Supplement to "Adaptive Approximate Policy Iteration"

In Section A, we present the detailed proofs of main results. In Section B, the linear value function approximation is considered. In Section C, some supporting lemmas are included.

A Proofs of main results

A.1 Proof of Theorem 4.5: main result

We combine the decomposition (4.1), (5.2) and (5.3) together and utilize the results in Lemmas 5.2, 5.3 and 5.7. Then we have

$$R_T \lesssim \underbrace{Kt_{\text{mix}} + 4Kt_{\text{mix}}\sqrt{2\tau\log(T/\delta)}}_{\text{Lemma 5.2}} + \underbrace{\widetilde{C}T\sqrt{\frac{\log(1/\delta)}{\tau}} + T\varepsilon_0}_{\text{Lemma 5.3}}$$
$$+ \underbrace{\frac{\tau t_{\max}^4\log_2^4(K)}{\mu_{\min}^*} + T\Big(\frac{\widetilde{C}^2\log(1/\delta)}{\tau} + \varepsilon_0^2\Big)}_{\text{Lemma 5.7}}.$$

We choose $\delta = 1/T$ and ignore any universal constant and logarithmic factor in the following. Since $K = T/\tau$, it holds that

$$\begin{aligned} R_T &\overset{\log}{\lesssim} t_{\text{mix}}K\tau^{1/2} + \widetilde{C}T\tau^{-1/2} + \frac{t_{\text{mix}}^4\tau}{\mu_{\min}^*} + T(\varepsilon_0^2 + \varepsilon_0) \\ &\overset{\log}{\lesssim} t_{\text{mix}}\widetilde{C}\tau^{-1/2}T + \frac{t_{\text{mix}}^4\tau}{\mu_{\min}^*} + T(\varepsilon_0^2 + \varepsilon_0), \end{aligned}$$

with probability at least $1 - 1/T$. With a little abuse of notations, we re-define $\varepsilon_0 = \varepsilon_0^2 + \varepsilon_0$. By optimizing τ such that the first two term above is equal, i.e., $t_{\text{mix}}\widetilde{C}\tau^{-1/2}T = t_{\text{mix}}^4\tau/\mu_{\min}^*$, we choose $\tau = (\widetilde{C}\mu_{\min}^*/t_{\text{mix}}^3)^{2/3}T^{2/3}$. Overall, we reach the final regret bound,

$$R_T = \widetilde{\mathcal{O}}\left(t_{\text{mix}}^2\rho^{1/3}\widetilde{C}^{2/3}T^{2/3} + T\varepsilon_0\right),$$

where $\rho = \max_\pi \max_{x:\mu_\pi(x)\neq 0}(1/\mu_\pi(x))$. This ends the proof.

■

A.2 Proof of Lemma 5.4: adaptive optimistic FTRL (AO-FTRL)

Lemma 5.4 states that the cumulative regret for AO-FTRL is upper-bounded by

$$R_T \le \Big(\frac{8}{\eta} + \eta\mathcal{R}(f^*)\Big)\sqrt{\sum_{t=2}^T \|q_t - M_t\|_*^2 - \sum_{t=1}^T \frac{\eta_t}{4}\|f_t - f_{t+1}\|^2 + g},$$

where $g = \langle M_{T+1}, f^* - f_{T+1}\rangle + \|q_1\|_*^2/\eta_1$.

First, at each round t, AO-FTRL has the form of

$$\begin{aligned} f_{t+1} &= \operatorname*{argmin}_{f\in\mathcal{F}}\langle f, \sum_{s=1}^t q_s + M_{t+1}\rangle + \eta_t\mathcal{R}(f) \\ &= \operatorname*{argmin}_{f\in\mathcal{F}}\langle f, \sum_{s=1}^t q_s\rangle + \sum_{s=1}^t\langle M_{s+1} - M_s, f\rangle + \eta_t\mathcal{R}(f), \end{aligned}$$

where $\eta_1 \le \cdots \le \eta_t$ are data-dependent learning rates. For simplicity, we assume $\eta_0 = 0$. For $s = 1, \ldots, t$, we define

$$h_s(f) = \langle M_{s+1} - M_s, f\rangle + (\eta_s - \eta_{s-1})\mathcal{R}(f). \tag{A.1}$$

We define $h_0(f) = 0$ for all $f \in \mathcal{F}$ and $h_{1:t}(f) = \sum_{s=1}^{t} h_s(f) = \langle M_{t+1}, f\rangle + \eta_t R(f)$. Since $\mathcal{R}(f)$ is 1-strongly convex with respect to norm $\|\cdot\|$, $h_s(f)$ is $(\eta_s - \eta_{s-1})$-strongly-convex with respect to $\|\cdot\|$. Then we could rewrite the AO-FTRL update as

$$f_{t+1} = \operatorname*{argmin}_{f\in\mathcal{F}} \langle f, \sum_{s=1}^{t} q_s\rangle + \sum_{s=1}^{t} h_s(f).$$

