

A TECHNICAL LEMMAS

Lemma 1. Let Z_t be a sequence of random variables, where each Z_t may depend on the previous observations $\mathcal{S}_{t-1} = [Z_1, \dots, Z_{t-1}] \in \mathcal{Z}^{t-1}$. Furthermore, we define a filtration $\{\mathcal{F}_t = \sigma(\mathcal{S}_t)\}$, which is also the natural filtration of $\{Z_t\}$. Consider a sequence of real-valued random (measurable) functions $\xi_1(\mathcal{S}_1), \dots, \xi_T(\mathcal{S}_T)$. Let $\tau \leq T$ be a stopping time so that $\mathbb{I}(t \leq \tau)$ is measurable in \mathcal{S}_t . We have

$$\mathbb{E}_{\mathcal{S}_T} \exp \left(\sum_{t=1}^{\tau} \xi_t - \sum_{t=1}^{\tau} \ln \mathbb{E}_{Z_t | \mathcal{S}_{t-1}} e^{\xi_t} \right) = 1.$$

Proof. This proof is a revised version of Lemma 13.1 in [Zhang \(2023\)](#). We prove this lemma by induction. When $T = 0$, the equality apparently holds. We then assume that the claim holds at $T - 1$ for some $T \geq 1$. Now we will prove the equation at time T using the induction hypothesis.

First we define $\tilde{\xi}_t = \xi_t \mathbb{I}(t \leq \tau)$ and notice that $\tilde{\xi}_t$ is measurable in \mathcal{S}_t so we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{S}_T} \exp \left(\sum_{t=1}^{\tau} \xi_t - \sum_{i=1}^{\tau} \ln \mathbb{E}_{Z_i | \mathcal{S}_{i-1}} e^{\xi_i} \right) \\ &= \mathbb{E}_{\mathcal{S}_T} \exp \left(\sum_{t=1}^T \tilde{\xi}_t - \sum_{i=1}^T \ln \mathbb{E}_{Z_i | \mathcal{S}_{i-1}} e^{\tilde{\xi}_i} \right) \\ &= \mathbb{E}_{\mathcal{S}_{T-1}} \left[\exp \left(\sum_{t=1}^{T-1} \tilde{\xi}_t - \sum_{i=1}^{T-1} \ln \mathbb{E}_{Z_i | \mathcal{S}_{i-1}} e^{\tilde{\xi}_i} \right) \mathbb{E}_{Z_T | \mathcal{S}_{T-1}} \exp \left(\tilde{\xi}_T - \ln \mathbb{E}_{Z_T | \mathcal{S}_{T-1}} e^{\tilde{\xi}_T} \right) \right] \\ &= \mathbb{E}_{\mathcal{S}_{T-1}} \left[\exp \left(\sum_{t=1}^{T-1} \tilde{\xi}_t - \sum_{i=1}^{T-1} \ln \mathbb{E}_{Z_i | \mathcal{S}_{i-1}} e^{\tilde{\xi}_i} \right) \right] \\ &= \mathbb{E}_{\mathcal{S}_{T-1}} \left[\exp \left(\sum_{t=1}^{\min(\tau, T-1)} \tilde{\xi}_t - \sum_{i=1}^{\min(\tau, T-1)} \ln \mathbb{E}_{Z_i | \mathcal{S}_{i-1}} e^{\tilde{\xi}_i} \right) \right] \\ &= 1, \end{aligned}$$

where the third equality exploits the fact that $\mathbb{E}_{Z_T(\cdot)} \exp \left(\tilde{\xi}_T - \ln \mathbb{E}_{Z_T | \mathcal{S}_{T-1}} e^{\tilde{\xi}_T} \right) = 1$; and the last equality is because we could treat $\min(\tau, T - 1)$ as a stopping time no more than $T - 1$ and we could use the induction hypothesis. \square

Lemma 2 (Martingale exponential inequality). For a sequence of real-valued random variables $\{X_t\}_{t \leq T}$ adapted to a filtration $\{\mathcal{F}_t\}_{t \leq T}$, the following holds with probability at least $1 - \delta$, for $\forall t \in [T]$,

$$-\sum_{s=1}^t X_s \leq \sum_{s=1}^t \ln \mathbb{E} [e^{-X_s} | \mathcal{F}_{s-1}] + \ln \frac{1}{\delta}.$$

And also

$$\sum_{s=1}^t X_s \leq \sum_{s=1}^t \ln \mathbb{E} [e^{X_s} | \mathcal{F}_{s-1}] + \ln \frac{1}{\delta}.$$

Proof. It only suffices to show the case when $\{\xi_i\}_{i=1}^T$ is a finite case. The statement implies the original lemma by pushing $T \rightarrow +\infty$. Let

$$U_\tau = -\sum_{s=1}^{\tau} X_s - \sum_{s=1}^{\tau} \ln \mathbb{E}_{\mathcal{S}_t} e^{-X_s},$$

where τ is some stopping time. By Lemma 1 we have $\mathbb{E}(\exp^{U_\tau}) = 1$. (In this case, we apply $Z_s = \xi_s = -X_s$ in Lemma 1). Now we define the stopping time τ as

$$\tau = \min (T, \min (n : U_n \geq -\ln \delta)).$$

Then it follows that

$$\mathbb{P}(\exists n : U_\tau \geq -\ln \delta) \leq \mathbb{E}[e^{U_\tau + \ln \delta}] = \delta \mathbb{E}[e^{U_\tau}] = \delta,$$

where the first inequality is by the famous Markov Inequality.

By considering the complementary event, we know with probability at least $1 - \delta$, the following inequality holds for any $t \in [T]$

$$-\sum_{s=1}^t X_s \leq \sum_{s=1}^t \ln \mathbb{E}[e^{-X_s} | \mathcal{F}_{s-1}] + \ln \frac{1}{\delta}.$$

□

Lemma 3 (Freedman's inequality). *Let $\{X_t\}_{t \leq T}$ be any sequence of real-valued random variables adapted to filtration $\{\mathcal{F}_t\}_{t \leq T}$. If $|X_t| \leq R$ almost surely, then for any $\eta \in (0, \frac{1}{2R}]$ it holds that with probability at least $1 - \delta$,*

$$\sum_{t=1}^T X_t \leq \sum_{t=1}^T \mathbb{E}(X_t | \mathcal{F}_{t-1}) + \eta \sum_{t=1}^T \text{Var}[X_t | \mathcal{F}_{t-1}] + \frac{\ln \frac{1}{\delta}}{\eta}.$$

Furthermore, we have

$$\sum_{t=1}^T \mathbb{E}(X_t | \mathcal{F}_{t-1}) \leq \sum_{s=1}^T X_s + \eta \sum_{s=1}^T \text{Var}[X_s | \mathcal{F}_{s-1}] + \frac{\ln \frac{1}{\delta}}{\eta}.$$

Proof. For any random variable X we assume $|X| \leq R$ almost surely, and let $X' = X - \mathbb{E}X$. We then get $|X'| \leq 2R$ almost surely, and we have

$$\begin{aligned} \ln \mathbb{E}[e^{\lambda X}] &= \lambda \mathbb{E}X + \ln \mathbb{E}e^{\lambda X'} \\ &\leq \lambda \mathbb{E}X + \mathbb{E}e^{\lambda X'} - 1 \\ &= \lambda \mathbb{E}X + \lambda^2 \mathbb{E}\left[\frac{e^{\lambda X' - \lambda X' - 1}}{(\lambda X')^2} (X')^2\right] \\ &\leq \lambda \mathbb{E}X + \lambda^2 \phi(\lambda 2R) \text{Var}[X], \end{aligned}$$

where $\phi(x) = \frac{e^x - x - 1}{x^2}$; the first inequality uses $\ln x \leq x - 1$; the second inequality exploits the fact that $\phi(x)$ is non-decreasing. Then, we consider the Taylor expansion: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, and we have

$$\phi(x) = \sum_{n=2}^{\infty} \left(\frac{x^{n-2}}{n!}\right) \leq \frac{1}{2} \sum_{n=0}^{\infty} \left(\frac{x}{2}\right)^n.$$

For any $\lambda \in (0, \frac{1}{2R}]$, we could get a finite upper bound for $\ln \mathbb{E}[e^{\lambda X}]$:

$$\ln \mathbb{E}[e^{\lambda X}] \leq \lambda \mathbb{E}X + \lambda^2 \frac{1}{2} \sum_{n=0}^{\infty} (\lambda R)^n \text{Var}[X] = \lambda \mathbb{E}X + \frac{\lambda^2 \text{Var}[X]}{2(1 - \lambda R)}. \quad (\text{A.1})$$

Similar to Lemma 2, we let

$$V_\tau(\lambda) = \lambda \sum_{s=1}^{\tau} X_s - \sum_{s=1}^{\tau} \ln \mathbb{E}_{S_t} e^{\lambda X_s},$$

where τ is some stopping time. By Lemma 1 we have $\mathbb{E}(\exp^{V_\tau(\lambda)}) = 1$. (In this case, we apply $Z_s = \xi_s = X_s$ in Lemma 1). Now we define the stopping time τ as

$$\tau = \min(T, \min(n : V_n(\lambda) \geq -\ln \delta)).$$

Then it follows that

$$\mathbb{P}(\exists n : V_\tau(\lambda) \geq -\ln \delta) \leq \mathbb{E}[e^{V_\tau(\lambda) + \ln \delta}] = \delta \mathbb{E}[V_\tau(\lambda)] = \delta,$$

where the first inequality is by the famous Markov Inequality.

