A. Proof of Results for Double Sampling (Theorem 4.1)

Throughout the supplementary materials, we omit the subscript ρ in population Rademacher complexty $\mathcal{R}_n^{\rho}(\cdot)$ if the distribution is clear from the context.

In this part, we prove Theorem 4.1 in Section 4. We first define some auxiliary notations to simplify the writing. We divide the dataset $\widetilde{\mathcal{D}}$ into $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}}_1 \cup \ldots \cup \widetilde{\mathcal{D}}_H$, where $\widetilde{\mathcal{D}}_h$ consists of n independent sample tuples collected at the h^{th} time step. For $f_h, g_h \in \mathcal{F}_h$, denote

$$\ell_{\text{DS}}(g_h, f_h)(s, a, r, s', \tilde{s}') := \big(g_h(s,a) - r - V_{f_h}(s')\big)^2 - \frac{1}{2}\big(V_{f_h}(s') - V_{f_h}(\tilde{s}')\big)^2.$$

Define an expected value $\mathbb{E}_{\mu_h}\ell_{\text{DS}}(g_h, f_h) := \mathbb{E}\big[\ell_{\text{DS}}(g_h, f_h)(s, a, r, s', \tilde{s}')\big]$ with $(s,a) \sim \mu_h$, $r = r_h(s,a)$, $s', \tilde{s}' \overset{i.i.d.}{\sim} \mathbb{P}_h(\cdot \mid s_h, a_h)$ and its empirical version $\hat{\ell}_{\text{DS}}(g_h, f_h) := \frac{1}{n}\sum_{(s,a,r,s',\tilde{s}',h)\in\widetilde{\mathcal{D}}_h} \ell_{\text{DS}}(g_h, f_h)(s, a, r, s', \tilde{s}')$. It is easy to see that $\mathbb{E}_{\mu_h}\ell_{\text{DS}}(g_h, f_h) = \|g_h - \mathcal{T}_h^{\star} f_h\|_{\mu_h}^2$. For any $f = (f_1, \ldots, f_H) \in \mathcal{F}$, we have

$$L_{\text{DS}}(f) := \frac{1}{H}\sum_{h=1}^{H} \ell_{\text{DS}}(f_h, f_{h+1}), \quad \mathbb{E}_{\mu} L_{\text{DS}}(f) = \mathcal{E}(f) \quad \text{and} \quad \hat{L}_{\text{DS}}(f) := \frac{1}{H}\sum_{h=1}^{H} \hat{\ell}_{\text{DS}}(f_h, f_{h+1}),$$

where $f_{H+1} := 0$. Note that the loss function $\hat{L}_{\text{DS}}(f)$ is an empirical estimation of $\mathcal{E}(f)$.

Theorem 4.1 provides an upper error bound for the BRM estimator $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{L}_{\text{DS}}(f)$, of which the proof is given below.

Theorem 4.1. *There exists an absolute constant $c > 0$, with probability at least $1-\delta$, the ERM estimator $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{L}_{DS}(f)$ satisfies the following:*

$$\begin{aligned}\mathcal{E}(\hat{f}) \leq & \min_{f\in\mathcal{F}} \mathcal{E}(f) + cH^2\sqrt{\frac{\log(1/\delta)}{n}} \\ & + c\sum_{h=1}^{H}\big(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + \mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})\big).\end{aligned}$$

Proof of Theorem 4.1. We apply the uniform concentration inequalites in Lemma G.1. Let $f^{\dagger}$ be a minimizer of the Bellman error within the function class $\mathcal{F}$, *i.e.* $f^{\dagger} \in \arg\min_{f\in\mathcal{F}} \mathcal{E}(f)$. By noting that $L_{\text{DS}}(f) \in \big[-2H^2, 4H^2\big]$, we have with probabliity at least $1-\delta$,

$$\mathbb{E}_{\mu} L_{\text{DS}}(\hat{f}) - \mathbb{E}_{\mu} L_{\text{DS}}(f^{\dagger}) \leq \big(\hat{L}_{\text{DS}}(\hat{f}) - \hat{L}_{\text{DS}}(f^{\dagger})\big) + 2\mathcal{R}_n\big(\{L_{\text{DS}}(f) \mid f \in \mathcal{F}\}\big) + 6H^2\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{10}$$

We use the relations $\mathbb{E}_{\mu} L_{\text{DS}}(\hat{f}) = \mathcal{E}(\hat{f})$, $\mathbb{E}_{\mu} L_{\text{DS}}(f^{\dagger}) = \mathcal{E}(f^{\dagger}) = \min_{f\in\mathcal{F}} \mathcal{E}(f)$ and $\hat{L}_{\text{DS}}(\hat{f}) \leq \hat{L}_{\text{DS}}(f^{\dagger})$ and reduce eq. (10) to

$$\mathcal{E}(\hat{f}) \leq \min_{f\in\mathcal{F}} \mathcal{E}(f) + 2\mathcal{R}_n\big(\{L_{\text{DS}} \mid f \in \mathcal{F}\}\big) + 6H^2\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{11}$$

It then remains to simplify the form of Rademacher complexity $\mathcal{R}_n\big(\{L_{\text{DS}}(f) \mid f \in \mathcal{F}\}\big)$.

Due to the sub-additivity of Rademacher complexity, we have

$$\mathcal{R}_n\big(\{L_{\text{DS}}(f) \mid f \in \mathcal{F}\}\big) \leq \frac{1}{H}\sum_{h=1}^{H} \mathcal{R}_n\big(\{\ell_{\text{DS}}(f_h, f_{h+1}) \mid f_h \in \mathcal{F}_h, f_{h+1} \in \mathcal{F}_{h+1}\}\big). \tag{12}$$

In order to tackle the term $\mathcal{R}_n\big(\{\ell_{\text{DS}}(f_h, f_{h+1}) \mid f_h \in \mathcal{F}_h, f_{h+1} \in \mathcal{F}_{h+1}\}\big)$ on the right hand side, we apply the vector-form contraction property of Rademacher complexity in Lemma G.7. By letting

$$\tilde{\phi}_{h,1} := f_h(s,a), \quad \tilde{\phi}_{h,2} := r_h + V_{f_{h+1}}(s') \quad \text{and} \quad \tilde{\phi}_{h,3} := r_h + V_{f_{h+1}}(\tilde{s}'),$$

we can write

$$\ell_{\mathrm{DS}}(f_h, f_{h+1}) = \frac{1}{2}\big(\tilde{\phi}_{h,1}, \tilde{\phi}_{h,2}, \tilde{\phi}_{h,3}\big)^\top \tilde{\boldsymbol{A}} \begin{pmatrix} \tilde{\phi}_{h,1} \\ \tilde{\phi}_{h,2} \\ \tilde{\phi}_{h,3} \end{pmatrix} \qquad \text{with } \tilde{\boldsymbol{A}} = \begin{pmatrix} 2 & -2 & 0 \\ -2 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

Since the spectral norm $\|\tilde{\boldsymbol{A}}\|_2 \le 4$ and $\big\|\big(\tilde{\phi}_{h,1}, \tilde{\phi}_{h,2}, \tilde{\phi}_{h,3}\big)^\top\big\|_2 \le \sqrt{3}H$ due to the boundedness of f_h and $\mathcal{T}_h^\star f_{h+1}$, we find that $\ell_{\mathrm{DS}}(f_h, f_{h+1})$ is $(4\sqrt{3}H)$-Lipschitz with respect to the vector $\big(\tilde{\phi}_{h,1}, \tilde{\phi}_{h,2}, \tilde{\phi}_{h,3}\big)^\top$. Lemma G.7 then implies

$$\mathcal{R}_n\big(\big\{\ell_{\mathrm{DS}}(f_h, f_{h+1}) \;\big|\; f_h \in \mathcal{F}_h, f_{h+1} \in \mathcal{F}_{h+1}\big\}\big) \le 10H\Big(\mathcal{R}_n\big(\big\{\tilde{\phi}_{h,1}\big\}\big) + \mathcal{R}_n\big\{\tilde{\phi}_{h,2}\big\}\big) + \mathcal{R}_n\big\{\tilde{\phi}_{h,3}\big\}\big)\Big). \tag{13}$$

Recalling that s' and $\tilde{s}'$ are *i.i.d.* conditioned on (s,a), we use the sub-additivity of Rademacher complexity and find that

$$\begin{aligned} &\mathcal{R}_n\big(\big\{\tilde{\phi}_{h,1}\big\}\big) \le \mathcal{R}_n^{\mu_h}(\mathcal{F}_h) \\ &\mathcal{R}_n\big(\big\{\tilde{\phi}_{h,2}\big\}\big) = \mathcal{R}_n\big(\big\{\tilde{\phi}_{h,3}\big\}\big) \le \mathcal{R}_n(\{r_h\}) + \mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}}), \end{aligned} \tag{14}$$

where ν_h is the marginal distribution of s' in the h^{th} step. Note that $\{r_h\}$ is a singleton, therefore, $\mathcal{R}_n(\{r_h\}) = 0$. It follows from eqs. (13) and (14) that

$$\mathcal{R}_n\big(\big\{\ell_{\mathrm{DS}}(f_h, f_{h+1}) \;\big|\; f_h \in \mathcal{F}_h, f_{h+1} \in \mathcal{F}_{h+1}\big\}\big) \le 10H\big(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + 2\mathcal{R}_n^{\nu_h}(V_{f_{h+1}})\big). \tag{15}$$

Combining eqs. (12) and (15), we learn that

$$\mathcal{R}_n\big(\big\{L_{\mathrm{DS}}(f) \;\big|\; f \in \mathcal{F}\big\}\big) \le 10\sum_{h=1}^{H}\big(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + 2\mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})\big). \tag{16}$$

Plugging eq. (16) into eq. (11), we finish the proof.

□

B. Proof of Results for FQI (Theorems 5.2 and 5.3)

In this section, we analyze the FQI estimator defined in Algorithm 1. For any $f_h \in \mathcal{F}_h$ and $f_{h+1} \in \mathcal{F}_{h+1}$, we denote

$$\ell(f_h, f_{h+1})(s,a,r,s') := \big(f_h(s,a) - r - V_{f_{h+1}}(s')\big)^2, \tag{17}$$

therefore, $\hat{\ell}_h(f_h, f_{h+1}) := \frac{1}{n}\sum_{(s,a,r,s',h)\in\mathcal{D}_h} \ell(f_h, f_{h+1})(s,a,r,s')$. Note that each iteration in FQI solves an empirical loss minimization problem $\hat{f}_h := \arg\min_{f_h\in\mathcal{F}_h} \hat{\ell}_h(f_h, \hat{f}_{h+1})$. The empirical loss $\hat{\ell}_h(f_h, \hat{f}_{h+1})$ approximates

$$\begin{aligned} \mathbb{E}_{\mu_h}\ell(f_h, \hat{f}_{h+1}) =& \mathbb{E}\big[\ell(f_h, \hat{f}_{h+1}) \;\big|\; (s,a) \sim \mu_h, s' \sim \mathbb{P}_h(\cdot \mid s,a)\big] \\ =& \|f_h - \mathcal{T}_h^\star \hat{f}_{h+1}\|_{\mu_h}^2 + \mathbb{E}_{\mu_h}\mathrm{Var}_{s'\sim\mathbb{P}_h(\cdot|s,a)}(V_{\hat{f}_{h+1}}(s')). \end{aligned}$$

Recall that

$$f_h^\dagger = \arg\min_{f_h\in\mathcal{F}_h} \|f_h - \mathcal{T}_h^\star \hat{f}_{h+1}\|_{\mu_h}. \tag{18}$$

$f_h^\dagger$ minimizes $\mathbb{E}_{\mu_h}\ell(f_h, \hat{f}_{h+1})$.

In the sequel, we develop upper bounds for Bellman error $\mathcal{E}(\hat{f})$ based on (local) Rademathcer complexities.

B.1. Analyzing FQI with Rademacher Complexity (Theorem 5.2)

Theorem 5.2 (FQI, Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumption 2, with probability at least $1-\delta$, the output of FQI $\hat{f}$ satisfies*

$$\mathcal{E}(\hat{f}) \le \epsilon + c\sum_{h=1}^{H}\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + cH^2\sqrt{\frac{\log(H/\delta)}{n}}. \tag{7}$$

Proof of Theorem 5.2. By Lemma G.1, with probability at least $1-\delta$, for any $f_h \in \mathcal{F}_h$,

$$\begin{aligned}\mathbb{E}_{\mu_h}\ell(f_h,\hat{f}_{h+1}) - \mathbb{E}_{\mu_h}\ell(f_h^\dagger,\hat{f}_{h+1}) \leq& \big(\hat{\ell}_h(f_h,\hat{f}_{h+1}) - \hat{\ell}_h(f_h^\dagger,\hat{f}_{h+1})\big) \\ &+ 2\mathcal{R}_n\big(\{\ell(f_h,\hat{f}_{h+1}) - \ell(f_h^\dagger,\hat{f}_{h+1}) \,|\, f_h\in\mathcal{F}_h\}\big) + 4H^2\sqrt{\frac{2\log(2/\delta)}{n}},\end{aligned} \tag{19}$$

where $f_h^\dagger$ is defined in eq. (18) and we have used $\ell(f_h,\hat{f}_{h+1}) - \ell(f_h^\dagger,\hat{f}_{h+1}) \in [-2H^2, 2H^2]$.

Specifically, we take $f_h = \hat{f}_h$ in eq. (19). Due to the optimality of $\hat{f}_h$, we have $\hat{\ell}_h(\hat{f}_h,\hat{f}_{h+1}) \leq \hat{\ell}_h(f_h^\dagger,\hat{f}_{h+1})$. We further use the relation

$$\|\hat{f}_h - \mathcal{T}_h^\star \hat{f}_{h+1}\|_{\mu_h}^2 = \big(\mathbb{E}_{\mu_h}\ell(\hat{f}_h,\hat{f}_{h+1}) - \mathbb{E}_{\mu_h}\ell(f_h^\dagger,\hat{f}_{h+1})\big) + \|f_h^\dagger - \mathcal{T}_h^\star\hat{f}_{h+1}\|_{\mu_h}^2. \tag{20}$$

and Assumption 2. It follows that

$$\big\|\hat{f}_h - \mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2 \leq 2\mathcal{R}_n\big(\{\ell(f_h,\hat{f}_{h+1}) - \ell(f_h^\dagger,\hat{f}_{h+1}) \,|\, f_h\in\mathcal{F}_h\}\big) + 4H^2\sqrt{\frac{2\log(2/\delta)}{n}} + \epsilon. \tag{21}$$

We now simplify the Rademacher complexity term in eq. (21). Due to the symmmetry of Rademacher random variables, we have $\mathcal{R}_n\big(\{\ell(f_h,\hat{f}_{h+1}) - \ell(f_h^\dagger,\hat{f}_{h+1}) \,|\, f_h\in\mathcal{F}_h\}\big) = \mathcal{R}_n\big(\{\ell(f_h,\hat{f}_{h+1}) \,|\, f_h\in\mathcal{F}_h\}\big)$. We also note that the loss function ℓ is $(4H)$-Lipschitz in its first argument. In fact, since $|f_h| \leq H$ for all $f_h \in \mathcal{F}_h$ and $r + V_{\hat{f}_{h+1}}(s') \in [-H, H]$, it holds that for any $f_h, f_h' \in \mathcal{F}_h$,

$$\begin{aligned}&\big|\ell(f_h,\hat{f}_{h+1})(s,a,r,s') - \ell(f_h',\hat{f}_{h+1})(s,a,r,s')\big| \\ =&|f_h(s,a) - f_h'(s,a)|\big|f_h(s,a) + f_h'(s,a) - 2r - 2V_{\hat{f}_{h+1}}(s')\big| \\ \leq& 4H|f_h(s,a) - f_h'(s,a)|.\end{aligned} \tag{22}$$

According to the contraction property of Rademacher complexity (see Lemma G.6), we have

$$\mathcal{R}_n\big(\{\ell(f_h,\hat{f}_{h+1}) - \ell(f_h^\dagger,\hat{f}_{h+1}) \,|\, f_h\in\mathcal{F}_h\}\big) = \mathcal{R}_n\big(\{\ell(f_h,\hat{f}_{h+1}) \,|\, f_h\in\mathcal{F}_h\}\big) \leq 2H\mathcal{R}_n^{\mu_h}(\mathcal{F}_h). \tag{23}$$

Plugging eq. (23) into eq. (21) and applying union bound, we find that with probability at least $1-\delta$,

$$\mathcal{E}(\hat{f}) = \frac{1}{H}\sum_{h=1}^H \big\|\hat{f}_h - \mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2 \leq 8\sum_{h=1}^H \mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + 4H^2\sqrt{\frac{2\log(2H/\delta)}{n}} + \epsilon,$$

which completes the proof. □

B.2. Analyzing FQI with Local Rademacher Complexity (Theorem 5.3)

Theorem 5.3 (FQI, local Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumption 2, with probability at least $1-\delta$, the output of FQI $\hat{f}$ satisfies*

$$\begin{aligned}\mathcal{E}(\hat{f}) &\leq \epsilon + c\sqrt{\epsilon\cdot\Delta} + c\Delta\,, \\ \Delta &:= H\sum_{h=1}^H r_h^\star + H^2\frac{\log(H/\delta)}{n}\,.\end{aligned} \tag{8}$$

Here $r_h^\star$ is the critical radius of local Rademacher complexity $\mathcal{R}_n^{\mu_h}(\{f_h \in \mathcal{F}_h \mid \|f_h - f_h^\dagger\|_{\mu_h}^2 \leq r\})$ with $f_h^\dagger := \arg\min_{f_h\in\mathcal{F}_h}\|f_h - \mathcal{T}_h^\star\hat{f}_{h+1}\|_{\mu_h}$.

Proof of Theorem 5.3. Recall that we have shown in eq. (22) that $\ell(f,g)$ is $(4H)$-Lipchitz in its first argument f. Under

Assumption 2, for $f_h^\dagger$ shown in eq. (18), we have

$$\begin{aligned}
&\mathrm{Var}\big[\ell(f_h,\hat{f}_{h+1})-\ell(f_h^\dagger,\hat{f}_{h+1})\big]\\
\le&\mathbb{E}\big[\big(\ell(f_h,\hat{f}_{h+1})-\ell(f_h^\dagger,\hat{f}_{h+1})\big)^2\big]\le 16H^2\mathbb{E}\Big[\big|f_h(s_h,a_h)-f_h^\dagger(s_h,a_h)\big|^2\Big]\\
=&16H^2\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le 32H^2\Big(\big\|f_h-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2+\big\|f_h^\dagger-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2\Big)\\
=&32H^2\Big[\Big(\big\|f_h-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2-\big\|f_h^\dagger-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2\Big)+2\big\|f_h^\dagger-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2\Big]\\
\le&32H^2\Big(\mathbb{E}_{\mu_h}\big[\ell(f_h,\hat{f}_{h+1})-\ell(f_h^\dagger,\hat{f}_{h+1})\big]+2\epsilon\Big).
\end{aligned}$$

When applying Theorem G.3, we are supposed to take a sub-root function larger than

$$\psi_{\mathsf{FQI}}(r):=32H^2\mathcal{R}_n\Big\{\ell(f_h,\hat{f}_{h+1})-\ell(f_h^\dagger,\hat{f}_{h+1})\,\Big|\,f_h\in\mathcal{F}_h,32H^2\Big(\mathbb{E}\big[\ell(f_h,\hat{f}_{h+1})-\ell(f_h^\dagger,\hat{f}_{h+1})\big]+2\epsilon\Big)\le r\Big\}.$$

Note that

$$\begin{aligned}
\psi_{\mathsf{FQI}}(r)\le&32H^2\mathcal{R}_n\Big(\Big\{\ell(f_h,\hat{f}_{h+1})-\ell(f_h^\dagger,\hat{f}_{h+1})\,\Big|\,f_h\in\mathcal{F}_h,16H^2\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le r\Big\}\Big)\\
\le&128H^3\mathcal{R}_n\Big(\Big\{f_h-f_h^\dagger\,\Big|\,f_h\in\mathcal{F}_h,16H^2\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le r\Big\}\Big)\\
=&128H^3\mathcal{R}_n\Big(\Big\{f_h\in\mathcal{F}_h\,\Big|\,16H^2\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le r\Big\}\Big)\le 128H^3\psi_h\Big(\frac{r}{16H^2}\Big)
\end{aligned}$$

where ψ_h is a sub-root function satisfying $\psi_h(r)\ge\mathcal{R}_n\big(\{f_h\in\mathcal{F}_h\mid\|f_h-f_h^\dagger\|_{\mu_h}^2\le r\}\big)$ and the positive fixed point $r_h^\star$ of ψ_h is the corresponding critical radius. In the second inequality, we have used the contraction property of Rademacher complexity (see Lemma G.6) and the Lipschitz continuity of ℓ. The equality in the last line is due to the symmetry of Rademacher random variables. According to Lemma G.5, the positive fixed point of $128H^3\psi_h\big(\frac{r}{16H^2}\big)$ is upper bounded by $1024H^4r_h^\star$.

We apply eq. (90) in Theorem G.3 and use the eq. (20) and $\hat{\ell}_h(\hat{f}_h,\hat{f}_{h+1})\le\hat{\ell}_h(f_h^\dagger,\hat{f}_{h+1})$. It follows that for a fixed parameter θ, with probability at least $1-\delta$,

$$\begin{aligned}
&\big\|\hat{f}_h-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2-\big\|f_h^\dagger-\mathcal{T}_h^\star\hat{f}_{h+1}\big\|_{\mu_h}^2\\
\le&cH^2r_h^\star+\frac{cH^2\log(1/\delta)}{n}+c(\theta-1)\Big(H^2r_h^\star+c\frac{H^2\log(1/\delta)}{n}\Big)+\frac{2\epsilon}{\theta-1},
\end{aligned}$$

where $c>0$ is a universal constant. By union bound and Assumption 2, we have

$$\mathcal{E}(\hat{f})\le\epsilon+cH\sum_{h=1}^H r_h^\star+cH^2\frac{\log(H/\delta)}{n}+c(\theta-1)\Big(H\sum_{h=1}^H r_h^\star+\frac{H^2\log(1/\delta)}{n}\Big)+\frac{2\epsilon}{\theta-1}.$$

We further take $\theta:=1+\frac{\sqrt{\epsilon}}{2H}\big(\frac{1}{H}\sum_{h=1}^H r_h^\star+\frac{\log(H/\delta)}{n}\big)^{-\frac{1}{2}}$ and find that

$$\mathcal{E}(\hat{f})\le\epsilon+cH\sum_{h=1}^H r_h^\star+cH^2\frac{\log(H/\delta)}{n}+c\sqrt{\epsilon\Big(H\sum_{h=1}^H r_h^\star+H^2\frac{\log(1/\delta)}{n}\Big)},$$

which completes the proof. □

C. Proof of Results for Minimax Algorithm (Theorems 5.4, 5.5 and C.1)

In this part, we prove the statistical guarantees for minimax algorithm in Section 5.3.