Second, let us define the forward linear regret R_T^+ as:

$$R_T^+ = \sum_{t=1}^{T} \big\langle q_t, f_{t+1} - f^*\big\rangle.$$

One could interpret R_T^+ as a cheating regret since it uses prediction f_{t+1} at round t. We decompose the cumulative regret based on the forward linear regret as follows,

$$R_T = \sum_{t=1}^{T}\langle q_t, f_t\rangle - \sum_{t=1}^{T}\langle q_t, f^*\rangle = R_T^+ + \sum_{t=1}^{T}\langle q_t, f_t - f_{t+1}\rangle. \tag{A.2}$$

The second term in the right side captures the regret by the algorithm's inability to accurately predict the future. We define the Bregman divergence between two vectors induced by a differentiable function R as follows:

$$\mathcal{D}_R(w, u) = R(w) - \Big(R(u) + \langle \nabla R(u), w - u\rangle\Big).$$

Next theorem is used to bound the forward regret.

Theorem A.1 (Theorem 3 in Joulani et al. [2017])**.** For any $f^* \in \mathcal{F}$ and any sequence of linear losses, the forward regret satisfies the following inequality:

$$R_T^+ \le \sum_{t=1}^{T}\Big(h_t(f^*) - h_t(f_{t+1})\Big) - \sum_{t=1}^{T}\mathcal{D}_{h_{1:t}}(f_{t+1}, f_t).$$

Recall that $h_{1:t}(f)$ is η_t-strongly convex. From the definitions of strong convexity and Bregman divergence, we have

$$\sum_{t=1}^{T}\mathcal{D}_{h_{1:t}}(f_{t+1}, f_t) \ge \sum_{t=1}^{T}\frac{\eta_t}{2}\|f_{t+1} - f_t\|^2. \tag{A.3}$$

Applying Theorem A.1 and Eq. (A.3), we have

$$R_T^+ \le \sum_{t=1}^{T}\Big(h_t(f^*) - h_t(f_{t+1})\Big) - \sum_{t=1}^{T}\frac{\eta_t}{2}\|f_{t+1} - f_t\|^2. \tag{A.4}$$

To bound the first term in Eq. (A.4), we expand it by the definition of Eq. (A.1),

$$\begin{aligned}
&\sum_{t=1}^{T}\Big(h_t(f^*) - h_t(f_{t+1})\Big)\\
=\;& \sum_{t=1}^{T}\langle M_{t+1} - M_t, f^* - f_{t+1}\rangle + \sum_{t=1}^{T}(\eta_t - \eta_{t-1})(\mathcal{R}(f^*) - \mathcal{R}(f_{t+1}))\\
\le\;& \sum_{t=1}^{T}\langle M_{t+1} - M_t, f^*\rangle - \sum_{t=1}^{T}\langle M_{t+1} - M_t, f_{t+1}\rangle + \eta_T\mathcal{R}(f^*)\\
=\;& \langle M_{T+1}, f^*\rangle - \sum_{t=1}^{T}\langle M_{t+1} - M_t, f_{t+1}\rangle + \eta_T\mathcal{R}(f^*),
\end{aligned} \tag{A.5}$$

where the first inequality is due to the fact that η_t is non-decreasing and $\eta_0 = 0$. We decompose the second term in Eq. (A.5) as follows,

$$\begin{aligned}\sum_{t=1}^{T} \langle M_{t+1} - M_t, f_{t+1} \rangle &= \sum_{t=2}^{T+1} \langle M_t, f_t \rangle - \sum_{t=1}^{T} \langle M_t, f_{t+1} \rangle \\ &= \sum_{t=1}^{T} \langle M_t, f_t \rangle - \sum_{t=1}^{T} \langle M_t, f_{t+1} \rangle + \langle M_{T+1}, f_{T+1} \rangle, \end{aligned} \tag{A.6}$$

since $M_1 = 0$. Combining Eq. (A.5) and Eq. (A.6) together,

$$\sum_{t=1}^{T} \Big(h_t(f^*) - h_t(f_{t+1}) \Big) = -\sum_{t=1}^{T} \langle M_t, f_t - f_{t+1} \rangle + \langle M_{T+1}, f^* - f_{T+1} \rangle + \eta_T \mathcal{R}(f^*). \tag{A.7}$$

Putting Eq. (A.2), Eq. (A.4) and Eq. (A.7) together, we reach

$$R_T \leq \sum_{t=1}^{T} \langle q_t - M_t, f_t - f_{t+1} \rangle - \sum_{t=1}^{T} \frac{\eta_t}{2} \|f_t - f_{t+1}\|^2 + \langle M_{T+1}, f^* - f_{T+1} \rangle + \eta_T \mathcal{R}(f^*). \tag{A.8}$$

