Setting $T \asymp m$ in (B.23), we obtain that

$$\varepsilon_{\max} = \max_{k\in[K]} \mathbb{E}_{\text{init}}\Big[\big\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\big\|_\sigma\Big] \le \mathcal{O}(B^{3/2}\cdot m^{-1/4} + B^{5/4}\cdot m^{-1/8}). \tag{B.24}$$

Combining (B.16) and (B.24), we conclude the proof of Theorem B.3. □

C Proofs of Auxiliary Results

In this section, we present the proofs for Theorems 6.1 and 6.2, which are used in the §6 to establish our main theorem.

C.1 Proof of Theorem 6.1

Proof. Before we present the proof, we introduce some notation. For any $k \in \{0,\ldots,K-1\}$, we denote $T\widetilde{Q}_k$ by Q_{k+1} and define

$$\varrho_k = Q_k - \widetilde{Q}_k. \tag{C.1}$$

Also, we denote by π_k the greedy policy with respect to $\widetilde{Q}_k$. In addition, throughout the proof, for two functions $Q_1, Q_2 \colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, we use the notation $Q_1 \ge Q_2$ if $Q_1(s,a) \ge Q_2(s,a)$ for any $s\in\mathcal{S}$ and any $a\in\mathcal{A}$, and define $Q_1 \le Q_2$ similarly. Furthermore, for any policy π, recall that in (2.4) we define the operator P^π by

$$(P^\pi Q)(s,a) = \mathbb{E}\big[Q(S',A')\,\big|\,S'\sim P(\cdot\,|\,s,a), A'\sim\pi(\cdot\,|\,S')\big]. \tag{C.2}$$

In addition, we define the operator T^π by

$$(T^\pi Q)(s,a) = r(s,a) + \gamma\cdot(P^\pi Q)(s,a).$$

Finally, we denote $R_{\max}/(1-\gamma)$ by $V_{\max}$. Now we are ready to present the proof, which consists of three key steps.

Step (i): In the first step, we establish a recursion that relates $Q^* - \widetilde{Q}_{k+1}$ with $Q^* - \widetilde{Q}_k$ to measure the sub-optimality of the value function $\widetilde{Q}_k$. In the following, we first establish an upper bound for $Q^* - \widetilde{Q}_{k+1}$ as follows. For each $k\in\{0,\ldots,K-1\}$, by the definition of ϱ_{k+1} in (C.1), we have

$$\begin{aligned} Q^* - \widetilde{Q}_{k+1} &= Q^* - (Q_{k+1} - \varrho_{k+1}) = Q^* - Q_{k+1} + \varrho_{k+1} = Q^* - T\widetilde{Q}_k + \varrho_{k+1} \\ &= Q^* - T^{\pi^*}\widetilde{Q}_k + (T^{\pi^*}\widetilde{Q}_k - T\widetilde{Q}_k) + \varrho_{k+1}, \end{aligned} \tag{C.3}$$

where π^* is the greedy policy with respect to Q^*. Now we leverage the following lemma to show $T^{\pi^*}\widetilde{Q}_k \le T\widetilde{Q}_k$.

Lemma C.1. For any action-value function $Q : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and any policy π, it holds that

$$T^{\pi_Q}Q = TQ \ge T^\pi Q.$$

Proof. Note that we have $\max_{a'} Q(s',a') \geq Q(s',a')$ for any $s' \in \mathcal{S}$ and $a' \in \mathcal{A}$. Thus, it holds that

$$\begin{aligned}(TQ)(s,a) &= r(s,a) + \gamma \cdot \mathbb{E}\bigl[\max_{a'} Q(S',a') \,\big|\, S' \sim P(\cdot \,|\, s,a)\bigr] \\ &\geq r(s,a) + \gamma \cdot \mathbb{E}\bigl[Q(S',A') \,\big|\, S' \sim P(\cdot \,|\, s,a), A' \sim \pi(\cdot \,|\, S')\bigr] = (T^{\pi}Q)(s,a).\end{aligned}$$

Recall that π_Q is the greedy policy with respect to Q such that

$$\mathbb{P}\bigl[A \in \operatorname*{argmax}_a Q(s,a) \,\big|\, A \sim \pi_Q(\cdot \,|\, s)\bigr] = 1,$$

which implies

$$\mathbb{E}\bigl[Q(s',A') \,\big|\, A' \sim \pi_Q(\cdot \,|\, s')\bigr] = \max_{a'} Q(s',a').$$

Consequently, we have

$$\begin{aligned}(T^{\pi_Q}Q)(s,a) &= r(s,a) + \gamma \cdot \mathbb{E}\bigl[Q(S',A') \,\big|\, S' \sim P(\cdot \,|\, s,a), A' \sim \pi_Q(\cdot \,|\, S')\bigr] \\ &= r(s,a) + \gamma \cdot \mathbb{E}\bigl[\max_{a'} Q(S',a') \,\big|\, S' \sim P(\cdot \,|\, s,a)\bigr] = (TQ)(s,a),\end{aligned}$$

which concludes the proof of Lemma C.1. □

By Lemma C.1, we have $T\widetilde{Q}_k \geq T^{\pi^*}\widetilde{Q}_k$. Also note that Q^* is the unique fixed point of T^{π^*}. Thus, by (C.3) we have

$$Q^* - \widetilde{Q}_{k+1} = (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_k) + (T^{\pi^*}\widetilde{Q}_k - T\widetilde{Q}_k) + \varrho_{k+1} \leq (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_k) + \varrho_{k+1}, \qquad \text{(C.4)}$$

In the following, we establish a lower bound for $Q^* - \widetilde{Q}_{k+1}$ based on $\widetilde{Q}^* - \widetilde{Q}_k$. Note that, by Lemma C.1, we have $T^{\pi_k}\widetilde{Q}_k = T\widetilde{Q}_k$ and $TQ^* \geq T^{\pi_k}Q^*$. Similar to (C.3), since Q^* is the unique fixed point of T, it holds that

$$\begin{aligned}Q^* - \widetilde{Q}_{k+1} = Q^* - T\widetilde{Q}_k + \varrho_{k+1} &= Q^* - T^{\pi_k}\widetilde{Q}_k + \varrho_{k+1} = Q^* - T^{\pi_k}Q^* + (T^{\pi_k}Q^* - T^{\pi_k}\widetilde{Q}_k) + \varrho_{k+1} \\ &= (TQ^* - T^{\pi_k}Q^*) + (T^{\pi_k}Q^* - T^{\pi_k}\widetilde{Q}_k) + \varrho_{k+1} \geq (T^{\pi_k}Q^* - T^{\pi_k}\widetilde{Q}_k) + \varrho_{k+1}. \qquad \text{(C.5)}\end{aligned}$$

Thus, combining (C.4) and (C.5) we obtain that, for any $k \in \{0, \ldots, K-1\}$,

$$T^{\pi_k}Q^* - T^{\pi_k}\widetilde{Q}_k + \varrho_{k+1} \leq Q^* - \widetilde{Q}_{k+1} \leq T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_k + \varrho_{k+1}. \qquad \text{(C.6)}$$