Notations We first introduce some notations that will be used later in the analyses. For any vector-valued function $f = (f_1, \ldots, f_H) \in L^2(\mu_1) \times \ldots \times L^2(\mu_H)$, we denote $\|f\|_\mu := \sqrt{\frac{1}{H}\sum_{h=1}^H \|f_h\|_{\mu_h}^2}$ for short. Parallel to the optimal Bellman operator $\mathcal{T}_h^\star$, we define $\mathcal{T}_h^\dagger$ and $\widehat{\mathcal{T}}_h$ as

$$\mathcal{T}_h^\dagger f_{h+1} := \arg\min_{g_h \in \mathcal{G}_h} \|g_h - \mathcal{T}_h^\star f_{h+1}\|_{\mu_h} \quad \text{and} \quad \widehat{\mathcal{T}}_h f_{h+1} := \arg\min_{g_h \in \mathcal{G}_h} \frac{1}{n} \sum_{(s,a,r,s',h)\in\mathcal{D}_h} \big(g_h(s,a) - r - V_{f_{h+1}}(s')\big)^2.$$

Let $\mathcal{T}^\star$, $\mathcal{T}^\dagger$, $\widehat{\mathcal{T}}$ be their vector form, given by

$$\begin{aligned} \mathcal{T}^\star f :=& (\mathcal{T}_1^\star f_2, \ldots, \mathcal{T}_H^\star f_{H+1}), \\ \mathcal{T}^\dagger f :=& (\mathcal{T}_1^\dagger f_2, \ldots, \mathcal{T}_H^\dagger f_{H+1}), \\ \widehat{\mathcal{T}} f :=& (\widehat{\mathcal{T}}_1 f_2, \ldots, \widehat{\mathcal{T}}_H f_{H+1}), \end{aligned} \tag{24}$$

for any $f \in \mathcal{F}$.

Similar to the definition of ℓ in eq. (17), for any $g_h \in \mathcal{G}_h \cup \mathcal{F}_h$ and $f_{h+1} \in \mathcal{F}_{h+1}$, we take

$$\ell(g_h, f_{h+1})(s,a,r,s') = \big(g_h(s,a) - r - V_{f_{h+1}}(s')\big)^2.$$

For any $f \in \mathcal{F}$, $g \in \mathcal{F} \cup \mathcal{G}$ and $\{(s_h, a_h, r_h, s_h')\}_{h=1}^H \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S})^H$, let

$$\ell(g,f)(\cdot) := \frac{1}{H}\sum_{h=1}^H \ell(g_h, f_{h+1})(s_h, a_h, r_h, s_h') = \frac{1}{H}\sum_{h=1}^H \big(g_h(s_h,a_h) - r_h - V_{f_{h+1}}(s_h')\big)^2.$$

Denote

$$\begin{aligned} \mathbb{E}_\mu \ell(g,f) := \frac{1}{H}\sum_{h=1}^H \mathbb{E}_{\mu_h} \ell(g_h, f_{h+1}) =& \frac{1}{H}\sum_{h=1}^H \mathbb{E}\big[\ell(g_h, f_{h+1})(s,a,r,s') \;\big|\; (s,a) \sim \mu_h, s' \sim \mathbb{P}_h(\cdot \mid s,a)\big] \\ =& \|g - \mathcal{T}^\star f\|_\mu^2 + \frac{1}{H}\sum_{h=1}^H \mathbb{E}_{\mu_h} \mathrm{Var}_{s'\sim\mathbb{P}_h(\cdot|s,a)}(V_{f_{h+1}}(s')) \end{aligned}$$

$$\text{and} \qquad \hat{\ell}(g,f) := \frac{1}{H}\sum_{h=1}^H \hat{\ell}_h(g_h, f_{h+1}) = \frac{1}{nH} \sum_{(s,a,r,s',h)\in\mathcal{D}} \big(g_h(s,a) - r - V_{f_{h+1}}(s')\big)^2.$$

The loss function in minimax algorithm then can be written as

$$\begin{aligned} L_{\text{MM}}(f,g) &:= \ell(f,f) - \ell(g,f), \\ \mathbb{E}_\mu L_{\text{MM}}(f,g) &:= \mathbb{E}_\mu \ell(f,f) - \mathbb{E}_\mu \ell(g,f) \\ \hat{L}_{\text{MM}}(f,g) &:= \hat{\ell}(f,f) - \hat{\ell}(g,f). \end{aligned} \tag{25}$$

Note that $\mathbb{E}_\mu L_{\text{MM}}(f,g) = \|f - \mathcal{T}^\star f\|_\mu^2 - \|g - \mathcal{T}^\star f\|_\mu^2 = \mathcal{E}(f) - \|g - \mathcal{T}^\star f\|_\mu^2$.

With our newly-defined notations, we formulate the minimax estimator as

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \hat{L}_{\text{MM}}(f,g) = \arg\min_{f\in\mathcal{F}} \hat{L}_{\text{MM}}(f, \widehat{\mathcal{T}} f). \tag{26}$$

In the analysis of minimax algorithm, we take $f^\dagger$ as the function in $\mathcal{F}$ that minimizes the Bellmen risk, *i.e.*

$$f^\dagger := \arg\min_{f\in\mathcal{F}} \mathcal{E}(f).$$

Main results

Theorem 5.4 (Minimax algorithm, Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumption 3, with probability at least $1-\delta$, the minimax estimator $\hat{f}$ satisfies:*

$$\mathcal{E}(\hat{f}) \leq \min_{f\in\mathcal{F}} \mathcal{E}(f) + \epsilon + cH^2\sqrt{\frac{\log(1/\delta)}{n}} + c\sum_{h=1}^{H}\left(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + \mathcal{R}_n^{\mu_h}(\mathcal{G}_h) + \mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})\right).$$

Theorem 5.5 (Minimax algorithm, local Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumptions 3 and 4, with probability at least $1-\delta$, the minimax estimator $\hat{f}$ satisfies:*

$$\mathcal{E}(\hat{f}) \leq \min_{f\in\mathcal{F}} \mathcal{E}(f) + \epsilon + c\sqrt{\left(\min_{f\in\mathcal{F}} \mathcal{E}(f) + \epsilon\right)\Delta} + c\Delta\,, \tag{9}$$

$$\Delta := H^3\sum_{h=1}^{H}\left[\widetilde{C}\left(r^\star_{f,h} + r^\star_{g,h} + \tilde{r}^\star_{f,h}\right) + \sqrt{\widetilde{C} r^\star_{g,h}\epsilon}\right] + H^2\frac{\log(H/\delta)}{n}\,.$$

where $\widetilde{C}$ is the concentrability coefficient in Assumption 4, and $r^\star_{f,h}, r^\star_{g,h}, \tilde{r}^\star_{f,h}$ are the critical radius of the following local Rademacher complexities respectively:

$$\mathcal{R}_n^{\mu_h}\left(\left\{f_h \in \mathcal{F}_h \;\middle|\; \|f_h - f_h^\dagger\|_{\mu_h}^2 \leq r\right\}\right),$$
$$\mathcal{R}_n^{\mu_h}\left(\left\{g_h \in \mathcal{G}_h \;\middle|\; \|g_h - g_h^\dagger\|_{\mu_h}^2 \leq r\right\}\right),$$
$$\mathcal{R}_n^{\nu_h}\left(\left\{V_{f_{h+1}} \;\middle|\; f_{h+1}\in\mathcal{F}_{h+1}, \|f_{h+1} - f_{h+1}^\dagger\|_{\nu_h\times \mathit{Unif}(\mathcal{A})}^2 \leq r\right\}\right).$$

Aside from Theorems 5.4 and 5.5, we also have an alternative statistical guarantee for $\mathcal{E}(\hat{f})$ using local Rademacher complexity for composite function $L_{\text{MM}}(f, \mathcal{T}^\dagger f)$. See Theorem C.1 below.

Theorem C.1 (Minimax algorithm, local Rademacher complexity, alternaltive)**.** *There exists an absolute constant $c > 0$, under Assumption 3, with probability at least $1-\delta$, the minimax estimator $\hat{f}$ satisfies:*

$$\mathcal{E}(\hat{f}) \leq \min_{f\in\mathcal{F}} \mathcal{E}(f) + \epsilon + c\sqrt{\left(\min_{f\in\mathcal{F}} \mathcal{E}(f) + \epsilon\right)\Delta} + c\Delta\,, \tag{27}$$

$$\Delta := H^2 r_L^\star + H\sum_{h=1}^{H} r^\star_{g,h} + H^2\frac{\log(H/\delta)}{n}\,.$$

where $r_L^\star$ and $r^\star_{g,h}$ are the critical radius of the following local Rademacher complexities respectively:

$$\mathcal{R}_n^{\mu_h}\left(\left\{L_{MM}(f,\mathcal{T}^\dagger f) \;\middle|\; f\in\mathcal{F}, \mathbb{E}\left[L_{MM}(f,\mathcal{T}^\dagger f)^2\right] \leq r\right\}\right),$$
$$\mathcal{R}_n^{\mu_h}\left(\left\{g_h \in \mathcal{G}_h \;\middle|\; \|g_h - g_h^\dagger\|_{\mu_h}^2 \leq r\right\}\right).$$

In contrast to Theorem 5.5, Theorem C.1 does not rely on the additional Assumption 4. In general, Theorem C.1 provides a tighter upper bound for $\mathcal{E}(\hat{f})$ than Theorem 5.5 when the function class $\left\{L_{\text{MM}}(f,\mathcal{T}^\dagger f) \;\middle|\; f\in\mathcal{F}\right\}$ has a clear structure and $r_L^\star$ is easy to estimate. For instance, this is the case if both f and $\mathcal{G}$ have finite elements. Based on Theorem C.1, we can recover the sharp results for finite function classes in Chen & Jiang (2019).

Assumption 3 used in our analysis of minimax algorithm can be relaxed to:

"There exist constants $\epsilon > 0$ and $\zeta \in [0,1)$ such that $\inf_{g\in\mathcal{G}} \|g - \mathcal{T}^\star f\|_\mu^2 \leq \epsilon + \zeta\mathcal{E}(f)$ for any $f\in\mathcal{F}$."

In this way, we only need a high-quality approximation of $\mathcal{T}^\star f$ in $\mathcal{G}$ when f lies within a neighborhood of the optimal Q-function. We can easily generalize our analyses to this case. However, in order to avoid unnecessary clutter, we stick to the current Assumption 3.

Proof outline Our analyses in this section are devoted to the proofs of Theorems 5.4, 5.5 and C.1.

1. We first translate the estimation of $\mathcal{E}(\hat{f})$ into deriving uniform concentration bounds for $\hat{L}_{\text{MM}}(f, \mathcal{T}^\dagger f) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)$ and $\hat{\ell}(g, f) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger)$ (Lemma C.2 in Appendix C.1). The error decomposition lemma is shared among the proofs of Theorems 5.4, 5.5 and C.1.
2. We then develop the desired uniform concentration bounds using Rademacher complexities (Appendix C.2) and local Rademacher complexities (Appendix C.3) separately. In particular, when tackling $\hat{L}_{\text{MM}}(f, \mathcal{T}^\dagger f) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)$, we have two alternative analyses involving local Rademacher complexities of different types of function classes. One leads to Theorem 5.5 and the other results in Theorem C.1.
3. In Appendix C.4, we integrate the error decomposition result and uniform concentration bounds, and finish the proofs of theorems.

C.1. Error Decomposition

We provide a decomposition of the Bellman error $\mathcal{E}(\hat{f})$ and upper bound the error using some uniform concentration inequalities.

Lemma C.2 (Error decomposition)**.** *Suppose there exist $\alpha > 0$ and $Err_f, Err_g > 0$ such that the following concentration inequalities hold simultaneously.*

1. *For any $f \in \mathcal{F}$,*
$$\mathbb{E}_\mu L_{MM}\big(f, \mathcal{T}^\dagger f\big) - \mathbb{E}_\mu L_{MM}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big) \le \alpha\big(\hat{L}_{MM}\big(f, \mathcal{T}^\dagger f\big) - \hat{L}_{MM}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big)\big) + Err_f. \tag{28}$$
2. *For any $g \in \mathcal{G}$,*
$$\mathbb{E}_\mu \ell\big(g, f^\dagger\big) - \mathbb{E}_\mu \ell\big(\mathcal{T}^\dagger f^\dagger, f^\dagger\big) \le \alpha\big(\hat{\ell}\big(g, f^\dagger\big) - \hat{\ell}\big(\mathcal{T}^\dagger f^\dagger, f^\dagger\big)\big) + Err_g. \tag{29}$$

Then under Assumption 3, the Bellman error satisfies
$$\mathcal{E}(\hat{f}) \le \min_{f \in \mathcal{F}} \mathcal{E}(f) + Err_f + Err_g + \epsilon. \tag{30}$$

Proof. By definition of function $\mathbb{E}_\mu L_{\text{MM}}(f, g)$ in eq. (25), we find that for any $f \in \mathcal{F}$,
$$\mathbb{E}_\mu L_{\text{MM}}\big(f, \mathcal{T}^\dagger f\big) = \mathbb{E}_\mu \ell(f, f) - \mathbb{E}_\mu \ell(\mathcal{T}^\dagger f, f) = \|f - \mathcal{T}^\star f\|_\mu^2 - \|\mathcal{T}^\dagger f - \mathcal{T}^\star f\|_\mu^2 = \mathcal{E}(f) - \|\mathcal{T}^\dagger f - \mathcal{T}^\star f\|_\mu^2.$$
We learn from Assumption 3 that $\|\mathcal{T}^\dagger f - \mathcal{T}^\star f\|_\mu^2 \le \epsilon$ for any $f \in \mathcal{F}$, therefore,
$$\begin{aligned}\mathbb{E}_\mu L_{\text{MM}}\big(f, \mathcal{T}^\dagger f\big) - \mathbb{E}_\mu L_{\text{MM}}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big) =& \mathcal{E}(f) - \mathcal{E}(f^\dagger) - \|\mathcal{T}^\dagger f - \mathcal{T}^\star f\|_\mu^2 + \|\mathcal{T}^\dagger f^\dagger - \mathcal{T}^\star f^\dagger\|_\mu^2 \\ \ge& \mathcal{E}(f) - \mathcal{E}(f^\dagger) - \epsilon,\end{aligned} \tag{31}$$
which implies
$$\mathcal{E}(\hat{f}) \le \min_{f \in \mathcal{F}} \mathcal{E}(f) + \Big(\mathbb{E}_\mu L_{\text{MM}}\big(\hat{f}, \mathcal{T}^\dagger \hat{f}\big) - \mathbb{E}_\mu L_{\text{MM}}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big)\Big) + \epsilon.$$
By virtue of eq. (28),
$$\mathcal{E}(\hat{f}) \le \min_{f \in \mathcal{F}} \mathcal{E}(f) + \alpha\Big(\hat{L}_{\text{MM}}\big(\hat{f}, \mathcal{T}^\dagger \hat{f}\big) - \hat{L}_{\text{MM}}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big)\Big) + Err_f + \epsilon. \tag{32}$$
In the following, we leverage eq. (29) to estimate $\hat{L}_{\text{MM}}\big(\hat{f}, \mathcal{T}^\dagger \hat{f}\big) - \hat{L}_{\text{MM}}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big)$.

We use the definition of L_{MM} and find that
$$\begin{aligned}&\hat{L}_{\text{MM}}\big(\hat{f}, \mathcal{T}^\dagger \hat{f}\big) - \hat{L}_{\text{MM}}\big(f^\dagger, \mathcal{T}^\dagger f^\dagger\big) = \hat{\ell}(\hat{f}, \hat{f}) - \hat{\ell}(\mathcal{T}^\dagger \hat{f}, \hat{f}) - \hat{\ell}(f^\dagger, f^\dagger) + \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger) \\ =&\big(\hat{L}_{\text{MM}}\big(\hat{f}, \widehat{\mathcal{T}} \hat{f}\big) + \hat{\ell}(\widehat{\mathcal{T}} \hat{f}, \hat{f})\big) - \hat{\ell}(\mathcal{T}^\dagger \hat{f}, \hat{f}) - \big(\hat{L}_{\text{MM}}\big(f^\dagger, \widehat{\mathcal{T}} f^\dagger\big) + \hat{\ell}(\widehat{\mathcal{T}} f^\dagger, f^\dagger)\big) + \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger) \\ =&\big(\hat{L}_{\text{MM}}\big(\hat{f}, \widehat{\mathcal{T}} \hat{f}\big) - \hat{L}_{\text{MM}}\big(f^\dagger, \widehat{\mathcal{T}} f^\dagger\big)\big) + \big(\hat{\ell}(\widehat{\mathcal{T}} \hat{f}, \hat{f}) - \hat{\ell}(\mathcal{T}^\dagger \hat{f}, \hat{f})\big) - \big(\hat{\ell}(\widehat{\mathcal{T}} f^\dagger, f^\dagger) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger)\big).\end{aligned} \tag{33}$$

Since $(f, g) := (\hat{f}, \widehat{\mathcal{T}}\hat{f})$ solves the minimax optimizaiton problem eq. (26), we have $\hat{L}_{\text{MM}}(\hat{f}, \widehat{\mathcal{T}}\hat{f}) \le \hat{L}_{\text{MM}}(f^\dagger, \widehat{\mathcal{T}}f^\dagger)$. Due to the optimality of $\widehat{\mathcal{T}}$, it also holds that $\hat{\ell}(\widehat{\mathcal{T}}\hat{f}, \hat{f}) \le \hat{\ell}(\mathcal{T}^\dagger\hat{f}, \hat{f})$. To this end, eq. (33) reduces to

$$\hat{L}_{\text{MM}}(\hat{f}, \mathcal{T}^\dagger\hat{f}) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \le -\big(\hat{\ell}(\widehat{\mathcal{T}}f^\dagger, f^\dagger) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger)\big). \tag{34}$$

Additionally, eq. (29) implies

$$\hat{\ell}(\widehat{\mathcal{T}}f^\dagger, f^\dagger) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger) \ge \alpha^{-1}\Big(\mathbb{E}_\mu \ell(\widehat{\mathcal{T}}f^\dagger, f^\dagger) - \mathbb{E}_\mu \ell(\mathcal{T}^\dagger f^\dagger, f^\dagger)\Big) - \alpha^{-1}Err_g.$$

Note that $\mathbb{E}_\mu \ell(\widehat{\mathcal{T}}f^\dagger, f^\dagger) - \mathbb{E}_\mu \ell(\mathcal{T}^\dagger f^\dagger, f^\dagger) = \|\widehat{\mathcal{T}}f^\dagger - \mathcal{T}^\star f^\dagger\|_\mu^2 - \|\mathcal{T}^\dagger f^\dagger - \mathcal{T}^\star f^\dagger\|_\mu^2$ and $\|\widehat{\mathcal{T}}f^\dagger - \mathcal{T}^\star f^\dagger\|_\mu \ge \|\mathcal{T}^\dagger f^\dagger - \mathcal{T}^\star f^\dagger\|_\mu$ by definition of $\mathcal{T}^\dagger$, therefore,

$$\hat{\ell}(\widehat{\mathcal{T}}f^\dagger, f^\dagger) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger) \ge -\alpha^{-1}Err_g.$$

It then follows from eq. (34) that

$$\hat{L}_{\text{MM}}(\hat{f}, \mathcal{T}^\dagger\hat{f}) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \le \alpha^{-1}Err_g. \tag{35}$$

Combining eq. (32) and eq. (35), we obtain eq. (30). □

C.2. Analyzing Minimax Algorithm with Rademacher Complexity

In what follows, we develop uniform concentration inequalities eqs. (28) and (29) using Rademacher complexities.

Lemma C.3. *With probability at least* $1-\delta$,

$$\mathbb{E}_\mu L_{MM}(f, \mathcal{T}^\dagger f) - \mathbb{E}_\mu L_{MM}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \le \big(\hat{L}_{MM}(f, \mathcal{T}^\dagger f) - \hat{L}_{MM}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\big) + Err_f \qquad \textit{for any } f \in \mathcal{F},$$

where

$$Err_f := c\sum_{h=1}^{H}\big(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + \mathcal{R}_n^{\mu_h}(\mathcal{G}_h) + \mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})\big) + 4H^2\sqrt{\frac{2\log(2/\delta)}{n}}$$

for some universal constant $c > 0$.

Proof. Note that $\big|L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\big| \le 8H^2$. We apply Lemma G.1 and find that

$$\begin{aligned}
&\mathbb{E}_\mu L_{\text{MM}}(f, \mathcal{T}^\dagger f) - \mathbb{E}_\mu L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\\
\le&\big(\hat{L}_{\text{MM}}(f, \mathcal{T}^\dagger f) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\big)\\
&+ 2\mathcal{R}_n\big(\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \mid f \in \mathcal{F}\}\big) + 16H^2\sqrt{\frac{2\log(2/\delta)}{n}}.
\end{aligned}$$

Due to the symmetry of Rademacher random variables, we have

$$\mathcal{R}_n\big(\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \mid f \in \mathcal{F}\}\big) = \mathcal{R}_n\big(\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) \mid f \in \mathcal{F}\}\big).$$

We now use Lemma G.7 to simplify the term $\mathcal{R}_n\big(\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) \mid f \in \mathcal{F}\}\big)$.