By considering the complementary event, we know with probability at least $1 - \delta$, the following inequality holds

$$\sum_{s=1}^T X_s \leq \frac{1}{\lambda} \left(\sum_{s=1}^T \ln \mathbb{E} [e^{\lambda X_s} | \mathcal{F}_{s-1}] + \ln \frac{1}{\delta} \right).$$

Then we take $\lambda = \eta \in (0, \frac{1}{2R}]$ and use equation (A.1) to prove the original statement:

$$\begin{aligned} \sum_{s=1}^T X_s &\leq \sum_{t=1}^T \mathbb{E}(X_t | \mathcal{F}_{t-1}) + \frac{\eta \sum_{s=1}^T \text{Var} [X_s | \mathcal{F}_{s-1}]}{2(1 - \eta R)} + \frac{\ln \frac{1}{\delta}}{\eta} \\ &\leq \sum_{t=1}^T \mathbb{E}(X_t | \mathcal{F}_{t-1}) + \eta \sum_{s=1}^T \text{Var} [X_s | \mathcal{F}_{s-1}] + \frac{\ln \frac{1}{\delta}}{\eta}. \end{aligned}$$

By letting $X'_s = -X_s$, we could easily get

$$\sum_{t=1}^T \mathbb{E}(X_t | \mathcal{F}_{t-1}) \leq \sum_{s=1}^T X_s + \eta \sum_{s=1}^T \text{Var} [X_s | \mathcal{F}_{s-1}] + \frac{\ln \frac{1}{\delta}}{\eta}.$$

□

Lemma 4 (Elliptical Potential Lemma). *Let $\{x_s\}_{s \in [k]}$ be a sequence of vectors with $x_s \in \mathcal{V}$ for some Hilbert space \mathcal{V} . Let Λ_0 be a positive definite matrix and define $\Lambda_k = \Lambda_0 + \sum_{s=1}^k x_s x_s^\top$. Then it holds that*

$$\sum_{s=1}^k \min \left\{ 1, \|x_s\|_{\Lambda_s^{-1}}^2 \right\}^2 \leq 2 \ln \left(\frac{\det(\Lambda_{k+1})}{\det(\Lambda_0)} \right).$$

Proof. This proof mainly follows Lemma 11 in Abbasi-Yadkori et al. (2011). By simple calculation, we have

$$\begin{aligned} \det(\Lambda_k) &= \det(\Lambda_{k-1} + x_k x_k^\top) = \det(\Lambda_{k-1}) \det(I + \Lambda_{k-1}^{-\frac{1}{2}} x_k (\Lambda_{k-1}^{-\frac{1}{2}} x_k)^\top) \\ &= \det(\Lambda_{k-1}) (1 + \|x_{n-1}\|_{\Lambda_{k-1}^{-1}}^2) = \det(\Lambda_0) \prod_{s=1}^k (1 + \|x_s\|_{\Lambda_{s-1}^{-1}}^2), \end{aligned}$$

where we use the fact that all eigenvalues of a matrix of the form $I + xx^\top$ are 1 except one eigenvalue, which is $1 + \|x\|_2^2$ and which corresponds to the eigenvector x . Using $\log(1 + t) \leq t$, we can bound $\log(\det(\Lambda_k))$ by

$$\log \det(\Lambda_k) \leq \log \det(\Lambda_0) + \sum_{s=1}^k \|x_s\|_{\Lambda_{s-1}^{-1}}^2.$$

Combining $x \leq 2 \log(1 + x)$ when $x \in [0, 1]$, we get

$$\sum_{s=1}^k \min \left(1, \|x_s\|_{\Lambda_{s-1}^{-1}}^2 \right) \leq 2 \sum_{t=1}^n \log \left(1 + \|x_s\|_{\Lambda_{s-1}^{-1}}^2 \right) = 2 \ln \left(\frac{\det(\Lambda_k)}{\det(\Lambda_0)} \right).$$

□

Lemma 5. (Lemma G.2 of Chen et al. (2023)) *We consider a fixed policy π and Let \tilde{Q} be an estimate of Q^π . We define a V-function V and an advantage function \tilde{A} by letting*

$$\tilde{V}_h(s) = \frac{1}{\eta} \log \left(\sum_{b \in \mathcal{B}} \exp(\eta \cdot \tilde{Q}_h(s, b)) \right), \quad \tilde{A}_h(s, a) = \tilde{Q}_h(s, b) - \tilde{V}_h(s).$$

Furthermore, we define a follower's policy \tilde{v} be letting $\tilde{v}_h(b|s) = \exp(\eta \cdot \tilde{A}_h(s, b))$. Then we have

$$\mathbb{D}_H(v^\pi, \tilde{v}) \geq \frac{\eta^2}{8(1 + \eta B_A)^2} \cdot \langle v^\pi, (\tilde{A} - A)^2 \rangle_{\mathcal{B}}.$$

where $B_A = 2(\eta^{-1} \log |\mathcal{B}| + 1)$.

Lemma 6. For any $h \in [H]$ and $(s_h, b_h) \in \mathcal{S} \times \mathcal{B}$, using the same notation as in Lemma 5 we have

$$A_h^\pi(s_h, b_h) - \tilde{A}_h(s_h, b_h) = (\mathbb{E}_{s_h, b_h} - \mathbb{E}_{s_h}) [Q_h(s_h, b_h)^\pi - \tilde{Q}_h(s_h, b_h)] + \frac{1}{\eta} \text{KL}(v_h^\pi \| \tilde{v}_h).$$

Proof. This proof mainly follows Lemma G.4 in Chen et al. (2023). At first, we notice the fact that

$$\frac{1}{\eta} \mathcal{H}(v_h^\pi) = -\frac{1}{\eta} \langle v_h^\pi, \log v_h^\pi \rangle_{\mathcal{B}} = -\langle v_h^\pi, Q_h^\pi(s_h, b_h) - V_h^\pi(s_h) \rangle_{\mathcal{B}}, \quad (\text{A.2})$$

$$\frac{1}{\eta} \mathcal{H}(\tilde{v}_h) = -\frac{1}{\eta} \langle \tilde{v}_h, \log \tilde{v}_h \rangle_{\mathcal{B}} = -\langle \tilde{v}_h, \tilde{Q}_h(s_h, b_h) - \tilde{V}_h(s_h) \rangle_{\mathcal{B}}. \quad (\text{A.3})$$

Then we could write the difference of V-functions as

$$\begin{aligned} & V_h^\pi(s_h) - \tilde{V}_h(s_h) \\ &= \langle v_h^\pi, V_h^\pi(s_h) \rangle_{\mathcal{B}} - \langle \tilde{v}_h, \tilde{V}_h(s_h) \rangle_{\mathcal{B}} \\ &= \langle v_h^\pi, Q_h^\pi(s_h, b_h) \rangle_{\mathcal{B}} + \frac{1}{\eta} \mathcal{H}(v_h^\pi) - \langle \tilde{v}_h, \tilde{Q}_h(s_h, b_h) \rangle_{\mathcal{B}} - \frac{1}{\eta} \mathcal{H}(\tilde{v}_h) \\ &= \langle v_h^\pi, Q_h^\pi(s_h, b_h) - \tilde{Q}_h(s_h, b_h) \rangle_{\mathcal{B}} + \langle v_h^\pi - \tilde{v}_h, \tilde{Q}_h(s_h, b_h) \rangle_{\mathcal{B}} \\ &\quad - \langle v_h^\pi, Q_h^\pi(s_h, b_h) - V_h^\pi(s_h) \rangle_{\mathcal{B}} + \langle \tilde{v}_h, \tilde{Q}_h(s_h, b_h) - \tilde{V}_h(s_h) \rangle_{\mathcal{B}}, \end{aligned}$$

where the first equality exploits the fact that $V_h(s_h)$ is constant w.r.t. $b_h \in \mathcal{B}$ and v_h^π, \tilde{v}_h are probability distributions on \mathcal{B} ; the second equality is by equation (A.2); the last equality is by simple algebraic tricks.

Then, by direct calculation and omitting (s_h, b_h) for Q_h^π, \tilde{Q}_h and (s_h) for V_h, \tilde{V}_h , we have

$$-\langle v_h^\pi, Q_h^\pi - V_h^\pi - (\tilde{Q}_h - \tilde{V}_h) \rangle_{\mathcal{B}} = \langle v_h^\pi - \tilde{v}_h, \tilde{Q}_h \rangle_{\mathcal{B}} - \langle v_h^\pi, Q_h^\pi - V_h^\pi \rangle_{\mathcal{B}} + \langle \tilde{v}_h, \tilde{Q}_h - \tilde{V}_h \rangle_{\mathcal{B}},$$

where we use the fact $\langle v_h^\pi, \tilde{V}_h \rangle_{\mathcal{B}} = \langle \tilde{v}_h, \tilde{V}_h \rangle_{\mathcal{B}}$, since \tilde{V}_h is a constant w.r.t. $b_h \in \mathcal{B}$. Therefore, we can write $V_h^\pi(s_h) - \tilde{V}_h(s_h)$ as

$$\begin{aligned} & V_h^\pi(s_h) - \tilde{V}_h(s_h) \\ &= \langle v_h^\pi, Q_h^\pi(s_h, b_h) - \tilde{Q}_h(s_h, b_h) \rangle_{\mathcal{B}} - \langle v_h, Q_h^\pi(s_h, b_h) - V_h^\pi(s_h) - (\tilde{Q}_h(s_h, b_h) - \tilde{V}_h(s_h)) \rangle_{\mathcal{B}} \\ &= \langle v_h^\pi, Q_h^\pi(s_h, b_h) - \tilde{Q}_h(s_h, b_h) \rangle_{\mathcal{B}} - \langle v_h, A_h^\pi(s_h, b_h) - \tilde{A}_h(s_h, b_h) \rangle_{\mathcal{B}} \\ &= \langle v_h^\pi, Q_h^\pi(s_h, b_h) - \tilde{Q}_h(s_h, b_h) \rangle_{\mathcal{B}} - \frac{1}{\eta} \text{KL}(v_h^\pi \| \tilde{v}_h)_{\mathcal{B}}. \end{aligned}$$

We notice the fact that $\text{KL}(v_h^\pi \| \tilde{v}_h) = \eta \langle v_h^\pi, A_h^\pi(s_h, b_h) - \tilde{A}_h(s_h, b_h) \rangle_{b \in \mathcal{B}}$. At last, we could get

$$A_h^\pi(s_h, b_h) - \tilde{A}_h(s_h, b_h) = (\mathbb{E}_{s_h, b_h} - \mathbb{E}_{s_h}) [Q_h^\pi(s_h, b_h) - \tilde{Q}_h(s_h, b_h)] + \frac{1}{\eta} \text{KL}(v_h^\pi \| \tilde{v}_h).$$

□

Lemma 7. We define a distance ρ_1 on Θ by letting

$$\rho_1(\theta, \tilde{\theta}) := \max_{\pi \in \Pi, s_h \in \mathcal{S}, h \in [H]} \left\{ D_H \left(v_h^{\pi, \theta}(\cdot | s_h), v_h^{\pi, \tilde{\theta}}(\cdot | s_h) \right), (1 + \eta) \left\| r_h^{\pi, \theta} - r_h^{\pi, \tilde{\theta}} \right\|_{\infty} \right\}. \quad (\text{A.4})$$