The inequalities in (C.6) show that the error $Q^* - \widetilde{Q}_{k+1}$ can be sandwiched by the summation of a term involving $Q^* - \widetilde{Q}_k$ and the error ϱ_{k+1}, which is defined in (C.1) and induced by approximating the action-value function. Using P^{π} defined in (C.2), we can write (C.6) in a more compact form,

$$\gamma \cdot P^{\pi^*}(Q^* - \widetilde{Q}_k) + \varrho_{k+1} \geq Q^* - \widetilde{Q}_{k+1} \geq \gamma \cdot P^{\pi_k}(Q^* - \widetilde{Q}_k) + \varrho_{k+1}. \qquad \text{(C.7)}$$

Meanwhile, note that P^{π} defined in (C.2) is a linear operator. In fact, P^{π} is the Markov transition operator for the Markov chain on $\mathcal{S} \times \mathcal{A}$ with transition dynamics

$$S_{t+1} \sim P(\cdot \,|\, S_t, A_t), \qquad A_{t+1} \sim \pi(\cdot \,|\, S_{t+1}).$$

By the linearity of the operator P^{π} and the one-step error bound in (C.6), we have the following characterization of the multi-step error.

Lemma C.2 (Error Propagation)**.** For any $k, \ell \in \{0, 1, \ldots, K-1\}$ with $k < \ell$, we have

$$Q^* - \widetilde{Q}_\ell \le \sum_{i=k}^{\ell-1} \gamma^{\ell-1-i} \cdot (P^{\pi^*})^{\ell-1-i} \varrho_{i+1} + \gamma^{\ell-k} \cdot (P^{\pi^*})^{\ell-k} (Q^* - \widetilde{Q}_k), \tag{C.8}$$

$$Q^* - \widetilde{Q}_\ell \ge \sum_{i=k}^{\ell-1} \gamma^{\ell-1-i} \cdot (P^{\pi_{\ell-1}} P^{\pi_{\ell-2}} \cdots P^{\pi_{i+1}}) \varrho_{i+1} + \gamma^{\ell-k} \cdot (P^{\pi_{\ell-1}} P^{\pi_{\ell-2}} \cdots P^{\pi_k})(Q^* - \widetilde{Q}_k). \tag{C.9}$$

Here ϱ_{i+1} is defined in (C.1) and we use $P^\pi P^{\pi'}$ and $(P^\pi)^k$ to denote the composition of operators.

Proof. Note that P^π is a linear operator for any policy π. We obtain (C.8) and (C.9) by iteratively applying the inequalities in (C.7). □

Lemma C.2 gives the upper and lower bounds for the propagation of error through multiple iterations of Algorithm 1, which concludes the first step of our proof.

Step (ii): The results in the first step only concern the propagation of error $Q^* - \widetilde{Q}_k$. In contrast, the output of Algorithm 1 is the greedy policy π_k with respect to $\widetilde{Q}_k$. In the second step, our goal is to quantify the suboptimality of Q^{π_k}, which is the action-value function corresponding to π_k. In the following, we establish an upper bound for $Q^* - Q^{\pi_k}$.

To begin with, we have $Q^* \ge Q^{\pi_k}$ by the definition of Q^* in (2.5). Note that we have $Q^* = T^{\pi^*} Q^*$ and $Q^{\pi_k} = T^{\pi_k} Q^{\pi_k}$. Hence, it holds that

$$\begin{aligned} Q^* - Q^{\pi_k} &= T^{\pi^*} Q^* - T^{\pi_k} Q^{\pi_k} = T^{\pi^*} Q^* + (-T^{\pi^*} \widetilde{Q}_k + T^{\pi^*} \widetilde{Q}_k) + (-T^{\pi_k} \widetilde{Q}_k + T^{\pi_k} \widetilde{Q}_k) - T^{\pi_k} Q^{\pi_k} \\ &= (T^{\pi^*} \widetilde{Q}_k - T^{\pi_k} \widetilde{Q}_k) + (T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_k) + (T^{\pi_k} \widetilde{Q}_k - T^{\pi_k} Q^{\pi_k}). \end{aligned} \tag{C.10}$$

Now we quantify the three terms on the right-hand side of (C.10) respectively. First, by Lemma C.1, we have

$$T^{\pi^*} \widetilde{Q}_k - T^{\pi_k} \widetilde{Q}_k = T^{\pi^*} \widetilde{Q}_k - T \widetilde{Q}_k \le 0. \tag{C.11}$$

Meanwhile, by the definition of the operator P^π in (C.2), we have

$$T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_k = \gamma \cdot P^{\pi^*} (Q^* - \widetilde{Q}_k), \qquad T^{\pi_k} \widetilde{Q}_k - T^{\pi_k} Q^{\pi_k} = \gamma \cdot P^{\pi_k} (\widetilde{Q}_k - Q^{\pi_k}). \tag{C.12}$$

Plugging (C.11) and (C.12) into (C.10), we obtain

$$\begin{aligned} Q^* - Q^{\pi_k} &\le \gamma \cdot P^{\pi^*} (Q^* - \widetilde{Q}_k) + \gamma \cdot P^{\pi_k} (\widetilde{Q}_k - Q^{\pi_k}) \\ &= \gamma \cdot (P^{\pi^*} - P^{\pi_k})(Q^* - \widetilde{Q}_k) + \gamma \cdot P^{\pi_k} (Q^* - Q^{\pi_k}), \end{aligned}$$

which further implies that

$$(I - \gamma \cdot P^{\pi_k})(Q^* - Q^{\pi_k}) \le \gamma \cdot (P^{\pi^*} - P^{\pi_k})(Q^* - \widetilde{Q}_k).$$

Here I is the identity operator. Since T^π is a γ-contractive operator for any policy π, $I - \gamma \cdot P^\pi$ is invertible. Thus, we obtain

$$0 \le Q^* - Q^{\pi_k} \le \gamma \cdot (I - \gamma \cdot P^{\pi_k})^{-1} \bigl[P^{\pi^*} (Q^* - \widetilde{Q}_k) - P^{\pi_k} (Q^* - \widetilde{Q}_k) \bigr], \tag{C.13}$$

which relates $Q^* - Q^{\pi_k}$ with $Q^* - \widetilde{Q}_k$. In the following, we plug Lemma C.2 into (C.13) to obtain the multiple-step error bounds for Q^{π_k}. First note that, by the definition of P^π in (C.2), for any functions $f_1, f_2 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ satisfying $f_1 \geq f_2$, we have $P^\pi f_1 \geq P^\pi f_2$. Combining this inequality with the upper bound in (C.8) and the lower bound in (C.9), we have that, for any $k < \ell$,

$$P^{\pi^*}(Q^* - \widetilde{Q}_\ell) \leq \sum_{i=k}^{\ell-1} \gamma^{\ell-1-i} \cdot (P^{\pi^*})^{\ell-i} \varrho_{i+1} + \gamma^{\ell-k} \cdot (P^{\pi^*})^{\ell-k+1}(Q^* - \widetilde{Q}_k), \tag{C.14}$$

$$\begin{aligned} P^{\pi_\ell}(Q^* - \widetilde{Q}_\ell) \geq & \sum_{i=k}^{\ell-1} \gamma^{\ell-1-i} \cdot (P^{\pi_\ell} P^{\pi_{\ell-1}} \cdots P^{\pi_{i+1}}) \varrho_{i+1} \\ & + \gamma^{\ell-k} \cdot (P^{\pi_\ell} P^{\pi_{\ell-1}} \cdots P^{\pi_k})(Q^* - \widetilde{Q}_k). \end{aligned} \tag{C.15}$$

