\,\widetilde{\mathcal{F}}^k_h\big]=0,$$

with $\mathcal{F}^k_h$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s^k_h, and $\widetilde{\mathcal{F}}^k_h$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s^k_h, a^k_h, b^k_h.

The equality (32) forms a recursion for $\overline{V}^k_h(s^k_h) - V^{\mu^k,\nu^k}_h(s^k_h)$. We also have $\overline{V}^k_{H+1}(\cdot) = \mathbf{0}$ and $V^{\mu^k,\nu^k}_{H+1}(\cdot)=\mathbf{0}$. Thus, recursively apply (32) from $h=1$ to H leads to the following equality

$$\overline{V}^k_1(s_1) - V^{\mu^k,\nu^k}_1(s_1) = \sum_{h=1}^H \zeta^k_h + \sum_{h=1}^H \xi^k_h - \sum_{h=1}^H \overline{\iota}^k_h(s^k_h,a^k_h,b^k_h).\tag{33}$$

Moreover, by (17) and Line 1 of Algorithm 1, we have

$$-\overline{\iota}_h^k(s_h^k, a_h^k, b_h^k) = -r_h(s_h^k, a_h^k, b_h^k) - \big\langle \mathcal{P}_h(\cdot \,|\, s_h, a_h, b_h), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \min\big\{\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) + \big\langle \widehat{\mathcal{P}}_h^k(\cdot|s_h, a_h, b_h), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \beta_h^k(s_h^k, a_h^k, b_h^k), H\big\}.$$

Then, we can further bound $-\overline{\iota}_h^k(s_h^k, a_h^k, b_h^k)$ as follows

$$\begin{aligned}
-\overline{\iota}_h^k(s_h^k, a_h^k, b_h^k) &\le -r_h(s_h^k, a_h^k, b_h^k) - \big\langle \mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \widehat{r}_h^k(s_h^k, a_h^k, b_h^k) \\
&\quad + \big\langle \widehat{\mathcal{P}}_h^k(\cdot|s_h^k, a_h^k, b_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \beta_h^k(s_h^k, a_h^k, b_h^k) \\
&\le \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| \\
&\quad + \Big|\big\langle \mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \,|\, s_h^k, a_h^k, b_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big| + \beta_h^k(s_h^k, a_h^k, b_h^k),
\end{aligned}$$

where the first inequality is due to $\min\{x, y\} \le x$. Additionally, we have

$$\begin{aligned}
&\Big|\big\langle \mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \,|\, s_h^k, a_h^k, b_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big| \\
&\qquad \le \big\|\overline{V}_{h+1}^k(\cdot)\big\|_\infty \big\|\mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \,|\, s_h^k, a_h^k, b_h^k)\big\|_1 \\
&\qquad \le H\big\|\mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \,|\, s_h^k, a_h^k, b_h^k)\big\|_1,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality and the second inequality is by (57). Thus, putting the above together, we obtain

$$\begin{aligned}
-\overline{\iota}_h^k(s_h^k, a_h^k, b_h^k) &\le \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| + H\big\|\mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k) - \mathcal{P}_h(\cdot \,|\, s_h^k, a_h^k, b_h^k)\big\|_1 + \beta_h^k(s_h^k, a_h^k, b_h^k) \\
&\le 2\beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\beta_h^{\mathcal{P},k}(s_h^k, a_h^k, a_h^k),
\end{aligned}$$

where the second inequality is by Lemma B.3, Lemma B.4, and the decomposition of the bonus term β_h^k as (9). Due to Lemma B.3 and Lemma B.4, by union bound, for any $h \in [H], k \in [K]$ and $(s_h, a_h, b_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, the above inequality holds with probability with probability at least $1 - 2\delta$. Therefore, by (33), with probability at least $1 - 2\delta$, we have

$$\begin{aligned}
&\sum_{k=1}^K \big[\overline{V}_1^k(s_1) - V_1^{\mu^k, \nu^k}(s_1)\big] \\
&\qquad \le \sum_{k=1}^K \sum_{h=1}^H \zeta_h^k + \sum_{k=1}^K \sum_{h=1}^H \xi_h^k + 2\sum_{k=1}^K \sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\sum_{k=1}^K \sum_{h=1}^H \beta_h^{\mathcal{P},k}(s_h^k, a_h^k, b_h^k).
\end{aligned} \tag{34}$$

By Azuma-Hoeffding inequality, with probability at least $1 - \delta$, the following inequalities hold

$$\sum_{k=1}^K \sum_{h=1}^H \zeta_h^k \le \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta}}\right),$$

$$\sum_{k=1}^K \sum_{h=1}^H \xi_h^k \le \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta}}\right),$$

where we use the facts that $|\overline{Q}_h^k(s_h^k, a_h^k, b_h^k) - Q_h^{\mu^k, \nu^k}(s_h^k, a_h^k, b_h^k)| \le 2H$ and $|\overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mu^k, \nu^k}(s_{h+1}^k)| \le 2H$. Next, we need to bound $\sum_{k=1}^K \sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k)$ and $\sum_{k=1}^K \sum_{h=1}^H \beta_h^{\mathcal{P},k}(s_h^k, a_h^k, b_h^k)$ in (34). We show that

$$\begin{aligned}
\sum_{k=1}^K \sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) &= C \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k), 1\}}} \\
&= C \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{N_h^k(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k)}} \\
&\le C \sum_{h=1}^H \sum_{\substack{(s^1, s^2, a, b) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{A} \times \mathcal{B} \\ N_h^K(s^1, s^2, a, b) > 0}} \sum_{n=1}^{N_h^K(s^1, s^2, a, b)} \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}},
\end{aligned}$$

where the second equality is because $(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k)$ is visited such that $N_h^k(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k) \geq 1$. In addition, we have

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{\substack{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B} \\ N_h^K(s^1,s^2,a,b)>0}} \sum_{n=1}^{N_h^K(s^1,s^2,a,b)} \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}} \\
&\quad \leq \sum_{h=1}^{H} \sum_{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}} \mathcal{O}\left(\sqrt{N_h^K(s^1,s^2,a,b)\log\frac{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right) \\
&\quad \leq \mathcal{O}\left(H\sqrt{K|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}_1||\mathcal{S}_2|\mathcal{A}||\mathcal{B}|HK}{\delta}}\right),
\end{aligned}$$

where the last inequality is based on the consideration that $\sum_{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}} N_h^K(s^1,s^2,a,b) = K$ such that $\sum_{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}} \sqrt{N_h^K(s^1,s^2,a,b)} \leq \mathcal{O}\left(\sqrt{K|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|}\right)$ when K is sufficiently large. Putting the above together, we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) \leq \mathcal{O}\left(H\sqrt{K|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}_1||\mathcal{S}_2|\mathcal{A}||\mathcal{B}|HK}{\delta}}\right).$$

Similarly, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{\mathcal{P},k}(s_h^k, a_h^k, b_h^k) &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sqrt{\frac{2H^2|\mathcal{S}_1|\log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s_h^{1,k}, a_h^k), 1\}}} + \sqrt{\frac{2H^2|\mathcal{S}_2|\log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s_h^{2,k}, b_h^k), 1\}}}\right) \\
&\leq \mathcal{O}\left(H\sqrt{K|\mathcal{S}_1|^2|\mathcal{A}|H^2\log\frac{2|\mathcal{S}_1||\mathcal{A}|HK}{\delta}} + H\sqrt{K|\mathcal{S}_2|^2|\mathcal{B}|H^2\log\frac{2|\mathcal{S}_2||\mathcal{B}|HK}{\delta}}\right).
\end{aligned}$$

Thus, by (34), with probability at least $1-\delta$, we have

$$\sum_{k=1}^{K} \overline{V}_1^k(s_1) - \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1) \leq \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}_1|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}_2|^2|\mathcal{B}|H^4K} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H^2K}),$$

where $\widetilde{\mathcal{O}}$ hides logarithmic terms. This completes the proof. □

Before presenting the next lemma, we first show the following definition of confidence set for the proof of the next lemma.

Definition B.7 (Confidence Set for Player 2)**.** *Define the following confidence set for transition models for Player 2*

$$\begin{aligned}
\Upsilon^{2,k} := \Big\{\widetilde{\mathcal{P}} : \Big|\widetilde{\mathcal{P}}_h(s^{2\prime}|s^2,b) - \widehat{\mathcal{P}}_h^{2,k}(s^{2\prime}|s^2,b)\Big| \leq \epsilon_h^{2,k},\ \|\widetilde{\mathcal{P}}_h(\cdot|s^2,b)\|_1 = 1, \\
\text{and } \widetilde{\mathcal{P}}_h(s^{2\prime}|s^2,b) \geq 0,\ \forall(s^2,b,s^{2\prime}) \in \mathcal{S}_2\times\mathcal{B}\times\mathcal{S}_2, \forall k\in[K]\Big\}
\end{aligned}$$

where we define

$$\epsilon_h^{2,k} := 2\sqrt{\frac{\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime}|s^2,b)\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2,b)-1,1\}}} + \frac{14\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{3\max\{N_h^k(s^2,b)-1,1\}}$$

with $N_h^k(s^2,b) := \sum_{\tau=1}^k \mathbf{1}\{(s^2,b) = (s_h^{2,\tau}, b_h^\tau)\}$*, and* $\widehat{\mathcal{P}}^{2,k}$ *being the empirical transition model for Player 2.*

Lemma B.8. *With probability at least* $1-\delta$*, the difference between* $q_h^{\nu^k,\mathcal{P}^2}$ *and* $d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}$ *are bounded as*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\left|q_h^{\nu^k,\mathcal{P}^2}(s^2) - d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right| \leq \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}\right).$$