To bound the first term in Eq. (A.8), we first use Hölder's inequality such that

$$\begin{aligned}\langle q_t - M_t, f_t - f_{t+1} \rangle &= \frac{2}{\eta_t} (q_t - M_t)^\top \frac{\eta_t}{2} (f_t - f_{t+1}) \\ &\leq \| \frac{2}{\eta_t} (q_t - M_t) \|_* \| \frac{\eta_t}{2} (f_t - f_{t+1}) \| \\ &\leq \frac{1}{\eta_t} \|q_t - M_t\|_*^2 + \frac{\eta_t}{4} \|f_t - f_{t+1}\|^2, \end{aligned}$$

where the last inequality is due to $2ab \leq a^2 + b^2$. Thus we have

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} \frac{1}{\eta_t} \|q_t - M_t\|_*^2 + \sum_{t=1}^{T} \frac{\eta_t}{4} \|f_t - f_{t+1}\|^2 - \sum_{t=1}^{T} \frac{\eta_t}{2} \|f_t - f_{t+1}\|^2 \\ &\quad + \langle M_{T+1}, f^* - f_{T+1} \rangle + \eta_T \mathcal{R}(f^*) \\ &= \sum_{t=1}^{T} \frac{1}{\eta_t} \|q_t - M_t\|_*^2 - \sum_{t=1}^{T} \frac{\eta_t}{4} \|f_t - f_{t+1}\|^2 + \langle M_{T+1}, f^* - f_{T+1} \rangle + \eta_T \mathcal{R}(f^*). \end{aligned}$$

By choosing $\eta_t = \eta \sqrt{\sum_{s=1}^{t} \|q_s - M_s\|_*^2}$ for some absolute constant η, we have

$$\begin{aligned} R_T \leq \sum_{t=1}^{T} \frac{\|q_t - M_t\|_*^2}{\eta \sqrt{\sum_{s=1}^{t} \|q_t - M_t\|_*^2}} + \eta \sqrt{\sum_{t=1}^{T} \|q_t - M_t\|_*^2 \mathcal{R}(f^*)} \\ - \sum_{t=1}^{T} \frac{\eta_t}{4} \|f_t - f_{t+1}\|^2 + \langle M_{T+1}, f^* - f_{T+1} \rangle. \end{aligned} \tag{A.9}$$

Lemma 4 in McMahan [2017] states that for any non-negative real numbers $a_1, \ldots, a_T$,

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}} \leq 2 \sqrt{\sum_{t=1}^{T} a_t}.$$

Applying this inequality to the first term in Eq. (A.9) with $a_t = \|q_t - M_t\|_*^2$, we have

$$R_T \leq \left(\frac{2}{\eta} + \eta \mathcal{R}(f^*) \right) \sqrt{\sum_{t=1}^{T} \|q_t - M_t\|_*^2} - \sum_{t=1}^{T} \frac{\eta_t}{4} \|f_t - f_{t+1}\|^2 + \langle M_{T+1}, f^* - f_{T+1} \rangle.$$

Letting $\eta = \sqrt{2/\mathcal{R}(f^*)}$ and $R_{\max} = \max_f \mathcal{R}(f)$, this concludes the proof. ■

A.3 Proof of Lemma 5.7: online learning reduction

Step 1. We utilize Lemma 5.4 for each individual state x. Recall that

$$\begin{aligned} R_{2T} &= \tau\sum_{k=1}^K\Big\langle \mu_{\pi^*}, \widehat{Q}_{\pi_k}(\cdot,\pi^*)-\widehat{Q}_{\pi_k}(\cdot,\pi_k)\Big\rangle \\ &= \tau\sum_{x\in\mathcal{X}}\mu_{\pi^*}(x)\sum_{k=1}^K\Big\langle \pi^*(\cdot|x)-\pi_k(\cdot|x), \widehat{Q}_{\pi_k}(x,\cdot)\Big\rangle. \end{aligned}$$

Applying Lemma 5.4 with $f_k=\pi_k(\cdot|x)$, $q_k=\widehat{Q}_{\pi_k}(x,\cdot)$ and $M_k=\widehat{Q}_{\pi_{k-1}}(x,\cdot)$, we have

$$\begin{aligned} R_{2T}\leq \tau\sum_{x\in\mathcal{X}}\mu_{\pi^*}(x)\Big(&\sqrt{2R_{\max}}\sqrt{\sum_{k=1}^K\big\|\widehat{Q}_{\pi_k}(x,\cdot)-\widehat{Q}_{\pi_{k-1}}(x,\cdot)\big\|_\infty^2} \\ &-\sum_{k=1}^K\frac{\eta_k(x)}{4}\big\|\pi_k(\cdot|x)-\pi_{k+1}(\cdot|x)\big\|_1^2+2(b+Q_{\max})\Big), \end{aligned} \tag{A.10}$$

since $\widehat{Q}_{\pi_K}(x,a)\in[b,b+Q_{\max}]$ from Condition 4.1. Here, $\eta_k(x)=\eta\sqrt{\sum_{s=1}^k\|\widehat{Q}_{\pi_s}(x,\cdot)-\widehat{Q}_{\pi_{s-1}}(x,\cdot)\|_\infty^2}$.