Then we plug (C.14) and (C.15) into (C.13) and obtain

$$\begin{aligned} 0 \leq Q^* - Q^{\pi_\ell} \leq (I - \gamma \cdot P^{\pi_\ell})^{-1} \Big\{ & \sum_{i=k}^{\ell-1} \gamma^{\ell-i} \cdot \big[(P^{\pi^*})^{\ell-i} - (P^{\pi_\ell} P^{\pi_{\ell-1}} \cdots P^{\pi_{i+1}})\big] \varrho_{i+1} \\ & + \gamma^{\ell+1-k} \cdot \big[(P^{\pi^*})^{\ell-k+1} - (P^{\pi_\ell} P^{\pi_{\ell-1}} \cdots P^{\pi_k})\big](Q^* - \widetilde{Q}_k) \Big\} \end{aligned} \tag{C.16}$$

for any $k < \ell$. To quantify the error of Q^{π_K}, we set $\ell = K$ and $k = 0$ in (C.16) to obtain

$$\begin{aligned} 0 \leq Q^* - Q^{\pi_K} \leq (I - \gamma P^{\pi_K})^{-1} \Big\{ & \sum_{i=0}^{K-1} \gamma^{K-i} \cdot \big[(P^{\pi^*})^{K-i} - (P^{\pi_K} P^{\pi_{K-1}} \cdots P^{\pi_{i+1}})\big] \varrho_{i+1} \\ & + \gamma^{K+1} \cdot \big[(P^{\pi^*})^{K+1} - (P^{\pi_K} P^{\pi_{K-1}} \cdots P^{\pi_0})\big](Q^* - \widetilde{Q}_0) \Big\}. \end{aligned} \tag{C.17}$$

For notational simplicity, we define

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}}, \quad \text{for} \quad 0 \leq i \leq K-1, \quad \text{and} \quad \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}. \tag{C.18}$$

One can show that $\sum_{i=0}^K \alpha_i = 1$. Meanwhile, we define $K+1$ linear operators $\{O_k\}_{k=0}^K$ by

$$\begin{aligned} O_i &= (1-\gamma)/2 \cdot (I - \gamma P^{\pi_K})^{-1} \big[(P^{\pi^*})^{K-i} + (P^{\pi_K} P^{\pi_{K-1}} \cdots P^{\pi_{i+1}})\big], \quad \text{for} \quad 0 \leq i \leq K-1, \\ O_K &= (1-\gamma)/2 \cdot (I - \gamma P^{\pi_K})^{-1} \big[(P^{\pi^*})^{K+1} + (P^{\pi_K} P^{\pi_{K-1}} \cdots P^{\pi_0})\big]. \end{aligned}$$

Using this notation, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, by (C.17) we have

$$\begin{aligned} & \big|Q^*(s,a) - Q^{\pi_K}(s,a)\big| \\ & \quad \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \cdot \bigg[\sum_{i=0}^{K-1} \alpha_i \cdot \big(O_i |\varrho_{i+1}|\big)(s,a) + \alpha_K \cdot \big(O_K |Q^* - \widetilde{Q}_0|\big)(s,a)\bigg], \end{aligned} \tag{C.19}$$

where both $O_i|\varrho_{i+1}|$ and $O_K|Q^* - \widetilde{Q}_0|$ are functions defined on $\mathcal{S} \times \mathcal{A}$. Here (C.19) gives a uniform upper bound for $Q^* - Q^{\pi_K}$, which concludes the second step.

Step (iii): In this step, we conclude the proof by establishing an upper bound for $\|Q^* - Q^{\pi_K}\|_{1,\mu}$ based on (C.19). Here $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ is a fixed probability distribution. To simplify the notation, for any measurable function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we denote $\mu(f)$ to be the expectation of f under μ, that is, $\mu(f) = \int_{\mathcal{S}\times\mathcal{A}} f(s,a)\mathrm{d}\mu(s,a)$. Using this notation, by (C.19) we bound $\|Q^* - Q^{\pi_\ell}\|_{1,\mu}$ by

$$\begin{aligned}\|Q^* - Q^{\pi_K}\|_{1,\mu} &= \mu\big(|Q^* - Q^{\pi_K}|\big) \\ &\leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \cdot \mu\biggl[\sum_{i=0}^{K-1} \alpha_i \cdot \big(O_i|\varrho_{i+1}|\big) + \alpha_K \cdot \big(O_K|Q^* - \widetilde{Q}_0|\big)\biggr]. \end{aligned} \tag{C.20}$$

By the linearity of expectation, (C.20) implies

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \cdot \biggl[\sum_{i=0}^{K-1} \alpha_i \cdot \mu\big(O_i|\varrho_{i+1}|\big) + \alpha_K \cdot \mu\big(O_K|Q^* - \widetilde{Q}_0|\big)\biggr]. \tag{C.21}$$

Furthermore, since both Q^* and $\widetilde{Q}_0$ are bounded by $V_{\max} = R_{\max}/(1-\gamma)$ in ℓ_∞-norm, we have

$$\mu\big(O_K|Q^* - \widetilde{Q}_0|\big) \leq 2 \cdot R_{\max}/(1-\gamma). \tag{C.22}$$

Moreover, for any $i \in \{0, \ldots, K-1\}$, by expanding $(1 - \gamma P^{\pi_K})^{-1}$ into a infinite series, we have

$$\begin{aligned}\mu\big(O_i|\varrho_{i+1}|\big) &= \mu\biggl\{\frac{1-\gamma}{2} \cdot (1-\gamma P^{\pi_K})^{-1}\big[(P^{\pi^*})^{K-i} + (P^{\pi_K} P^{\pi_{K-1}} \cdots P^{\pi_{i+1}})\big]|\varrho_{i+1}|\biggr\} \\ &= \frac{1-\gamma}{2} \cdot \mu\biggl\{\sum_{j=0}^{\infty} \gamma^j \cdot \big[(P^{\pi_K})^j (P^{\pi^*})^{K-i} + (P^{\pi_K})^{j+1}(P^{\pi_{K-1}} \cdots P^{\pi_{i+1}})\big]|\varrho_{i+1}|\biggr\}. \end{aligned} \tag{C.23}$$

To upper bound the right-hand side of (C.23), we consider the following quantity

$$\mu\big[(P^{\pi_K})^j (P^{\tau_m} P^{\tau_{m-1}} \cdots P^{\tau_1}) f\big] = \int_{\mathcal{S}\times\mathcal{A}} \big[(P^{\pi_K})^j (P^{\tau_m} P^{\tau_{m-1}} \cdots P^{\tau_1}) f\big](s,a)\mathrm{d}\mu(s,a). \tag{C.24}$$