Proof. By the definition of state distribution for Player 2, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\left|q_h^{\nu^k,\mathcal{P}^2}(s^2)-d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right| = \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\left|\sum_{b\in\mathcal{B}}w_h^{2,k}(s^2,b)-\sum_{b\in\mathcal{B}}\widehat{w}_h^{2,k}(s^2,b)\right|$$
$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\sum_{b\in\mathcal{B}}\left|w_h^{2,k}(s,a)-\widehat{w}_h^{2,k}(s^2,b)\right|.$$

where $\widehat{w}_h^{2,k}(s^2,b)$ is the occupancy measure under the empirical transition model $\widehat{\mathcal{P}}^{2,k}$ and the policy ν^k. Then, since $\widehat{\mathcal{P}}^{2,k}\in\Upsilon^{2,k}$ always holds for any k, by Lemma B.11, we can bound the last term of the bound inequality such that with probability at least $1-6\delta'$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\left|q_h^{\nu^k,\mathcal{P}^2}(s^2)-d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right|\leq\mathcal{E}_1+\mathcal{E}_2.$$

Then, we compute $\mathcal{E}_1$ by Lemma B.10. With probability at least $1-2\delta'$, we have

$$\begin{aligned}
\mathcal{E}_1 &= \mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sum_{k=1}^{K}\sum_{s^2\in\mathcal{S}_2}\sum_{b\in\mathcal{B}}w_h^k(s^2,b)\left(\sqrt{\frac{|\mathcal{S}_2|\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2,b),1\}}}+\frac{\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2,b),1\}}\right)\right]\\
&= \mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sqrt{|\mathcal{S}_2|}\left(\sqrt{|\mathcal{S}_2||\mathcal{B}|K}+|\mathcal{S}_2||\mathcal{B}|\log K+\log\frac{H}{\delta'}\right)\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}\right]\\
&= \mathcal{O}\left[\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}+H^2|\mathcal{S}_2|^{3/2}|\mathcal{B}|\log K+H^2\sqrt{|\mathcal{S}_2|}\log\frac{H}{\delta'}\right)\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}\right]\\
&= \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}\right),
\end{aligned}$$

where we ignore $\log K$ when K is sufficiently large such that $\sqrt{K}$ dominates, and $\widetilde{\mathcal{O}}$ hides logarithm dependence on $|\mathcal{S}_2|$, $|\mathcal{B}|$, H, K, and $1/\delta'$. In addition, $\mathcal{E}_2$ depends on $\mathrm{ploy}(H,|\mathcal{S}_2|,|\mathcal{B}|)$ except the factor $\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}$ as shown in Lemma B.11. Thus, $\mathcal{E}_2$ can be ignored comparing to $\mathcal{E}_1$ if K is sufficiently large. Therefore, we obtain that with probability at least $1-8\delta'$, the following inequality holds

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\left|q_h^{\nu^k,\mathcal{P}^2}(s^2)-d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\right|\leq\widetilde{\mathcal{O}}\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}\right).$$

We further let $\delta=8\delta'$ such that $\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}=\log\frac{8|\mathcal{S}_2||\mathcal{B}|HK}{\delta}$ which does not change the order as above. Then, with probability at least $1-\delta$, we have $\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}|q_h^{\nu^k,\mathcal{P}^2}(s^2)-d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)|\leq\widetilde{\mathcal{O}}(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K})$. This completes the proof. □

B.1. Other Supporting Lemmas

The following lemmas are adapted from the recent papers (Efroni et al., 2020; Jin & Luo, 2019), where we can find their detailed proofs.

Lemma B.9. *With probability at least $1-4\delta'$, the true transition model $\mathcal{P}^2$ satisfies that for any $k\in[K]$,*

$$\mathcal{P}\in\Upsilon^{2,k}.$$

This lemma indicates that the estimated transition model $\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime}|s^2,b)$ for Player 2 by (11) is closed to the true transition model $\mathcal{P}_h^2(s^{2\prime}|s^2,b)$ with high probability. The upper bound is by empirical Bernstein's inequality and the union bound.

The next lemma is adapted from Lemma 10 in Jin & Luo (2019).

Lemma B.10. *We let $w_h^{2,k}(s^2, b)$ denote the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathcal{P}^2$ and the current policy ν^k. Then, with probability at least $1 - 2\delta'$ we have for all $h \in [H]$, the following inequalities hold*

$$\sum_{k=1}^{K} \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \frac{w_h^k(s^2, b)}{\max\{N_h^k(s^2, b), 1\}} = \mathcal{O}\left(|\mathcal{S}_2||\mathcal{B}| \log K + \log \frac{H}{\delta'}\right),$$

and

$$\sum_{k=1}^{K} \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \frac{w_h^k(s^2, b)}{\sqrt{\max\{N_h^k(s^2, b), 1\}}} = \mathcal{O}\left(\sqrt{|\mathcal{S}_2||\mathcal{B}|K} + |\mathcal{S}_2||\mathcal{B}| \log K + \log \frac{H}{\delta'}\right).$$

By Lemma B.9 and Lemma B.10, we have the following lemma to show the difference of two occupancy measures, which is modified from parts of the proof of Lemma 4 in Jin & Luo (2019).

Lemma B.11. *For Player 2, we let $w_h^{2,k}(s^2, b)$ be the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathcal{P}^2$ and the current policy ν^k, and $\widetilde{w}_h^{2,k}(s^2, b)$ be the occupancy measure at the h-th step of the k-th episode under any transition model $\widetilde{\mathcal{P}}^{2,k} \in \Upsilon^k$ and the current policy ν^k for any k. Then, with probability at least $1 - 6\delta'$ we have for all $h \in [H]$, the following inequalities hold*

$$\sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{s \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \left|\widetilde{w}_h^{2,k}(s^2, b) - w_h^{2,k}(s^2, b)\right| \leq \mathcal{E}_1 + \mathcal{E}_2,$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are in the level of

$$\mathcal{E}_1 = \mathcal{O}\left[\sum_{h=2}^{H} \sum_{h'=1}^{h-1} \sum_{k=1}^{K} \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} w_h^k(s^2, b) \left(\sqrt{\frac{|\mathcal{S}_2| \log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2, b), 1\}}} + \frac{\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2, b), 1\}}\right)\right]$$

and

$$\mathcal{E}_2 = \mathcal{O}\left(\text{poly}(H, |\mathcal{S}_2|, |\mathcal{B}|) \cdot \log \frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}\right),$$

where $\text{poly}(H, |\mathcal{S}_2|, |\mathcal{B}|)$ denotes the polynomial dependency on $H, |\mathcal{S}_2|, |\mathcal{B}|$.

C. Proofs for Section 4

Lemma C.1. *At the k-th episode, the difference between value functions $V_1^{\mu^*,\nu^k}(s_1)$ and $V_1^{\mu^k,\nu^k}(s_1)$ is*

$$\begin{aligned} V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1) &= \overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1) \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}\left[\left\langle \mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h), U_h^k(s_h, \cdot)\right\rangle_{\mathcal{A}} \,\middle|\, s_1\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\left[\overline{\varsigma}_h^k(s_h, a_h, b_h) \,\middle|\, s_1\right]. \end{aligned} \tag{35}$$

where s_h, a_h, b_h are random variables for state and actions, $U_h^k(s, a) := \langle \overline{Q}_h^k(s, a, \cdot), \nu_h^k(\cdot \mid s)\rangle_{\mathcal{B}}$, and we define the model prediction error of Q-function as

$$\overline{\varsigma}_h^k(s, a, b) = r_h(s, a, b) + \mathcal{P}_h \overline{V}_{h+1}^k(s, a) - \overline{Q}_h^k(s, a, b). \tag{36}$$

Proof. We start the proof by decomposing the value function difference as

$$V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1) = V_1^{\mu^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) + \overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1). \tag{37}$$

Note that the term $\overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1)$ is the bias between the estimated value function $\overline{V}_1^k(s_1)$ generated by Algorithm 2 and the value function $V_1^{\mu^k,\nu^k}(s_1)$ under the true transition model $\mathcal{P}$ at the k-th episode.

We focus on analyzing the other term $V_1^{\mu^*,\nu^k}(s_1) - \overline{V}_1^k(s_1)$ in this proof. For any h and s, we have

$$
\begin{aligned}
&V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s) \\
&\qquad = [\mu_h^*(\cdot|s)]^\top Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad = [\mu_h^*(\cdot|s)]^\top Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - [\mu_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad\quad + [\mu_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad = [\mu_h^*(\cdot|s)]^\top \big[Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \\
&\qquad\quad + \big[\mu_h^*(\cdot|s) - \mu_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s),
\end{aligned} \tag{38}
$$

where the first inequality is by the definition of $V_h^{\mu^*,\nu^k}$ in (1) and the definition of $\overline{V}_h^k$ in Line 2 of Algorithm 2. Moreover, by the definition of $Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot)$ in (2) and the model prediction error $\overline{\varsigma}_h^k$ for Player one in (36), we have

$$
\begin{aligned}
&[\mu_h^*(\cdot|s)]^\top \big[Q_h^{\mu^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \\
&\quad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\bigg[\sum_{s'\in\mathcal{S}} \mathcal{P}_h(s'|s,a)\big[V_{h+1}^{\mu^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] + \overline{\varsigma}_h^k(s,a,b)\bigg]\nu_h^k(b|s) \\
&\quad = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \mu_h^*(a|s)\mathcal{P}_h(s'|s,a)\big[V_{h+1}^{\mu^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\overline{\varsigma}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned}
$$

where the last equality holds due to $\sum_{b\in\mathcal{B}} \nu_h^k(b\,|\,s) = 1$. Combining this equality with (38) gives

$$
\begin{aligned}
V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s) &= \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \mu_h^*(a|s)\mathcal{P}_h(s'|s,a)\big[V_{h+1}^{\mu^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] \\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^*(a|s)\overline{\varsigma}_h^k(s,a,b)\nu_h^k(b|s) \\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \big[\mu_h^*(a|s) - \mu_h^k(a|s)\big]\overline{Q}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned} \tag{39}
$$

Note that (39) indicates a recursion of the value function difference $V_h^{\mu^*,\nu^k}(s) - \overline{V}_h^k(s)$. Since we define $V_{H+1}^{\mu^*,\nu^k}(s) = 0$ and $\overline{V}_{H+1}^k(s) = 0$, by recursively applying (39) from $h = 1$ to H, we obtain

$$
\begin{aligned}
&V_1^{\mu^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) \\
&\qquad = \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}}\big\{[\mu_h^*(\cdot|s_h)]^\top \overline{\varsigma}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\,s_1\big\} \\
&\qquad\quad + \sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}}\big\{\big[\mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\,s_1\big\},
\end{aligned} \tag{40}
$$

where s_h are a random variables denoting the state at the h-th step following a distribution determined jointly by $\mu^*, \mathcal{P}$.

