A. Proof of Supporting Lemmas

In this section, we present the omitted proof of the technical lemmas used in out main theorems.

A.1. Proof of Lemma 6.1

Before we prove the error bound for the local linearization, we first present some useful lemmas from recent studies of overparameterized deep neural networks. Note that in the following lemmas, $\{C_i\}_{i=1,\dots}$ are universal constants that are independent of problem parameters such as $d, \boldsymbol{\theta}, m, L$ and their values can be different in different contexts. The first lemma states the uniform upper bound for the gradient of the deep neural network. Note that by definition, our parameter $\boldsymbol{\theta}$ is a long vector containing the concatenation of the vectorization of all the weight matrices. Correspondingly, the gradient $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \mathbf{x})$ is also a long vector.

Lemma A.1 (Lemma B.3 in Cao & Gu (2019b))**.** Let $\boldsymbol{\theta} \in \mathbb{B}(\boldsymbol{\theta}_0, \omega)$ with the radius satisfying $C_1 d^{3/2} L^{-1} m^{-3/2} \leq \omega \leq C_2 L^{-6} (\log m)^{-3/2}$. Then for all unit vectors in $\mathbb{R}^d$, i.e., $\mathbf{x} \in S^{d-1}$, the gradient of the neural network f defined in (4.2) is bounded as $\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \mathbf{x})\|_2 \leq C_3 \sqrt{m}$ with probability at least $1 - L^2 \exp(-C_4 m \omega^{2/3} L)$.

The second lemma provides the perturbation bound for the gradient of the neural network function. Note that the original theorem holds for any fixed d dimensional unit vector $\mathbf{x}$. However, due to the choice of ω and its dependency on m and d, it is easy to modify the results to hold for all $\mathbf{x} \in S^{d-1}$.

Lemma A.2 (Theorem 5 in Allen-Zhu et al. (2019b))**.** Let $\boldsymbol{\theta} \in \mathbb{B}(\boldsymbol{\theta}_0, \omega)$ with the radius satisfying

$$C_1 d^{3/2} L^{-3/2} m^{-3/2} (\log m)^{-3/2} \leq \omega \leq C_2 L^{-9/2} (\log m)^{-3}.$$

Then for all $\mathbf{x} \in S^{d-1}$, with probability at least $1 - \exp(-C_3 m \omega^{2/3} L)$ over the randomness of $\boldsymbol{\theta}_0$, it holds that

$$\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \mathbf{x}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; \mathbf{x})\|_2 \leq C_4 \omega^{1/3} L^3 \sqrt{\log m} \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; \mathbf{x})\|_2.$$

Now we are ready to bound the linearization error.

Proof of Lemma 6.1. Recall the definition of $\mathbf{g}_t(\boldsymbol{\theta}_t)$ and $\mathbf{m}_t(\boldsymbol{\theta}_t)$ in (4.5) and (6.2) respectively. We have

$$\begin{aligned}
\|\mathbf{g}_t(\boldsymbol{\theta}_t) - \mathbf{m}_t(\boldsymbol{\theta}_t)\|_2 &= \big\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t; s_t, a_t) \Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t) - \nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t) \widehat{\Delta}(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big\|_2 \\
&\leq \big\|(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t; s_t, a_t) - \nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)) \Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big\|_2 \\
&\quad + \big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t) \big(\Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t) - \widehat{\Delta}(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big)\big\|_2. \qquad \text{(A.1)}
\end{aligned}$$

Since $\widehat{f}(\boldsymbol{\theta}) \in \mathcal{F}_{\boldsymbol{\Theta}, m}$, we have $\widehat{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_0) + \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0), \boldsymbol{\theta} - \boldsymbol{\theta}_0 \rangle$ and $\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0)$. Then with probability at least $1 - 2L^2 \exp(-C_1 m \omega^{2/3} L)$, we have

$$\begin{aligned}
&\big\|(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t; s_t, a_t) - \nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)) \Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big\|_2 \\
&= |\Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)| \cdot \big\|(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t; s_t, a_t) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s_t, a_t))\big\|_2 \\
&\leq C_2 \omega^{1/3} L^3 \sqrt{m \log m} |\Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)|,
\end{aligned}$$

where the inequality comes from Lemmas A.1 and A.2. By Lemma 6.4, with probability at least $1 - \delta$, it holds that

$$|\Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)| = \Big|f(\boldsymbol{\theta}_t; s_t, a_t) - r_t - \gamma \max_{b \in \mathcal{A}} f(\boldsymbol{\theta}_t; s_{t+1}, b)\Big| \leq (2 + \gamma) C_3 \sqrt{\log(T/\delta)},$$

which further implies that with probability at least $1 - \delta - 2L^2 \exp(-C_1 m \omega^{2/3} L)$, we have

$$\begin{aligned}
&\big\|(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t; s_t, a_t) - \nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)) \Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big\|_2 \\
&\leq (2 + \gamma) C_2 C_3 \omega^{1/3} L^3 \sqrt{m \log m \log(T/\delta)}.
\end{aligned}$$