Here $\tau_1, \ldots, \tau_m$ are m policies. Recall that P^π is the transition operator of a Markov process defined on $\mathcal{S} \times \mathcal{A}$ for any policy π. Then the integral on the right-hand side of (C.24) corresponds to the expectation of the function $f(X_t)$, where $\{X_t\}_{t\geq 0}$ is a Markov process defined on $\mathcal{S} \times \mathcal{A}$. Such a Markov process has initial distribution $X_0 \sim \mu$. The first m transition operators are $\{P^{\tau_j}\}_{j\in[m]}$, followed by j identical transition operators P^{π_K}. Hence, $(P^{\pi_K})^j (P^{\tau_m} P^{\tau_{m-1}} \cdots P^{\tau_1})\mu$ is the marginal distribution of X_{j+m}, which we denote by $\widetilde{\mu}_j$ for notational simplicity. Hence, (C.24) takes the form

$$\mu\big[(P^{\pi_K})^j (P^{\tau_m} P^{\tau_{m-1}} \cdots P^{\tau_1}) f\big] = \mathbb{E}\big[f(X_{j+m})\big] = \widetilde{\mu}_j(f) = \int_{\mathcal{S}\times\mathcal{A}} f(s,a)\mathrm{d}\widetilde{\mu}_j(s,a) \tag{C.25}$$

for any measurable function f on $\mathcal{S} \times \mathcal{A}$. By Cauchy-Schwarz inequality, we have

$$\widetilde{\mu}_j(f) \leq \biggl[\int_{\mathcal{S}\times\mathcal{A}} \biggl|\frac{\mathrm{d}\widetilde{\mu}_j}{\mathrm{d}\sigma}(s,a)\biggr|^2 \mathrm{d}\sigma(s,a)\biggr]^{1/2} \biggl[\int_{\mathcal{S}\times\mathcal{A}} |f(s,a)|^2 \mathrm{d}\sigma(s,a)\biggr]^{1/2}, \tag{C.26}$$

in which $\mathrm{d}\widetilde{\mu}_j/\mathrm{d}\sigma\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the Radon-Nikodym derivative. Recall that the $(m+j)$-th order concentration coefficient $\kappa(m+j;\mu,\sigma)$ is defined in (4.4). Combining (C.25) and (C.26), we obtain

$$\widetilde{\mu}_j(f) \le \kappa(m+j;\mu,\sigma)\cdot \|f\|_\sigma.$$

Thus, by (C.23) we have

$$\begin{aligned}\mu\bigl(O_i|\varrho_{i+1}|\bigr) &= \frac{1-\gamma}{2}\cdot\sum_{j=0}^{\infty}\gamma^j\cdot\Bigl\{\mu\bigl[(P^{\pi_K})^j(P^{\pi^*})^{K-i}|\varrho_{i+1}|\bigr]+\mu\bigl[(P^{\pi_K})^{j+1}(P^{\pi_{K-1}}\cdots P^{\pi_{i+1}})|\varrho_{i+1}|\bigr]\Bigr\}\\ &\le (1-\gamma)\cdot\sum_{j=0}^{\infty}\gamma^j\cdot\kappa(K-i+j;\mu,\sigma)\cdot\|\varrho_{i+1}\|_\sigma. \qquad (C.27)\end{aligned}$$

Now we combine (C.21), (C.22), and (C.27) to obtain

$$\begin{aligned}\|Q^*-Q^{\pi_K}\|_{1,\mu} &\le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\cdot\Biggl[\sum_{i=0}^{K-1}\alpha_i\cdot\mu\bigl(O_i|\varrho_{i+1}|\bigr)+\alpha_K\cdot\mu\bigl(O_K|Q^*-\widetilde{Q}_0|\bigr)\Biggr]\\ &\le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)}\cdot\Biggl[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\cdot\gamma^j\cdot\kappa(K-i+j;\mu,\sigma)\cdot\|\varrho_{i+1}\|_\sigma\Biggr]+\frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3}\cdot\alpha_K\cdot R_{\max}.\end{aligned}$$

Recall that in Theorem 6.1 and (C.1) we define $\varepsilon_{\max}=\max_{i\in[K]}\|\varrho_i\|_\sigma$. We have that $\|Q^*-Q^{\pi_K}\|_{1,\mu}$ is further upper bounded by

$$\begin{aligned}&\|Q^*-Q^{\pi_K}\|_{1,\mu} \qquad (C.28)\\ &\quad\le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)}\cdot\Biggl[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\cdot\gamma^j\cdot\kappa(K-i+j;\mu,\sigma)\Biggr]\cdot\varepsilon_{\max}+\frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3}\cdot\alpha_K\cdot R_{\max}\\ &\quad= \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)}\cdot\Biggl[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}}\cdot\gamma^j\cdot\kappa(K-i+j;\mu,\sigma)\Biggr]\cdot\varepsilon_{\max}+\frac{4\gamma^{K+1}}{(1-\gamma)^2}\cdot R_{\max},\end{aligned}$$

where the last equality follows from the definition of $\{\alpha_i\}_{0\le i\le K}$ in (C.18). We simplify the summation on the right-hand side of (C.28) and use Assumption 4.3 to obtain

$$\begin{aligned}&\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}}\cdot\gamma^j\cdot\kappa(K-i+j;\mu,\sigma)\\ &\qquad=\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{j=0}^{\infty}\sum_{i=0}^{K-1}\gamma^{K-i+j-1}\cdot\kappa(K-i+j;\mu,\sigma)\\ &\qquad\le\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{m=0}^{\infty}\gamma^{m-1}\cdot m\cdot\kappa(m;\mu,\sigma)\le\frac{\phi_{\mu,\sigma}}{(1-\gamma^{K+1})(1-\gamma)}, \qquad (C.29)\end{aligned}$$

where the last inequality follows from (4.5) in Assumption 4.3. Finally, combining (C.28) and (C.29), we obtain

$$\|Q^*-Q^{\pi_K}\|_{1,\mu}\le\frac{2\gamma\cdot\phi_{\mu,\sigma}}{(1-\gamma)^2}\cdot\varepsilon_{\max}+\frac{4\gamma^{K+1}}{(1-\gamma)^2}\cdot R_{\max},$$

which concludes the third step and hence the proof of Theorem 6.1. □

C.2 Proof of Theorem 6.2

Proof. Recall that in Algorithm 1 we define $Y_i = R_i + \gamma \cdot \max_{a \in \mathcal{A}} Q(S_{i+1}, a)$, where Q is any function in $\mathcal{F}$. By definition, we have $\mathbb{E}(Y_i \,|\, S_i = s, A_i = a) = (TQ)(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Thus, TQ can be viewed as the underlying truth of the regression problem defined in (6.2), where the covariates and responses are $\{(S_i, A_i)\}_{i \in [n]}$ and $\{Y_i\}_{i \in [n]}$, respectively. Moreover, note that TQ is not necessarily in function class $\mathcal{F}$. We denote by Q^* the best approximation of TQ in $\mathcal{F}$, which is the solution to

$$\underset{f \in \mathcal{F}}{\text{minimize}} \, \|f - TQ\|_\sigma^2 = \mathbb{E}\Big\{ \big[f(S_i, A_i) - Q(S_i, A_i) \big]^2 \Big\}. \tag{C.30}$$