Further combining (40) with (37), we eventually have

$$
\begin{aligned}
&V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1) \\
&\quad = \overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1) + \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}\big\{[\mu_h^*(\cdot|s_h)]^\top \overline{\varsigma}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,\big|\, s_1\big\} \\
&\qquad + \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}\big\{\big[\mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,\big|\, s_1\big\} \\
&\quad = \overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1) + \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big[\overline{\varsigma}_h^k(s_h,a_h,b_h) \,\big|\, s_1\big] \\
&\qquad + \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}\Big[\big\langle \mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h), U_h^k(s_h,\cdot)\big\rangle_{\mathcal{A}} \,\Big|\, s_1\Big],
\end{aligned}
$$

where s_h, a_h, b_h are a random variables denoting the state and actions at the h-th step following a distribution determined jointly by $\mu^*, \mathcal{P}, \nu^k$, and $U_h^{k-1}(s,a) := \langle \overline{Q}_h^{k-1}(s,a,\cdot), \nu_h^{k-1}(\cdot\,|\,s)\rangle_{\mathcal{B}}$. This completes our proof. □

Lemma C.2. *At the k-th episode, with probability at least $1-2\delta$, the difference between the value functions $V_1^{\mu^k,\nu^k}(s_1)$ and $V_1^{\mu^k,\nu^*}(s_1)$ is bound as*

$$
\begin{aligned}
V_1^{\mu^k,\nu^k}(s_1) - V_1^{\mu^k,\nu^*}(s_1) &\le 2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \Big| q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s) \Big| \\
&\quad + \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^k(s) \big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} \\
&\quad + 2\sum_{h=1}^{H} \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\big[\beta_h^{r,k}(s_h,a_h,b_h) \,\big|\, s_1\big],
\end{aligned} \tag{41}
$$

where s_h, a_h, b_h are random variables for state and actions, and $W_h^k(s,b) = \langle \widetilde{r}_h^k(s,\cdot,b), \mu_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}}$.

Proof. We start our proof from analyzing the difference for any h and s as follows

$$
\begin{aligned}
&V_h^{\mu^k,\nu^k}(s) - V_h^{\mu^k,\nu^*}(s) \\
&\quad = \big[\mu_h^k(\cdot|s)\big]^\top Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top Q_h^{\mu^k,\nu^*}(s,\cdot,\cdot)\nu_h^*(\cdot|s) \\
&\quad = \big[\mu_h^k(\cdot|s)\big]^\top Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot)\nu_h^*(\cdot|s) \\
&\qquad + [\mu_h^k(\cdot|s)]^\top Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot)\nu_h^*(\cdot|s) - \big[\mu_h^k(\cdot|s)\big]^\top Q_h^{\mu^k,\nu^*}(s,\cdot,\cdot)\nu_h^*(\cdot|s) \\
&\quad = \big[\mu_h^k(\cdot|s)\big]^\top Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big] \\
&\qquad + \big[\mu_h^k(\cdot|s)\big]^\top \big[Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot) - Q_h^{\mu^k,\nu^*}(s,\cdot,\cdot)\big]\nu_h^*(\cdot|s),
\end{aligned} \tag{42}
$$

where the first equality is by the Bellman equation for $V_h^{\mu,\nu}(s)$ in (1). Moreover, by the Bellman equation for $Q_h^{\mu,\nu}$ in (2), we can expand the last term in (42) as

$$
\begin{aligned}
&\big[\mu_h^k(\cdot|s)\big]^\top \big[Q_h^{\mu^k,\nu^k}(s,\cdot,\cdot) - Q_h^{\mu^k,\nu^*}(s,\cdot,\cdot)\big]\nu_h^*(\cdot|s) \\
&\quad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_h^k(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}_h(s'|s,a)\big[V_{h+1}^{\mu^k,\nu^k}(s') - V_{h+1}^{\mu^k,\nu^*}(s')\big]\nu_h^*(b|s) \\
&\quad = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \mu_h^k(a|s)\mathcal{P}_h(s'|s,a)\big[V_{h+1}^{\mu^k,\nu^k}(s') - V_{h+1}^{\mu^k,\nu^*}(s')\big].
\end{aligned} \tag{43}
$$

where the last equality holds due to $\sum_{b\in\mathcal{B}}\nu_h^*(b\,|\,s)=1$. Combining (43) with (42) gives

$$\begin{aligned} V_h^{\mu^k,\nu^k}(s)-V_h^{\mu^k,\nu^*}(s)&=\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\mu_h^k(a|s)Q_h^{\mu^k,\nu^k}(s,a,b)\big[\nu_h^k(b|s)-\nu_h^*(b|s)\big]\\ &\quad+\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\mu_h^k(a|s)\mathcal{P}_h(s'|s,a)\big[V_{h+1}^{\mu^k,\nu^k}(s')-V_{h+1}^{\mu^k,\nu^*}(s')\big]. \end{aligned} \tag{44}$$

Note that (44) indicates a recursion of the value function difference $V_h^{\mu^k,\nu^k}(s)-V_h^{\mu^k,\nu^*}(s)$. Since we define $V_{H+1}^{\mu,\nu}(s)=0$ for any μ and ν, by recursively applying (44) from $h=1$ to H, we obtain

$$V_1^{\mu^k,\nu^k}(s_1)-V_1^{\mu^k,\nu^*}(s_1)=\sum_{h=1}^H\mathbb{E}_{\mu^k,\mathcal{P}}\big\{\big[\mu_h^k(\cdot|s_h)\big]^\top Q_h^{\mu^k,\nu^k}(s_h,\cdot,\cdot)\big[\nu_h^k(\cdot|s_h)-\nu_h^*(\cdot|s_h)\big]\,\big|\,s_1\big\}, \tag{45}$$

where s_h are a random variables following a distribution determined jointly by $\mu^k,\mathcal{P}$. Note that since we have defined the distribution of s_h under μ^k and $\mathcal{P}$ as

$$q_h^{\mu^k,\mathcal{P}}(s)=\Pr\big(s_h=s\,\big|\,\mu^k,\mathcal{P},s_1\big),$$

we can rewrite (45) as

$$V_1^{\mu^k,\nu^k}(s_1)-V_1^{\mu^k,\nu^*}(s_1)=\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}q_h^{\mu^k,\mathcal{P}}(s)\mu_h^k(a|s)Q_h^{\mu^k,\nu^k}(s,a,b)\big[\nu_h^k(b|s)-\nu_h^*(b|s)\big]. \tag{46}$$

By plugging the Bellman equation for Q-function as (2) into (46), we further expand (46) as

$$\begin{aligned} &V_1^{\mu^k,\nu^k}(s_1)-V_1^{\mu^k,\nu^*}(s_1)\\ &\quad=\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}q_h^{\mu^k,\mathcal{P}}(s)\mu_h^k(a|s)\big[r_h(s,a,b)+\big\langle\mathcal{P}_h(\cdot|s,a),V_{h+1}^{\mu^k,\nu^k}(\cdot)\big\rangle\big][\nu_h^k(b|s)-\nu_h^*(b|s)]\\ &\quad=\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}q_h^{\mu^k,\mathcal{P}}(s)\mu_h^k(a|s)\,[r_h(s,a,b)]\,[\nu_h^k(b|s)-\nu_h^*(b|s)]\\ &\quad=\sum_{h=1}^H\sum_{s\in\mathcal{S}}q_h^{\mu^k,\mathcal{P}}(s)[\mu_h^k(\cdot|s)]^\top r_h(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s)-\nu_h^*(\cdot|s)\big], \end{aligned}$$

where the second equality by

$$\begin{aligned} &\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}q_h^{\mu^k,\mathcal{P}}(s)\mu_h^k(a|s)\big\langle\mathcal{P}_h(\cdot|s,a),V_{h+1}^{\mu^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}}[\nu_h^k(b|s)-\nu_h^*(b|s)]\\ &\qquad=\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}q_h^{\mu^k,\mathcal{P}}(s)\mu_h^k(a|s)\big\langle\mathcal{P}_h(\cdot|s,a),V_{h+1}^{\mu^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}}\sum_{b\in\mathcal{B}}[\nu_h^k(b|s)-\nu_h^*(b|s)]\\ &\qquad=0. \end{aligned}$$

In particular, the last equality above is due to

$$\sum_{b\in\mathcal{B}}\big[\nu_h^k(b|s)-\nu_h^*(b|s)\big]=1-1=0.$$

Thus, we have

$$V_1^{\mu^k,\nu^k}(s_1)-V_1^{\mu^k,\nu^*}(s_1)=\sum_{h=1}^H\sum_{s\in\mathcal{S}}q_h^{\mu^k,\mathcal{P}}(s)\big[\mu_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s)-\nu_h^*(\cdot|s)\big]. \tag{47}$$

Now we define the following term associated with estimation $\widehat{\mathcal{P}}^k$, $\widehat{r}^h$, policies μ^k, ν^k, and the initial state s_1 as

$$\underline{V}_1^k(s_1) := \sum_{h=1}^H \sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s),$$

with $\widetilde{r}$ defined in Line 3 of Algorithm 3, which is

$$\widetilde{r}_h^k(s,a,b) = \max\big\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b),\ 0\big\}.$$