Step 2. It remains to bound the cumulative change of estimated Q-values in Eq. (A.10). We first decompose it by substrating the true Q-function and using the triangle inequality and $2ab\leq a^2+b^2$:

$$\begin{aligned} &\sum_{k=1}^K\big\|\widehat{Q}_{\pi_k}(x,\cdot)-\widehat{Q}_{\pi_{k-1}}(x,\cdot)\big\|_\infty^2 \\ \leq&\sum_{k=1}^K 2\big\|\widehat{Q}_{\pi_k}(x,\cdot)-Q_{\pi_k}(x,\cdot)\big\|_\infty^2+\sum_{k=1}^K 2\big\|Q_{\pi_{k-1}}(x,\cdot)-\widehat{Q}_{\pi_{k-1}}(x,\cdot)\big\|_\infty^2 \\ &+\sum_{k=1}^K 2\big\|Q_{\pi_k}(x,\cdot)-Q_{\pi_{k-1}}(x,\cdot)\big\|_\infty^2. \end{aligned} \tag{A.11}$$

The first two terms in Eq. (A.11) measure the estimation error. By Condition 4.1, we have,

$$\big\|\widehat{Q}_{\pi_k}(x,\cdot)-Q_{\pi_k}(x,\cdot)\big\|_\infty^2\leq\frac{2\widetilde{C}^2\log(1/\delta)}{\tau}+2\varepsilon_0^2, \tag{A.12}$$

with probability at least $1-\delta$ for each $k\in[K]$ and for problem-dependent constants $\widetilde{C}$. Putting Eq. (A.11), Eq. (A.12) and Lemma 5.6 together, the following holds with probability $1-K\delta$,

$$\begin{aligned} &\sum_{k=1}^K\big\|\widehat{Q}_{\pi_k}(x,\cdot)-\widehat{Q}_{\pi_{k-1}}(x,\cdot)\big\|_\infty^2 \\ \leq\;& \frac{8\widetilde{C}^2K\log(1/\delta)}{\tau}+8K\varepsilon_0^2 \\ &+2t_{\text{mix}}^4\log_2^4(K)\sum_{k=1}^K\max_x\big\|\pi_k(\cdot|x)-\pi_{k-1}(\cdot|x)\big\|_1^2+\frac{4K}{K^6}. \end{aligned} \tag{A.13}$$

Step 3. Finally, by our choice of the data-dependent learning rate $\eta_k(x)$, we are able to cancel out the positive term in Eq. (A.10) such that the regret is greatly sharpened. For notation simplicity, we denote $d_k(x)=\|\pi_k(\cdot|x)-\pi_{k-1}(\cdot|x)\|_1$. Putting Eq. (A.10) and Eq. (A.13) together, with a union bound, we have

$$\begin{aligned} \frac{R_{2T}}{\tau}\leq C_1\sum_{x\in\mathcal{X}}\mu_{\pi^*}(x)\Big(&\sqrt{R_{\max}}\sqrt{t_{\text{mix}}^4\log_2^4(K)\sum_{k=1}^K\max_x d_k^2(x)+\frac{\widetilde{C}^2K\log(K\mu_{\pi^*}(x)\delta^{-1})}{\tau}+K\varepsilon_0^2} \\ &-\sum_{k=1}^K\frac{\eta_k(x)}{4}d_{k+1}^2(x)+2(b+Q_{\max})\Big) \end{aligned} \tag{A.14}$$

holds with probability at least $1-\delta$. Assuming $\eta_0(x)=\eta_1(x)$, we have

$$\sum_{k=1}^{K}\frac{\eta_k(x)}{4}d_{k+1}^2(x)\geq\sum_{k=1}^{K}\frac{\eta_{k-1}(x)}{4}d_k^2(x).$$

Moreover, we denote $\mu_{\min}^*=\min_{x:\mu_{\pi^*}(x)>0}\mu_{\pi^*}(x)$ and

$$\begin{aligned}
g_1 &= R_{\max}t_{\text{mix}}^4\log_2^4(K)\\
g_2 &= \widetilde{C}^2R_{\max}K\log(K\mu_{\min}^*\delta)^{-1}/\tau+K\varepsilon_0^2R_{\max}\\
g_3 &= 2(b+Q_{\max}).
\end{aligned}$$

Then we simplify Eq. (A.14) as

$$\begin{aligned}
\frac{R_{2T}}{\tau} &\leq \sum_{x\in\mathcal{X}}\mu_{\pi^*}(x)\Big(\sqrt{g_1\sum_{k=1}^{K}\max_x d_k^2(x)+g_2}-\sum_{k=1}^{K}\frac{\eta_{k-1}(x)}{4}d_k^2(x)+g_3\Big)\\
&= \sqrt{g_1\sum_{k=1}^{K}\max_x d_k^2(x)+g_2}-\sum_{x\in\mathcal{X}}\mu_{\pi^*}(x)\sum_{k=1}^{K}\frac{\eta_{k-1}(x)}{4}d_k^2(x)+g_3.
\end{aligned}$$