For notational simplicity, in the sequel we denote (S_i, A_i) by X_i for all $i \in [n]$. For any $f \in \mathcal{F}$, we define $\|f\|_n^2 = 1/n \cdot \sum_{i=1}^n [f(X_i)]^2$. Since both $\widehat{Q}$ and TQ are bounded by $V_{\max} = R_{\max}/(1-\gamma)$, we only need to consider the case where $\log N_\delta \le n$. Here N_δ is the cardinality of $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$. Moreover, let $f_1, \ldots, f_{N_\delta}$ be the centers of the minimal δ-covering of $\mathcal{F}$. Then by the definition of δ-covering, there exists $k^* \in [N_\delta]$ such that $\|\widehat{Q} - f_{k^*}\|_\infty \le \delta$. It is worth mentioning that k^* is a random variable since $\widehat{Q}$ is obtained from data.

In the following, we prove (6.3) in two steps, which are bridged by $\mathbb{E}[\|\widehat{Q} - TQ\|_n^2]$.

Step (i): We relate $\mathbb{E}[\|\widehat{Q} - TQ\|_n^2]$ with its empirical counterpart $\|\widehat{Q} - TQ\|_n^2$. Recall that we define $Y_i = R_i + \gamma \cdot \max_{a \in \mathcal{A}} Q(S_{i+1}, a)$ for each $i \in [n]$. By the definition of $\widehat{Q}$, for any $f \in \mathcal{F}$ we have

$$\sum_{i=1}^n \big[Y_i - \widehat{Q}(X_i) \big]^2 \le \sum_{i=1}^n \big[Y_i - f(X_i) \big]^2. \tag{C.31}$$

For each $i \in [n]$, we define $\xi_i = Y_i - (TQ)(X_i)$. Then (C.31) can be written as

$$\|\widehat{Q} - TQ\|_n^2 \le \|f - TQ\|_n^2 + \frac{2}{n} \sum_{i=1}^n \xi_i \cdot \big[\widehat{Q}(X_i) - f(X_i) \big]. \tag{C.32}$$

Since both f and Q are deterministic, we have $\mathbb{E}(\|f - TQ\|_n^2) = \|f - TQ\|_\sigma^2$. Moreover, since $\mathbb{E}(\xi_i \,|\, X_i) = 0$ by definition, we have $\mathbb{E}[\xi_i \cdot g(X_i)] = 0$ for any bounded and measurable function g. Thus, it holds that

$$\mathbb{E}\bigg\{ \sum_{i=1}^n \xi_i \cdot \big[\widehat{Q}(X_i) - f(X_i) \big] \bigg\} = \mathbb{E}\bigg\{ \sum_{i=1}^n \xi_i \cdot \big[\widehat{Q}(X_i) - (TQ)(X_i) \big] \bigg\}. \tag{C.33}$$

In addition, by triangle inequality and (C.33), we have

$$\begin{aligned} &\bigg| \mathbb{E}\bigg\{ \sum_{i=1}^n \xi_i \cdot \big[\widehat{Q}(X_i) - (TQ)(X_i) \big] \bigg\} \bigg| \\ &\qquad \le \bigg| \mathbb{E}\bigg\{ \sum_{i=1}^n \xi_i \cdot \big[\widehat{Q}(X_i) - f_{k^*}(X_i) \big] \bigg\} \bigg| + \bigg| \mathbb{E}\bigg\{ \sum_{i=1}^n \xi_i \cdot \big[f_{k^*}(X_i) - (TQ)(X_i) \big] \bigg\} \bigg|, \end{aligned} \tag{C.34}$$

where f_{k^*} satisfies $\|f_{k^*} - \widehat{Q}\|_\infty \le \delta$. In the following, we upper bound the two terms on the right-hand side of (C.34) respectively. For the first term, by applying Cauchy-Schwarz inequality twice, we have

$$\begin{aligned}\biggl|\mathbb{E}\biggl\{\sum_{i=1}^n \xi_i \cdot \bigl[\widehat{Q}(X_i) - f_{k^*}(X_i)\bigr]\biggr\}\biggr| &\le \sqrt{n} \cdot \biggl|\mathbb{E}\biggl[\biggl(\sum_{i=1}^n \xi_i^2\biggr)^{1/2} \cdot \|\widehat{Q} - f_{k^*}\|_n\biggr]\biggr| \\ &\le \sqrt{n} \cdot \biggl[\mathbb{E}\biggl(\sum_{i=1}^n \xi_i^2\biggr)\biggr]^{1/2} \cdot \bigl[\mathbb{E}(\|\widehat{Q} - f_{k^*}\|_n^2)\bigr]^{1/2} \le n\delta \cdot \bigl[\mathbb{E}(\xi_i^2)\bigr]^{1/2}, \end{aligned} \tag{C.35}$$

where we use the fact that $\{\xi_i\}_{i\in[n]}$ have the same marginal distributions and $\|\widehat{Q} - f_{k^*}\|_n \le \delta$. Since both Y_i and TQ are bounded by $V_{\max}$, ξ_i is a bounded random variable by its definition. Thus, there exists a constant $C_\xi > 0$ depending on ξ such that $\mathbb{E}(\xi_i^2) \le C_\xi^2 \cdot V_{\max}^2$. Then (C.35) implies

$$\biggl|\mathbb{E}\biggl\{\sum_{i=1}^n \xi_i \cdot \bigl[\widehat{Q}(X_i) - f_{k^*}(X_i)\bigr]\biggr\}\biggr| \le C_\xi \cdot V_{\max} \cdot n\delta. \tag{C.36}$$