Thus, we have the following decomposition

$$\begin{aligned}
&V_1^{\mu^k,\nu^k}(s_1) - V_1^{\mu^k,\nu^*}(s_1)\\
&\quad = V_1^{\mu^k,\nu^k}(s_1) - \underline{V}_1^k(s_1) + \underline{V}_1^k(s_1) - V_1^{\mu^k,\nu^*}(s_1)\\
&\quad = \underbrace{\sum_{h=1}^H \sum_{s\in\mathcal{S}} \Big\{ q_h^{\mu^k,\mathcal{P}}(s)\big[\mu_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^k(\cdot|s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\Big\}}_{\text{Term(I)}}\\
&\quad\quad + \underbrace{\sum_{h=1}^H \sum_{s\in\mathcal{S}} \Big\{ q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) - q_h^{\mu^k,\mathcal{P}}(s)\big[\mu_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^*(\cdot|s)\Big\}}_{\text{Term(II)}}.
\end{aligned} \tag{48}$$

We first bound Term(I) as

$$\begin{aligned}
\text{Term(I)} &= \sum_{h=1}^H \sum_{s\in\mathcal{S}} \Big\{ q_h^{\mu^k,\mathcal{P}}(s)\big[\mu_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^k(\cdot|s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\Big\}\\
&= \sum_{h=1}^H \sum_{s\in\mathcal{S}} q_h^{\mu^k,\mathcal{P}}(s)\big[\mu_h^k(\cdot|s)\big]^\top \big[r_h(s,\cdot,\cdot) - \widetilde{r}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s)\\
&\quad + \sum_{h=1}^H \sum_{s\in\mathcal{S}} \Big[q_h^{\mu^k,\mathcal{P}}(s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\Big]\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\\
&\le 2\sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\big[\beta_h^{r,k}(s,a,b)\big] + \sum_{h=1}^H \sum_{s\in\mathcal{S}} \Big|q_h^{\mu^k,\mathcal{P}}(s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\Big|,
\end{aligned}$$

where the inequality is due to $|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)| \le \beta_h^{r,k}(s,a,b)$ with probability at least $1-\delta$ because of Lemma C.4 such that we have

$$\begin{aligned}
r_h(s,a,b) - \widetilde{r}_h^k(s,a,b) &= r_h(s,a,b) - \max\big\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b), 0\big\}\\
&= \min\big\{r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \beta_h^{r,k}(s,a,b), r_h(s,a,b)\big\}\\
&\le r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \beta_h^{r,k}(s,a,b) \le 2\beta_h^{r,k}(s,a,b),
\end{aligned}$$

which yields

$$\sum_{s\in\mathcal{S}} q_h^{\mu^k,\mathcal{P}}(s)\big[\mu_h^k(\cdot|s)\big]^\top \big[r_h(s,\cdot,\cdot) - \widetilde{r}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \le 2\mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\big[\beta_h^{r,k}(s,a,b)\big],$$

and we also have

$$\begin{aligned}
\Big|\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\Big| &\le \Big|\sum_a\sum_b \mu_h^k(a|s)\widetilde{r}_h^k(s,a,b)\nu_h^k(b|s)\Big|\\
&\le \sum_a\sum_b \mu_h^k(a|s)\cdot\big|\widetilde{r}_h^k(s,a,b)\big|\cdot\nu_h^k(b|s) \le 1,
\end{aligned}$$

because of $\widetilde{r}_h^k(s,a,b) = \max\big\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b), 0\big\} \le \widehat{r}_h^k(s,a,b) \le 1$. Therefore, with probability at least $1-\delta$, we have

$$\text{Term(I)} \le 2\sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\big[\beta_h^{r,k}(s_h,a_h,b_h)\big] + \sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\mu^k,\mathcal{P}}(s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\Big|. \tag{49}$$

Next, we bound Term(II) in the following way

$$\begin{aligned}\text{Term(II)} &= \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big]\\ &\quad + \sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big[q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s) - q_h^{\mu^k,\mathcal{P}}(s)\Big]\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\\ &\quad + \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top\big[\widetilde{r}_h^k(s,\cdot,\cdot) - r_h(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s).\end{aligned}$$

Here the first term in the above equality is associated with the mirror descent step in Algorithm 3. The second term can be similarly bounded by $\sum_{h=1}^H\sum_{s\in\mathcal{S}}|q_h^{\mu^k,\mathcal{P}}(s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)|$. With probability at least $1-\delta$, the third term is bounded as

$$\begin{aligned}&\sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)[\mu_h^k(\cdot|s)]^\top\big[\widetilde{r}_h^k(s,\cdot,\cdot) - r_h(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s)\\ &\quad= \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\sum_{a,b}\mu_h^k(a|s)\big[\widetilde{r}_h^k(s,a,b) - r_h(s,a,b)\big]\nu_h^k(b|s)\\ &\quad= \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\sum_{a,b}\mu_h^k(a|s)\max\big\{\widehat{r}_h^{k-1}(s,a,b) - r_h(s,a,b) - \beta_h^{r,k-1}, -r_h(s,a,b)\big\}\nu_h^k(b|s)\\ &\quad\le 0,\end{aligned}$$

since $\widehat{r}_h^{k-1}(s,a,b) - r_h(s,a,b) - \beta_h^{r,k-1} \le 0$ with probability at least $1-\delta$ by Lemma C.4, which reflects the 'optimism' of the algorithm. Thus, with probability at least $1-\delta$, we have

$$\begin{aligned}\text{Term(II)} &\le \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big]\\ &\quad + \sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\mu^k,\mathcal{P}}(s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\Big|.\end{aligned} \tag{50}$$

Combining (49), (50) with (48), we obtain that with probability at least $1-2\delta$, the following inequality holds

$$\begin{aligned}V_1^{\mu^k,\nu^k}(s_1) - V_1^{\mu^k,\nu^*}(s_1) &\le \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\big[\mu_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big]\\ &\quad + 2\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\mu^k,\mathcal{P}}(s) - q_h^{\mu^k,\widehat{\mathcal{P}}^k}(s)\Big| + 2\sum_{h=1}^H\mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\big[\beta_h^{r,k}(s_h,a_h,b_h)\big].\end{aligned}$$

This completes our proof. □

Lemma C.3. *With setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, the mirror ascent steps of Algorithm 2 lead to*

$$\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P}}\Big[\big\langle \mu_h^*(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big] \le \mathcal{O}\Big(\sqrt{H^4K\log|\mathcal{A}|}\Big).$$

Proof. As shown in (13), the mirror ascent step at the k-th episode is to solve the following maximization problem

$$\underset{\mu\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{maximize}} \sum_{h=1}^{H} \big\langle \mu_h(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} - \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\mu_h(\cdot|s), \mu_h^k(\cdot|s)\big),$$

with $U_h^k(s,a) = \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot|s)\rangle_{\mathcal{B}}$. We can further equivalently rewrite this maximization problem to a minimization problem as

$$\underset{\mu\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{minimize}} -\sum_{h=1}^{H} \big\langle \mu_h(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} + \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\mu_h(\cdot|s), \mu_h^k(\cdot|s)\big).$$

Note that the closed-form solution $\mu_h^{k+1}(a|s) = (Y_h^k)^{-1}\mu_h^k(a\,|\,s)\exp\{\eta\langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot\,|\,s)\rangle_{\mathcal{B}}\}$ to this minimization problem is guaranteed to stay in the relative interior of its feasible set $\Delta(\mathcal{A}\,|\,\mathcal{S},H)$ when initialize $\mu_h^0(\cdot|s) = \mathbf{1}/|\mathcal{A}|$. Thus, we can apply Lemma C.12 and obtain that for any $\mu = \{\mu_h\}_{h=1}^H$, the following inequality holds

$$\begin{aligned}
&-\eta\big\langle \mu_h^{k+1}(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} + \eta\big\langle \mu_h(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\quad \le D_{\mathrm{KL}}\big(\mu_h(\cdot|s), \mu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h(\cdot|s), \mu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s), \mu_h^k(\cdot|s)\big).
\end{aligned}$$

Then, by rearranging the terms, we have

$$\begin{aligned}
&\eta\big\langle \mu_h^*(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\quad \le D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s), \mu_h^k(\cdot|s)\big) \\
&\qquad + \eta\big\langle \mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}.
\end{aligned} \tag{51}$$

Due to Pinsker's inequality, we have

$$-D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s), \mu_h^k(\cdot|s)\big) \le -\frac{1}{2}\big\|\mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s)\big\|_1^2.$$

Moreover, by Cauchy-Schwarz inequality, we have

$$\eta\big\langle \mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \le \eta H\big\|\mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s)\big\|_1.$$

Thus, we have

$$\begin{aligned}
&-D_{\mathrm{KL}}\big(\mu_h^{k+1}(\cdot|s), \mu_h^k(\cdot|s)\big) + \eta\big\langle \mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\quad \le -\frac{1}{2}\big\|\mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s)\big\|_1^2 + \eta H\big\|\mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s)\big\|_1 \\
&\quad \le \frac{1}{2}\eta^2 H^2,
\end{aligned} \tag{52}$$

where the last inequality is by viewing $\big\|\mu_h^{k+1}(\cdot|s) - \mu_h^k(\cdot|s)\big\|_1$ as a variable x and finding the maximal value of $-1/2\cdot x^2 + \eta H x$ to obtain the upper bound $1/2\cdot\eta^2 H^2$.