Let $N_{\rho_1}(\theta, \epsilon)$ be the ϵ -covering number of Θ with respect to the distance ρ_1 . For any $\delta \in (0, 1)$, we set $\beta_1 = 2 \ln(H \cdot N(\Theta, T^{-1})/\delta) + 8$. For $\forall \theta \in \Theta, \forall h \in [H]$,

$$\sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \text{Var}_{s_h}^{\pi^i, \theta^*} \left[r_h^{\pi^i, \theta}(s_h, b_h) - r_h^{\pi^i, \theta^*}(s_h, b_h) \right] \leq 4C_\eta^2 (L_{h,1}^t(\theta) - L_{h,1}^t(\theta^*)) + \beta,$$

where we define

$$\begin{aligned} \text{Var}_{s_h}^{\pi, \theta} [Z] &= \text{Var}^{\pi, \theta} [Z | s_h] = \mathbb{E}^{\pi, v_h^{\pi, \theta}} [(Z - \mathbb{E}^{\pi, v_h^{\pi, \theta}} [Z | s_h])^2 | s_h], \\ C_\eta &= \frac{1}{\eta} + B_A, B_A = 2(\eta^{-1} \log |\mathcal{B}| + 1). \end{aligned}$$

Proof. We first exploit Lemma 2 with $X_t^h = \frac{1}{2}(\log v_h^{\pi_i, \theta}(s_h^t | b_h^t) - \log v_h^{\pi_i, \theta^*}(s_h^t | b_h^t))$. We choose filtration to be $\mathcal{F}_{h:t-1} \{X_i^h : i \in [t-1]\}$. Let $\mathcal{N}_{\rho_1}(\Theta, \epsilon)$ be the covering number for the ϵ -covering net of Θ with respect to norm ρ_1 defined in A.4. Let Θ_ϵ be the ϵ -covering net of Θ . By Lemma 2 w.p. at least $1 - \delta$, for a fixed $\theta \in \Theta_\epsilon$ and a fixed $h \in [H]$, we have

$$\begin{aligned} \sum_{t=1}^{t-1} X_t^h &= \frac{1}{2}(L_{h,1}^t(\theta^*) - L_{h,1}^t(\theta)) \\ &\stackrel{(a)}{\leq} \sum_{t=1}^{t-1} \log \mathbb{E}(e^{X_t} | \mathcal{F}_{t-1}) + \frac{1}{\delta} \\ &\stackrel{(b)}{=} \sum_{i=1}^{t-1} \log \mathbb{E}^{\pi^i} \left[\sqrt{\frac{v^{\pi_i, \theta}(\cdot | s_h)}{v^{\pi_i, \theta^*}(\cdot | s_h)}} \right] + \frac{1}{\delta} \\ &\stackrel{(c)}{\leq} - \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \left[\mathbb{D}_{\mathbb{H}}^2(v^{\pi_i, \theta}(\cdot | s_h), v^{\pi_i, \theta^*}(\cdot | s_h)) \right] + \frac{1}{\delta}, \end{aligned}$$

where the first equality is by the definition of $L_{h,1}^t$; (a) is by Lemma 2; (b) is by the definition of X_t ; (c) is by the fact that $\log(x) \leq x - 1$ and the definition of Hellinger distance.

By taking union bound on $\theta \in \Theta_\epsilon$ and $h \in [H]$, we have for any $\theta \in \Theta$, any $h \in [H]$, with probability at least $1 - \delta$, for $\forall t \in [T]$

$$\frac{1}{2}(L_{h,1}^t(\theta^*) - L_{h,1}^t(\theta)) \leq - \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \left[\mathbb{D}_{\mathbb{H}}^2(v^{\pi_i, \theta}(\cdot | s_h), v^{\pi_i, \theta^*}(\cdot | s_h)) \right] + \frac{\log(H\mathcal{N}_{\rho}(\Theta, \epsilon))}{\delta}. \quad (\text{A.5})$$

On the other hand, by the definition of ρ_1 in equation A.4, for any $\theta, \tilde{\theta} \in \Theta$, we have

$$\begin{aligned} &\left| \mathbb{D}_{\mathbb{H}}^2(v_h^{\pi, \theta}, v_h^{\pi, \theta^*}) - \mathbb{D}_{\mathbb{H}}^2(v_h^{\pi, \tilde{\theta}}, v_h^{\pi, \theta^*}) \right| \\ &\stackrel{(a)}{=} \left| \mathbb{D}_{\mathbb{H}}(v_h^{\pi, \theta}, v_h^{\pi, \theta^*}) + \mathbb{D}_{\mathbb{H}}(v_h^{\pi, \tilde{\theta}}, v_h^{\pi, \theta^*}) \right| \cdot \left| \mathbb{D}_{\mathbb{H}}(v_h^{\pi, \theta}, v_h^{\pi, \theta^*}) - \mathbb{D}_{\mathbb{H}}(v_h^{\pi, \tilde{\theta}}, v_h^{\pi, \theta^*}) \right| \\ &\stackrel{(b)}{\leq} 2\mathbb{D}_{\mathbb{H}}(v_h^{\pi, \tilde{\theta}}, v_h^{\pi, \theta}) \\ &\stackrel{(c)}{\leq} 2\rho_1(\theta, \tilde{\theta}), \end{aligned}$$

where (a) is by the fact that $a^2 - b^2 = (a+b)(a-b) \leq |a+b||a-b|$; (b) is by the fact that $\mathbb{D}_{\mathbb{H}}(\cdot, \cdot) \leq 1$; (c) is by the definition of ρ_1 . Then noting that $L_{h,1}^t(\theta) = - \sum_{i=1}^t \eta A_h^{\pi^i, \theta}(s_h^i, b_h^i)$, for any $\theta, \tilde{\theta} \in \Theta$, we have

$$\begin{aligned} \left| L_{h,1}^t(\theta) - L_{h,1}^t(\tilde{\theta}) \right| &\leq \eta T \max_{i \in [t-1]} \left| A_h^{\pi^i, \theta}(s_h^i, b_h^i) - A_h^{\pi^i, \tilde{\theta}}(s_h^i, b_h^i) \right| \\ &\leq 2\eta T \max_{i \in [t-1]} \left\| r_h^{\pi^i, \theta} - r_h^{\pi^i, \tilde{\theta}} \right\|_{\infty} \\ &\leq 2T \cdot \rho_1(\theta, \tilde{\theta}), \end{aligned}$$

where the second inequality uses the fact that $\left| (V_h^{\pi, \theta} - V_h^{\pi, \tilde{\theta}})(s_h) \right| \leq \left\| r_h^{\pi, \theta} - r_h^{\pi, \tilde{\theta}} \right\|_{\infty}$; and the last inequality is by the definition of ρ_1 . Therefore, all the error terms in $\mathbb{D}_{\mathbb{H}}^2(\cdot, \cdot)$, $L_{h,1}^t(\theta^*)$ and $L_{h,1}^t(\theta)$ induced by ϵ -net could be bounded by $2T\epsilon$. By adding an extra $4T\epsilon$ in equation A.5, we have for all $\theta \in \Theta, h \in [H], t \in [T]$, w.p. $1 - \delta$,

$$\frac{1}{2}(L_{h,1}^t(\theta^*) - L_{h,1}^t(\theta)) \leq - \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \left[\mathbb{D}_{\mathbb{H}}^2(v^{\pi_i, \theta}(\cdot | s_h), v^{\pi_i, \theta^*}(\cdot | s_h)) \right] + \frac{\log(H\mathcal{N}_{\rho}(\Theta, \epsilon))}{\delta} + 4T\epsilon. \quad (\text{A.6})$$

In the rest of the proof we take $\epsilon = \frac{1}{T}$ and let $\beta_1 = 2 \log(HN_\rho(\Theta, T^{-1})/\delta) + 8$. By Lemma 5 we have

$$\begin{aligned} 8D_H^2 \left(v^{\pi_t, \theta}(\cdot|s_h), v^{\pi_t, \theta^*}(\cdot|s_h) \right) &\geq \left(\frac{\eta}{1 + \eta B_A} \right)^2 \cdot \left\langle v_h^{\pi_t, \theta^*}, (A_h^{\pi_t, \theta} - A_h^{\pi_t, \theta^*})^2 \right\rangle \\ &\geq \left(\frac{\eta}{1 + \eta B_A} \right)^2 \cdot \mathbb{E}_{s_h}^{\pi_t, \theta^*} \left(A_h^{\pi_t, \theta} - A_h^{\pi_t, \theta^*} \right)^2 \\ &\geq \left(\frac{\eta}{1 + \eta B_A} \right)^2 \cdot \mathbb{E}_{s_h}^{\pi_t, \theta^*} \left((\mathbb{E}_{s_h, b_h}^{\pi_t, \theta^*} - \mathbb{E}_{s_h}^{\pi_t, \theta^*}) [r_h^{\pi_t, \theta} - r_h^{\pi_t, \theta^*}] \right)^2 \\ &= \left(\frac{\eta}{1 + \eta B_A} \right)^2 \cdot \text{Var}_{s_h}^{\pi_t, \theta^*} [r_h^{\pi_t, \theta}(s_h, b_h) - r_h^{\pi_t, \theta^*}(s_h, b_h)], \end{aligned}$$

where the second inequality is by Jensen's inequality of x^2 ; the last inequality is by Lemma 6; the last equality is by the definition of $\text{Var}_{s_h}^{\pi_t, \theta^*}(\cdot)$. Therefore, by letting $C_\eta = \frac{1}{\eta} + B_A$ and insert the above result back to equation (A.6), we have

$$\sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \text{Var}_{s_h}^{\pi^i, \theta^*} \left[r_h^{\pi^i, \theta}(s_h, b_h) - r_h^{\pi^i, \theta^*}(s_h, b_h) \right] \leq 4C_\eta^2 (L_{h,1}^t(\theta) - L_{h,1}^t(\theta^*)) + \beta.$$

□

Lemma 8. Let $\mathcal{F}_h = U_h \times U_{h+1} \times \Theta$, we define the following distance on for $f, \tilde{f} \in \mathcal{F}_h$:

$$\rho_2(f, \tilde{f}) := \max_{h \in [H]} \left\{ \|U_h - \tilde{U}_h\|_\infty, \left\| T_{h+1}^{\star, \theta} U(h+1)(\cdot) - T_{h+1}^{\star, \tilde{\theta}} \tilde{U}(h+1)(\cdot) \right\|_\infty \right\}. \quad (\text{A.7})$$

Let $N_{\rho_2}(\theta, \epsilon)$ be the ϵ -covering number of \mathcal{F} with respect to the distance ρ_2 . For any $\delta \in (0, 1)$, we set $\beta_2 = 4H^2 \ln\left(\frac{HN_{\rho_2}(\mathcal{F}, \epsilon)}{\delta}\right) + 5$. For $\forall \{f_h^t\}_{h \in [H], t \in [T]} \subseteq \mathcal{F}$

$$L_{h,2}^{t-1}(f_h^*) - L_{h,2}^{t-1}(f_h^t) \leq -\frac{1}{2} \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \left[\left(U_h - \mathbb{T}_{h+1}^{\star, \theta^i} U_{h+1} \right) (s_h, a_h, b_h)^2 \right] + \beta_2.$$