It remains to upper bound the second term on the right-hand side of (C.34). We first define N_δ self-normalized random variables

$$Z_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \cdot \bigl[f_j(X_i) - (TQ)(X_i)\bigr] \cdot \|f_j - (TQ)\|_n^{-1} \tag{C.37}$$

for all $j \in [N_\delta]$. Here recall that $\{f_j\}_{j\in[N_\delta]}$ are the centers of the minimal δ-covering of $\mathcal{F}$. Then we have

$$\begin{aligned}\biggl|\mathbb{E}\biggl\{\sum_{i=1}^n \xi_i \cdot \bigl[f_{k^*}(X_i) - (TQ)(X_i)\bigr]\biggr\}\biggr| &= \sqrt{n} \cdot \mathbb{E}\bigl[\|f_{k^*} - TQ\|_n \cdot |Z_{k^*}|\bigr] \\ &\le \sqrt{n} \cdot \mathbb{E}\Bigl\{\bigl[\|\widehat{Q} - TQ\|_n + \|\widehat{Q} - f_{k^*}\|_n\bigr] \cdot |Z_{k^*}|\Bigr\} \le \sqrt{n} \cdot \mathbb{E}\Bigl\{\bigl[\|\widehat{Q} - TQ\|_n + \delta\bigr] \cdot |Z_{k^*}|\Bigr\}, \end{aligned} \tag{C.38}$$

where the first inequality follows from triangle inequality and the second inequality follows from the fact that $\|\widehat{Q} - f_{k^*}\|_\infty \le \delta$. Then applying Cauchy-Schwarz inequality to the last term on the right-hand side of (C.38), we obtain

$$\begin{aligned}\mathbb{E}\Bigl\{\bigl[\|\widehat{Q} - TQ\|_n + \delta\bigr] \cdot |Z_{k^*}|\Bigr\} &\le \Bigl(\mathbb{E}\Bigl\{\bigl[\|\widehat{Q} - TQ\|_n + \delta\bigr]^2\Bigr\}\Bigr)^{1/2} \cdot \bigl[\mathbb{E}(Z_{k^*}^2)\bigr]^{1/2} \\ &\le \Bigl(\Bigl\{\mathbb{E}\bigl[\|\widehat{Q} - TQ\|_n^2\bigr]\Bigr\}^{1/2} + \delta\Bigr) \cdot \Bigl[\mathbb{E}\bigl(\max_{j\in[N]} Z_j^2\bigr)\Bigr]^{1/2}, \end{aligned} \tag{C.39}$$

where the last inequality follows from

$$\mathbb{E}\bigl[\|\widehat{Q} - TQ\|_n\bigr] \le \Bigl\{\mathbb{E}\bigl[\|\widehat{Q} - TQ\|_n^2\bigr]\Bigr\}^{1/2}.$$

Moreover, since ξ_i is centered conditioning on $\{X_i\}_{i\in[n]}$ and is bounded by $2V_{\max}$, ξ_i is a sub-Gaussian random variable. In specific, there exists an absolute constant $H_\xi > 0$ such that $\|\xi_i\|_{\psi_2} \le H_\xi \cdot V_{\max}$ for each $i \in [n]$. Here the ψ_2-norm of a random variable $W \in \mathbb{R}$ is defined as

$$\|W\|_{\psi_2} = \sup_{p\ge 1} p^{-1/2}\bigl[\mathbb{E}(|W|^p)\bigr]^{1/p}.$$

By the definition of Z_j in (C.37), conditioning on $\{X_i\}_{i\in[n]}$, $\xi_i \cdot [f_j(X_i) - (TQ)(X_i)]$ is a centered and sub-Gaussian random variable with

$$\big\|\xi_i \cdot \big[f_j(X_i) - (TQ)(X_i)\big]\big\|_{\psi_2} \le H_\xi \cdot V_{\max} \cdot \big|f_j(X_i) - (TQ)(X_i)\big|.$$

Moreover, since Z_j is a summation of independent sub-Gaussian random variables, by Lemma 5.9 of Vershynin (2010), the ψ_2-norm of Z_j satisfies

$$\|Z_j\|_{\psi_2} \le C \cdot H_\xi \cdot V_{\max} \cdot \|f_j - TQ\|_n^{-1} \cdot \biggl[\frac{1}{n}\sum_{i=1}^n \big|[f_j(X_i) - (TQ)(X_i)]\big|^2\biggr]^{1/2} \le C \cdot H_\xi \cdot V_{\max},$$

where $C > 0$ is an absolute constant. Furthermore, by Lemmas 5.14 and 5.15 of Vershynin (2010), Z_j^2 is a sub-exponential random variable, and its the moment-generating function is bounded by

$$\mathbb{E}\big[\exp(t \cdot Z_j^2)\big] \le \exp(C \cdot t^2 \cdot H_\xi^4 \cdot V_{\max}^4) \tag{C.40}$$

for any t satisfying $C' \cdot |t| \cdot H_\xi^2 \cdot V_{\max}^2 \le 1$, where C and C' are two positive absolute constants. Moreover, by Jensen's inequality, we bound the moment-generating function of $\max_{j\in[N_\delta]} Z_j^2$ by

$$\mathbb{E}\Bigl[\exp\bigl(t \cdot \max_{j\in[N_\delta]} Z_j^2\bigr)\Bigr] \le \sum_{j\in[N_\delta]} \mathbb{E}\big[\exp(t \cdot Z_j^2)\big]. \tag{C.41}$$

Combining (C.40) and (C.41), we have

$$\mathbb{E}\bigl(\max_{j\in[N]} Z_j^2\bigr) \le C^2 \cdot H_\xi^2 \cdot V_{\max}^2 \cdot \log N_\delta, \tag{C.42}$$

where $C > 0$ is an absolute constant. Hence, plugging (C.42) into (C.38) and (C.39), we upper bound the second term of the right-hand side of (C.33) by

$$\begin{aligned}
&\biggl|\mathbb{E}\biggl\{\sum_{i=1}^n \xi_i \cdot \big[f_{k^*}(X_i) - (TQ)(X_i)\big]\biggr\}\biggr| \\
&\qquad \le \biggl(\Bigl\{\mathbb{E}\big[\|\widehat{Q} - TQ\|_n^2\big]\Bigr\}^{1/2} + \delta\biggr) \cdot C \cdot H_\xi \cdot V_{\max} \cdot \sqrt{n \cdot \log N_\delta}.
\end{aligned} \tag{C.43}$$

Finally, combining (C.32), (C.36) and (C.43), we obtain the following inequality

$$\begin{aligned}
\mathbb{E}\big[\|\widehat{Q} - TQ\|_n^2\big] &\le \inf_{f\in\mathcal{F}} \mathbb{E}\big[\|f - TQ\|_n^2\big] + C_\xi \cdot V_{\max} \cdot \delta \\
&\qquad + \Bigl(\bigl\{\mathbb{E}\big[\|\widehat{Q} - (TQ)\|_n^2\big]\bigr\}^{1/2} + \delta\Bigr) \cdot C \cdot H_\xi \cdot V_{\max} \cdot \sqrt{\log N_\delta / n} \\
&\le C \cdot V_{\max}\sqrt{\log N_\delta / n} \cdot \Bigl\{\mathbb{E}\big[\|\widehat{Q} - (TQ)\|_n^2\big]\Bigr\}^{1/2} + \inf_{f\in\mathcal{F}} \mathbb{E}\big[\|f - TQ\|_n^2\big] + C' \cdot V_{\max}\delta,
\end{aligned} \tag{C.44}$$

where C and C' are two positive absolute constants. Here in the first inequality we take the infimum over $\mathcal{F}$ because (C.31) holds for any $f \in \mathcal{F}$, and the second inequality holds because $\log N_\delta \le n$.