Let us denote $x_k^*=\text{argmax}_x\,d_k^2(x)$. Noting that

$$\sum_{x\in\mathcal{X}}\mu_{\pi^*}(x)\sum_{k=1}^{K}\frac{\eta_{k-1}(x)}{4}d_k^2(x)\geq\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*),\tag{A.15}$$

we have

$$\begin{aligned}
\frac{R_{2T}}{\tau} &\leq \sqrt{g_1\sum_{k=1}^{K}d_k^2(x_k^*)+g_2}-\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+g_3\\
&= \sqrt{g_1\sum_{k=1}^{K}\frac{\mu_{\pi^*}(x_k^*)}{\mu_{\pi^*}(x_k^*)}d_k^2(x_k^*)+g_2}-\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+g_3\\
&\leq \sqrt{\frac{4g_1}{\eta_1(x_k^*)\mu_{\min}^*}\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+g_2}-\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+g_3\\
&= 2\sqrt{\frac{g_1}{\eta_1(x_k^*)\mu_{\min}^*}\Big(\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+\frac{\mu_{\min}^*\eta_1g_2}{4g_1}\Big)}\\
&\quad -\Big(\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+\frac{\mu_{\min}^*\eta_1(x_k^*)g_2}{4g_1}\Big)+\frac{\mu_{\min}^*\eta_1(x_k^*)g_2}{4g_1}+g_3,
\end{aligned}$$

where the second inequality we use the fact that η_k is monotone increasing. Letting

$$a=\frac{g_1}{\eta_1(x_k^*)\mu_{\min}^*},\,b=\sum_{k=1}^{K}\mu_{\pi^*}(x_k^*)\frac{\eta_{k-1}(x_k^*)}{4}d_k^2(x_k^*)+\frac{\mu_{\min}^*\eta_1(x_k^*)g_2}{4g_1},$$

and using the fact that $2\sqrt{ab}-b\leq a$, we reach

$$\frac{R_{2T}}{\tau}\leq\frac{g_1}{\eta_1(x_k^*)\mu_{\min}^*}+\frac{\mu_{\min}^*\eta_1(x_k^*)g_2}{4g_1}+g_3.\tag{A.16}$$

Plugging in back the definition of g_1,g_2,g_2, we have with probability at least $1-\delta$,

$$R_{2T}\leq\frac{R_{\max}t_{\text{mix}}^4\log_2^4(K)\tau}{\eta_1(x_k^*)\mu_{\min}^*}+\frac{\eta_1(x_k^*)(\widetilde{C}^2K\log(K\mu_{\min}^*\delta)^{-1}+T\varepsilon_0^2)}{4t_{\text{mix}}^4\log_2^4(K)}+2(b+Q_{\max}).\tag{A.17}$$

By definition, $\eta_1(x_k^*) = \sqrt{2R_{\max}\|\widehat{Q}_1(x_k^*,\cdot)\|_\infty}$. Since $\widehat{Q}_1(x,a) \in [b, b+Q_{\max}]$ from Condition 4.1, we have $\eta_1(x_k^*)$ is lower and upper bounded by some constant. Denote $\rho = \max_\mu \max_{x:\mu(x)\neq 0}(1/\mu_\pi(x))$. Based on this, we simplify the upper bound (A.17) as

$$R_{2T} \lesssim \tau t_{\text{mix}}^4 \rho \log_2^4(K) + \widetilde{C}^2 K \log(K/\delta) + T\varepsilon_0^2,$$

where $\lesssim$ hides constant factors. This ends the proof. ■

A.4 Proof of Lemma 5.6: relative Q-function error

We first introduce a lemma that illustrates the true Q-value can be bounded by the mixing time.

Lemma A.2 (Lemma 3 in Neu et al. [2014])**.** For any policy π and any state-action pair $(x,a) \in \mathcal{X}\times\mathcal{A}$, for any reward function $r \in [0,1]$, we have

$$|Q_\pi(x,a)| \le 2t_{\text{mix}} + 3. \tag{A.18}$$

From the Bellman equation Eq. (2.2),

$$Q_{\pi_k}(x,a) - Q_{\pi_{k-1}}(x,a) = \sum_{x'} \mathcal{P}(x'|x,a)\Big(V_{\pi_k}(x') - V_{\pi_{k-1}}(x')\Big) + \lambda_{\pi_{k-1}} - \lambda_{\pi_k}. \tag{A.19}$$

We first bound $\lambda_{\pi_{k-1}} - \lambda_{\pi_k}$. By Lemma C.1 (performance difference lemma),

$$\lambda_{\pi_{k-1}} - \lambda_{\pi_k} = \sum_x \mu_{\pi_{k-1}}(x)\Big(\sum_a (\pi_{k-1}(a|x) - \pi_k(a|x))\Big)Q_{\pi_k}(x,a).$$