Note that

$$\begin{aligned}
L_{\text{MM}}(f, \mathcal{T}^\dagger f) =& \frac{1}{H}\sum_{h=1}^{H}\boldsymbol{\phi}_h(f)^\top \boldsymbol{A}\boldsymbol{\phi}_h(f), \quad \text{where } \boldsymbol{A} := \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix},\\
&\boldsymbol{\phi}_h(f) := \big(f_h(s_h, a_h) - \mathcal{T}_h^\dagger f_{h+1}(s_h, a_h), r_h + V_{f_{h+1}}(s_h') - \mathcal{T}_h^\dagger f_{h+1}(s_h, a_h)\big)^\top.
\end{aligned}$$

Since $\|\boldsymbol{\phi}_h(f)\|_2 \le \sqrt{2}H$ and $\|\boldsymbol{A}\|_2 = \frac{\sqrt{5}+1}{2}$, we learn that $\boldsymbol{\phi}_h(f)^\top \boldsymbol{A}\boldsymbol{\phi}_h(f)$ is $\big(\frac{\sqrt{5}+1}{\sqrt{2}}H\big)$-Lipschitz with respect to $\boldsymbol{\phi}_h(f)$. According to Lemma G.7,

$$\begin{aligned}
&\mathcal{R}_n\big(\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) \mid f \in \mathcal{F}\}\big) = \frac{1}{H}\sum_{h=1}^{H}\mathcal{R}_n\big(\{\boldsymbol{\phi}_h(f)^\top \boldsymbol{A}\boldsymbol{\phi}_h(f) \mid f \in \mathcal{F}\}\big)\\
\le&(\sqrt{5}+1)\sum_{h=1}^{H}\Big(\mathcal{R}_n\big(\{\phi_{h,1}(f) \mid f \in \mathcal{F}\}\big) + \mathcal{R}_n\big(\{\phi_{h,2}(f) \mid f \in \mathcal{F}\}\big)\Big).
\end{aligned}$$

Here,

$$\begin{aligned}\mathcal{R}_n\big(\{\phi_{h,1}(f)\,|\,f\in\mathcal{F}\}\big)=&\mathcal{R}_n\big(\{f_h-\mathcal{T}_h^\dagger f_{h+1}\,|\,f_h\in\mathcal{F}_h, f_{h+1}\in\mathcal{F}_{h+1}\}\big)\\ \le&\mathcal{R}_n\big(\{f_h-g_h\,|\,f_h\in\mathcal{F}_h, g_h\in\mathcal{G}_h\}\big)\le\mathcal{R}_n^{\mu_h}(\mathcal{F}_h)+\mathcal{R}_n^{\mu_h}(\mathcal{G}_h),\\ \mathcal{R}_n\big(\{\phi_{h,2}(f)\,|\,f\in\mathcal{F}\}\big)=&\mathcal{R}_n\big(\{r_h+V_{f_{h+1}}-\mathcal{T}_h^\dagger f_{h+1}\,|\,f_{h+1}\in\mathcal{F}_{h+1}\}\big)\\ \le&\mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})+\mathcal{R}_n\big(\{\mathcal{T}_h^\dagger f_{h+1}\,|\,f_{h+1}\in\mathcal{F}_{h+1}\}\big)\le\mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})+\mathcal{R}_n^{\mu_h}(\mathcal{G}_h).\end{aligned}$$

Integrating the pieces, we finish the proof of Lemma C.3. □

Lemma C.4. *With probability at least $1-\delta$, for any $g\in\mathcal{G}$,*

$$\mathbb{E}_\mu\ell(g,f^\dagger)-\mathbb{E}_\mu\ell(\mathcal{T}^\dagger f^\dagger,f^\dagger)\le\big(\hat{\ell}(g,f^\dagger)-\hat{\ell}(\mathcal{T}^\dagger f^\dagger,f^\dagger)\big)+Err_g,$$

where

$$Err_g:=8\sum_{h=1}^H\mathcal{R}_n^{\mu_h}(\mathcal{G}_h)+4H^2\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Proof. Note that $\big|\ell(g,f^\dagger)-\ell\big(\mathcal{T}^\dagger f^\dagger,f^\dagger\big)\big|\le 2H^2$. By Lemma G.1, with probability at least $1-\delta$, for any $g\in\mathcal{G}$,

$$\begin{aligned}\mathbb{E}_\mu\ell(g,f^\dagger)-\mathbb{E}_\mu\ell(\mathcal{T}^\dagger f^\dagger,f^\dagger)\le&\big(\hat{\ell}(g,f^\dagger)-\hat{\ell}(\mathcal{T}^\dagger f^\dagger,f^\dagger)\big)\\ &+2\mathcal{R}_n\big(\{\ell(g,f^\dagger)-\ell(\mathcal{T}^\dagger f^\dagger,f^\dagger)\;|\;g\in\mathcal{G}\}\big)+4H^2\sqrt{\frac{2\log(2/\delta)}{n}}.\end{aligned}\tag{36}$$

We observe that

$$\mathcal{R}_n\big(\{\ell(g,f^\dagger)-\ell(\mathcal{T}^\dagger f^\dagger,f^\dagger)\,|\,g\in\mathcal{G}\}\big)=\mathcal{R}_n\big(\{\ell(g,f^\dagger)\,|\,g\in\mathcal{G}\}\big)\le\frac{1}{H}\sum_{h=1}^H\mathcal{R}_n\big(\{\ell(g_h,f_{h+1}^\dagger)\,|\,g_h\in\mathcal{G}_h\}\big).\tag{37}$$

Similar to eq. (22), we can show that $\ell(g_h,f_{h+1}^\dagger)$ is $(4H)$-Lipschitz with respect to g_h, therefore,

$$\mathcal{R}_n\big(\{\ell(g_h,f_{h+1}^\dagger)\,|\,g_h\in\mathcal{G}_h\}\big)\le 4H\mathcal{R}_n^{\mu_h}(\mathcal{G}_h).\tag{38}$$

Combining eq. (36) - eq. (38), we complete the proof. □

C.3. Analyzing Minimax Algorithm with Local Rademacher Complexity

In this part, Lemmas C.5 and C.6 are devoted to the uniform concentration of $\hat{L}_{\text{MM}}(f,\mathcal{T}^\dagger f)-\hat{L}_{\text{MM}}(f^\dagger,\mathcal{T}^\dagger f^\dagger)$ and Lemma C.7 is concerned with $\hat{\ell}(g,f^\dagger)-\hat{\ell}(\mathcal{T}^\dagger f^\dagger,f^\dagger)$. The proof of Theorem C.1 uses Lemmas C.5 and C.7, while Theorem 5.5 uses Lemmas C.6 and C.7.

Concentration inequality eq. (28), $\hat{L}_{\text{MM}}(f,\mathcal{T}^\dagger f)-\hat{L}_{\text{MM}}(f^\dagger,\mathcal{T}^\dagger f^\dagger)$ Lemma C.5 below will be used as a buiding block of the proof of Theorem C.1.

Lemma C.5. *There exists a universal constant $c>0$ such that under Assumption 3, for any fixed parameter $\theta>1$, with probability at least $1-\delta$, we have*

$$\mathbb{E}_\mu L_{MM}(f,\mathcal{T}^\dagger f)-\mathbb{E}_\mu L_{MM}(f^\dagger,\mathcal{T}^\dagger f^\dagger)\le\frac{\theta}{\theta-1}\big(\hat{L}_{MM}(f,\mathcal{T}^\dagger f)-\hat{L}_{MM}(f^\dagger,\mathcal{T}^\dagger f^\dagger)\big)+Err_f\tag{39}$$

for any $f\in\mathcal{F}$, with

$$Err_f:=c\theta H^2r_L^\star+c\theta H^2\frac{\log(1/\delta)}{n}+\frac{c}{\theta-1}\big(\mathcal{E}(f^\dagger)+\epsilon\big).$$

Proof. We consider using Theorem G.3 to analyze the concentration of $\hat{L}_{\text{MM}}(f, \mathcal{T}^\dagger f) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)$. Similar to eq. (22), we can show that for any $f \in \mathcal{F}$,

$$\left|L_{\text{MM}}(f, \mathcal{T}^\dagger f)\right| \le 2\sum_{h=1}^{H} \left|f_h(s_h, a_h) - \mathcal{T}_h^\dagger f_{h+1}(s_h, a_h)\right|.$$

By Cauchy-Schwarz inequality,

$$\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f)^2\right] \le 4H^2 \|f - \mathcal{T}^\dagger f\|_\mu^2 \le 8H^2\left(\|f - \mathcal{T}^\star f\|_\mu^2 + \|\mathcal{T}^\dagger f - \mathcal{T}^\star f\|_\mu^2\right) \le 8H^2\big(\mathcal{E}(f) + \epsilon\big), \tag{40}$$

where we have used Assumption 3. It follows that

$$\begin{aligned}
&\text{Var}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\right] \le \mathbb{E}\left[(L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger))^2\right] \\
\le& 2\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f)^2\right] + 2\mathbb{E}\left[L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)^2\right] \le 16H^2\big(\mathcal{E}(f) + \mathcal{E}(f^\dagger) + 2\epsilon\big).
\end{aligned}$$

We also learn from eq. (31) that

$$\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\right] \ge \mathcal{E}(f) - \mathcal{E}(f^\dagger) - \epsilon. \tag{41}$$

We combine eq. (40) and eq. (41) and find that

$$\text{Var}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\right] \le 16H^2\big(\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}[f^\dagger, \mathcal{T}^\dagger f^\dagger]\right] + 2\mathcal{E}(f^\dagger) + 3\epsilon\big).$$

We now apply Theorem G.3 and aim to find a sub-root function ψ_L such that $\psi_L(r) \ge \widetilde{\psi}(r)$ for

$$\begin{aligned}
\widetilde{\psi}(r) :=& 16H^2\mathcal{R}_n\Big(\Big\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \;\Big|\; f \in \mathcal{F}, \\
&\qquad 16H^2\big(\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\right] + 2\mathcal{E}(f^\dagger) + 3\epsilon\big) \le r\Big\}\Big) \\
=& 16H^2\mathcal{R}_n\Big(\Big\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) \;\Big|\; f \in \mathcal{F}, \\
&\qquad 16H^2\big(\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\right] + 2\mathcal{E}(f^\dagger) + 3\epsilon\big) \le r\Big\}\Big).
\end{aligned} \tag{42}$$

Note that by eqs. (40) and (41), we have

$$16H^2\big(\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f) - L_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\right] + 2\mathcal{E}(f^\dagger) + 3\epsilon\big) \ge 2\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f)^2\right],$$

therefore,

$$\widetilde{\psi}(r) \le 16H^2\mathcal{R}_n\Big(\Big\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) \;\Big|\; f \in \mathcal{F}, 2\mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f)^2\right] \le r\Big\}\Big) \le 16H^2\psi_L\Big(\frac{r}{2}\Big),$$

where

$$\psi_L(r) = \mathcal{R}_n\Big(\Big\{L_{\text{MM}}(f, \mathcal{T}^\dagger f) \;\Big|\; f \in \mathcal{F}, \mathbb{E}\left[L_{\text{MM}}(f, \mathcal{T}^\dagger f)^2\right] \le r\Big\}\Big).$$

Let $r_L^\star$ be the positive fixed point of ψ_L. Lemma G.5 implies the positive fixed point of mapping $r \mapsto 16H^2\psi_L\big(r/2\big)$ is upper bounded by $128H^4 r_L^\star$. We then obtain eq. (39) by applying eq. (90) in Theorem G.3. □

While Lemma C.5 above uses the local Rademacher complexity of a composite function $L_{\text{MM}}(f, \mathcal{T}^\dagger f)$, Lemma C.6 below provides an alternative concentration inequality for $\hat{L}_{\text{MM}}(f, \mathcal{T}^\dagger f) - \hat{L}_{\text{MM}}(f^\dagger, \mathcal{T}^\dagger f^\dagger)$, which involves the complexities of $\mathcal{F}_h$, $\mathcal{G}_h$ and $V_{\mathcal{F}_{h+1}}$.

Lemma C.6. *Suppose Assumptions 3 and 4 hold. There exists a universal constant $c > 0$ such that for any fixed parameter $\theta > 1$, with probability at least $1 - \delta$,*

$$\mathbb{E}_\mu L_{MM}(f, \mathcal{T}^\dagger f) - \mathbb{E}_\mu L_{MM}(f^\dagger, \mathcal{T}^\dagger f^\dagger) \le \frac{\theta}{\theta - 1}\big(\hat{L}_{MM}(f, \mathcal{T}^\dagger f) - \hat{L}_{MM}(f^\dagger, \mathcal{T}^\dagger f^\dagger)\big) + Err_f \tag{43}$$

for any $f \in \mathcal{F}$, with

$$Err_f := c\theta\widetilde{C}H^3\sum_{h=1}^{H}\Big(r^\star_{f,h}+r^\star_{g,h}+\widetilde{r}^\star_{f,h+1}+\sqrt{\epsilon r^\star_{g,h}/\widetilde{C}}\Big)+c\theta H^2\frac{\log(1/\delta)}{n}+\frac{c}{\theta-1}\big(\mathcal{E}(f^\dagger)+\epsilon\big).$$

Here, $\widetilde{C}$ is the concentrability coefficient in Assumption 4.

Proof. In this proof, we estimate the critical radius of $\widetilde{\psi}(r)$ in eq. (42) in an alternative way. In particular, we use parameters $r^\star_{f,h}$, $r^\star_{g,h}$ and $\widetilde{r}^\star_{f,h}$ defined in the statement of Theorem 5.5. The key step is to upper bound $\widetilde{\psi}(r)$ by the local Rademacher complexities $\mathcal{R}_n^{\mu_h}\big(\{f_h\in\mathcal{F}_h \mid \|f_h-f_h^\dagger\|_{\mu_h}^2\le r\}\big)$, $\mathcal{R}_n^{\mu_h}\big(\{g_h\in\mathcal{G}_h \mid \|g_h-g_h^\dagger\|_{\mu_h}^2\le r\}\big)$ and $\mathcal{R}_n^{\nu_h}\big(\{V_{f_{h+1}} \mid f_{h+1}\in\mathcal{F}_{h+1},\|f_{h+1}-f_{h+1}^\dagger\|^2_{\nu_h\times\mathrm{Unif}(\mathcal{A})}\le r\}\big)$.

We take a shorthand $\mathcal{F}(r):=\big\{f\in\mathcal{F}\,\big|\,16H^2\big(\mathbb{E}\big[L_{\mathrm{MM}}\big(f,\mathcal{T}^\dagger f\big)-L_{\mathrm{MM}}\big(f^\dagger,\mathcal{T}^\dagger f^\dagger\big)\big]+2\mathcal{E}(f^\dagger)+3\epsilon\big)\le r\big\}$ and rewrite $\widetilde{\psi}(r)$ as $\widetilde{\psi}(r)=16H^2\mathcal{R}_n\big(\{L_{\mathrm{MM}}\big(f,\mathcal{T}^\dagger f\big)\mid f\in\mathcal{F}(r)\}\big)$. Similar to Lemma C.3, one can show that there exists a univeral constant $c>0$ such that

$$\widetilde{r}\le cH^2\sum_{h=1}^{H}\big(\psi_{h,1}(r)+\psi_{h,2}(r)+\psi_{h,3}(r)\big).$$

where $\psi_{h,1}(r):=\mathcal{R}_n^{\mu_h}\big(\{f_h\mid f\in\mathcal{F}(r)\}\big)$, $\psi_{h,2}(r):=\mathcal{R}_n^{\mu_h}\big(\{\mathcal{T}_h^\dagger f_{h+1}\mid f\in\mathcal{F}(r)\}\big)$ and $\psi_{h,3}(r):=\mathcal{R}_n^{\nu_h}\big(\{V_{f_{h+1}}\mid f\in\mathcal{F}(r)\}\big)$. In the sequel, we simplify $\psi_{h,1}$, $\psi_{h,2}$ and $\psi_{h,3}$.

For any $f\in\mathcal{F}(r)$, due to eq. (41), we have

$$\begin{aligned}\big\|(f-\mathcal{T}^\star f)-(f^\dagger-\mathcal{T}^\star f^\dagger)\big\|_\mu^2\le&2\big\|f-\mathcal{T}^\star f\big\|_\mu^2+2\big\|f^\dagger-\mathcal{T}^\star f^\dagger\big\|_\mu^2=2\mathcal{E}(f)+2\mathcal{E}(f^\dagger)\\ \le&2\mathbb{E}\big[L_{\mathrm{MM}}\big[f,\mathcal{T}^\dagger f\big]-L_{\mathrm{MM}}\big(f^\dagger,\mathcal{T}^\dagger f^\dagger\big)\big]+4\mathcal{E}(f^\dagger)+2\epsilon\le\frac{r}{8H^2}.\end{aligned}$$

We use Lemma F.1 and find that under Assumptions 3 and 4, for any $f\in\mathcal{F}$,

$$\begin{aligned}&\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le\frac{\widetilde{C}r}{8},\\ &\big\|\mathcal{T}_h^\dagger f_{h+1}-\mathcal{T}_h^\dagger f_{h+1}^\dagger\big\|_{\mu_h}^2\le\big(\big\|\mathcal{T}_h^\star f_{h+1}-\mathcal{T}_h^\star f_{h+1}^\dagger\big\|_{\mu_h}+2\sqrt{\epsilon}\big)^2\\ &\qquad\qquad\le2\big\|\mathcal{T}_h^\star f_{h+1}-\mathcal{T}_h^\star f_{h+1}^\dagger\big\|_{\mu_h}^2+8\epsilon\le\frac{\widetilde{C}r}{4}+8\epsilon,\\ &\big\|f_{h+1}-f_{h+1}^\dagger\big\|^2_{\nu_h\times\mathrm{Unif}(\mathcal{A})}\le\frac{\widetilde{C}r}{8}.\end{aligned}$$

It follows that

$$\begin{aligned}&\psi_{h,1}(r)=\mathcal{R}_n^{\mu_h}\big(\{f_h\,\big|\,f\in\mathcal{F}(r)\}\big)\le\mathcal{R}_n^{\mu_h}\Big(\Big\{f_h\in\mathcal{F}_h\,\Big|\,\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le\frac{\widetilde{C}r}{8}\Big\}\Big),\\ &\psi_{h,2}(r)=\mathcal{R}_n^{\mu_h}\big(\{\mathcal{T}_h^\dagger f_{h+1}\mid f\in\mathcal{F}(r)\}\big)\le\mathcal{R}_n^{\mu_h}\Big(\Big\{g_h\in\mathcal{G}_h\,\Big|\,\big\|g_h-\mathcal{T}_h^\dagger f_{h+1}^\dagger\big\|_{\mu_h}^2\le\frac{\widetilde{C}r}{4}+8\epsilon\Big\}\Big),\\ &\psi_{h,3}(r)=\mathcal{R}_n^{\nu_h}\big(\{V_{f_{h+1}}\mid f\in\mathcal{F}(r)\}\big)\le\mathcal{R}_n^{\nu_h}\Big(\Big\{V_{f_{h+1}}\,\Big|\,f_{h+1}\in\mathcal{F}_{h+1},\big\|f_{h+1}-f_{h+1}^\dagger\big\|^2_{\nu_h\times\mathrm{Unif}(\mathcal{A})}\le\frac{\widetilde{C}r}{8}\Big\}\Big).\end{aligned}$$

Recall that $r^\star_{f,h}$, $r^\star_{g,h}$ and $\widetilde{r}^\star_{f,h+1}$ are respectively the fixed points of

$$\begin{aligned}&\psi_{f,h}(r)=\mathcal{R}_n^{\mu_h}\big(\{f_h\in\mathcal{F}_h\mid\big\|f_h-f_h^\dagger\big\|_{\mu_h}^2\le r\}\big),\\ &\psi_{g,h}(r)=\mathcal{R}_n^{\mu_h}\big(\{g_h\in\mathcal{G}_h\mid\big\|g_h-\mathcal{T}_h^\dagger f_{h+1}^\dagger\big\|_{\mu_h}^2\le r\}\big)\qquad\text{and}\\ &\widetilde{\psi}_{f,h}(r)=\mathcal{R}_n^{\nu_h}\big(\{V_{f_{h+1}}\mid f_{h+1}\in\mathcal{F}_{h+1},\big\|f_{h+1}-f_{h+1}^\dagger\big\|_{\widetilde{\mu}_{h+1}}^2\le r\}\big).\end{aligned}$$

According to Lemma G.5, the positive fixed points of $\psi_{h,1}$, $\psi_{h,2}$ and $\psi_{h,3}$ are upper bounded by $8\widetilde{C}r^\star_{f,h}$, $4\widetilde{C}r^\star_{g,h} + \sqrt{32\epsilon\widetilde{C}r^\star_{g,h}}$ and $8\widetilde{C}\widetilde{r}_{f,h}$, therefore, the critical radius $\widetilde{r}^\star$ of $\widetilde{\psi}(r)$ satisfies

$$\begin{aligned}\widetilde{r}^\star \leq & c^2H^4\left(\sum_{h=1}^H\left(\sqrt{8\widetilde{C}r^\star_{f,h}}+\sqrt{4\widetilde{C}r^\star_{g,h}}+\sqrt[4]{32\epsilon\widetilde{C}r^\star_{g,h}}+\sqrt{8\widetilde{C}\widetilde{r}^\star_{f,h}}\right)\right)^2\\ \leq & c'\widetilde{C}H^5\sum_{h=1}^H\left(r^\star_{f,h}+r^\star_{g,h}+\widetilde{r}^\star_{f,h}+\sqrt{\epsilon r^\star_{g,h}/\widetilde{C}}\right),\end{aligned}$$

where $c, c' > 0$ are universal constants.

We then apply eq. (90) in Theorem G.3 and obtain eq. (43). □

Concentration inequality eq. (29), $\hat{\ell}(g, f^\dagger) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger)$

Lemma C.7. *Suppose Assumption 3 holds. Then there exists a universal constant $c > 0$ such that for any fixed parameter $\theta > 1$, with probability at least $1-\delta$,*

$$\mathbb{E}_\mu\ell(g,f^\dagger) - \mathbb{E}_\mu\ell(\mathcal{T}^\dagger f^\dagger, f^\dagger) \leq \frac{\theta}{\theta-1}\left(\hat{\ell}(g,f^\dagger) - \hat{\ell}(\mathcal{T}^\dagger f^\dagger, f^\dagger)\right) + Err_g, \tag{44}$$

$$\text{with} \qquad Err_g := c\theta H\sum_{h=1}^H r^\star_{g,h} + c\theta H^2\frac{\log(H/\delta)}{n} + \frac{c\epsilon}{\theta-1}.$$

Proof. Note that

$$\ell(g,f^\dagger) - \ell(\mathcal{T}^\dagger f^\dagger, f^\dagger) = \frac{1}{H}\sum_{h=1}^H\left(\ell(g_h, f^\dagger_{h+1}) - \ell(\mathcal{T}^\dagger_h f^\dagger_{h+1}, f^\dagger_{h+1})\right).$$

We can analyze the concentration of $\ell(g_h, f^\dagger_{h+1}) - \ell(\mathcal{T}^\dagger_h f^\dagger_{h+1}, f^\dagger_{h+1})$ in a way similar to Theorem 5.3. It follows that for any $h \in [H]$, with probability at least $1-\delta$,

$$\begin{aligned}&\mathbb{E}_\mu\ell(g_h, f^\dagger_{h+1}) - \mathbb{E}_\mu\ell(\mathcal{T}^\dagger_h f^\dagger_{h+1}, f^\dagger_{h+1})\\ \leq & \frac{\theta}{\theta-1}\left(\hat{\ell}(g_h, f^\dagger_{h+1}) - \hat{\ell}(\mathcal{T}^\dagger_h f^\dagger_{h+1}, f^\dagger_{h+1})\right) + 8c_1\theta H^2 r^\star_{g,h} + (2c_2+8c_3\theta)H^2\frac{\log(1/\delta)}{n} + \frac{2\epsilon}{\theta-1},\end{aligned}$$

for any $g_h \in \mathcal{G}_h$, where c_1, c_2, c_3 are the constants in Theorem G.3. By union bound, we can further derive eq. (44). □

C.4. Proof of Theorems 5.4, 5.5 and C.1

Proof of Theorem 5.4. Combining Lemmas C.2 to C.4, we obtain Theorem 5.4. □

Proof of Theorems 5.5 and C.1. Plugging Lemmas Lemmas C.5 and C.7 into Lemma C.2 yields that with probability at least $1-\delta$,

$$\mathcal{E}(\hat{f}) \leq \min_{f\in\mathcal{F}}\mathcal{E}(f) + \epsilon + c\theta H^2\left(r^\star_L + \frac{1}{H}\sum_{h=1}^H r^\star_{g,h} + \frac{\log(H/\delta)}{n}\right) + \frac{c}{\theta-1}\left(\mathcal{E}(f^\dagger)+\epsilon\right)$$

for a universal constant $c > 0$. By letting

$$\theta := 1 + \sqrt{\frac{\mathcal{E}(f^\dagger)+\epsilon}{cH^2\left(r^\star_L + \frac{1}{H}\sum_{h=1}^H r^\star_{g_h} + \frac{\log(H/\delta)}{n}\right)}},$$

we have

$$\begin{aligned}\mathcal{E}(\hat{f}) \leq & \min_{f\in\mathcal{F}}\mathcal{E}(f) + \epsilon + cH^2\left(r^\star_L + \frac{1}{H}\sum_{h=1}^H r^\star_{g_h} + \frac{\log(H/\delta)}{n}\right)\\ & + cH\sqrt{\left(\min_{f\in\mathcal{F}}\mathcal{E}(f^\dagger)+\epsilon\right)\left(r^\star_L + \frac{1}{H}\sum_{h=1}^H r^\star_{g_h} + \frac{\log(H/\delta)}{n}\right)},\end{aligned}$$

which finishes the proof of Theorem C.1.

Similarly, by combining Lemmas C.2, C.6 and C.7, we prove Theorem 5.5. □

D. Examples (Propositions 6.1 to 6.4)

In this part, we provide estimates for the (local) Rademacher complexities of four special function spaces, namely function class with finite elements, linear function space, kernel class and sparse linear space. The results presented here slightly generalize Propositions 6.1 to 6.4.

D.1. Function class with finite elements (Proposition 6.1)

Lemma D.1 (Full version of Proposition 6.1)**.** *Suppose $\mathcal{F}$ is a discrete function class with $|\mathcal{F}| < \infty$ and $f \in [0, D]$ for any $f \in \mathcal{F}$. Then for any distribution ρ,*

$$\mathcal{R}_n^\rho(\mathcal{F}) \le 2D \max\left\{\sqrt{\frac{\log|\mathcal{F}|}{n}}, \frac{\log|\mathcal{F}|}{n}\right\}. \tag{45}$$

For any function f° with range in $[0, D]$, we have

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \le \psi(r), \quad \text{where } \psi(r) := 2\max\left\{\sqrt{\frac{r\log|\mathcal{F}|}{n}}, \frac{D\log|\mathcal{F}|}{n}\right\}. \tag{46}$$

ψ is a sub-root function with positive fixed point

$$r^\star = \frac{2(D \vee 2)\log|\mathcal{F}|}{n}.$$

We remark that Proposition 6.1 is a corollary of Lemma D.1 with $D := H$.