Proof. At first we verify our loss l_h^t satisfies generalized Bellman completeness and boundedness, which is defined as follows:

Assumption 4. The function $l : \mathcal{U}_h \times \mathcal{U}_{h+1} \times \Theta \times (\mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathbb{R} \times \mathcal{S}) \rightarrow \mathbb{R}$ satisfies:

1. (Generalized Bellman Completeness) There exists a functional operator $\mathcal{P}_h : \mathcal{H}_{h+1} \rightarrow \mathcal{H}_h$ such that for any $(U_h, U_{h+1}, \theta) \in \mathcal{H}_h \times \mathcal{H}_{h+1} \times \Theta$ and $D_h = (s_h, a_h, b_h, s_{h+1}) \in (\mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathbb{R} \times \mathcal{S})$.

$$l(U_h, U_{h+1}, \theta; D_h) - l(\mathcal{P}_h U_{h+1}, U_{h+1}, \theta; D_h) = \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, b_h)} [l(U_h, U_{h+1}, \theta; D_h)],$$

where we require $\mathcal{P}_h U_{h+1}^* = U_h^*$ and that $\mathcal{P}_h U_{h+1} \in \mathcal{H}_h$ for any $U_{h+1} \in \mathcal{U}_{h+1}$ and $h \in [H]$;

2. (Boundedness) It holds that $|l(U_h, U_{h+1}, \theta; D_h)| \leq B_l$ for some $B_l > 0$ and for any $(U_h, U_{h+1}, \theta) \in \mathcal{H}_h \times \mathcal{H}_{h+1} \times \Theta$ and $D_h = (s_h, a_h, b_h, s_{h+1}) \in (\mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathbb{R} \times \mathcal{S})$.

First we verify the Generalized Bellman Completeness:

$$\begin{aligned} l_h^t(U_h, U_{h+1}, \theta; D_h^t) - \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, b_h)} [l_h^t(U_h, U_{h+1}, \theta; D_h)] \\ = [(U_h - u_h)(s_h, a_h, b_h) - T^{\star, \theta}(s_{h+1})] - [U_h(s_h, a_h, b_h) - \mathbb{T}_{h+1}^{\star, \theta}(U_{h+1})] \\ = \mathbb{T}_{h+1}^{\star, \theta}(U_{h+1}) - T^{\star, \theta}(U_{h+1})(s_{h+1}) \\ = (\mathbb{E}_{s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h, b_h)} [T^{\star, \theta}(U_{h+1})(s_{h+1})] - u_h)(s_h, a_h, b_h) - T^{\star, \theta}(U_{h+1})(s_{h+1}). \end{aligned}$$

Therefore, the operator \mathcal{P}_h is $\mathbb{E}_{s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h, b_h)} [T^{\star, \theta}(\cdot)(s_{h+1})]$ and the generalized Bellman completeness holds. To check boundedness, we only need to notice that $u_h \in [0, 1], \forall h \in [H]$, so

$|l_h^t(U_h, U_{h+1}, \theta; D_h^t)| \leq H, \forall h \in [H]$. Then we generalize the proof of Proposition 5.1 in Liu et al. (2024a) to show our wanted result.

We define the random variables $X_{h,f}^t$ as

$$X_{h,f}^t = l_h^t(U_h, U_{h+1}, \theta; D_h^t)^2 - l_h^t(\mathcal{P}_h U_{h+1}, U_{h+1}, \theta; D_h^t)^2. \quad (\text{A.8})$$

For any $f = (U_h, U_{h+1}, \theta) \in \mathcal{U}_h \times \mathcal{U}_{h+1} \times \Theta$ and the operator \mathcal{P}_h is defined as above. We first show $X_{h,f}^t$ is an unbiased estimator of the discrepancy function $d_h^t(U_h, U_{h+1}; D_h^t)^2$, which is defined as

$$d_h^t(f; D_h^t) = \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [l_h^t(f; D_h^t)] = U_h - \mathbb{T}_h^{\star, \theta}(U_{h+1}).$$

For simplicity we also let $f_{\mathcal{P}} = (\mathcal{P}_h U_{h+1}, U_{h+1}, \theta)$. Consider that

$$\begin{aligned} l_h^t(f; D_h^t)^2 &= (l_h^t(f; D_h^t) - l_h^t(f_{\mathcal{P}}; D_h^t) + l_h^t(\mathcal{P}_h U_{h+1}, U_{h+1}, \theta; D_h^t))^2 \\ &= (d_h^t(f; D_h^t) + l_h^t(f_{\mathcal{P}}; D_h^t))^2 \\ &= (d_h^t(f; D_h^t))^2 + l_h^t(f_{\mathcal{P}}; D_h^t)^2 + 2d_h^t(f; D_h^t) \cdot l_h^t(f_{\mathcal{P}}; D_h^t), \end{aligned} \quad (\text{A.9})$$

where the second equality is by the generalized Bellman completeness. Exploiting the completeness again, we have

$$\begin{aligned} &\mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [d_h^t(f; D_h^t) \cdot l_h^t(f_{\mathcal{P}}; D_h^t)] \\ &= d_h^t(f; D_h^t) \cdot \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [l_h^t(\mathcal{P}_h U_{h+1}, U_{h+1}, \theta; D_h^t)] \\ &= d_h^t(f; D_h^t) \cdot \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [d_h^t(f; D_h^t) - l_h^t(f; D_h^t)] \\ &= 0. \end{aligned}$$

Inserting the result back to A.9 we have

$$\mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [X_{h,f}^t] = d_h^t(f; D_h^t)^2. \quad (\text{A.10})$$

Then for each time step h , we define the filtration $\{\mathcal{F}_{h,t}\}_{t=1}^T$ with

$$\mathcal{F}_{h,t} = \sigma \left(\sum_{s=1}^k \sum_{h=1}^H D_h^s \right),$$

where $D_h^t = \{s_h^t, a_h^t, b_h^t, u_h^t, s_{h+1}^t\}$. From the previous arguments, we can derive that

$$\mathbb{E}[X_{h,f}^t | \mathcal{F}_{h,t-1}] = \mathbb{E} \left[\mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [X_{h,f}^t] | \mathcal{F}_{h,t-1} \right] = \mathbb{E}^{\pi^t} [d_h^t(f; D_h^t)^2], \quad (\text{A.11})$$

$$\text{Var} [X_{h,f}^t | \mathcal{F}_{h,t-1}] \leq \mathbb{E} [X_{h,f}^t | \mathcal{F}_{h,t-1}] \leq B_l^2 \mathbb{E} [X_{h,f}^t | \mathcal{F}_{h,t-1}] = B_l^2 \mathbb{E}^{\pi^t} [d_h^t(f; D_h^t)^2], \quad (\text{A.12})$$

where \mathbb{E}^{π^t} means the data D_h^t is generated by measure $(\pi^t, \nu_h^{\pi^t}, P_h)$. By Lemma 3, $(|X_{h,f}^t| \leq B_l^2)$ and we set $\eta = \frac{1}{2B_l^2}$, for any fixed $h \in [H], t \in [T], U \in \mathcal{U}$, we have

$$\begin{aligned} \left| \sum_{s=1}^{t-1} \mathbb{E} [X_{h,f}^s | \mathcal{F}_{h,t-1}] - \sum_{s=1}^{t-1} X_{h,f}^s \right| &\leq \frac{1}{2B_l^2} \sum_{s=1}^{t-1} \text{Var} [X_{h,f}^s | \mathcal{F}_{h,s-1}] + 2B_l^2 \log \left(\frac{1}{\delta} \right) \\ &\leq \frac{1}{2} \sum_{s=1}^{t-1} \mathbb{E}^{\pi^s} [d_h^s(f; D_h^s)^2] + 2B_l^2 \log \left(\frac{1}{\delta} \right). \end{aligned}$$

Rearranging the above terms, we can get

$$-\sum_{s=1}^{t-1} X_{h,f}^s \leq -\frac{1}{2} \mathbb{E}^{\pi^t} [d_h^t(f; D_h^t)^2] + 2B_l^2 \log \left(\frac{1}{\delta} \right).$$

By the definition of $X_{h,f}^t$ and the loss function $L_{h,2}^t$ in (4.8), we have

$$\begin{aligned} \sum_{s=1}^{t-1} X_{h,f}^s &= \sum_{s=1}^{k-1} l_h^s(f; D_h^s)^2 - \sum_{s=1}^{k-1} l_h^s(\mathcal{P}_h U_{h+1}, U_{h+1}, \theta; D_h^s)^2 \\ &\leq \sum_{s=1}^{k-1} l_h^s(f; D_h^s)^2 - \inf_{U'_h \in \mathcal{U}_h} l_h^s(U'_h, U_{h+1}, \theta; D_h^s)^2 \\ &= L_{h,2}^t(f). \end{aligned}$$

Then we can derive that, for any fixed $h \in [H], t \in [T], f \in \mathcal{U}_h \times \mathcal{U}_{h+1} \times \Theta$.

$$-L_{h,2}^t(f) \leq -\frac{1}{2} \mathbb{E}^{\pi^t} [d_h^t(f; D_h^t)^2] + 2H^2 \log\left(\frac{1}{\delta}\right). \quad (\text{A.13})$$

Then we consider $L_{h,2}^t(f^*)$. We first define the random variables $Y_{h,f}^t$ as

$$Y_{h,f}^t = l_h^t(U_h, U_{h+1}^*, \theta^*; D_h^t)^2 - l_h^t(f^*; D_h^t)^2.$$

Similarly, we could show

$$\mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} [Y_{h,f}^t] = (d_h^t(U_h, U_{h+1}^*, \theta^*; D_h^t))^2.$$

Under the filtration $\{\mathcal{F}_{h,t}\}_{t=1}^T$, we can derive that

$$\begin{aligned} \mathbb{E}[Y_{h,f}^t | \mathcal{F}_{h,t-1}] &= \mathbb{E}^{\pi^t} [d_h^t(U_h, U_{h+1}^*, \theta^*; D_h^t)^2], \\ \text{Var} [Y_{h,f}^t | \mathcal{F}_{h,t-1}] &\leq B_l^2 \mathbb{E}^{\pi^t} [d_h^t(U_h, U_{h+1}^*, \theta^*; D_h^t)^2]. \end{aligned}$$