By Lemma A.2, it implies

$$\lambda_{\pi_{k-1}} - \lambda_{\pi_k} \le (2t_{\text{mix}}+3)\max_x \big\|\pi_{k-1}(\cdot|x) - \pi_k(\cdot|x)\big\|_1. \tag{A.20}$$

Next we bound $V_{\pi_k}(x) - V_{\pi_{k-1}}(x)$. In an ergodic MDP, the expected average reward λ_π can be written as $\lambda_\pi = \mu_\pi^\top r_\pi$, where $r_\pi(x) = \sum_a \pi(a|x) r(x,a)$. Let $\boldsymbol{e}_x$ be an indicator vector for state x. For all π,

$$\begin{aligned} V_\pi(x) &= \sum_{t=0}^{\infty}\Big(\boldsymbol{e}_x^\top(\mathcal{P}^\pi)^t - \mu_\pi^\top\Big)r_\pi \\ &= \sum_{t=0}^{N-1}\Big(\boldsymbol{e}_x^\top(\mathcal{P}^\pi)^t - \mu_\pi^\top\Big)r_\pi + \sum_{t=N}^{\infty}\Big(\boldsymbol{e}_x^\top(\mathcal{P}^\pi)^t - \mu_\pi^\top\Big)r_\pi, \end{aligned} \tag{A.21}$$

Corollary 13.2 of Wei et al. [2019] shows that for an ergodic MDP with mixing time t_{mix} and $N = \lceil 4t_{\text{mix}}\log_2(K)\rceil$, for all π,

$$\sum_{t=N}^{\infty}\Big\|\boldsymbol{e}_x^\top(\mathcal{P}^\pi)^t - \mu_\pi\Big\|_1 \le \sum_{t=N}^{\infty} 2^{1-\frac{t}{t_{\text{mix}}}} = \frac{2^{1-\frac{N}{t_{\text{mix}}}}}{1-2^{-\frac{1}{t_{\text{mix}}}}} \le \frac{2t_{\text{mix}}}{\ln 2}2^{1-\frac{N}{t_{\text{mix}}}} = \frac{2t_{\text{mix}}}{\ln 2}\frac{2}{K^4} \le \frac{1}{K^3}.$$

Thus, the second term in Eq. (A.21) can be bounded by

$$\Big|\sum_{t=N}^{\infty}\Big(\boldsymbol{e}_x^\top(\mathcal{P}^\pi)^t - \mu_\pi^\top\Big)r_\pi\Big| \le \sum_{t=N}^{\infty}\Big\|\boldsymbol{e}_x^\top(\mathcal{P}^\pi)^t - \mu_\pi\Big\|_1 \le \frac{1}{K^3}.$$

The following steps are similar to the proof of Lemma 7 in Wei et al. [2019]. For the sake of completeness, we present a full proof here. The difference between $V_{\pi_k}(x)$ and $V_{\pi_{k-1}}(x)$ can be bounded by

$$\begin{aligned} &\Big|V_{\pi_k}(x) - V_{\pi_{k-1}}(x)\Big| \\ =\ &\Big|\sum_{t=0}^{N-1}\boldsymbol{e}_x^\top\Big((\mathcal{P}^{\pi_k})^t - (\mathcal{P}^{\pi_{k-1}})^t\Big)r_{\pi_k} + \sum_{t=0}^{N-1}\boldsymbol{e}_x^\top(\mathcal{P}^{\pi_k})^t(r_{\pi_k} - r_{\pi_{k-1}}) - N\lambda_{\pi_k} + N\lambda_{\pi_{k-1}}\Big| + \frac{2}{K^3} \\ \le\ &\sum_{t=0}^{N-1}\Big\|\Big((\mathcal{P}^{\pi_k})^t - (\mathcal{P}^{\pi_{k-1}})^t\Big)r_{\pi_k}\Big\|_\infty + \sum_{t=0}^{N-1}\|r_{\pi_k} - r_{\pi_{k-1}}\|_\infty + N|\lambda_{\pi_k} - \lambda_{\pi_{k-1}}| + \frac{2}{K^3}. \end{aligned} \tag{A.22}$$

Next, we will derive a recursive form for the first term as follows:

$$\begin{aligned}
& \left\|\left((\mathcal{P}^{\pi_k})^t - (\mathcal{P}^{\pi_{k-1}})^t\right) r_{\pi_k}\right\|_\infty \\
= \; & \left\|\left(\mathcal{P}^{\pi_k}(\mathcal{P}^{\pi_k})^{t-1} - \mathcal{P}^{\pi_k}(\mathcal{P}^{\pi_{k-1}})^{t-1} + \mathcal{P}^{\pi_k}(\mathcal{P}^{\pi_{k-1}})^{t-1} - \mathcal{P}^{\pi_{k-1}}(\mathcal{P}^{\pi_{k-1}})^{t-1}\right) r_{\pi_k}\right\|_\infty \\
\leq \; & \left\|\mathcal{P}^{\pi_k}\left((\mathcal{P}^{\pi_k})^{t-1} - (\mathcal{P}^{\pi_{k-1}})^{t-1}\right) r_{\pi_k}\right\|_\infty + \left\|(\mathcal{P}^{\pi_k} - \mathcal{P}^{\pi_{k-1}})(\mathcal{P}^{\pi_{k-1}})^{t-1} r_{\pi_k}\right\|_\infty \\
\leq \; & \left\|\left((\mathcal{P}^{\pi_k})^{t-1} - (\mathcal{P}^{\pi_{k-1}})^{t-1}\right) r_{\pi_k}\right\|_\infty + \max_x \left\|\boldsymbol{e}_x^\top (\mathcal{P}^{\pi_k} - \mathcal{P}^{\pi_{k-1}})(\mathcal{P}^{\pi_{k-1}})^{t-1}\right\|_1 \\
\leq \; & \left\|\left((\mathcal{P}^{\pi_k})^{t-1} - (\mathcal{P}^{\pi_{k-1}})^{t-1}\right) r_{\pi_k}\right\|_\infty + \max_x \left\|\boldsymbol{e}_x^\top (\mathcal{P}^{\pi_k} - \mathcal{P}^{\pi_{k-1}})\right\|_1 \\
\leq \; & \left\|\left((\mathcal{P}^{\pi_k})^{t-1} - (\mathcal{P}^{\pi_{k-1}})^{t-1}\right) r_{\pi_k}\right\|_\infty + \max_x \left(\sum_{x'} \left|\sum_a \left(\pi_k(a|x) - \pi_{k-1}(a|x)\right) \mathcal{P}(x'|x,a)\right|\right) \\
\leq \; & \left\|\left((\mathcal{P}^{\pi_k})^{t-1} - (\mathcal{P}^{\pi_{k-1}})^{t-1}\right) r_{\pi_k}\right\|_\infty + \max_x \left\|\pi_k(a|x) - \pi_{k-1}(a|x)\right\|_1 .
\end{aligned}$$

By induction, it holds that

$$\left\|\left((\mathcal{P}^{\pi_k})^t - (\mathcal{P}^{\pi_{k-1}})^t\right) r_{\pi_k}\right\|_\infty \leq t \max_x \left\|\pi_k(a|x) - \pi_{k-1}(a|x)\right\|_1 .$$

Thus,

$$\sum_{t=0}^{N-1} \left\|\left((\mathcal{P}^{\pi_k})^t - (\mathcal{P}^{\pi_{k-1}})^t\right) r_{\pi_k}\right\|_\infty \leq N^2 \max_x \left\|\pi_k(a|x) - \pi_{k-1}(a|x)\right\|_1 . \tag{A.23}$$

In addition,

$$\sum_{t=0}^{N-1} \|r_{\pi_k} - r_{\pi_{k-1}}\|_\infty \leq N \max_x \left\|\pi_k(a|x) - \pi_{k-1}(a|x)\right\|_1 . \tag{A.24}$$

Plugging Eq. (A.20), Eq. (A.23) and Eq. (A.24) into Eq. (A.22) yields

$$\left|V_{\pi_k}(x) - V_{\pi_{k-1}}(x)\right| \leq \left(N^2 + N + (2t_{\text{mix}} + 3)N\right) \max_x \left\|\pi_k(a|x) - \pi_{k-1}(a|x)\right\|_1 + \frac{2}{K^3}, \tag{A.25}$$

where $N = \lceil 4t_{\text{mix}} \log_2(K) \rceil$. Together with Eq. (A.19), we reach the result. ■

B Linear value function approximation

In this section, we show that with linear value function approximation and under similar assumptions as in Abbasi-Yadkori et al. [2019a], the estimation error in each state can be bounded in the ℓ_∞ norm. Note that we consider an unrealizable case such that Q-function could be approximated linear represented with an irreducible approximation error ε_0. This is in contrast of many existing works [Yang and Wang, 2019a,b, Jin et al., 2019] who consider realizable cases.

Suppose $\phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$ is a feature map chosen by the user. Consider $\widehat{Q}_{\pi_k}(x,a) = \phi(x,a)^\top \widehat{w}_k$ be the linear value function estimate where $\widehat{w}_k$ is the estimated weight vector. Let Ψ be a $|\mathcal{X}||\mathcal{A}| \times d$ feature matrix whose rows correspond to state-action feature vectors. We make the regularity assumption on Ψ and assume that for all policies π, the following feature excitation condition holds.

Assumption B.1 (Linearly independent features)**.** The columns of the matrix $[\Psi, \mathbf{1}]$ are linearly independent.

Assumption B.2 (Uniformly excited features, Assumption A4 in Abbasi-Yadkori et al. [2019a])**.** There exists a positive real σ such that for any policy π, $\lambda_{\text{min}}(\Psi^\top \text{diag}(\mu_\pi \otimes \pi)\Psi) \geq \sigma$.

Furthermore, we assume that the following error bound holds.