In order to prove Lemma D.1, we first present a preliminary lemma that will be used later. See Lemma D.2.

Lemma D.2. *Suppose a random variable X satisfies $|X| \le D$ and $\mathbb{E}[X] = 0$. Then for any $\lambda > 0$, we have*

$$\mathbb{E}[e^{\lambda X}] \le \exp\left\{\lambda^2 \mathrm{Var}[X]\left(\frac{e^{\lambda D} - 1 - \lambda D}{\lambda^2 D^2}\right)\right\}. \tag{47}$$

Proof. Note that $X \le D$ and the mapping $x \mapsto \frac{e^x - 1 - x}{x^2}$ is nondecreasing, therefore, $\frac{e^{\lambda X} - 1 - \lambda X}{\lambda^2 X^2} \le \frac{e^{\lambda D} - 1 - \lambda D}{\lambda^2 D^2}$. It follows that

$$\mathbb{E}[e^{\lambda X}] = 1 + \lambda\mathbb{E}[X] + \lambda^2\mathbb{E}\left[X^2\left(\frac{e^{\lambda X} - 1 - \lambda X}{\lambda^2 X^2}\right)\right] \le 1 + \lambda^2\mathrm{Var}[X]\left(\frac{e^{\lambda D} - 1 - \lambda D}{\lambda^2 D^2}\right), \tag{48}$$

where we have used the fact $\mathbb{E}[X] = 0$. Since $1 + x \le e^x$ for any $x \in \mathbb{R}$, eq. (48) implies eq. (47). □

We are now ready to prove Lemma D.1.

Proof of Lemma D.1. We can easily see that eq. (45) is a corollary of eq. (46) by letting $f^\circ = 0$ and $r = D^2$, therefore, we focus on proving eq. (46). By definition of Rademacher complexity and the symmetry of Rademacher variables, we have

$$\begin{aligned}\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) =& \mathcal{R}_n^\rho\big(\{f - f^\circ \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \\ =& \mathbb{E}\max\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i\big(f(X_i) - f^\circ(X_i)\big) \,\middle|\, f \in \mathcal{F}, \|f - f^\circ\|_\rho^2 \le r\right\}.\end{aligned}$$

For any $\lambda > 0$, it holds that

$$\begin{aligned}\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) =& \frac{1}{\lambda n}\mathbb{E}\log \max_{\substack{f\in\mathcal{F}:\\ \|f-f^\circ\|_\rho^2\le r}} \exp\left\{\lambda\sum_{i=1}^n \sigma_i\big(f(X_i)-f^\circ(X_i)\big)\right\}\\ \le& \frac{1}{\lambda n}\mathbb{E}\log \sum_{\substack{f\in\mathcal{F}:\\ \|f-f^\circ\|_\rho^2\le r}} \exp\left\{\lambda\sum_{i=1}^n \sigma_i\big(f(X_i)-f^\circ(X_i)\big)\right\}\\ \le& \frac{1}{\lambda n}\log \sum_{\substack{f\in\mathcal{F}:\\ \|f-f^\circ\|_\rho^2\le r}} \mathbb{E}\exp\left\{\lambda\sum_{i=1}^n \sigma_i\big(f(X_i)-f^\circ(X_i)\big)\right\},\end{aligned} \tag{49}$$

where the last line is due to Jensen's inequality. Since $(\sigma_1, X_1), \ldots, (\sigma_n, X_n)$ are *i.i.d.* samples,

$$\mathbb{E}\exp\left\{\lambda\sum_{i=1}^n \sigma_i\big(f(X_i)-f^\circ(X_i)\big)\right\} = \Big(\mathbb{E}\exp\big\{\lambda\sigma_1\big(f(X_1)-f^\circ(X_1)\big)\big\}\Big)^n. \tag{50}$$

Note that $\big|\sigma_1\big(f(X_1)-f^\circ(X_1)\big)\big| \le D$ and $\mathbb{E}\big[\sigma_1\big(f(X_1)-f^\circ(X_1)\big)\big] = 0$ since $\mathbb{E}[\sigma_1] = 0$. For any $f \in \mathcal{F}$ such that $\|f - f^\circ\|_\rho^2 \le r$, we have $\mathrm{Var}\big[\sigma_1\big(f(X_1)-f^\circ(X_1)\big)\big] = \mathbb{E}\big[\big(f(X_1)-f^\circ(X_1)\big)^2\big] = \|f-f^\circ\|_\rho^2 \le r$. We apply Lemma D.2 and derive that

$$\mathbb{E}\exp\big\{\lambda\sigma_1\big(f(X_1)-f^\circ(X_1)\big)\big\} \le \exp\left\{\lambda^2 r\left(\frac{e^{\lambda D}-1-\lambda D}{\lambda^2 D^2}\right)\right\}. \tag{51}$$

Combining eqs. (49) to (51), we obtain

$$\begin{aligned}\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \le& \frac{1}{\lambda n}\log \sum_{\substack{f\in\mathcal{F}:\\ \|f-f^\circ\|_\rho^2\le r}} \Big(\mathbb{E}\exp\big\{\lambda\sigma_1\big(f(X_1)-f^\circ(X_1)\big)\big\}\Big)^n\\ \le& \frac{1}{\lambda n}\log\left(|\mathcal{F}|\exp\left\{n\lambda^2 r\left(\frac{e^{\lambda D}-1-\lambda D}{\lambda^2 D^2}\right)\right\}\right)\\ =& \frac{\log|\mathcal{F}|}{\lambda n} + \lambda r\left(\frac{e^{\lambda D}-1-\lambda D}{\lambda^2 D^2}\right).\end{aligned} \tag{52}$$

For $r \ge \frac{D^2\log|\mathcal{F}|}{n}$, by letting $\lambda := \sqrt{\frac{\log|\mathcal{F}|}{rn}}$, eq. (52) implies $\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \le 2\sqrt{\frac{r\log|\mathcal{F}|}{n}}$, where we have used the fact $\frac{e^x-1-x}{x^2} \le 1$ for any $x \le 1$. When $0 \le r < \frac{D^2\log|\mathcal{F}|}{n}$, by letting $\lambda := \frac{1}{D}$, eq. (52) ensures $\mathcal{R}_n\big(\{f \in \mathcal{F} \mid P(f-f^\circ)^2 \le r\}\big) \le \frac{2D\log|\mathcal{F}|}{n}$. Integrating the pieces, we complete the proof of eq. (46).

It is easy to see that the right hand side of eq. (46) is a sub-root function with positive fixed point $\frac{2(D\vee 2)\log|\mathcal{F}|}{n}$. □

D.2. Linear Space (Proposition 6.2)

Lemma D.3 (Full version of Proposition 6.2)**.** *Let $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ be a feature map to d-dimensional Euclidean space and ρ be a distribution over $\mathcal{S}\times\mathcal{A}$. Consider a function class*

$$\mathcal{F} = \big\{f = w^\top\phi \mid w \in \mathbb{R}^d, \|f\|_\rho^2 \le B\big\},$$

where $B > 0$. It holds that

$$\mathcal{R}_n^\rho(\mathcal{F}) \le \sqrt{\frac{2Bd}{n}}.$$

For any $f^\circ \in \mathcal{F}$, we have

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \le \psi(r), \quad \textit{where } \psi(r) := \sqrt{\frac{2rd}{n}}.$$

ψ is sub-root and has a positive fixed point

$$r^\star = \frac{2d}{n}.$$

Proposition 6.2 in Section 6 is a corollary to Lemma D.3. In Proposition 6.2, conditions $\|w\| \le H$ and $\|\phi(s,a)\| \le 1$ ensure $\|f\|_\infty \le H$ for $f = w^\top \phi$ and therefore $\|f\|_\rho^2 \le H^2$. By letting $B := H^2$ in Lemma D.3, we obtain Proposition 6.2.

Proof of Lemma D.3. Lemma D.3 can be viewed as a consequence of Lemma D.4 in Appendix D.3. Without loss of generality, suppose that ϕ is orthonormal in $L^2(\rho)$, that is, $\int_{\mathcal{S}\times\mathcal{A}} \phi_i(s,a)\phi_j(s,a)\rho(s,a)\mathrm{d}s\mathrm{d}a = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$ Define a kernel function $k\big((s,a),(s',a')\big) = \phi(s,a)^\top \phi(s',a')$. The RKHS associated with kernel k is the linear space spanned by ϕ endorsed with inner product $\langle f, f'\rangle_{\mathcal{K}} := w^\top w'$ for $f = \phi^\top w$, $f' = \phi^\top w'$. In this way, we have $\|\cdot\|_\rho = \|\cdot\|_{\mathcal{K}}$. For any $f \in \mathcal{F}$, $\|f\|_\rho^2 \le B$ implies $\|f\|_{\mathcal{K}} \le \sqrt{B}$. We apply the results in Lemma D.4 with $D = \sqrt{B}$. It follows that $\mathcal{R}_n^\rho(\mathcal{F}) \le \sqrt{\frac{2B}{n}\sum_{i=1}^\infty 1 \wedge (4\lambda_i)} \le \sqrt{\frac{2Bd}{n}}$ and $\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \le \sqrt{\frac{2}{n}\sum_{i=1}^\infty r \wedge \big(4B\lambda_k\big)} \le \sqrt{\frac{2rd}{n}}$ since $\lambda_i = 0$ for $i > d$. □

D.3. Kernel Class (Proposition 6.3)

We now consider kernel class, that is, a sphere in an RKHS $\mathcal{H}$ associated with a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In our paper, $\mathcal{X} = \mathcal{S} \times \mathcal{A}$. Let ρ be a distribution over $\mathcal{X}$. We are interested in Rademacher complexities of function class

$$\mathcal{F} = \big\{f \in \mathcal{H} \bigm| \|f\|_{\mathcal{K}} \le D, \|f\|_\rho^2 \le B\big\}. \tag{53}$$

Here, $\|\cdot\|_{\mathcal{K}}$ denotes the RKHS norm and $D, B \ge 0$ are some constants. Suppose that $\mathbb{E}_\rho k(X,X) < \infty$ for $X \sim \rho$. We define an integral operator $\mathscr{T} : L^2(\rho) \to L^2(\rho)$ as

$$\mathscr{T} f = \int k(\cdot, y) f(y) \rho(y) \mathrm{d}y.$$

It is easy to see that $\mathscr{T}$ is positive semidefinite and trace-class. Let $\{\lambda_i\}_{i=1}^\infty$ be the eigenvalues of $\mathscr{T}$, arranging in a nonincreasing order. By using these eigenvalues, we have an estimate for (local) Rademacher complexities of $\mathcal{F}$ in Lemma D.4 below.

Lemma D.4 (Full version of Proposition 6.3)**.** *For function class $\mathcal{F}$ defined in eq. (53), we have*

$$\mathcal{R}_n^\rho(\mathcal{F}) \le \sqrt{\frac{2}{n}\sum_{i=1}^\infty B \wedge (4D^2\lambda_i)}. \tag{54}$$

Let f° be an arbitrary function in $\mathcal{F}$. The local Rademacher complexity around f° satisfies

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \bigm| \|f - f^\circ\|_\rho^2 \le r\}\big) \le \psi(r), \quad \textit{where } \psi(r) := \sqrt{\frac{2}{n}\sum_{i=1}^\infty r \wedge \big(4D^2\lambda_i\big)}. \tag{55}$$

ψ is a sub-root function with positive fixed point

$$r^\star \le 2\min_{j\in\mathbb{N}} \left\{ \frac{j}{n} + D\sqrt{\frac{2}{n}\sum_{i=j+1}^\infty \lambda_i} \right\}. \tag{56}$$

In Proposition 6.3, we assume that $k(x,x) \le 1$ for any $x \in \mathcal{X}$ and $\|f\|_{\mathcal{K}} \le H$ for any $f \in \mathcal{F}$. It is then guaranteed that $|f(x)| = \big|\langle f, k(\cdot,x)\rangle_{\mathcal{K}}\big| \le \|f\|_{\mathcal{K}}\big\|k(\cdot,x)\big\|_{\mathcal{K}} = \|f\|_{\mathcal{K}}\sqrt{k(x,x)} \le H$, which further implies $\|f\|_\rho^2 \le H^2$. To this end, Proposition 6.3 is a consequence of lemma D.4 by taking $D := H$ and $B := H^2$.

We remark on the rate of $r^\star$ with respect to sample size n. Firstly, it is evident that $r^\star \lesssim n^{-\frac{1}{2}}$. When $\lambda_i \lesssim i^{-\alpha}$ for $\alpha > 1$, $r_h^\star$ has order $n^{-\frac{\alpha}{1+\alpha}}$ which is typical in nonparametric estimation. When the eigenvalues $\{\lambda_i\}_{i=1}^\infty$ decay exponentially quickly, *i.e.* $\lambda_i \lesssim \exp(-\beta i^\alpha)$ for $\alpha, \beta > 0$, $r^\star$ can be of order $n^{-1}(\log n)^{1/\alpha}$.

Our proof of Lemma D.4 is based on a classical result shown in Theorem D.5.

Theorem D.5 (Theorem 41 in Mendelson (2002)). *For every $r > 0$, we have*

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{H} \;\big|\; \|f\|_\mathcal{K} \le 1, \|f\|_\rho^2 \le r\}\big) \le \sqrt{\frac{2}{n}\sum_{i=1}^\infty r \wedge \lambda_i}.$$

Now we are ready to prove Lemma D.4.

Proof of Lemma D.4. Since eq. (54) is a corollary of eq. (55) by setting $r = B$, we only consider eqs. (55) and (56).

Due to the symmetry of Rademacher random variables,

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \;\big|\; \|f - f^\circ\|_\rho^2 \le r\}\big) = \mathcal{R}_n^\rho\big(\{f - f^\circ \;\big|\; f \in \mathcal{F}, \|f - f^\circ\|_\rho^2 \le r\}\big). \tag{57}$$

Since $\|f\|_\mathcal{K} \le D$ implies $\|f - f^\circ\|_\mathcal{K} \le 2D$, we have $\mathcal{F} \subseteq \big\{f \in \mathcal{H} \;\big|\; \|f - f^\circ\|_\mathcal{K} \le 2D\big\}$. It follows that

$$\begin{aligned}
\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \;\big|\; \|f - f^\circ\|_\rho^2 \le r\}\big) \le& \mathcal{R}_n^\rho\big(\{f - f^\circ \;\big|\; f \in \mathcal{H}, \|f - f^\circ\|_\mathcal{K} \le 2D, \|f - f^\circ\|_\rho^2 \le r\}\big) \\
=& \mathcal{R}_n^\rho\big(\{f \in \mathcal{H} \;\big|\; \|f\|_\mathcal{K} \le 2D, \|f\|_\rho^2 \le r\}\big) \\
\overset{f_h' := f_h/(2D)}{=}& 2D \cdot \mathcal{R}_n^\rho\left(\left\{f' \in \mathcal{H} \;\middle|\; \|f'\|_\mathcal{K} \le 1, \|f'\|_\rho^2 \le \frac{r}{4D^2}\right\}\right),
\end{aligned}$$

where we have used the translational symmetry of RKHS $\mathcal{H}$. We apply Theorem D.5 and derive that

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F} \;\big|\; \|f - f^\circ\|_\rho^2 \le r\}\big) \le 2D\sqrt{\frac{2}{n}\sum_{i=1}^\infty \frac{r}{4D^2} \wedge \lambda_i} = \sqrt{\frac{2}{n}\sum_{i=1}^\infty r \wedge \big(4D^2\lambda_i\big)} = \psi(r).$$

It is evident that ψ is sub-root. In the following, we estimate the positive fixed point $r^\star$ of ψ.

If $r \le r^\star$, then $r \le \psi(r)$, which implies

$$r^2 \le \frac{2}{n}\sum_{i=1}^\infty r \wedge \big(4D^2\lambda_i\big) \le \frac{2}{n}\left(jr + 4D^2\sum_{i=j+1}^\infty \lambda_i\right) \qquad \text{for any } j \in \mathbb{N}.$$

Solving the quadratic inequality yields

$$r \le \frac{2j}{n} + 2D\sqrt{\frac{2}{n}\sum_{i=j+1}^\infty \lambda_i} \qquad \text{for any } j \in \mathbb{N}.$$

It ensures that

$$r^\star \le 2\min_{j \in \mathbb{N}}\left\{\frac{j}{n} + D\sqrt{\frac{2}{n}\sum_{i=j+1}^\infty \lambda_i}\right\}.$$

□

D.4. Sparse Linear Class (Proposition 6.4)

Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a d-dimensional feature map and ρ be a distribution over $\mathcal{S} \times \mathcal{A}$. We are interested in function class

$$\mathcal{F}_s = \big\{f = w^\top\phi \;\big|\; w \in \mathbb{R}^d, \|w\|_0 \le s, \|f\|_\rho^2 \le B\big\}.$$

In the following, we provide an estimate for (local) Rademacher complexities of $\mathcal{F}_s$ based on the transportation T_2 inequality. Proposition 6.4 would be a special case of our result in this part since Gaussian distributions always satisfy T_2 inequality.

Notations We denote by $\alpha \subseteq [d]$ an index set with s elements. Let $\mathcal{I} := \{\alpha \subseteq [d] \mid |\alpha| = s\}$. Note that $|\mathcal{I}| \le d^s$. For any $\alpha \in \mathcal{I}$, let ϕ_α be the subvector of ϕ with $\phi_\alpha := (\phi_{\alpha_1}, \phi_{\alpha_2}, \ldots, \phi_{\alpha_s})^\top$. Denote covariance matrix $\Sigma := \mathbb{E}_\rho[\phi\phi^\top] \in \mathbb{R}^{d\times d}$. Let $\Sigma_\alpha := \mathbb{E}_\rho[\phi_\alpha\phi_\alpha^\top] \in \mathbb{R}^{s\times s}$ be the principal submatrix of Σ with indices given by α.

We use Orlicz norms $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ in the spaces of random variables. For a real-valued random variable X, define $\|X\|_{\psi_1} := \inf\{c > 0 \mid \mathbb{E}[\exp(|X|/c) - 1] \le 1\}$ and $\|X\|_{\psi_2} := \inf\{c > 0 \mid \mathbb{E}[\exp(X^2/c^2) - 1] \le 1\}$. For a random vector $X \in \mathbb{R}^d$, define $\|X\|_{\psi_1} := \sup_{u\in\mathbb{S}^{d-1}} \|u^\top X\|_{\psi_1}$ and $\|X\|_{\psi_2} := \sup_{u\in\mathbb{S}^{d-1}} \|u^\top X\|_{\psi_2}$.

For any positive semidefinite (PSD) matrix $M \in \mathbb{R}^{d\times d}$, let $M^\dagger$ denote its Moore–Penrose inverse and $\sqrt{M^\dagger} \in \mathbb{R}^{d\times d}$ be the unique PSD matrix such that $\left(\sqrt{M^\dagger}\right)^2 = M^\dagger$. We define a $M^\dagger$-weighted vector norm $\|\cdot\|_{M^\dagger}$ as $\|\mathbf{x}\|_{M^\dagger} = \sqrt{\mathbf{x}^\top M^\dagger \mathbf{x}} := \left\|\sqrt{M^\dagger}\mathbf{x}\right\|_2$ for any $\mathbf{x} \in \mathbb{R}^d$.

For any two distributions μ and ν on a same metric space $(\mathbb{X}, d)$, we say a measure $p(X, Y)$ over $\mathbb{X} \times \mathbb{X}$ is a coupling of μ and ν if the marginal distributions of p are μ and ν respectively, *i.e.* $p(\cdot, \mathbb{X}) = \mu$ and $p(\mathbb{X}, \cdot) = \nu$. The *quadratic Wasserstein metric* of μ and ν is defined as

$$W_2(\mu, \nu) := \inf_{p(X,Y)\in\mathcal{C}(\mu,\nu)} \sqrt{\mathbb{E}[d(X, Y)^2]},$$

where $\mathcal{C}(\mu, \nu)$ is the collection of all couplings of μ, ν.

Main results Before the statement of main results, we first introduce the notion of T_2 property. See Definition D.6 below.

Definition D.6 ($T_2(\sigma)$ distribution)**.** Suppose that a probability measure ρ on metric space $(\mathbb{X}, d)$ satisfy the *quadratic transportation cost* (T_2) inequality

$$W_2(\rho, \nu) \le \sqrt{2\sigma^2 KL(\nu, \rho)} \qquad \text{for all measures } \nu \text{ on } \mathbb{X},$$

then we say ρ is a $T_2(\sigma)$ distribution.

We remark that T_2 is a broad class that contains many common distributions as special cases. For example, Gaussian distribution $\mathcal{N}(\cdot, M)$ satisfies $T_2\left(\sqrt{\|M\|_2}\right)$-inequality. Strongly log-concave distributions are T_2. Suppose ρ is a continuous measure with a convex and compact support set. If its smallest density is lower bounded within the support, then ρ is T_2.

We have an estimate of the (local) Rademacher complexities of $\mathcal{F}_s$ in Lemma D.7.

Lemma D.7 (Full version of Proposition 6.4)**.** *Suppose that for $X \sim \rho$, the distribution of $\phi_\alpha(X) \in \mathbb{R}^s$ satisfies $T_2\big(\sigma(\alpha)\big)$-inequality for any $\alpha \in \mathcal{I}$. Let $\sigma^2_{\min}(\alpha)$ be the smallest positive eigenvalue of $\Sigma_\alpha = \mathbb{E}_\rho[\phi_\alpha\phi_\alpha^\top]$. Let η_s be a constant such that $\eta_s \ge \sigma(\alpha)/\sigma_{\min}(\alpha)$ for any $\alpha \in \mathcal{I}$. There exists a universal constant $c > 0$ such that when $n \ge cs\log d$,*

$$\mathcal{R}_n^\rho(\mathcal{F}_s) \le c(1 + \eta_s)\sqrt{\frac{Bs\log d}{n}}.$$

Moreover, when $n \ge cs\log d$, for any $f^\circ \in \mathcal{F}_s$, the local Rademacher complexity of $\mathcal{F}_s$ satisfies

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f - f^\circ\|_\rho^2 \le r\}\big) \le \psi(r), \qquad \textit{with } \psi(r) := c\sqrt{r}(1 + \eta_s)\sqrt{\frac{s\log d}{n}}.$$

Here, $\psi(r)$ is a sub-root function with a unique positive fixed point

$$r^\star = c^2(1 + \eta_s)^2 \cdot \frac{s\log d}{n}.$$

When $\phi(X)$ follows a non-degenerated Gaussian distribution with covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have $\sigma(\alpha) \le \sqrt{\lambda_{\max}(\Sigma_\alpha)}$. Since $\mathbb{E}_\rho[\phi_\alpha\phi_\alpha^\top] \succeq \Sigma_\alpha$, it also holds that $\sigma_{\min}(\alpha) = \sqrt{\lambda_{\min}\big(\mathbb{E}_\rho[\phi_\alpha\phi_\alpha^\top]\big)} \ge \sqrt{\lambda_{\min}(\Sigma_\alpha)}$. According to Lemma D.7, we take a parameter $\kappa_s(\Sigma)$ such that $\kappa_s(\Sigma) \ge \lambda_{\max}(\Sigma_\alpha)/\lambda_{\min}(\Sigma_\alpha) \ge 1$ for all $\alpha \in \mathcal{I}$. In this way, the result in Lemma D.7 holds for $\eta_s = \sqrt{\kappa_s(\Sigma)}$ and reduces to Proposition 6.4.