For the second term in (A.1), we have

$$\big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t) \big(\Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t) - \widehat{\Delta}(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big)\big\|_2$$

$$
\begin{aligned}
&\leq \big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\big(f(\boldsymbol{\theta}_t; s_t, a_t) - \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\big)\big\|_2 \\
&\quad + \Big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\Big(\max_{b\in\mathcal{A}} f(\boldsymbol{\theta}_t; s_{t+1}, b) - \max_{b\in\mathcal{A}} \widehat{f}(\boldsymbol{\theta}_t; s_{t+1}, b)\Big)\Big\|_2 \\
&\leq \big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\big\|_2 \cdot \big|f(\boldsymbol{\theta}_t; s_t, a_t) - \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\big| \\
&\quad + \big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\|_2 \max_{b\in\mathcal{A}} \big|f(\boldsymbol{\theta}_t; s_{t+1}, b) - \widehat{f}(\boldsymbol{\theta}; s_{t+1}, b)\big|. \qquad \text{(A.2)}
\end{aligned}
$$

By Lemma 6.4, with probability at least $1-\delta$ we have

$$
|f(\boldsymbol{\theta}_t; s_t, a_t) - \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)| \leq \omega^{4/3} L^{11/3}\sqrt{m\log m} + C_4\omega^2 L^4\sqrt{m},
$$

for all $(s_t, a_t) \in \mathcal{S}\times\mathcal{A}$ such that $\|\phi(s_t, a_t)\|_2 = 1$. Substituting the above result into (A.2) and applying the gradient bound in Lemma A.1, we obtain with probability at least $1-\delta-L^2\exp(-C_1 m\omega^{2/3}L)$ that

$$
\begin{aligned}
&\big\|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta}_t; s_t, a_t)\big(\Delta(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t) - \widehat{\Delta}(s_t, a_t, s_{t+1}; \boldsymbol{\theta}_t)\big)\big\|_2 \\
&\leq C_5\omega^{4/3}L^{11/3} m\sqrt{\log m} + C_6\omega^2 L^4 m.
\end{aligned}
$$

Note that the above results require that the choice of ω should satisfy all the constraints in Lemmas A.1, 6.4 and A.2, of which the intersection is

$$
C_7 d^{3/2} L^{-1} m^{-3/4} \leq \omega \leq C_8 L^{-6}(\log m)^{-3}.
$$

Therefore, the error of the local linearization of $\mathbf{g}_t(\boldsymbol{\theta}_t)$ can be upper bounded by

$$
\begin{aligned}
|\langle \mathbf{g}_t(\boldsymbol{\theta}_t) - \mathbf{m}_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}^*\rangle| &\leq (2+\gamma) C_2 C_3 \omega^{1/3} L^3 \sqrt{m\log m \log(T/\delta)}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 \\
&\quad + \big(C_5\omega^{4/3}L^{11/3} m\sqrt{\log m} + C_6\omega^2 L^4 m\big)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2,
\end{aligned}
$$

which holds with probability at least $1-2\delta-3L^2\exp(-C_1 m\omega^{2/3}L)$ over the randomness of the initial point. For the upper bound of the norm of $\mathbf{g}_t$, by Lemmas 6.4 and A.1, we have

$$
\begin{aligned}
\|\mathbf{g}_t(\boldsymbol{\theta}_t)\|_2 &= \Big\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t; s_t, a_t)\Big(f(\boldsymbol{\theta}_t; s_t, a_t) - r_t - \gamma \max_{b\in\mathcal{A}} f(\boldsymbol{\theta}_t; s_{t+1}, b)\Big)\Big\|_2 \\
&\leq (2+\gamma) C_9\sqrt{m\log(T/\delta)}
\end{aligned}
$$