Thus, combing (52) with (51), the policy improvement step in Algorithm 2 implies

$$\eta\big\langle \mu_h^*(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \le D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^{k+1}(\cdot|s)\big) + \frac{1}{2}\eta^2 H^2,$$

which further leads to

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}\Big[\big\langle \mu_h^*(\cdot|s) - \mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big] \\
&\quad \le \frac{1}{\eta}\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s), \mu_h^{k+1}(\cdot|s)\big)\big] + \frac{1}{2}\eta H^3.
\end{aligned}$$

Moreover, we take summation from $k=1$ to K of both sides and then obtain

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P}}\Big[\big\langle \mu_h^*(\cdot|s)-\mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big]\\
&\qquad \le \frac{1}{\eta}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P}}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s),\mu_h^1(\cdot|s)\big)-D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s),\mu_h^{K+1}(\cdot|s)\big)\big]+\frac{1}{2}\eta K H^3\\
&\qquad \le \frac{1}{\eta}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P}}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s),\mu_h^1(\cdot|s)\big)\big]+\frac{1}{2}\eta K H^3,
\end{aligned}$$

where the last inequality is non-negativity of KL divergence. By the initialization in Algorithm 2, it is guaranteed that $\mu_h^1(\cdot|s)=\mathbf{1}/|\mathcal{A}|$, which thus leads to $D_{\mathrm{KL}}\left(\mu_h^*(\cdot|s),\mu_h^1(\cdot|s)\right)\le \log|\mathcal{A}|$. Then, with setting $\eta=\sqrt{\log|\mathcal{A}|/(KH^2)}$, we bound the last term as

$$\frac{1}{\eta}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P}}\big[D_{\mathrm{KL}}\big(\mu_h^*(\cdot|s),\mu_h^1(\cdot|s)\big)\big]+\frac{1}{2}\eta K H^3\le \mathcal{O}\left(\sqrt{H^4K\log|\mathcal{A}|}\right),$$

which gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P}}\Big[\big\langle \mu_h^*(\cdot|s)-\mu_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big]\le \mathcal{O}\left(\sqrt{H^4K\log|\mathcal{A}|}\right),$$

This completes the proof. □

Lemma C.4. *For any $k\in[K]$, $h\in[H]$ and all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, with probability at least $1-\delta$, we have*

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}.$$

Proof. The proof for this theorem is a direct application of Hoeffding's inequality. For $k\ge 1$, the definition of $\widehat{r}_h^k$ in (11) indicates that $\widehat{r}_h^k(s,a,b)$ is the average of $N_h^k(s,a,b)$ samples of the observed rewards at (s,a,b) if $N_h^k(s,a,b)>0$. Then, for fixed $k\in[K],h\in[H]$ and state-action tuple $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, when $N_h^k(s,a,b)>0$, according to Hoeffding's inequality, with probability at least $1-\delta'$ where $\delta'\in(0,1]$, we have

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{\log(2/\delta')}{2N_h^k(s,a,b)}},$$

where we also use the facts that the observed rewards $r_h^k\in[0,1]$ for all k and h, and $\mathbb{E}\big[\widehat{r}_h^k\big]=r_h$ for all k and h. For the case where $N_h^k(s,a,b)=0$, by (11), we know $\widehat{r}_h^k(s,a,b)=0$ such that $|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)|=|r_h(s,a,b)|\le 1$. On the other hand, we have $\sqrt{2\log(2/\delta')}\ge 1>|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)|$. Thus, combining the above results, with probability at least $1-\delta'$, for fixed $k\in[K],h\in[H]$ and state-action tuple $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, we have

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{2\log(2/\delta')}{\max\{N_h^k(s,a,b),1\}}}.$$

Moreover, by the union bound, letting $\delta=|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK\delta'/2$, assuming $K>1$, with probability at least $1-\delta$, for any $k\in[K],h\in[H]$ and any state-action tuple $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, we have

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}.$$

This completes the proof. □

In (12), we set $\beta_h^{r,k}(s,a,b) = \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}$, which equals the bound in Lemma C.4.

Lemma C.5. *For any $k \in [K]$, $h \in [H]$ and all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$, we have*

$$\left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1 \le \sqrt{\frac{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}}.$$

Proof. For $k \ge 1$, we have $\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\|_1 = \max_{\|\mathbf{z}\|_\infty \le 1} \langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}\rangle_{\mathcal{S}}$ by the duality. We construct an ε-covering net for the set $\{\mathbf{z} \in \mathbb{R}^{|\mathcal{S}|} : \|\mathbf{z}\|_\infty \le 1\}$ with the distance induced by $\|\cdot\|_\infty$, denoted as $\mathcal{N}_\varepsilon$, such that for any $\mathbf{z} \in \mathbb{R}^{|\mathcal{S}|}$, there always exists $\mathbf{z}' \in \mathcal{N}_\varepsilon$ satisfying $\|\mathbf{z} - \mathbf{z}'\|_\infty \le \varepsilon$. The covering number is $|\mathcal{N}_\varepsilon| = 1/\varepsilon^{|\mathcal{S}|}$. Thus, we know that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $\mathbf{z}$ with $\|\mathbf{z}\|_\infty \le 1$, there exists $\mathbf{z}' \in \mathcal{N}_\varepsilon$ such that $\|\mathbf{z}' - \mathbf{z}\|_\infty \le \varepsilon$ and

$$\begin{aligned}
&\left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}\right\rangle_{\mathcal{S}} \\
&\quad = \left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}'\right\rangle_{\mathcal{S}} + \left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z} - \mathbf{z}'\right\rangle_{\mathcal{S}} \\
&\quad \le \left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}'\right\rangle_{\mathcal{S}} + \varepsilon \left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1,
\end{aligned}$$

such that we further have

$$\begin{aligned}
&\left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1 \\
&\quad = \max_{\|\mathbf{z}\|_\infty \le 1} \left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}\right\rangle_{\mathcal{S}} \\
&\quad \le \max_{\mathbf{z}' \in \mathcal{N}_\varepsilon} \left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}'\right\rangle_{\mathcal{S}} + \varepsilon \left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1.
\end{aligned} \tag{53}$$

By Hoeffding's inequality and union bound over all $\mathbf{z}' \in \mathcal{N}_\varepsilon$, when $N_h^k(s,a) > 0$, with probability at least $1-\delta'$ where $\delta' \in (0,1]$,

$$\max_{\mathbf{z}' \in \mathcal{N}_\varepsilon} \left\langle \widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a), \mathbf{z}'\right\rangle_{\mathcal{S}} \le \sqrt{\frac{|\mathcal{S}|\log(1/\varepsilon) + \log(1/\delta')}{2N_h^k(s,a)}}. \tag{54}$$

Letting $\varepsilon = 1/2$, by (53) and (54), with probability at least $1-\delta'$, we have

$$\left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1 \le 1\sqrt{\frac{|\mathcal{S}|\log 2 + \log(1/\delta')}{2N_h^k(s,a)}}.$$

When $N_h^k(s,a) = 0$, we have $\left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1 = \|\mathcal{P}_h(\cdot \mid s,a)\|_1 = 1$ such that $2\sqrt{\frac{|\mathcal{S}|\log 2 + \log(1/\delta')}{2}} > 1 = \left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1$ always holds. Thus, with probability at least $1-\delta'$,

$$\left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1 \le 2\sqrt{\frac{|\mathcal{S}|\log 2 + \log(1/\delta')}{2\max\{N_h^k(s,a),1\}}} \le \sqrt{\frac{2|\mathcal{S}|\log(2/\delta')}{\max\{N_h^k(s,a),1\}}}.$$

Then, by union bound, assuming $K > 1$, letting $\delta = |\mathcal{S}||\mathcal{A}|HK\delta'/2$, with probability at least $1-\delta$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $h \in [H]$ and $k \in [K]$, we have

$$\left\|\widehat{\mathcal{P}}_h^k(\cdot \mid s,a) - \mathcal{P}_h(\cdot \mid s,a)\right\|_1 \le \sqrt{\frac{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}},$$

This completes the proof. □

In (12), we set $\beta_h^{\mathcal{P},k}(a,b) = \sqrt{\frac{2H^2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}}$, which equals the product of the upper bound in Lemma C.5 and the factor H.

Lemma C.6. *With probability at least $1-2\delta$, Algorithm 2 ensures that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big[\overline{\varsigma}_h^k(s_h,a_h,b_h)\,\big|\,s_1\big]\leq 0.$$

Proof. We prove the upper bound of the model prediction error term. We can decompose the instantaneous prediction error at the h-step of the k-th episode as

$$\overline{\varsigma}_h^k(s,a,b)=r_h(s,a,b)+\big\langle\mathcal{P}_h(\cdot\,|\,s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\overline{Q}_h^k(s,a,b), \tag{55}$$

where the equality is by the definition of the prediction error in (36). By plugging in the definition of $\overline{Q}_h^k$ in Line (2) of Algorithm 2, for any (s,a,b), we bound the following term as

$$\begin{aligned}
&r_h(s,a,b)+\big\langle\mathcal{P}_h(\cdot\,|\,s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\overline{Q}_h^k(s,a,b)\\
&\quad\leq r_h(s,a,b)+\big\langle\mathcal{P}_h(\cdot\,|\,s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\min\Big\{\widehat{r}_h^k(s,a,b)+\big\langle\widehat{\mathcal{P}}_h^k(\cdot|s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\beta_h^k,H-h+1\Big\}\\
&\quad\leq \max\Big\{r_h(s,a,b)-\widehat{r}_h^k(s,a,b)+\big\langle\mathcal{P}_h(\cdot\,|\,s,a)-\widehat{\mathcal{P}}_h^k(\cdot|s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\beta_h^k,0\Big\},
\end{aligned}\tag{56}$$

where the inequality holds because

$$\begin{aligned}
&r_h(s,a,b)+\big\langle\mathcal{P}_h(\cdot\,|\,s_h,a_h),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\\
&\qquad\leq r_h(s,a,b)+\big\|\mathcal{P}_h(\cdot\,|\,s_h,a_h)\big\|_1\|\overline{V}_{h+1}^k(\cdot)\|_\infty\leq 1+\max_{s'\in\mathcal{S}}\big|\overline{V}_{h+1}^k(s')\big|\leq 1+H-h,
\end{aligned}$$

since $\big\|\mathcal{P}_h(\cdot\,|\,s_h,a_h)\big\|_1=1$ and also the truncation step as shown in Line 2 of Algorithm 2 for $\overline{Q}_{h+1}^k$ such that for any $s'\in\mathcal{S}$

$$\begin{aligned}
\big|\overline{V}_{h+1}^k(s')\big|&=\Big|\big[\mu_{h+1}^k(\cdot|s')\big]^\top\overline{Q}_{h+1}^k(s',\cdot,\cdot)\nu_{h+1}^k(\cdot|s')\Big|\\
&\leq\big\|\mu_{h+1}^k(\cdot|s')\big\|_1\big\|\overline{Q}_{h+1}^k(s',\cdot,\cdot)\nu_{h+1}^k(\cdot|s')\big\|_\infty\\
&\leq\max_{a,b}\big|\overline{Q}_{h+1}^k(s',a,b)\big|\\
&\leq H-h.
\end{aligned}\tag{57}$$