Now we invoke a simple fact to obtain the final bound for $\mathbb{E}[\|\widehat{Q}-TQ\|_n^2]$ from (C.44). Let a, b, and c be positive numbers satisfying $a^2 \le 2ab+c$. For any $\epsilon \in (0,1]$, since $2ab \le \epsilon \cdot a^2/(1+\epsilon)+(1+\epsilon)\cdot b^2/\epsilon$, we have

$$a^2 \le (1+\epsilon)^2 \cdot b^2/\epsilon + (1+\epsilon)\cdot c. \tag{C.45}$$

Therefore, applying (C.45) to (C.44) with $a^2 = \mathbb{E}[\|\widehat{Q}-TQ\|_n^2]$, $b = C \cdot V_{\max} \cdot \sqrt{\log N_\delta / n}$, and $c = \inf_{f\in\mathcal{F}} \mathbb{E}[\|f-TQ\|_n^2] + C' \cdot V_{\max} \cdot \delta$, we obtain

$$\mathbb{E}\big[\|\widehat{Q}-TQ\|_n^2\big] \le (1+\epsilon)\cdot \inf_{f\in\mathcal{F}} \mathbb{E}\big[\|f-TQ\|_n^2\big] + C \cdot V_{\max}^2 \cdot \log N_\delta/(n\epsilon) + C' \cdot V_{\max}\cdot\delta, \tag{C.46}$$

where C and C' are two positive absolute constants. Now we conclude the first step.

Step (ii). In this step, we relate the population risk $\|\widehat{Q}-TQ\|_\sigma^2$ with $\mathbb{E}[\|\widehat{Q}-TQ\|_n^2]$, which is characterized in the first step. To begin with, we generate n i.i.d. random variables $\{\widetilde{X}_i = (\widetilde{S}_i, \widetilde{A}_i)\}_{i\in[n]}$ following σ, which are independent of $\{(S_i, A_i, R_i, S_i')\}_{i\in[n]}$. Since $\|\widehat{Q}-f_{k^*}\|_\infty \le \delta$, for any $x \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned} &\Big|\big[\widehat{Q}(x)-(TQ)(x)\big]^2 - \big[f_{k^*}(x)-(TQ)(x)\big]^2\Big| \\ &\qquad = \big|\widehat{Q}(x)-f_{k^*}(x)\big|\cdot\big|\widehat{Q}(x)+f_{k^*}(x)-2(TQ)(x)\big| \le 4V_{\max}\cdot\delta, \end{aligned} \tag{C.47}$$

where the last inquality follows from the fact that $\|TQ\|_\infty \le V_{\max}$ and $\|f\|_\infty \le V_{\max}$ for any $f \in \mathcal{F}$. Then by the definition of $\|\widehat{Q}-TQ\|_\sigma^2$ and (C.47), we have

$$\begin{aligned} \|\widehat{Q}-TQ\|_\sigma^2 &= \mathbb{E}\bigg\{\frac{1}{n}\sum_{i=1}^n \big[\widehat{Q}(\widetilde{X}_i)-(TQ)(\widetilde{X}_i)\big]^2\bigg\} \\ &\le \mathbb{E}\bigg\{\|\widehat{Q}-TQ\|_n^2 + \frac{1}{n}\sum_{i=1}^n\big[f_{k^*}(\widetilde{X}_i)-(TQ)(\widetilde{X}_i)\big]^2 - \frac{1}{n}\sum_{i=1}^n\big[f_{k^*}(X_i)-(TQ)(\widetilde{X}_i)\big]^2\bigg\} + 8V_{\max}\cdot\delta \\ &= \mathbb{E}\big(\|\widehat{Q}-TQ\|_n^2\big) + \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n h_{k^*}(X_i,\widetilde{X}_i)\bigg] + 8V_{\max}\cdot\delta, \end{aligned} \tag{C.48}$$

where we apply (C.47) to obtain the first inequality, and in the last equality we define

$$h_j(x,y) = \big[f_j(y)-(TQ)(y)\big]^2 - \big[f_j(x)-(TQ)(x)\big]^2, \tag{C.49}$$

for any $(x,y) \in \mathcal{S}\times\mathcal{A}$ and any $j \in [N_\delta]$. Note that h_{k^*} is a random function since k^* is random. By the definition of h_j in (C.49), we have $|h_j(x,y)| \le 4V_{\max}^2$ for any $(x,y) \in \mathcal{S}\times\mathcal{A}$ and $\mathbb{E}[h_j(X_i,\widetilde{X}_i)] = 0$ for any $i \in [n]$. Moreover, the variance of $h_j(X_i,\widetilde{X}_i)$ is upper bounded by

$$\begin{aligned} \mathrm{Var}\big[h_j(X_i,\widetilde{X}_i)\big] &= 2\,\mathrm{Var}\Big\{\big[f_j(X_i)-(TQ)(X_i)\big]^2\Big\} \\ &\le 2\mathbb{E}\Big\{\big[f_j(X_i)-(TQ)(X_i)\big]^4\Big\} \le 8\Upsilon^2\cdot V_{\max}^2, \end{aligned}$$

where we define Υ by letting

$$\Upsilon^2 = \max\Big(4V_{\max}^2\cdot\log N_\delta/n,\ \max_{j\in[N_\delta]}\mathbb{E}\Big\{\big[f_j(X_i)-(TQ)(X_i)\big]^2\Big\}\Big). \tag{C.50}$$

Furthermore, we define

$$T = \sup_{j\in[N_\delta]} \biggl| \sum_{i=1}^n h_j(X_i, \widetilde{X}_i)/\Upsilon \biggr|. \tag{C.51}$$

Combining (C.48) and (C.51), we obtain

$$\|\widehat{Q} - TQ\|_\sigma^2 \le \mathbb{E}\bigl[\|\widehat{Q} - TQ\|_n^2\bigr] + \Upsilon/n \cdot \mathbb{E}(T) + 8V_{\max} \cdot \delta. \tag{C.52}$$

In the sequel, we utilize Bernstein's inequality to establish an upper bound for $\mathbb{E}(T)$, which is stated as follows for completeness.

Lemma C.3 (Bernstein's Inequality)**.** Let $U_1, \ldots U_n$ be n independent random variables satisfying $\mathbb{E}(U_i) = 0$ and $|U_i| \le M$ for all $i \in [n]$. Then for any $t > 0$, we have

$$\mathbb{P}\biggl(\biggl|\sum_{i=1}^n U_i\biggr| \ge t\biggr) \le 2\exp\biggl(\frac{-t^2}{2M \cdot t/3 + 2\sigma^2}\biggr),$$

where $\sigma^2 = \sum_{i=1}^n \mathrm{Var}(U_i)$ is the variance of $\sum_{i=1}^n U_i$.