By Lemma 3 ($|Y_{h,f}^t| \leq B_l^2$ and we set $\eta = \frac{1}{2B_l^2}$), for any fixed $h \in [H], t \in [T], f \in \mathcal{F}$, we have

$$-\sum_{s=1}^{t-1} Y_{h,f}^s \leq -\frac{1}{2} \sum_{s=1}^{t-1} \mathbb{E}^{\pi^s} [d_h^s(U_h, U_{h+1}^*, \theta^*; D_h^s)^2] + 2B_l^2 \log\left(\frac{1}{\delta}\right) \leq 2B_l^2 \log\left(\frac{1}{\delta}\right).$$

By the definition of $Y_{h,f}^t$ and the loss function $L_{h,2}^t$ in (4.8), we have

$$-\sum_{s=1}^{t-1} Y_{h,f}^s = \sum_{s=1}^{k-1} l_h^s(f^*; D_h^s)^2 - \sum_{s=1}^{k-1} l_h^s(U_h, U_{h+1}^*, \theta^*; D_h^s)^2.$$

Since such inequality holds for any $U_h \in \mathcal{U}_h$, we have

$$L_{h,2}^t(f^*) = \sup_{U_h \in \mathcal{U}_h} \left(-\sum_{s=1}^{t-1} Y_{h,f}^s \right) \leq 2B_l^2 \log\left(\frac{1}{\delta}\right).$$

Combining the above result with (A.13), for any fixed $h \in [H], t \in [T], f \in \mathcal{F}$, we have

$$L_{h,2}^t(f^*) - L_{h,2}^t(f) \leq -\frac{1}{2} \sum_{s=1}^{t-1} \mathbb{E}^{\pi^s} [U_h^s(s_h, a_h, b_h) - \mathbb{T}^{\pi^s, \theta^s}(U_{h+1}^s)(s_h, a_h, b_h)] + 4H^2 \log\left(\frac{1}{\delta}\right). \quad (\text{A.14})$$

Then we generalize this result on a ϵ -net \mathcal{F}_ϵ of \mathcal{F} . By taking union bound over all $h \in [H], \tilde{f} \in \mathcal{F}_\epsilon$ and a $\tilde{f}^* \in \Theta_\epsilon$ such that $\rho_2(f^*, \tilde{f}^*) \leq \epsilon$, with probability $1 - \delta$ we have for any $t \in [T]$

$$\begin{aligned} &L_{h,2}^t(\tilde{f}^*) - L_{h,2}^t(\tilde{f}) \\ &\leq -\frac{1}{2} \sum_{s=1}^{t-1} \mathbb{E}^{\pi^s} [\tilde{U}_h^s(s_h, a_h, b_h) - \mathbb{T}^{\pi^s, \tilde{\theta}^s}(\tilde{U}_{h+1}^s)(s_h, a_h, b_h)] + 4H^2 \log\left(\frac{HN_{\rho_2}(\mathcal{F}, \epsilon)}{\delta}\right). \quad (\text{A.15}) \end{aligned}$$

By the definition of ρ_2 , we know

$$\begin{aligned}
& \left| L_{h,2}^t(\tilde{f}^\star) - L_{h,2}^t(f^\star) \right| \\
&= \left| \sum_{s=1}^t [(\tilde{U}_h - u_h^s)(s_h^s, a_h^s, b_h^s) - T_{h+1}^{\star, \tilde{\theta}}(s_{h+1}^s)] - [U_h(s_h^s, a_h^s, b_h^s) - T_{h+1}^{\star, \theta}(U_{h+1})] \right| \\
&\leq \left| \sum_{s=1}^t [(\tilde{U}_h - U_h)(s_h^s, a_h^s, b_h^s) - (T_{h+1}^{\star, \tilde{\theta}} - T_{h+1}^{\star, \theta})(s_{h+1}^s)] \right| \\
&\leq t(\|\tilde{U}_h - U_h\|_\infty + \|T_{h+1}^{\star, \tilde{\theta}} - T_{h+1}^{\star, \theta}\|_\infty) \\
&\leq 2T\rho_2(f, \tilde{f}).
\end{aligned}$$

Similarly we could get

$$\begin{aligned}
& \left| L_{h,2}^t(\tilde{f}) - L_{h,2}^t(f) \right| \leq 2T\rho_2(f, \tilde{f}), \\
& \left| [\tilde{U}_h^s - \mathbb{T}^{\star, \tilde{\theta}^s}(\tilde{U}_{h+1}^s)](s_h, a_h, b_h) - [U_h^s - \mathbb{T}^{\star, \theta^s}(U_{h+1}^s)](s_h, a_h, b_h) \right| \leq 2\rho_2 T(f, \tilde{f}).
\end{aligned}$$

Then we could generate equation (A.15) from \mathcal{F}_ϵ to \mathcal{F} only paying an extra cost of $5T\epsilon$. By setting $\epsilon = 1/T$, for any $h \in [H]$, $t \in [T]$, $f \in \mathcal{F}$, with probability $1 - \delta$ we have

$$\begin{aligned}
& L_{h,2}^t(f^\star) - L_{h,2}^t(f) \\
&\leq -\frac{1}{2} \sum_{s=1}^{t-1} \mathbb{E}^{\pi^s} [(U_h^s - \mathbb{T}^{\star, \theta^s}(U_{h+1}^s))(s_h^s, a_h^s, b_h^s)] + 4H^2 \ln\left(\frac{HTN_{\rho_2}(\mathcal{F}, \epsilon)}{\delta}\right) + 5.
\end{aligned}$$

Let $\beta_2 = 4H^2 \ln\left(\frac{HTN_{\rho_2}(\mathcal{F}, \epsilon)}{\delta}\right) + 5$, then we are done. \square

Lemma 9. (Lemma B.2 in (Chen et al., 2023)) For any fixed policy π and a fixed s_1 , let \tilde{v} be an estimate of the quantal response v^π and let \tilde{U} and \tilde{W} be estimates of U^π and W^π respectively. Based on \tilde{U} and \tilde{W} , we can estimate $J(\pi)$ by $\tilde{W}(s_1)$. Then the error of these estimators can be bounded as follows:

$$\tilde{W}(s_1) - J(\pi) \leq \sum_{h=1}^H \mathbb{E} [\tilde{U}_h(s_h, a_h, b_h) - (\mathbb{T}_h^{\pi, \tilde{v}} \tilde{U}_{h+1})] + H \sum_{h=1}^H \mathbb{E} [\|(v_h^\pi - \tilde{v})(\cdot | s_h)\|_1].$$

where we define

$$\begin{aligned}
\mathbb{T}_h^{\pi, \theta} U(s_h, a_h, b_h) &= u_h(s_h, a_h, b_h) + \mathbb{E}_{s_{h+1} \sim P_h(\cdot | s_h, a_h, b_h)} [(T_{h+1}^{\pi, \theta} U_{h+1})(s_{h+1})], \\
T_h^{\pi, \theta}(U_h)(s_h) &= \langle U_h(s_h, \cdot, \cdot), \pi_h \otimes v_h^{\pi, \theta}(\cdot, \cdot | s_h) \rangle.
\end{aligned}$$

Furthermore, by $\mathbb{T}_h^{\pi, \tilde{v}} \tilde{U}_{h+1} \leq \mathbb{T}_h^{\star, \tilde{v}} \tilde{U}_{h+1}$, we have

$$\tilde{W}(s_1) - J(\pi) \leq \sum_{h=1}^H \mathbb{E} [\tilde{U}_h(s_h, a_h, b_h) - (\mathbb{T}_h^{\star, \tilde{v}} \tilde{U}_{h+1})] + H \sum_{h=1}^H \mathbb{E} [\|(v_h^\pi - \tilde{v})(\cdot | s_h)\|_1].$$

Lemma 10. (Lemma B.1 in (Chen et al., 2023)) We consider a fixed policy π and let \tilde{r} be an estimate of r . We define a V-function \tilde{V} and an advantage function \tilde{A} by letting

$$\tilde{V}_h(s) = \frac{1}{\eta} \log \left(\sum_{b \in \mathcal{B}} \exp(\eta \cdot \tilde{r}_h^\pi(s, b)) \right), \quad \tilde{A}_h(s, a) = \tilde{r}_h^\pi(s, b) - \tilde{V}_h(s).$$

Furthermore, we define a follower's policy \tilde{v} by letting $\tilde{v}_h(b | s) = \exp(\eta \cdot \tilde{A}_h(s, b))$. Then the difference between \tilde{v} and v^π can be bounded by

$$\begin{aligned}
& H \sum_{h=1}^H \mathbb{E} [\|v_h^\pi - \tilde{v}(\cdot | s_h)\|_1] \\
&\leq C_0 \sum_{h=1}^H \mathbb{E} [\|\mathcal{T}_h^\pi(\tilde{r}_h - r_h)\|] + C_1 \sum_{h=1}^H \mathbb{E} [\mathcal{T}_h^\pi(\tilde{r}_h - r_h)^2],
\end{aligned}$$

where C_1 is defined as

$$C_1 = \frac{\eta^2 \exp(2\eta B_A)}{2} (2 + \eta B_A \cdot 2\eta B_A),$$

and \mathcal{T}_h^π has been defined in equation (5.1).