holds with probability at least $1-\delta-L^2\exp(-C_1 m\omega^{2/3}L)$. □

A.2. Proof of Lemma 6.2

Let us define $\zeta_t(\boldsymbol{\theta}) = \langle \mathbf{m}_t(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle$, which characterizes the bias of the data. Different from the similar quantity ζ_t in Bhandari et al. (2018), our definition is based on the local linearization of f, which is essential to the analysis in our proof. It is easy to verify that $\mathbb{E}[\mathbf{m}_t(\boldsymbol{\theta})] = \overline{\mathbf{m}}(\boldsymbol{\theta})$ for any fixed and deterministic $\boldsymbol{\theta}$. However, it should be noted that $\mathbb{E}[\mathbf{m}_t(\boldsymbol{\theta}_t)|\boldsymbol{\theta}_t = \boldsymbol{\theta}] \neq \overline{\mathbf{m}}(\boldsymbol{\theta})$ because $\boldsymbol{\theta}_t$ depends on all historical states and actions $\{s_t, a_t, s_{t-1}, a_{t-1}, \ldots\}$ and $\mathbf{m}_t(\cdot)$ depends on the current observation $\{s_t, a_t, s_{t+1}\}$ and thus also depends on $\{s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2}, \ldots\}$. Therefore, we need a careful analysis of Markov chains to decouple the dependency between $\boldsymbol{\theta}_t$ and $\mathbf{m}_t(\cdot)$.

The following lemma uses data processing inequality to provide an information theoretic control of coupling.

Lemma A.3 (Control of coupling (Bhandari et al., 2018))**.** Consider two random variables X and Y that form the following Markov chain:

$$
X \to s_t \to s_{t+\tau} \to Y,
$$

where $t \in \{0, 1, 2, \ldots\}$ and $\tau > 0$. Suppose Assumption 5.2 holds. Let X' and Y' be independent copies drawn from the marginal distributions of X and Y respectively, i.e., $\mathbb{P}(X' = \cdot, Y' = \cdot) = \mathbb{P}(X = \cdot)\otimes\mathbb{P}(Y = \cdot)$. Then for any bounded function $h : \mathcal{S}\times\mathcal{S} \to \mathbb{R}$, it holds that

$$
|\mathbb{E}[h(X, Y)] - \mathbb{E}[h(X', Y')]| \leq 2\sup_{s,s'} |h(s, s')|\lambda\rho^{\tau}.
$$

Proof of Lemma 6.2. The proof of this lemma is adapted from (Bhandari et al., 2018), where the result was originally proved for linear function approximation of temporal difference learning. We first show that $\zeta_t(\boldsymbol{\theta})$ is Lipschitz. For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{B}(\boldsymbol{\theta}_0, \omega)$, we have

$$\begin{aligned}\zeta_t(\boldsymbol{\theta}) - \zeta_t(\boldsymbol{\theta}') &= \langle \mathbf{m}_t(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle - \langle \mathbf{m}_t(\boldsymbol{\theta}') - \overline{\mathbf{m}}(\boldsymbol{\theta}'), \boldsymbol{\theta}' - \boldsymbol{\theta}^* \rangle \\ &= \langle \mathbf{m}_t(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta}) - (\mathbf{m}_t(\boldsymbol{\theta}') - \overline{\mathbf{m}}(\boldsymbol{\theta}')), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \\ &\quad + \langle \mathbf{m}_t(\boldsymbol{\theta}') - \overline{\mathbf{m}}(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle,\end{aligned}$$

which directly implies

$$\begin{aligned}|\zeta_t(\boldsymbol{\theta}) - \zeta_t(\boldsymbol{\theta}')| &\le \|\mathbf{m}_t(\boldsymbol{\theta}) - \mathbf{m}_t(\boldsymbol{\theta}')\|_2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 + \|\overline{\mathbf{m}}(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta}')\|_2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \\ &\quad + \|\mathbf{m}_t(\boldsymbol{\theta}') - \overline{\mathbf{m}}(\boldsymbol{\theta}')\|_2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.\end{aligned}$$