We first apply Bernstein's inequality by setting $U_i = h_j(X_i, \widetilde{X}_i)/\Upsilon$ for each $i \in [n]$. Then we take a union bound for all $j \in [N_\delta]$ to obtain

$$\mathbb{P}(T \ge t) = \mathbb{P}\biggl[\sup_{j\in[N_\delta]} \frac{1}{n}\biggl|\sum_{i=1}^n h_j(X_i, \widetilde{X}_i)/\Upsilon\biggr| \ge t\biggr] \le 2N_\delta \cdot \exp\biggl\{\frac{-t^2}{8V_{\max}^2 \cdot [t/(3\Upsilon) + n]}\biggr\}. \tag{C.53}$$

Since T is nonnegative, we have $\mathbb{E}(T) = \int_0^\infty \mathbb{P}(T \ge t)\mathrm{d}t$. Thus, for any $u \in (0, 3\Upsilon \cdot n)$, by (C.53) it holds that

$$\begin{aligned}\mathbb{E}(T) &\le u + \int_u^\infty \mathbb{P}(T \ge t)\mathrm{d}t \le u + 2N_\delta \int_u^{3\Upsilon\cdot n} \exp\biggl(\frac{-t^2}{16V_{\max}^2 \cdot n}\biggr)\,\mathrm{d}t + 2N_\delta \int_{3\Upsilon\cdot n}^\infty \exp\biggl(\frac{-3\Upsilon\cdot t}{16V_{\max}^2}\biggr)\,\mathrm{d}t \\ &\le u + 32N_\delta \cdot V_{\max}^2 \cdot n/u \cdot \exp\biggl(\frac{-u^2}{16V_{\max}^2 \cdot n}\biggr) + 32N_\delta \cdot V_{\max}^2/(3\Upsilon) \cdot \exp\biggl(\frac{-9\Upsilon^2 \cdot n}{16V_{\max}^2}\biggr),\end{aligned} \tag{C.54}$$

where in the second inequality we use the fact that $\int_s^\infty \exp(-t^2/2)\mathrm{d}t \le 1/s \cdot \exp(-s^2/2)$ for all $s > 0$. Now we set $u = 4V_{\max}\sqrt{n \cdot \log N_\delta}$ in (C.54) and plug in the definition of Υ in (C.50) to obtain

$$\mathbb{E}(T) \le 4V_{\max}\sqrt{n \cdot \log N_\delta} + 8V_{\max}\sqrt{n/\log N_\delta} + 6V_{\max}\sqrt{n/\log N_\delta} \le 8V_{\max}\sqrt{n \cdot \log N_\delta}, \tag{C.55}$$

where the last inequality holds when $\log N_\delta \ge 4$. Moreover, the definition of Υ in (C.50) implies that $\Upsilon \le \max[2V_{\max}\sqrt{\log N_\delta/n}, \|\widehat{Q} - TQ\|_\sigma + \delta]$. In the following, we only need to consider the case where $\Upsilon \le \|\widehat{Q} - TQ\|_\sigma + \delta$, since we already have (6.3) if $\|\widehat{Q} - TQ\|_\sigma + \delta \le 2V_{\max}\sqrt{\log N_\delta/n}$, which concludes the proof.

Then, when $\Upsilon \le \|\widehat{Q} - TQ\|_\sigma + \delta$ holds, combining (C.52) and (C.55) we obtain

$$\begin{aligned}\|\widehat{Q} - TQ\|_\sigma^2 &\le \mathbb{E}\bigl[\|\widehat{Q} - TQ\|_n^2\bigr] + 8V_{\max}\sqrt{\log(N)/n} \cdot \|\widehat{Q} - TQ\|_\sigma + 8V_{\max}\sqrt{\log N_\delta/n} \cdot \delta + 8V_{\max} \cdot \delta \\ &\le \mathbb{E}\bigl[\|\widehat{Q} - TQ\|_n^2\bigr] + 8V_{\max}\sqrt{\log N_\delta/n} \cdot \|\widehat{Q} - TQ\|_\sigma + 16V_{\max} \cdot \delta.\end{aligned} \tag{C.56}$$

We apply the inequality in (C.45) to (C.56) with $a = \|\widehat{Q} - TQ\|_\sigma$, $b = 8V_{\max}\sqrt{\log N_\delta / n}$, and $c = \mathbb{E}[\|\widehat{Q} - TQ\|_n^2] + 16V_{\max} \cdot \delta$. Hence we finally obtain that

$$\begin{aligned}\|\widehat{Q} - TQ\|_\sigma^2 &\leq (1+\epsilon) \cdot \mathbb{E}\big[\|\widehat{Q} - TQ\|_n^2\big] \\ &\qquad + (1+\epsilon)^2 \cdot 64V_{\max} \cdot \log(N)/(n \cdot \epsilon) + (1+\epsilon) \cdot 18V_{\max} \cdot \delta, \end{aligned} \tag{C.57}$$

which concludes the second step of the proof.

Finally, combining these two steps together, namely, (C.46) and (C.57), we conclude that

$$\|\widehat{Q} - TQ\|_\sigma^2 \leq (1+\epsilon)^2 \cdot \inf_{f \in \mathcal{F}} \mathbb{E}\big[\|f - TQ\|_n^2\big] + C_1 \cdot V_{\max}^2 \cdot \log N_\delta / (n \cdot \epsilon) + C_2 \cdot V_{\max} \cdot \delta,$$

where C_1 and C_2 are two absolute constants. Moreover, since $Q \in \mathcal{F}$, we have

$$\inf_{f \in \mathcal{F}} \mathbb{E}\big[\|f - TQ\|_n^2\big] \leq \sup_{Q \in \mathcal{F}} \Big\{ \inf_{f \in \mathcal{F}} \mathbb{E}\big[\|f - TQ\|_n^2\big] \Big\},$$

which concludes the proof of Theorem 6.2. □

D Proof of Theorem 5.4

In this section, we present the proof of Theorem 5.4. The proof is similar to that of Theorem 4.4, which is presented in §6 in details. In the following, we follow the proof in §6 and only highlight the differences for brevity.

Proof. The proof requires two key ingredients, namely the error propagation and the statistical error incurred by a single step of Minimax-FQI. We note that Pérolat et al. (2015) establish error propagation for the state-value functions in the approximate modified policy iteration algorithm, which is more general than the FQI algorithm.

Theorem D.1 (Error Propagation)**.** Recall that $\{\widetilde{Q}_k\}_{0 \leq k \leq K}$ are the iterates of Algorithm 2 and (π_K, ν_K) is the equilibrium policy with respect to $\widetilde{Q}_K$. Let Q_K^* be the action-value function corresponding to $(\pi_K, \nu_{\pi_K}^*)$, where $\nu_{\pi_K}^*$ is the best-response policy of the second player against π_K. Then under Assumption 5.3, we have

$$\|Q^* - Q_K^*\|_{1,\mu} \leq \frac{2\phi_{\mu,\rho} \cdot \gamma}{(1-\gamma)^2} \cdot \varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{\max}, \tag{D.1}$$

where we define the maximum one-step approximation error $\varepsilon_{\max} = \max_{k \in [K]} \|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\|_\sigma$, and constant $\phi_{\mu,\nu}$ is specified in Assumption 5.3.

Proof. We note that the proof of Theorem 6.1 cannot be directly applied to prove this theorem. The main reason is that here we also need to consider the role played by the opponent, namely player two. Different from the MDP setting, here Q_K^* is a fixed point of a nonlinear operator due to the fact that player two adopts the optimal policy against π_K. Thus, we need to conduct a more refined analysis. See §D.1 for a detailed proof. □