B PROOF OF THEOREM 1

At first, we could decompose the regret into two terms:

$$\begin{aligned} \text{Reg}(T) &= \sum_{t=1}^T W_1^{U^*, \theta^*}(s_1) - W_1^{\pi^t}(s_1) \\ &\leq \underbrace{\sum_{t=1}^T \left(W_1^{U^*, \theta^*}(s_1) - W_1^{U^t, \theta^t}(s_1) \right)}_{I_1} + \underbrace{\sum_{t=1}^T \left(W_1^{U^t, \theta^t}(s_1) - W_1^{\pi^t}(s_1) \right)}_{I_2}. \end{aligned}$$

By the definition of U^t, θ^t in algorithm 1 we have

$$\begin{aligned} W_1^{U^t, \theta^t}(s_1) - \eta_1 \sum_{h=1}^H L_{h,1}^t(\theta_h^t) - \eta_2 \sum_{h=1}^H L_{h,2}^t(U_h^t, \theta_h^t) \\ \geq W_1^{U^*, \theta^*}(s_1) - \eta_1 \sum_{h=1}^H L_{h,1}^t(\theta_h^*) - \eta_2 \sum_{h=1}^H L_{h,2}^t(U_h^*, \theta_h^*), \end{aligned}$$

which implies that

$$W_1^{U^*, \theta^*}(s_1) - W_1^{U^t, \theta^t}(s_1) \leq \eta_1 \sum_{h=1}^H (L_{h,1}^t(\theta_h^*) - L_{h,1}^t(\theta_h^t)) + \eta_2 \sum_{h=1}^H (L_{h,2}^t(U_h^*, \theta_h^*) - L_{h,2}^t(U_h^t, \theta_h^t)).$$

By the lemma 7, set $\beta_1 = 2 \ln \mathcal{N}_\rho(\Theta, 1/T)/\delta + 8$ with the distance ρ defined in Lemma 7, and let $C_\eta = \eta^{-1} + B_A$, $B_A = 2(\eta^{-1} \log |\mathcal{B}| + 1)$, then with probability at least $1 - \delta$,

$$\begin{aligned} &\sum_{h=1}^H (L_{h,1}^t(\theta_h^*) - L_{h,1}^t(\theta_h^t)) \\ &\leq \frac{-1}{4C_\eta^2} \sum_{h=1}^H \sum_{i=1}^t \mathbb{E}^{\pi^i} \text{Var}_{s_h}^{\pi^i, \theta^*} \left[r_h^{\pi^i, \theta^t}(s_h, b_h) - r_h^{\pi^i, \theta^*}(s_h, b_h) \right] + H\beta_1. \end{aligned} \quad (\text{B.1})$$

For the variance term, we have:

$$\begin{aligned} &\mathbb{E}^{\pi^i} \text{Var}_{s_h}^{\pi^i, \theta^*} \left[r_h^{\pi^i, \theta^t}(s_h, b_h) - r_h^{\pi^i, \theta^*}(s_h, b_h) \right] \\ &= \mathbb{E}^{\pi^i} \text{Var}^{\pi^i, \theta^*} \left[r_h^{\pi^i, \theta^t}(s_h, b_h) - r_h^{\pi^i, \theta^*}(s_h, b_h) \mid s_h \right] \\ &\stackrel{(a)}{=} \mathbb{E}^{\pi^i} \mathbb{E}^{\pi^i, \theta^*} \left[\left((r_h^{\pi^i, \theta^t} - r_h^{\pi^i, \theta^*})(s_h, b_h) - \mathbb{E}^{\pi^i, \theta^*} \left[(r_h^{\pi^i, \theta^t} - r_h^{\pi^i, \theta^*})(s_h, b_h) \mid s_h \right] \right)^2 \mid s_h \right] \\ &\stackrel{(b)}{=} \mathbb{E}^{\pi^i} \left[\left(\mathcal{T}_h^{\pi^i}(r_h^{\theta^t} - r_h^*) \right)^2 (s_h, b_h) \right], \end{aligned}$$

where (a) follows from the definition of $\text{Var}_{s_h}^{\pi^i, \theta^*}(\cdot)$, and (b) follows from the definition of $\mathcal{T}_h^\pi(\cdot)$. Insert the last term back to equation (B.1), we have:

$$\sum_{h=1}^H (L_{h,1}^t(\theta_h^*) - L_{h,1}^t(\theta_h^t)) \leq \frac{-1}{4C_\eta^2} \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} [\mathcal{T}_h^{\pi^i}(r_h^{\theta^t} - r_h^*)^2] + H\beta_1.$$

By Lemma 8, set $\beta_2 = 4H^2 \ln(\frac{HN\rho_2(\mathcal{F}, \epsilon)}{\delta}) + 5$ with the distance ρ_2 defined in Lemma 8, we have

$$\sum_{h=1}^H (L_{h,2}^t(U_h^*, \theta_h^*) - L_{h,2}^t(U_h^t, \theta_h^t)) \leq -\frac{1}{2} \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \left[\left(U_h - \mathbb{T}_{h+1}^{*, \theta^t} U_{h+1} \right) (s_h, a_h, b_h)^2 \right] + H\beta_2. \quad (\text{B.2})$$

We then have

$$\begin{aligned} I_1 &\leq \sum_{t=1}^T \left(\eta_1 \sum_{h=1}^H (L_{h,1}^t(\theta_h^*) - L_{h,1}^t(\theta_h^t)) + \eta_2 \sum_{h=1}^H (L_{h,2}^t(U_h^*, \theta_h^*) - L_{h,2}^t(U_h^t, \theta_h^t)) \right) \\ &\leq \eta_1 \cdot \left(-\frac{1}{4C_\eta^2} \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} [\mathcal{T}_h^{\pi^i} (r_h^{\theta^t} - r_h^*)^2] + HT\beta_1 \right) \\ &\quad + \eta_2 \cdot \left(-\frac{1}{2} \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} \left[\left(U_h - \mathbb{T}_{h+1}^{*, \theta^t} U_{h+1} \right) (s_h, a_h, b_h)^2 \right] + HT\beta_2 \right) \end{aligned} \quad (\text{B.3})$$

To bound I_2 , we exploit Lemma 9 and Lemma 10,

$$\begin{aligned} I_2 &\stackrel{(a)}{\leq} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[(U_h^t)(s_h, a_h, b_h) - \mathbb{T}_{h+1}^{*, \theta^t} U_{h+1}^t(s_{h+1}) \right] + H \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[\|(\tilde{v}_h - v_h^\pi)(\cdot | s_h)\|_1 \right] \quad (\text{B.4}) \\ &\stackrel{(b)}{\leq} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[(U_h^t)(s_h, a_h, b_h) - \mathbb{T}_{h+1}^{*, \theta^t} U_{h+1}^t(s_{h+1}) \right] \\ &\quad + \sum_{t=1}^T \sum_{h=1}^H C_0 \cdot \mathbb{E}^{\pi^t} \left[|\mathcal{T}_h^{\pi^t} (r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)| \right] \\ &\quad + \sum_{t=1}^T \sum_{h=1}^H C_1 \cdot \mathbb{E}^{\pi^t} \left[\mathcal{T}_h^{\pi^t} (r_h^{\theta^t} - r_h^*)^2(s_h^t, b_h^t) \right] \end{aligned} \quad (\text{B.5})$$

Where (a) is from Lemma 9, (b) is by Lemma 10 and Notice that $X_t^h = |\mathcal{T}_h^{\pi^t} (r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)| \leq 1$, by Lemma 3 (setting $\eta = \frac{1}{2}$), we have

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}^{\pi^t} [X_t^h] &\leq \sum_{t=1}^T X_t^h + \frac{1}{2} \text{Var}^{\pi^t} [X_t^h | \mathcal{F}_{t-1}] + 2 \log \frac{1}{\delta} \\ &\stackrel{(a)}{\leq} \sum_{t=1}^T X_t^h + \frac{1}{2} \text{Var}^{\pi^t} [X_t^h] + 2 \log \frac{1}{\delta} \\ &\stackrel{(b)}{\leq} \sum_{t=1}^T X_t^h + \frac{1}{2} \mathbb{E}^{\pi^t} [(X_t^h)^2] + 2 \log \frac{1}{\delta} \\ &\stackrel{(c)}{\leq} \sum_{t=1}^T X_t^h + \frac{1}{2} \mathbb{E}^{\pi^t} [X_t^h] + 2 \log \frac{1}{\delta}, \end{aligned}$$

where (a) is by the property of conditional variance; (b) is by $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$; (c) is by the fact that $0 \leq X_t \leq 1$. Hence, we get

$$\sum_{t=1}^T \mathbb{E}^{\pi^t} [X_t^h] \leq 2 \sum_{t=1}^T X_t^h + 4 \log \frac{1}{\delta}.$$

By taking a union bound over all $h \in [H]$, we know for any $h \in [H]$, with probability $1 - \delta$,

$$\sum_{t=1}^T \mathbb{E}^{\pi^t} [X_t^h] \leq 2 \sum_{t=1}^T X_t^h + 4 \log \frac{H}{\delta}.$$

Summing over $h \in [H]$ and considering $X_t^h = |\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)|$, we get

$$\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[|\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)| \right] \leq 2 \sum_{t=1}^T \sum_{h=1}^H |\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)| + 4H \log \frac{H}{\delta}.$$

Similarly, we could also get

$$\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[|\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)|^2 \right] \leq 2 \sum_{t=1}^T \sum_{h=1}^H |\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)|^2 + 4H \log \frac{H}{\delta}.$$

Inserting the above result back to equation (B.5), we have

$$\begin{aligned} I_2 &\leq \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[(U_h^t)(s_h^t, a_h^t, b_h^t) - \mathbb{T}_{h+1}^{*, \theta^t} U_{h+1}^t(s_{h+1}^t) \right] \\ &\quad + \sum_{t=1}^T \sum_{h=1}^H 2C_0 \cdot \left[|\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)| \right] \\ &\quad + \sum_{t=1}^T \sum_{h=1}^H 2C_1 \cdot \left[\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t) \right] + O(H \log(H/\delta)). \end{aligned}$$

Then using the fact that $|\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, a_h^t, b_h^t)| \leq |\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, a_h^t, b_h^t)|$, we can further have

$$\begin{aligned} I_2 &\leq \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^t} \left[(U_h^t)(s_h, a_h, b_h) - \mathbb{T}_{h+1}^{*, \theta^t} U_{h+1}^t(s_{h+1}) \right] \\ &\quad + \sum_{t=1}^T \sum_{h=1}^H 2(C_0 + C_1) \cdot \left[|\mathcal{T}_h^{\pi^t}(r_h^{\theta^t} - r_h^*)(s_h^t, b_h^t)| \right] + O(H \log(H/\delta)). \end{aligned}$$

Furthermore, using decoupling-coefficient assumption I with the definition of d_1 and d_2 , we can get

$$\begin{aligned} I_2 &\leq \mu_1 \cdot \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} [(U_h - \mathbb{T}_{h+1}^{*, \theta^i} U_{h+1})(s_h, a_h, b_h)^2] + \frac{d_1}{\mu_1} \\ &\quad + 2(C_0 + C_1) \cdot \mu_2 \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{t-1} [\mathcal{T}_h^{\pi^i}((r_h^{\theta^t} - r_h^*)(s_h^i, b_h^i))^2] + 2(C_0 + C_1) \cdot \frac{d_2}{\mu_2} \\ &\quad + O(H \log(H/\delta)). \end{aligned}$$

At last, we exploit the Lemma 3 again, and with probability at least $1 - \delta$, we have

$$\begin{aligned} I_2 &\leq \mu_1 \cdot \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} [(U_h - \mathbb{T}_{h+1}^{*, \theta^i} U_{h+1})(s_h, a_h, b_h)^2] + \frac{d_1}{\mu_1} \\ &\quad + 4(C_0 + C_1) \cdot \mu_2 \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{t-1} \mathbb{E}^{\pi^i} [\mathcal{T}_h^{\pi^i}((r_h^{\theta^t} - r_h^*))^2] + 2(C_0 + C_1) \cdot \frac{d_1}{\mu_2} \\ &\quad + O(H \log(H/\delta)). \end{aligned} \tag{B.6}$$

Now note that $\eta_1 = \eta_2 = 1/\sqrt{T}$, and by choosing $\mu_1 = \frac{\eta_1}{4C_\eta^2}$, $\mu_2 = \frac{\eta_2}{8(C_0+C_1)}$, combining (B.3), and (B.6), with probability at least $1 - 3\delta$, we can have

$$\begin{aligned} \text{Reg}(T) &= I_1 + I_2 \\ &\leq \frac{1}{\sqrt{T}} \cdot HT \cdot (\beta_1 + \beta_2) + \frac{d_1}{\mu_1} + 2(C_0 + C_1) \cdot \frac{d_2}{\mu_2} + O(H \log(H/\delta)) \\ &= \sqrt{T}H(\beta_1 + \beta_2) + 4C_\eta^2 d_1 \sqrt{T} + 16(C_0 + C_1)^2 d_2 \sqrt{T} + O(H \log(H/\delta)) \\ &= \left(H(\beta_1 + \beta_2) + 4C_\eta^2 d_1 + 16(C_0 + C_1)^2 d_2 \right) \sqrt{T} + O(H \log(H/\delta)) \end{aligned}$$

C PROOF OF DECOUPLING COEFFICIENT BOUNDS

We mainly generalize the proof of Proposition 1-3 in [Xiong et al. \(2022\)](#) in this section.