Combining (55) and (56) gives

$$\overline{\varsigma}_h^k(s,a,b)\leq\max\Big\{r_h(s,a,b)-\widehat{r}_h^k(s,a,b)+\big\langle\mathcal{P}_h(\cdot\,|\,s,a)-\widehat{\mathcal{P}}_h^k(\cdot|s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\beta_h^k,0\Big\}. \tag{58}$$

Note that as shown in (12), we have

$$\beta_h^k(s,a,b)=\beta_h^{r,k}(s,a,b)+\beta_h^{\mathcal{P},k}(s,a).$$

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

$$\begin{aligned}
&r_h(s,a,b)-\widehat{r}_h^k(s,a,b)-\beta_h^{r,k}(s,a,b)\\
&\qquad\leq\big|r_h(s,a,b)-\widehat{r}_h^k(s,a,b)\big|-\beta_h^{r,k}(s,a,b)\\
&\qquad\leq\beta_h^{r,k}(s,a,b)-\beta_h^{r,k}(s,a,b)=0,
\end{aligned}$$

where the last inequality is by Lemma C.4 and the setting of the bonus for the reward. Moreover, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\big\langle\mathcal{P}_h(\cdot\,|\,s,a)-\widehat{\mathcal{P}}_h^k(\cdot|s,a),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}-\beta_h^{\mathcal{P},k}(s,a)\\
&\qquad\leq\big\|\mathcal{P}_h(\cdot\,|\,s,a)-\widehat{\mathcal{P}}_h^k(\cdot|s,a)\big\|_1\big\|\overline{V}_{h+1}^k(\cdot)\big\|_\infty-\beta_h^{\mathcal{P},k}(s,a)\\
&\qquad\leq H\big\|\mathcal{P}_h(\cdot\,|\,s,a)-\widehat{\mathcal{P}}_h^k(\cdot|s,a)\big\|_1-\beta_h^{\mathcal{P},k}(s,a)\\
&\qquad\leq\beta_h^{\mathcal{P},k}(s,a)-\beta_h^{\mathcal{P},k}(s,a)=0,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality is due to $\max_{s'\in\mathcal{S}} \|\overline{V}_{h+1}^k(s')\|_\infty \leq H$ as shown in (57), and the last inequality is by the setting of $\beta_h^{\mathcal{P},k}$ and also Lemma C.5. Thus, with probability at least $1-2\delta$, the following inequality holds

$$r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \big\langle \mathcal{P}_h(\cdot\,|\,s,a) - \widehat{\mathcal{P}}_h^k(\cdot|s,a), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} - \beta_h^k(s,a,b) \leq 0.$$

Combining the above inequality with (58), we have that with probability at least $1-2\delta$, for any $h\in[H]$ and $k\in[K]$, the following inequality holds

$$\overline{\varsigma}_h^k(s,a,b) \leq 0, \;\; \forall (s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B},$$

which leads to

$$\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big[\overline{\varsigma}_h^k(s_h,a_h,b_h)\,\big|\,s_1\big] \leq 0.$$

This completes the proof. ◻

Lemma C.7. *With probability at least $1-\delta$, Algorithm 2 ensures that*

$$\sum_{k=1}^K \overline{V}_1^k(s_1) - \sum_{k=1}^K V_1^{\mu^k,\nu^k}(s_1) \leq \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K}\right).$$

Proof. We assume that a trajectory $\{(s_h^k,a_h^k,b_h^k,s_{h+1}^k)\}_{h=1}^H$ for all $k\in[K]$ is generated according to the policies μ^k, ν^k, and the true transition model $\mathcal{P}$. Thus, we expand the bias term at the h-th step of the k-th episode, which is

$$\begin{aligned}
&\overline{V}_h^k(s_h^k) - V_h^{\mu^k,\nu^k}(s_h^k)\\
&\quad= \big[\mu_h^k(\cdot|s_h^k)\big]^\top\big[\overline{Q}_h^k(s_h^k,\cdot,\cdot) - Q_h^{\mu^k,\nu^k}(s_h^k,\cdot,\cdot)\big]\nu_h^k(\cdot|s_h^k)\\
&\quad= \zeta_h^k + \overline{Q}_h^k(s_h^k,a_h^k,b_h^k) - Q_h^{\mu^k,\nu^k}(s_h^k,a_h^k,b_h^k)\\
&\quad= \zeta_h^k + \big\langle \mathcal{P}_h(\cdot\,|\,s_h^k,a_h^k), \overline{V}_{h+1}^k(\cdot) - V_{h+1}^{\mu^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}} - \overline{\varsigma}_h^k(s_h^k,a_h^k,b_h^k)\\
&\quad= \zeta_h^k + \xi_h^k + \overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mu^k,\nu^k}(s_{h+1}^k) - \overline{\varsigma}_h^k(s_h^k,a_h^k,b_h^k),
\end{aligned} \tag{59}$$

where the first equality is by Line 2 of Algorithm 2 and (1), the third equality is by plugging in (2) and (36). Specifically, in the above equality, we introduce two martingale difference sequence, namely, $\{\zeta_h^k\}_{h\geq 0,k\geq 0}$ and $\{\xi_h^k\}_{h\geq 0,k\geq 0}$, which are defined as

$$\begin{aligned}
&\zeta_h^k := \big[\mu_h^k(\cdot|s_h^k)\big]^\top\big[\overline{Q}_h^k(s_h^k,\cdot,\cdot) - Q_h^{\mu^k,\nu^k}(s_h^k,\cdot,\cdot)\big]\nu_h^k(\cdot|s_h^k) - \big[\overline{Q}_h^k(s_h^k,a_h^k,b_h^k) - Q_h^{\mu^k,\nu^k}(s_h^k,a_h^k,b_h^k)\big],\\
&\xi_h^k := \big\langle \mathcal{P}_h(\cdot\,|\,s_h^k,a_h^k), \overline{V}_{h+1}^k(\cdot) - V_{h+1}^{\mu^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}} - \big[\overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mu^k,\nu^k}(s_{h+1}^k)\big],
\end{aligned}$$

such that

$$\mathbb{E}_{a_h^k\sim\mu_h^k(\cdot|s_h^k),b_h^k\sim\nu_h^k(\cdot|s_h^k)}\big[\zeta_h^k\,\big|\,\mathcal{F}_h^k\big] = 0, \qquad \mathbb{E}_{s_{h+1}^k\sim\mathcal{P}_h(\cdot\,|\,s_h^k,a_h^k)}\big[\xi_h^k\,\big|\,\widetilde{\mathcal{F}}_h^k\big] = 0,$$

with $\mathcal{F}_h^k$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s_h^k, and $\widetilde{\mathcal{F}}_h^k$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s_h^k, a_h^k, b_h^k.

We can observe that the equality (59) construct a recursion for $\overline{V}_h^k(s_h^k) - V_h^{\mu^k,\nu^k}(s_h^k)$. Moreover, we also have $\overline{V}_{H+1}^k(\cdot) = \mathbf{0}$ and $V_{H+1}^{\mu^k,\nu^k}(\cdot) = \mathbf{0}$. Thus, recursively apply (59) from $h=1$ to H leads to the following equality

$$\overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1) = \sum_{h=1}^H \zeta_h^k + \sum_{h=1}^H \xi_h^k - \sum_{h=1}^H \overline{\varsigma}_h^k(s_h^k,a_h^k,b_h^k). \tag{60}$$

Moreover, by (36) and Line 2 of Algorithm 2, we have

$$-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k) = -r_h(s_h^k, a_h^k, b_h^k) - \big\langle \mathcal{P}_h(\cdot \mid s_h, a_h), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \min\big\{\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) + \big\langle \widehat{\mathcal{P}}_h^k(\cdot|s_h, a_h), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \beta_h^k(s_h^k, a_h^k, b_h^k), H-h+1\big\}.$$

Then, we can further bound $-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k)$ as follows

$$\begin{aligned}
-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k) &\le -r_h(s_h^k, a_h^k, b_h^k) - \big\langle \mathcal{P}_h(\cdot \mid s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \widehat{r}_h^k(s_h^k, a_h^k, b_h^k) \\
&\quad + \big\langle \widehat{\mathcal{P}}_h^k(\cdot|s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \beta_h^k(s_h^k, a_h^k, b_h^k) \\
&\le \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| \\
&\quad + \Big|\big\langle \mathcal{P}_h(\cdot \mid s_h^k, a_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \mid s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big| + \beta_h^k(s_h^k, a_h^k, b_h^k),
\end{aligned}$$

where the first inequality is due to $\min\{x, y\} \le x$. Additionally, we have

$$\begin{aligned}
&\Big|\big\langle \mathcal{P}_h(\cdot \mid s_h^k, a_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \mid s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big| \\
&\quad \le \big\|\overline{V}_{h+1}^k(\cdot)\big\|_\infty \big\|\mathcal{P}_h(\cdot \mid s_h^k, a_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \mid s_h^k, a_h^k)\big\|_1 \\
&\quad \le H\big\|\mathcal{P}_h(\cdot \mid s_h^k, a_h^k) - \widehat{\mathcal{P}}_h^k(\cdot \mid s_h^k, a_h^k)\big\|_1,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality and the second inequality is by (57). Thus, putting the above together, we obtain

$$\begin{aligned}
-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k) &\le \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| + H\big\|\overline{V}_{h+1}^k(\cdot) - \overline{V}_{h+1}^k(\cdot)\big\|_1 + \beta_h^k(s_h^k, a_h^k, b_h^k) \\
&\le 2\beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\beta_h^{\mathcal{P},k}(s_h^k, a_h^k),
\end{aligned}$$

where the second inequality is by Lemma C.4, Lemma C.5, and the decomposition of the bonus term β_h^k as (12). Due to Lemma C.4 and Lemma C.5, by union bound, for any $h \in [H], k \in [K]$ and $(s_h, a_h, b_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, the above inequality holds with probability with probability at least $1 - 2\delta$. Therefore, by (60), with probability at least $1 - 2\delta$, we have

$$\begin{aligned}
&\sum_{k=1}^K \big[\overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1)\big] \\
&\quad \le \sum_{k=1}^K\sum_{h=1}^H \zeta_h^k + \sum_{k=1}^K\sum_{h=1}^H \xi_h^k + 2\sum_{k=1}^K\sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\sum_{k=1}^K\sum_{h=1}^H \beta_h^{\mathcal{P},k}(s_h^k, a_h^k).
\end{aligned} \tag{61}$$