By the definition of $\mathbf{m}_t$, we have

$$\begin{aligned}&\|\mathbf{m}_t(\boldsymbol{\theta}) - \mathbf{m}_t(\boldsymbol{\theta}')\|_2 \\ &= \Big\| \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0) \Big(\big(f(\boldsymbol{\theta}; s, a) - f(\boldsymbol{\theta}'; s, a) \big) - \gamma \Big(\max_{b \in \mathcal{A}} f(\boldsymbol{\theta}; s', b) - \max_{b \in \mathcal{A}} f(\boldsymbol{\theta}'; s', b) \Big) \Big) \Big\|_2 \\ &\le (1+\gamma) C_3^2 m \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2,\end{aligned}$$

which holds with probability at least $1 - L^2 \exp(-C_4 m \omega^{2/3} L)$, where we used the fact that the neural network function is Lipschitz with parameter $C_3\sqrt{m}$ by Lemma A.1. Similar bound can also be established for $\|\overline{\mathbf{m}}_t(\boldsymbol{\theta}) - \overline{\mathbf{m}}_t(\boldsymbol{\theta}')\|$ in the same way. Note that for $\boldsymbol{\theta} \in \mathbb{B}(\boldsymbol{\theta}_0, \omega)$ with ω and m satisfying the conditions in Lemma 6.1, we have by the definition in (6.2) that

$$\begin{aligned}\|\mathbf{m}_t(\boldsymbol{\theta})\|_2 &\le \Big(|\widehat{f}(\boldsymbol{\theta}; s, a)| + r(s, a) + \gamma \big| \max_b \widehat{f}(\boldsymbol{\theta}; s', b) \big| \Big) \|\nabla_{\boldsymbol{\theta}} \widehat{f}(\boldsymbol{\theta})\|_2 \\ &\le (2+\gamma)(|f(\boldsymbol{\theta}_0)| + \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0)\|_2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2) \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0)\|_2 \\ &\le (2+\gamma) C_3 (C_8 \sqrt{m} \sqrt{\log(T/\delta)} + C_3 m \omega). \qquad \text{(A.3)}\end{aligned}$$

The same bound can be established for $\|\bar{\mathbf{m}}_t\|$ in a similar way. Therefore, we have $|\zeta_t(\boldsymbol{\theta}) - \zeta_t(\boldsymbol{\theta}')| \le \ell_{m,L} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, where $\ell_{m,L}$ is defined as

$$\ell_{m,L} = 2(1+\gamma) C_3^2 m \omega + (2+\gamma) C_3 (C_8 \sqrt{m} \sqrt{\log(T/\delta)} + C_3 m \omega).$$

Applying the above inequality recursively, for all $\tau = 0, \ldots, t$, we have

$$\begin{aligned}\zeta_t(\boldsymbol{\theta}_t) &\le \zeta_t(\boldsymbol{\theta}_{t-\tau}) + \ell_{m,L} \sum_{i=t-\tau}^{t-1} \|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\|_2 \\ &\le \zeta_t(\boldsymbol{\theta}_{t-\tau}) + (2+\gamma) C_3 (C_8 \sqrt{m} \sqrt{\log(T/\delta)} + C_3 m \omega) \ell_{m,L} \sum_{i=t-\tau}^{t-1} \eta_i. \qquad \text{(A.4)}\end{aligned}$$

Next, we need to bound $\zeta_t(\boldsymbol{\theta}_{t-\tau})$. Define the observed tuple $O_t = (s_t, a_t, s_{t+1})$ as the collection of the current state and action and the next state. Note that $\boldsymbol{\theta}_{t-\tau} \to s_{t-\tau} \to s_t \to O_t$ forms a Markov chain induced by the target policy π. Recall that $\mathbf{m}_t(\cdot)$ depends on the observation O_t. Let's rewrite $\mathbf{m}(\boldsymbol{\theta}, O_t) = \mathbf{m}_t(\boldsymbol{\theta})$. Similarly, we can rewrite $\zeta_t(\boldsymbol{\theta})$ as $\zeta(\boldsymbol{\theta}, O_t)$. Let $\boldsymbol{\theta}'_{t-\tau}$ and O'_t be independently drawn from the marginal distributions of $\boldsymbol{\theta}_{t-\tau}$ and O_t respectively. Applying Lemma A.3 yields

$$\mathbb{E}[\zeta(\boldsymbol{\theta}_{t-\tau}, O_t)] - \mathbb{E}[\zeta(\boldsymbol{\theta}'_{t-\tau}, O'_t)] \le 2 \sup_{\boldsymbol{\theta}, O} |\zeta(\boldsymbol{\theta}, O)| \lambda \rho^{\tau},$$

where we used the uniform mixing result in Assumption 5.2. By definition $\boldsymbol{\theta}'_{t-\tau}$ and O'_t are independent, which implies $\mathbb{E}[\mathbf{m}(\boldsymbol{\theta}'_{t-\tau}, O'_t) | \boldsymbol{\theta}'_{t-\tau}] = \overline{\mathbf{m}}(\boldsymbol{\theta}'_{t-\tau})$ and

$$\mathbb{E}[\zeta(\boldsymbol{\theta}'_{t-\tau}, O'_t)] = \mathbb{E}[\mathbb{E}[\langle \mathbf{m}(\boldsymbol{\theta}'_{t-\tau}, O'_t) - \overline{\mathbf{m}}(\boldsymbol{\theta}'_{t-\tau}), \boldsymbol{\theta}'_{t-\tau} - \boldsymbol{\theta}^* \rangle] | \boldsymbol{\theta}'_{t-\tau}] = 0.$$