Proof of main results In the sequel, we prove Lemma D.7. We first present some preliminary results.

Lemma D.8. *For arbitrary random variables $X_1, X_2, \ldots, X_m \geq 0$ ($m \geq 2$) satisfying $\|X_i \mathbb{1}_{\{|X_i| \leq R\}}\|_{\psi_2} \leq \kappa_2$ and $\|X_i \mathbb{1}_{\{|X_i| > R\}}\|_{\psi_1} \leq \kappa_1$ for $i = 1, 2, \ldots, m$ and some parameter $R \geq 1$, we have*

$$\mathbb{E} \max_{1 \leq i \leq m} X_i \leq c\Big(\kappa_2 \sqrt{\log m} + m(\kappa_1 + R) e^{-cR/\kappa_1}\Big),$$

where $c > 0$ is a universal constant.

Proof. We first note that $\mathbb{E}\big[\max_{1 \leq i \leq m} X_i\big] \leq U + V$ with $U := \mathbb{E}\big[\max_{1 \leq i \leq m} X_i \mathbb{1}_{\{X_i \leq R\}}\big]$ and $V := \mathbb{E}\big[\max_{1 \leq i \leq m} X_i \mathbb{1}_{\{|X_i| \geq R\}}\big]$. In what follows, we analyze U and V separately.

By definition of ψ_2-norm and our assumption $\big\|X_i \mathbb{1}_{\{|X_i| \leq R\}}\big\|_{\psi_2} \leq \kappa_2$, we have $\mathbb{E}\big[\exp(X_i^2 \mathbb{1}_{\{|X_i| \leq R\}} / \kappa_2^2) - 1\big] \leq 1$ for $i = 1, 2, \ldots, m$. It follows that

$$\begin{aligned}
\mathbb{E}\left[\max_{1 \leq i \leq m} \frac{X_i^2 \mathbb{1}_{\{|X_i| \leq R\}}}{\kappa_2^2}\right] &\overset{\text{Jensen's inequality}}{\leq} \log \mathbb{E}\left[\max_{1 \leq i \leq m} \exp\Big(\frac{X_i^2 \mathbb{1}_{\{|X_i| \leq R\}}}{\kappa_2^2}\Big)\right] \\
&\leq \log\left(\sum_{i=1}^m \mathbb{E}\left[\exp\Big(\frac{X_i^2 \mathbb{1}_{\{|X_i| \leq R\}}}{\kappa_2^2}\Big)\right]\right) \leq \log(2m) \leq 2 \log m.
\end{aligned}$$

Therefore, by Jensen's inequality $U = \mathbb{E}\big[\max_{1 \leq i \leq m} X_i \mathbb{1}_{\{|X_i| \leq R\}}\big] \leq \sqrt{\mathbb{E}\big[\max_{1 \leq i \leq m} X_i^2 \mathbb{1}_{\{|X_i| \leq R\}}\big]} \leq \kappa_2 \sqrt{2 \log m}$.

Recall that $\|X_i \mathbb{1}_{\{|X_i| > R\}}\|_{\psi_1} \leq \kappa_1$, which implies there exists a universal constant $c \geq 1$ such that $\mathbb{P}\big(|X_i| \mathbb{1}_{\{|X_i| \geq R\}} > t\big) \leq c e^{-ct/\kappa_1}$. Using this fact, we find that

$$\begin{aligned}
V \leq& \mathbb{E}\left[\max_{1 \leq i \leq m} |X_i| \mathbb{1}_{\{|X_i| > R\}}\right] = \left(\int_0^R + \int_R^\infty\right) \mathbb{P}\Big(\max_{1 \leq i \leq m} |X_i| \mathbb{1}_{\{|X_i| > R\}} \geq t\Big) \mathrm{d}t \\
=& R \mathbb{P}\Big(\max_{1 \leq i \leq m} |X_i| \mathbb{1}_{\{|X_i| > R\}} \geq R\Big) + \int_R^\infty \mathbb{P}\Big(\max_{1 \leq i \leq m} |X_i| \mathbb{1}_{|X_i| > R\}} \geq t\Big) \mathrm{d}t \\
\overset{\text{union bound}}{\leq}& mR \mathbb{P}\big(|X_i| \mathbb{1}_{\{|X_i| > R\}} \geq R\big) + m \int_R^\infty \mathbb{P}\big(|X_i| \mathbb{1}_{\{|X_i| > R\}} \geq t\big) \mathrm{d}t \\
\leq& mR \cdot c e^{-cR/\kappa_1} + m \int_R^\infty c e^{-ct/\kappa_1} \mathrm{d}t = cmR e^{-cR/\kappa_1} + m\kappa_1 e^{-cR/\kappa_1} \leq cm(\kappa_1 + R) e^{-cR/\kappa_1}.
\end{aligned}$$

Integrating the pieces, we finish the proof. □

Lemma D.9. *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors satisfying $T_2(\sigma)$-inequality and $\mathbb{E}[X_1 X_1^\top] = M \in \mathbb{R}^{d \times d}$. Suppose that $n \geq d$. Let $\sigma_1, \sigma_2, \ldots, \sigma_n$ be Rademacher random variables independent of $X_1, X_2, \ldots, X_n$. Then $Y := \big\|\frac{1}{n} \sum_{k=1}^n \sigma_k X_k\big\|_{M^\dagger}$ satisfies*

$$\Big\||Y - \mathbb{E}[Y]| \mathbb{1}\big\{|Y - \mathbb{E}[Y]| \leq \big(1 + \sigma \sqrt{\|M^\dagger\|_2}\big)\big\}\Big\|_{\psi_2} \leq c\left(\frac{1}{\sqrt{n}} + \sigma \sqrt{\frac{\|M^\dagger\|_2}{n}}\right)$$

and

$$\Big\||Y - \mathbb{E}[Y]| \mathbb{1}\big\{|Y - \mathbb{E}[Y]| > \big(1 + \sigma \sqrt{\|M^\dagger\|_2}\big)\big\}\Big\|_{\psi_1} \leq c\left(\frac{1}{n} + \frac{\sigma \sqrt{\|M^\dagger\|_2}}{n}\right).$$

Proof. We take shorthands $\mathbf{X} := [X_1, X_2, \ldots, X_n] \in \mathbb{R}^{d \times n}$, $\boldsymbol{\sigma} := (\sigma_1, \ldots, \sigma_n)^\top \in \mathbb{R}^n$ and rewrite Y as $Y = \frac{1}{n} \|\mathbf{X}\boldsymbol{\sigma}\|_{M^\dagger}$. Note that $Y - \mathbb{E}Y = \big(Y - \mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}]\big) + \big(\mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}] - \mathbb{E}Y\big)$. In the following, we analyze these two terms separately.

Note that $\nabla_\sigma Y = n^{-1} \|\mathbf{X}\boldsymbol{\sigma}\|_{M^\dagger}^{-1} \mathbf{X}^\top M^\dagger \mathbf{X}\boldsymbol{\sigma}$ and $\|\nabla_{\boldsymbol{\sigma}} Y\|_2 \leq n^{-1} \big\|\sqrt{M^\dagger}\mathbf{X}\big\|_2$, therefore, Y is $(n^{-1} \big\|\sqrt{M^\dagger}\mathbf{X}\big\|_2)$-Lipschitz with respect to σ in the Euclidean norm. Moreover, Y is convex in $\boldsymbol{\sigma}$ and the Rademacher random variables are independent

and bounded. We use Talagrand's inequality (See Theorem 4.20 and Corollary 4.23 in (van Handel, 2014).) and obtain that there exists a universal constant $c > 0$ such that

$$\mathbb{P}\Big(\big|Y - \mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}]\big| \geq t_1 n^{-1} \big\|\sqrt{M^\dagger}\mathbf{X}\big\|_2 \;\Big|\; \mathbf{X}\Big) \leq ce^{-ct_1^2} \qquad \text{for any } t_1 > 0. \tag{58}$$

We next consider the concentration of $\big\|\sqrt{M^\dagger}\mathbf{X}\big\|_2$. For random vector X, we define

$$\|X\|_{\psi_2} := \sup_{\mathbf{u}\in\mathbb{R}^d, \|\mathbf{u}\|_2 \leq 1} \|\mathbf{u}^\top X\|_{\psi_2}.$$

Since X satisfies $T_2(\sigma)$-inequality, according to Gozlan's theorem (Theorem 4.31 in van Handel (2014)), we find that $\big\|\sqrt{M^\dagger}(X - \mathbb{E}X)\big\|_{\psi_2} \leq c\sigma\sqrt{\|M^\dagger\|_2}$ for some universal constant $c > 0$. Additionally, we have $\big\|\sqrt{M^\dagger}\mathbb{E}X\big\|_2 \leq \sqrt{\big\|\mathbb{E}\big[\sqrt{M^\dagger}XX^\top\sqrt{M^\dagger}\big]\big\|_2} = 1$. Therefore, $\big\|\sqrt{M^\dagger}X\big\|_{\psi_2} \leq \big\|\sqrt{M^\dagger}(X - \mathbb{E}X)\big\|_{\psi_2} + \big\|\sqrt{M^\dagger}\mathbb{E}X\big\|_2 \leq 1 + c\sigma\sqrt{\|M^\dagger\|_2}$. We now apply Theorem 5.39 in Vershynin (2010) and obtain that

$$\mathbb{P}\Big(\big\|\sqrt{M^\dagger}\mathbf{X}\big\|_2 \geq \sqrt{n} + c(\sqrt{d} + t)\big\|\sqrt{M^\dagger}X\big\|_{\psi_2}\Big) \leq ce^{-ct^2},$$

which further implies

$$\mathbb{P}\Big(\big\|\sqrt{M^\dagger}\mathbf{X}\big\|_2 \geq \sqrt{n} + c\big(\sqrt{d} + t_2\big)\big(1 + \sigma\sqrt{\|M^\dagger\|_2}\big)\Big) \leq ce^{-ct_2^2} \qquad \text{for all } t_2 > 0. \tag{59}$$

Combining eq. (58) and eq. (59), we learn that

$$\mathbb{P}\Big(\big|Y - \mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}]\big| \geq t_1 n^{-\frac{1}{2}} + ct_1 n^{-1}\big(\sqrt{d} + t_2\big)\big(1 + \sigma\sqrt{\|M^\dagger\|_2}\big)\Big) \leq c\big(e^{-ct_1^2} + e^{-ct_2^2}\big). \tag{60}$$

As for the second term $\mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}] - \mathbb{E}Y$, we use the $T_2(\sigma)$ property of sample distribution and Gozlan's theorem (Theorem 4.31 in van Handel (2014)). We first show that $\mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}]$ is $\sqrt{\frac{\|M^\dagger\|_2}{n}}$-Lipschitz with respect to Frobenius norm $\|\cdot\|_F$. In fact,

$$\begin{aligned}
&\big|\mathbb{E}_\sigma[Y \mid \mathbf{X}] - \mathbb{E}_{\boldsymbol{\sigma}}[Y \mid \mathbf{X}']\big| = \frac{1}{n}\big|\mathbb{E}_\sigma\|\mathbf{X}\boldsymbol{\sigma}\|_{M^\dagger} - \mathbb{E}_{\boldsymbol{\sigma}}\|\mathbf{X}'\sigma\|_{M^\dagger}\big| \leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\big|\|\mathbf{X}\boldsymbol{\sigma}\|_{M^\dagger} - \|\mathbf{X}'\boldsymbol{\sigma}\|_{M^\dagger}\big| \\
\leq &\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\big\|(\mathbf{X} - \mathbf{X}')\boldsymbol{\sigma}\big\|_{M^\dagger} \leq \frac{1}{n}\sqrt{\|M^\dagger\|_2}\|\mathbf{X} - \mathbf{X}'\|_2\mathbb{E}_{\boldsymbol{\sigma}}\|\boldsymbol{\sigma}\|_2 \leq \sqrt{\frac{\|M^\dagger\|_2}{n}}\|\mathbf{X} - \mathbf{X}'\|_F.
\end{aligned}$$

We then apply Gozlan's theorem and find that there exists a universal constant $c > 0$ such that

$$\mathbb{P}\left(\big|\mathbb{E}_\sigma[Y \mid \mathbf{X}] - \mathbb{E}[Y]\big| \geq t_1\sigma\sqrt{\frac{\|M^\dagger\|_2}{n}}\right) \leq ce^{-ct_1^2} \qquad \text{for any } t_1 > 0. \tag{61}$$

Integrating eq. (60) and eq. (61) and using the condition $n \geq d$, we find that

$$\mathbb{P}\Big(\big|Y - \mathbb{E}[Y]\big| \geq t_1 n^{-\frac{1}{2}}\big(1 + cn^{-\frac{1}{2}}t_2\big)\big(1 + \sigma\sqrt{\|M^\dagger\|_2}\big)\Big) \leq c\big(e^{-ct_1^2} + e^{-ct_2^2}\big).$$

If $0 \leq t_1 \leq \sqrt{n}$, then by letting $t_2 = \sqrt{n}$, we have

$$\mathbb{P}\Big(\big|Y - \mathbb{E}[Y]\big| \geq ct_1 n^{-\frac{1}{2}}\big(1 + \sigma\sqrt{\|M^\dagger\|_2}\big)\Big) \leq ce^{-ct_1^2}.$$

Otherwise, when $t_1 > \sqrt{n}$, we take $t_2 = t_1$ and obtain

$$\mathbb{P}\Big(\big|Y - \mathbb{E}[Y]\big| \geq ct_1^2 n^{-1}\big(1 + \sigma\sqrt{\|M^\dagger\|_2}\big)\Big) \leq ce^{-ct_1^2}.$$

We then finish the proof by combining these two cases. □

We are now ready to prove Lemma D.7.

Proof of Lemma D.7. Note that $\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f - f^\circ\|_\rho^2 \le r\}\big) = \mathcal{R}_n^\rho\big(\{f - f^\circ \mid f \in \mathcal{F}_s, \|f - f^\circ\|_\rho^2 \le r\}\big) \le \mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_{2s} \mid \|f\|_\rho^2 \le r\}\big)$. Therefore, we can easily obtain upper bounds for $\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f - f^\circ\|_\rho^2 \le r\}\big)$ by analyzing $\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f\|_\rho^2 \le r\}\big)$. To this end, in the following, we focus on the local Rademacher complexity

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f\|_\rho^2 \le r\}\big).$$

To simplify the notation, we write $x := (s, a)$. Note that

$$\begin{aligned}\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f\|_\rho^2 \le r\}\big) =& \mathbb{E}\sup\left\{\frac{1}{n}\sum_{k=1}^K \sigma_k f(x_k) \;\middle|\; f \in \mathcal{F}, \|f\|_\rho^2 \le r\right\} \\ =& \mathbb{E}\sup\left\{\frac{1}{n}\sum_{k=1}^n \sigma_k \phi_\alpha(x_k)^\top w \;\middle|\; \alpha \in \mathcal{I}, w \in \mathbb{R}^s, w^\top \Sigma_\alpha w \le r\right\}.\end{aligned}$$

We fix α, $\{\sigma_k\}_{k=1}^n$ and $\{x_k\}_{k=1}^n$ and then optimize $w \in \mathbb{R}^s$. Since $x_k \in \mathrm{supp}(\rho)$, one always has $\frac{1}{n}\sum_{k=1}^K \sigma_k \phi_\alpha(x_k) \in \mathrm{range}(\Sigma_\alpha)$ with probability one. The supremum is therefore acheived at

$$w := \frac{\sqrt{r}\Sigma_\alpha^\dagger\big[\frac{1}{n}\sum_{k=1}^n \sigma_k \phi_\alpha(x_k)\big]}{\big\|\frac{1}{n}\sum_{k=1}^n \sigma_k \phi_\alpha(x_k)\big\|_{\Sigma_\alpha^\dagger}}.$$

It follows that

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f\|_\rho^2 \le r\}\big) = \sqrt{r}\mathbb{E}\max_{\alpha\in\mathcal{I}} Y_\alpha, \qquad \text{where } Y_\alpha := \left\|\frac{1}{n}\sum_{k=1}^n \sigma_k \phi_\alpha(x_k)\right\|_{\Sigma_\alpha^\dagger}.$$

We further upper bound the local Rademacher complexity by

$$\mathcal{R}_n^\rho\big(\{f \in \mathcal{F}_s \mid \|f\|_\rho^2 \le r\}\big) \le \sqrt{r}\Bigg(\underbrace{\max_{\alpha\in\mathcal{I}}\mathbb{E}[Y_\alpha]}_{E_1} + \underbrace{\mathbb{E}\Big[\max_{\alpha\in\mathcal{I}}\big(Y_\alpha - \mathbb{E}[Y_\alpha]\big)\Big]}_{E_2}\Bigg). \tag{62}$$

In the following, we estimate the two terms in the right hand side of eq. (62) separately.

Define $\boldsymbol{\sigma} := (\sigma_1, \ldots, \sigma_n)^\top \in \mathbb{R}^n$ and $\Phi_\alpha := \big[\phi_\alpha(x_1), \ldots, \phi_\alpha(x_n)\big] \in \mathbb{R}^{s\times n}$. We reform Y_α as $Y_\alpha = n^{-1}\|\Phi_\alpha\boldsymbol{\sigma}\|_{\Sigma_\alpha^\dagger}$. It follows that

$$\mathbb{E}[Y_\alpha^2] = \frac{1}{n^2}\mathbb{E}\big[\|\Phi_\alpha\boldsymbol{\sigma}\|_{\Sigma_\alpha^\dagger}^2\big] = \frac{1}{n^2}\mathbb{E}\big[(\Phi_\alpha\boldsymbol{\sigma})^\top\Sigma_\alpha^\dagger(\Phi_\alpha\boldsymbol{\sigma})\big] = \frac{1}{n^2}\mathbb{E}\big[\mathrm{Tr}(\Sigma_\alpha^\dagger\Phi_\alpha\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\Phi_\alpha^\top)\big].$$

We use the relations $\frac{1}{n}\mathbb{E}\big[\Phi_\alpha\Phi_\alpha^\top\big] = \Sigma_\alpha$ and $\mathbb{E}[\boldsymbol{\sigma}\boldsymbol{\sigma}^\top] = I_s$ where I_r represents the identity matrix in $\mathbb{R}^{s\times s}$. The inequality above is then reduced to

$$\mathbb{E}[Y_\alpha] \le \sqrt{\mathbb{E}[Y_\alpha^2]} \le \frac{1}{\sqrt{n}}\sqrt{\mathrm{rank}(\Sigma_\alpha)} \le \sqrt{\frac{s}{n}}. \tag{63}$$

To this end, we have $E_1 \le \sqrt{s/n}$.

Now we focus on E_2. Since $\phi_\alpha(x)$ satisfies $T_2\big(\sigma(\alpha)\big)$-inequality. Applying Lemma D.9, we find that if $n \ge s$,

$$\big\||Y_\alpha - \mathbb{E}[Y_\alpha]|\mathbb{1}\big\{|Y_\alpha - \mathbb{E}[Y_\alpha]| \le \big(1 + \sigma(\alpha)/\sigma_{\min}(\alpha)\big)\big\}\big\|_{\psi_2} \le \frac{c}{\sqrt{n}}\left(1 + \frac{\sigma(\alpha)}{\sigma_{\min}(\alpha)}\right) \le \frac{c}{\sqrt{n}}(1 + \eta_s)$$

and

$$\big\||Y_\alpha - \mathbb{E}[Y_\alpha]|\mathbb{1}\big\{|Y_\alpha - \mathbb{E}[Y_\alpha]| > \big(1 + \sigma(\alpha)/\sigma_{\min}(\alpha)\big)\big\}\big\|_{\psi_1} \le \frac{c}{n}\left(1 + \frac{\sigma(\alpha)}{\sigma_{\min}(\alpha)}\right) \le \frac{c}{n}(1 + \eta_s).$$

We further use Lemma D.9 and obtain

$$\begin{aligned}\mathbb{E}\max_{\alpha\in\mathcal{I}}|Y_\alpha-\mathbb{E}[Y_\alpha]| \leq& c(1+\eta_s)\Big(n^{-\frac{1}{2}}\sqrt{\log|\mathcal{I}|}+|\mathcal{I}|e^{-cn}\Big)\\ \leq& c(1+\eta_s)\Big(n^{-\frac{1}{2}}\sqrt{s\log d}+\exp\big(-cn+s\log d\big)\Big).\end{aligned}$$

If $n\geq c's\log d$ for some sufficiently large constant c', then

$$E_2=\mathbb{E}\max_{\alpha\in\mathcal{I}}\big|Y_\alpha-\mathbb{E}[Y_\alpha]\big|\leq c(1+\eta_s)\sqrt{\frac{s\log d}{n}}. \tag{64}$$

Plugging eqs. (63) and (64) into eq. (62), we complete our proof. □

E. Proof of Lower Bound (Theorem 5.1)

In this section, we will prove a stronger version of Theorem 5.1, which is Theorem E.1. In Theorem E.1, we show that in the same setting as Theorem 5.1, even if additionally assuming Assumption 1 holds with $C=1$, i.e., μ_h is the true marginal distribution of the single-action MDP, and the algorithm knows $\{\mu_h\}_{h=1}^H$, it still takes $\Omega(\frac{\sqrt{S}}{\varepsilon^2})$ samples for the learning algorithm $\mathfrak{A}$ to achieve ε optimality gap for Bellman error. This further justifies the necessity of Assumption 2 and Assumption 3 in the single sampling regime.

Theorem 5.1. *Let $\mathfrak{A}$ be an arbitrary algorithm that takes any dataset $\mathcal{D}$ and function class $\mathcal{F}$ as input and outputs an estimator $\hat{f}\in\mathcal{F}$. For any $S\in\mathbb{N}^+$ and sample size $n\geq 0$, there exists an S-state, single-action MDP paired with a function class $\mathcal{F}$ with $|\mathcal{F}|=2$ such that the $\hat{f}$ output by algorithm $\mathfrak{A}$ satisfies*

$$\mathbb{E}\mathcal{E}(\hat{f})\geq\min_{f\in\mathcal{F}}\mathcal{E}(f)+\Omega\left(\min\left\{1,\frac{S^{1/2}}{n}\right\}\right). \tag{6}$$

Here, the expectation is taken over the randomness in $\mathcal{D}$.

Theorem E.1. *For any $\varepsilon<0.5$ and $S\geq 2$, there is a family of single-action, $S+5$-state MDPs ($H=3$) with the same underlying distributions μ_h (satisfying Assumption 1 with $C=1$) and the same reward function (thus the MDPs only differ in probabiilty transition matrices) and a function class $\mathcal{F}$ of size 2, such that all learning algorithm $\mathfrak{A}$ that takes n pairs of states (s,a,r,s') and output a value function in $\mathcal{F}$ must suffer $\Omega(\varepsilon^2)$ expected optimality gap in terms of mean-squared bellman error w.r.t μ if $n=O(\frac{\sqrt{S}}{\varepsilon^2})$.*

Mathematically, it means for any learning algorithm $\mathfrak{A}$, there is a single-action, $S+5$-state MDP defined above, such that for $D=\cup_h\{(s_i,a_i,r_i,s_i',h)\}_{i=1}^n$ sampled from $\mathcal{M}$ and μ, if $n=O(\frac{\sqrt{S}}{\varepsilon^2})$, we have

$$\mathbb{E}_D\left[\mathcal{E}_{\mathcal{M}}\left(\mathfrak{A}(D)\right)\right]\geq\min_{f\in\mathcal{F}}\mathcal{E}_{\mathcal{M}}\left(f\right)+\Omega(\varepsilon^2).$$

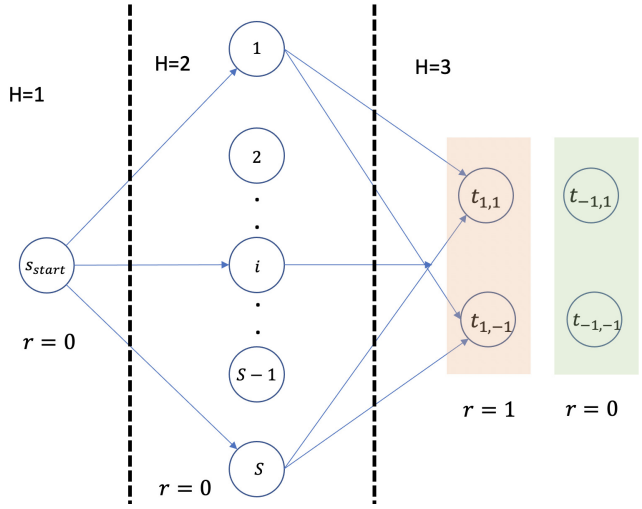
Below we will prove Theorem E.1. To better illustrate the idea of the hard instance, we will first prove a slightly weaker version with $C=2$ (Theorem E.2) in Appendix E.1 and in Appendix E.2 we will prove Theorem E.1 by slightly twisting the proof in Appendix E.1.