By Azuma-Hoeffding inequality, with probability at least $1 - \delta$, the following inequalities hold

$$\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \zeta_h^k &\le \mathcal{O}\left(\sqrt{H^3 K \log\frac{1}{\delta}}\right), \\
\sum_{k=1}^K\sum_{h=1}^H \xi_h^k &\le \mathcal{O}\left(\sqrt{H^3 K \log\frac{1}{\delta}}\right),
\end{aligned}$$

where we use the facts that $|\overline{Q}_h^k(s_h^k, a_h^k, b_h^k) - Q_h^{\mu^k,\nu^k}(s_h^k, a_h^k, b_h^k)| \le 2H$ and $|\overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\mu^k,\nu^k}(s_{h+1}^k)| \le 2H$. Next, we need to bound $\sum_{k=1}^K\sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k)$ and $\sum_{k=1}^K\sum_{h=1}^H \beta_h^{\mathcal{P},k}(s_h^k, a_h^k)$ in (61). We show that

$$\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) &= C\sum_{k=1}^K\sum_{h=1}^H \sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s_h^k, a_h^k, b_h^k), 1\}}} \\
&= C\sum_{k=1}^K\sum_{h=1}^H \sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{N_h^k(s_h^k, a_h^k, b_h^k)}} \\
&\le C\sum_{h=1}^H \sum_{\substack{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B} \\ N_h^K(s,a,b)>0}} \sum_{n=1}^{N_h^K(s,a,b)} \sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}},
\end{aligned}$$

where the second equality is because (s_h^k, a_h^k, b_h^k) is visited such that $N_h^k(s_h^k, a_h^k, b_h^k) \geq 1$. In addition, we have

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{\substack{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B} \\ N_h^K(s,a,b)>0}} \sum_{n=1}^{N_h^K(s,a,b)} \sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}} \\
&\quad \leq \sum_{h=1}^{H} \sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} \mathcal{O}\left(\sqrt{N_h^K(s,a,b)\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right) \\
&\quad \leq \mathcal{O}\left(H\sqrt{K|\mathcal{S}||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right),
\end{aligned}$$

where the last inequality is based on the consideration that $\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} N_h^K(s,a,b) = K$ such that $\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} \sqrt{N_h^K(s,a,b)} \leq \mathcal{O}\left(\sqrt{K|\mathcal{S}||\mathcal{A}||\mathcal{B}|}\right)$ when K is sufficiently large. Putting the above together, we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) \leq \mathcal{O}\left(H\sqrt{K|\mathcal{S}||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right).$$

Similarly, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{\mathcal{P},k}(s_h^k, a_h^k) &= \sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\frac{H^2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s_h^k, a_h^k), 1\}}} \\
&\leq \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{O}\left(\sqrt{N_h^K(s,a)H^2|\mathcal{S}|\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta}}\right) \\
&\leq \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{O}\left(\sqrt{\sum_{b\in B} N_h^K(s,a,b)H^2|\mathcal{S}|\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta}}\right) \\
&\leq \mathcal{O}\left(H\sqrt{K|\mathcal{S}|^2|\mathcal{A}|H^2\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta}}\right),
\end{aligned}$$

where the second inequality is due to $\sum_{b\in\mathcal{B}} N_h^K(s,a,b) = N_h^K(s,a)$, and the last inequality is based on the consideration that $\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} N_h^K(s,a,b) = K$ such that $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{\sum_{b\in\mathcal{B}} N_h^K(s,a,b)} \leq \mathcal{O}(\sqrt{K|\mathcal{S}||\mathcal{A}|})$ when K is sufficiently large.

Thus, by (61), with probability at least $1-\delta$, we have

$$\sum_{k=1}^{K} \overline{V}_1^k(s_1) - \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1) \leq \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K})$$

where $\widetilde{\mathcal{O}}$ hides logarithm terms. This completes the proof. □

Lemma C.8. *With setting $\gamma = \sqrt{|\mathcal{S}|\log|\mathcal{B}|/K}$, the mirror descent steps of Algorithm 3 lead to*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle \leq \mathcal{O}\left(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|}\right).$$

Proof. Similar to the proof of Lemma C.3, and also by Lemma C.12, for any $\nu = \{\nu_h\}_{h=1}^H$ and $s\in\mathcal{S}$, the mirror descent step in Algorithm 3 leads to

$$\begin{aligned}
&\gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^{k+1}(\cdot|s)\big\rangle_{\mathcal{B}} - \gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h(\cdot|s)\big\rangle_{\mathcal{B}} \\
&\quad \leq D_{\mathrm{KL}}\big(\nu_h(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h(\cdot|s), \nu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big),
\end{aligned}$$

according to (14), where $W_h^k(s,a) = \langle \nu_h^k(\cdot|s), \widetilde{r}_h^k(s,a,\cdot)\rangle$. Then, by rearranging the terms, we have

$$\begin{aligned}
&\gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} \\
&\quad\le D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big) \\
&\qquad - \gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\rangle_{\mathcal{B}}.
\end{aligned} \tag{62}$$

Due to Pinsker's inequality, we have

$$- D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big) \le -\frac{1}{2}\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1^2. \tag{63}$$

Moreover, we have

$$\begin{aligned}
&- \gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^{k+1}(\cdot|s)\big\rangle_{\mathcal{B}} \\
&\quad\le \gamma d_h^k(s)\big\|W_h^k(s,\cdot)\big\|_\infty \big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1 \\
&\quad\le \gamma d_h^k(s)\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1,
\end{aligned} \tag{64}$$

where the last inequality is by

$$\begin{aligned}
\|W_h^k(s,\cdot)\|_\infty &= \max_{b\in\mathcal{B}} W_h^k(s,b) \\
&\le \max_{s\in\mathcal{S}, b\in\mathcal{B}} W_h^k(s,b) \\
&\le \max_{s\in\mathcal{S}, b\in\mathcal{B}} \big\langle \widetilde{r}_h^{k-1}(s,\cdot,b), \mu_h^k(\cdot\,|\,s)\big\rangle \\
&\le \max_{s\in\mathcal{S}, b\in\mathcal{B}} \big\|\widetilde{r}_h^{k-1}(s,\cdot,b)\big\|_\infty \big\|\mu_h^k(\cdot\,|\,s)\big\|_1 \le 1.
\end{aligned}$$

due to the definition of W_h^k and $\widetilde{r}_h^k(s,a,b) = \max\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}, 0\} \le \widehat{r}_h^k(s,a,b) \le 1$. Combining (63) and (64) gives

$$\begin{aligned}
&- D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big) - \gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^{k+1}(\cdot|s)\big\rangle \\
&\quad\le -\frac{1}{2}\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1^2 + \gamma d_h^k(s)\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1 \\
&\quad\le \frac{1}{2}\big[d_h^k(s)\big]^2\gamma^2 \le \frac{1}{2} d_h^k(s)\gamma^2,
\end{aligned}$$

where the second inequality is obtained via solving $\max_x\{-1/2\cdot x^2 + \gamma d_h^k(s)\cdot x\}$ if letting $x = \|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\|_1$. Plugging the above inequality into (62) gives

$$\gamma d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} \le D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big) + \frac{1}{2} d_h^k(s)\gamma^2.$$

Thus, the policy improvement step implies

$$\begin{aligned}
&\sum_{h=1}^H \sum_{s\in\mathcal{S}} d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} \\
&\quad\le \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \big[D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big)\big] + \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \frac{1}{2} d_h^k(s)\gamma^2 \\
&\quad\le \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \big[D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big)\big] + \frac{1}{2}H\gamma.
\end{aligned}$$

Further summing on both sides of the above inequality from $k = 1$ to K gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}$$
$$\leq \frac{1}{\gamma}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\big[D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^1(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{K+1}(\cdot|s)\big)\big] + \frac{1}{2}HK\gamma$$
$$\leq \frac{1}{\gamma}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^1(\cdot|s)\big) + \frac{1}{2}HK\gamma.$$

Note that by the initialization in Algorithm 3, it is guaranteed that $\nu_h^1(\cdot|s) = \mathbf{1}/|\mathcal{B}|$, which thus leads to $D_{\mathrm{KL}}\left(\mu_h^*(\cdot|s), \mu_h^1(\cdot|s)\right) \leq \log|\mathcal{B}|$. By setting $\gamma = \sqrt{|\mathcal{S}|\log|\mathcal{B}|/K}$, we further bound the term as

$$\frac{1}{\gamma}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^1(\cdot|s)\big) + \frac{1}{2}HK\gamma \leq \mathcal{O}\left(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|}\right),$$

which gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^k(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} \leq \mathcal{O}\left(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|}\right).$$

This completes the proof. □

Before giving the next lemma, we first present the following definition for the proof of the next lemma.