Moreover, by the definition of $\zeta(\cdot,\cdot)$, we have

$$|\zeta(\boldsymbol{\theta}, O) \le \|\mathbf{m}_t(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta})\|_2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le 2(2+\gamma)C_3(C_8\sqrt{m}\sqrt{\log(T/\delta)} + C_3 m\omega)\omega,$$

where the second inequality is due to (A.3) and that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le \omega$. Therefore, for any $\tau = 0, \ldots, t$, we have

$$\begin{aligned}\mathbb{E}[\zeta_t(\boldsymbol{\theta}_t)] &\le \mathbb{E}\zeta_t(\boldsymbol{\theta}_{t-\tau}) + (2+\gamma)C_3(C_8\sqrt{m}\sqrt{\log(T/\delta)} + C_3 m\omega)\ell_{m,L}\sum_{i=t-\tau}^{t-1}\eta_i \\ &\le (2+\gamma)C_3(C_8\sqrt{m}\sqrt{\log(T/\delta)} + C_3 m\omega)(\omega\lambda\rho^\tau + \ell_{m,L}\tau\eta_{t-\tau}). \end{aligned} \tag{A.5}$$

Define τ^* as the mixing time of the Markov chain that satisfies

$$\tau^* = \min\{t = 0, 1, 2, \ldots | \lambda\rho^t \le \eta_T\}.$$

When $t \le \tau^*$, we choose $\tau = t$ in (A.5). Note that η_t is nondecreasing. We obtain

$$\begin{aligned}\mathbb{E}[\zeta_t(\boldsymbol{\theta}_t)] &\le \mathbb{E}[\zeta_t(\boldsymbol{\theta}_0)] + 2(2+\gamma)C_3(C_8\sqrt{m}\sqrt{\log(T/\delta)} + C_3 m\omega)\ell_{m,L}\tau^*\eta_0 \\ &= 2(2+\gamma)C_3(C_8\sqrt{m}\sqrt{\log(T/\delta)} + C_3 m\omega)\ell_{m,L}\tau^*\eta_0,\end{aligned}$$

where we used the fact that the initial point $\boldsymbol{\theta}_0$ is independent of $\{s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0\}$ and thus independent of $\zeta_t(\cdot)$. When $t > \tau^*$, we can choose $\tau = \tau^*$ in (A.5) and obtain

$$\begin{aligned}\mathbb{E}[\zeta_t(\boldsymbol{\theta}_t)] &\le (2+\gamma)C_3(C_8\sqrt{m}\sqrt{\log(T/\delta)} + C_3 m\omega)(\omega\eta_T + \ell_{m,L}\tau^*\eta_{t-\tau^*}) \\ &\le \widetilde{C}(m\log(T/\delta) + m^2\omega^2)\tau^*\eta_{t-\tau^*},\end{aligned}$$

where $\widetilde{C} > 0$ is a universal constant, which completes the proof. □

A.3. Proof of Lemma 6.3

Proof of Lemma 6.3. To simplify the notation, we use $\mathbb{E}_\pi$ to denote $\mathbb{E}_{\mu,\pi,\mathcal{P}}$, namely, the expectation over $s \in \mu, a \sim \pi(\cdot|s)$ and $s' \sim \mathcal{P}(\cdot|s,a)$, in the rest of the proof. By the definition of $\overline{\mathbf{m}}$ in (6.2), we have

$$\begin{aligned}&\langle \overline{\mathbf{m}}(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle \\ &= \mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\widehat{\Delta}(s,a,s';\boldsymbol{\theta}) - \widehat{\Delta}(s,a,s';\boldsymbol{\theta}^*)\big)\langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\big] \\ &= \mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\widehat{f}(\boldsymbol{\theta};s,a) - \widehat{f}(\boldsymbol{\theta}^*;s,a)\big)\langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\big] \\ &\quad - \gamma\mathbb{E}_{\mu,\pi,\mathcal{P}}\Big[\Big(\max_{b\in\mathcal{A}}\widehat{f}(\boldsymbol{\theta};s',b) - \max_{b\in\mathcal{A}}\widehat{f}(\boldsymbol{\theta}^*;s',b)\Big)\langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\Big],\end{aligned}$$