E.1. Warm-up with $C=2$

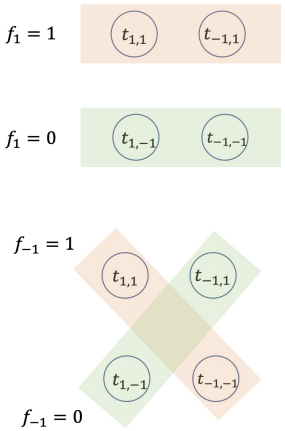
We construct the hard instances for single sampling in the following way.

Hard Instance Construction: We first generate a uniform random bit $c\in\{-1,1\}$, and a Radamacher vector $\boldsymbol{\sigma}\in\{\pm1\}^S$. For each $c,\boldsymbol{\sigma}$, we define MDP $\mathcal{M}_{c,\boldsymbol{\sigma}}^{\varepsilon}=(\mathcal{S},\mathcal{A},H,\mathbb{P}_{c,\boldsymbol{\sigma}}^{\varepsilon},r)$ below, where $0<\varepsilon<0,5$. The claim is the distribution of $\mathcal{M}_{c,\boldsymbol{\sigma}}^{\varepsilon}$ serves as the distribution of hard instances. Note that only $\mathbb{P}_{c,\boldsymbol{\sigma}}^{\varepsilon}$ in the tuple defining $\mathcal{M}_{c,\boldsymbol{\sigma}}^{\varepsilon}$ depends on c and $\boldsymbol{\sigma}$. Here the probability transition matrix $\mathcal{M}_{c,\boldsymbol{\sigma}}^{\varepsilon}$ is the same for all $h=1,2,\ldots,H$.

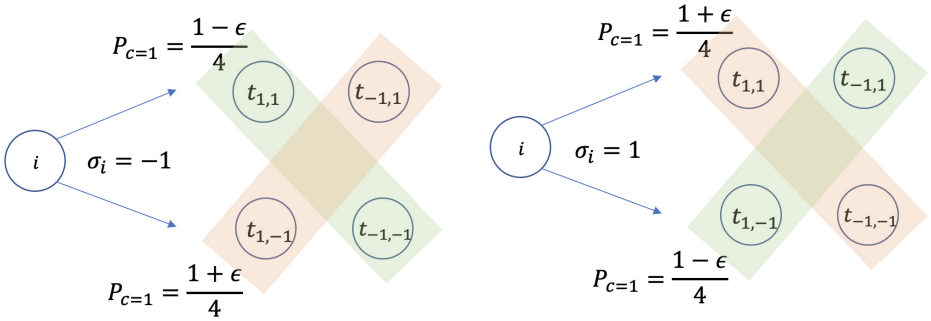
Let $\mathcal{S}=\{s_{\text{start}}\}\cup\{1,\ldots,S\}\cup\{t_{j,k}\}_{j,k\in\{-1,1\}}$, $H=2$, $|\mathcal{A}|=1$ and the initial state is s_{start}. Since there's only one action, below we will just drop the dependence on action and thus simplify the notation. We will always define the probability transition matrix in the way such that in the 2nd step, we will reach some state among $1,\ldots,S$ and in the 3rd step, we will reach some state among $t_{j,k}$.



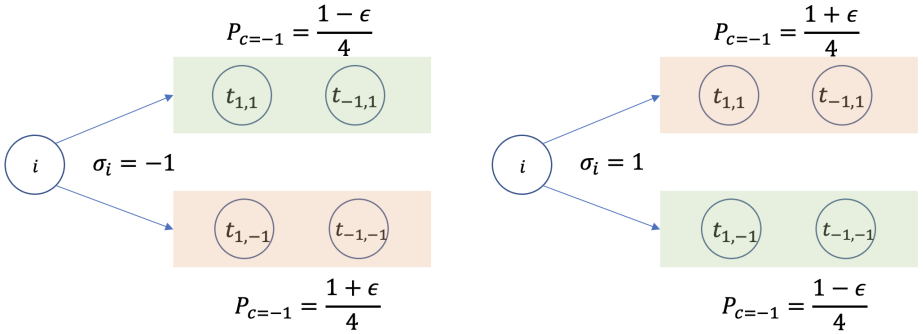
(a) Illustration of the 3-stage, single action MDP. Each state can be visited for at most one $h = 1, 2, 3$. r is the reward for each state.(action omitted since there's only one)



(b) Illustration of f_1 and f_{-1}. They only differ on $t_{-1,1}$ and $t_{-1,-1}$. For $h = 3$, the Bellman error $\|f_{c'} - \mathcal{T}^3_{c,\boldsymbol{\sigma}} f_{c'}\|^2_{2,\mu_3} = 0.5$, regardless of c' and c.



(c) Illustration of $\overline{\mathbb{P}}^{\varepsilon}_{1,\sigma_i}$. When $c = 1$, there are two different but equally likely types of state i, depending on their probability transition matrix for the next step.



(d) Illustration of $\overline{\mathbb{P}}^{\varepsilon}_{-1,\sigma_i}$. When $c = 1$, there are two different but equally likely types of state i, depending on their probability transition matrix for the next step.

Figure 1. Graphical illustration of the hard instances $\mathcal{M}^{\varepsilon}_{c,\boldsymbol{\sigma}}$. As shown in Equation (65), the total Bellman error is only determined by the Bellman error for $h = 2$, which is equal to optimal error + $\frac{\varepsilon^2}{12}\mathbb{1}[c \neq c']$ if $f_{c'}$ is the returned function. The main idea of the proof is to show it's difficult to guess c via the observed dataset D if D only contains single-sampled data. As a sanity check, for any c and sample $(i, t_{j,k})$, if $\sigma_i \overset{unif}{\sim} \{\pm 1\}$, the marginal distribution of $t_{j,k}$ is always uniform, but for double sampling of form $(i, t_{j,k}, t_{j',k'})$, we can decide c by simply looking at histogram of $(t_{j,k}, t_{j',k'})$.

Function class: $\mathcal{F} = \{f_1, f_{-1}\}$, where $f_c(s_{\text{start}}) = \frac{1}{2}$, $f_c(i) = \frac{1}{2}, \forall 1 \le i \le S$ and $f_c(t_{j,k}) = \frac{k\max(c,j)+1}{2}$, $\forall c, j, k \in \{\pm 1\}$. Compared to the notation in the main paper, we drop the dependency on h for $f \in \mathcal{F}$. This is because the MDP will reach a disjoint set of states for each step h (see below).

Probability Transition Matrix: We define the probability transition matrix below. Specifically, for $i \in \{1, \dots, S\}$ and $j, k \in \{\pm 1\}$, $\mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}(t_{j,k} \mid i) \equiv \overline{\mathbb{P}}^{\varepsilon}_{c,\sigma_i}(t_{j,k}) := 0.25(1 + \varepsilon k \max(-c, j)\sigma_i)$.

From \ To	s_{start}	$i (i = 1, \dots, S)$	$t_{j,k}$	s_{end}
s_{start}	0	$\frac{1}{S}$	0	0
$i (i = 1, \dots, S)$	0	0	$\overline{\mathbb{P}}^{\varepsilon}_{c,\sigma_i}(t_{j,k}) := 0.25(1 + \varepsilon k \max(-c, j)\sigma_i)$	0
$t_{j,k}$	0	0	0	1

Table 1. Probability Transition Matrix $\mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}$ for MDP $\mathcal{M}^{\varepsilon}_{c,\boldsymbol{\sigma}}$. Starting from s_{start}, the process terminates as it reaches s_{end} in the 4th step.

Reward Function: $r(s_{\text{start}}) = 0, r(i) = 0, \forall 1 \le i \le S, r(t_{j,k}) = \frac{j+1}{2}, \forall j, k \in \{-1, 1\}$.

Underlying distribution: We define the underlying distribution for batch data μ as $\mu_2(i) = \frac{1}{S}$ and $\mu_3(t_{j,k}) = \frac{1}{4}$, we can check that Assumption 1 is satisfied with $C = 2$ as $\varepsilon < 0.5$. Define $\mathcal{T}^1_{c,\boldsymbol{\sigma}}, \mathcal{T}^2_{c,\boldsymbol{\sigma}}, \mathcal{T}^3_{c,\boldsymbol{\sigma}}$ be the Bellman operator of $\mathcal{M}^{\varepsilon}_{c,\sigma}$, we have $\forall \boldsymbol{\sigma} \in \{-1, 1\}^S, \forall c, c' \in \{-1, 1\}$,

$$\|f_{c'} - \mathcal{T}^3_{c,\boldsymbol{\sigma}} f_{c'}\|^2_{2,\mu_3} = \|f_{c'} - r\|^2_{2,\mu_3} = \mathbb{P}\left[j \ne k \max(c', j)\right] = 0.5,$$

$$\begin{aligned}\|f_{c'} - \mathcal{T}^2_{c,\boldsymbol{\sigma}} f_{c'}\|^2_{2,\mu_2} =& \left\| \sum_{j,k \in \{\pm 1\}} \mathbb{P}_{c,\boldsymbol{\sigma}}(t_{j,k} \mid i) f_{c'}(t_{j,k}) \right\|^2_{2,\mu_2} \\ =& \frac{1}{64} \left\| \sum_{j,k} \varepsilon \sigma_i k^2 \max(j, c) \max(j, -c') \right\|^2_{2,\mu_2} = \frac{\varepsilon^2}{4} \mathbb{1}\left[c \ne c'\right],\end{aligned}$$

$$\|f_{c'} - \mathcal{T}^1_{c,\boldsymbol{\sigma}} f_{c'}\|^2_{2,\mu_3} = \|f_c(s_{\text{start}}) - f_c(i)\|^2_{2,\mu_1} = 0.$$

Thus

$$\mathcal{E}_{c,\boldsymbol{\sigma}}(f_{c'}) \equiv \mathcal{E}_{\mathcal{M}_{c,\boldsymbol{\sigma}}}(f_{c'}) = \frac{1}{3} \sum_{h=1}^{3} \|f_{c'} - \mathcal{T}^h_{c,\boldsymbol{\sigma}} f_{c'}\|^2_{2,\mu_h} = \frac{1}{3}(0.5 + \frac{\varepsilon^2}{4} \mathbb{1}\left[c \ne c'\right]). \tag{65}$$

From eq. (65) we can see minimizing Bellman error in this case is equivalent to predict $-c$. And any algorithm predicts c wrongly, i.e., outputs $f_{c'}$ with $c' \ne c$ with constant probability, will suffer $\Omega(\varepsilon^2)$ expected optimality gap. More specifically, we can show that for random $\boldsymbol{\sigma}$, it's information-theoretically hard to predict c correctly given D, which leads to the following theorem.

Theorem E.2. *For $c \overset{iid}{\sim} \{-1, 1\}$, $\boldsymbol{\sigma} \overset{iid}{\sim} \{-1, 1\}^S$, $D = \cup_{h=1}^{3} \{(s_i, a_i, r_i, s_i', h)\}_{i=1}^{n}$ sampled from $\mathcal{M}^{\varepsilon}_{c,\boldsymbol{\sigma}}$ and μ, we have for any learning algorithm $\mathfrak{A}$ with $n = O(\frac{\sqrt{S}}{\varepsilon^2})$ samples,*

$$\mathbb{E}_{c,\boldsymbol{\sigma}} \mathbb{E}_D \left[\mathcal{E}_{c,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right] \ge \mathbb{E}_{c,\boldsymbol{\sigma}} \left[\min_{c' \in \{-1,1\}} \mathcal{E}_{c,\boldsymbol{\sigma}}\left(f_{c'}\right)\right] + \Omega(\varepsilon^2).$$

Or equivalently (and more specifically), if we view $\widetilde{\mathfrak{A}}(D)$ as the modified version of $\mathfrak{A}$, whose range is $\{-1, 1\}$ and satisfies $\mathfrak{A} = f_c$ with $c = \widetilde{\mathfrak{A}}(D)$. Then we have

$$\mathbb{E}_{c,\boldsymbol{\sigma}} \mathbb{E}_D \left[\mathbb{1}\left[\widetilde{\mathfrak{A}}(D) \ne c\right]\right] \ge \Omega(\varepsilon^2).$$

Towards proving Theorem E.2, we need the following lower bound, where $\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}$ is defined as the joint distribution of (s, s'), where $s \sim \mu_2$ and $s' \sim \mathbb{P}_{c,\boldsymbol{\sigma}|s}$. Note that when $\varepsilon = 0$, $\mathbb{P}^0_{c,\boldsymbol{\sigma}}(\cdot \mid i)$ becomes uniform distribution for every $1 \le i \le S$, and thus is independent of $c, \boldsymbol{\sigma}$, which could be denoted by $\mathbb{P}^0$ therefore.

Lemma E.3. *If* $n \leq 0.1\frac{S^{0.5}}{\varepsilon^2}$, *then* $\|\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n - \left(\mu_2 \circ \mathbb{P}^0\right)^n\|_{TV} \leq 0.1$, *for all* $c \in \{-1, 1\}$.

Proof. For convenience, we denote $\left(\mu_2 \circ \mathbb{P}^0\right)^n$ by P and $\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n$ by Q. By Pinsker's inequality, we have $\|P - Q\|_{TV} \leq \sqrt{2KL(P,Q)}$, for any distribution P, Q. Thus it suffices to upper bound $KL(P,Q)$ by 0.05.

We define E_i as a random subset, i.e., $E_i = \{l | 1 \leq l \leq n, s_l = i\}$, given $D = \{(s_i, s'_i)\}_{i=1}^n$. Then for both $\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n$ and $\left(\mu_2 \circ \mathbb{P}^0\right)^n$, $s_1, \ldots, s_n$ are i.i.d. distributed by μ_2. Note that

$$\begin{aligned}
&Q(s'_1, \ldots, s'_n \mid s_1, \ldots, s_n) \\
=& \sum_{\boldsymbol{\sigma} \in \{-1,1\}^S} p(\boldsymbol{\sigma}) Q(s'_1, \ldots, s'_n \mid s_1, \ldots, s_n, \boldsymbol{\sigma}) \\
=& \sum_{\boldsymbol{\sigma} \in \{-1,1\}^S} \prod_{i=1}^S p(\sigma_i) \prod_{i=1}^S Q(s'_{E_i} | E_i, \sigma_i) \\
=& \prod_{i=1}^S \left(\sum_{\sigma_i \in \{-1,1\}} p(\sigma_i) Q(s'_{E_i} | E_i, \sigma_i) \right),
\end{aligned} \tag{66}$$

and

$$P(s'_1, \ldots, s'_n \mid s_1, \ldots, s_n) = \prod_{i=1}^S P(s'_{E_i} | E_i). \tag{67}$$

For any tuple $(s_1, \ldots, s_n)$ and subset $E \subset \{1, \ldots, n\}$, we define s_E as the sub-tuple of s with length $|E|$ selected by E. Define P_{E_i}, Q_{E_i} as the distribution of s'_{E_i} conditioned on E_i. In detail, $Q_{E_i}(s'_{E_i}) = \sum_{\sigma_i \in \{-1,1\}} p(\sigma_i) Q(s'_{E_i} | E_i, \sigma_i)$ and $P_{E_i}(s'_{E_i}) = P(s'_{E_i} | E_i)$. Note that for Q, s'_{E_i} are i.i.d. conditioned on E_i and σ_i, i,.e., $Q(s'_{E_i} | E_i, \sigma_i) = \prod_{l \in E_i} \overline{\mathbb{P}}^{\varepsilon}_{c,\sigma_i}(s'_l)$. Therefore the distribution Q_{E_i} only depends on $|E_i|$, so does P_{E_i}.

Thus we can write the KL divergence as:

$$\begin{aligned}
&KL\left(\left(\mu_2 \circ \mathbb{P}^0\right)^n, \mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n\right) = KL\left(P, Q\right) = \mathop{\mathbb{E}}_{D \sim P}\left[\log \frac{P(D)}{Q(D)}\right] \\
=& \mathop{\mathbb{E}}_{D \sim P}\left[\log \frac{P(s'_1, \ldots, s'_n \mid s_1, \ldots, s_n)}{Q(s'_1, \ldots, s'_n \mid s_1, \ldots, s_n)} + \log \frac{P(s_1, \ldots, s_n)}{Q(s_1, \ldots, s_n)}\right] \\
=& \mathop{\mathbb{E}}_{D \sim P}\left[\log \frac{\prod_{i=1}^S P(s'_{E_i} | E_i)}{\prod_{i=1}^S Q(s'_{E_i} | E_i)} + \log \frac{P(s_1, \ldots, s_n)}{Q(s_1, \ldots, s_n)}\right] \quad (P(s_1, \ldots, s_n) = Q(s_1, \ldots, s_n)) \\
=& \mathop{\mathbb{E}}_{D \sim P}\left[\sum_{i=1}^S \log \frac{P_{E_i}(s'_{E_i})}{Q_{E_i}(s'_{E_i})}\right] \\
=& \sum_{i=1}^S \mathop{\mathbb{E}}_{D \sim P}\left[\log \frac{P_{E_i}(s'_{E_i})}{Q_{E_i}(s'_{E_i})}\right].
\end{aligned} \tag{68}$$

By the definition of P and Q, given $c \in \pm 1$ and $\varepsilon > 0$, we can see that $\mathbb{E}_{D \sim P}\left[\log \frac{P_{E_i}(s'_{E_i})}{Q_{E_i}(s'_{E_i})}\right]$ only a function of $|E_i|$, and we denote it by $G_{c,\varepsilon}(|E_i|)$. Thus we have

$$
\begin{aligned}
&KL\left(\left(\mu_2\circ\mathbb{P}^0\right)^n,\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n\right)\\
=&KL\left(P,Q\right)\\
=&\sum_{i=1}^{S}\sum_{m=0}^{n}\mathop{\mathbb{E}}_{D\sim P}\left[\sum_{i=1}^{S}\log\frac{P_{E_i}(s'_{E_i})}{Q_{E_i}(s'_{E_i})}\middle||E_i|=m\right]P(|E_i|=m)\\
=&\sum_{i=1}^{S}\sum_{m=0}^{n}G_{c,\varepsilon}(m)P(|E_i|=m)\\
=&S\sum_{m=0}^{n}G_{c,\varepsilon}(m)P(|E_1|=m).
\end{aligned}
\tag{69}
$$

The last step is because E_i are i.i.d. distributed. For convenience, we will denote $P(|E_1|=m)$ by $P_N(m)$.

It can be shown that $G_{c,\varepsilon}(m)$ is independent of c, and thus we drop c in the subscription. We could even simplify the expression of $G_\varepsilon(m)$ by defining $R_{\sigma,\varepsilon}(j)=0.5(1+j\sigma\varepsilon)$ over $\{-1,1\}$ (For $c=1$, this is effectively grouping $(t_{1,1},t_{-1,-1})$ into a state, say 1, and $(t_{1,-1},t_{-1,1})$ into another state, say -1.)

$$
G_\varepsilon(m)=KL\left(\left(\text{Unif}\{-1,1\}\right)^m,\frac{\left(R'_{-1,\varepsilon}\right)^m+\left(R'_{1,\varepsilon}\right)^m}{2}\right).
$$

Below are some basic properties of $G_\varepsilon(m)$.

- $G_\varepsilon(0)=0$.
- $G_\varepsilon(1)=0$.
- $G_\varepsilon(m)\le\frac{6m^2+m}{8}\varepsilon^4+\frac{m}{2}\frac{\varepsilon^4}{1-\varepsilon^2}\le 2m\varepsilon^4$, for $\varepsilon^2\le\frac{1}{2}$.

The first two properties can be verified by direct calculation, and the third property is proved in Lemma E.4.

Now it remains to calculate $P_N(1)$ and $\mathbb{E}_P\left[|E_1|^2\right]$. We have

$$
P_N(1)=n\frac{1}{S}(1-\frac{1}{S})^{n-1}\le\frac{n}{S}(1-\frac{n}{S}),
$$

and

$$
\mathop{\mathbb{E}}_{P}\left[|E_1|^2\right]=\mathop{\mathbb{E}}_{P}\left[(\sum_{i=1^n}\mathbb{1}\left[s_i=1\right])^2\right]=\mathop{\mathbb{E}}_{P}\left[\sum_{i=1}^{n}\mathbb{1}\left[s_i=1\right]+\sum_{i,j=1,1\neq j}^{n}\mathbb{1}\left[s_i=s_j=1\right]\right]=\frac{n}{S}+\frac{n(n-1)}{S^2}.
$$

Thus we conclude that

$$
\begin{aligned}
KL\left(P,Q\right)=&S\sum_{m=2}^{n}P_N(m)G_\varepsilon(m)\le S\sum_{m=2}^{n}P_N(m)\cdot 2m^2\varepsilon^4=2\big(\sum_{m=2}^{n}P_N(m)m^2\big)S\varepsilon^4\\
=&2\left(\mathop{\mathbb{E}}_{P}\left[|E_1|^2\right]-P_N(1)\right)S\varepsilon^4=2\left(\frac{n(n-1)}{S^2}+\frac{n^2}{S^2}\right)S\varepsilon^4\le\frac{4n^2\varepsilon^4}{S}.
\end{aligned}
$$

Since $n\le 0.1\frac{S^{0.5}}{\varepsilon^2}$, we have $\|P-Q\|_{TV}\le\sqrt{2KL(P,Q)}\le\sqrt{0.08}\le 0.1$, which completes the proof. □

Lemma E.4. *For $\varepsilon^2 \le \frac{1}{2}$, we have*

$$G_\varepsilon(m) \le \frac{6m^2+m}{8}\varepsilon^4 + \frac{m}{2}\frac{\varepsilon^4}{1-\varepsilon^2} \le 2m^2\varepsilon^4.$$

Proof of Lemma E.4. Let $x_1, \dots, x_n \overset{i.i.d.}{\sim} \{-1, 1\}$, we have

$$G_\varepsilon(m) = -\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log\left(\prod_{i=1}^m(1-x_i\varepsilon) + \prod_{i=1}^m(1+x_i\varepsilon)\right)\right].$$

For convenience, we define $|\boldsymbol{x}| := |\sum_{i=1}^m x_i|$. Note that

$$\prod_{i=1}^m(1-x_i\varepsilon) + \prod_{i=1}^m(1+x_i\varepsilon) = \left((1-\varepsilon)^{|\boldsymbol{x}|} + (1+\varepsilon)^{|\boldsymbol{x}|}\right)(1-\varepsilon^2)^{\frac{m-|\boldsymbol{x}|}{2}}.$$

Thus

$$G_\varepsilon(m) = -\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log\left((1-\varepsilon)^{|\boldsymbol{x}|} + (1+\varepsilon)^{|\boldsymbol{x}|}\right)\right] - \frac{m-|\boldsymbol{x}|}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log(1-\varepsilon^2)\right].$$

For the first term, we have

$$\begin{aligned}
&-\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log\left((1-\varepsilon)^{|\boldsymbol{x}|} + (1+\varepsilon)^{|\boldsymbol{x}|}\right)\right] \\
\le& -\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log\left(1 + \frac{|\boldsymbol{x}|(|\boldsymbol{x}|-1)}{2}\varepsilon^2\right)\right] \\
\le& \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[-\frac{|\boldsymbol{x}|(|\boldsymbol{x}|-1)}{2}\varepsilon^2 + \frac{1}{2}\left(\frac{|\boldsymbol{x}|(|\boldsymbol{x}|-1)}{2}\varepsilon^2\right)^2\right] \\
\le& \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[-\frac{|\boldsymbol{x}|(|\boldsymbol{x}|-1)}{2}\varepsilon^2 + \frac{|\boldsymbol{x}|^4}{8}\varepsilon^4\right] \\
=& -\frac{m}{2}\varepsilon^2 + \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\frac{|\boldsymbol{x}|}{2}\right]\varepsilon^2 + \frac{6m^2+m}{8}\varepsilon^4.
\end{aligned}$$

For the second term, we have

$$-\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log(1-\varepsilon^2)\right] = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\log(1+\frac{\varepsilon^2}{1-\varepsilon^2})\right] \le \frac{\varepsilon^2}{1-\varepsilon^2} = \varepsilon^2 + \frac{\varepsilon^4}{1-\varepsilon^2}.$$

Thus $G_\varepsilon(m)$ only contains ε^4 terms, i.e.,

$$G_\varepsilon(m) \le \frac{6m^2+m}{8}\varepsilon^4 + \frac{m}{2}\frac{\varepsilon^4}{1-\varepsilon^2} \le 2m^2\varepsilon^4,$$

the last step is by assumption $\varepsilon^2 \le \frac{1}{2}$. □

Proof of Theorem E.2. In our case, since r is known and $|\mathcal{A}| = 1$, we can simplify the each data in D into the form of (s, s', h). Further since the probability transition matrix for $h = 1$ and $h = 3$ are known, below we will assume D only contains n pairs of $(s, s', 2)$, and we will call these states by $\{s_i\}_{i=1}^n$ and $\{s_i'\}_{i=1}^n$. Since $\|f_{c'} - \mathcal{T}^3_{c,\boldsymbol{\sigma}} f_{c'}\|^1_{2,\mu_2}$ and $\|f_{c'} - \mathcal{T}^3_{c,\boldsymbol{\sigma}} f_{c'}\|^1_{2,\mu_2}$ are constant for all c, c', we only need to consider $\|f_{c'} - \mathcal{T}^2_{c,\boldsymbol{\sigma}} f_{c'}\|^2_{2,\mu_3}$ as our loss.