Definition C.9 (Confidence Set)**.** *Define the following confidence set for transition models*

$$\Upsilon^k := \Big\{\widetilde{\mathcal{P}} : \Big|\widetilde{\mathcal{P}}_h(s'|s,a) - \widehat{\mathcal{P}}_h^k(s'|s,a)\Big| \leq \epsilon_h^k,\ \|\widetilde{\mathcal{P}}_h(\cdot|s,a)\|_1 = 1,$$
$$\textit{and}\ \ \widetilde{\mathcal{P}}_h(s'|s,a) \geq 0,\ \forall(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}, \forall k \in [K]\Big\}$$

where we define

$$\epsilon_h^k := 2\sqrt{\frac{\widehat{\mathcal{P}}_h^k(s'|s,a)\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a)-1, 1\}}} + \frac{14\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{3\max\{N_h^k(s,a)-1, 1\}}$$

with $N_h^k(s,a) := \sum_{\tau=1}^{k}\mathbf{1}\{(s,a) = (s_h^\tau, a_h^\tau)\}$*, and* $\widehat{\mathcal{P}}^k$ *being the empirical transition model.*

Lemma C.10. *With probability at least* $1-\delta$*, the difference between* $q^{\mu^k,\mathcal{P}}$ *and* d^k *are bounded as*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s)\right| \leq \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K}\right).$$

Proof. By the definition of state distribution, we first have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s)\right| = \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|\sum_{a\in\mathcal{A}} w_h^k(s,a) - \sum_{a\in\mathcal{A}}\widehat{w}_h^k(s,a)\right|$$
$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left|w_h^k(s,a) - \widehat{w}_h^k(s,a)\right|.$$

where $\widehat{w}_h^k(s,a)$ is the occupancy measure under the empirical transition model $\widehat{\mathcal{P}}^k$ and the policy μ^k. Then, since $\widehat{\mathcal{P}}^k \in \Upsilon^k$ always holds for any k, by Lemma C.15, we can bound the last term of the bound inequality such that with probability at least $1-6\delta'$,

$$\sum_{k=1}^K \sum_{h=1}^H \sum_{s\in\mathcal{S}} \left| q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s) \right| \le \mathcal{E}_1 + \mathcal{E}_2.$$

Next, we compute the order of $\mathcal{E}_1$ by Lemma C.14. With probability at least $1-2\delta'$, we have

$$\begin{aligned}
\mathcal{E}_1 &= \mathcal{O}\left[\sum_{h=2}^H \sum_{h'=1}^{h-1} \sum_{k=1}^K \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w_h^k(s,a) \left(\sqrt{\frac{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a),1\}}} + \frac{\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a),1\}} \right) \right] \\
&= \mathcal{O}\left[\sum_{h=2}^H \sum_{h'=1}^{h-1} \sqrt{|\mathcal{S}|} \left(\sqrt{|\mathcal{S}||\mathcal{A}|K} + |\mathcal{S}||\mathcal{A}|\log K + \log\frac{H}{\delta'} \right) \log \frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'} \right] \\
&= \mathcal{O}\left[\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K} + H^2|\mathcal{S}|^{3/2}|\mathcal{A}|\log K + H^2\sqrt{|\mathcal{S}|}\log\frac{H}{\delta'} \right) \log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'} \right] \\
&= \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K} \right),
\end{aligned}$$

where we ignore $\log K$ terms when K is sufficiently large such that $\sqrt{K}$ dominates, and $\widetilde{\mathcal{O}}$ hides logarithm dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, H, K, and $1/\delta'$. On the other hand, $\mathcal{E}_2$ also depends on $\mathrm{ploy}(H, |\mathcal{S}|, |\mathcal{A}|)$ except the factor $\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}$ as shown in Lemma C.15. Thus, $\mathcal{E}_2$ can be ignored comparing to $\mathcal{E}_1$ if K is sufficiently large. Therefore, we eventually obtain that with probability at least $1-8\delta'$, the following inequality holds

$$\sum_{k=1}^K \sum_{h=1}^H \sum_{s\in\mathcal{S}} \left| q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s) \right| \le \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K} \right).$$

We let $\delta = 8\delta'$ such that $\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'} = \log\frac{8|\mathcal{S}||\mathcal{A}|HK}{\delta}$ without changing the order as shown above. Then, with probability at least $1-\delta$, we have $\sum_{k=1}^K\sum_{h=1}^H\sum_{s\in\mathcal{S}} |q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s)| \le \widetilde{\mathcal{O}}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K})$. This completes the proof. ☐

Lemma C.11. *With probability at least $1-\delta$, the following inequality holds*

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h) \,\middle|\, s_1 \right] \le \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K} \right).$$

Proof. Since we have

$$\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h) \,\middle|\, s_1 \right] \\
&\qquad = \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\left[C\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{N_h^k(s,a,b)}} \right] \\
&\qquad = C\sqrt{\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}} \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\left[\sqrt{\frac{1}{N_h^k(s,a,b)}} \right],
\end{aligned}$$

then we can apply Lemma C.16 and obtain

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h) \,\middle|\, s_1 \right] \le \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K} \right),$$

with probability at least $1-\delta$. Here $\widetilde{\mathcal{O}}$ hides logarithm dependence on $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|, H, K$, and $1/\delta$. This completes the proof. ☐

C.1. Other Supporting Lemmas

Lemma C.12. *Let $f : \Lambda \mapsto \mathbb{R}$ be a convex function, where Λ is the probability simplex defined as $\Lambda := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = 1$ and $\mathbf{x}_i \geq 0, \forall i \in [d]\}$. For any $\alpha \geq 0$, $\mathbf{z} \in \Lambda$, and $\mathbf{y} \in \Lambda^o$ where $\Lambda^o \subset \Lambda$ with only relative interior points of Λ, supposing $\mathbf{x}^{\mathrm{opt}} = \mathrm{argmin}_{\mathbf{x}\in\Lambda} f(\mathbf{x}) + \alpha D_{\mathrm{KL}}(\mathbf{x}, \mathbf{y})$, then the following inequality holds*

$$f(\mathbf{x}^{\mathrm{opt}}) + \alpha D_{\mathrm{KL}}(\mathbf{x}^{\mathrm{opt}}, \mathbf{y}) \leq f(\mathbf{z}) + \alpha D_{\mathrm{KL}}(\mathbf{z}, \mathbf{y}) - \alpha D_{\mathrm{KL}}(\mathbf{z}, \mathbf{x}^{\mathrm{opt}}).$$

This lemma is for mirror descent algorithms, whose proof can be found in existing works (Tseng, 2008; Nemirovski et al., 2009; Wei et al., 2019).

Lemma C.13. *With probability at least $1 - 4\delta'$, the true transition model $\mathcal{P}$ satisfies that for any $k \in [K]$,*

$$\mathcal{P} \in \Upsilon^k.$$

This lemma implies that the estimated transition model $\widehat{\mathcal{P}}_h^k(s'|s,a)$ by (11) is closed to the true transition model $\mathcal{P}_h(s'|s,a)$ with high probability. The upper bound for their difference is by empirical Bernstein's inequality and the union bound.

The next lemma is modified from Lemma 10 in Jin & Luo (2019).

Lemma C.14. *We let $w_h^k(s,a)$ denote the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathcal{P}$ and the current policy μ^k. Then, with probability at least $1-2\delta'$ we have for all $h \in [H]$, the following inequalities hold*

$$\sum_{k=1}^K \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \frac{w_h^k(s,a)}{\max\{N_h^k(s,a), 1\}} = \mathcal{O}\left(|\mathcal{S}||\mathcal{A}| \log K + \log \frac{H}{\delta'}\right),$$

and

$$\sum_{k=1}^K \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \frac{w_h^k(s,a)}{\sqrt{\max\{N_h^k(s,a), 1\}}} = \mathcal{O}\left(\sqrt{|\mathcal{S}||\mathcal{A}|K} + |\mathcal{S}||\mathcal{A}| \log K + \log \frac{H}{\delta'}\right).$$

Furthermore, by Lemma C.13 and Lemma C.14, we give the following lemma to characterize the difference of two occupancy measures, which is modified from parts of the proof of Lemma 4 in Jin & Luo (2019).

Lemma C.15. *Let $w_h^k(s,a)$ be the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathcal{P}$ and the current policy μ^k, and $\widetilde{w}_h^k(s,a)$ be the occupancy measure at the h-th step of the k-th episode under any transition model $\widetilde{\mathcal{P}}^k \in \Upsilon^k$ and the current policy μ^k for any k. Then, with probability at least $1 - 6\delta'$ we have for all $h \in [H]$, the following inequalities hold*

$$\sum_{k=1}^K \sum_{h=1}^K \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left|\widetilde{w}_h^k(s,a) - w_h^k(s,a)\right| \leq \mathcal{E}_1 + \mathcal{E}_2,$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are in the level of

$$\mathcal{E}_1 = \mathcal{O}\left[\sum_{h=2}^H \sum_{h'=1}^{h-1} \sum_{k=1}^K \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} w_h^k(s,a) \left(\sqrt{\frac{|\mathcal{S}| \log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a), 1\}}} + \frac{\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a), 1\}}\right)\right]$$

and

$$\mathcal{E}_2 = \mathcal{O}\left(\mathrm{poly}(H, |\mathcal{S}|, |\mathcal{A}|) \cdot \log \frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}\right),$$

where $\mathrm{poly}(H, |\mathcal{S}|, |\mathcal{A}|)$ denotes the polynomial dependency on $H, |\mathcal{S}|, |\mathcal{A}|$.

Lemma C.16. *With probability at least $1 - \delta$, the following inequality hold*

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k, \mathcal{P}, \nu^k}\left[\sqrt{\frac{1}{\max\{N_h^k(s,a,b), 1\}}}\right] \leq \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K} + |\mathcal{S}||\mathcal{A}||\mathcal{B}|H\right),$$

where $\widetilde{\mathcal{O}}$ hides logarithm terms.

Proof. The zero-sum Markov game with single controller in this paper can interpreted as a regular MDP learning problem with policies $w_h^k(a, b \mid s) = \mu_h^k(a|s)\nu_h^k(b|s)$ and a transition model $\mathcal{P}_h(s'|s, a, b) = \mathcal{P}_h(s'|s, a)$ with a joint action (a, b) in the action space of size $|\mathcal{A}||\mathcal{B}|$. Thus, we apply Lemma 19 of Efroni et al. (2020), which extends lemmas in Zanette & Brunskill (2019); Efroni et al. (2019) to MDP with non-stationary dynamics by adding a factor of H, to obtain our lemma. This completes the proof. $\square$