Proof of Proposition 1. We first note that the completeness assumption is satisfied in linear MSG case whose proof can be found in [Huang et al. \(2021\)](#); [Chen et al. \(2023\)](#). Now we consider two arbitrary vector $\omega_h, \omega_{h+1} \in \mathbb{R}^d$ whose norms are bounded $H\sqrt{d}$. We define a function $\tilde{U} \in \mathcal{U}$ such that $\tilde{U}_h = \phi^\top \omega_h$ and $\tilde{U}_{h+1} = \phi^\top \omega_{h+1}$. Furthermore more we take arbitrary $\theta = \{\theta_h\}_{h \in H} \subset \mathbb{R}^d$ such that $\|\theta_h\| \leq \sqrt{d}$. Then we could find $r = \{r_h\}_{h \in [H]} \subseteq \mathcal{F}_r$ and $r_h = \phi(s, a, b)^\top \theta_h, \forall h \in [H], (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Then by Assumption 3 we can find some $U \in \mathcal{U}$ and the corresponding vector $\omega_h(U) \in \mathbb{R}^d$ such that $\|\omega_h(U)\| \leq H\sqrt{d}$ and $\mathbb{T}_h^{s, \theta}(\phi(s, a, b)^\top \omega_{h+1}) = \phi(s, a, b)^\top \omega_h(U) = U_h \in \mathcal{U}_h$. Therefore, we have

$$l_h(\tilde{U}, \theta, s, a, b) = \tilde{U}_h(s, a, b) - \mathbb{T}_h^{s, \theta}(\tilde{U}_{h+1}) = \phi(s, a, b)^\top (\omega_h - \omega_h(U)) = \phi(s, a, b)^\top \Delta_h(U, \tilde{U})$$

where $\Delta_h(U, \tilde{U}) \in \mathbb{R}^d$ and $\|\Delta_h\| \leq 2H\sqrt{d}$.

For any $\{\rho^s\}_{s \in [t]} \subset \mathcal{Q}_1$, i.e. we take any sequence of the leader and follower's joint policies $\{(\pi^s, \nu^{\pi^s, \theta^s})\}_{s \in [t]} \subset \Pi$, we denote as $\phi_h^s = \mathbb{E}^{\rho^s}[\phi(s_h, a_h, b_h)]$ and denote $\Phi_t^h = \lambda I + \sum_{s=1}^t \mathbb{E}^{\rho^s}[\phi(s_h, a_h, b_h)\phi(s_h, a_h, b_h)^\top]$, where $\lambda \geq 1$ is a tuning parameter. We further have

$$\begin{aligned} & \mathbb{E}^{\rho^s}[l_h^t(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)] - \mu \sum_{s=1}^{t-1} \mathbb{E}^{\rho^s}[l_h^t(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)^2] \\ &= \Delta_h(\tilde{U}^t, U_t)^\top \phi_h^t - \mu \Delta_h(\tilde{U}^t, U_t)^\top \sum_{s=1}^{t-1} \mathbb{E}^{\rho^s}[\phi(s_h^s, a_h^s, b_h^s)\phi(s_h^s, a_h^s, b_h^s)^\top] \Delta_h(\tilde{U}_t, U_t) \\ &\leq \Delta_h(\tilde{U}^t, U_t)^\top \phi_h^t - \mu \Delta_h(\tilde{U}^t, U_t)^\top \Phi_{t-1}^h \Delta_h(\tilde{U}_t, U_t) + 4\mu\lambda H^2 d \\ &\leq \frac{1}{4\mu} (\phi_h^t)^\top (\Phi_{t-1}^h)^{-1} \phi_h^t + 4\mu\lambda H^2 d \end{aligned}$$

where the first inequality uses Jensen's inequality and $\|\Delta_h(\tilde{U}_t, U_t)\| \leq 2H\sqrt{d}$ and the second inequality exploits the fact that

$$a^\top b \leq (\|a\|_{\Phi_{t-1}^h} \|b\|_{(\Phi_{t-1}^h)^{-1}}) \leq \frac{1}{2} (\|a\|_{\Phi_{t-1}^h}^2 + \|b\|_{(\Phi_{t-1}^h)^{-1}}^2)$$

Summing over $t \in [T]$ and $h \in [H]$, we have

$$\begin{aligned} & \sum_{t=1}^T \sum_{h=1}^H \left(\mathbb{E}^{\rho^s}[l_h(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)] - \mu \sum_{s=1}^{t-1} \mathbb{E}^{\rho^s}[l_h(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)^2] \right) \\ &\leq \sum_{h=1}^H \left(\frac{\ln(\det(\Phi_T^h)) - d \ln \lambda}{2\mu} + 4\mu\lambda d H^2 T \right) \\ &\leq \left(\frac{dH \ln(1 + \frac{T}{d\lambda})}{2\mu} + 4\mu\lambda d H^3 T \right) \end{aligned}$$

where the first inequality exploit Lemma 4 and the second inequality uses

$$\ln \det(\Phi_T^h) \leq d \ln \frac{\text{tr}(\Phi_T^h)}{d}, \quad \text{where } \text{tr}(\Phi_T^h) \leq \lambda d + T$$

By setting $\lambda = \min\{1, \frac{1}{\mu^2 H^2 T}\}$, we have

$$d_1 \leq 2dH(2 + \ln(2HT))$$

Similarly, for d_2 , notice we could still write

$$m_h(\tilde{\theta}, s, a, b) = r_h^{\tilde{\theta}}(s, b) - r_h(s, b) = \phi(s, a, b)^\top (\tilde{\theta}_h - \theta_h) = \phi(s, a, b)^\top \delta_h(\tilde{\theta}, \tilde{\theta})$$

Then we could repeat the above process to generate the similar bound. Another way to get an upper bound for d_2 is to write $r_h^{\tilde{\theta}}(s, b) - r_h(s, b)$ as a bilinear form and then use the classical decoupling coefficient results on this class. The readers could see [Dann et al. \(2021\)](#); [Chen et al. \(2023\)](#) for reference.

Proof of Proposition 2 We first note that the completeness assumption is also satisfied in generalized linear MSG [\(Huang et al. 2021\)](#) [Chen et al. \(2023\)](#). Similarly, we consider two arbitrary vector $\omega_h, \omega_{h+1} \in \mathbb{R}^d$ whose norms are bounded $H\sqrt{d}$. We define a function $\tilde{U} \in \mathcal{U}$ such that $\tilde{U}_h = \phi^\top \omega_h$ and $\tilde{U}_{h+1} = \phi^\top \omega_{h+1}$. Furthermore more we take arbitrary $\theta = \{\theta_h\}_{h \in H} \subset \mathbb{R}^d$ such that $\|\theta_h\| \leq \sqrt{d}$. Then we could find $r \in \mathcal{F}_r$ and $r_h = \sigma(\phi(s, a, b)^\top \theta_h), \forall h \in [H], (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Then by Assumption 3 we can find some $U \in \mathcal{U}$ and the corresponding vector $\omega_h(U) \in \mathbb{R}^d$ such that $\|\omega_h(U)\| \leq H\sqrt{d}$ and $\mathbb{T}_h^{*,\theta}(\phi(s, a, b)^\top \omega_{h+1}) = \phi(s, a, b)^\top \omega_h(U) = U_h \in \mathcal{U}_h$. Therefore, we have

$$l_h(\tilde{U}, \theta, s, a, b) = \tilde{U}_h(s, a, b) - \mathbb{T}_h^{*,\theta}(\tilde{U}_{h+1}) = \sigma(\phi^\top \omega_h) - \sigma(\phi^\top \omega_h(U))$$

By the Lipschitz condition we have

$$c_1 |\phi^\top \Delta_h(U, \tilde{U})| \leq |l_h(\tilde{U}, \theta, s, a, b)| \leq c_2 |\phi^\top \Delta_h(U, \tilde{U})|$$

where $\Delta_h(U, \tilde{U}) \in \mathbb{R}^d$ and $\|\Delta_h\| \leq 2H\sqrt{d}$.