Recall we define $\mu_2 \circ \mathbb{P}^\varepsilon_{c,\boldsymbol{\sigma}}$ as the joint distribution of (s, s'), where $s \sim \mu_2$ and $s' \sim \mathbb{P}_{c,\boldsymbol{\sigma}|s}$. Thus the dataset D can be viewed as sampled from $\mathbb{E}_{c,\boldsymbol{\sigma}}\left(\mu_2 \circ \mathbb{P}_{c,\boldsymbol{\sigma}}\right)^n$, i.e., D is sampled from a mixture of product measures.

By Lemma E.3, we know

$$\begin{aligned}
&\left\|\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{\varepsilon}_{1,\boldsymbol{\sigma}}\right)^n-\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{\varepsilon}_{-1,\boldsymbol{\sigma}}\right)^n\right\|_{TV}\\
\leq&\left\|\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{\varepsilon}_{1,\boldsymbol{\sigma}}\right)^n-\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{0}\right)^n\right\|_{TV}+\left\|\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{\varepsilon}_{-1,\boldsymbol{\sigma}}\right)^n-\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{0}\right)^n\right\|_{TV}\\
\leq&0.2.
\end{aligned}$$

Thus if we denote the distribution of $widetilde\mathfrak{A}(D)$ by X_c, where $D\sim\mathbb{E}_{\boldsymbol{\sigma}}\left(\mu_2\circ\mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n$ and $\boldsymbol{\sigma}\sim\{-1,1\}^S$, and $\widetilde{\mathfrak{A}}$ can be random, the above inequality implies $\mathbb{P}\left[X_{-1}\neq X_1\right]\leq 0.2$, and therefore we have

$$\begin{aligned}
\mathbb{E}_{c,\boldsymbol{\sigma}}\mathbb{E}_D\left[\mathcal{E}_{c,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right]=&\frac{1}{2}\left(\mathbb{E}_{\boldsymbol{\sigma},D}\left[\mathcal{E}_{1,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right]+\mathbb{E}_{\boldsymbol{\sigma},D}\left[\mathcal{E}_{-1,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right]\right)\\
=&\frac{1}{6}+\frac{\varepsilon^2}{24}\left(\mathbb{P}\left[X_1\neq 1\right]+\mathbb{P}\left[X_{-1}\neq -1\right]\right)\\
=&\frac{1}{6}+\frac{\varepsilon^2}{24}\left(\mathbb{P}\left[X_1\neq 1\right]+\mathbb{P}\left[X_{-1}\neq -1\right]+\mathbb{P}\left[X_1\neq X_{-1}\right]\right)-\frac{\varepsilon^2}{24}\mathbb{P}\left[X_1\neq X_{-1}\right]\\
\geq&\frac{1}{6}+\frac{\varepsilon^2}{24}-\frac{\varepsilon^2}{24}\mathbb{P}\left[X_1\neq X_{-1}\right]\\
\geq&\frac{1}{6}+\frac{\varepsilon^2}{24}-\frac{\varepsilon^2}{24}\times 0.2\\
=&\frac{1}{6}+\frac{\varepsilon^2}{30}\\
=&\mathbb{E}_{c,\boldsymbol{\sigma}}\left[\min_{c'\in\{-1,1\}}\mathcal{E}_{c,\boldsymbol{\sigma}}\left(f_{c'}\right)\right]+\frac{\varepsilon^2}{30}.
\end{aligned}\tag{70}$$

□

E.2. Proof of Theorem E.1

Now we will prove Theorem E.1 by slightly twisting the distribution of hard instances (MDPs) constructed in the previous subsection.

Proof of Theorem E.1. W.O.L.G, we can assume S is even and $S=2S'$ (o.w. we can just abandon one state.) The only modification from the previous lower bound with $C=2$ is now the distribution of σ is defined as the conditional distribution of P on $\sum_{i=1}^S=0$, i.e., $P'(\boldsymbol{\sigma})=P(\boldsymbol{\sigma}|\sum_{i=1}^S\sigma_i=0)$, where P is the uniform distribution on $\{-1,1\}^S$. The main idea is that the data distribution (i.e., distribution of (s,s')) shouldn't be very different even if we add this additional 'balancedness' restriction. We further define a metric d on $\{-1,1\}^S$. In detail, for $\boldsymbol{\sigma},\boldsymbol{\sigma}'\in\{-1,1\}^d$, we define $d(\boldsymbol{\sigma},\boldsymbol{\sigma}')=\frac{\sum_{i=1}^S|\sigma_i-\sigma_i'|}{2S}$. We have the following lemma:

Lemma E.5.

$$W_1^d(P,P')=\frac{1}{2S}\mathbb{E}_P\left|\sum_{i=1}^S\sigma_i\right|,\tag{71}$$

where $W_1^d(P,P')$ is defined as $\min_{\boldsymbol{\sigma}\sim P,\boldsymbol{\sigma}'\sim P'}\mathbb{E}[d(\boldsymbol{\sigma},\boldsymbol{\sigma}')]$.

By Cauchy Inequality, we have

$$W_1^d(P,P')=\frac{1}{2S}\mathbb{E}_P\left|\sum_{i=1}^S\sigma_i\right|\leq\frac{1}{2S}\sqrt{\mathbb{E}_P\left(\sum_{i=1}^S\sigma_i\right)^2}=\frac{1}{2\sqrt{S}}\tag{72}$$

Proof. For even S, we define B as the set of the "balanced" $\boldsymbol{\sigma}$, i.e., $B=\{\boldsymbol{\sigma}|\sum_{i=1}^S\sigma_i'=0\}$. For every $\boldsymbol{\sigma}\in\{-1,1\}^S$, we define $Q_{\boldsymbol{\sigma}}$ as the uniform distribution on $U_{\boldsymbol{\sigma}}=\{\boldsymbol{\sigma}'\mid d(\boldsymbol{\sigma},\boldsymbol{\sigma}')=\frac{|\sum_{i=1}^S\sigma_i|}{2S}\}\cap B$, i.e. $\boldsymbol{\sigma}'\in U_{\boldsymbol{\sigma}}$ if and only if $\boldsymbol{\sigma}'\in B$ and $d(\boldsymbol{\sigma},\boldsymbol{\sigma}')=\min_{\boldsymbol{\sigma}'\in B}d(\boldsymbol{\sigma},\boldsymbol{\sigma}')$.

Now we define $\Gamma(\boldsymbol{\sigma}, \boldsymbol{\sigma'}) = P(\boldsymbol{\sigma})Q_{\boldsymbol{\sigma}}(\boldsymbol{\sigma'})$. By definition the marginal distribution of Γ on $\boldsymbol{\sigma}$ is P. By symmetry, the marginal distribution of $\boldsymbol{\sigma'}$ is P'. Thus by definition of W_1,

$$W_1^d(P, P') \leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{\sigma'} \sim \Gamma} [d(\boldsymbol{\sigma}, \boldsymbol{\sigma'})] = \frac{1}{2S} \mathbb{E}_P \left| \sum_{i=1}^{S} \sigma_i \right|.$$

□

Lemma E.6.

$$\left\| \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n - \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right)^n \right\|_{TV} \leq C\varepsilon\sqrt{nd(\boldsymbol{\sigma}, \boldsymbol{\sigma'})}. \tag{73}$$

Proof. First, note that

$$\begin{aligned}
&KL\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}, \mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right) \\
=&KL\left(\mu_2, \mu_2\right) + \mathop{\mathbb{E}}_{i \sim \mu_2} \left[KL\left(\mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}(\cdot \mid i), \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}(\cdot \mid i)\right)\right] \\
=&0 + \mathop{\mathbb{E}}_{i \sim \mu_2} \left[KL\left(\mathbb{P}^{\varepsilon}_{c,\sigma_i}, \mathbb{P}^{\varepsilon}_{c,\sigma'_i}\right)\right] \\
=&\mathop{\mathbb{P}}_{i \sim \mu_2} \left[\sigma_i \neq \sigma'_i\right] \cdot \left(\frac{1+\varepsilon}{2} \log \frac{1+\varepsilon}{1-\varepsilon} + \frac{1-\varepsilon}{2} \log \frac{1-\varepsilon}{1+\varepsilon}\right) \\
=&\mathop{\mathbb{P}}_{i \sim \mu_2} \left[\sigma_i \neq \sigma'_i\right] \cdot \varepsilon \log \frac{1+\varepsilon}{1-\varepsilon} \\
=&\mathop{\mathbb{P}}_{i \sim \mu_2} \left[\sigma_i \neq \sigma'_i\right] \cdot \frac{2\varepsilon^2}{1-\varepsilon} \\
\leq&4d(\boldsymbol{\sigma}, \boldsymbol{\sigma'})\varepsilon^2.
\end{aligned} \tag{74}$$

Thus we have

$$\begin{aligned}
&\left\| \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n - \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right)^n \right\|_{TV} \\
\leq&\sqrt{2KL\left(\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n, \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right)^n\right)} \\
\leq&\sqrt{2nKL\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}, \mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right)} \\
\leq&\varepsilon\sqrt{8md(\boldsymbol{\sigma}, \boldsymbol{\sigma'})}.
\end{aligned} \tag{75}$$

□

Let $\Gamma(\boldsymbol{\sigma}, \boldsymbol{\sigma'})$ be the joint probabilistic distribution on $\{-1, 1\}^S \times \{-1, 1\}^S$ which attains the eq. (71). Therefore the marginal distribution of Γ is P and P'. And thus we have for any $c \in \{-1, 1\}$,

$$\begin{aligned}
&\| \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim P} \left[\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n\right] - \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim P'} \left[\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right)^n\right] \|_{TV} \\
\leq& \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{\sigma'} \sim \Gamma} \left[\| \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n - \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma'}}\right)^n \|_{TV}\right] \\
\leq& \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{\sigma'} \sim \Gamma} \left[\varepsilon\sqrt{8nd(\boldsymbol{\sigma}, \boldsymbol{\sigma'})}\right] \\
\leq&\varepsilon\sqrt{n \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, \boldsymbol{\sigma'} \sim \Gamma} \left[8d(\boldsymbol{\sigma}, \boldsymbol{\sigma'})\right]} \\
=&\varepsilon\sqrt{8nW_1^d(P, P')} \\
\leq&2\varepsilon n^{0.5} S^{-0.25}.
\end{aligned} \tag{76}$$

Therefore, when $n \leq \frac{\sqrt{S}}{400\varepsilon^2}$, for any $c \in \{-1, 1\}$,

$$\| \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim P} \left[\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n \right] - \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim P'} \left[\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}'}\right)^n \right] \|_{TV} \leq 0.1.$$

By Lemma E.3, we have

$$\| \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim P'} \left[\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{1,\boldsymbol{\sigma}'}\right)^n \right] - \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim P'} \left[\left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{-1,\boldsymbol{\sigma}'}\right)^n \right] \|_{TV} \leq 0.1 + 0.1 + 0.1 + 0.1 = 0.4.$$

Thus using the same argument in eq. (70), In detail, denote the distribution of $\mathfrak{A}(D)$ by X_c, where $D \sim \mathbb{E}_{\boldsymbol{\sigma}} \left(\mu_2 \circ \mathbb{P}^{\varepsilon}_{c,\boldsymbol{\sigma}}\right)^n$, $\boldsymbol{\sigma} \sim \{-1, 1\}^S$, the above inequality implies $\mathbb{P}\left[X_{-1} \neq X_1\right] \leq 0.4$, and therefore we have

$$\begin{aligned}
\mathbb{E}_{c,\boldsymbol{\sigma}}\mathbb{E}_D \left[\mathcal{E}_{c,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right] =& \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{\sigma},D}\left[\mathcal{E}_{1,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right] + \mathbb{E}_{\boldsymbol{\sigma},D}\left[\mathcal{E}_{-1,\boldsymbol{\sigma}}\left(\mathfrak{A}(D)\right)\right]\right) \\
=& \frac{1}{6} + \frac{\varepsilon^2}{24}\left(\mathbb{P}\left[X_1 \neq 1\right] + \mathbb{P}\left[X_{-1} \neq -1\right]\right) \\
=& \frac{1}{6} + \frac{\varepsilon^2}{24}\left(\mathbb{P}\left[X_1 \neq 1\right] + \mathbb{P}\left[X_{-1} \neq -1\right] + \mathbb{P}\left[X_1 \neq X_{-1}\right]\right) - \frac{\varepsilon^2}{24}\mathbb{P}\left[X_1 \neq X_{-1}\right] \\
\geq& \frac{1}{6} + \frac{\varepsilon^2}{24} - \frac{\varepsilon^2}{24} \times 0.4 \\
=& \frac{1}{6} + \frac{1}{40}\varepsilon^2 \\
=& \mathbb{E}_{c,\boldsymbol{\sigma}}\left[\min_{c' \in \{-1,1\}} \mathcal{E}_{c,\boldsymbol{\sigma}}\left(f_{c'}\right)\right] + \frac{\varepsilon^2}{40}.
\end{aligned} \tag{77}$$

□

F. Auxiliary Results

In this section, we prove some auxiliary lemmas. Appendix F.1 considers the relation between Bellman error and suboptimality in values (Lemma 3.2). Appendix F.2 provides a supporting lemma used in the proof of Theorem 5.5. Appendix F.3 presents a full version of Proposition 5.6.

F.1. Connections between Bellman error and suboptimality in value (Lemma 3.2)

In this part, we present several possible ways to connect Bellman error $\mathcal{E}(f)$ with the suboptimality gap $V_1^\star(s_1) - V_1^{\pi_f}(s_1)$.

Via concentrability coefficient

Lemma 3.2 (Bellman error to value suboptimality)**.** *Under Assumption 1, for any $f \in \mathcal{F}$, we have that ,*

$$V_1^\star(s_1) - V_1^{\pi_f}(s_1) \leq 2H\sqrt{C \cdot \mathcal{E}(f)}, \tag{3}$$

where C is the concentrability coefficient in Assumption 1.

Lemma 3.2 gives a feasible method to upper bound $V_1^\star(s_1) - V_1^{\pi_f}(s_1)$ with $\mathcal{E}(f)$ using the concentrability coefficient introduced in Assumption 1. We provide the proof of Lemma 3.2 below.

Proof of Lemma 3.2. The proof of Lemma 3.2 is analogous to Theorem 2 in (Xie & Jiang, 2020b). We place it here for the self-containedness of our paper. In discussions below, we omit the subscript h in policy π_{f_h} and simply write π_f to ease the notation. We first note that since π_f is greedy w.r.t f, therefore,

$$V_1^\star(s_1) - V_1^{\pi_f}(s_1) \leq V_1^\star(s_1) - f_1\big(s_1, \pi^\star(s_1)\big) + f_1\big(s_1, \pi_f(s_1)\big) - V_1^{\pi_f}(s_1). \tag{78}$$

Consider any policy π. Since $f_{H+1} = 0$ and $V_1^\pi(s_1) = \mathbb{E}\big[\sum_{h=1}^H r_h \,\big|\, s_1, \pi\big]$ by definition, we have

$$f_1\big(s_1, \pi(s_1)\big) - V_1^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^H \Big(f_h(s_h, a_h) - \mathbb{E}_h^\pi\big[r_h + f_{h+1}\big(s_{h+1}, a_{h+1}\big) \,\big|\, s_h, a_h\big]\Big) \,\middle|\, s_1, \pi\right].$$

Therefore, combined with the fact π_f is the greedy policy w.r.t. f, we can show that

$$f_1\big(s_1, \pi^\star(s_1)\big) - V_1^\star(s_1) \geq \mathbb{E}\left[\sum_{h=1}^H \big(f_h - \mathcal{T}_h^\star f_{h+1}\big)(s_h, a_h) \,\middle|\, s_1, \pi^\star\right], \tag{79}$$

$$f_1(s_1, \pi_f(s_1)) - V_1^{\pi_f}(s_1) = \mathbb{E}\left[\sum_{h=1}^H \big(f_h - \mathcal{T}_h^\star f_{h+1}\big)(s_h, a_h) \,\middle|\, s_1, \pi_f\right]. \tag{80}$$

Plugging eqs. (79) and (80) into eq. (78) yields

$$V_1^\star(s_1) - V_1^{\pi_f}(s_1) \leq -\mathbb{E}\left[\sum_{h=1}^H \big(f_h - \mathcal{T}_h^\star f_{h+1}\big)(s_h, a_h) \,\middle|\, s_1, \pi^\star\right] + \mathbb{E}\left[\sum_{h=1}^H \big(f_h - \mathcal{T}_h^\star f_{h+1}\big)(s_h, a_h) \,\middle|\, s_1, \pi_f\right].$$

Under Assumption 1, by Cauchy-Swartz inequality, it holds that for any policy π:

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{h=1}^H \big(f_h - \mathcal{T}_h^\star f_{h+1}\big)(s_h, a_h) \,\middle|\, s_1, \pi\right]\right| \leq& \sqrt{H \sum_{h=1}^H \mathbb{E}\left[\big(f_h - \mathcal{T}_h^\star f_{h+1}\big)^2(s_h, a_h) \,\middle|\, s_1, \pi\right]} \\ \leq& \sqrt{C} H \sqrt{\frac{1}{H}\sum_{h=1}^H \big\|f_h - \mathcal{T}_h^\star f_{h+1}\big\|_{\mu_h}^2},\end{aligned}$$

which finishes the proof. □

Via a weaker concentrability assumption We observe that Lemma 3.2 does not necessarily need an assumption as strong as Assumption 1. In fact, the inequality $V_1^\star(s_1) - V_1^{\pi_f}(s_1) \leq 2H\sqrt{C \cdot \mathcal{E}(f)}$ still holds if

$$\mathbb{E}\big[\big(f_h - \mathcal{T}_h^\star f_{h+1}\big)(s_h, a_h) \,\big|\, s_1, \pi\big] \leq \sqrt{C}\big\|f_h - \mathcal{T}_h^\star f_{h+1}\big\|_{\mu_h} \qquad \text{for } \pi = \pi^\star \text{ or } \pi = \pi_f \text{ for } f \in \mathcal{F}. \tag{81}$$

If the function class $\mathcal{F}$ and $\mathcal{T}^\star\mathcal{F} = \big\{\mathcal{T}^\star f = (\mathcal{T}_1^\star f_2, \ldots, \mathcal{T}_H^\star f_{H+1}) \,\big|\, f \in \mathcal{F}\big\}$ have good structures, we may have a tighter estimate of the required C. For illustrative purpose, we take a simple example where $\mathcal{F}_h$ is a subset of a finite dimensional linear space and $\mathcal{T}_h^\star f_{h+1} \in \mathcal{F}_h$ for any $f_{h+1} \in \mathcal{F}_{h+1}$. Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a basis of $\mathcal{F}_h$ with $\|\phi(s,a)\|_2 \leq 1$. Define $\Sigma_h := \mathbb{E}_{\mu_h}[\phi\phi^\top] \in \mathbb{R}^{d\times d}$. For any $f = w^\top\phi \in \mathcal{F}_h$, $\|f\|_\infty \leq \|w\|_2 \leq \|\Sigma_h^{\frac{1}{2}} w\|_2 \sqrt{1/\lambda_{\min}(\Sigma_h)} = \|f\|_{\mu_h}\sqrt{1/\lambda_{\min}(\Sigma_h)}$. Therefore, eq. (81) holds for $C = \max_{h\in[H]}\{1/\lambda_{\min}(\Sigma_h)\}$.