where in the first equation we used the fact that $\nabla_{\boldsymbol{\theta}}\widehat{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\widehat{f} \in \mathcal{F}_{\boldsymbol{\Theta},m}$. Further by the property of the local linearization of f at $\boldsymbol{\theta}_0$, we have

$$\widehat{f}(\boldsymbol{\theta};s,a) - \widehat{f}(\boldsymbol{\theta}^*;s,a) = \langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle, \tag{A.6}$$

which further implies

$$\begin{aligned}&\mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\widehat{f}(\boldsymbol{\theta};s,a) - \widehat{f}(\boldsymbol{\theta}^*;s,a)\big)\langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle|\boldsymbol{\theta}_0\big] \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top\mathbb{E}\big[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a)\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a)^\top|\boldsymbol{\theta}_0\big](\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= m\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi}^2.\end{aligned}$$

where $\boldsymbol{\Sigma}_\pi$ is defined in Assumption 5.3. For the other term, we define $b_{\max}^{\boldsymbol{\theta}} = \operatorname{argmax}_{b\in\mathcal{A}}\widehat{f}(\boldsymbol{\theta};s',b)$ and $b_{\max}^{\boldsymbol{\theta}^*} = \operatorname{argmax}_{b\in\mathcal{A}}\widehat{f}(\boldsymbol{\theta}^*;s',b)$. Then we have

$$\mathbb{E}_{\mu,\pi,\mathcal{P}}\Big[\Big(\max_{b\in\mathcal{A}}\widehat{f}(\boldsymbol{\theta};s',b) - \max_{b\in\mathcal{A}}\widehat{f}(\boldsymbol{\theta}^*;s',b)\Big)\langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0;s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\Big]$$

$$= \mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\widehat{f}(\boldsymbol{\theta}; s', b_{\max}^{\boldsymbol{\theta}}) - \widehat{f}(\boldsymbol{\theta}^*; s', b_{\max}^{\boldsymbol{\theta}^*})\big)\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\big]. \tag{A.7}$$

Applying Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\widehat{f}(\boldsymbol{\theta}; s', b_{\max}^{\boldsymbol{\theta}}) - \widehat{f}(\boldsymbol{\theta}^*; s', b_{\max}^{\boldsymbol{\theta}^*})\big)\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\big] \\
&\leq \sqrt{\mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\max_b |(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s', b)|\big)^2\big]}\sqrt{\mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, a)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\big)^2\big]} \\
&= m\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi},
\end{aligned}$$

where we used the fact that $\boldsymbol{\Sigma}_\pi^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = 1/m\mathbb{E}_{\mu,\pi,\mathcal{P}}[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, \widetilde{b}_{\max})\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, \widetilde{b}_{\max})^\top]$ and $\widetilde{b}_{\max} = \mathrm{argmax}_{b\in\mathcal{A}} |\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, b), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle|$ according to (5.6). Substituting the above results into (A.7), we obtain

$$\begin{aligned}
&\mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\max_{b\in\mathcal{A}} \widehat{f}(\boldsymbol{\theta}; s', b) - \max_{b\in\mathcal{A}} \widehat{f}(\boldsymbol{\theta}^*; s', b)\big)\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0; s, a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\big] \\
&\leq m\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi},
\end{aligned}$$

which immediately implies

$$\begin{aligned}
\langle \overline{\mathbf{m}}(\boldsymbol{\theta}) - \overline{\mathbf{m}}(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle &\geq m\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi} \cdot \Big(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi} - \gamma\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}\Big) \\
&= m\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi} \cdot \frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi}^2 - \gamma^2\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}^2}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi} + \gamma\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}} \\
&\geq m(1 - \alpha^{-1/2})\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_\pi}^2 \\
&= (1 - \alpha^{-1/2})\mathbb{E}_{\mu,\pi,\mathcal{P}}\big[\big(\widehat{f}(\boldsymbol{\theta}) - \widehat{f}(\boldsymbol{\theta}^*)\big)^2 | \boldsymbol{\theta}_0\big],
\end{aligned}$$

where the second inequality is due to Assumption 5.3 and the last equation is due to (A.6) and the definition of $\boldsymbol{\Sigma}_\pi$ in (5.5). □