For any $\{\rho^s\}_{s \in [t]} \subset \varrho_1$, i.e. we take sequence of $\{\pi^s\}_{s \in [t]} \subset \Pi$, we let $\phi_h^s = \mathbb{E}^{\rho^s}[\phi(s_h, a_h, b_h)]$ and let $\Phi_t^h = \lambda I + \sum_{s=1}^t E^{\rho^s}[\phi(s_h, a_h, b_h)\phi(s_h, a_h, b_h)^\top]$, where $\lambda \geq 1$ is a tuning parameter. We further have

$$\begin{aligned} & \mathbb{E}^{\rho^s}[l_h^t(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)] - \mu \sum_{s=1}^{t-1} \mathbb{E}^{\rho^s}[l_h^t(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)^2] \\ & \leq c_2 |\Delta_h(\tilde{U}^t, U_t)^\top \phi_h^t| - \mu c_1^2 \Delta_h(\tilde{U}^t, U_t)^\top \sum_{s=1}^{t-1} \mathbb{E}^{\rho^s}[\phi(s_h^s, a_h^s, b_h^s)\phi(s_h^s, a_h^s, b_h^s)^\top] \Delta_h(\tilde{U}_t, U_t) \\ & \leq c_2 \Delta_h(\tilde{U}^t, U_t)^\top \phi_h^t - \mu c_1^2 \Delta_h(\tilde{U}^t, U_t)^\top \Phi_{t-1}^h \Delta_h(\tilde{U}_t, U_t) + 4\mu c_1^2 \lambda H^2 d \\ & \leq \frac{c_2^2}{4\mu c_1^2} (\phi_h^t)^\top (\Phi_{t-1}^h)^{-1} \phi_h^t + 4\mu c_1^2 \lambda H^2 d \end{aligned}$$

Summing over $t \in [T]$ and $h \in [H]$, we have

$$\begin{aligned} & \sum_{t=1}^T \sum_{h=1}^H \left(\mathbb{E}^{\rho^s}[l_h(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)] - \mu \sum_{s=1}^{t-1} \mathbb{E}^{\rho^s}[l_h(\tilde{U}^t, \theta^t, s_h^t, a_h^t, b_h^t)^2] \right) \\ & \leq \sum_{h=1}^H c_2^2 \left(\left(\frac{\ln(\det(\Phi_T^h)) - d \ln \lambda}{2\mu c_1^2} + 4\mu \lambda c_1^2 d H^2 T \right) \right) \\ & \leq d H c_2^2 \left(\frac{\ln(1 + \frac{T}{d\lambda})}{2\mu c_1^2} + 4\mu c_1^2 \lambda H^2 T \right) \end{aligned}$$

By setting $\lambda = \min\{1, \frac{1}{\mu^2 c_1^2 H^2 T}\}$, we have

$$d_1 \leq 2 \frac{c_2^2}{c_1^2} d H (2 + \ln(2HT))$$

Similarly, for d_2 , notice we could still write

$$m_h(\tilde{\theta}, s, a, b) = r_h^{\tilde{\theta}}(s, b) - r_h(s, b) = \phi(s, a, b)^\top (\tilde{\theta}_h - \theta_h) = \phi(s, a, b)^\top \delta_h(\tilde{\theta}, \tilde{\theta})$$

Then we could repeat the above process to generate the upper bound. Similarly, another way to get an upper bound for d_2 is to exploit Lipschitz condition to upper and lower bound $r_h^{\tilde{\theta}}(s, b) - r_h(s, b)$ by two bilinear forms and then use the classical decoupling coefficient results on this class. The readers could see [Dann et al. \(2021\)](#); [Chen et al. \(2023\)](#) for reference.

D PROOF OF THEOREM 2

Proof. At first, we could decompose the regret into three terms:

$$\begin{aligned}
 \text{Reg}(T) &= \sum_{t=1}^T J(\pi^*) - J(\pi^t) \\
 &= \underbrace{\sum_{t=1}^T \left(\mathbb{E}_{x \sim \rho, a \sim \pi^*} [u^*(x, a)] - \mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^{\theta^t}(x, a)] \right)}_{I_1} \\
 &\quad + \underbrace{\sum_{t=1}^T \left(\mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^{\theta^t}(x, a)] - \mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^*(x, a)] \right)}_{I_2} \\
 &\quad - \underbrace{\sum_{t=1}^T \beta \cdot (\mathbb{D}_{\text{KL}}(\pi^* \parallel \pi_{\text{ref}}) - (\mathbb{D}_{\text{KL}}(\pi^t \parallel \pi_{\text{ref}})))}_{I_3}.
 \end{aligned}$$

First, we compute the upper bound of I_1 . By the definition of π^t and θ^t , we can get

$$\begin{aligned}
 &\mathbb{E}_{x \sim \rho, a \sim \pi^*} [u^*(x, a)] - \beta \mathbb{D}_{\text{KL}}[\pi^* \parallel \pi_{\text{ref}}] - \eta_1 L^t(\theta^*) \\
 &\leq \mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^{\theta^t}(x, a)] - \beta \mathbb{D}_{\text{KL}}[\pi^t \parallel \pi_{\text{ref}}] - \eta_1 L^t(\theta^t),
 \end{aligned}$$

which is equivalent to

$$\begin{aligned}
 &\mathbb{E}_{x \sim \rho, a \sim \pi^*} [u^*(x, a)] - \mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^{\theta^t}(x, a)] \\
 &\leq \beta \mathbb{D}_{\text{KL}}[\pi^* \parallel \pi_{\text{ref}}] - \beta \mathbb{D}_{\text{KL}}[\pi^t \parallel \pi_{\text{ref}}] + \eta_1 \cdot (L^t(\theta^*) - L^t(\theta^t)).
 \end{aligned}$$

Now we introduce the Lemma 2 and Lemma 4 in [Cen et al. \(2024\)](#) to further bound the cross-entropy loss:

Lemma 11 (Lemma 2 and 4 in [Cen et al. \(2024\)](#) when $0 \leq R(x, y) \leq 1$). *The following inequality holds with probability at least $1 - \delta$ that*

$$L^t(\theta^*) - L^t(\theta^t) \leq -(3 + e^2)^{-2} \eta^2 \sum_{i=1}^{t-1} \mathbb{E}_{x \sim \rho, a \sim \pi^i} \left[|\delta^*(x^t, a^t) - \delta^t(x^t, a^t)|^2 \right] + 2 \log \left(\frac{|\mathcal{R}|}{\delta} \right),$$

where $\delta^*(x, a) = R^*(x, y_1) - R^*(x, y_2)$, $\delta^t(x, a) = R^{\theta^t}(x, y_1) - R^{\theta^t}(x, y_2)$.

Then, we compute the upper bound of I_2 .

$$\begin{aligned}
 I_2 &= \sum_{t=1}^T \left(\mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^{\theta^t}(x, a)] - \mathbb{E}_{x \sim \rho, a \sim \pi^t} [u^*(x, a)] \right) \\
 &= 2 \sum_{t=1}^T \left(\mathbb{E}_{x \sim \rho, y \sim \pi^t} [R^{\theta^t}(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \pi^t} [R^*(x, y)] \right) \\
 &\quad - 2 \sum_{t=1}^T \left(\mathbb{E}_{x \sim \rho, y \sim \pi_{\text{base}}} [R^{\theta^t}(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \pi_{\text{base}}} [R^*(x, y)] \right) \\
 &\leq 2 \sum_{t=1}^T \left(\mathbb{E}_{x \sim \rho, y_1 \sim \pi^t, y_2 \sim \pi_{\text{base}}} [\delta^t(x, y_1, y_2) - \delta^*(x, y_1, y_2)] \right).
 \end{aligned}$$

By Multi-agent Decoupling Coefficient, we can further derive

$$\begin{aligned}
I_2/2 &\leq \mu \cdot \sum_{t=1}^T \sum_{i=1}^{t-1} \left(\mathbb{E}_{x \sim \rho, y_1 \sim \pi^i, y_2 \sim \pi_{\text{base}}} [(\delta^t(x, y_1, y_2) - \delta^*(x, y_1, y_2))^2] \right) + \frac{d}{4\mu} \\
&\leq \mu \cdot \sup_{x, y, i} \frac{\pi_{\text{base}}(y | x)}{\pi^i(y | x)} \cdot \sum_{t=1}^T \sum_{i=1}^{t-1} \left(\mathbb{E}_{x \sim \rho, y_1 \sim \pi^i, y_2 \sim \pi^i} [(\delta^t(x, y_1, y_2) - \delta^*(x, y_1, y_2))^2] \right) + \frac{d}{4\mu} \\
&= \mu \cdot \sup_{x, y, i} \frac{\pi_{\text{base}}(y | x)}{\pi^i(y | x)} \cdot \sum_{t=1}^T \sum_{i=1}^{t-1} \left(\mathbb{E}_{x \sim \rho, a \sim \pi^i} [(\delta^t(x, a) - \delta^*(x, a))^2] \right) + \frac{d}{4\mu}.
\end{aligned}$$

Note that

$$\frac{\pi_{\text{base}}(y | x)}{\pi^i(y | x)} = \frac{\pi_{\text{base}}(y | x)}{\pi_{\text{ref}}(y | x)} \cdot \frac{\pi_{\text{ref}}(y | x)}{\pi^i(y | x)} = \kappa \cdot \frac{\pi_{\text{ref}}(y | x)}{\pi^i(y | x)}$$

Then by $\pi^i(y | x) \propto \pi_{\text{ref}}(y | x) \exp(R^i(x, y)/\beta)$ in Rafailov et al. (2024), we can derive $|\log \pi^i(y | x) - \log \pi_{\text{ref}}(y | x)| \leq 2\|R^i(x, \cdot)/\beta\|_{\infty} \leq 2/\beta$ (Cen et al. (2022), Appendix A.2), then $\frac{\pi_{\text{ref}}(y | x)}{\pi^i(y | x)} \leq \exp(2/\beta)$. Then

$$\sup_{x, y, i} \frac{\pi_{\text{base}}(y | x)}{\pi^i(y | x)} = \kappa \exp(2/\beta).$$

Now we sum over I_1, I_2 and I_3 . Thus, we can get

$$\begin{aligned}
\text{Reg}(T) &= I_1 + I_2 + I_3 \\
&= \sum_{t=1}^T (\eta_1 \cdot (L^t(\theta^*) - L^t(\theta^t))) + I_2 \\
&\leq -(3 + e^2)^{-2} \eta_1 \cdot \eta^2 \cdot \sum_{t=1}^T \sum_{i=1}^{t-1} \mathbb{E}_{x \sim \rho, a \sim \pi^i} [|\delta^*(x^t, a^t) - \delta^t(x^t, a^t)|^2] + 2\eta_1 T \log \left(\frac{|\mathcal{R}|}{\delta} \right) \\
&\quad + 2\mu \cdot \kappa \cdot \exp(2/\beta) \cdot \sum_{t=1}^T \sum_{i=1}^{t-1} \left(\mathbb{E}_{x \sim \rho, a \sim \pi^i} [(\delta^t(x, a) - \delta^*(x, a))^2] \right) + \frac{d}{2\mu}.
\end{aligned}$$

Now we choose $\eta_1 = 2\mu\kappa \exp(2/\beta) \cdot (3 + e^2)^2 \cdot \eta^{-2} = 1/\sqrt{T}$, then the inequality above will become

$$\text{Reg}(T) \leq 2\sqrt{T} \log \frac{|\mathcal{R}|}{\delta} + 2 \cdot (3 + e^2)^2 \eta^{-2} d\kappa \exp(2/\beta) \sqrt{T}.$$

□