F.2. Proof of Supporting Lemmas in Minimax Algorithm Analysis

Lemma F.1. *Suppose Assumption 4 holds. Denote $f^\dagger := \min_{f\in\mathcal{F}} \mathcal{E}(f)$. For $h \in [H]$, it holds that*

$$\big\|f_h - f_h^\dagger\big\|_{\rho_h}^2 \leq \widetilde{C} H(H-h+1)\big\|\big(f - \mathcal{T}^\star f\big) - \big(f^\dagger - \mathcal{T}^\star f^\dagger\big)\big\|_\mu^2 \quad \textit{for } \rho_h = \mu_h \textit{ or } \nu_h \times \textit{Unif}(\mathcal{A}), \tag{82}$$

$$\textit{and} \qquad \big\|\mathcal{T}_h^\star f_{h+1} - \mathcal{T}_h^\star f_{h+1}^\dagger\big\|_{\mu_h}^2 \leq \widetilde{C} H(H-h)\big\|\big(f - \mathcal{T}^\star f\big) - \big(f^\dagger - \mathcal{T}^\star f^\dagger\big)\big\|_\mu^2. \tag{83}$$

Proof. 1. Let π_f be the greedy policy associated with $f \in \mathcal{F}$. Since $f_{H+1} = f_{H+1}^\dagger = 0$, we have

$$\begin{aligned}f_h(s,a) - f_h^\dagger(s,a) =& \mathbb{E}\left[\sum_{\tau=h}^H \Big[\Big(f_\tau\big(s_\tau, a_\tau\big) - \mathbb{E}\big[r_\tau + f_{\tau+1}\big(s_{\tau+1}, a_{\tau+1}\big) \,\big|\, s_\tau, a_\tau\big]\Big)\right. \\ &\left. - \Big(f_\tau^\dagger\big(s_\tau, a_\tau\big) - \mathbb{E}\big[r_\tau + f_{\tau+1}^\dagger\big(s_{\tau+1}, a_{\tau+1}\big) \,\big|\, s_\tau, a_\tau\big]\Big)\Big] \,\middle|\, s_h = s, a_h = a, \pi_f\right].\end{aligned} \tag{84}$$

Note that

$$\begin{aligned}\mathbb{E}\big[r_\tau + f_{\tau+1}\big(s_{\tau+1}, \pi_{f_{\tau+1}}(s_{\tau+1})\big) \,\big|\, s_\tau, a_\tau\big] &= \mathcal{T}_\tau^\star f_{\tau+1}(s_\tau, a_\tau),\\ \mathbb{E}\big[r_\tau + f_{\tau+1}^\dagger\big(s_{\tau+1}, \pi_{f_{\tau+1}}(s_{\tau+1})\big) \,\big|\, s_\tau, a_\tau\big] &\le \mathcal{T}_\tau^\star f_{\tau+1}^\dagger(s_\tau, a_\tau).\end{aligned} \tag{85}$$

Combining eqs. (84) and (85), we learn that

$$f_h(s,a) - f_h^\dagger(s,a) \le \mathbb{E}\left[\sum_{\tau=h}^{H}\Big[\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\Big](s_\tau, a_\tau)\,\middle|\, s_h = s, a_h = a, \pi_f\right]. \tag{86}$$

By symmetry, it also holds that

$$f_h^\dagger(s,a) - f_h(s,a) \le \mathbb{E}\left[\sum_{\tau=h}^{H}\Big[\big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big) - \big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big)\Big](s_\tau, a_\tau)\,\middle|\, s_h = s, a_h = a, \pi_{f^\dagger}\right]. \tag{87}$$

Under Assumption 4, by Cauchy-Swartz inequality, for any policy π:

$$\begin{aligned}&\mathbb{E}_{(s_h,a_h)\sim\mu_h}\left(\mathbb{E}\left[\sum_{\tau=h}^{H}\Big[\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\Big](s_\tau, a_\tau)\,\middle|\, s_h, a_h, \pi\right]\right)^2\\ \le&(H-h+1)\mathbb{E}\left[\sum_{\tau=h}^{H}\Big[\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\Big]^2(s_\tau, a_\tau)\,\middle|\, (s_h,a_h)\sim\mu_h, \pi\right]\\ \le&\widetilde{C}(H-h+1)\sum_{\tau=h}^{H}\big\|\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\big\|_{\mu_\tau}^2\\ \le&\widetilde{C}H(H-h+1)\big\|\big(f - \mathcal{T}^\star f\big) - \big(f^\dagger - \mathcal{T}^\star f^\dagger\big)\big\|_\mu^2.\end{aligned}$$

Therefore, eqs. (86) and (87) imply eq. (82).

2. We now consider $\big\|\mathcal{T}_h^\star f_{h+1} - \mathcal{T}_h^\star f_{h+1}^\dagger\big\|_{\mu_h}$. Take $\widetilde{\pi}_{h+1}(s) := \arg\max_{a\in\mathcal{A}}\big\{f_{h+1}(s,a) \vee f_{h+1}^\dagger(s,a)\big\}$. Then we have

$$\big|V_{f_{h+1}}(s) - V_{f_{h+1}^\dagger}(s)\big| \le \big|f_{h+1}\big(s, \widetilde{\pi}_{h+1}(s)\big) - f_{h+1}^\dagger\big(s, \widetilde{\pi}_{h+1}(s)\big)\big|.$$

It follows that

$$\begin{aligned}\big\|\mathcal{T}_h^\star f_{h+1} - \mathcal{T}_h^\star f_{h+1}^\dagger\big\|_{\mu_h} =&\big\|\mathbb{E}\big[V_{f_{h+1}}(s') - V_{f_{h+1}^\dagger}(s') \,\big|\, s, a\big]\big\|_{\mu_h}\\ \le&\big\|V_{f_{h+1}} - V_{f_{h+1}^\dagger}\big\|_{\nu_h} \le \big\|f_{h+1} - f_{h+1}^\dagger\big\|_{\nu_h\times\widetilde{\pi}_{h+1}}.\end{aligned}$$

Similar to eqs. (86) and (87), we find that

$$\begin{aligned}&\big\|f_{h+1} - f_{h+1}^\dagger\big\|_{\nu_h\times\widetilde{\pi}_{h+1}}^2\\ \le&\max_{\pi=\pi_f \text{ or } \pi_{f^\dagger}}\mathbb{E}_{(s_{h+1},a_{h+1})\sim\nu_h\times\widetilde{\pi}_{h+1}}\left(\mathbb{E}\left[\sum_{\tau=h+1}^{H}\Big[\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\Big](s_\tau, a_\tau)\,\middle|\, s_{h+1}, a_{h+1}, \pi\right]\right)^2\\ \le&(H-h)\max_{\pi=\pi_f \text{ or } \pi_{f^\dagger}}\mathbb{E}_{(s_{h+1},a_{h+1})\sim\nu_h\times\widetilde{\pi}_{h+1}}\mathbb{E}\left[\sum_{\tau=h+1}^{H}\Big[\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\Big]^2(s_\tau, a_\tau)\,\middle|\, s_{h+1}, a_{h+1}, \pi\right]\\ \le&\widetilde{C}(H-h)\sum_{\tau=h+1}^{H}\big\|\big(f_\tau - \mathcal{T}_\tau^\star f_{\tau+1}\big) - \big(f_\tau^\dagger - \mathcal{T}_\tau^\star f_{\tau+1}^\dagger\big)\big\|_{\mu_\tau}^2 \le \widetilde{C}H(H-h)\big\|\big(f - \mathcal{T}^\star f\big) - \big(f^\dagger - \mathcal{T}^\star f^\dagger\big)\big\|_\mu^2.\end{aligned}$$

Therefore, we conclude that $\big\|\mathcal{T}_h^\star f_{h+1} - \mathcal{T}_h^\star f_{h+1}^\dagger\big\|_{\mu_h}^2 \le \widetilde{C}H(H-h)\big\|\big(f - \mathcal{T}^\star f\big) - \big(f^\dagger - \mathcal{T}^\star f^\dagger\big)\big\|_\mu^2$. □

F.3. Proof of Proposition 5.6

Lemma F.2 (Full version of Proposition 5.6)**.** *Let $\widetilde{\mathcal{F}}_{h+1}$ be any subset of $\mathcal{F}_{h+1}$. We have the following inequality,*

$$\mathcal{R}_n^{\mu_h}\left(\left\{\mathcal{T}_h^\star f_{h+1} \,\middle|\, f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}\right\}\right) \le \mathcal{R}_n^{\nu_h}\left(V_{\widetilde{\mathcal{F}}_{h+1}}\right) \le \sqrt{2}A\mathcal{R}_n^{\nu_h\times \mathit{Unif}(\mathcal{A})}\left(\widetilde{\mathcal{F}}_{h+1}\right).$$

Proof. 1. Due to the symmetry of Rademacher random variables,

$$\begin{aligned}\mathcal{R}_n^{\mu_h}\left(\left\{\mathcal{T}_h^\star f_{h+1} \,\middle|\, f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}\right\}\right) =&\mathcal{R}_n^{\mu_h}\left(\left\{r_h+\mathbb{E}\left[V_{f_{h+1}}(s_h')\,\middle|\, s_h,a_h\right] \,\middle|\, f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}\right\}\right)\\ =&\mathcal{R}_n^{\mu_h}\left(\left\{\mathbb{E}\left[V_{f_{h+1}}(s_h')\,\middle|\, s_h,a_h\right] \,\middle|\, f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}\right\}\right).\end{aligned}$$

By definition,

$$\mathcal{R}_n^{\mu_h}\left(\left\{\mathbb{E}\left[V_{f_{h+1}}(s_h')\,\middle|\, s_h,a_h\right] \,\middle|\, f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}\right\}\right) = \mathbb{E}_{\mu_h}\left[\sup_{f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}}\sum_{k=1}^n \sigma_k\mathbb{E}\left[V_{f_{h+1}}(s_{k,h}')\,\middle|\, s_{k,h},a_{k,h}\right]\right].$$

Switching the order of supremum and the inner expectation, we derive that

$$\mathcal{R}_n^{\mu_h}\left(\left\{\mathbb{E}\left[V_{f_{h+1}}(s_h')\,\middle|\, s_h,a_h\right] \,\middle|\, f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}\right\}\right) \le \mathbb{E}\left[\sup_{f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}}\sum_{k=1}^n \sigma_k V_{f_{h+1}}(s_{k,h}')\right] = \mathcal{R}_n^{\nu_h}\left(V_{\widetilde{\mathcal{F}}_{h+1}}\right).$$

2. For notational convenience, let $\mathcal{A}=[A]$. Consider a vector function $\vec{f}_{h+1}:\mathcal{S}\to\mathbb{R}^A$ defined as $\vec{f}_{h+1}(s) := \big(f_{h+1}(s,1), f_{h+1}(s,2),\dots,f_{h+1}(s,A)\big)^\top\in\mathbb{R}^A$. Then for any $f_{h+1},f_{h+1}'\in\mathcal{F}_{h+1}$, $\left|V_{f_{h+1}}(s)-V_{f_{h+1}'}(s)\right| \le \|\vec{f}_{h+1}-\vec{f}_{h+1}'\|_\infty \le \|\vec{f}_{h+1}-\vec{f}_{h+1}'\|_2$, *i.e.* the mapping $\mathbb{R}^A\ni\vec{f}_{h+1}(s)\mapsto V_{f_{h+1}}(s)$ is 1-Lipschitz. By Lemma G.7, we have

$$\mathcal{R}_n^{\nu_h}\left(V_{\widetilde{\mathcal{F}}_{h+1}}\right) \le \sqrt{2}\mathbb{E}\left[\sup_{f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}}\sum_{k=1}^n\sum_{a\in\mathcal{A}}\sigma_{k,a}f_{h+1}(s_k',a)\right],$$

where $s_1',s_2',\dots,s_n'$ are *i.i.d.* samples generated from ν_h. Let $a_1',a_2',\dots,a_n'\in\mathcal{A}$ be random variables such that $\mathbb{P}(a_k'=a\,|\,s_k')=A^{-1}$ for $a\in\mathcal{A}$. It follows that

$$\begin{aligned}&\mathbb{E}\left[\sup_{f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}}\sum_{k=1}^n\sum_{a\in\mathcal{A}}\sigma_{k,a}f_{h+1}(s_k',a)\right] \le A\mathbb{E}\left[\frac{1}{A}\sum_{a\in\mathcal{A}}\sup_{f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}}\sum_{k=1}^n\sigma_{k,a}f_{h+1}(s_k',a)\right]\\ =&A\mathbb{E}\left[\sup_{f_{h+1}\in\widetilde{\mathcal{F}}_{h+1}}\sum_{k=1}^n\sigma_{k,a_k'}f_{h+1}(s_k',a_k')\right] = A\mathcal{R}_n^{\nu_h\times\mathrm{Unif}(\mathcal{A})}\left(\widetilde{\mathcal{F}}_{h+1}\right).\end{aligned}$$

Therefore, $\mathcal{R}_n^{\nu_h}\left(V_{\widetilde{\mathcal{F}}_{h+1}}\right) \le \sqrt{2}A\mathcal{R}_n^{\nu_h\times\mathrm{Unif}(\mathcal{A})}\left(\widetilde{\mathcal{F}}_{h+1}\right)$. □

G. Useful Results for (Local) Rademacher Complexity

In this section, we sumarize some useful results for (local) Rademacher complexity that are used throughout our analysis.

G.1. Concentration with Rademacher Complexity

Lemma G.1 below shows some uniform concentration inequalities with Rademacher complexity.

Lemma G.1. *Let $\mathcal{F}$ be a class of functions with ranges in $[a,b]$. With probability at least $1-\delta$,*

$$Pf \le P_nf + 2\mathcal{R}_n(\mathcal{F}) + (b-a)\sqrt{\frac{2\log(2/\delta)}{n}}, \qquad \textit{for any } f\in\mathcal{F}.$$

Also, with probability at least $1-\delta$,

$$P_nf \le Pf + 2\mathcal{R}_n(\mathcal{F}) + (b-a)\sqrt{\frac{2\log(2/\delta)}{n}}, \qquad \textit{for any } f\in\mathcal{F}.$$

Proof. Consider the empirical process $\sup_{f\in\mathcal{F}}(Pf - P_nf)$. By McDiarmid's inequality, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\big(Pf - P_nf\big) \le \mathbb{E}\sup_{f\in\mathcal{F}}\big(Pf - P_nf\big) + (b-a)\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{88}$$

The basic property of Rademacher complexity ensures that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\big(Pf - P_nf\big) \le 2\mathcal{R}_n(\mathcal{F}). \tag{89}$$

Combining eqs. (88) and (89), we finish the proof. □

G.2. Concentration with Local Rademacher complexity

In this part, we present some auxiliary results regarding local Rademacher complexity. In particular, Lemma G.2 guarantees the well-definedness of critical radius, Theorem G.3 provides concentration inequalities and Lemma G.5 gives some useful properties of sub-root functions.

G.2.1. Well-definedness of critical radius

Recall that in Definition 2.3, the critical radius $r^\star$ of local Rademacher complexity $\mathcal{R}_n^\rho(\{f\in\mathcal{F}\mid T(f)\le r\})$ is defined as the possitive fixed point of some sub-root functions $\psi(r)$. The following Lemma G.2 ensures that $r^\star$ exists and is unique.

Lemma G.2 (Lemma 3.2 in Bartlett et al. (2005))**.** *If $\psi:[0,\infty)\to[0,\infty)$ is a nontrivial sub-root function, then it is continuous on $[0,\infty)$ and the equation $\psi(r)=r$ has a unique positive solution $r^\star$. Moreover, for all $r>0$, $r\ge\psi(r)$ if and only if $r^\star\le r$.*

G.2.2. Concentration inequalities

Throughout the paper, we use Theorem G.3 below to prove uniform concentration with local Rademacher complexity. Theorem G.3 is a variant of Theorem 3.3 in Bartlett et al. (2005).

Theorem G.3 (Corollary of Theorem 3.3 in Bartlett et al. (2005))**.** *Let $\mathcal{F}$ be a class of functions with ranges in $[a,b]$ and assume that there are some functional $T:\mathcal{F}\to\mathbb{R}^+$ and some constants B and η such that for every $f\in\mathcal{F}$, $\mathrm{Var}[f]\le T(f)\le B(Pf+\eta)$. Let ψ be a sub-root function and let $r^\star$ be the fixed point of ψ. Assume that ψ satisfies, for any $r\ge r^\star$, $\psi(r)\ge B\mathcal{R}_n\big(\{f\in\mathcal{F}\,\big|\,T(f)\le r\}\big)$. Then for any $\theta>1$, with probability at least $1-\delta$,*

$$Pf \le \frac{\theta}{\theta-1}P_nf + \frac{c_1\theta}{B}r^\star + \big(c_2(b-a)+c_3B\theta\big)\frac{\log(1/\delta)}{n} + \frac{\eta}{\theta-1}, \qquad \textit{for any } f\in\mathcal{F}. \tag{90}$$

Also, with probability at least $1-\delta$,

$$P_nf \le \frac{\theta+1}{\theta}Pf + \frac{c_1\theta}{B}r^\star + \big(c_2(b-a)+c_3B\theta\big)\frac{\log(1/\delta)}{n} + \frac{\eta}{\theta}, \qquad \textit{for any } f\in\mathcal{F}.$$

Here, $c_1,c_2,c_3>0$ are some universal constants.

Proof. Theorem G.3 is proved in the same way as the first part of Theorem 3.3 in Bartlett et al. (2005), by applying the following Lemma G.4 instead of Lemma 3.8 in Bartlett et al. (2005). □

Given a class $\mathcal{F}$, $\lambda>1$ and $r>0$, let $w(f):=\min\big\{r\lambda^k\,\big|\,k\in\mathbb{N}, r\lambda^k\ge T(f)\big\}$ and set $\mathcal{G}_r:=\big\{\frac{r}{w(f)}f\,\big|\,f\in\mathcal{F}\big\}$. Define $V_r^+:=\sup_{g\in\mathcal{G}_r}Pg-P_ng$ and $V_r^-:=\sup_{g\in\mathcal{G}_r}P_ng-Pg$.

Lemma G.4 (Corollary of Lemma 3.8 in Bartlett et al. (2005))**.** *Assume that there is a constant $B>0$ such that for every $f\in\mathcal{F}$, $T(f)\le B(Pf+\eta)$. Fix $\theta>1$, $\lambda>0$ and $r>0$. If $V_r^+\le\frac{r}{\lambda B\theta}$, then $Pf\le\frac{\theta}{\theta-1}P_nf+\frac{r}{\lambda B\theta}+\frac{\eta}{\theta-1}$. Also, if $V_r^-\le\frac{r}{\lambda B\theta}$, then $P_nf\le\frac{\theta+1}{\theta}Pf+\frac{r}{\lambda B\theta}+\frac{\eta}{\theta}$.*

Proof. When $V_r^+\le\frac{r}{\lambda B\theta}$, following the same reasoning as Lemma 3.8 in Bartlett et al. (2005), we derive that $Pf\le P_nf+\theta^{-1}(Pf+\eta)$ under the modified condition $T(f)\le B(Pf+\eta)$. It immediately implies the first statement. Similarly, the second part is proved by showing that $P_nf\le Pf+\theta^{-1}(Pf+\eta)$. □

G.2.3. PROPERTIES OF SUB-ROOT FUNCTIONS

We apply the following Lemma G.5 to simplify the forms of critical radii.

Lemma G.5. *If $\psi:[0,\infty)\to[0,\infty)$ is a nontrivial sub-root function and $r^\star$ is its positive fixed point, then*

1. *$\psi(r)\le\sqrt{r^\star r}$ for any $r\ge r^\star$.*
2. *For any $c>0$, $\widetilde{\psi}(r):=c\psi(c^{-1}r)$ is sub-root and its positive fixed point $\widetilde{r}^\star$ satisfies $\widetilde{r}^\star=cr^\star$.*
3. *For any $C>0$, $\widetilde{\psi}(r):=C\psi(r)$ is sub-root and its positive fixed point $\widetilde{r}^\star$ satisfies $\widetilde{r}^\star\le(C^2\vee 1)r^\star$.*
4. *For any $\Delta r>0$, $\widetilde{\psi}(r):=\psi(r+\Delta r)$ is sub-root and its positive fixed point $\widetilde{r}^\star$ satisfies $\widetilde{r}^\star\le r^\star+\sqrt{r^\star\Delta r}$.*

If $\psi_i{:}[0,\infty)\to[0,\infty), i=1,\ldots,n$ are nontrivial sub-root functions and $r_i^\star$ is the positive fixed point of ψ_i, then

5. *$\widetilde{\psi}(r)=\sum_{i=1}^n\psi_i(r)$ is sub-root and its positive fixed point $\widetilde{r}^\star$ satisfies $\widetilde{r}^\star\le\big(\sum_{i=1}^n\sqrt{r_i^\star}\big)^2$.*

Proof. 1. Since ψ is a sub-root function, we have $\frac{\psi(r)}{\sqrt{r}}\le\frac{\psi(r^\star)}{\sqrt{r^\star}}$ for any $r\ge r^\star$. Note that $r^\star>0$ is the fixed point and $\frac{\psi(r^\star)}{\sqrt{r^\star}}=\sqrt{r^\star}$. Therefore, $\psi(r)\le\sqrt{r^\star r}$ for $r\ge r^\star$.

2. It is evident that $\widetilde{\psi}$ is sub-root. Additionally, if $r\ge cr^\star$, then by Lemma G.2, we have $\widetilde{\psi}(r)=c\psi(c^{-1}r)\le c(c^{-1}r)=r$. In contrast, if $0<r<cr^\star$, then $\widetilde{\psi}(r)=c\psi(c^{-1}r)>c(c^{-1}r)=r$. To this end, we can conclude that $\widetilde{r}^\star=cr^\star$.

3. We use part 1 and derive that if $\widetilde{r}^\star\ge r^\star$ then $\widetilde{r}^\star=\widetilde{\psi}(\widetilde{r}^\star)=C\psi(\widetilde{r}^\star)\le C\sqrt{r^\star\widetilde{r}^\star}$, which further implies $\widetilde{r}^\star\le C^2r^\star$. Therefore, $\widetilde{r}^\star\le(C^2\vee 1)r^\star$.

4. If $\widetilde{r}^\star+\Delta r\ge r^\star$, then we have $\widetilde{r}^\star=\widetilde{\psi}(\widetilde{r}^\star)=\psi(\widetilde{r}^\star+\Delta r)\le\sqrt{r^\star(\widetilde{r}^\star+\Delta r)}$ due to part 1. It follows that $\widetilde{r}^\star\le\frac{1}{2}\big(r^\star+\sqrt{(r^\star)^2+4r^\star\Delta r}\big)\le r^\star+\sqrt{r^\star\Delta r}$.

5. If $\widetilde{r}^\star\ge\max_{i\in[n]}r_i^\star$, then we apply part 1 and obtain $\widetilde{r}^\star=\widetilde{\psi}(\widetilde{r}^\star)=\sum_{i=1}^n\psi_i(\widetilde{r}^\star)\le\sum_{i=1}^n\sqrt{r_i^\star\widetilde{r}^\star}$. Hence, $\widetilde{r}^\star\le\big(\sum_{i=1}^n\sqrt{r_i^\star}\big)^2$. □

G.3. Contraction property of Rademacher complexity

Our analyses use contraction properties of Rademacher complexity. See Lemmas G.6 and G.7.

Lemma G.6 (Contraction property of Rademacher complexity, Ledoux & Talagrand (2013), Theorem A.6 in Bartlett et al. (2005))**.** *Suppose $\mathcal{F}\subseteq\{f:\mathcal{X}\to\mathbb{R}\}$. Let $\phi:\mathbb{R}\to\mathbb{R}$ be a contraction such that $|\phi(x)-\phi(y)|\le|y-y'|$ for any $y,y'\in\mathbb{R}$. Then for any $X_1,X_2,\ldots,X_n\in\mathcal{X}$,*

$$\widehat{\mathcal{R}}_X(\phi\circ\mathcal{F})=\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i\phi\big(f(X_i)\big)\right]\le\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_if(X_i)\right]=\widehat{\mathcal{R}}_X(\mathcal{F}).$$

Lemma G.7 (Vector-form contraction property of Rademacher complexity, Maurer (2016))**.** *Suppose $\boldsymbol{\mathcal{F}}$ is a collection of vector-valued functions $\boldsymbol{f}:\mathcal{X}\to\mathbb{R}^d$ and $h:\mathbb{R}^d\to\mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm, i.e. $\big|h(y)-h(y')\big|\le L\|y-y'\|_2$ for any $y,y'\in\mathbb{R}^d$. Then for any $X_1,X_2,\ldots,X_n\in\mathcal{X}$,*

$$\begin{aligned}\widehat{\mathcal{R}}_X(h\circ\boldsymbol{\mathcal{F}})=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^n\sigma_ih\big(\boldsymbol{f}(X_i)\big)\right]\\ \le&\sqrt{2}L\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d\sigma_{i,j}f_j(X_i)\right]\le\sqrt{2}L\sum_{j=1}^d\widehat{\mathcal{R}}_X\big(\{f_j\,|\,\boldsymbol{f}\in\boldsymbol{\mathcal{F}}\}\big).\end{aligned}$$