Assumption B.3 (Estimation error in $\mu_\pi \otimes \pi$-norm)**.** For all $k \in [K]$, with probability at least $1 - \delta$, the value error is bounded in the $\mu_\pi \otimes \pi$-norm.

$$\left\|\widehat{Q}_{\pi_k} - Q_{\pi_k}\right\|_{\mu_\pi \otimes \pi} \leq C_2 \sqrt{\frac{\log(1/\delta)}{\tau}} + \varepsilon_0 \, ,$$

where C_2 is a problem-dependent constant and ε_0 is the irreducible approximation error.

The above error Assumption B.3 can be satisfied, for example, by the LSPE algorithm of Bertsekas and Ioffe [1996], as shown in Theorem 5.1 in Abbasi-Yadkori et al. [2019a]. The same authors show that Assumptions B.1, B.2 and B.3 also suffice to bound the error in $\mu^* \otimes \pi_k$ and $\mu^* \otimes \pi^*$-norms, as required by our Lemma 5.3. Here we additionally prove that under same assumptions, the error in each state is bounded in the ℓ_∞-norm, as required by Lemma B.4.

Lemma B.4 (Estimation error in ℓ_∞-norm)**.** Under Assumptions B.2 and B.3, the following holds with probability at least $1-\delta$,

$$\big\|\widehat{Q}_{\pi_k}(x,\cdot) - Q_{\pi_k}(x,\cdot)\big\|_\infty \le C_\psi\Big(C_2\sqrt{\sigma\frac{\log(1/\delta)}{\tau}} + \varepsilon_0\Big),$$

where $C_\Psi = \max_{x,a}\|\phi(x,a)\|_2$.

Proof. Note that under Assumption B.2, $||\Psi(\widehat{w}_k - w_k)||^2_{\mu_\pi\otimes\pi} \ge \sigma||\widehat{w}_k - w_k||_2^2$. We have the following:

$$\begin{aligned}\|\widehat{Q}_{\pi_k}(x,\cdot) - Q_{\pi_k}(x,\cdot)\|_\infty &= \max_a|\phi(x,a)^\top(\widehat{w}_k - w_k)| \\ &\le C_\Psi||\widehat{w}_k - w_k||_2 \\ &\le C_\Psi||\Psi(\widehat{w}_k - w_k)||_{\mu_\pi\otimes\pi}/\sqrt{\sigma} \\ &= C_\Psi\|\widehat{Q}_{\pi_k} - Q_{\pi_k}\|_{\mu_\pi\otimes\pi}/\sqrt{\sigma} \\ &\le C_\Psi C_2\sqrt{\log(1/\delta)/(\sigma\tau)} + C_\Psi/\sqrt{\sigma}\varepsilon_0\,.\end{aligned}$$

■

C Supporting lemmas

Lemma C.1 (Performance difference lemma)**.** Consider an MDP specified by the transition probability kernel $\mathcal{P}$ and reward function r. For any policy $\pi, \widehat{\pi}$, it holds that

$$\lambda_\pi - \lambda_{\widehat{\pi}} = \sum_{x,a}\mu_\pi(x)(\pi(a|x) - \widehat{\pi}(a|x))Q_{\widehat{\pi}}(x,a),$$

where $\mu_\pi(x)$ is the stationary distribution of a policy π.

Proof. Based on average reward Bellman equation, we have

$$\begin{aligned}\sum_{x,a}\mu_\pi(x)\pi(a|x)Q_{\widehat{\pi}}(x,a) &= \sum_{x,a}\mu_\pi(x)\pi(a|x)\Big[r(x,a) - \lambda_{\widehat{\pi}} + \sum_{x'}\mathcal{P}(x'|x,a)V_{\widehat{\pi}}(x')\Big] \\ &= \lambda_\pi - \lambda_{\widehat{\pi}} + \sum_x \mu_\pi(x)V_{\widehat{\pi}}(x),\end{aligned}$$

where the second equation is due to $\sum_{x,a}\mu_\pi(x)\pi(a|x)\mathcal{P}(x'|x,a) = \mu_\pi(x')$. Therefore,

$$\begin{aligned}\lambda_\pi - \lambda_{\widehat{\pi}} &= \sum_{x,a}\mu_\pi(x)\pi(a|x)Q_{\widehat{\pi}}(x,a) - \sum_x\mu_\pi(x)V_{\widehat{\pi}}(x) \\ &= \sum_{x,a}\mu_\pi(x)\Big(\pi(a|x)Q_{\widehat{\pi}}(x,a) - \widehat{\pi}(a|x)Q_{\widehat{\pi}}(x,a)\Big).\end{aligned}$$

This ends the proof. ■

Lemma C.2 (Lemma 4 in McMahan [2017])**.** For any non-negative real numbers $a_1, \ldots, a_T$, the following holds

$$\sum_{t=1}^T\frac{a_t}{\sqrt{\sum_{s=1}^t a_s}} \le 2\sqrt{\sum_{t=1}^T a_t}.$$